IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management pour la validation du cursus à la Formation d'Actuaire de l'Institut du Risk Management et l'admission à l'Institut des actuaires le

Par : Olivier Danneaux

Titre : Pricing segmentation of a travel insurance heterogeneous portfolio

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1an ☒ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : Europ Assistance Group

Nom : Etienne Bonnet

Signature et Cachet :

europ assistance Holding
S.A.S. au capital de 20 270 528,00 €
Siège Social : 2 rue Pillet-Will
75309 PARIS CEDEX 9
01 58 34 23 00
SIRET 632 016 382 00178

Directeur de mémoire en entreprise :

Nom : Matthieu Quilfen

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature(s) du candidat(s)

Pricing segmentation of a travel insurance heterogeneous portfolio

Résumé

Mots clés : Assurance Non-Vie, Assurance Voyage, Couverture Annulation, B2B2C, Segmentation tarifaire, portefeuille hétérogène, MLG[1].

Historiquement, les sociétés d'assurance Voyage en Europe (la situation est différente aux Etats Unis) ont peu développé des modèles de tarification segmentés car peu de données étaient disponibles (notamment à la suite de la vente intermédiée de ces produits) et aussi car le montant moyen de primes est très faible par rapport à d'autres produits d'assurance. Néanmoins, ces dernières années, certains modèles ont commencé à être développés sur des périmètres restreints et ne s'appliquant pas à la totalité des portefeuilles B2B2C assurés étant donné la diversité des portefeuilles d'Assurance Voyage.

L'objet du présent mémoire est donc d'élargir l'usage de la tarification segmentée sur la couverture Annulation (qui représente la majorité des primes d'assurance voyage) en adressant les problèmes soulevés par l'hétérogénéité des portefeuilles à modéliser. Afin d'atteindre cet objectif, deux principales problématiques sont adressées dans ce mémoire.

La première problématique adressée est de tester différentes méthodologies de modélisation de la fréquence. En effet, les quelques modèles développés jusque-là utilise la loi de Poisson comme loi de modélisation alors que la nature de la couverture Annulation tend à faire penser que la loi Binomiale serait plus appropriée.

La deuxième problématique adressée sera de définir, sur base des données disponibles, dans quelle mesure des modèles de tarification plus globaux (moins spécifiques à un périmètre donné) peuvent être construits sans diminuer trop la qualité de la prédiction du modèle. A cette fin, nous construirons différents MLGs à des niveaux de granularités différents (par partenaire, par segment de marché et global), comparerons les mesures appropriées de performance des modèles obtenus et définirons ainsi à quel niveau de granularité les modèles de tarification doivent être construits.

[1] Modèle Linéaire Généralisé

Abstract

Keywords : Non-Life insurance, Travel insurance, B2B2C, pricing segmentation, heterogeneous portfolio, GLM[2]

Historically, Travel Insurance companies in Europe (the situation is different in the United States) have merely developed segmented pricing models as little data were made available for the Insurance Company (notably because of the intermediated sales process of Travel Insurance products) and as well, because of the low average premium amount per policy of Travel Insurance products. Nevertheless, over the last years, some models started to be developed on restricted perimeters without allowing to be applied across a whole portfolio given the heterogeneity of the Travel Insurance portfolios.

The purpose of this thesis is thus to extend the use of segmented pricing models on the Cancellation cover (that represents most of the Travel Insurance Premiums) by addressing the issues raised by the heterogeneity of the portfolio to be modeled. In order to achieve this objective, two main issues will be addressed in this thesis.

First, different modelling methodologies will be tested for the frequency models. Indeed, the GLM models developed so far were using a Poisson distribution whereas the nature of the Travel Cancellation cover would make us think that the binomial distribution would be best suited.

Secondly, based on data available, we will see to what extent we are able to build more global pricing segmentation models without loosing too much accuracy of the models. In that respect, we will build different GLMs at different levels of granularity (per partner, per segment, global), compare their performance based on appropriate performance metrics and define the appropriate level of granularity at which the models should be built moving forward.

[2] Generalized Linear Model

Note de synthèse

Mots clés : Assurance Non-Vie, Assurance Voyage, Couverture Annulation, B2B2C, Segmentation tarifaire, portefeuille hétérogène, MLG[3].

L'objectif de cette thèse est d'étendre l'utilisation des modèles de tarification segmentée pour les produits d'assurance voyage pour lesquels leur utilisation est moins étendue que pour d'autres produits d'assurance. L'enjeu qu'il faudra adresser dans cette étude est l'hétérogénéité car le portefeuille qui sera analysé est hétérogène en termes de produits vendus, de risques sous-jacents (différents types de voyages assurés) et de canaux de distribution (différents types d'acteurs du tourisme).

Le marché de l'assurance voyage en Europe

Les produits d'assurance voyage en Europe sont principalement vendus par les acteurs du tourisme, tels que les compagnies aériennes, les agences de voyage en ligne (OTA), les croisiéristes, … On dit que ces produits d'assurance voyage sont vendus en « B2B2C ».

La plupart du temps, ces produits d'assurance ne couvrent uniquement que le voyage qui a été acheté par l'assuré via le site Internet ou le canal de distribution de l'acteur du tourisme. Nous appelons ce type de produits d'assurance des polices « court terme » car l'exposition ne s'étend que du moment de la souscription jusqu'à la date de départ (dans le cas de la couverture annulation) et entre la date de départ et la date de retour pour les autres couvertures d'assurance voyage (comme l'assistance médicale).

En Europe, contrairement à ce qui se passe aux États-Unis, les modèles linéaires généralisés ne sont pas largement utilisés pour tarifer les couvertures d'assurance voyage. Cela peut s'expliquer par différents éléments :

- La vente de ces produits est majoritairement intermédiée, ce qui a un impact négatif sur la disponibilité des données permettant à l'assureur de tarifer les couvertures d'assurance
- La prime moyenne d'un produit d'assurance voyage est relativement faible par rapport aux autres produits d'assurance et cela diminue l'intérêt d'avoir des MLGs en place
- Les partenaires à travers lesquels les produits d'assurance voyage sont vendus, ont peu d'appétit pour mettre en œuvre sur leurs canaux de vente des modèles de tarification complexes
- Il existe généralement une grande hétérogénéité du portefeuille assuré, en termes de produits vendus, de voyages assurés (billets d'avion vs forfaits, voyages low-cost vs premium, domestique vs international, …)

Néanmoins, les premier et troisième points mentionnés ci-dessus évoluent et l'utilisation des MLGs sur les produits d'assurance voyage augmente progressivement.

La couverture la plus importante en termes de montant de prime est la couverture annulation et cette étude se concentrera donc uniquement sur cette couverture.

[3] Modèle Linéaire Généralisé

Collecte de données, préparation et validation

Dans ce contexte et dans le cadre de ce mémoire, beaucoup d'efforts ont été consacrés à la construction d'une base de données appropriée pour mener cette analyse. En effet, dans le passé, la plupart des MLGs construits pour les produits d'assurance voyage étaient construits sur des périmètres restreints (généralement limités à un partenaire) et l'objectif de cette étude est donc de construire une base de données incluant un périmètre plus large de polices afin de construire un ou plusieurs MLGs permettant d'estimer le prix de l'assurance sur l'ensemble du portefeuille. Le défi était donc d'intégrer dans la base de données autant de polices que possible tout en assurant la cohérence des données afin de pouvoir exécuter les MLGs appropriés.

Le périmètre retenu est celui des contrats internationaux (contrats sur lesquels Europ Assistance souscrit des polices dans plusieurs pays) car il incluait une grande variété de partenaires avec des données relativement cohérentes entre les différents partenaires.

Pour ce périmètre, des bases de données centrales étaient disponibles : une pour les données de ventes, une pour les données de sinistres et une pour les détails des produits. Cela a conduit à deux préoccupations majeures : 1 / la qualité des données pour fusionner les données de ventes et de sinistres était très différente d'un sous-périmètre (partenaire) à l'autre et j'ai dû exclure un certain nombre de partenaires de l'étude afin de maintenir un niveau acceptable de qualité des données ; 2 / la richesse et la qualité des données de la base « détails produits » n'ont pas été aussi bonnes que prévu. Cela m'a de nouveau obligé à exclure plusieurs partenaires pour lesquels je ne pouvais pas nettoyer suffisamment les données. J'ai également enrichi la base de données avec de nouveaux champs (type de produits, …) afin d'obtenir plus de variables de segmentation à utiliser pour nos modèles de tarification.

La base de données initiale (avant nettoyage) représentait un montant cumulé de 503 M€ de primes brutes émises (TTC) et la base de données finale conservée ne représentait que 160 M€ de primes brutes émises (TTC). Je n'ai pas eu d'autres choix afin de garantir la qualité de la modélisation à faire plus tard. Néanmoins, la base de données finale contient malgré tout une variété de partenaires (16 partenaires différents sur 5 segments différents de partenaires) pour un nombre total de 9,6M de polices.

La base de données a été divisée en deux échantillons aléatoires : un échantillon de modélisation (contenant 80% des polices) et un échantillon de validation (contenant les 20% restants des polices). La répartition de la base de données entre les deux échantillons a été faite à l'aide d'un package spécifique R appelé « caret » permettant d'assurer une répartition équilibrée des données entre l'échantillon de modélisation et l'échantillon de validation.

La stratégie de modélisation

Tout d'abord, je voulais remettre en question la méthodologie traditionnelle utilisée par d'autres études sur la segmentation des prix des produits d'assurance voyage. En effet, la plupart de ces études utilisent la loi de Poisson (qui est bien adaptée pour « dénombrer » le nombre attendu de sinistres). Néanmoins, dans le cas des couvertures annulation pour les polices à court terme, une fois qu'un sinistre est ouvert, l'exposition sur la police prend fin et aucun autre sinistre ne peut être ouvert sur cette police. Le nombre de sinistres par police est donc soit de 0 soit de 1 et dans ce type de situation, la loi binomiale semble plus appropriée.

Par ailleurs, il existe deux façons de calculer la fréquence d'une police à court terme :

1 / basé sur le nombre de polices : Nombre de sinistres / Nombre de polices

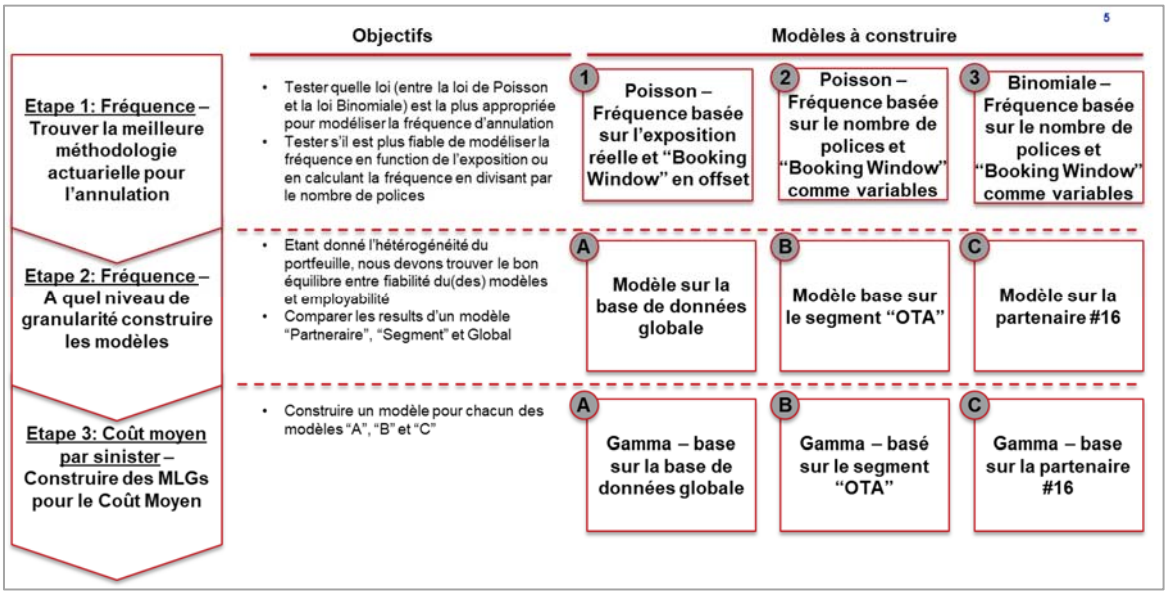
2 / basé sur l'exposition : Nombre de sinistres / Nombre de « polices annuelles »

Habituellement, dans d'autres études sur le sujet, la première méthodologie est utilisée. J'ai donc voulu vérifier si changer pour la deuxième méthodologie apporterait de la valeur.

Troisièmement, jusqu'à présent, les prix établis en Europe sur les couvertures annulation étaient basés sur un grand partenaire en particulier. L'objectif du présent mémoire est d'analyser si des modèles plus globaux pouvaient être construits (afin d'avoir moins de modèles à construire et à maintenir). Il a donc fallu enrichir la base de données initiale de champs supplémentaires afin de capter l'hétérogénéité du portefeuille comme expliqué précédemment. J'ai ainsi construit trois types de modèles (Partenaire, Segment et Global) et comparé la perte d'information en passant d'un modèle « Partenaire » à un modèle « Segment » ou « Global ».

La stratégie pour le processus de modélisation peut être résumée comme suit :

Figure 1 – Stratégie pour le processus de modélisation



Pour les étapes 1 et 2 décrites ci-dessus, les modèles ont été comparés à l'aide des indicateurs de performance clés suivants : AIC, BIC, déviance, taux de déviance attendu (EDR), Gini normalisé et RMSE.

Avant d'entrer dans les trois étapes mentionnées dans la figure ci-dessus, j'ai effectué un certain nombre d'analyses univariées afin de me familiariser avec les différents candidats prédicteurs disponibles dans la base de données. Nous avions 11 candidats prédicteurs dans la base de données. J'ai également effectué une analyse de corrélation entre ces 11 candidats prédicteurs en utilisant le V de Cramer (car la plupart des variables sont des variables catégorielles).

Enfin, avant d'entrer dans les trois étapes mentionnées dans la figure ci-dessus, j'ai construit un MLG en utilisant la méthodologie traditionnelle sur un grand partenaire sélectionné afin de tester le processus complet et de m'assurer que j'étais capable de passer par tout le processus. Cela a été fait

en utilisant la méthodologie « forward » pour la sélection des variables et cela a conduit à garder l'ensemble des candidats prédicteurs dans le modèle final, car l'AIC a continué à s'améliorer lorsque j'ai ajouté les derniers prédicteurs.

Validation de la méthodologie à utiliser pour modéliser la fréquence

Après avoir exécuté les trois modèles comme prévu sur la base de données globale, j'ai produit les différentes mesures de performance mentionnées ci-dessus afin de comparer les modèles et décider la méthodologie à utiliser à l'avenir. Les résultats ont montré que la méthodologie 1 (distribution de Poisson basée sur l'exposition réelle avec Booking Window comme offset) sous-performait nettement les autres méthodologies en termes de performance prédictive et cette méthodologie a donc été mise de côté rapidement.

La méthodologie n°2 (distribution de Poisson avec fréquence calculée sur le nombre de polices, méthodologie traditionnelle utilisée pour les produits d'assurance voyage) et la méthodologie n°3 (distribution binomiale, méthodologie « challenger ») étaient très proches en ce qui concerne les différentes mesures de performance décrites ci-dessus (avec un petit avantage pour la méthodologie n°2) et il était donc difficile de prendre position sur la base de ces indicateurs uniquement. Les deux méthodologies semblaient bien fonctionner sur la base de données globale et d'autres raisons ont ensuite été utilisées pour opter pour la méthodologie n°2 comme la méthodologie à appliquer pour les étapes suivantes. Néanmoins, il sera important de tester à nouveau la méthodologie n°3 à des étapes ultérieures lorsque de nouvelles bases de données seront mises à disposition afin de valider ou non la décision prise étant donné la proximité des deux méthodologies en termes de mesures de performance.

Définition du niveau de granularité optimal pour les modèles

J'ai lancé les trois modélisations : une sur le partenaire uniquement, une sur le segment auquel appartient le partenaire et une dernière sur la base de données globale. Les métriques de performance ont montré que la perte d'information liée au passage à des modèles « Segment » ou « Global » n'était pas significative et que l'on pouvait donc utiliser un modèle global au lieu de construire un modèle par partenaire (comme cela avait été fait jusqu'à présent). De plus, la perte d'information en passant du modèle « Partenaire » au modèle « Global » était très proche de celle du passage du modèle Partenaire au modèle Segment : contrairement à l'hypothèse initiale, cela indiquait que le modèle Segment n'apportait pas beaucoup de valeur dans le processus.

Étant donné que les résultats n'étaient pas conformes à l'intuition initiale, j'ai décidé de contre-vérifier les résultats en exécutant les mêmes types de comparaison sur d'autres partenaires. Cela a conduit aux résultats suivants.

Figure 2 – Niveau de granularité optimal pour la modélisation – comparaison des résultats clés

1st partner				
Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,3867	0,3946		
Segment	0,3770	0,3866	-2,5%	-2,0%
Global	0,3749	0,3849	-3,0%	-2,4%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,08742	0,08788		
Segment	0,08744	0,08789	0,0%	0,0%
Global	0,08744	0,08789	0,0%	0,0%

2nd partner				
Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,5404	0,5443		
Segment	0,5240	0,5222	-3,0%	-4,1%
Global	0,5184	0,5150	-4,1%	-5,4%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,06914	0,06859		
Segment	0,06917	0,06861	0,0%	0,0%
Global	0,06918	0,06862	0,1%	0,0%

3rd partner				
Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,3213	0,1697		
Segment	0,2223	0,2197	-30,8%	29,5%
Global	0,2135	0,2223	-33,5%	31,0%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,11637	0,11574		
Segment	0,11780	0,11703	1,2%	1,1%
Global	0,11768	0,11681	1,1%	0,9%

4th partner				
Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,7465	0,7419		
Segment	0,5520	0,5466	-26,1%	-26,3%
Global	0,5230	0,5131	-29,9%	-30,8%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,14289	0,16402		
Segment	0,14800	0,14854	3,6%	-9,4%
Global	0,14830	0,14885	3,8%	-9,3%

Ces résultats sont très différents d'un partenaire à l'autre. Cela indique que le passage à un modèle « Segment » ou « Global » peut conduire à une perte d'information importante et doit donc être fait avec prudence. Cela signifie probablement qu'il existe d'autres prédicteurs qui n'ont pas été pris en compte dans mes modèles et expliqueraient les différences entre les partenaires. Néanmoins, dans certains cas, nous avons vu que les modèles plus globaux apportent également de la valeur dans le cas des petits partenaires pour lesquels, nous avons peu de données (et sur lesquels les MLGs ont tendance à « overfitter »). Une caractéristique commune aux différents modèles est que, contrairement à l'hypothèse initiale, le modèle Segment n'apporte pas beaucoup d'avantages par rapport au modèle global.

Conclusion et améliorations

Avoir un modèle par partenaire n'est pas quelque chose de très pratique car cela demande des efforts importants en termes de modélisation et de maintenance. Par ailleurs, cela nécessite aussi une décision humaine importante au moment de la tarification d'une nouvelle opportunité (quel modèle est le plus approprié pour la tarification de cette nouvelle affaire). Néanmoins, avoir construit des modèles plus globaux apporte également de la valeur dans le cas de partenaires plus petits pour lesquels nous n'avons pas suffisamment de polices à modéliser.

Aussi, plusieurs actions peuvent être prises à la fois d'un point de vue commercial et actuariel afin d'améliorer notre capacité à construire des modèles plus globaux à l'avenir.

D'un point de vue commercial, les principales actions que je voudrais favoriser dans cette situation sont les suivantes :

- Encourager la standardisation de nos produits d'assurance voyage. Cela facilitera la compréhension de notre portefeuille. Il est probable qu'une part importante de l'hétérogénéité du portefeuille soit liée à l'hétérogénéité de nos produits.

- Entamer une discussion avec les partenaires sur les avantages de partager plus de données avec leur assureur pour alimenter notre base de modélisation

D'un point de vue actuariel, je me concentrerais sur les éléments suivants :

- Nous pourrions essayer d'utiliser des méthodologies de classification afin de créer des clusters de partenaires différents des clusters « Segment » que nous avons créés artificiellement

- Investir dans l'enrichissement de nos bases de données existantes avec de nouveaux champs qui décrivent mieux nos produits et permettent d'inclure dans les modèles plus de prédicteurs reflétant les différences entre les produits (âge du voyageur, …)

- Investiguer si des bases de données externes pourraient alimenter nos modèles de tarification.

- J'investirais dans un logiciel tel qu'Emblem ou Akur8 afin d'industrialiser la production de GLM tant qu'il faudra faire face à cette hétérogénéité sans pouvoir s'appuyer sur des modèles plus globaux

Synthesis note

Keywords : Non-Life insurance, Travel insurance, B2B2C, pricing segmentation, heterogeneous portfolio, GLM[4]

The objective of this thesis is to extend the use of segmented pricing models for Travel Insurance products where their use is less extended than for other insurance products. The challenge that will have to be addressed is the heterogeneity of the portfolio as the portfolio that will be considered is heterogeneous in terms of products sold, underlying risks (different types of travels insured) and distribution channels (different types of Tourism players).

The Travel Insurance Market in Europe

Travel Insurance products in Europe are mostly sold through Tourism players, such as Airlines companies, Online Travel Agencies (OTAs), Cruises, … We say that these travel insurance products are sold on a B2B2C channel.

These insurance products, most of the time, only cover the specific trip that was bought by the insured through the Tourism player's website or distribution channel. We call this type of Insurance products "Short Term" policies as the exposure only lasts from the time of subscription till the departure date (in the case of the cancellation cover) and between the departure date and the return date for other travel insurance covers (such as Medical Assistance).

In Europe, to the contrary of what happens in the United States, the use of Generalized Linear Models to price travel insurance covers is not widely used. This can be explained by different elements:

- The sale of these products is mostly intermediated, and this has a negative impact on the availability of data for the insurer to price the insurance covers
- The average premium of travel insurance product is relatively low compared to other insurance products and it decreases the interest of having GLMs in place
- Partners through which the travel insurance products are sold have little appetite for implementing on their sales channels complicated pricing models
- There is usually a wide heterogeneity of the portfolio insured, in terms products sold, trips insured (flight tickets vs packaged trip, low-cost vs premium trips, domestic vs International, …).

Nevertheless, the first and third points mentioned above are changing and the use of GLMs on Travel Insurance products is now progressively rising.

The most important cover in terms of premiums amount is the cancellation cover and this study will thus focus on this cover only.

[4] Generalized Linear Model

Data collection, preparation and validation

In this context and as part of this study, a significant amount of energy has been put on building an appropriate database to run this analysis. Indeed, in the past, most of the GLMs built for Travel Insurance products were built on restricted perimeters and the objective of this study was to build a database that includes a broader perimeter of policies in order to build one or several GLMs that allow to predict the insurance price across the full portfolio. The challenge was thus to bring into the database as much policies as possible while keeping consistency in order to be able to run proper GLMs.

The perimeter that has been chosen is International contracts (contracts on which Europ Assistance underwrites policies in several countries) as it was including a wide variety of partners with relatively consistent data across the different partners.

For this perimeter, central databases were available: one for sales data, one for claims data and one for products details. This has led to two key concerns for this step: 1/ the quality of the data to merge the sales and claims data was very different from one sub-perimeter (partner) to another and I had to exclude a number of partners of the study in order to keep an acceptable level of quality of data; 2/ the richness and the quality of the data of the "products details" table was as well not as good as expected. This has again forced me to exclude several partners for which I could not clean sufficiently the data. I have also enriched the database with new fields (type of products, …) in order to get more segmentation variables in the database.

The initial database that I was looking at was representing a cumulated amount of 503M€ of Gross Written Premiums (incl. taxes) and the final database kept only represented 160M€ of Gross Written Premiums (incl. taxes). This had to be done in order to secure the quality of modelling to be done later but at the same time, I still kept a variety of partners (16 different partners kept the database across 5 different segment of partners) for a total number of 9,6M policies.

The database was split into two sets: one "Train" set (containing 80% of the policies) and one "Test" set (containing the remaining 20% of the policies). The split of the database between the "Train" and the "Test" sets was made using a R specific package called "*caret"* which allows to ensure a balanced split of the data between the "Train" set and the "Test" set.

The strategy for the model building process

First, I wanted to challenge the traditional methodology used by other studies on pricing segmentation of travel insurance products. Indeed, most of these studies use a Poisson distribution (that is well suited to "count" the expected number of claims). Nevertheless, in the case of cancellation covers within Short Term policies, once a claim is opened, the exposure on the policy ends and no other claim can be opened on that policy. The number of claims per policy is thus either 0 or 1 and in this type of situation a binomial distribution seems more appropriate.

Secondly, there are two ways to calculate frequency for a Short-Term policy:

1/ based on the number of policies: Number of claims / Number of policies

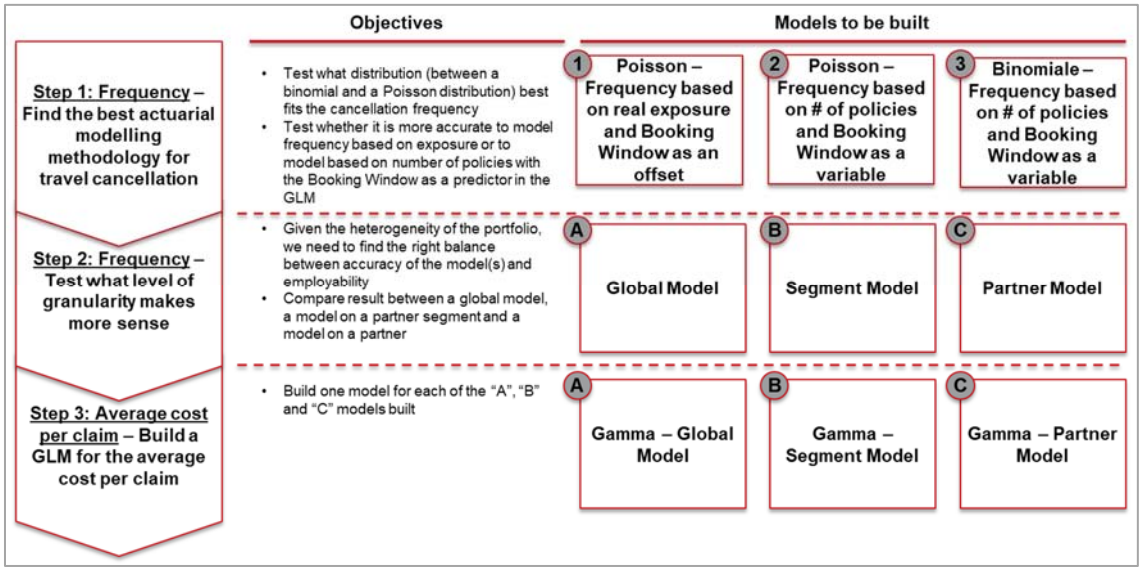
2/ based on the exposure: Number of claims / Number of "annual policies"

Usually, in other studies on the topic, the first methodology is used. I thus wanted to verify whether changing for the second methodology would bring value.

Thirdly, so far, pricings made in Europe on cancellation covers were built on a given large partner. The objective of this thesis was to see if more global models could be used (in order to have less models to be built and maintained). This thus required to enrich the initial database with additional fields in order to capture the heterogeneity of the portfolio as explained earlier. I thus built three types of models ("Partner", "Segment" and "Global") and compare the loss of information by moving from a Partner model to a Segment or Global model.

The strategy for the model building process can be summarized as follows:

Figure 3 - Strategy for GLM building process



Both for step 1 and step 2 described above, models were compared using the following key performance metrics: AIC, BIC, Deviance, Expected Deviance Ratio (EDR), Normalized Gini and RMSE.

Before entering the three steps mentioned in the figure above, I ran a number of univariate analyses in order to get familiar with the different predictor candidates available in the database. We had 11 predictor candidates in the database. I also ran a correlation analyses between these 11 predictor candidates using Cramer's V (as most of the variables are categorical variables).

Lastly before entering the three steps mentioned in the figure above, I built a GLM using the traditional methodology on a selected large partner in order to test the full process and make sure that I was able to go through the whole process. This was done using the forward methodology for variables selection and it led to keeping all the predictor candidates in the final model as the AIC continued to improve when I was adding the last predictors.

Validation of the methodology to be used for frequency

After running the three models as planned on the Global database, I produced the different performance metrics mentioned above in order to compare the models and decide what methodology should be used moving forward. The results showed that the methodology 1 (Poisson distribution based on real exposure with Booking Window as an offset) was really underperforming the other methodologies in terms of predictive performance and this methodology was thus put aside quickly.

Methodology #2 (Poisson distribution with frequency calculated on number of policies, traditional methodology used for travel insurance products) and Methodology #3 (binomial distribution, challenger methodology) were very close in terms of the various performance metrics described above with a little advantage for Methodology #2. It was thus difficult to take a position based on these metrics only. Both methodologies seemed to work fine on the global database and other more "business-oriented" reasons were then used to choose Methodology #2 as the one to be pursued. Nevertheless, it will be important to test again Methodology #3 at further stages when new databases are made available in order to challenge decision taken given how close the two methodologies were in terms of performance metrics.

Definition of the optimal level of granularity for frequency models

I ran the three models: one for the partner only, one for the segment to which the partner belongs and a last one on the global database. The performance metrics showed that the loss of information by moving to segment or global models were not significant and that we could thus use a global model instead of building one model per partner (as it was done so far). Also, the loss of information by moving from the Partner to the Global model was very close to the one of moving from the Partner to the Segment model : contrary to the initial assumption, this indicated that the Segment model was not bringing much value into the process.

Given the results were not in line with the initial intuition, I decided to cross-check the results by running the same types of comparison with other partners. It led to the following results.

Figure 4 – optimal level of granularity for frequency models - Key results comparison

1st partner

Gini

	Train	Test	%gap train	%gap test
Partner	0,3867	0,3946		
Segment	0,3770	0,3866	-2,5%	-2,0%
Global	0,3749	0,3849	-3,0%	-2,4%

RMSE

	Train	Test	%gap train	%gap test
Partner	0,08742	0,08788		
Segment	0,08744	0,08789	0,0%	0,0%
Global	0,08744	0,08789	0,0%	0,0%

2nd partner

Gini

	Train	Test	%gap train	%gap test
Partner	0,5404	0,5443		
Segment	0,5240	0,5222	-3,0%	-4,1%
Global	0,5184	0,5150	-4,1%	-5,4%

RMSE

	Train	Test	%gap train	%gap test
Partner	0,06914	0,06859		
Segment	0,06917	0,06861	0,0%	0,0%
Global	0,06918	0,06862	0,1%	0,0%

3rd partner

Gini

	Train	Test	%gap train	%gap test
Partner	0,3213	0,1697		
Segment	0,2223	0,2197	-30,8%	29,5%
Global	0,2135	0,2223	-33,5%	31,0%

RMSE

	Train	Test	%gap train	%gap test
Partner	0,11637	0,11574		
Segment	0,11780	0,11703	1,2%	1,1%
Global	0,11768	0,11681	1,1%	0,9%

4th partner

Gini

	Train	Test	%gap train	%gap test
Partner	0,7465	0,7419		
Segment	0,5520	0,5466	-26,1%	-26,3%
Global	0,5230	0,5131	-29,9%	-30,8%

RMSE

	Train	Test	%gap train	%gap test
Partner	0,14289	0,16402		
Segment	0,14800	0,14854	3,6%	-9,4%
Global	0,14830	0,14885	3,8%	-9,3%

These results are very different from one partner to another. This indicates that switching to a "Segment" or "Global" model can lead to significant information loss and should therefore be done with caution. This probably means that there are other predictors that were not taken into account in

my models and would explain the differences between the partners. Nevertheless, in some cases, we have seen that more global models also bring value in the case of small partners for which we have little data (and on which GLMs tend to overfit). A characteristic common to the different models is that, contrary to the initial assumption, the Segment model does not provide many advantages over the overall model.

