Recherche opérationnelle en data science appliquée en assurance au cas d'un portefeuille assurance-vie

~ Retour d'expérience d'un assureur vie ~

Présentation du 26 novembre 2015

COLLOQUE « ACTUARIAT & DATA SCIENCE

Anani OLYMPIO
MsBA, Actuaire Certifié de l'Institut des Actuaires, Expert ERM CERA
Responsable R&D et Data'Lab Groupe
Groupe CNP Assurances
Tel : +33.1.42.18.77.20
Email : anani.olympio@cnp.fr

AGENDA

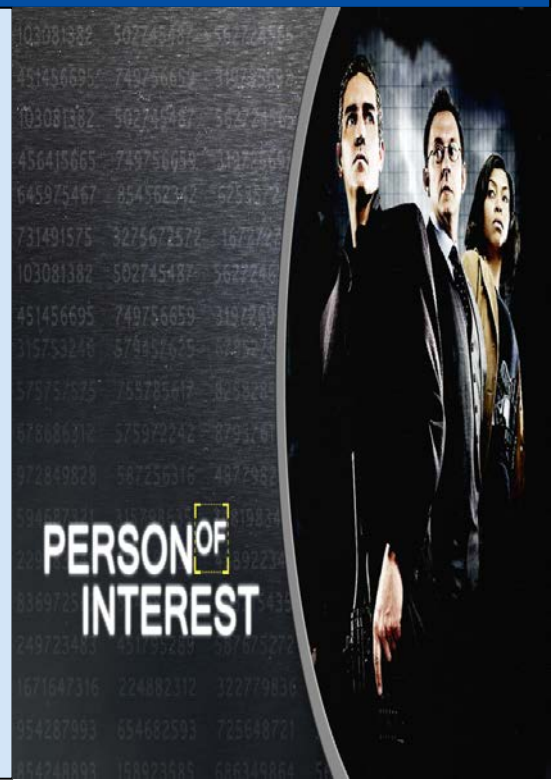
1	Contexte et enjeux pour CNP Assurances
2	De la théorie à la pratique, quelques repères…
3	Quelques exemples de *Proof of Concept*
4	Conclusion

Contexte :
Quelques faits et chiffres clés du Big Data…

DES ALGORITHMES POUR PREVOIR ET COMPRENDRE LA CRIMINALITE

❑ **Initiative de la Police de Los Angeles**

- ✓ La police de Los Angeles a rassemblé les données relatives à plus de **130 millions de crimes ces 80 dernières années** et continue de mettre à jour le logiciel en ajoutant les nouveaux crimes
- ✓ Les résultats obtenus ont notamment montré :
  - o *réduction de 33% du nombre de vols*
  - o *réduction de 21% des crimes violents*
  - o *réduction de 12% des crimes contre les biens*
  - o *meilleure compréhension des raisons pour lesquelles des crimes sont commis dans certaines zones*
- ✓ Actuellement, ce logiciel est utilisé dans de nombreuses villes (Royaume Uni, Etats-Unis, Canada, Australie, France, Italie, Chine…)



Contexte :

La Data science est un changement de perspective sur les données, mais nécessite de faire la différence entre mythes et réalité!

People thinks Big data is :

Analyse social media data !	The major value is inside your database. Look at external data after
Analyse unstructured data !	You can store unstructured data but algorithm still need structure to work
An explosion of the volume	Only if you change the way you describe the object of the analysis
Business intelligence 2.0	BI focus on reporting. Big Data shall focus on prediction
Analyse petabytes of data	Bigger is better but smart analysis on a « not so big » dataset can give very impressive results

La révolution Big Data & Data Science, c'est d'abord une amélioration continue de tous les processus d'affaires, grâce à une utilisation intelligente des données, tout le temps, partout et à toutes les fins…

Enjeux pour CNP assurances :
Quelles sont les enjeux business du Big data pour l'industrie des assurances

QUESTION N°1	Quelles sont nos analyses des faits générateurs de profits et de pertes sur le business?
QUESTION N°2	Quel est notre compréhension du rôle des données et de les enjeux économiques que cela peut représenter pour le business?
QUESTION N°3	Quels sont nos accès à nos données clients ? A quelle fréquence?