Conclusion and improvements

Having a model per partner is not something very convenient as it requires significant efforts in terms of modelling and maintenance and it still requires significant human decision at the time of pricing a new business. Nevertheless, having built the more global models also bring value in the case of smaller partners for which we do not have enough policies to be modelled.

Also, several actions can be taken both from a business and actuarial prospective in order to improve our ability to build more global models moving forward.

From a business point, the main actions that I would like to foster in this situation are the followings:

- I would encourage the standardization of our travel insurance products. This will ease the understanding of our portfolio
- I would discuss with partners the benefits of sharing more data with the insurer in order to feed our database

From an actuarial point of view, I would focus on the following elements:

- We could try to use classification methodologies in order to create cluster of partners that are different from the “Segment” clusters that we have artificially created
- I would invest in enriching our existing databases with new fields that better describe our products and allow to include in the models more predictors reflecting differences across products
- I would invest in a software such as Emblem or Akur8 in order to industrialize the production of GLMs as long as we need to deal with this heterogeneity without being able to rely on more global models

Greetings

First of all, I would like to thank Antoine Parisi, *Europ Assistance Group CEO*, who first gave me the chance to enter the insurance world six years ago, who then proposed me to follow him in the Europ Assistance adventure and who finally encouraged me to take this challenge to start my actuarial studies. It has been a very interesting and challenging journey so far and I am looking forward for the next steps.

Secondly, I would like to thank Etienne Bonnet, *Europ Assistance Group Chief Insurance Officer* for his encouragements and guidance throughout the process of writing this thesis. I am proud of the work done as part of this thesis. I truly believe it will support our day-to-day work within the Travel Pricing team and I am impatient to see the benefits.

Thirdly, I would like to warmly thank Matthieu Quilfen, *Travel Actuarial Manager within Europ Assistance Group*, for his strong support on this study. His knowledge and support have been detrimental in this study. Matthieu is a very nice professional and it is a pleasure to work with him on a daily basis.

Finally, I would like to thank my wife and our three daughters (Léa, Elinor and Salomé) for their encouragement, patience and support in the past 5 years that have allowed me to succeed in my actuarial studies. It has been a significant commitment from the whole family.

Table des matières

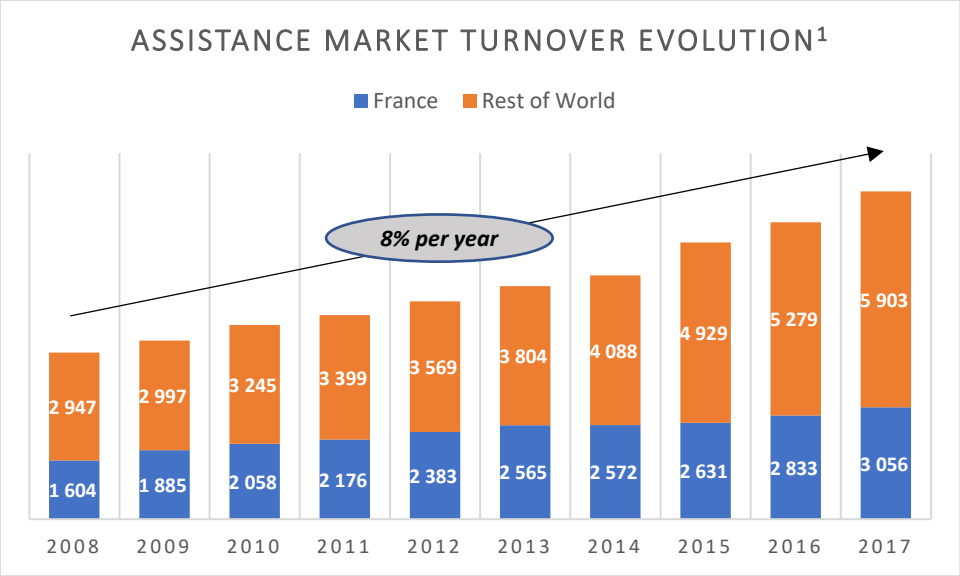
1. The Travel Insurance Market in Europe

a. The Assistance market and the main players

i. Assistance companies are multi-liners

The Travel Insurance market in Europe is mostly taken by big International Assistance Companies. The Assistance market has been growing at a relatively good pace as it grew at an average 8% per year over the last 10 years.

Figure 5 - Assistance market turnover evolution



The Assistance Companies usually are multi-liners in the sense that they do not only address the Travel Insurance Market. They are, most of the time, also present on all the assistance traditional Lines of Business, such as Roadside Assistance, Home Assistance or Elderly care.

Roadside Assistance represents in France in 2017 61% of the turnover of the French Assistance Companies[5]. This is thus by far the largest Line of Business of French Assistance Companies. Roadside Assistance products are mostly sold through Insurance Companies (this cover is included or is sold on an optional basis within Auto Insurance contracts) or through Car Manufacturers that must offer a 2 years assistance cover to a client buying a new car. This cover can be sold as an insurance product or a "service product". The Roadside Assistance market is a relatively mature market and the evolution of the activity on this market is mostly driven by weather conditions.

Home Assistance represents in France in 2017 almost 20% of the turnover of the French Assistance Companies[1]. This is a growing Line of Business as it grew by 12% between 2016 and 2017. These products are mostly sold in inclusion to Home Insurance policies or through Utilities companies that offer to their customers the possibility to subscribe Home Assistance Services.

[5] https://www.argusdelassurance.com/acteurs/assisteurs/assistance-des-resultats-en-hausse-pour-la-profession.129052

Elderly care is still marginal in terms of turnover but most of the Assistance Companies are investing on this segment of business as the perceived potential is significant and also because Assistance Companies have most of the capabilities to develop new offers on this market.

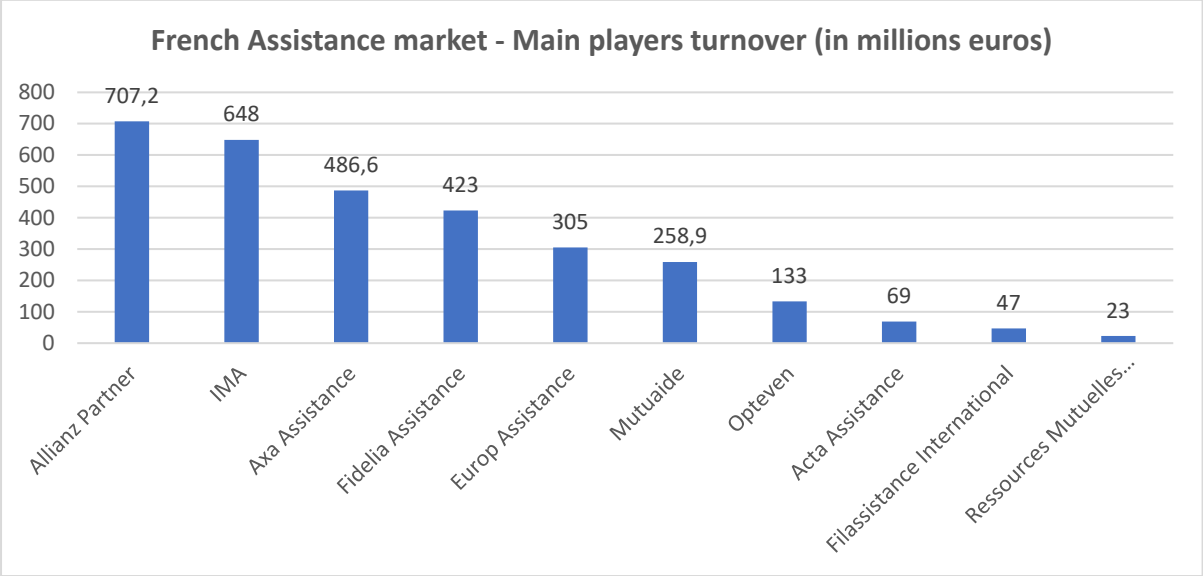
ii. The main players

In Europe, the big International Assistance Companies are Allianz Partner (that includes Allianz Global Assistance), Europ Assistance and Axa Partner (that includes Axa Assistance). All these Assistance companies belong to big Insurance Companies: Europ Assistance being fully integrated in the Generali Group.

It is difficult to rank these Assistance companies in terms of size as most of these companies (being incorporated in larger entities such as Allianz Partner or Axa Partner) do no longer publish their numbers on the Assistance market. Nevertheless, it is well-known on the market that, among these three players, Allianz Global Assistance is by far the biggest player on the Travel Insurance market then followed by Europ Assistance.

In local markets, there are as well local Assistance players that can be big on the local market. For example, in France, despite being a large international player and having been created in France, Europ Assistance is only ranked as the 5th largest player on the French Assistance market by the SNSA[6].

Figure 6 - French assistance market - Main players' turnover[7]



[6] SNSA stands for « Syndicat National des Sociétés d'Assistance ».

[7] https://www.argusdelassurance.com/classements/classement-assistance-2018-l-assistance-dans-l-urgence-de-l-innovation.128944

b. Zoom on Europ Assistance

Europ Assistance was created in Paris in 1963 by Pierre Desnos. At the same time, the Travel Assistance business was created and Europ Assistance has often been seen as the pioneer on the Travel Assistance business. The idea of the creation of Europ Assistance was to offer protection to travelers abroad. Travel assistance was thus at the heart of the creation of Europ Assistance and it is only later that it entered other segments of business such as Roadside Assistance or Travel Insurance.

Europ Assistance was created in France. Over the years, Europ Assistance has extended its international footprint and now, has 41 assistance centers across the world. Europ Assistance now counts almost 8000 employees and a network of around 750.000 medical, travel and roadside assistance providers.

Since the launch of the company in 1963, Generali was part of the shareholders. In 2001, Generali acquired 100% of the stakes of the Europ Assistance Group.

In 2014, Antoine Parisi joined Europ Assistance and was appointed Chief Executive Officer of the Group. A few months later, beginning of 2015, he launched a new strategic plan, called the WeConnect plan. The main purposes of the plan were:

1/ to transform Europ Assistance from a federation of entities into a real Group. Several initiatives were launched at Global level in order to leverage group synergies and significant efforts were made to align the strategies of the different entities with the Group strategy

2/ to better connect with the Generali Group. The ambition was to better support the Generali Group strategy and become a strategic differentiator for Generali.

3/ to relaunch growth. As we have seen, the Assistance market has been growing constantly over the last 10 years. In that context, during the period 2010-2014, Europ Assistance was relatively flat in terms of turnover. A significant growth plan was thus put in place with the objective to go from 1,3Bn€ of total turnover in 2014 to 2,0Bn€ in 2020. To achieve that growth, the Travel Business has been identified as the key growth factor and the objective was to double the premiums on this Line of Business over the period 2014-2020.

4/ to innovate. Besides growing on Europ Assistance's traditional Lines of Business (Roadside Assistance, Travel Assistance & Insurance), it was also decided to build on Europ Assistance core capabilities (Management of networks, Management of platforms and Technology) to grow into new areas of business such as Home Assistance or Elderly care.

Europ Assistance is on well on track with this strategic plan. New offers have been launched on Home Assistance and Elderly Care; targeted acquisitions were made in Eldery Care. A significant acquisition was also made in the Travel Insurance Business in the USA, making of Europ Assistance the third largest Travel Insurer in the USA.

In terms of financial targets, Europ Assistance is well in track with the strategic plan and was expected to reach the 2Bn€ turnover target (before the Covid crisis started).

c. The distribution channels

i. B2B2C

This distribution channel often represents a vast majority of the sales of travel insurance products made by travel insurance companies. The travel insurance Companies thus contract with a B-partner so that Travel insurance products are sold on the website (most of the time) or through other sales channels of the B-partner.

In this case, the products that are sold, are, in almost all cases, Short Term products that only cover the trip bought by the final customer on the sales channel of the B-partner.

There are a wide variety of B-partners through which a travel insurance company can sell Travel insurance products. Hereafter, a non-exhaustive list of the main segments of B-partner:

- Credit Cards issuers or Financial Institutions
- Cruise, airline or train Companies
- Travel agencies, Tour Operators, Online Travel Agencies
- Vacation Rentals companies

These segments can be very different in terms of products sold, conversion rate and also, level of sinistrality observed.

ii. B2C

Travel Insurance products can also be sold on the Travel Insurer website. This represents a small portion of the sale of the Travel Insurance Companies. On this channel, it is more usual to sell annual policies that will cover the different trips made by the customer throughout the year. Also, it will often be bundled with other products, such as Roadside assistance covers.

d. The Travel insurance covers

i. The cancellation cover

The purpose of the cover is to reimburse the Insured for the expenses that he/she incurred directly due to the cancellation of the covered travel.

There is a wide variety of events for which the Insured may ask for the reimbursement of his/her travel costs. These events must be defined in the policy wordings. The most frequent covered events are the followings:

- Serious Illness, Serious Injury or death of: an Insured, a Family Member, the person designated for the custody of minors or disabled persons the Insured is responsible for or the Professional Substitute
- Serious Damage to the Home or Professional Premises of an Insured
- Redundancy of the Insured
- Commencement of employment in a new company

- Summons of an Insured to appear as a party, witness, jury member in court or any other public authority

In the case of a Short-Term product, the risk of the cancellation cover starts at the time of the booking of the trip and ends at the time of departure. Therefore, it can only be bought at the moment of the booking of the trip (or right after).

There is a set of exclusions that are usually put in Cancellation covers to mitigate the risks. The most commonly used exclusions that can be observed on the market are:

- The consequences of an event intentionally caused by an Insured, a Family Member or a Travel Companion
- Illnesses or injuries derived from the consumption of alcoholic beverages
- Consumption of narcotics, drugs or medicine, other than those which have been prescribed by a doctor.
- Suicide, attempted suicide or self-harm on the part of an Insured, a Family Member or Travel Companion
- The consequences of a Serious illness of the Insured diagnosed to the Insured before the start of the Membership of the Group Insurance Policy
- The consequences of an Accident occurring before the Membership to the Group Insurance Policy

In some countries, there is an increasing demand from the market to offer "Cancellation for Any Reasons". This obviously provides peace of mind to the Insured. Nevertheless, in some countries, it raises some Compliance issues as it is no longer considered as an Insurance product (as there is not anymore, an Insurable risk). This is the case in all European countries, but it is not the case in the USA for example.

ii. The Medical Assistance covers

Under the medical assistance cover, there is a wide variety of sub-covers but the main purpose of all these covers is basically to take care of the Insured who suffer from an illness or an accident during a trip abroad. Taking care of the Insured encompass mostly three elements:

- the reimbursement of medical expenses
- the organization of the medical care abroad and potentially the repatriation of the Insured
- taking care of the family of the Insured (in the Home country or abroad)

In the case of Short-Term products, the risk of the medical assistance covers starts when the Insured starts his/her trip and finishes when the Insured finishes the trip. It can thus be bought from the booking of the trip till the departure date.

1/ Medical expenses abroad

The purpose of the cover is to reimburse to the Insured the differences between the medical costs incurred abroad and what is already reimbursed by the Insured's Home Country Social Security or Private Health Plan. This cover does not require that the Assistance Company manages (from an Assistance point of view) the medical case.

Depending on the market in which the product is sold, the limit for medical expenses abroad can vary significantly. In some markets (such as France or Spain for example), it is not unusual to have limits at

10-15k€/event. In the UK, the limits can be as high as 2M£ and in Germany, the market practice is to sell unlimited medical expenses abroad.

2/ Repatriation

If the Insured suffers an illness or has an accident during his/her travel abroad and provided that this event prevents the Insured from continuing his/her Travel, the Insurer will organize the repatriation of the Insured in close collaboration with Medical teams in the country where the Insured is.

The Assistance Company medical team will be the decision-maker in authorizing or not the transfer to a better-equipped or specialized hospital close to the home of the Insured and will organize that transfer:

- in accordance with the degree of severity of the condition, and
- using the most appropriate means of transport.

There are usually no limits for these types of products and the Insurance Company will bear 100% of the incurred costs.

Offering this type of protection to the customers require two elements that are very specific to the Assistance Companies:

- An internal medical team that will be in touch with the Medical team abroad to decide when and how repatriate the Insured

- A network of medical correspondents across the Globe to facilitate the management of the medical case

3/ Other medical assistance covers

There are a variety of assistance services that can be offered to the Insured in order to support him/her while being abroad ill or injured:

- if the Insured is abroad alone and is hospitalized for more than x days, the Assistance Company will organize and pay the costs so that a family member can join the Insured abroad while being ill or injured

- if the Insured is abroad with children or a disabled person, the Assistance Company will organize and pay the costs of sending a family member to take care of the children or disabled person

- if the nature of the illness or injury prevents the Insured of travelling back to the home country on the expected date of return, the Assistance Company will pay for the cost of the extra hotel nights to be paid

iii. The luggage covers

In the case of Short-Term products, most of the time, the duration of the luggage benefits covers the insured from the departure date until the return date.

There are two main luggage covers:

- Delayed delivery of luggage
- Loss, damage, and Violent Robbery of luggage

1/ Delayed delivery of luggage

If there is a delay of more than 24 hours in the delivery of the luggage that was checked-in, the Insured will have the ability to claim the costs of any necessary purchases (clothes, food and toiletries). These costs will be reimbursed by the Insurer if the costs were incurred:

- At a destination of the covered travel
- At a location where the covered travel involves a stop-over between connecting flights

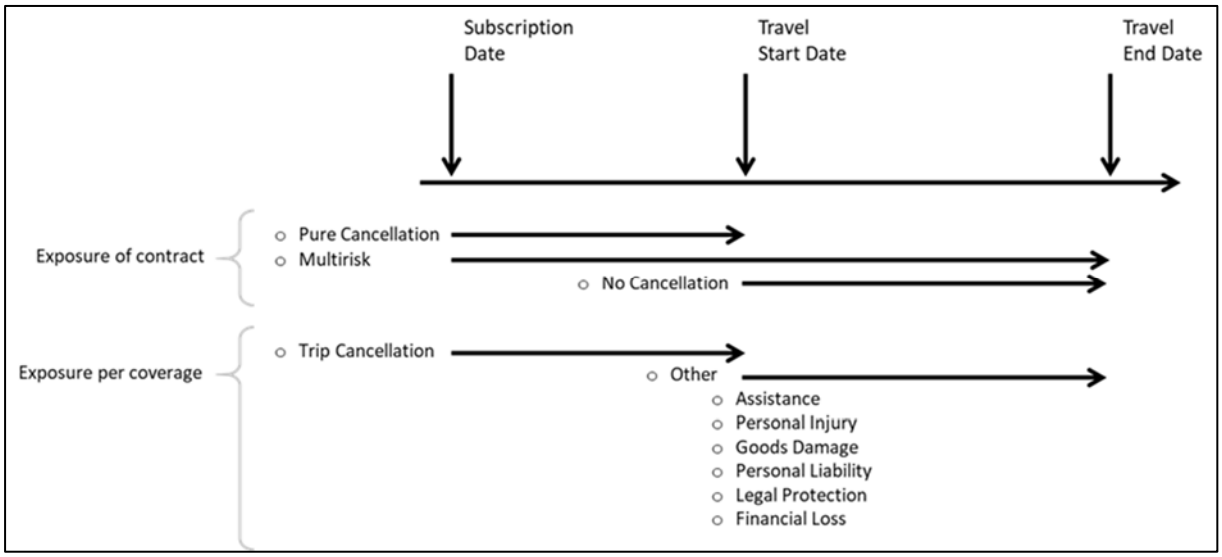
2/ Loss, Damage, and Violent Robbery of luggage

If during the course of the Travel, the luggage of the Insured is stolen or is lost definitively or if it suffers damages for causes attributable to the carrier, the insurer will reimburse the costs of the goods included in the luggage up to a certain amount determined in the Terms and Conditions and after deduction of an excess.

iv. Travel insurance covers and exposure

To summarize this section on travel insurance covers, I wanted to recap when the policy is exposed depending on the type of covers included in the Travel insurance Short-Term products.

Figure 7 - Recap of policy exposure by type of product and type of cover



As we can see from the "Figure 3" above, there are three important dates to have in mind when speaking about Travel Insurance Short Term products:

- The Subscription date
- The Travel start date
- The Travel end date

The cancellation cover exposure lasts from the subscription date till the departure date whereas the other covers start at the travel start date and ends at the travel end date.

Contrary to many P&C policies, the travel Short-Term policies are thus not annual policies. This will have to be taken into account when we will be modelling these products.

e. Main trends and challenges moving forward

i. EIOPA report on Travel Insurance

Since a few years, regulators are concerned by the practices surrounding the sale of Travel Insurance products and the value-added by these products. In 2018, EIOPA has thus launched a thematic review in order to better understand the market practices on Travel Insurance. This started by asking the main players of the Member States to reply to a questionnaire.

In October 2019, the EIOPA published the report of the thematic review as well as a warning to the Industry. The main outputs of the report could be summarized as follows:

- **Travel Insurance products remain valuable** for the end-customers, but the market is going through changes that may bring opportunities as well as new risks

- **Commissions can be extremely high** in this sector and customers may pay more for commission than insurance

- In terms of **Product design**, wordings are complex, products may not fit the needs of some consumers and consumers are not enough informed of some exclusions (e.g. pre-existing medical conditions)

- **Claims ratios are low** (below 40% as an average but some products can reach less than 20%)

ii. New products

The main covers sold for Travel Insurance exist for a long time but, recently, some new parametric insurance products are emerging on the market, leveraging data such as:

- Flights incidents (delays, cancellations)
- Weather conditions and weather forecasts

Based on the flight incidents, Travel Insurers can develop new products such as:

- **Pro-active flight delay**: in case of a flight delay on which an Insured is booked, the insurer can trigger an automatic payment of an indemnity to the customer so that the customer can buy food/drinks while waiting at the airport

- **Missed connection**: in case of connecting flights where the first flight is delayed, the insurer can proactively identify a new flight to get the customer to his/her destination and pay the cost of the rebooked flight

Based on weather data, Travel Insurers can develop new products such as:

- **Bad weather cover**: in case of bad weather conditions during the stay, the Insurer can generate an automatic payment of an indemnity (e.g. to cover the costs of new activities that the customer needs to pay)

- **Bad weather cancellation**: in case of bad weather forecast 2 days before the departure date, the customer can choose to cancel his/her trip and get reimbursed for the cost of cancelling the trip

iii. The impact of the Covid19 crisis on the Travel Insurance industry

The Travel insurance industry is obviously deeply interconnected with the tourism industry and, therefore, the Covid19 crisis has severely impacted travel insurance companies as well. The crisis has impacted both the revenues and the claims of the travel insurance companies.

The biggest impact of the Covid19 crisis is probably on the premiums side and can be decomposed in several sub-effects:

- Customers are booking less travels as there are a lot of uncertainties regarding what travels will be possible when. This impacts insurers revenues as well.
- Lockdowns or borders closures have severely reduced the number of travels made in the first half of 2020 creating a loss of premiums for insurers through
    1. Reimbursement of already underwritten premiums alongside the reimbursement of the trips (packages) booked by the customers
    2. Postponement of premiums earned later in the year or to the next accounting year due to vouchers emitted by our partners on cancelled flights or bookings
- In the second half of 2020, lighter travel restrictions (eg. Quarantine upon arrival at destination, quarantine upon return in the country of origin, test result requirements to cross borders) and more importantly, all the uncertainty surrounding these rules are still impacting very negatively the demand for travel. Research on the topic says that the Travel market will probably not be back to “pre-covid” level before at best 2023.

\+ The crisis has increased the awareness for customers on the need to be insured when travelling abroad and this tends to positively impact conversion rate[8] on our partners’ websites.

In terms of claims, the impact for most insurers has been less important than on the revenues side but the situation varies significantly from one partner to another depending on the following elements:

- Existence of an exclusion for epidemics/pandemics or not: not all products do exclude epidemics
- Reinsurance treaty: not all travel insurers were properly covered by their reinsurance treaty for epidemics. Notably, some travel insurers had restrictive hours clause related to the

[8] Conversion rate here refers, on the B2B2C channel, to the number of travel policies sold divided by the number of trips sold

epidemics cover which has reduced significantly their ability to claim under their reinsurance treaty

Given the situation, there is an increased request from customers and partners to be covered for epidemics/pandemics. From an insurance point of view, it makes sense to cover travelers infected with this new disease. Nevertheless, it is difficult to meet customers' expectations to be covered for all consequences of epidemics and pandemics without exposing the travel insurers to significant accumulation risks. At the same time, given the expected impact on their financials in 2020, reinsurers are unlikely to continue to include epidemics/pandemics in the reinsurance agreements.

2. Generalized Linear Models (GLMs) in Non-Life pricing[9, 10]

a. Overview of GLMs – Three components

i. The Random variable – Exponential components

In a GLM, the variable that we want to model (the "target variable") is considered as a random variable. This random variable follows a probability distribution and this probability distribution needs to be a member of the Exponential family.

There are a number of well-known probability distributions that are part of the Exponential family, such as the Normal, Gamma, Poisson or binomial distributions.

A probability distribution is part of the Exponential family if it admits a probability density function that can be written under the form:

$$f(y_i, \theta_i, \varphi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i, \varphi)\right) \qquad (3.1)$$

Where:

- the functions a, b and c are specified depending on the type of exponential functions
- θ_i is unknown, varies based on *i* and is a function of the expected value $\theta_i = g(\mu_i)$
- φ is the dispersion parameter (it is related to the variance, but it is not the variance), is supposed to be the same for all *i* and is supposed to be known
- ω_i is a weight

Table 1 - Components of some common exponential family distributions

Distribution	Probability Density Function	$\boldsymbol{\theta(\mu)}$	$\mathbf{b(\theta)}$	$\boldsymbol{\varphi_0}$
Normale	$f(y) = \frac{\exp(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2)}{\sigma\sqrt{2\pi}}$	μ	$\frac{\theta^2}{2}$	$\frac{\exp(-\frac{y^2}{2\sigma^2})}{\sigma\sqrt{2\pi}}$
Poisson	$f(y) = \exp(-\mu)\ \frac{\mu^y}{y!}$	$\mathrm{Ln}(\mu)$	$\exp(\theta)$	$\frac{1}{y!}$

The probability distributions included in the Exponential family include several properties that are very useful for GLMs:

1/ $\mu_i = E(Y_i) = b'(\theta_i)$ (3.2)

2/ $Var(Y_i) = \varphi_0\, b''(\theta_i)$ (3.3)

μ_i is the mean of the distribution, it represents the expected value of the random variable. This expected value is specific to each record and is the ultimate output of the GLM.

[9] D. ANDERSON *A Practitioner's guide to Generalized Linear Models (3rd Edition).* 2017

[10] M. GOLDBURG, A. KHARE and D. TEVET *Generalized Linear Models for insurance ratings.* 2016 (CAS Monographe Series, Number 5)

Therefore, we can write the expected value and the variance for each of the three distributions taken above as follows

Table 2 - Expected value and variance of some common Exponential family distributions

Distribution	$E(Y) = b'(\theta)$	$Var(Y) = \mathrm{a}(\varphi_0)\, b''(\theta_i)$
Normale	$\mu = \theta$	σ^2
Poisson	$\mu = \exp(\theta)$	μ

ii. The systematic component

In a GLM, the relation between the expected value of the random variable and the predictors can be written as follows:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (3.4)$$

Where:

- β_0 is known as the intercept
- x_{ij} are the predictors
- β_j are the coefficients that will be predicted by the model
- g(…) is called the "Link function" (see next paragraph)

iii. The link function

The link function is a great asset of the GLMs as it allows to have more flexibility in the definition of the relation between the expected value of the random variable and the predictors. When calculating the value that we will need to consider in our pricing exercises, we will thus need to take the inverse of the function to the results.

Applying the inverse function to the equation (3.4) can be written as follows:

$$\mu_i = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \quad (3.5)$$

Very commonly used link functions are the log function and the logit functions.

The log function has some specificities that make the interpretation of the outputs of the model even easier to understand and manipulate. Indeed, it will transform the output of the model into a multiplicative model.

$$ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad (3.6)$$

Therefore:

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) = \exp(\beta_0)\, x \exp(\beta_1 x_{i1})\, x \ldots x \exp(\beta_p x_{ip}) \quad (3.7)$$

Logit models use as link function:

$$g(\mu_i) = \log\left(\frac{\mu}{(1-\mu)}\right) \quad (3.8)$$

Using the logit model is of interest when the expected value of the random variable is comprised between 0 and 1 (such as the probability of success of a binomial).

b. The most-commonly used distribution

Some of the distributions of the Exponential family are particularly well-suited to model severity of claims or frequencies.

i. Distributions used for severity of claims

1/ The Gamma distribution

The Gamma distribution is the most widely used distribution to model the severity of claims as

1/ it is a distribution of positive random variables

2/ it is right-skewed (with a long tail to the right)

The Gamma distribution is characterized by two parameters, the shape parameter (*k)* and the scale parameter (ϑ). The Gamma distribution has the following probability density function

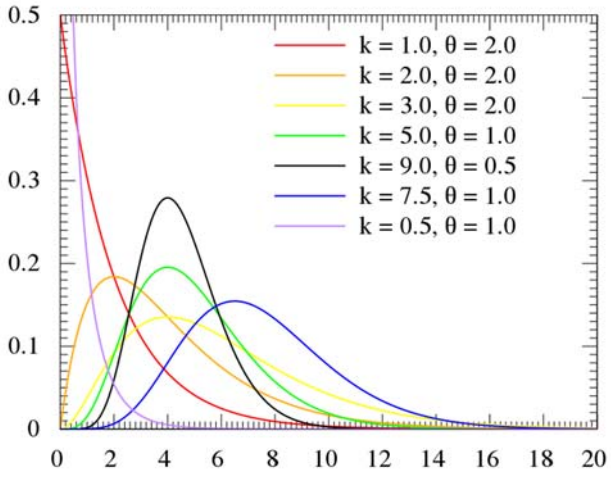
$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \qquad (3.9)$$

The first two moments of the Gamma distribution are the following:

- $E(X) = k\theta$
- $Var(X) = k\theta^2$

In the below graphic representation of the Gamma distribution, we can observe that when the scale parameter decreases the variance decreases as well.

Figure 8 - Probability density function of the Gamma distribution[11]



[11] https://en.wikipedia.org/wiki/Gamma_distribution

2/ The Inverse Gaussian distribution

The Inverse Gaussian distribution is relatively like the Gamma distribution in the sense that it is right-skewed and has a sharp peak. The difference with the Gamma distribution is that the tail is longer for the Inverse Gaussian distribution which makes it even more appropriate for variables with more extreme values.