➔ **L'ambition de Big Data c'est d'aider l'entreprise à répondre à ses *Business Case* par les moyens innovants proposés par la Data science…**

Enjeux pour CNP assurances :

La démarche Big Data & Data science est une question de stratégie !

VISION

Etre une compagnie Digitale de référence avec une démarche Big Data & data science orientée *Business Driven*

MISSIONS

Mener des initiatives dans le cadre des démarches « **Ambiton #Digital** » et « **Data'Lab** » **intégrées** aux processus de décision et de production afin d'accompagner le Groupe dans son ensemble pour un développement durable et rentable.

- Connaissance clients / Opérations marketing
- Produits / Rentabilité des portefeuilles
- Optimisation des Processus / Gestion de la relation client, des contrats et des sinistres

VALEURS

- Respect des règles d'éthique concernant l'utilisation des données personnelles
- Client au cœur, Inventivité, Initiative, Confiance

Enjeux pour CNP assurances :

La démarche Big Data & Data science c'est la création d'une Data'Lab!

ROLE

Vecteur idéal de diffusion de la démarche Big data & Data science au sein de l'entreprise dans son ensemble

ACTIONS

APPROCHE EXPERIMENTALE ORIENTEE OPTIMISATION ET RESULTATS

➔ CONCEVOIR DES **PROOF OF CONCEPT** (POC) =

- **Use Case**
- **Délais raisonnable de réalisation (3 mois)**
- **Contributeurs (Data'Lab, métiers, Directions…)**
- **Engagement et appuis (au plus haut niveau)**
- **Processus itératif (méthode AGILE…)**

VALORISATION

Communication (interne/externe)

Mise en production des succès (sous contrainte de rentabilité)

AGENDA

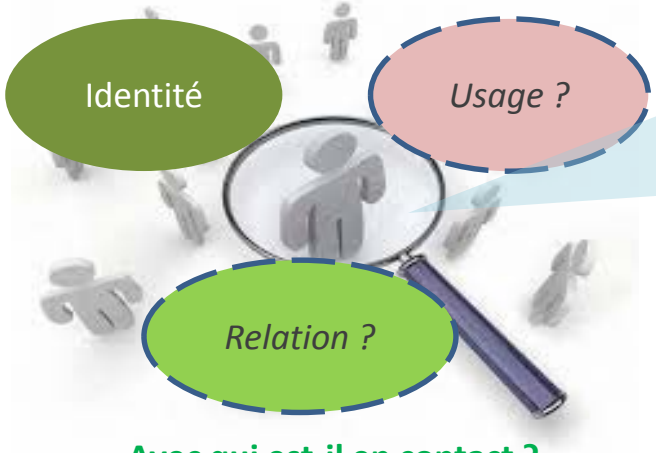
1	Contexte et enjeux pour CNP Assurances
2	De la théorie à la pratique, quelques repères…
3	Quelques exemples de *Proof of Concept*
4	Conclusion

De la théorie à la pratique, quelques repères…:
Démarche globale de la modélisation data science

❑ **Comment modéliser les comportements des clients ? Quelles données utilisées ?**

Nom, prénom, adresse, CSP, etc.

Comment utilise-t-il mes produits et services ?



Avec qui est-il en contact ?
Qu'est-ce qui l'influence ?
Qui influence-t-il ?

Historiquement dans les services financiers, le client est principalement décrit par sa donnée identitaire.

➔ L'approche analytics permet d'utiliser les apports des variables comportementales et d'influences.

❑ **Comment représenter les caractéristiques des clients ?**

✓ Un **CLIENT est représenté par une ligne de la matrice composée des variables suivantes :** caractéristiques **statiques** (identité) + **dynamiques** (transactions, conjoncture, etc.) liées à une période donnée P (par exemple 5 ans)

De la théorie à la pratique, quelques repères…:
Focus sur les modèles de prédiction

❑ Quelles familles de modèles ?