The Inverse Gaussian distribution is characterized by two parameters, the mean ($\mu>0$) and the shape parameter ($\lambda>0$).

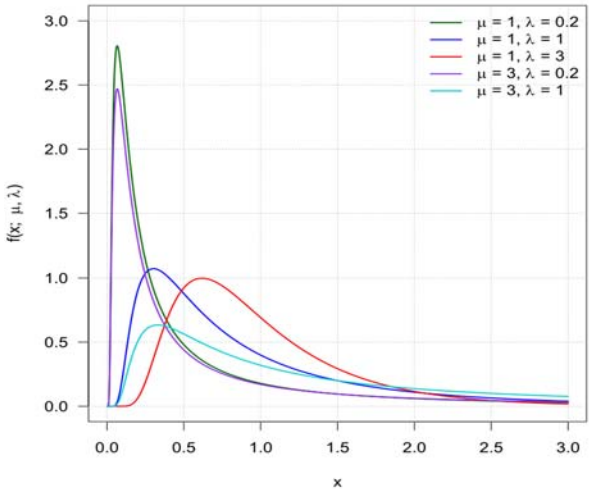
The Inverse Gaussian distribution has the following probability density function

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right) \quad (3.10)$$

The first two moments of the Gamma distribution are the following:

- $E(X) = \mu$
- $Var(X) = \frac{\mu^3}{\lambda}$

Figure 9 - Probability density function of the inverse Gaussian distribution[12]



ii. Distributions used for frequency of claims

1/ The Poisson distribution

The Poisson distribution counts the number of events that occur in a given time interval. It is thus a discrete distribution, but it can be used to model frequencies as it also accepts fractional values.

[12] https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution

The Probability Mass Function of the Poisson distribution is the following

$$\frac{\lambda^k e^{-\lambda}}{k!} \quad (3.11)$$

The first two moments of the Poisson distribution are:

- Mean = μ
- Variance = μ

The issue is that, most of the time, frequencies have variances that are greater than their mean. This is called overdispersion and it can be dealt with using the over-dispersed Poisson distribution or the Negative Binomial Dispersion.

2/ The Negative Binomial distribution

The Negative Binomial distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted r) occurs.

$$k ==> \binom{k+r-1}{k} (1-p)^r p^k \quad (3.12)$$

The first two moments of the Poisson distribution are:

- $Mean = \frac{pr}{(1-p)}$
- $Variance = \frac{pr}{(1-p)^2}$

c. The use of offset within GLMs[13]

Offsets in GLMs are usually used when the actuary does not update the full rating plan. Indeed, it can happen, that there is a need to update part of the rating plan while other elements remain identical (as for example, one of the rating factor will be calculated using a specific methodology while the others will be calculated through a GLM).

When doing so, the GLM will have a fixed element (to which the GLM will not assign a coefficient). Nevertheless, the fixed element will still be used in the rating plan and therefore, the GLM must be aware of that in order to calculate more appropriately the coefficients for the rest of the rating variables. This is allowed in GLMs by the use of an offset: the offset is a fixed element to which the GLM must not allocate a coefficient, but its existence must still be taken into account when calculating the coefficient for the rest of the rating variables.

This can be seen as follows

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + offset \quad (3.13)$$

As we will see later, we will have to use an offset in our case. Indeed, as we will see, we can calculate the frequency based on the number of policies or based on the real exposure of each policy. Given that all policies do not have the same exposure, when modelling the frequency based on the exposure of

[13] F. PLANCHET and A. MISERAY *Tarification IARD – Introduction aux techniques avancées – Version 1.3.* 2017 (ISFA)

the policy, it will have to be taken in the GLM. This is done with the use of an offset for the exposure. Let's assume in the coming equations that μ_i is the expected number of claims and let's assume that our link function will the log function.

$$log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \qquad (3.14)$$

Taking into this equation, the exposure of the policy, we would then have:

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta'_0 + \beta'_1 x_{i1} + \beta'_2 x_{i2} + \cdots + \beta'_p x_{ip} \qquad (3.15)$$

This equation can then be rewritten as:

$$log(\mu_i) = \beta'_0 + \beta'_1 x_{i1} + \beta'_2 x_{i2} + \cdots + \beta'_p x_{ip} + \log(t_i) \qquad (3.16)$$

Where $\log(t_i)$ is the offset.

d. Validation of the model[14]

i. Variables selection

For a given predictor, the output of the model will be an estimate of the coefficient. Nevertheless, this does not answer two important questions on this coefficient estimate:

1/ Is the estimate accurate? Is it close to the true value of the coefficient?

2/ Is the predictor really a predictor? In other words, does the predictor really have an impact on the target variable?

To answer these questions, we usually look at the following metrics: the standard error, the p-Value and the confidence interval.

1/ The standard error

As explained, the output of the model for a given predictor is an estimate of the coefficient as a result of a random process. The standard error is the estimated standard deviation of the estimation process of the coefficient. The smaller is the standard error, the more reliable will be the estimate of the coefficient and on the contrary, the bigger is the standard error, the less reliable will be the estimate of the coefficient.

2/ The pValue

The pValue is the estimated probability of having the value obtained as an estimate of the coefficient by pure chance. A pValue of "x" means that, in case the true value of the coefficient is 0 (and thus, has no impact on the target variable), the estimated probability of having the obtained value for the coefficient is "x".

[14] https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/

In case, the model returns a high pValue for a predictor, we cannot conclude that this predictor has no effect. It is just that our dataset cannot confirm the effect of the predictor on the target variable.

Significance test are usually expressed in terms of NULL hypothesis. In this case, the NULL hypothesis tested by the pValue is that the true value of the predictor is 0. We can thus reject the NULL hypothesis when the pValue is small. There is no definite rule to define a value of the pValue above which we should reject the NULL hypothesis, but the common practice is to consider that a pValue above 5% does not allow to confirm the significance a of predictor.

3/ Confidence interval

The issue of the pValue is that it only tests the NULL hypothesis where the true value of the predictor would be 0. The confidence interval defines the range of value that would not be rejected at a given pValue. Usually, the confidence interval is based on a 5% pValue and is then called the 95% confidence interval. The confidence interval thus gives a range of value that would not be rejected under a certain level of confidence, 95% in our example.

ii. How to measure the model fit?

1/ Log-Likelihood

Thanks to the specification of the GLM and the outputs of the GLM, it is possible to calculate for each record the probability that the GLM would assign the obtained value for that record. The likelihood is the multiplication of this probability for all the records. With a large set of records, the value of the likelihood will be very small and not easy to read and therefore, in order to ease its readiness, we take the log of the likelihood to obtain the log-Likelihood.

The Log-likehood itself is still difficult to interpret and as a consequence, we usually calculate the Log-likelihood of the *worst model* (the *null model*) and of *the best model* (the *saturated model*) as boundaries of the Log-likelihood and compare the likelihood of our models to these limits:

- The ***null model***: is a model in which we would not have any predictors. The model would only have an intercept (the mean of the records) and return systematically the same value.

- The ***saturated model***: is a model in which we would have as many predictors as records available in the dataset. This type of model can be solved as it would contains as many equations as unknowns. This model would be very complex and as such has no interest as it would perfectly replicate all the records included in the dataset.

2/ Deviance

The deviance is a more useful metric based on the log-likelihood of the model and of the saturated model:

$$Deviance = 2\ x\ (ll_{saturated} - ll_{Model}) \qquad (3.17)$$

Deviance will always reduce when we increase the number of parameters in the model. Nevertheless, our objective is to find the best compromise between having a very complex model that fits very well the data and a less complex model (less costly to implement, more easy to understand, …) that fits less well but still well the data.

In order to do that, we will use other metrics described in the next sections that penalize the measure of fit.

3/ AIC - Akaike Information Criterion

The AIC metrics penalizes the measure of fit by taking into account the number of parameters included in the model:

$$AIC = -2\ x\ ll_{Model} + 2\ p \qquad (3.17)$$

Where p is the number of parameters included in the model. The second part of this equation aims at penalizing the evaluation of the model for models including a larger number of parameters. The smaller is the AIC metric, the better we will consider the model. The only purpose of the AIC is to compare two models.

4/ BIC - Bayesian Information Criterion

The BIC is a similar metric that penalizes the complexity of the model taking into account the number of parameters of the model and as well the size of the dataset on which the model is built upon:

$$BIC = p\ log(n) - 2\ x\ Lll_{Model} \qquad (3.18)$$

Where p is the number of parameters included in the model and n is the number of records included in the dataset on which the model is built upon.

Given the large datasets that are usually required to build GLMs, the BIC tends to penalize significantly more than the AIC the more complex models.

iii. How to compare different models?

The previous methodologies work well when we have access to the details of the GLM built. It might not always be the case (eg. We want to compare our GLM with a rating plan for which we do not have access to the detailed GLM behind as it is proprietary of another company). In this case, the methodologies presented in the previous sections will not work. In these cases, we can use one of the methodologies described below.

1/ The Gini curve

When assessing the accuracy of a model, we often speak about the lift of a model. The lift of a model is the ability of the model to charge the right amount of premiums for the different insureds' segments and therefore to prevent from adverse selection. Model lift is a relative measure: it only makes sense when comparing two models and not in absolute terms to evaluate a unique model.

The Gini index (who was initially created to evaluate the inequality of income distribution in a given population) can also be used to compare GLMs. The Gini Index will evaluate the ability of the model to segment the insureds from the best risk to the worst risk. In order to calculate the Gini index, we must classify our dataset from the best risk (eg. The lowest frequency of cancellation) to the worst risk (eg. the higher frequency of cancellation) and then:

- Report on the x axis the cumulative percentage of exposure

- Report on the y axis the cumulative percentage of loss

The obtained curve is called the Lorenz curve. The Gini index will be twice the area between the curve and the bisector. The Gini index thus estimates the discriminatory power of the model versus model that would be totally random.

2/ The ROC curves

While a logistic model will give the probability of a given event to occur, in reality, it is often easier to translate this probability into two categories:

- Yes: the event occurs
- No: the event does not occur

The Receiver Operating Characteristic (ROC) curve is a tool to measure the performance of a binary classifier. A binary classifier is a model that separates elements into two groups according to the characteristics of each element.

With this type of model, we can face four different cases:

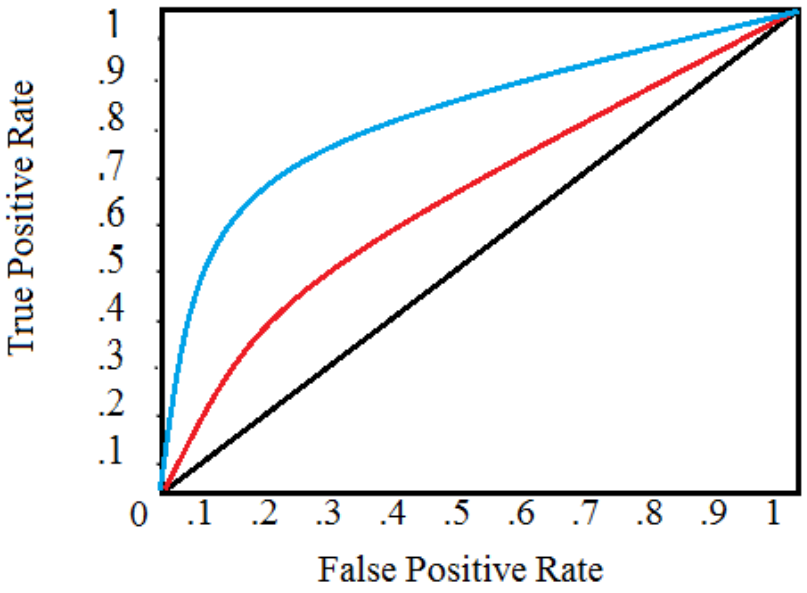
- True positive: the model predicts that the event occurs and indeed, the event occurs
- True negative: the model predicts that the event does not occur and indeed, the event does not occur
- False positive: the model predicts that the event occurs and, the event actually does not occur
- False negative: the model predicts that the event does not occur and, the event actually does occur

Starting from the probability of occurrence given by the predictive model, we will need to choose a probability (the discrimination threshold) above which we consider that the event occurs. When we select a lower discrimination threshold, it will translate into a higher number of true positive and fewer false negatives. For each discrimination threshold, we can then calculate two key metrics:

- the **sensitivity** (also called the true positive rate or the hit rate): this is the ratio between true positives and the total number of positives.
- the **specificity** (also called the false positive rate): this is the ratio between the true negatives and the total number of negatives. The complement of that ratio (1-Specificity) is called the false positive rate.

The ROC curve is the graphical representation of the false positives rate on the x-axis and the sensitivity on the y-axis. A model with a true predictive power will increase more quickly the sensitivity than the false-positive rate. The better will be the model's predictive power, the higher will be ROC curve: on the below "Figure 6", the blue curve has a better predictive power than the red curve.

Figure 10 - ROC curve illustration[15]



A metric linked to the ROC curve is the AUROC: the AUROC is the area under the curve. In case the model is on the equality line, the AUROC will be 0.5. There is a direct relationship between the Gini index and the AUROC and therefore, these two metrics should not be used in addition as they bring similar information on the predictive power of the model.

[15] Source : https://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/

3. Building a GLM on Travel Cancellation products

To the contrary of many other P&C covers, the Travel cancellation cover that we are analyzing in this thesis are not annual products. As explained earlier in this document, the exposure of the cancellation cover lasts between the date of subscription and the date at which the trip starts (for the rest of this thesis, let's call this time distance the "Booking Window"). The intuition is that the longer is the time between the date of subscription and the date at which the trip starts, the higher will be the risk (this has also been shown by other theses written on this topic where the Booking Window was one of the key variables used in the GLM of a cancellation cover).

There are several ways to take this specific feature at the time of modelling the cancellation frequency. As just mentioned, other theses on the topic used the Booking Window as a segmentation variable of their GLM. In this thesis, I suggest testing the following approaches and see which one is the most reliable one:

- Traditional approach for Cancellation cover: Build a GLM where we include the Booking Window as a segmentation variable, and we calculate the frequency as follows: Number of claims / Number of policies

- New proposed approach for Cancellation cover: Build a GLM where we offset the Booking Window and we calculate the frequency as follows: Number of claims / exposure, where:

  - Exposure of a given policy is "Start trip date – date of subscription"/365

A second element that we wanted to challenge with this thesis, is the way we usually model a cancellation cover within Europ Assistance[16]. Up to now, when applying GLM modelling to a cancellation cover, we were using the Poisson distribution to model the frequency. The Poisson distribution is indeed often used in P&C insurance to model the frequency but there is a fundamental difference between many classical P&C covers and a cancellation cover which is that once a customer has opened a cancellation claim, there is no other claim that can be opened on this policy. This indicates that the use of a binomial distribution would best work in this context. Therefore, we will model in this thesis with the two distributions, compare results and then, decide what to use for the rest of this work.

Finally, one of the big challenges of this thesis is that we address a portfolio of contracts which is heterogeneous. The portfolio is heterogeneous on several aspects: in terms of customers profiles, in terms of products and in terms of distribution channels.

As a consequence, one of the main value added of this thesis will be to identify the best possible way to deal with this heterogeneity and it will require to build several models and see what would best support our pricing work moving forward.

Indeed, in the past within Europ Assistance, a couple of GLMs were built on the cancellation products but each of these models were built based on the data of a given large partner. It was thus not dealing with the overall heterogeneity of our portfolio and the application of these models was thus mostly

[16] J. QIU *Travel pricing sophistication with GLM and Machine Learning approaches.* 2017. Mémoire EURIA

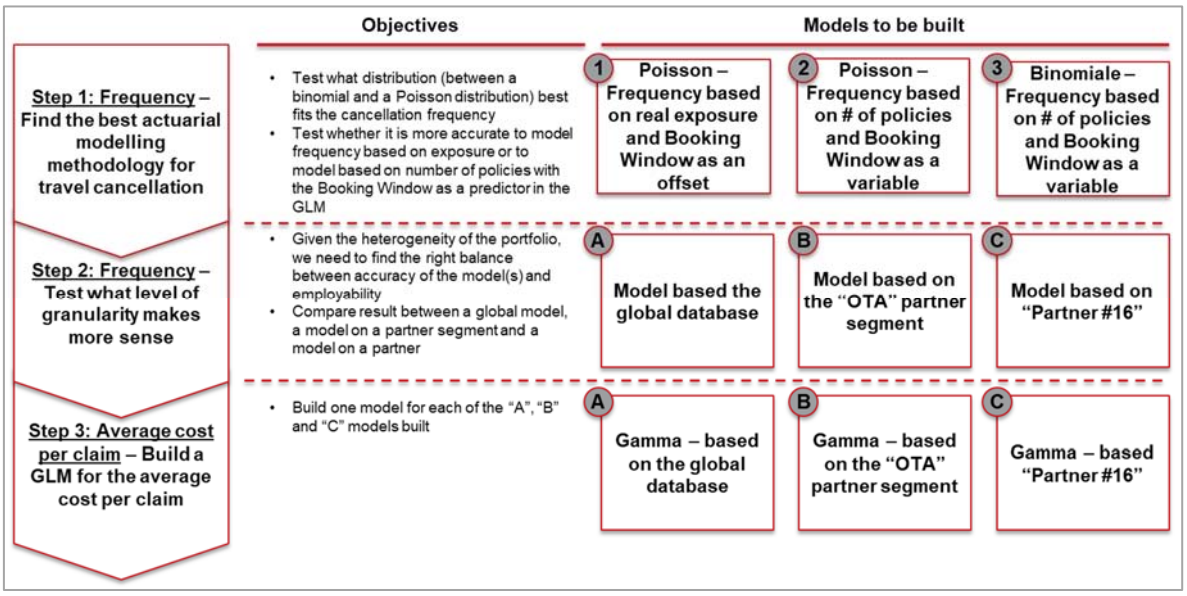
restricted to one partner or a limited number of very similar partners (for which we assume that the customers profiles are similar).

With this thesis, the objective is to build one model or a set of different models in order to increase the usage of GLMs in our pricing work. Indeed, as of today, we have one GLM available that is based on one partner and this GLM cannot be used on all partners or customer segments.

Dealing with this heterogeneity will require some extensive work on data cleaning and enrichment in order to get a set of comparable data on several partners. The next section will explain the different steps I have gone through in order to get to a clean and enriched set of data.

Combining these different elements that we would like to test/challenge in thesis, I propose to structure the modelling path as follows:

Figure 11 - Proposed approach for the modelling steps



a. Data collection, preparation and Validation

i. Context and perimeter of the analysis

The definition of the perimeter to be addressed in this thesis has been very important as I wanted at the same time to be as reliable as possible (which has required us to focus on some areas) and as broad as possible (in order to embrace as much as possible the diversity of our portfolio). In the below paragraphs, you will find the dimensions along which I have narrowed the perimeter of the analysis as well as the rationale for this choice

1/ Focus on B2B2C partnerships

As explained earlier in this document, we can distinguish two main distribution channels through which travel insurance products are sold: the B2C channel and the B2B2C channel.

Across those two different distribution channels, the products sold can be very different: the main difference in terms of products sold is that, on the B2B2C channel, mostly Short-Term single-trip policies are sold whereas on the B2C channel, mostly annul multi-trip policies are sold.

Also, as mentioned earlier as well, the B2B2C channel is by far the biggest distribution channel for the travel insurance companies. It represents 90-95% of the policies sold by Europ Assistance.

In order to ease the building process of the model, in order to be as representative as possible and in order to make the models as reliable as possible, I have thus made the choice to focus on the B2B2C distribution channel.

2/ Focus on International contracts

Within Europ Assistance, we make a difference between international contracts on one side and local contracts on the other side. International contracts refer to a global agreement signed with a travel partner and that will cover insureds from various countries whereas local contract will refer to an agreement signed with a local partner in order to cover insureds from one unique country.

For this exercise, I have made the choice to focus on International contracts for the following reasons:

- Existence of good and homogeneous (in the sense of having consistent variables available) database for this perimeter of contracts (we will come back more precisely, in the next chapter, to the qualification of this database)
- The portfolio of International contracts is growing within Europ Assistance: more and more travel partners are international companies or young local companies with plans on the very short term to become international. Only some key partner segments remain mostly local such the banks or some travel agencies networks
- Not all our Europ Assistance local entities have the capabilities to develop very detailed segmented pricings. It is thus beneficial for the company that we cover as many geographies as possible in our database so that our model can then be used by as many entities as possible (indeed, we have already observed different sinistrality levels from one geography to another one).

3/ Focus on pre-Covid trips

As explained earlier, the Covid19 crisis has deeply altered the travel insurance industry. At the time of writing this thesis, I did not have enough perspective on the impact of the crisis on our technical results, neither sufficient data to analyze.

The models developed in this thesis will thus probably not be applicable on the very short term (as long as the Covid Crisis affects travel behaviors) but should be applicable when we are back to normal, once the Covid19 crisis is terminated (for the travel industry, this will probably not be the case before 2022).

For all the policies sold, we have in the database various dates available:

- pol_dt_polSubscribed: date at which the policy was purchased by the customer
- pol_dt_extract: date at which the policy data has been uploaded into our database
- pol_dt_tripStart: date at which the customer's trip start
- pol_dt_tripEnd: date at which the customer's trip end

Almost all these dates (all except the "pol_dt_extract") will be needed in the modelling part.

For the selection of the perimeter to be addressed in this thesis, I selected all the policies with a "pol_dt_tripStart" prior to 31st of December 2019 in order to not distort the claims experience with effects related to the Covid19 crisis. This restricts the perimeter that can be analyzed but at the same time, increases the chances to build an accurate model.

4/ Bring partner diversity in the database

As outlined in the introduction of this thesis, one of the big objectives of this thesis is to understand the impact of the partner's segment on the sinistrality of our products and have models that can be more widely used than our existing model. It was thus very important to bring as many different partners as possible in the database in order to try and understand the impact of the partner segment (or any other variable) on the sinistrality.

Previous models made internally within Europ Assistance were mostly focusing on one partner or on a couple of partners and, therefore, did not bring that learning.

ii. Database used

Each entity within the Europ Assistance Group has its own databases in which policies and claims data of local contracts are stored. At Europ Assistance Group level, we have built over time central databases in which we store claims and sales data for International contracts. In these central databases, there are mostly three databases that are of interest for us.

1/ Sales data – "TRT_Sales"

This database includes the line by line policy data with many fields available to describe the detailed circumstances in which the policy was sold.

The main segmentation variables available in this database that we will keep in the modelization exercise will be the followings:

- ***pol_client***: name of the partner through which the policy has been sold to the final customer
- ***pol_dt_polSubscribed***: date at which the policy was subscribed
- ***pol_dt_tripStart***: date at which the trip for which the policy has been bought, starts
- ***pol_dt_tripEnd***: date at which the trip for which the policy has been bought, ends
- ***pol_countrySubcr***: country in which the policy has been subscribed (country of residence of the policy subscriber)
- ***pol_amtTripSold***: cost of the trip insured
- ***pol_numBenef***: number of travelers included in the policy
- ***pol_travel_duration:*** difference between the "pol_dt_tripEnd" and the "pol_dt_tripStart"

- ***pol_booking_window_subscri***: difference between the "pol_dt_tripStart" and the "pol_dt_polSubsribed"

2/ Claims data – "TRT_Claims"

TRT_Claims includes the line by line claims data with all the fields required to understand the type of claim, when the claim was opened/paid/…, the amounts paid and still reserved, the status of the claims, …

3/ Product information – "TRT_Product_Description"

TRT_Product_Description gives for each product the main features of the product, but mostly financial information related to the product (such as Gross Premium, IPT, Partner commissions, dates of validity of the product, …). This database, as it will be explained later, does not include all the data that would be required and I had to populate from other sources of data, additional information that were needed to perform the modelization.

These databases were used as the source of information for this pricing modelling exercise.

iii. Scope selection

1/ Scope selection based on sales data quality

In the database used, all policies sold are assigned a "product_code" (called "PPUC") and, in the "TRT_Product_description" database, there is a set of important information on the product that is given such:

- Tax rate to be applied
- Commission rate to be paid to the partner
- Commission rate to be paid to the broker
- Split of the premiums per SII Line of Business
- …

In the database, the initial number of products available on the perimeter selected was 1276 PPUCs. Some of these products were excluded as they were missing some key information:

- PPUC is null ➔ 3 PPUCs excluded
- Gross Written Premium incl. tax is null ➔ 15 PPUCs excluded
- Gross Written Premium excl. tax is null ➔ 45 PPUCs excluded

Then, as I was focusing on the cancellation cover only, I had to identify among all those products the ones that include a travel cancellation cover. Nevertheless, I did not have in the "TRT_Product_description" database a field that allowed me to identify the products that include a travel cancellation cover versus the others.

One information in this database could be used to partially identify the products that include a cancellation cover: this information was the split of the premium by SII Line of business (in which trip cancellation was separated from "Miscellaneous financial loss" which is the SII Line of business to be used for trip cancellation). Nevertheless, this information was not systematically filled in. When the information was not filled in, I proceeded as follows:

- Identification of an identical product sold in another geography in which the split of premiums by SII Line of Business was available
- Read of the Terms and Conditions of the product in order to see whether the product includes or not a trip cancellation cover

This methodology allowed me to correct the split of Premiums by SII Line of Business for 241 PPUCs that I was then able to keep in the database.

As a summary, my initial dataset was including 1276 PPUCs and the final scope that will be included in the database for the modelling will include 795 PPUCs.

Figure 12 - Waterfall from the initial number of products available to the final number of products kept in the database

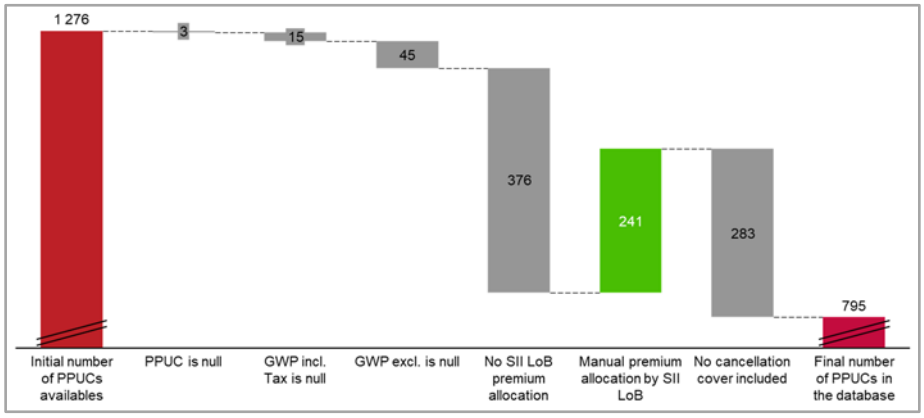
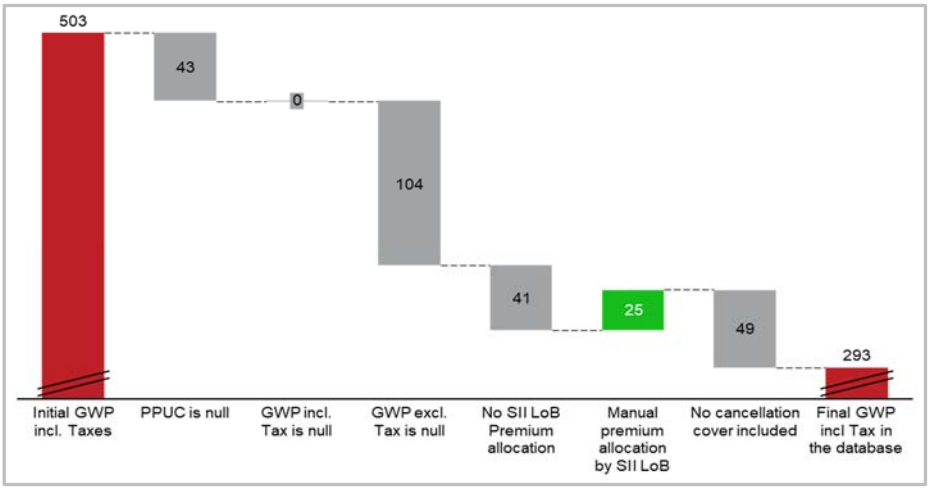


Figure 13 - Waterfall from the initial amount of Gross Written Premiums (excl. taxes) to the final amount of Gross Written Premiums kept in the database (in M€)



2/ Scope selection based on claims data quality

In terms of data quality issues to be addressed, the main issue was regarding the lack of unique ID between the sales database and the claims database. Indeed, not in all cases, we had a unique ID that allowed me to link a claim to the right policy.

Another situation that I faced was that, in some cases, there was a common identifier between the sales and the claims databases but, due to the scope selection that I had to do regarding the sales (see previous paragraph), I had to remove the policy from the database and then, the claim could not be merged with the right policy.

I have looked at these topics on a partner by partner basis and the situation is summarized in the table below. In the table below, the names of the partners have been anonymized as it represents confidential information that Europ Assistance do not want to disclose.

In the below table[17], I have highlighted 4 different situations:

- Pink partners: are partners for which we were not able to merge any claims available in the claims database
- Yellow partners: are partners for which the percentage of not merged/mismatched claims is higher than 20%
- Blue partners: are partners that were excluded from the sales database for data quality reasons (see previous paragraph)
- Red partners: are partners that were not excluded from the sales database but for which we do not have claims in our claims database

[17] In this table, the names of the different partners have anonymized for confidentiality reasons. One given partner will keep the same "anomymized name" throughout the thesis.

Table 3 - Claims data quality summary per partner

cla_clientGroup	in_sales	N_merged	N_mismatch	N_notmerged	N_total	N_pct_missing	Incurred_merged	Incurred_mismatch	Incurred_notmerged	Incurred_total
Partner #1	1	0	52	0	52	100%	0	35 432	0	35 432
Partner #2	1	135	13	0	148	9%	214 972	1 151 392	0	1 366 364
Partner #3	1	72	0	0	72	0%	36 799	0	0	36 799
Partner #4	NA	0	39	0	39	100%	0	42 878	0	42 878
Partner #5	1	28	3 445	584	4 057	99%	30 029	2 968 596	501 307	3 499 932
Partner #6	NA	0	4	3	7	100%	0	427	1 256	1 683
Partner #7	1	746	309	75	1 130	34%	713 216	326 471	47 827	1 087 514
Partner #8	1	1 861	103	0	1 964	5%	29 420	25 795	0	55 215
Partner #9	1	296	5	0	301	2%	848 633	15 236	0	863 869
Partner #10	1	0	548	660	1 208	100%	0	344 432	385 092	729 524
Partner #11	1	19	629	0	648	97%	12 372	361 891	0	374 264
Partner #12	NA	0	0	1	1	100%	0	0	4 095	4 095
Partner #13	1	42	30	55	127	67%	28 754	26 582	39 054	94 390
Partner #14	NA	0	0	23	23	100%	0	0	37 157	37 157
Partner #15	1	10 460	1 318	7	11 785	11%	4 374 145	613 636	5 302	4 993 083
Partner #16	1	26 617	3 625	444	30 686	13%	11 381 223	1 382 217	240 664	13 004 103
Partner #17	1	5 555	376	0	5 931	6%	2 934 575	198 733	0	3 133 308
Partner #18	1	7 067	129	0	7 196	2%	1 787 426	26 386	0	1 813 812
Partner #19	1	5	0	0	5	0%	985	0	0	985
Partner #20	1	1 626	1 254	0	2 880	44%	631 202	494 487	0	1 125 689
Partner #21	1	694	9	0	703	1%	312 469	4 267	0	316 735
Partner #22	1	403	3	0	406	1%	180 869	1 517	0	182 387
Partner #23	1	15	0	2	17	12%	11 759	0	2 917	14 676
Partner #24	1	10	1	0	11	9%	4 009	935	0	4 944
Partner #25	1	157	4	0	161	2%	100 194	2 030	0	102 224
Partner #26	1	747	84	53	884	15%	561 504	51 667	0	613 171
Partner #27	1	4 021	78	0	4 099	2%	4 233 320	99 891	0	4 333 211
Partner #28	NA	0	0	71	71	100%	0	0	28 197	28 197
Partner #29	NA	0	77	0	77	100%	0	8 901	0	8 901
Partner #30	1	1	195	0	196	99%	1 024	152 789	0	153 813
Partner #31	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Partner #32	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Partner #33	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Partner #34	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Partner #35	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Partner #36	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
Total	**30**	**60 577**	**12 330**	**1 978**	**74 885**	**1331%**	**28 428 897**	**8 336 589**	**1 292 867**	**38 058 353**

This table can be summarized as follows:

Table 4 - Claims data quality summary

Status	Claim number	Incurred Trip Cancellation (€)	Claim number %
merged	60 577	28 428 897	81%
mismatch	12 330	8 336 589	16%
notmerged	1 978	1 292 867	3%
Total	**74 885**	**38 058 353**	**100%**

Overall, from “table 4”, we can see that we were able to link 80% of the claims to the right policy. From “table 3”, we see that the quality of this data varies significantly from one partner to another.

I have excluded from my database all the “pink”, “yellow”, “blue” and “red” partners and, therefore, our scope of analysis will cover 16 partners instead of the 36 partners available initially in the database. This is a significant scope reduction compared to the data that was initially available in the databases but I believe that this scope reduction is needed in order to ensure quality of the modelling that will be done in this work and it still represents a wide variety of partners included in the database compared to previous models made internally.

Even for the partners kept, not all the claims are merged and thus, not all the claims will be taken into account into our modelling process. As a consequence, once the different models are built, we should not forget at the end to load our pure premiums by the percentage of claims that were not taken into account in the modelling process as we were not able to attach them to the right policy.

As a summary of scope selection based on sales and claims data quality, we had to reduce significantly the scope of analysis compared to what was initially available but we do believe that the variety of partners included in the database will still allow to draw learnings compared to our existing models.

Figure 14 - Summary of the scope selection based on sales & claims data quality - PPUC waterfall

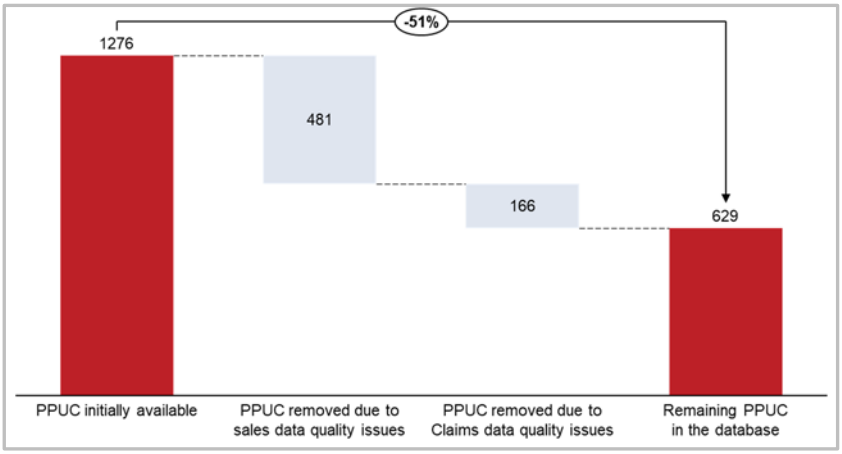
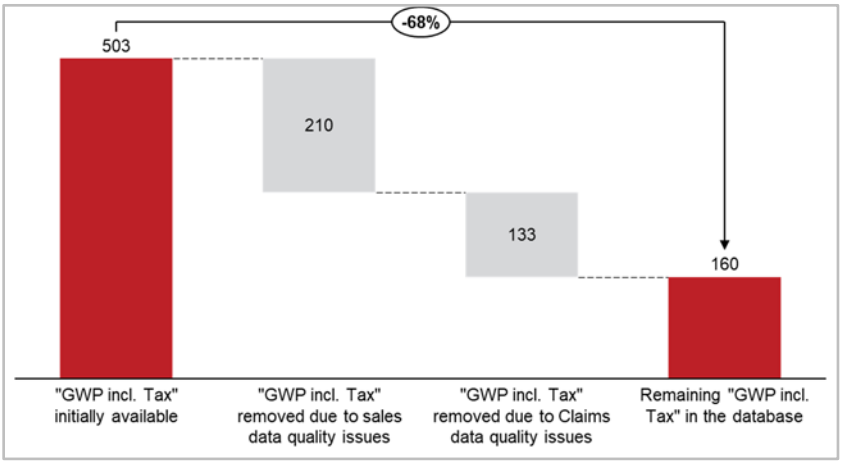


Figure 15- Summary of the scope selection based on sales and claims data quality – GWP waterfall (in M€)



iv. Data completeness of the different variables

I have explained earlier in this document the different variables that we were planning to keep in the database in order to do the modelling exercise. Now that we have clearly defined the scope of partners and products that we wanted to keep in our database, it is time to look at the quality of this data in order to make sure that data completeness is ok for the different fields.

In order to identify potential data incompleteness issues on some variables, I have counted, for each variable, and per partner, the number of "NA", of "0" and of negative values (as all the variables included in the database should have positive values). In the table below, you will find the result of this analysis.

Table 5 - Data completeness analysis by variable

pol_client	Number policies	Value NA						Value 0						Value Negative					
		pol_travel_duration	pol_countrySubcr	pol_amtTripSold	pol_numBenef	benef_numBenef	pol_booking_window_subscri	pol_travel_duration	pol_countrySubcr	pol_amtTripSold	pol_numBenef	benef_numBenef	pol_booking_window_subscri	pol_travel_duration	pol_countrySubcr	pol_amtTripSold	pol_numBenef	benef_numBenef	pol_booking_window_subscri
Partner #2	36 136	0	0	0	0	0	0	0	0	35	0	0	0	0	0	0	0	0	7
Partner #3	19 002	0	1	0	0	41	0	0	0	16	4	0	282	0	0	0	0	0	9
Partner #8	109 760	0	0	0	0	0	0	0	0	0	32	0	98	0	0	0	0	0	6
Partner #9	23 056	0	0	0	0	0	0	0	0	1	0	0	13	0	0	0	0	0	27
Partner #15	2 215 902	0	0	0	0	3 146	0	0	0	238 093	0	0	186 859	0	0	1	0	0	13 331
Partner #16	3 494 767	0	49	0	0	3	0	0	0	11	0	0	49 199	0	0	0	0	0	69
Partner #17	555 781	0	0	0	0	0	0	0	0	2 869	0	0	1 529	0	0	1	1	0	54
Partner #19	1 816	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Partner #18	2 885 138	0	0	0	0	80 893	0	0	0	2 885 138	0	0	38 047	0	0	0	0	0	208
Partner #21	44 637	0	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	39
Partner #23	716	0	0	0	0	3	0	0	0	289	11	0	2	0	0	0	0	0	0
Partner #24	3 026	0	0	0	0	390	0	0	0	2 636	0	0	48	0	0	0	0	0	0
Partner #25	20 623	0	0	0	0	0	0	0	0	0	0	0	170	0	0	0	0	0	176
Partner #22	14 999	0	0	0	0	0	0	0	0	4 769	0	0	4	0	0	0	0	0	0
Partner #26	57 930	0	0	1	0	4 807	0	0	0	0	0	0	30	0	0	0	0	0	1
Partner #27	200 665	0	0	0	0	9 181	0	0	0	200 665	393	0	5	0	0	0	0	0	4
Total	**9 683 954**	0	**50**	**1**	0	**98 464**	0	0	0	**3 334 522**	**440**	0	**276 314**	0	0	**2**	**1**	0	**13 931**

Based on this table, the major concern in terms of data completeness was regarding the variable "pol_amtTripSold".

1/ Pol_amtTripSold

The "pol_amtTripSold" is the value of the trip insured. For some partners, the "pol_amtTripSold" information is not populated in the database : as seen in "Table 7", this represents almost a third of the policies underwritten. The reason for that is that some of our partners are not able to provide us with this information when they send us the policy data in their regular batch of data.

The "pol_amtTripSold" is nevertheless a key information for us from a technical point of view as:

- The value of the trip insured is obviously a key driver of the average cost of a cancellation. Indeed, the amount of a cancellation claim is mainly driven by
    1. The value of the trip insured
    2. The date at which the customer cancels the trip: this has an impact on the share of the cost of the trip already paid by the customer (and a travel insurance product only covers the financial loss incurred by the customer, not the total amount of the trip)
    3. The potential application of the excess defined in the insurance policy
- The value of the trip insured is a good metric to have in the database in order to:

1. Identify anomalies (eg. Cases where the cancellation claim amount would be higher than the "pol_amtTripSold")
2. Calculate the ratio "average cancellation claim amount/average pol_amtTripSold" which is a key assumption taken at the time of the pricing by the underwriter and as well a good indicator to follow up and identify potential changes in the penalties schemes applied by the partner

When the data "Pol_amtTripSold" was missing in the databases, I applied three different methodologies:

- **Methodology 1** – The "Pol_amtTripSold" is a variable used to calculate the insurance premium
    1. In this case, most of the time, the premium grid is composed of different layers of "cost of trip".
    2. Based on the premium paid by the customer (concatenated with other information such the country of subscription of the insurance product bought), I was able to retro-engineer an approximate of the "Pol_amtTripSold" (taking the midpoint between the bottom and the top limits of the price bands).
    3. This is not perfect as it does not fully capture the variability of the variable "Pol_amtTripSold" for these partners but it was still better than putting totally aside these partners.
    4. This methodology was mostly used for "Partner #27".
- **Methodology 2** – The "Pol_amtTripSold" is not a variable used to calculate the insurance premium
    1. This is not the most common case for products that include a trip cancellation cover and, in these cases, the price is most of the time dependent on the length of the trip.
    2. In this situation, I was not able to retro-engineer a different "Pol_amtTripSold" by policy but I calculated an average cost of cancellation claim for each of these partners that I used as a proxy of the "Pol_amtTripSold" for all the policies of this partner.
    3. This is even less ideal than *"Methodology 1"* but allowed me to keep in the database some partners that were very useful for the rest of the variables.
    4. This methodology was mostly used for "Partner #18".
- **Methodology 3** – The "Pol_amtTripSold" is not a variable used to calculate the insurance premium and the number of policies is lower than 5000 policies
    1. In this case, I have replaced the "Pol_amtTripSold" by the average "Pol_amtTripSold" for all the policies for which the information was available
    2. This is, of course, the least ideal methodologies among the three methodologies but it has allowed me to keep these accounts and leverage available data for the other variables as well as the available data for "Pol_amtTripSold" for these accounts

2/ Other variables

For the rest of the variables, as the quality was rather good, I took a simpler approach as the added value for our modelling exercise was minimal:

- "pol_countrySubcr" = "NA": in this case, I have removed these policies from the database.
- "pol_numBenef" = "0": I have looked at the other variable "benef_numBenef". In case, it was different from "NA", I have replaced it by the number put for the variable "benef_numBenef".
- "pol_booking_window_subscri" is negative: this should not happen as the policy should always be subscribed before the departure date. Given there was a limited number of policies in this situations (and that the number of claims attached to these policies was also very limited, only 11 claims), I have removed these policies from the database

With these policies removed from the database, the number of policies on which we will build our GLMs has decreased from 9 683 954 policies to 9 669 972 policies.

v. Enrichment of the data with new variables

1/ Characteristics of the cover

The portfolio of international contracts included in my database is relatively heterogeneous in terms of cancellation products. Indeed, this portfolio of contracts include contracts that are international (as explained earlier) but each of these contracts were initiated by a given Europ Assistance entity. Each Europ Assistance entity has over time developed different products. Differences across the products mainly revolve around the following aspects:

- Different levels of excess
- Different levels of limit
- Different sets of cancellation reasons included in the cover

In order to capture part of the heterogeneity of the portfolio (and as this information was not captured in a proper product database as explained earlier), despite the initial idea was to add these three additional features (excess, limit and type of cancellation product), we only added in our sales data:

- Type of cancellation product: Cancellation for listed causes ("CFLC") vs Cancellation for Any Justified Reasons ("CFAJR")

Indeed, we did not add the two others for different reasons:

- The level of excess was not added as the way it is formulated in the different products made it quite complicated to summarize it into a couple of additional fields as:
    1. In some cases, it is expressed in percentage of the trip cost, with or without a minimum amount
    2. In other cases, it is expressed in absolute value
    3. In other cases, there are different levels of excess depending on what type of cancellation cause you use (medical vs professional vs …)

We will see later in this thesis whether the variable "type of cancellation product" has an impact on frequencies and/or average cost.

2/ Type of partner

One of the objectives of this modelling exercise was to capture as much heterogeneity as possible given the variance of claims experiences that we observe across different partners. I have thus created two additional variables to capture the type of partners through which our products are distributed:

- **Partner segment**: Here the objective is to capture the type of trip that is sold as the assumption is that it could have an impact on the loss experience. This type of variable was not used in previous models used within Europ Assistance on the European perimeter. The different types of partner segments are listed below
    1. Airlines
    2. Cruises
    3. Online Travel Agencies
    4. Tour Operators

- **Client target:** For a given "partner segment", the type of trip sold might be very different in terms of costs for the final customer. I have thus created a new variable to reflect this specific feature.
    1. In order to build that additional feature, I have calculated for each of the partners, the following KPIs:
        - Average Amount trip sold
        - Average amount trip sold per beneficiary
        - Average trip sold divided by the trip duration
    2. These different KPIs were useful. Indeed, as we have partner that sell "flights-only" and others that sell package "flights + hotels", the meaning of the KPI "Average trip sold divided by the trip duration" would not be the same for all partners.
    3. Based on these three KPIs, I have then grouped them into three categories of partners[18]
        - Low-Cost: 4 partners included
        - Medium: 7 partners included
        - Premium: 5 partners included

vi. Treatment of large claims

When building GLMs, it is usually convenient to exclude from the database large losses in order to not disturb the modelization process by these large losses that are hardly modelled by a GLM.

In travel insurance, large losses are often observed on medical assistance cases, but we do rarely speak about large losses in the case of Travel cancellation products. Indeed, as explained earlier, the cost of the cancellation insurance products is most of the time based on the cost of the trip insured and the expected claim amount is directly set by the cost of the insured trip. Also, as it will be seen in the next section on exploratory analyses, the frequency tends to be higher on higher trip costs.

Also, as it can be seen in the following table, there is not a lot of extreme values in terms of claims cost.

Figure 16 - Claim cost : quantiles analyses

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1,00	165,40	315,90	495,20	592,90	16 620,00

90%	91%	92%	93%	94%	95%	96%	97%	98%	99%	100%
1 062,9	1 119,0	1 184,9	1 262,5	1 359,8	1 495,3	1 619,5	1 829,1	2 159,5	2 923,2	16 620,0

99,0%	99,1%	99,2%	99,3%	99,4%	99,5%	99,6%	99,7%	99,8%	99,9%	100,0%
2 923,2	3 032,6	3 147,5	3 345,1	3 511,5	3 770,1	4 083,3	4 537,9	5 163,9	7 348,6	16 620,0

Finally, given the nature of the travel industry, defining a threshold even high would translate into removing a higher share of claims on some partners than others as some partners are more focused on Premium trips versus others that are focusing on low-cost trips.

As a consequence, I decided to not define a large loss threshold above which we should exclude claims from the claims database. All the claims were kept in the database in order to not distort the analysis.

[18] See appendixes for detailed table of this classification

b. Exploratory data analyses

i. Global overview of the database

As explained in previous sections, we had to do an extensive data selection/cleaning work to get to a clean set of data. This database is finally composed as follows:

- 9 669 972 Short Term policies
- Exposure: 1 655 114 "annual policies"[19]
- 159,7 M€ of GWP including taxes
- These policies are sold through 16 different partners
- 54 536 claims taken (either paid or opened but still in reserve) on these policies
- The corresponding claims amount is: 27,0 M€

The overall key technical metrics can be summarized as follows:

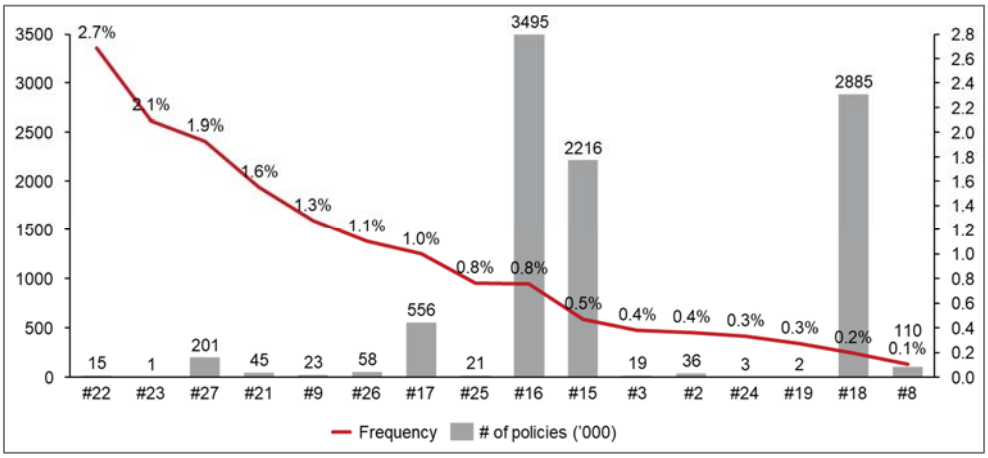
Table 6 - Key Technical metrics of the overall database

	Frequency	Average Cost (€)	Pure premium (€)
Based on number of policies	0,56%	495€	2,79€
Based on exposure	3,29%		16,32€

As already explained, one of the key challenges to be addressed in this thesis is the heterogeneity of sinistrality that we observed across our different partners. This heterogeneity can easily be seen in the following two charts that show the average cancellation frequency per partner and the average cost per claim per partner.

[19] As explained earlier, exposure is the transformation of the Short-Term policies into an equivalent of "annual policies". For example, a trip cancellation cover bought on the 1st of January with a departure date on the 1st of April will count for 0,25 in terms of exposure

Figure 17 - Average cancellation frequency (# of claims/# of policies) per partner

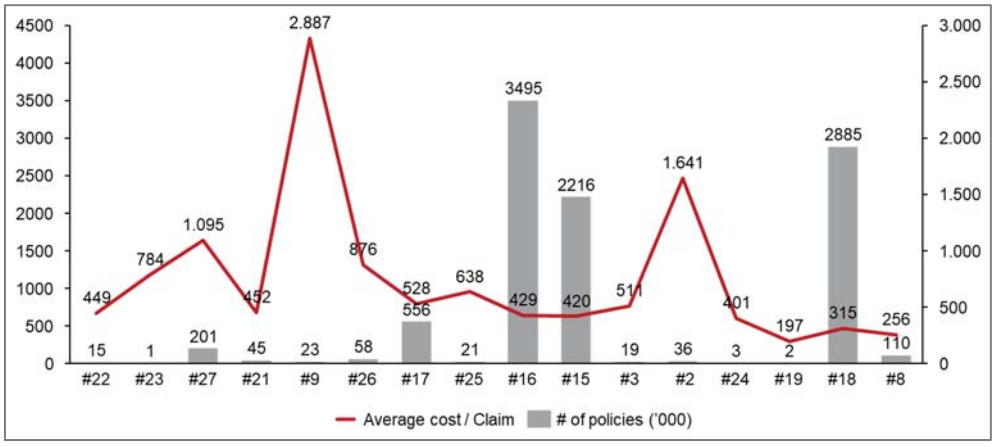


Two main learnings can be drawn from “Figure 12" on the observed frequency (# of claims / # of policies):

- Even if this is not really what we wanted to show with this chart, we can see from this chart that some of the partners concentrate a big portion of the policies sold. It would have been better to have a more balanced portfolio from that point of view, but this is the best compromise I was able to get from the data cleaning exercise.

- In terms of frequency, we see that the observed frequency goes from 0,1% to 2,7%: it is a “1 to 27” difference. This emphasizes clearly the need of understanding underlying elements that explain these differences across partners. Existing models within EA were mostly based on “Partner #16” which even if close to the average does not represent the full variety of the portfolio. We will try in the rest of this thesis to explain as much as possible these differences in terms of observed frequency by partner.

Figure 18 - Average cost per claims (Claims amount/# of claims) per partner



On “Figure 19”, I wanted to make three comments:

- Firstly, as for the frequency, we observe a significant variance in the average cost of claims per partner, as it goes from 197€/claim for “Partner #19” to 2887€/claim for “Partner #9”. This gap is more easily understandable by the fact that trips sold by these partners are very different from one to another one as the variety of trips covered goes from “one-way ticket sold by a low-cost airline” to the “Deluxe several weeks travel package sold by an Online Travel Agency”.

- Secondly, I have on purpose kept the same order of the partners between “Figure 12” and “Figure 13” to highlight that there was no evident relation between the average frequency observed on a given partner and the average cost of claims. This is important as, later, we will model separately frequency and average cost. We thus need to assume that frequency and average cost are independent variables

- Finally, I wanted to mention that, despite both frequency and average cost/claim vary significantly from one partner to another, the one on which the biggest modelling effort should be made, is the frequency. Indeed, as we will see later, the main driver that explains the average cost of claim is the value of the trip insured and it is common market practice to have the price of the cancellation insurance dependent on the value of the trip insured.

ii. Exposure vs number of policies

As explained earlier in this document, given travel insurance products in this database are not annual products, there are two ways we could calculate the exposure and the frequency in such portfolios. One of the contributions of this thesis will be to look whether, from a modelling point of view, one way is better than the other. There are two metrics that can be used to calculated.

Table 7 - Two ways to calculate exposure and frequency with non-annual policies

	Exposure	**Frequency**
Classical approach	# of policies	# of claims / # of policies
Proposed new approach	“Real exposure”: each policy counts for the portion of the year insured, eg. From the subscription date of the policy to the departure date	# of claims / “Real Exposure”

This will have an impact on the variables that we should consider in our modelling as it will be illustrated by the two following charts.

Figure 19 - Univariate analysis – Observed frequency (based on "Real exposure") by Booking Window[20]

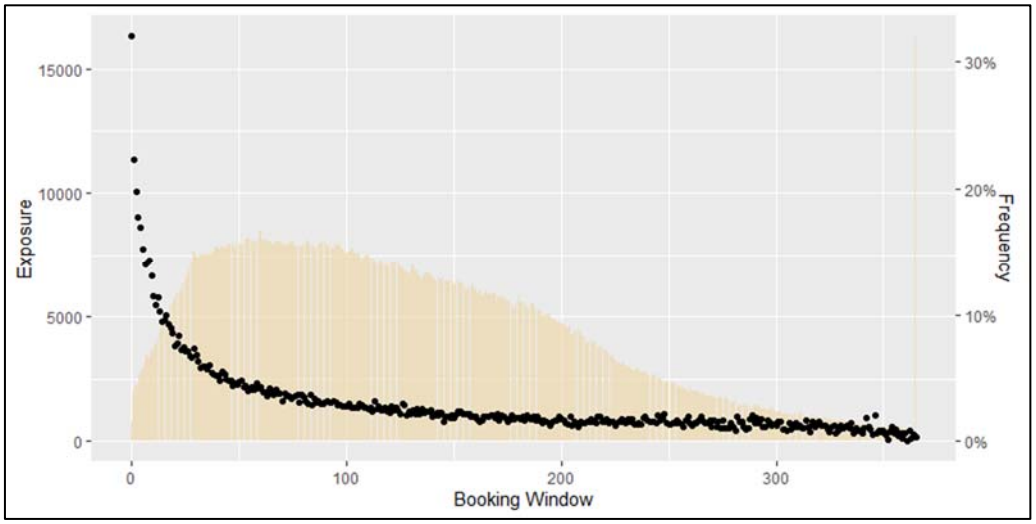
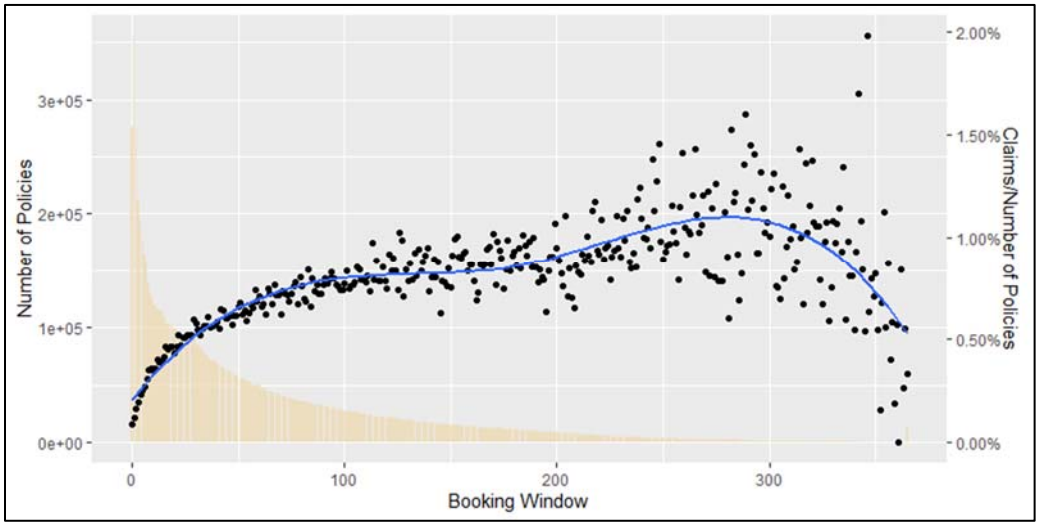


Figure 20 - Univariate analysis – Observed frequency (based on number of policies) by Booking Window



The first chart indicates that the frequency tends to decrease when the Booking Window increases whereas the second chart indicates that the frequency tends to increase when the Booking Window increases. The intuition would be to agree with "Figure 21" as the longer is the insured period, the higher is the risk to open a claim. Looking at extreme cases, it is very unlikely that someone booking a trip only three days in advance will have to cancel that trip whereas, it is much more likely for someone who would book several months in advance.

This indicates that if we use the "real exposure" to measure the frequency, we cannot have the Booking Window in the model as this variable is already taken into account into the way the frequency is calculated. What "Figure 20" tells us is that when the Booking Window increases, the frequency

[20] The Booking Window is the time distance between the subscription date and the date of start of the trip

increases to a lower extent than the Booking Window: this is why we see a negative trend for the frequency line on “Figure 20”.

iii. Univariate analyses

Before entering the modelling steps, it is important to get more familiar with the dataset. This is a way of checking quality of data but mostly a way to start understanding what will matter in terms of segmentation variables. It is important to check whether the results obtained in these univariate analyses are in line with initial assumptions we could make on it.

In the following sections, we will look at different variables that could impact the risk in order to get a feeling on whether they could be potential predictors in our models. For each of these potential predictors, we will look at the relation between the variable and both the frequency and the average cost per claim.

1/ By number of policy beneficiaries (“pol_numBenef”)

The number of policy beneficiaries is the number of persons who are booked on the trip. It can be seen as the number of insured persons. They are most of the time the number of persons who will be part of the trip.

Figure 21 - Univariate analysis – Observed frequency by number of policy beneficiaries

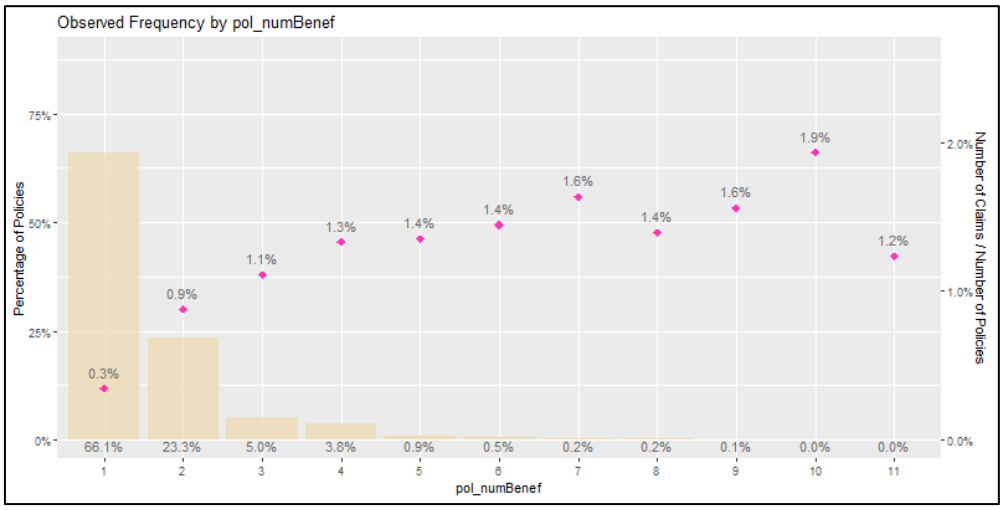
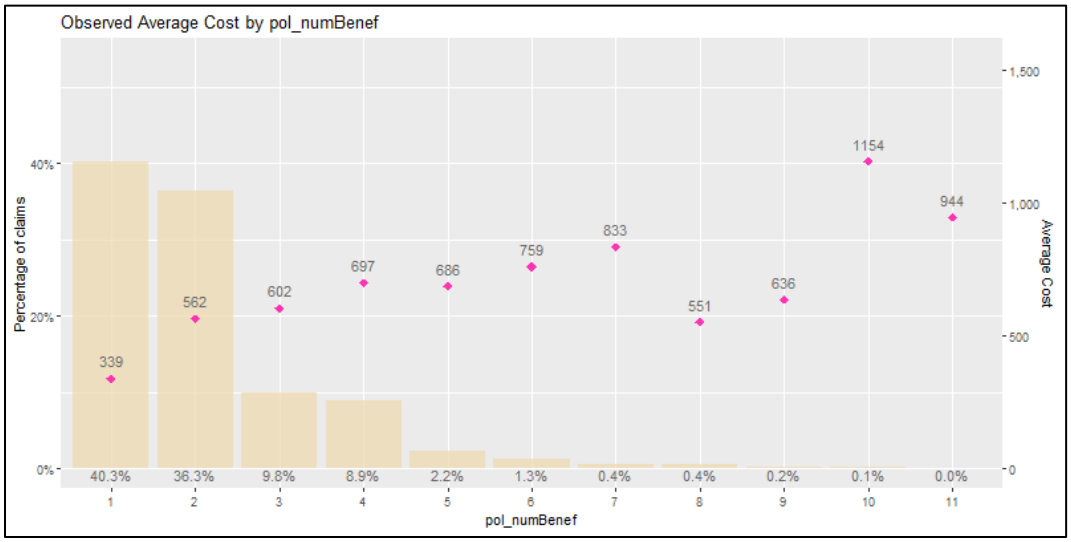


Figure 22 - Univariate analysis – Observed average cost/claim by number of policy beneficiaries



As it could have been anticipated, we see from "Figure 22" that the number of beneficiaries on a policy has a positive relationship with the frequency. This can easily be understandable as the more beneficiaries on a policy, the higher is the chance that someone gets ill or needs to cancel for professional reasons or … Nevertheless, what is surprising is that when the number of beneficiaries is doubled (from "1" to "2"), the observed frequency is almost tripled : this is probably explained by other predictors that are not captured in this univariate analysis and this highlights the limit of univariate analysis.

Looking at "Figure 23", this is again consistent with what could have been expected as we see a positive relation between the number of beneficiaries and the average cost per claim. This can be easily understood by the fact that the higher is the number of beneficiaries, the higher will be the value of the trip insured and hence the cost of the claim.

2/ By partner segment

The partner segment describes the type of partner through which the policies are sold. In our database, we have 5 types of partners: Online Travel Agencies ("OTAs"), Airlines, Cruises, Resorts and Others. This "Others" category was created for one partner that could not really fit in any of the other categories.

Figure 23 - Univariate analysis – Observed frequency by partner segment

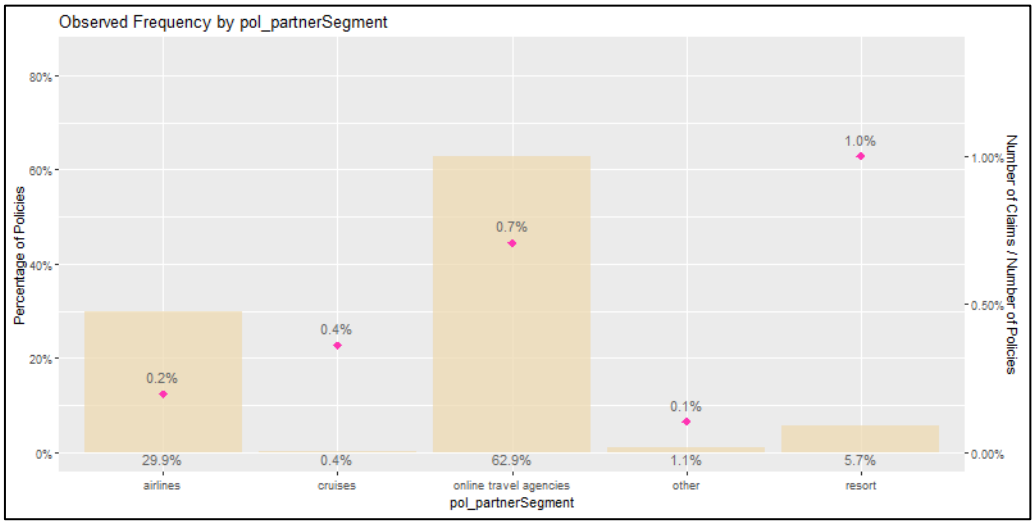
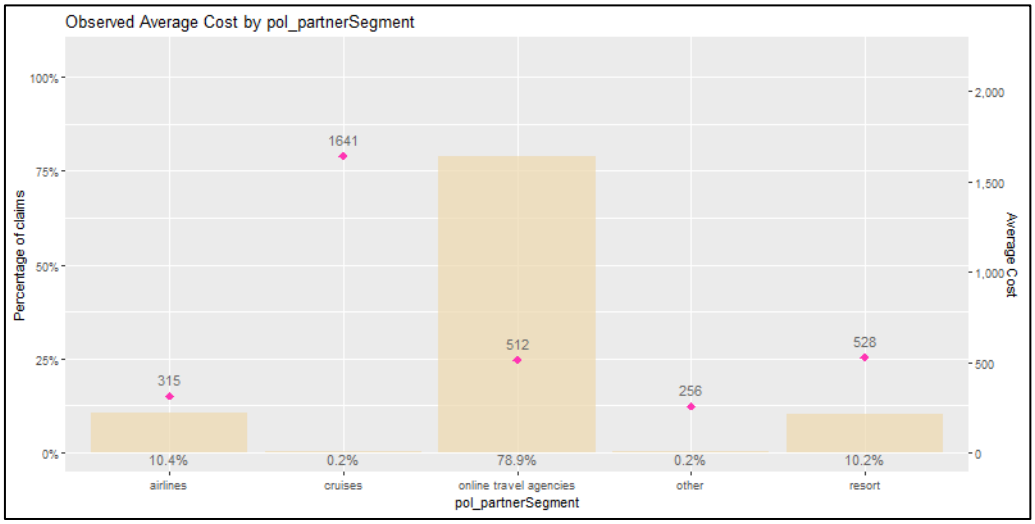


Figure 24 - Univariate analysis – Observed average cost/claim by partner segment



There seems to be significant differences in terms of frequency or average cost per claim between the different partner segments. Some could have been anticipated, eg. It seems to make sense that average cost per claim for cruises is the highest ones among the different partner segments. Nevertheless, given the limited number of partners included by category, we will see whether this variable adds value into the modelling or if this is only linked to differences specific to each partner included in our portfolio.

3/ By Booking Window

As already discussed, the Booking Window is the distance between the date of subscription of the policy (usually the date at which the travel is booked) and the date of departure of the trip.

Figure 25 - Univariate analysis – Observed frequency by Booking Window

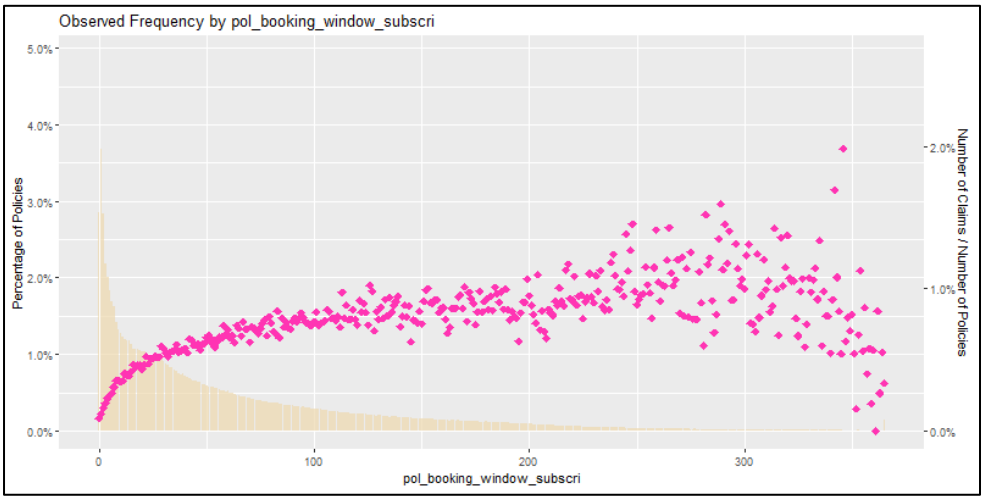
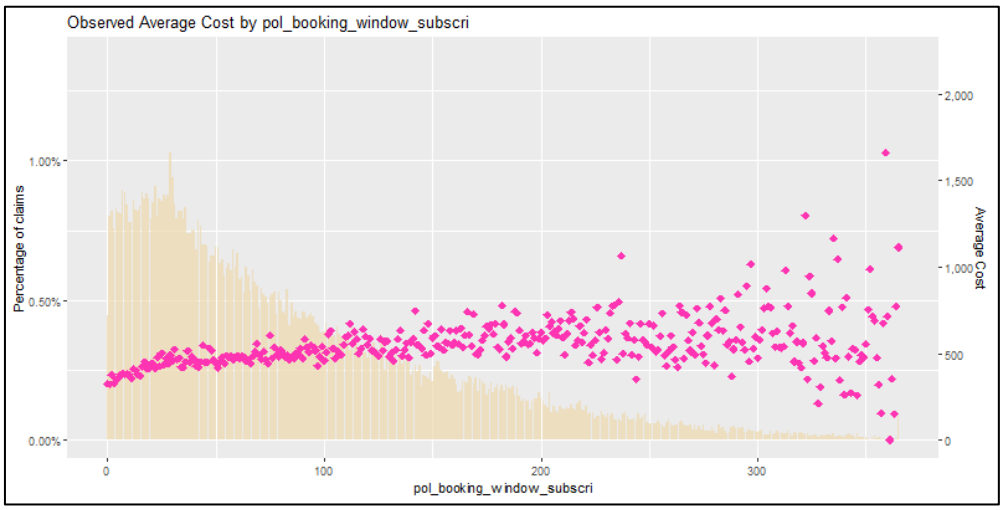


Figure 26 - Univariate analysis – Observed average cost/claim by Booking Window



In terms of frequency, as already seen and as anticipated, the longer is the Booking Window, the higher is the frequency. Regarding the average cost per claim, there seems to be a slightly positive relationship between the average cost per claim and the Booking Window. This could be interpreted by the fact that travelers probably book more in advance when it is for a more expensive trip. It will be interesting to look later at the correlation between these two variables.

4/ By Travel duration

The "Travel duration" is the time distance between the date at which the trip starts ("pol_dt_tripStart") and the date at which the trip ends ("pol_dt_tripEnd").

Figure 27 - Univariate analysis – Observed frequency by Travel Duration

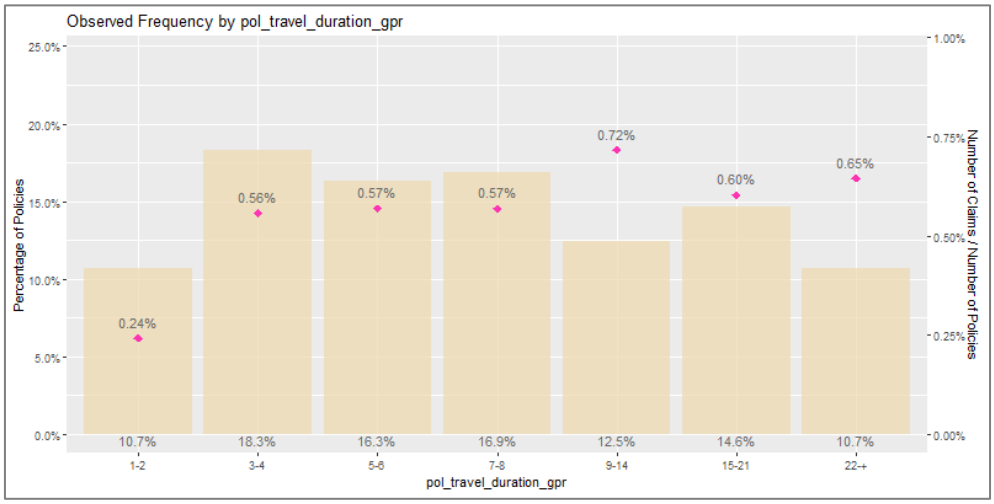
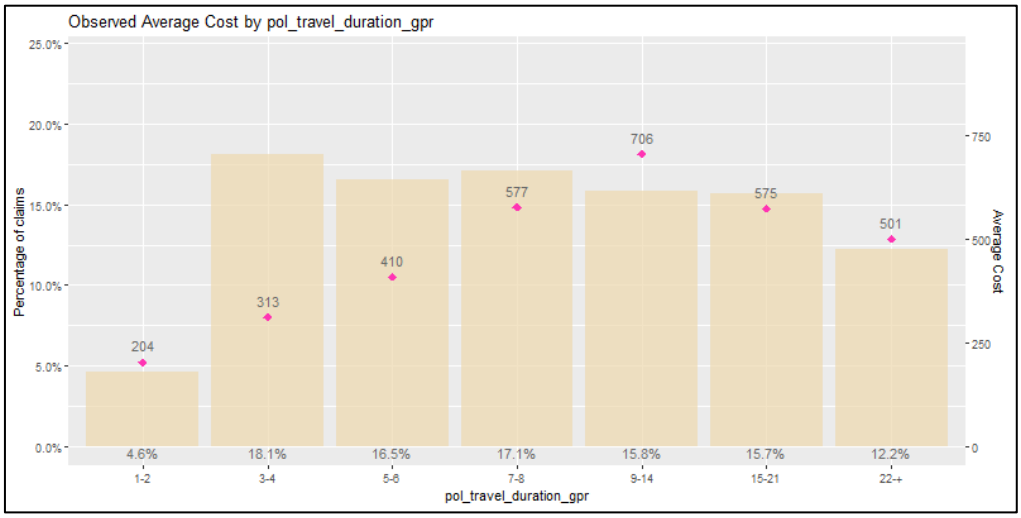


Figure 28 - Univariate analysis – Observed average cost/claim by Travel Duration



The risk (both in terms of frequency and average cost per claim) seems to increase when the travel duration increases. This seems to be in line with what we could have expected regarding the frequency: the longer is the trip, the most difficult it is to leave one of our family members who is sick. Regarding the average cost per claim, the positive relation can easily be explained: the longer is the trip, the more expensive it will be. It seems obvious but we might observe significant differences across partner as, for example, the cost of the trip is not linked to the duration of the trip when you book only flights (while it is well the case when you book both flights and hotel nights or only hotel nights). A bivariate analysis on the topic could be interesting.

5/ By month of subscription[21]

The month of subscription is the month during which the policy is subscribed: in most of the cases, it is the month during which the trip is purchased (as most of time, both are purchased at the same time).

Figure 29 - Univariate analysis – Observed frequency by month of subscription

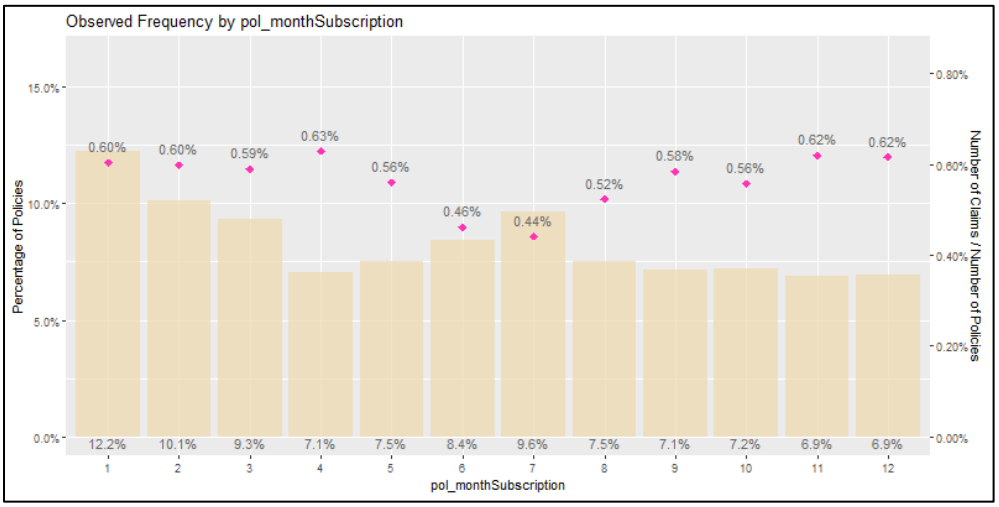
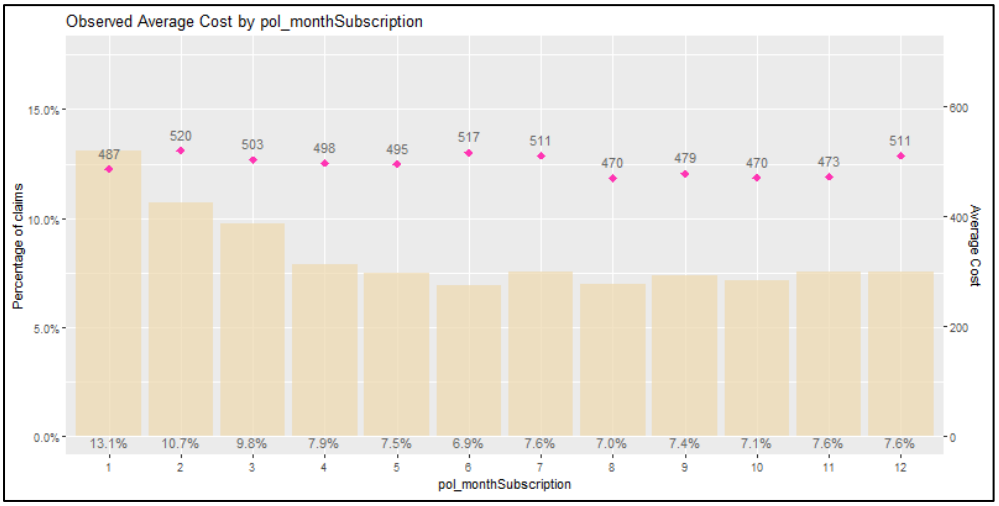


Figure 30 - Univariate analysis – Observed average cost/claim by month of subscription



This analysis is very interesting as I was not expecting much from this analysis. Nevertheless, what we can see is that some months of subscription have a higher frequency than others. The gap between the high-sinistrality months and low-sinistrality months is not neutral as there is a 43% difference between the month with the highest frequency and the month with the lowest frequency.

Regarding the average cost/claim, there are small differences across the different months but these differences are relatively marginal.

[21] See appendixes for the same analysis by month of departure

6/ By country of sales

The country of sales is the country in which the insured person is based. Our portfolio is focused on European countries.

Figure 31 - Univariate analysis – Observed frequency by country of sales

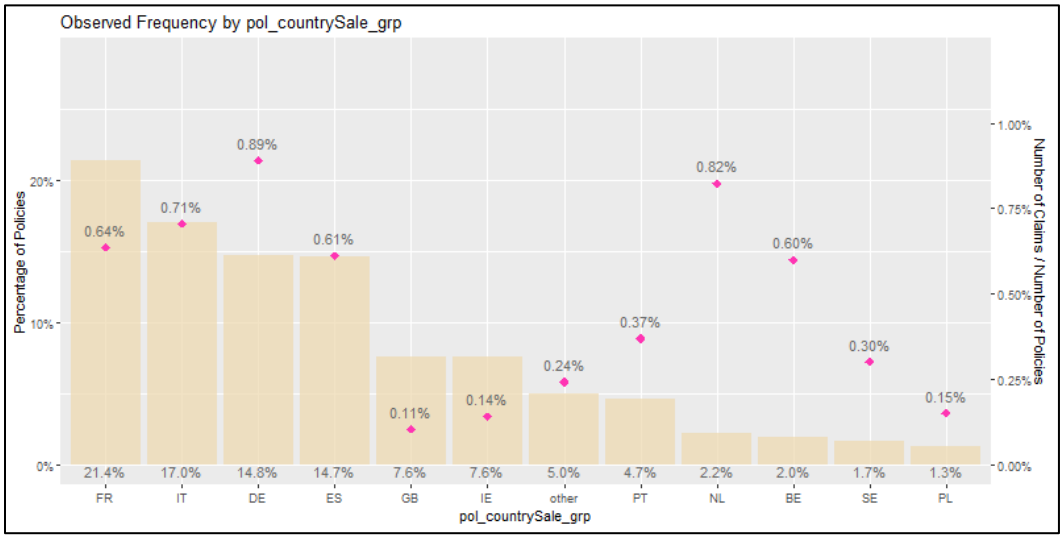
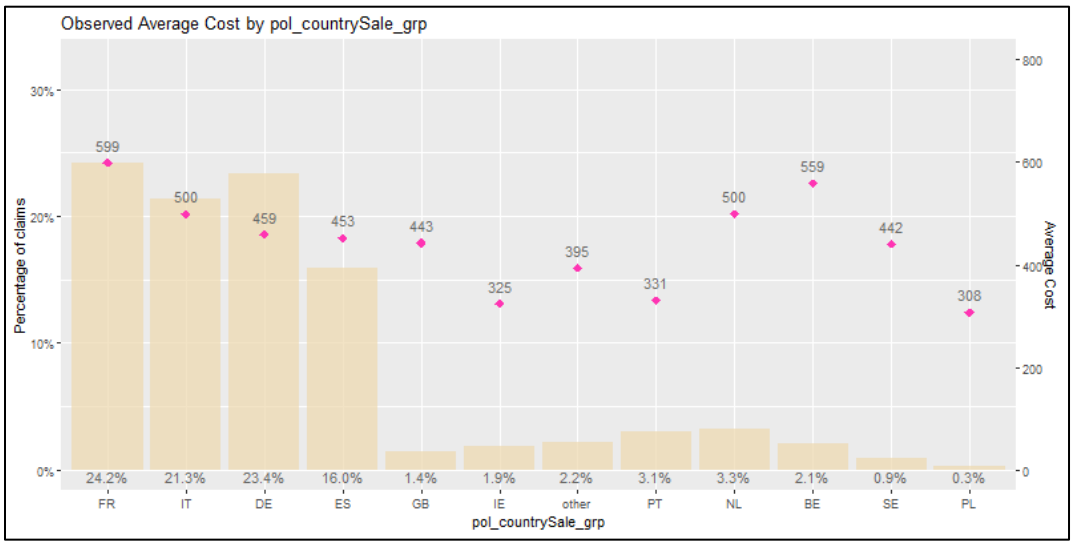


Figure 32 - Univariate analysis – Observed average cost/claim by country of sales



As we can see on these two charts, there are significant differences across the European countries for both the frequency and the average cost/claim, but mostly for the frequency. This could be explained by different market awareness regarding the travel insurance products.

7/ Other univariate analyses

Other univariate analyses are available in the appendixes. In particular, the following are important to note as they seem to have significant impact on the target variable:

- Univariate analysis on the Amount trip sold
- Univariate analysis on the type of cancellation product
- Univariate analysis on the client target

iv. Correlation between variables[22]

1/ Introduction

Before entering into the modelization, it is important to understand if there are correlations between variables that we intent to use as predictors. Indeed, when two strongly correlated variables are used in a GLM as predictors, it is like entering twice the same information in the model. It then becomes harder for the GLM to allocate between the two predictors this effect and the coefficients associated to the two variables then become uncertain.

A metric that is often used to measure the association between two variables is the Cramer's V. The values of the Cramer's V are comprised between 0 and 1 and is based on the Pearson's chi-squared statistic.

Let's consider a contingency table of two variables A and B. The size of the sample is n, n_{ij} represents the number of occurences of the couple (A_i, B_j) where:

- $i \in [1, r]$ where r is the number of lines
- $j \in [1, p]$ where p is the number of columns

The Pearson Khi statistic is then:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \frac{n_{i.}\, n_{.j}}{n})^2}{\frac{n_{i.}\, n_{.j}}{n}} \qquad (4.1)$$

The Cramer's V is computed as follows:

$$V = \sqrt{\frac{\varphi^2}{\min(r-1, p-1)}} = \sqrt{\frac{\chi^2/n}{\min(r-1, p-1)}} \qquad (4.2)$$

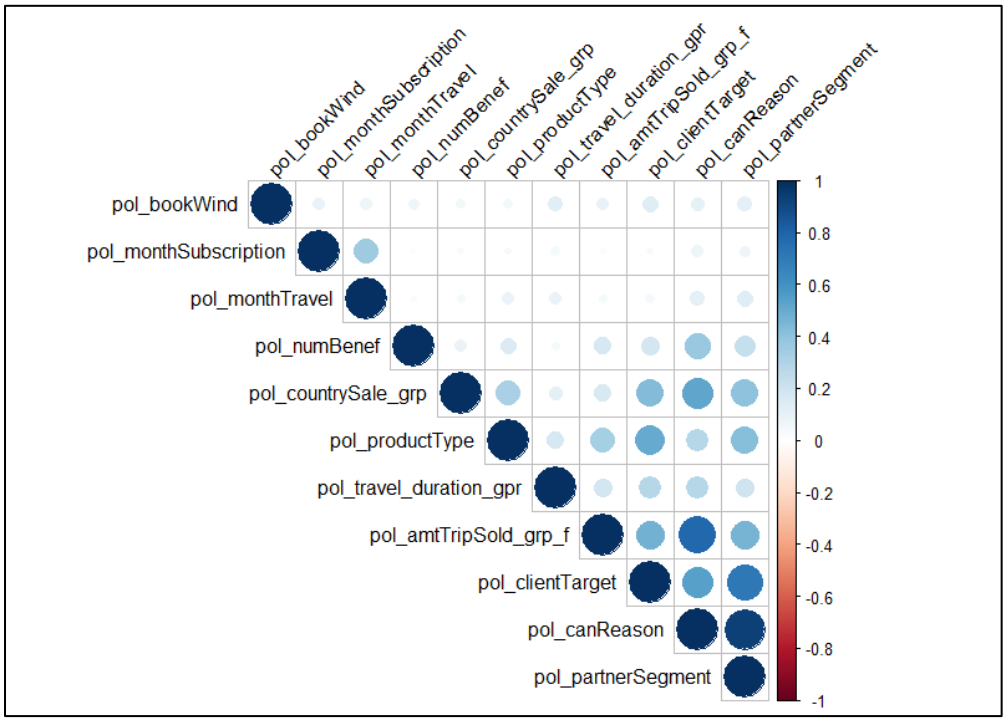
The Cramer's V varies from 0 when the association between the variables is null (they are independent) and 1 when the variables association is maximum (they are fully dependent).

[22] M. QUILFEN *Classification des véhicules en assurance automobile.* Mémoire ISFA

2/ Correlation analysis on the dataset

In the graph below, we can visually see the Cramer's V described above. Both the size and the colour of the bubble indicate the same information. This chart has been produced on R.

Figure 33 - Correlation analysis between the variables (Cramer's V)



From this chart, we mostly see three strong correlations between variables that we will need to keep in mind at the time of building the model:

- Between the type of cancellation product ("Pol_canReason") and the partner segment ("Pol_partnerSegment")
- Between the type of cancellation product ("Pol_canReason") and the amount trip sold
- Between the partner segment ("Pol_partnerSegment") and the type of customers targeted by the partner ("Pol_clientTarget")

One thing to be noted on these strong correlations is that, in each case, it includes at least one of the variables that we have added into the dataset ("Pol_partnerSegment", "Pol_canReason" or "Pol_clientTarget"). In particular, these variables are variables that only include a limited number of modalities (maximum 5 in the case of the variable "Pol_partnerSegment")

c. Final preparatory work and modelling strategy

i. Data partitioning

Before starting the modelling process, it is important to split our database into two sets of databases:

- The Train set: that will be used to perform all the modelling steps.

- The Test set: on which we will assess the performance of the model.

Usually the way this is done is by allocating a random number to each of the policies included in the policy, classify them according to this random number and make the selection based on that number (e.g. taking the first 80% for the Train set and the remaining 20% for the Test set).

In this work, I have used a R specific package called "*caret*" that allows this data partition to be more sophisticated as it allows to ensure a balanced split of the data between the *Train set* and the *Test set*. Using this package will allow to respect in both the *Train set* and the *Test set* the overall distribution of the data between classes. Doing so should improve the results of our model when we will test the performance of the model on the *Test set* and should avoid the model overfitting the *Train set.*

For this thesis, I have split the database as follows: 80% for the *Train set* and 20% for the *Test set*

ii. Modelling strategy

As defined, earlier in this chapter (see Figure 7), we propose to structure the modelling section around three main steps in order:

1/ To test different modelling methodologies for frequency and challenge the most common used methodology by two alternatives methodologies

Table 8 - Reminder of the three methodologies to be tested for cancellation

	Poisson based on exposure	**Poisson based on number of policies**	**Binomial**
Description	Alternative methodology	Classical methodology	Alternative Methodology
Distribution used	Poisson	Poisson	Binomial
Frequency definition	# of claims / exposure	# of claims / # of policies	# of claims / # of policies

2/ To find the best way to address the heterogeneity of the portfolio by finding the right trade-offs between performance of the model and overfitting of the model. In order to do that, we consider three different levels of granularity at which our modelization could be done

- One "global" model including all the partners of the database

- Five "segment" models (one per segment: OTA, Airline, Resorts, Cruise, Others)

- Sixteen "partner" models (one per partner)

Once these three types of models are done, we will be able to address the last step of the modelling section which will be to model the average cost/claim.

Testing all the combinations of models above would lead to a very high number of models to be run for the frequency modelling:

Table 9 - Number of models to be run for cancellation

	Number of models	Methodologies	Final number of models to be run
Global Model	1	3	3
Per partner-segment models	5	3	15
Per segment models	16	3	48
			66

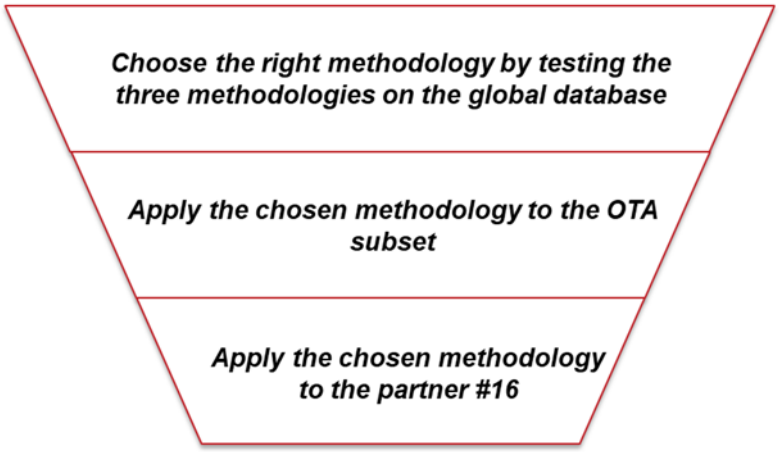
Therefore, in order to remain within a reasonable number of models to be run, I have taken the following funnel approach where

1/ we will test the three methodologies on the global database (Step 1 below)

2/ we will apply the best methodology to the "Online Travel Agency" segment (step 2 below)

3/ we will apply the best methodology to one of the major partners of the OTA segment, "Partner #16" (step 2 below)

Table 10 - Modelling funnel strategy



Once this is done, we will be able to analyze the results of the three models (each one built on a different set of data) and propose for the right balance between:

- High number of models to be built/maintained with a risk of over-fitting to one partner
- Lower number of models to be built/maintained with models that are less performant

A next step for us at Europ Assistance (that will not be presented in this thesis) will be eventually to run remaining models based the decision taken:

- Other models per partner-segment in case the per partner-segment models seem adequate
- Other models per partner in case the per partner models seem adequate

Also, in order to ease the read of this thesis, we will not go through the various steps for each of the the models but I will rather illustrate the methodology on one partner and then show the results (having followed the same methodology) for the rest of the models explained above

d. Full process illustration

As just explained, in order to ease the read of this thesis (and given the repetition of models to be handled), I propose to illustrate the full methodology on one partner ("Partner #16") and then, to focus on the results in the next sections of this thesis. Partner "#16" is an Online Travel Agency with a large number of policies.

i. Variables selection – the Forward methodology

The forward methodology consists of starting the modelling with the *null_model* described above in which we do not have any predictors included and then, we add the predictor that allows the better increase of the performance of the model (measured by the decrease of the AIC or the BIC). We repeat this operation as long as the AIC (and/or the BIC) decreases and we stop when the AIC stops decreasing.

In this case, we have run the *null_model* and the AIC of the *null-model* was 212366,7. We have then the following potential variable that can potentially be added:

- pol_amtTripSold_grp_f

- pol_bookWind

- pol_countrysale_grp

- pol_travel_duration_grp

- pol_numbenef

- pol_monthSubscription

-pol_monthTravel

-pol_ProductType

I ran the first iteration with R and it provided me with the following results:

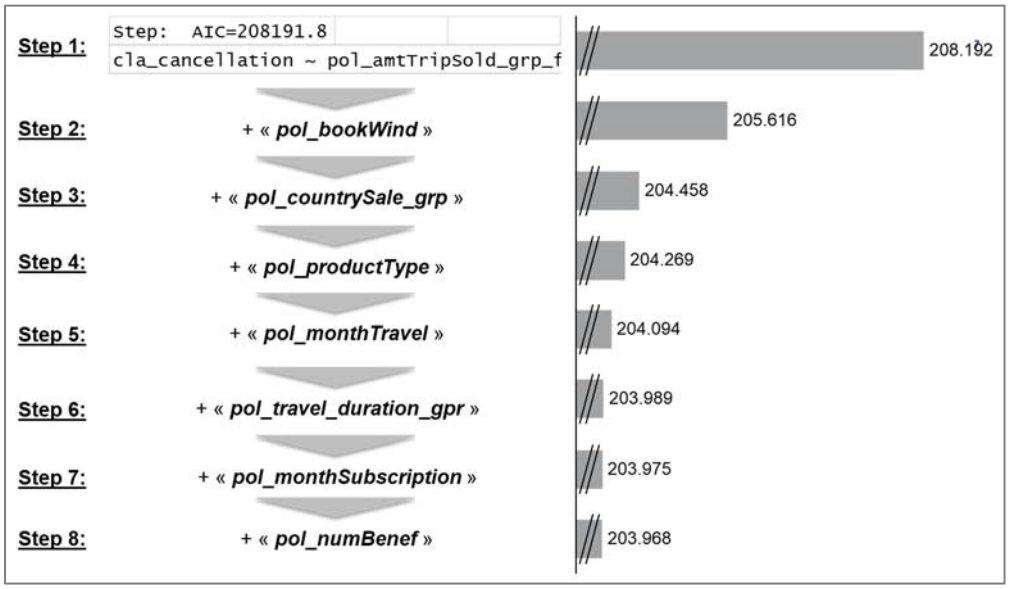
Table 11 - Forward selection of variables - Step 1 result

	Df	Deviance	AIC
+ pol_amtTripSold_grp_f	14	171938	208192
+ pol_bookWind	9	172047	208291
+ pol_countrySale_grp	9	174333	210577
+ pol_travel_duration_gpr	6	174919	211157
+ pol_numBenef	1	174937	211165
+ pol_monthSubscription	11	176036	212284
+ pol_monthTravel	11	176105	212353
+ pol_productType	1	176130	212358
<none>		176141	212367

This indicates that the best improvement of the model is achieved by adding to the model the variable "pol_amtTripSold_grp_f" leading to a decrease of the AIC from 212367 to 208192. It is not surprising that this variable comes first: it was identified in the univariate analyses made earlier that this variable had an important effect on the observed frequency and it is well in-line with what we could have expected.

I then ran similar iterations as long as the AIC was decreasing and it led to the following result:

Figure 34 - Results by step of the forward variables selection



As we can see, the second variable to be added was the Booking Window. This is again well in line with what could have expected as:

- The Booking Window was clearly identified in the univariate analyses as having a significant impact on the expected frequency which indicates that this variable would add value to the model

- Our Cramer's V analysis showed little correlation between the variable "pol_amtTripSold_grp_f" and the Booking Window which indicates that this variable would add marginal value compared to the first variable added into the model (the "pol_amtTripSold_grp_f")

Second learning from Figure 29 is that we have kept, at the end, all the available variables in the final model as the AIC was continuously decreasing till adding the last variable. This might look surprising but it is not that surprising given the limited number of available variables. Also, some could argue that the last variables added bring very little value but given the total number of values to be included in the model, I still decided to keep them. On the limited number of variables available, I would like to remind that:

- The policies that we are looking at are all sold through B2B2C channels. This means that we are dependent to a third party to get these variables. They usually have little interest in sharing too much data with their Insurance partner.

- The average value of the premiums that we are looking at is relatively low compared to other more usual insurance products which minimize the need to have a very large number of predictors to be used for pricing segmentation.

The output of this model including all the variables is the following:

Figure 35 - Output of Partner#16 (Model 2)[23]

[23] For confidentiality reasons, the coefficient of this table have been multiplied by a constant

```
glm(formula = cla_cancellation ~ pol_amtTripSold_grp_f + pol_bookWind +
    pol_countrySale_grp + pol_productType + pol_monthTravel +
    pol_travel_duration_gpr + pol_monthSubscription + pol_numBenef,
    family = poisson(link = "log"), data = trainDB)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3301  -0.1463  -0.1127  -0.0817   3.5992

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -6.499291   0.076803 -86.793  < 2e-16 ***
pol_amtTripSold_grp_f200         0.616193   0.045850  13.784  < 2e-16 ***
pol_amtTripSold_grp_f300         0.936295   0.045823  20.957  < 2e-16 ***
pol_amtTripSold_grp_f400         1.086154   0.047350  23.527  < 2e-16 ***
pol_amtTripSold_grp_f500         1.247253   0.047823  26.749  < 2e-16 ***
pol_amtTripSold_grp_f600         1.257737   0.048795  26.437  < 2e-16 ***
pol_amtTripSold_grp_f700         1.301928   0.050631  26.373  < 2e-16 ***
pol_amtTripSold_grp_f800         1.276677   0.052939  24.734  < 2e-16 ***
pol_amtTripSold_grp_f900         1.322465   0.055335  24.512  < 2e-16 ***
pol_amtTripSold_grp_f1000        1.329299   0.057403  23.751  < 2e-16 ***
pol_amtTripSold_grp_f1100        1.354334   0.059024  23.534  < 2e-16 ***
pol_amtTripSold_grp_f1200        1.365732   0.060526  23.143  < 2e-16 ***
pol_amtTripSold_grp_f1400        1.457094   0.054866  27.238  < 2e-16 ***
pol_amtTripSold_grp_f2300        1.482662   0.050143  30.327  < 2e-16 ***
pol_amtTripSold_grp_f_2300       1.561639   0.058417  27.418  < 2e-16 ***
pol_bookWind2w                   0.573066   0.042891  13.704  < 2e-16 ***
pol_bookWind3w                   0.732744   0.043352  17.336  < 2e-16 ***
pol_bookWind4w                   0.794868   0.044204  18.443  < 2e-16 ***
pol_bookWind5_6w                 0.879753   0.039501  22.843  < 2e-16 ***
pol_bookWind7_8w                 0,981428   0.040721  24.719  < 2e-16 ***
pol_bookWind3m                   1.07562  0.038069  28.979  < 2e-16 ***
pol_bookWind4m                   1.19726   0.039734  30.904  < 2e-16 ***
pol_bookWind5_6m                 1.28596   0.038870  33.932  < 2e-16 ***
pol_bookWind7m_                  1.433212   0.040281  36.492  < 2e-16 ***
pol_countrySale_grpES           -0.060422   0.022385  -2.768 0.005633 **
pol_countrySale_grpFR           -0.227963   0.020255 -11.543  < 2e-16 ***
pol_countrySale_grpGB           -1.192501   0.056115 -21.796  < 2e-16 ***
pol_countrySale_grpIT            0.025423   0.023868   1.092 0.274638
pol_countrySale_grpNL           -0.921434   0.142082  -6.651 2.90e-11 ***
pol_countrySale_grpother        -0.472334   0.050216  -9.647  < 2e-16 ***
pol_countrySale_grpPL            0.19549   0.140787   1.424 0.154400
pol_countrySale_grpPT           -0.638725   0.035515 -18.446  < 2e-16 ***
pol_countrySale_grpSE           -0.259002   0.069624  -3.815 0.000136 ***
pol_productTypepure_cancellation  0.208212   0.015304  13.954  < 2e-16 ***
pol_monthTravel2                 0.004103   0.044366   0.095 0.924439
pol_monthTravel3                -0.074502   0.043877  -1.742 0.081594 .
pol_monthTravel4                -0.232455   0.043871  -5.434 5.50e-08 ***
pol_monthTravel5                -0.201922   0.044601  -4.643 3.43e-06 ***
pol_monthTravel6                -0.250112   0.043999  -5.830 5.53e-09 ***
pol_monthTravel7                -0.397272   0.042825  -9.515  < 2e-16 ***
pol_monthTravel8                -0.385813   0.042135  -9.391  < 2e-16 ***
pol_monthTravel9                -0.191286   0.043208  -4.541 5.61e-06 ***
pol_monthTravel10               -0.220999   0.043913  -5.162 2.45e-07 ***
pol_monthTravel11               -0.19218   0.046131  -4.273 1.93e-05 ***
pol_monthTravel12               -0.234098   0.042367  -5.667 1.45e-08 ***
pol_travel_duration_gpr3_4      -0.092435   0.048839  -1.941 0.052236 .
pol_travel_duration_gpr5_6      -0.071543   0.049619  -1.479 0.139196
pol_travel_duration_gpr7_8      -0.041419   0.049979  -0.850 0.395330
pol_travel_duration_gpr9_14      0.074544   0.048706   1.570 0.116478
pol_travel_duration_gpr15_21    -0.057664   0.047348  -1.249 0.211626
pol_travel_duration_gpr22__      0.159279   0.049460   3.303 0.000957 ***
pol_monthSubscription2           0.007147   0.031172   0.235 0.814088
pol_monthSubscription3           0.036154   0.031873   1.163 0.244676
pol_monthSubscription4           0.099659   0.034367   2.974 0.002938 **
pol_monthSubscription5           0.086764   0.035440   2.511 0.012041 *
pol_monthSubscription6           0.085511   0.036989   2.371 0.017736 *
pol_monthSubscription7           0.135883   0.036682   3.799 0.000145 ***
pol_monthSubscription8           0.146743   0.037901   3.971 7.16e-05 ***
pol_monthSubscription9           0.09243   0.038230   2.480 0.013148 *
pol_monthSubscription10         -0.02119   0.039383  -0.552 0.581067
pol_monthSubscription11          0.026583   0.038390   0.710 0.477563
pol_monthSubscription12          0.045316   0.038760   1.199 0.230484
pol_numBenef                     0.021316   0.007454   2.933 0.003355 **
```

ii. Regrouping of variables

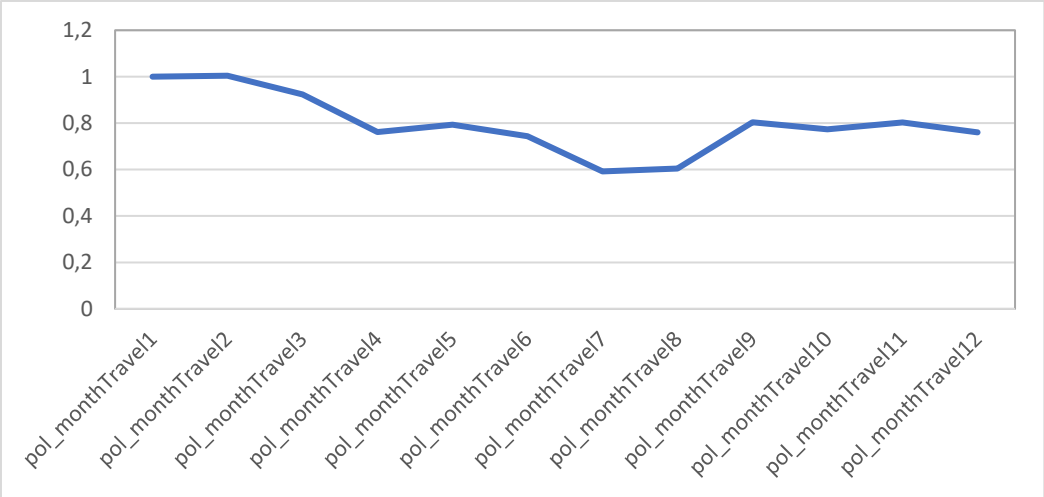
Some of the variables used in the model was grouped by default in the initial dataset based on common sense and layers that are usually used internally. One improvement to this would be to challenge this regrouping that were pre-defined in our database and that I did not challenge as part of this modelling exercise.

Nevertheless, based on the above figure, we can see that almost all the modalities of the different variables have meaningful coefficients except in the case of the following two variables:

- pol_monthTravel
- pol_travel_duration_grp

In order to decide what kind of regrouping could be made on the variable “pol_monthTravel”, I produce the following chart in order to visualize what months were similar to others.

Figure 36 - Partner #16 (Model 2) - "pol_monthTravel" Coefficients before regrouping



From that figure, we see that the month of July and August are the ones with the lowest sinistrality which is again consistent with what was observed in the univariate analysis. Other models built internally were considering two groups (summer vs non-summer months). Based on that figure, I considered that I could make the following regroupings:

- Group 1: January - March
- Group 2: April -June
- Group 3: July, August
- Group 4: September – December

. I re-run the modelling with the four proposed groups and it still showed high pValues for many of the modalities. I then decided to keep the variable with its initial grouping (by month) as:

- Further regrouping is not bringing value
- Having these twelve modalities do not bring too much complexity
- It is easily implementable in systems and can easily be explained to partners

iii. Model validation

In order to feed the evaluation of the performance of the model, we will systematically produce the following set of metrics.

Before looking at the actual result for this model, let's quickly define each of the metrics that will be used

- ***Null_Deviance**:* the *Null_Deviance* is the deviance (see definition above) of the *null_model*. The *null_model* is a model in which we would not have any predictors.

- ***Deviance:*** the *Deviance* is the actual deviance of the model that we have built

- ***Explained_Deviance_ratio***: the *Explained_Deviance_ratio* is given by the following formula. The higher is the Explained_Deviance_ratio, the better is the model as it means that the share of explained deviance by the model is higher.

$$Explained\ Deviance\ ratio = \frac{(Null\ Deviance - Deviance)}{Null\ Deviance} \qquad (4.3)$$

The following metrics are mostly used to compare models between them.

- ***AIC***: This metric was already defined above.

- ***BIC***: This metric was already defined above.

- ***Normalised Gini***: We defined earlier what the Gini index was. The normalized Gini is the ratio between the Gini index of the model and the Gini index of the perfect model. The higher is this ratio, the better it is.

- ***RMSE***: RMSE stands for Root Mean Square Error. It is one of the most commonly used distance to assess the performance of a simple or multiple regression. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (4.4)$$

The following table summarizes these metrics for the model that was built for partner #16.

Table 12 - Performance metrics Model 2 (Partner #16)

Partner #16	**Model 2**
Running time	2.4 mins
Null_Deviance	176 141
Deviance	167 619
Explained_Deviance_Ratio	**4,8%**
AIC	203 969
BIC	204 767
Normalised_Gini_Train	**0,3867**
RMSE_Train	**0,0874**
Normalised_Gini_Test	**0,3946**
RMSE_Test	**0,0879**

As can be seen in Table 13, I have also run the normalized Gini both in the Train set and the Test set in order to check that the model was not overfitted to the train set. Both results are similar and it tends to confirm that the model is not overfitted to the Train set and still valid on the Test set. To make this statement more visual, we can also look at lift curves both on the train set and on the test set.

Figure 37 - Lift curve - Train set: Model 2 Partner#16

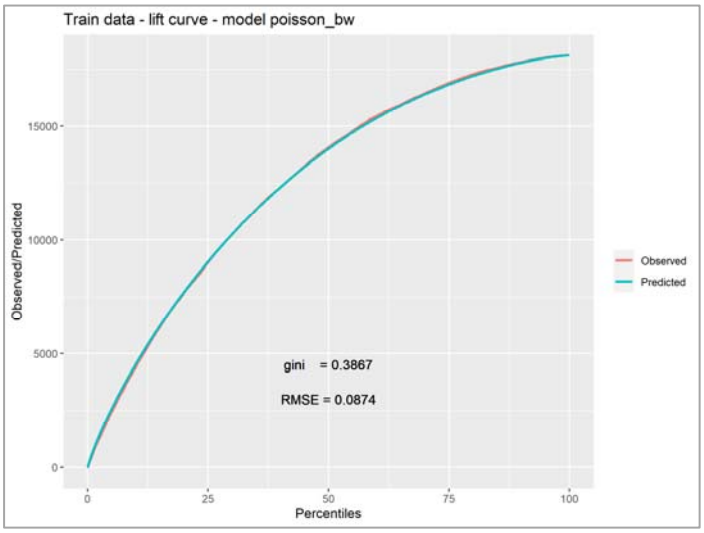
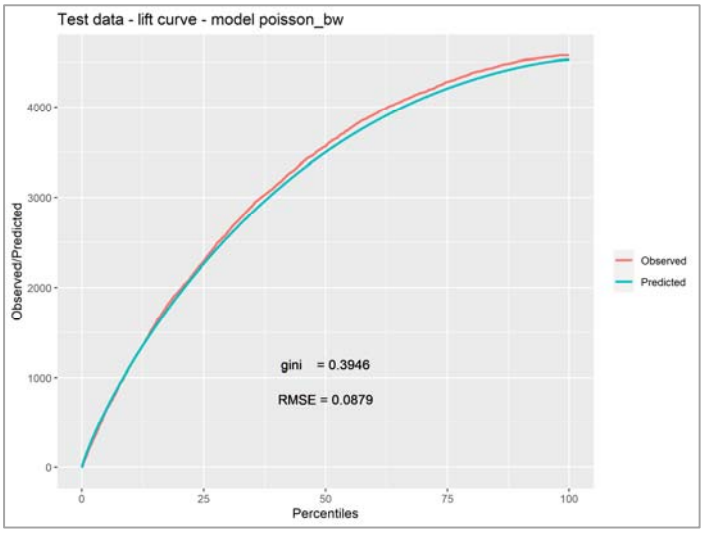


Figure 38 - Lift curve - Test set: Model 2 Partner #16



We see from these two figures that the fit on the test set is slightly less good than on the train set. Nevertheless, we still see that modelling reflects very well the modelling of the test set as the observed and predicted lines remain very close.

e. Step 1 – Frequency modelling: finding the right methodology

As already explained several times in the previous chapters, the usual methodology used to model the frequency of a cancellation product is to model it with a Poisson distribution, and defining frequency as the number of claims divided by the number of policies. Nevertheless, this is not the most obvious methodology to be used.

Indeed, once a cancellation claim is opened no additional claim can be opened on that policy. This indicates that a binomial distribution would be best suited to model that type of claim pattern (whereas the Poisson distribution would be best suited to count a number of claims).

Also, the Short Term cancellation policies that we are looking at are not annual policies: these policies are exposed between the subscription date of the policy and the departure date of the trip. As a consequence, calculating a frequency by dividing the number of claims by the number of policies might again be counter intuitive. Indeed, it would make more sense to calculate the frequency by dividing the number of claims by the exposure of the policy in years (eg. A policy with a 6-months distance between the subscription date and the departure date would count for 0,5). This would allow to take into account the fact that the longer is the duration between the subscription date and the departure date, the higher is the risk).

Combining these elements, we end up with three potential methodologies:

Table 13 - Reminder of the three methodologies to be tested for cancellation

	Methodology 1	**Methodology 2**	**Methodology 3**
Description	Alternative methodology	Classical methodology	Alternative Methodology
Distribution used	Poisson	Poisson	Binomial
Frequency definition	# of claims / exposure	# of claims / # of policies	# of claims / # of policies

As explained earlier, I have run the three methodologies on the global database using the process described in the previous section on Partner #16. The results of the three modellings made on the global database can be summarized as follows:

Table 14 - Performance metrics of the three methodologies

	Model 1 Poisson Expo	**Model 2 Poisson BW**	**Model 3 Binomial**
Null_Deviance	394 104	388 663	464 742
Deviance	375 416	358 942	434 795
Explained_Deviance_Ratio	4,7%	**7,6%**	6,4%
AIC	451 834	435 378	434 917
BIC	452 543	436 210	435 749
Normalised_Gini_Train	0,4301	**0,4855**	**0,4855**
RMSE_Train	0,0781	**0,0779**	**0,0779**
Normalised_Gini_Test	0,4332	**0,4822**	**0,4822**
RMSE_Test	0,0780	**0,0778**	**0,0778**

Based on these metrics, we can draw several learnings:

- The model 2 is the one with the highest *Explained Deviance Ratio*

- The model 3 has the lowest AIC and BIC

- In terms of normalized Gini and RMSE on the Train set, models 2 and 3 are totally equivalent

- The three models do not seem to overfit on the train set as the three models have similar normalized Gini and RMSE on the test set, compared to the Train set

The outputs of the model (with all the coefficients for the different variables and modalities) are shown in appendix[24].

Between the three models, model 1 can easily be excluded as there is no metrics on which Model 1 overperforms the other models. A potential explanation for that is that Model 1 forces a purely linear relationship between exposure and frequency whereas other models capture the same information (through the Booking Window predictor) without imposing a purely linear relationship between this predictor and the target variable (and thus benefiting of the advantages of the GLM modelling for this predictor as well).

Between model 2 and model 3, the choice is more difficult as they are very close on many metrics (Normalized Gini and RMSE); the model 2 shows a better Explained Deviance ratio than model 3. Nevertheless, the difference in the performance metrics are not big and the decision must thus also take into account other elements that go beyond pure actuarial metrics:

- Model 2 is the most common used so far and as a consequence, it will be less disruptive in the internal organization to keep the existing model.

- Some other cancellation products are annual products for which several claims could be open on a given policy for which Model 2 might be better suited.

For these two reasons, we will pursue the rest of the thesis focusing on Model 2.

f. Step 2 – Frequency modelling: finding the right granularity for our models

Now that we have validated the methodology that we wanted to apply moving forward, the second element that I wanted to challenge in this thesis is at what level of granularity we should model our database. Indeed, so far within Europ Assistance, models were made on one or two of the biggest partners separately. Nevertheless, there are big differences across partners on sinistrality and the partners on which models were made, are not representative of the full portfolio and hence, their models cannot be used across the board.

The question that we are trying to address is thus whether we can build more global models without losing too much accuracy given the variety of partners (and hence underlying risks) included in our database.

As explained above, we will focus on three models, at a first stage, comparing results on three different perimeters

- "Global" model: on the whole database
- "Segment" model: Online Travel Agency (OTA)
- "Partner" model: Partner #16 (which belongs to the OTA segment)

Using the process described above and the same metrics to compare models between them, the results of the three modelings can be summarized as follows.

[24] See appendix 3

Table 15 – Performance metrics comparaison by model

	Global	OTA	Partner #16
Null_Deviance	388 663	297 325	176 141
Deviance	358 942	275 923	167 619
Explained_Deviance_Ratio	**7,6%**	**7,2%**	**4,8%**
AIC	435 378	336 389	203 969
BIC	436 210	337 290	204 767
Normalised_Gini_Train	**0,4855**	**0,4627**	**0,3867**
RMSE_Train	**0,0779**	**0,0845**	**0,0874**
Normalised_Gini_Test	**0,4822**	**0,4617**	**0,3946**
RMSE_Test	**0,0778**	**0,0847**	**0,0879**

The results of two of these three models were already discussed previously in this paper:

- "Partner #16" results were already discussed in section "d. Full process illustration"
- "Global Model" results were already discussed as well in section "Step 1 – Frequency modelling: finding the right methodology"

Focusing on the "OTA" model, the model seems to perform well as we observe a similar level of fit both on the train and test set which means again that the model is not overfitting the Train set and works well as well on the Test set.

In order to compare the models between them and better evaluate the loss of information that we have by making more global models (OTA, Global) compared to the per partner Model, it is easier to look at the following table. The Gini and RMSE reported values in the below table (for the "Segment" and "Global" models) are not equivalent to the previous tables as, in the below table, they have been calculated on the restricted perimeter of "Partner #16" for a better comparison (whereas, in the above tables, they were calculated on the perimeters respectively in-scope, the segment OTA for the "Segment" model and the overall database for "Global" model).

Table 16 - Gini and RMSE: information lost vs "partner #16" model

Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,3867	0,3946		
Segment	0,3770	0,3866	-2,5%	-2,0%
Global	0,3749	0,3849	-3,0%	-2,4%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,08742	0,08788		
Segment	0,08744	0,08789	0,0%	0,0%
Global	0,08744	0,08789	0,0%	0,0%

Looking at the above table, it looks like that the level of information lost by using a "global" or "segment" models instead of the "partner" model is not major. Indeed, the RMSE remains equivalent and the normalized Gini is decreasing by only 3% moving from the "Partner" model to the "Global"

model. This is an overall picture of the different models but it is important to check as well that there is no major discrepancy between the different models in terms of the way the different predictors are fitted to the observed data. This can be easily visualized by plotting into the same chart the predicted frequency by the three models compared to the observed frequency. Here below, I have produced a number of these charts[25] on the Train set.

Figure 39 - Observed vs predicted: Amount_TripSold (Train set)

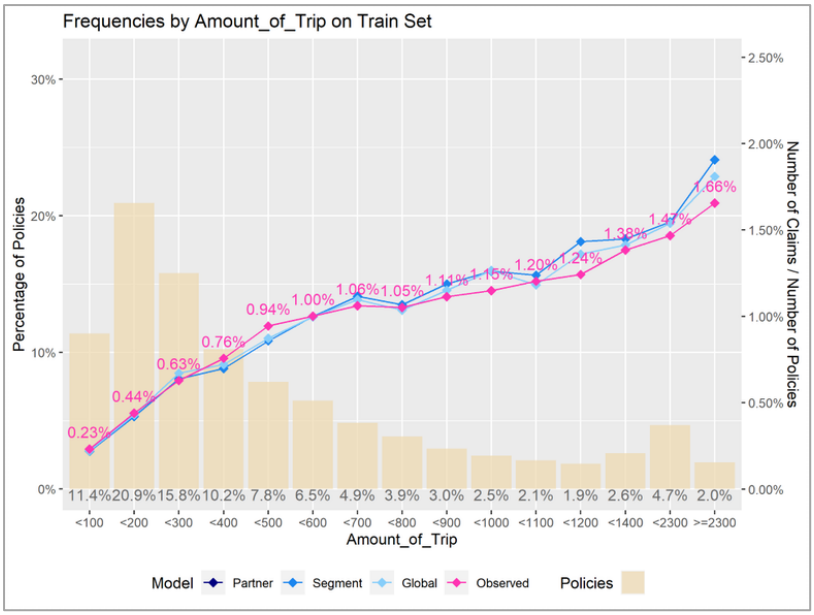
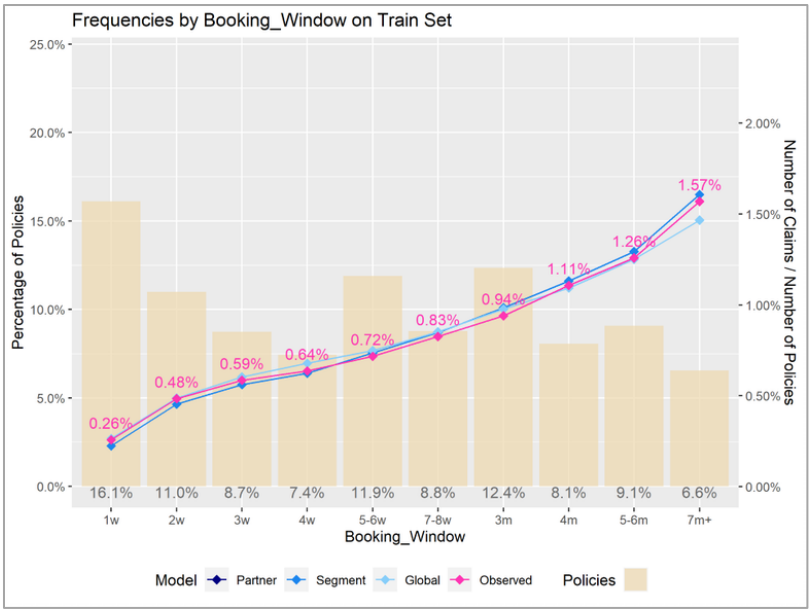


Figure 40 - Observed vs predicted : Booking_Window (Train set)



Looking at these charts (and other in appendixes), it looks like the different models work well not only at global level but, as well on the different predictors. These charts are built on the "Train set" and we

[25] Additional charts on other predictors can be found in appendix 4 for information.

can produce similar charts on the "Test set" in order to check whether there is not over-fitting on some predictors on the "Train set". Some of these charts are produced her below[26].

Figure 41 - Obsverved vs predicted: Amount_TripSold (Test set)

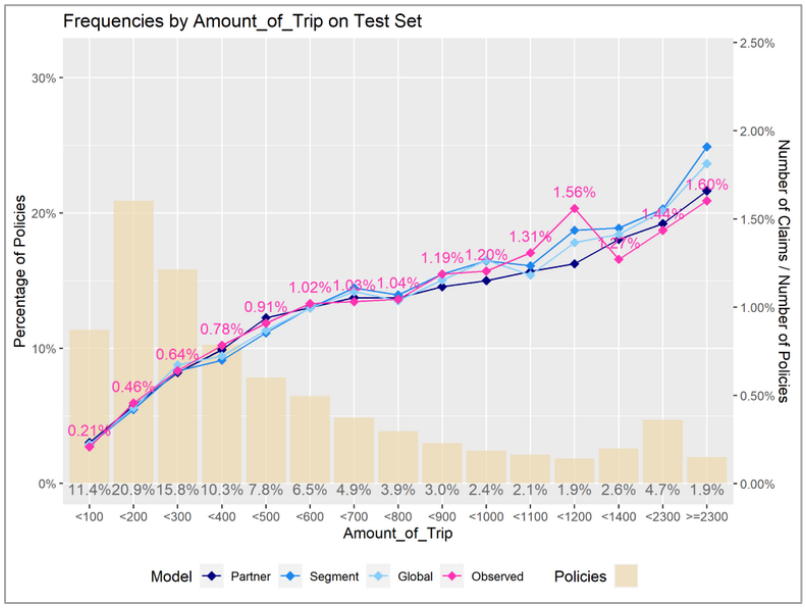
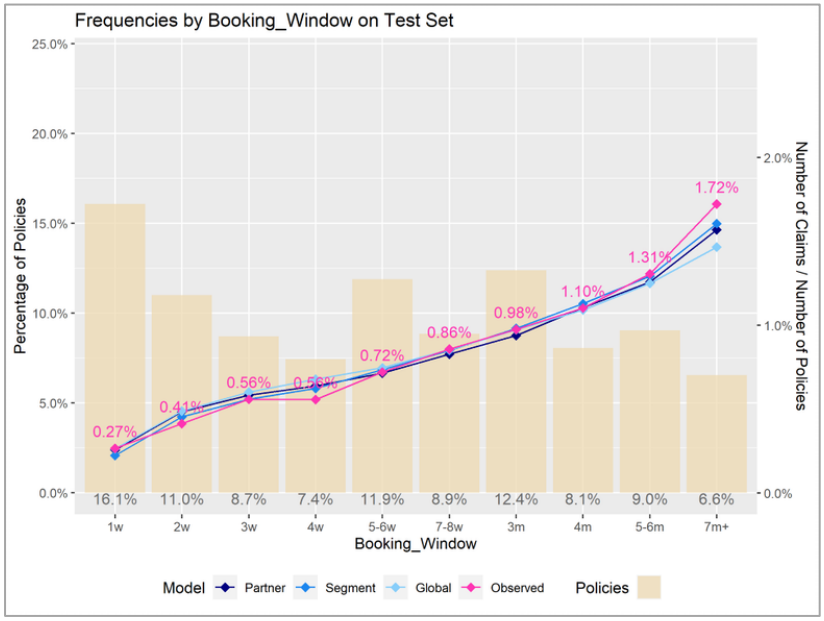


Figure 42 - Observed vs predicted : Booking_Window (Test set)



From this chart again, we do see that the different models continue to fit properly with the observed data. This means again that the model is not overfitting the train set.

This ends the methodology that we wanted to apply in order to test whether we can continue to use a global model without losing too much accuracy and the above analyses tends to show that this is ok.

[26] Additional charts on other predictors can be found in appendix 5 for information.

Nevertheless, this is not in line with the initial expectations where we were expecting, that given the heterogeneity of the partners, one "global" model would probably not be the chosen solution.

Thinking further at the potential flaws into the reasoning, the partner that we have chosen to run the comparison happens to be the largest partner in our database and actually represents 36% of the policies included in our initial database. There is thus a risk that our conclusion so far is biased by the importance this partner has played in the GLM building process.

As a consequence, I have gone slightly beyond the initial methodology that was planned in order to test this on other partners included in the database and I have thus run similar comparisons on three other "OTA" partners included in the database:

- Partner #15
- Partner #26
- Partner #27

These two additional partners have been chosen as they are two other significant partners of the OTA segment that we have been looking it so far. Let's reproduce the same analyses that have been produced for "Partner #16" above in order to compare amount of information lost by moving from a "Partner" model to a "Segment" model or to a "Global" model.

Table 17 - Gini and RMSE: information lost vs "partner #15" model

Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,5404	0,5443		
Segment	0,5240	0,5222	-3,0%	-4,1%
Global	0,5184	0,5150	-4,1%	-5,4%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,06914	0,06859		
Segment	0,06917	0,06861	0,0%	0,0%
Global	0,06918	0,06862	0,1%	0,0%

Table 18 - Gini and RMSE: information lost vs "partner #26" model

Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,3213	0,1697		
Segment	0,2223	0,2197	-30,8%	29,5%
Global	0,2135	0,2223	-33,5%	31,0%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,11637	0,11574		
Segment	0,11780	0,11703	1,2%	1,1%
Global	0,11768	0,11681	1,1%	0,9%

Table 19 - Gini and RMSE: information lost vs "partner #27" model

Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,7465	0,7419		
Segment	0,5520	0,5466	-26,1%	-26,3%
Global	0,5230	0,5131	-29,9%	-30,8%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,14289	0,16402		
Segment	0,14800	0,14854	3,6%	-9,4%
Global	0,14830	0,14885	3,8%	-9,3%

The main conclusions that we can draw from the above tables are the followings:

- From Partner #15, we have very similar conclusions to what we have observed for Partner #16 previously. We also observed that Normalized Gini is even better than for Partner #16 which means that the level of fit of our model is even better on that partner than on Partner #16. This tends to confirm that the information lost by moving from a “Partner” to a “Global” model is not very significant.

- From Partner #26, on the contrary, conclusions that can be drawn are unfortunately not in line with what we have observed on both Partner #15 and Partner #16 as:

- The Partner model does not seem to work well as the normalized Gini is significantly higher for the Train dataset compared to the Test dataset. This means that we are overfitting the Train dataset and the level of normalized Gini for the Test dataset is really to low to consider that the model can be used
- The information lost by moving from the “Partner” model to the “Global” model is significant and is already significant when moving from the “Partner” model to the “Segment” model.
- In this case, despite the loss of information between the “Partner” model and more global models, we should consider using the “Partner” or “Global” models in order to avoid overfitting the “Train” dataset.

- What we can learn from Partner #27 is again different from what we have observed for the other partners

- In this case, we have a very good “Partner” model as the normalized Gini is significantly higher that the ones of “Partner” models for other partners
- Nevertheless, we have a significant loss of information by moving from the “Partner” model to the “Segment” or “Global” models. Here again, the loss of information is already significant by moving from the “Partner” to the “Segment” model. Again in this case, the “Segment” model does not seem to bring much value into the modelling process.

As a conclusion to this section, the objective was to go away from pure "Partner" models by trying to build "Segment" models and a "Global" model. There was not much hope from the beginning that the global model would fit the different partners given heterogeneity across partners. The initial hypothesis was that the "segment" models could be a good compromise between the global model and the partners models. Unfortunately, it seems that the "Segment" models do not bring much value compared to the "Global" model but lead to non-negligible loss of information compared to the Partner models. Nevertheless, as it has been seen above, having built the "Global" model will certainly be useful, in some cases, in particular for smaller partners where we would not have enough data to model and where we would face a risk to overfit the "train" data.

g. Step 3 – Average cost per claim modelling

Now that we have modelled the frequency, in order to be able to model the pure premium, we still need to model the average cost per claim. As explained earlier, we do expect simpler model in the case of the average cost/claim given the direct relationship between the amount trip cost insured and the cost of the claim.

The variables that we have available in the database in order to predict the average cost per claim are the same as the ones that we used for the frequency models. I used a Gamma distribution in order to model the average cost / claim and I used a forward methodology in order to select the predictors.

In order to have complete models that match with the models that were built for the frequency, I built again three average cost/claim models:

- One for Partner #16
- One for the Segment OTA
- One for the Global database

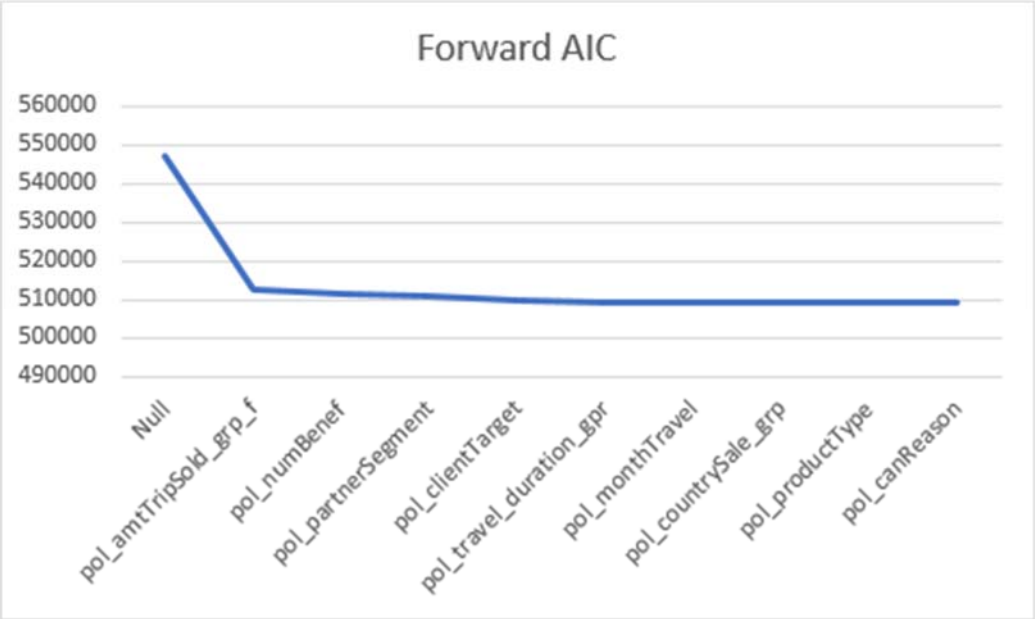
In all cases and as expected, the first variable that we added to the model was the "pol_amtTripSold_grp_f", representing the cost of the trip insured. Then, depending on the perimeter looked at (Partner #16, Segment or global models), the number of final variables kept in the model varies:

- In the case of the "Partner" model, the following variables were kept:
    1. pol_amtTripSold_grp_f
    2. pol_countrySale_grp
    3. pol_numBenef
    4. pol_travel_duration_gpr
- In the case of the "Segment" model, the following variables were kept:
    1. pol_amtTripSold_grp_f
    2. pol_clientTarget
    3. pol_numBenef
    4. pol_travel_duration_gpr
    5. pol_countrySale_grp
    6. pol_productType
- In the case of the Global model, the following variables were kept:
    1. pol_amtTripSold_grp_f
    2. pol_numBenef
    3. pol_partnerSegment
    4. pol_clientTarget

5. pol_travel_duration_gpr
6. pol_monthTravel
7. pol_countrySale_grp
8. pol_productType
9. pol_canReason

Nevertheless, looking at the evolution of the AIC when adding variables (see figure below), we could argue that keeping only the cost of the trip insured as predictor would be enough.

Figure 43 - Average cost per claim - Global model - Method forward: AIC evolution



Similarly to what we have done earlier for frequency models, I have then calculated the same key performance metric of the three models in order to evaluate their performance. The results of the three models are summarized below.

Table 20- Average cost / claim : key performance metrics

	Partner	Segment	Global
Null_Deviance	14 702	29 275	34 532
Deviance	7 229	9 896	13 848
Explained_Deviance_Ratio	50,8%	66,2%	59,9%
AIC	239 863	399 838	509 397
BIC	240 113	400 138	509 850
Normalised_Gini_Train	0,7487	0,8476	0,8083
RMSE_Train	349,9481	454,1731	428,5249
Normalised_Gini_Test	0,7594	0,8411	0,8176
RMSE_Test	383,8871	404,4306	425,7199

As we can see from this table, we have in all case very good Explained Deviance ratio compared to what we were having in the case of the frequency models. On top of that, what is important to note is

that, in the case of the three models, the Normalized Gini on the "Train" dataset is very close to the one on the "Test" dataset which indicates that we are not overfitting on the "Train" data.

If we want to run a similar analysis to what we have done for frequency in order to compare what granularity of model should be used, we can look at the below table where the Normalized Gini has been calculated for the three models on the perimeter of the Partner #16.

Table 21 - Comparaison of the average cost/claim models

Gini				
	Train	**Test**	**%gap train**	**%gap test**
Partner	0,7487	0,7594		
Segment	0,7473	0,7587	-0,2%	-0,1%
Global	0,7422	0,7528	-0,9%	-0,9%
RMSE				
	Train	**Test**	**%gap train**	**%gap test**
Partner	350	384		
Segment	358	385	2,3%	0,4%
Global	354	386	1,3%	0,6%

What we see from this table is that

1/ we have consistent level of normalized Gini between the Train and the test dataset for the different models

2/ The amount of information lost by moving from a "Partner" model to a "Segment" or "Global" model is negligible.

Based on the different models that we have built for frequency and average cost/claim, I am now in a position to build pure premium models. Doing so will still require some loadings as we should take into account the share of the claim that we were not able to link to the right policy. These must obviously be taken into account in order to not under-estimate the pure premium.

These last steps will not be illustrated in this thesis.

h. Further areas of improvement

1/ Frequency modelling

The purpose of this thesis was to challenge existing actuarial methodology in place on one side and on the other side, to build more global models and see in what cases they could be used.

On the first objective, the analysis tends to validate the methodology that was historically used on travel cancellation covers. Nevertheless, the results were very close between this methodology (Poisson distribution based on frequency) and the binomial distribution. Therefore, I would suggest testing these two methodologies moving forward in order to make sure we do not miss anything. Also, from an actuarial methodology standpoint, in the models that we have built, we have not much

reviewed the regrouping of the variables (eg. Booking Window) as they were already existing in the database used. They make sense from a business standpoint, but, moving forward, we could challenge this type of regrouping by using hierarchical ascending classification.

On the second objective, this study has enlarged the possibilities in terms of models available and also started to set ideas on what models to be used on what type of partners. Nevertheless, I see a number of further improvements that could bring value moving forward:

1/ Create new cluster of partners

What I have done is that I have created an intermediate level of models (the "Segment" models") based on the idea that each of the partners included in the database was belonging to one partner segment and using a segmentation used on the market.

Nevertheless, instead of doing so, we could use some classification methodologies to create new clusters of partners based on actuarial analyses instead of business rules. It could potentially then create adequate more homogeneous group of partners that would allow to more easily move away from a "Partner" model to more global model.

2/ Standardize the product offering

As said at the beginning of this thesis, there is quite a lot of heterogeneity in the products that are sold. These are not big differences but differences in the way deductibles are expressed, differences in terms of exclusions, differences in terms of cancellation reasons covered. All these small differences are the fruit of the past but do not bring much value to the final customer and create some issues in our modelling capability as well as in terms of efficiency on the claims management platforms.

3/ Enrich data available

As said earlier, with the fields available in the database, we do not necessarily capture the full heterogeneity of the portfolio. There is probably additional variables that might be captured and would be useful to model the sinistrality. There might be different sources of information:

- Partners might be willing to share more data. Some of them are very data driven and probably have many data that we could test in order to see the potential impact on our sinistrality
- External sources of data, eg. Official vacation dates. One of the things that we have observed is that frequency of cancellation seems to be lower in the months where there is school holidays.
- Increase the granularity at which we describe our products in our "product" tables.

4/ Industrialize the GLM building process

Given the number of models that we will have to build, we might benefit from existing market tools, such as EMBLEM or Akur8, in order to gain efficiency in creating new models.

2/ Average cost per claim modelization

In order to build the models for average cost/claim, we have used the same type of predictors candidates as the ones that we have used for modelling the frequency. This is the methodology that was used so far in this type of modelling exercise of the cancellation cover and it was not challenged in this thesis.

What we have seen is that, as expected, the most important predictor is the cost of the trip insured. What we have also learnt is that the rest of the predictors bring limited value compared to the cost of the trip insured.

Having these results in mind, what I would suggest to do moving forward is to proceed differently. Indeed, besides the cost of the trip insured, what would matter is when the customer cancels compared to the departure date. Indeed, depending on when the customer cancels, the penalty scheme in place in different and as a consequence, the financial loss incurred by the insured (to be indemnified) is different : the closer you get to the departure date, the higher is the share of the trip that will not be reimbursed.

Moving forward, I would thus try to model two things:

- Either integrate in the database the penalty scheme in place or model, on average, the share of trip cost that is not reimbursed depending on when the customer cancels
- Model the date at which the customer will probably cancel

Conclusion

This final aim of this study was to try expanding the use of segmented pricing models for the cancellation cover. These models are not widely used in Europe on Travel insurance products for multiple reasons and most of the models developed so far were built on restricted perimeters and therefore not suited across the full portfolio.

In order to achieve that goal, we tried to answer two main questions:

- Is the usual actuarial methodology used on travel cancellation the right methodology or should we try alternative methodologies?
- Is it possible to build more global models or is the portfolio too heterogeneous to do this at this stage?

Before answering these questions, I had to build the database. I focused on the perimeter of international contracts within Europe assistance as it was including a large variety of partners, with policies underwritten in many countries and with relatively consistent data across partners. Nevertheless, in building this database, I faced some cleaning challenges as, often for this type of exercise, the database was not as clean as expected. Main challenges faced at this level revolves around 1/ the ability to properly link the claims with the right policies and 2/ to enrich the database with new fields in order to capture as much heterogeneity of portfolio as possible thanks to added fields in the database.

In order to answer the first question, I tested and compared three different methodologies. The results of the comparison showed that one methodology was clearly underperforming the two others. At the end between the two methodologies, I decided to stick with the classical methodology as it was less disruptive compared to existing methods, but the binomial methodology needs to be investigated again when new databases will be made available.

Regarding the second question, I compared the key performance metrics of different models in order to understand the loss of information faced by moving from a “Partner” model to a “Segment” model or a “Global” model. The results of this comparison were heterogeneous from a partner to another: large partners tend to be well represented by the global model whereas small partners are not well represented by the global model. The good news is that we start having a set of models that we will be able to use for new pricings. Nevertheless, there are probably additional variables that could be added into the model in order to further capture differences across partners, products and segments.

Moving forward, in terms of improvements, there are several business and actuarial levers that could be activated in order to increase our ability to build more global models. From an actuarial point, of view, we could see 1/ if we could use classification methodologies to build clusters of partners (differently than what we have done by creating segment); 2/ what external data could be leveraged in order to capture features that we are not capturing through variables received from our partners and 3/ we could work on industrializing our GLM building processes by leveraging softwares such as EMBLEM or Akur8.

From a business standpoint, we would benefit from having more standardized products in order to have less heterogeneity in our portfolios and increase our ability to capture in our database proper description of the different products sold. Also, we should engage into discussions with partners in order to get more data from them that can then be leveraged to better understand the insured risks.

List of Figures

List of Tables

Bibliography

[1] https://www.argusdelassurance.com/acteurs/assisteurs/assistance-des-resultats-en-hausse-pour-la-profession.129052 (French assistance market analysis), consulted in May 2019

[2] https://www.argusdelassurance.com/classements/classement-assistance-2018-l-assistance-dans-l-urgence-de-l-innovation.128944 (Main players on the French assistance market analysis), consulted in May 2019

[3] M. GOLDBURG, A. KHARE and D. TEVET *Generalized Linear Models for insurance ratings.* 2016 (CAS Monographe Series, Number 5)

[4] F. PLANCHET and A. MISERAY *Tarification IARD – Introduction aux techniques avancées – Version 1.3.* 2017 (ISFA)

[5] *Technical Pricing Process v3.0.* 2019 (Generali Group – Internal publication)

[6] M. QUILFEN *Classification des véhicules en assurance automobile.* Mémoire ISFA

[7] J. QIU *Travel pricing sophistication with GLM and Machine Learning approaches.* 2017. Mémoire EURIA

[8] J. WANG *Tariff Segmentation for some travel insurance products in Europe.* 2014 Actuarial Science Thesis – Université Paris Dauphine

[9] M. AOUICHI *Analyse de l'assurance voyage à travers la tarification du risque.* 2016. Mémoire ISFA

[10] D. ANDERSON *A Practitioner's guide to Generalized Linear Models (3rd Edition).* 2017

[11] https://en.wikipedia.org/wiki/Gamma_distribution (Presentation/description of the Gamma distribution), website visited in October 2019

[12] https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution (Presentation/description of the Gaussian distribution), website visited in October 2019

[13] https://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/ (illustration of ROC Curve), website visited in November 2020

[14] https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/ (methodologies on the GLMs interpretation), website visited in October 2020)

[15] https://medium.com/@HolmesLaurence/modelling-large-insurance-claims-using-extreme-value-theory-in-r-18d84cb2742f (Methodology to define threshold for large losses), website visited in October 2020

[16] O. DECOURT *Le langage R au quotidien – Traitement et analyse de données volumineuses.* 2018, Editions Dunod

Appendixes

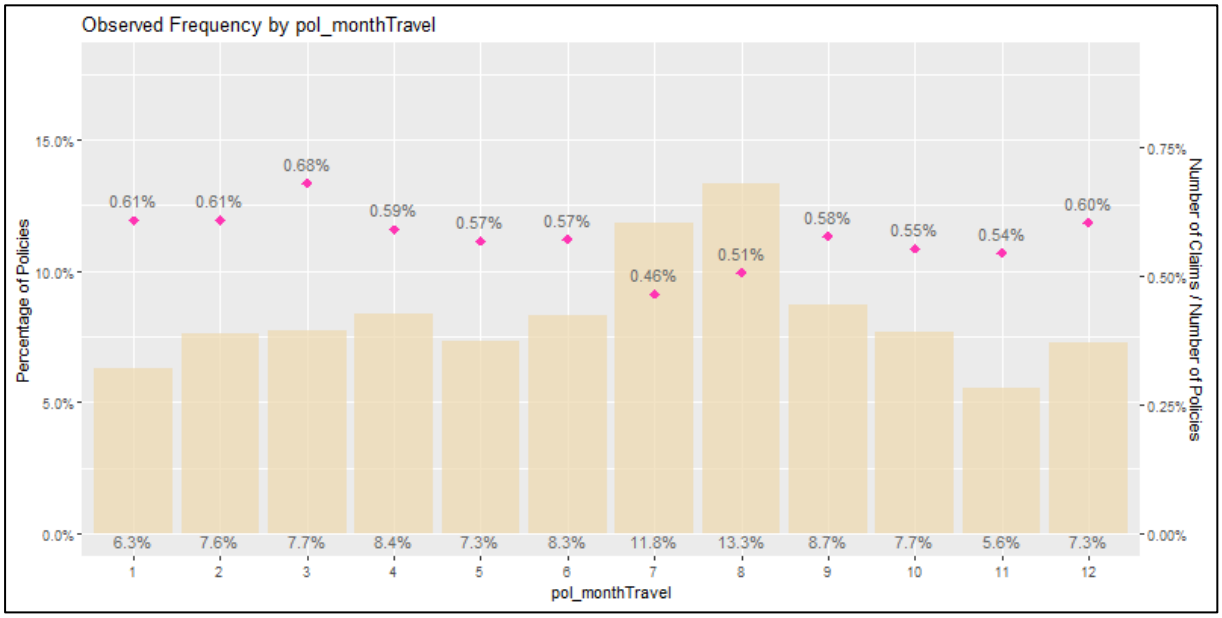
1. Additional variable added in the database – Client target

pol_clientGroup	Policies	mean__TripSold	TripSold_Benef	TripSold_Dura	rank1	rank2	rank3	rank_tot	target
Partner #19	1 816	332	64	16	3	1	1	2	low_cost
Partner #8	109 754	276	169	37	1	2	6	2	low_cost
Partner #24	3 026	296	183	34	2	3	5	3	low_cost
Partner #15	2 202 571	436	233	65	5	5	8	6	low_cost
Partner #18	2 884 930	406	404	32	4	8	4	6	low_cost
Partner #16	3 494 649	522	308	31	7	7	3	6	medium
Partner #21	44 598	540	242	125	9	6	10	8	medium
Partner #3	18 992	448	447	101	6	10	9	8	medium
Partner #22	14 999	864	592	30	11	12	2	10	medium
Partner #25	20 447	824	426	211	10	9	14	10	medium
Partner #27	200 661	539	223	58	8	4	7	6	premium
Partner #26	57 928	1 157	452	148	12	11	11	11	premium
Partner #17	555 727	1 447	970	168	14	13	12	13	premium
Partner #2	36 129	1 358	1 352	204	13	15	13	14	premium
Partner #23	716	3 004	1 243	256	15	14	15	15	premium
Partner #9	23 029	5 342	1 949	392	16	16	16	16	premium

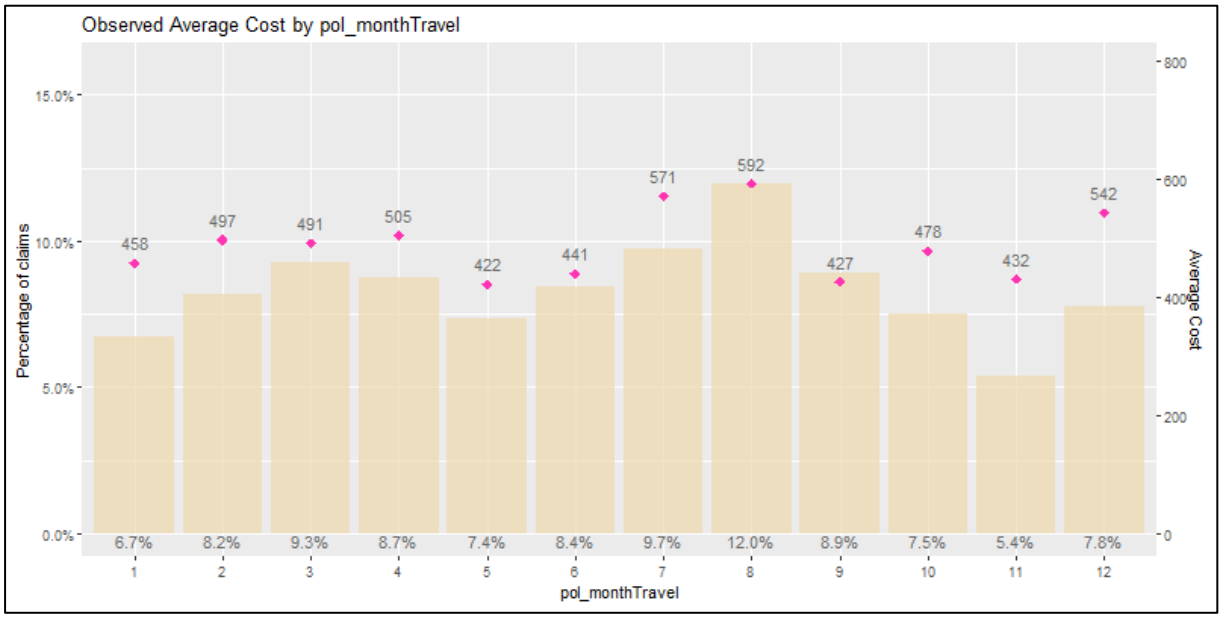
2. Additional univariate analyses

2.1 By month of departure

Univariate analysis – Observed frequency by month of departure

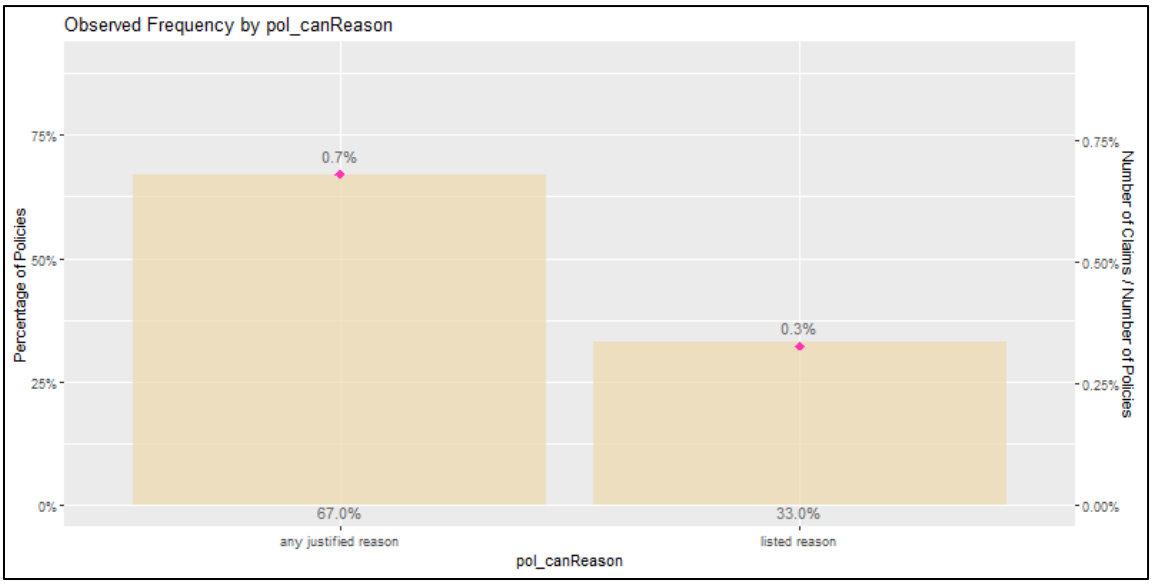


Univariate analysis – Observed average cost by month of departure

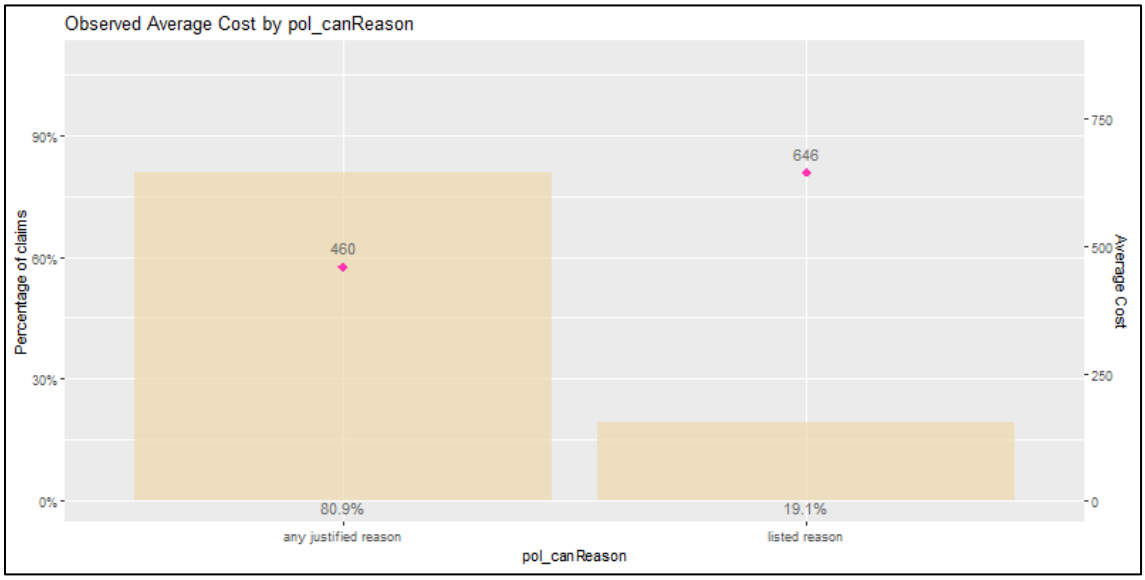


2.2 By type of cancellation product sold

Univariate analysis – Observed frequency by type of cancellation product

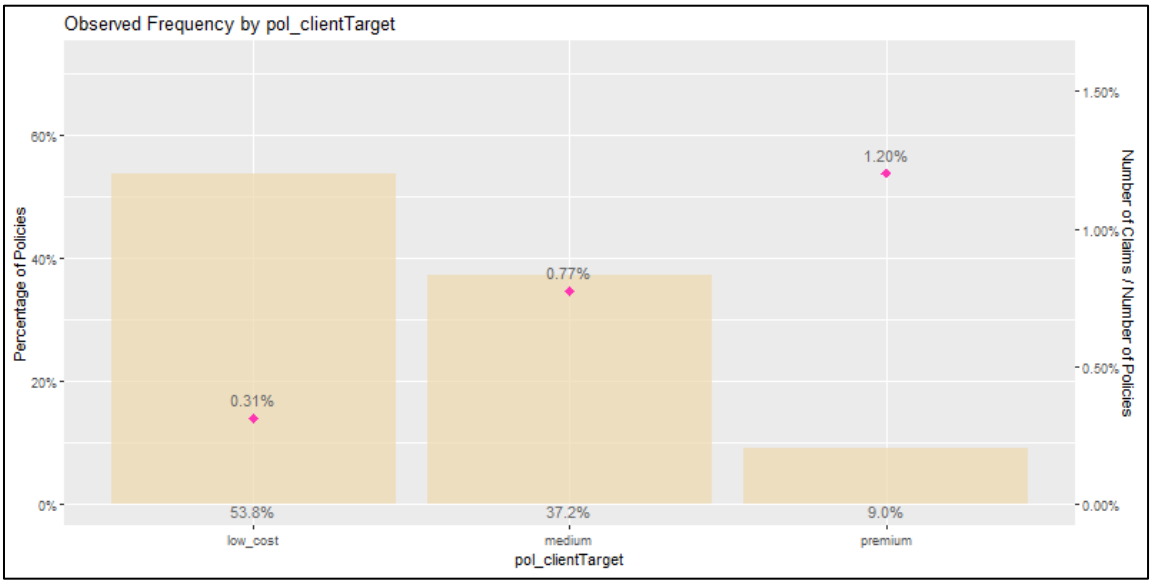


Univariate analysis – Observed average cost by type of cancellation product

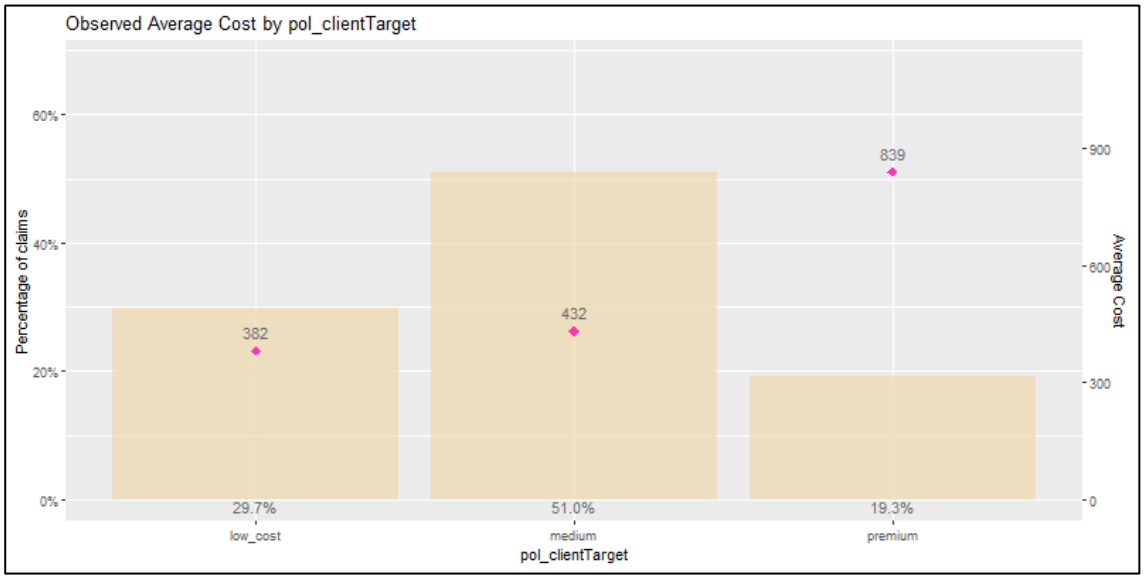


2.3 By client target

Univariate analysis – Observed frequency by client target

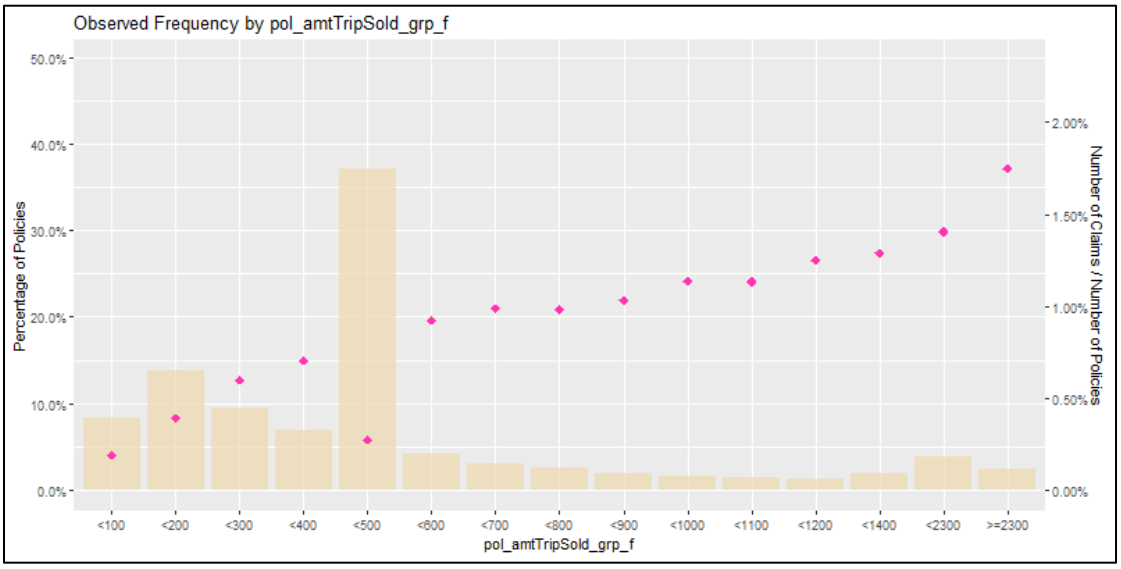


Univariate analysis – Observed average cost by client target

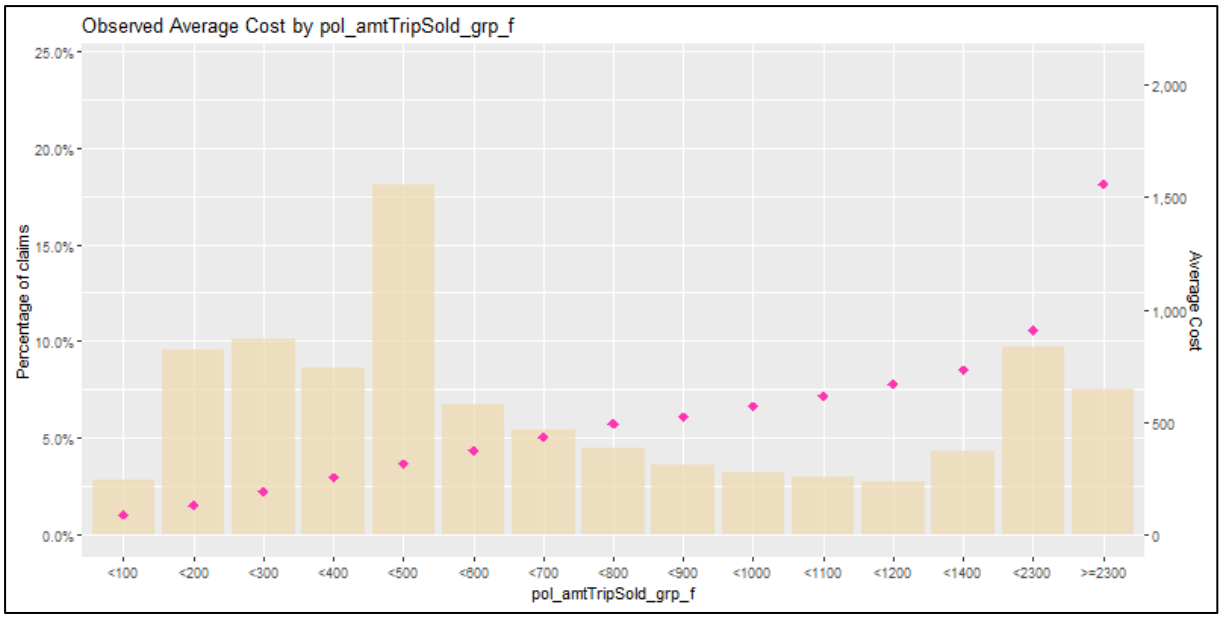


2.4 By amount trip sold

Univariate analysis – Observed frequency by amount trip sold



Univariate analysis – Observed average cost by amount trip sold



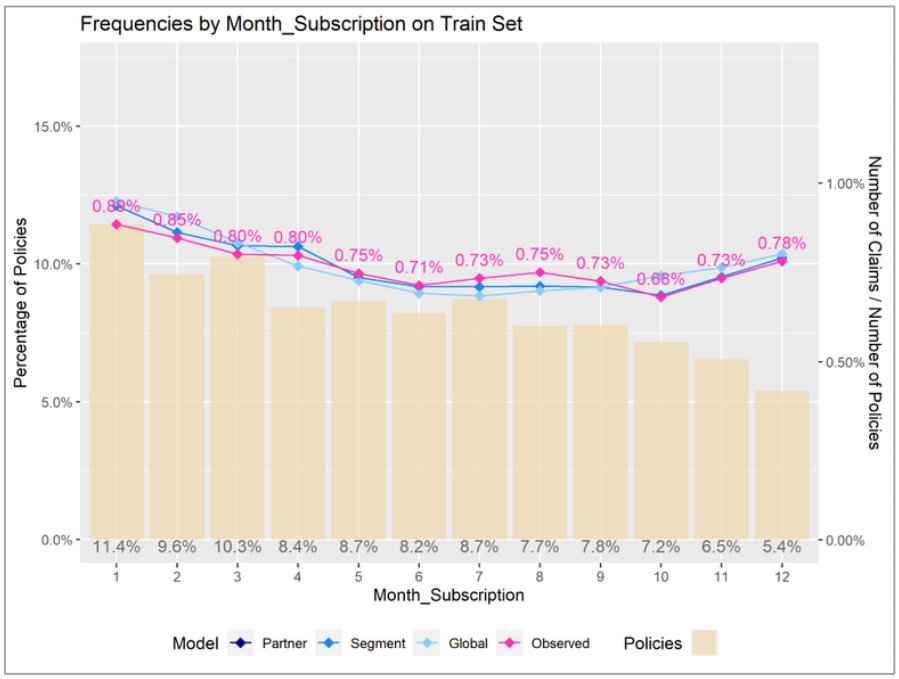
3. Outputs of Model 2 (Poisson distribution) on the global database

```
Call:
glm(formula = formula_2, family = poisson(link = "log"), data = trainDB)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4944  -0.1298  -0.0901  -0.0616   3.8637

Coefficients:
```

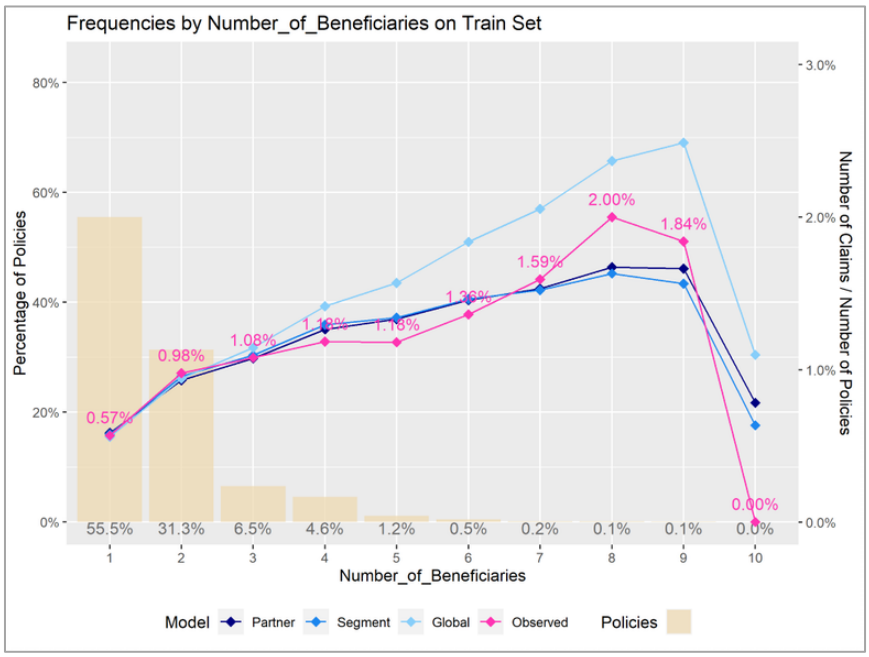
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8,382226	0,081097	-106,01	< 2e-16	***
pol_monthTravel2	0,005951	0,031782	0,192	0,8477	
pol_monthTravel3	-0,033461	0,030555	-1,123	0,2614	
pol_monthTravel4	-0,202205	0,029815	-6,956	3,50E-12	***
pol_monthTravel5	-0,158501	0,03031	-5,363	8,16E-08	***
pol_monthTravel6	-0,236852	0,029651	-8,193	2,55E-16	***
pol_monthTravel7	-0,444783	0,029056	-15,7	< 2e-16	***
pol_monthTravel8	-0,38631	0,02814	-14,08	< 2e-16	***
pol_monthTravel9	-0,155743	0,029007	-5,507	3,66E-08	***
pol_monthTravel10	-0,230308	0,029756	-7,938	2,05E-15	***
pol_monthTravel11	-0,191256	0,031662	-6,195	5,81E-10	***
pol_monthTravel12	-0,238879	0,029576	-8,284	< 2e-16	***
pol_bookWind2w	0,540279	0,029968	18,49	< 2e-16	***
pol_bookWind3w	0,71042	0,03002	24,272	< 2e-16	***
pol_bookWind4w	0,796575	0,030119	27,126	< 2e-16	***
pol_bookWind5_6w	0,857719	0,026823	32,797	< 2e-16	***
pol_bookWind7_8w	0,939471	0,027807	34,651	< 2e-16	***
pol_bookWind3m	1,030103	0,025969	40,684	< 2e-16	***
pol_bookWind4m	1,096669	0,027225	41,315	< 2e-16	***
pol_bookWind5_6m	1,18074	0,026475	45,742	< 2e-16	***
pol_bookWind7m_	1,244816	0,027546	46,349	< 2e-16	***
pol_travel_duration_gpr3_4	0,356429	0,027772	13,163	< 2e-16	***
pol_travel_duration_gpr5_6	0,328825	0,028389	11,88	< 2e-16	***
pol_travel_duration_gpr7_8	0,256646	0,028649	9,188	< 2e-16	***
pol_travel_duration_gpr9_14	0,336638	0,029094	11,867	< 2e-16	***
pol_travel_duration_gpr15_21	0,268227	0,029122	9,447	< 2e-16	***
pol_travel_duration_gpr22__	0,45793	0,030539	15,38	< 2e-16	***
pol_amtTripSold_grp_f200	0,630364	0,034804	18,576	< 2e-16	***
pol_amtTripSold_grp_f300	1,034544	0,034818	30,475	< 2e-16	***
pol_amtTripSold_grp_f400	1,073556	0,035862	30,703	< 2e-16	***
pol_amtTripSold_grp_f500	1,208636	0,036378	34,076	< 2e-16	***
pol_amtTripSold_grp_f600	1,300875	0,03733	35,742	< 2e-16	***
pol_amtTripSold_grp_f700	1,382661	0,038607	36,732	< 2e-16	***
pol_amtTripSold_grp_f800	1,315062	0,040095	33,64	< 2e-16	***
pol_amtTripSold_grp_f900	1,405745	0,041864	34,44	< 2e-16	***
pol_amtTripSold_grp_f1000	1,471985	0,042812	35,264	< 2e-16	***
pol_amtTripSold_grp_f1100	1,385037	0,044111	32,204	< 2e-16	***
pol_amtTripSold_grp_f1200	1,507882	0,044854	34,48	< 2e-16	***
pol_amtTripSold_grp_f1400	1,527475	0,040853	38,348	< 2e-16	***
pol_amtTripSold_grp_f2300	1,572495	0,036995	43,596	< 2e-16	***
pol_amtTripSold_grp_f_2300	1,67135	0,039911	42,951	< 2e-16	***
pol_countrySale_grpDE	0,250878	0,041417	6,213	5,21E-10	***
pol_countrySale_grpES	0,091502	0,043362	2,164	0,0304	*
pol_countrySale_grpFR	-0,010009	0,042216	-0,243	0,8079	
pol_countrySale_grpGB	-1,137173	0,058482	-19,943	< 2e-16	***
pol_countrySale_grpIE	-0,594466	0,056665	-10,76	< 2e-16	***
pol_countrySale_grpIT	0,369732	0,043035	8,812	< 2e-16	***
pol_countrySale_grpNL	0,080804	0,046974	1,764	0,0777	,
pol_countrySale_grpother	-0,655817	0,05446	-12,351	< 2e-16	***
pol_countrySale_grpPL	-0,479025	0,095746	-5,131	2,88E-07	***
pol_countrySale_grpPT	-0,41084	0,050954	-8,27	< 2e-16	***
pol_countrySale_grpSE	-0,505496	0,06924	-7,488	7,00E-14	***
pol_canReasonlisted_reason	0,465932	0,043651	10,948	< 2e-16	***
pol_productTypepure_cancella	0,205975	0,013438	15,721	< 2e-16	***
pol_clientTargetmedium	0,131343	0,016166	8,333	< 2e-16	***
pol_clientTargetpremium	0,354418	0,045132	8,054	7,99E-16	***
pol_partnerSegmentcruises	-0,281905	0,14112	-2,049	0,0405	*
pol_partnerSegmentonline_tra	1,177598	0,052454	23,026	< 2e-16	***
pol_partnerSegmentother	-0,941089	0,172248	-5,604	2,10E-08	***
pol_partnerSegmentresort	0,896036	0,092101	9,978	< 2e-16	***
pol_numBenef	0,072056	0,004722	15,651	< 2e-16	***

4. Observed vs predicted – Partner #16 (Train set)

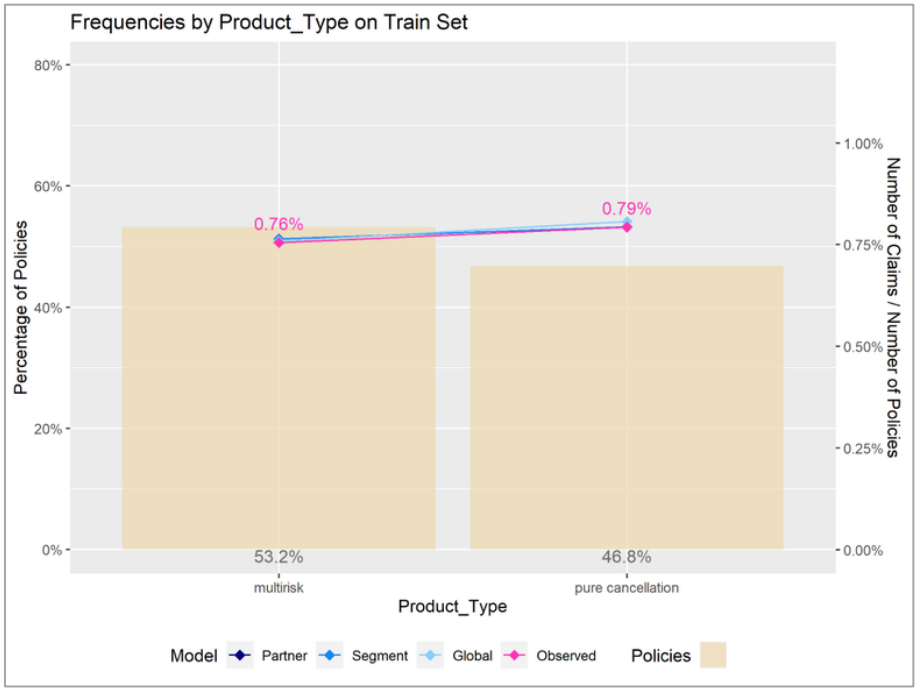
4.1 By month of subscription



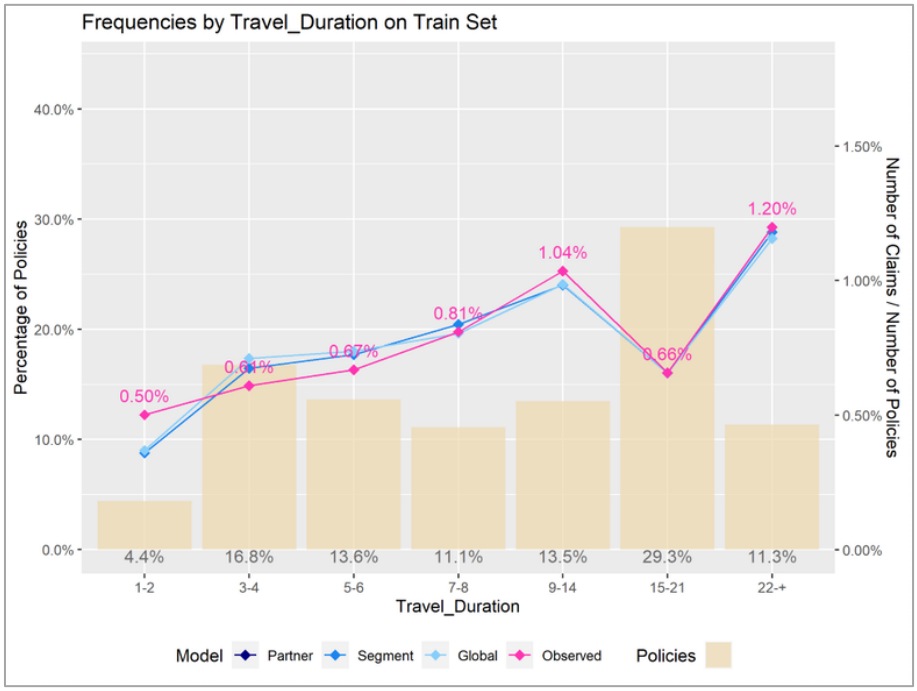
4.2 By number of beneficiaries



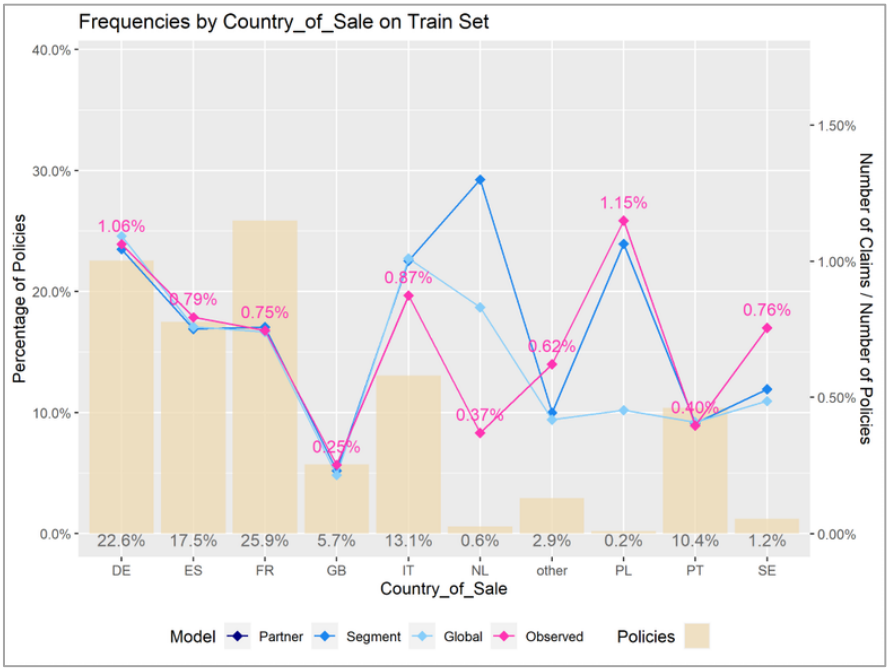
4.3 By type of cancellation product



4.4 By travel duration

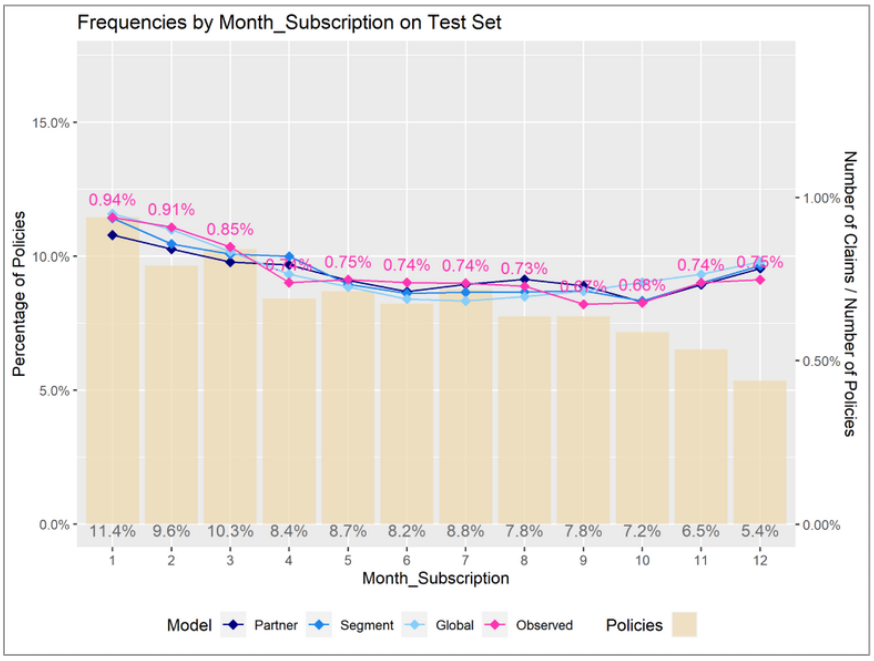


4.5 By country of subscription

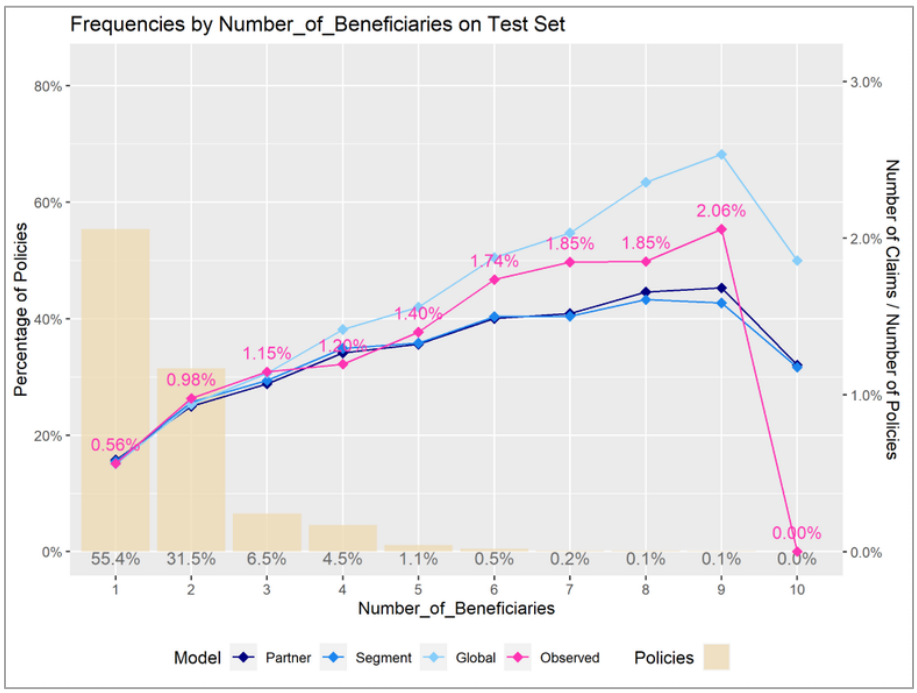


5. Observed vs predicted – Partner #16 (Test set)

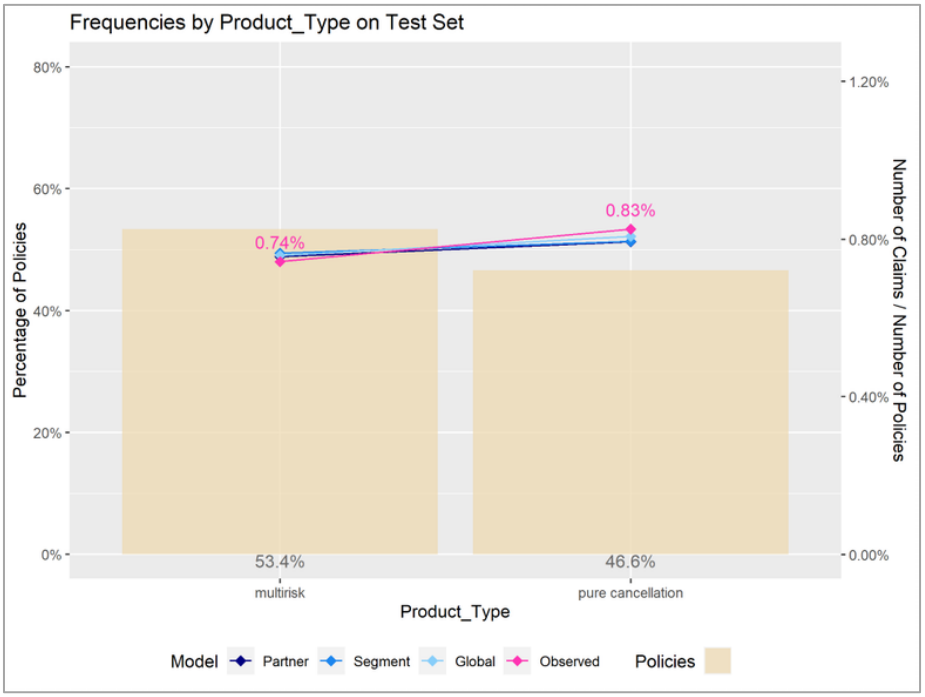
5.1 By month of subscription



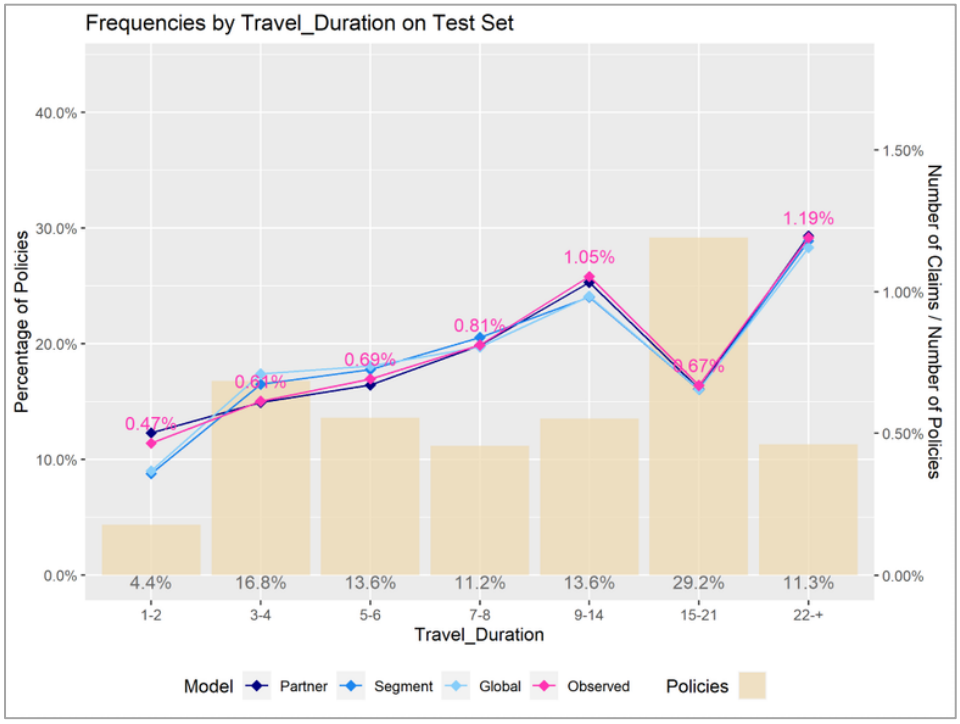
5.2 By number of beneficiaries



5.3 By type of cancellation product



5.4 By travel duration



5.5 By country of subscription