✓ La modélisation en Data Science est fondée sur la règle simple suivante : « **il n'y a pas a priori de meilleurs modèles ou méthodes, mais une multitude, pour un use case donné!** »

✓ Se qui en pratique se matérialise par une association de :

- ➢ *Une quinzaine de familles de modèles disponibles, avec de multiples variantes possibles…*
- ➢ *possibilité de créer ses propres modèles ou d'optimiser des modèles existants…*

Familles de modèles… ✚ **Variantes possibles…** ✚ **Optimisation / Modèles propres…**

- ➔ K plus proches voisins
- ➔ Support Vector Machines
- ➔ Réseaux de neurones
- ➔ Arbres de décisions
- ➔ Régressions logistiques
- ➔ …

- ➔ Gradient Boosting Machine
- ➔ Bagging
- ➔ Boosting
- ➔ Forêts aléatoires
- ➔ Stacking / blending
- ➔ …

Exemple :
Partir d'une approche Boosting en surpondérant les tirages sur les mauvais classements uniquement.

De la théorie à la pratique, quelques repères…:
Focus sur les modèles de prédiction

❑ Place prépondérantes des données

- ✓ Dans la réussite d'un projet Data Science, le **poids des données est bien plus important que celui des modèles**

- ✓ Ces poids peuvent beaucoup varier, mais **un rapport « moyen » pourrait être de 70% / 30%** :
  - ➢ *entre 70% et 80% sur les données (collecte, traitements, analyses, futures engineering…)*
  - ➢ *entre 20% et 30% sur les modèles (construction, calibrage, optimisation …)*

- ✓ Les **deux axes d'analyse** du poids des données :
  - ➢ ***La puissance intrinsèque du signal*** *(avec un* ***signale très faible*** *et une* ***volumétrie insuffisante*** *de données, même un* ***modèle optimal aura des performances médiocres****)*

  - ➢ ***L'amplification du signal*** *(en* ***augmentant le volume des données****, un* ***modèle non optimal****, quel que soit la famille à laquelle il appartient, obtiendra de* ***bien meilleurs résultats*** *que dans le cas d'un signal faible)*

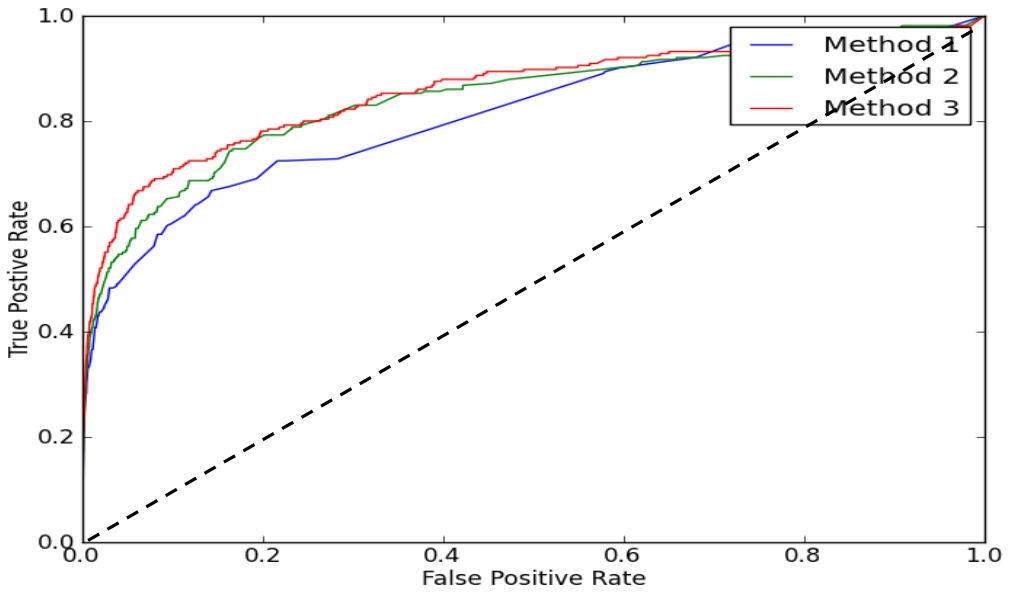
De la théorie à la pratique, quelques repères…:
Focus sur les modèles de prédiction

❑ Evaluation et performance (les métriques utilisées)

- ✓ **Aire sous la courbe ROC (AUC)** :
  - ➢ *La courbe ROC est un espace (taux de vrai positif, taux de faux positif) de représentation de modèle à sorties binaires pour expliquer son comportement en fonction du seuil de décision*
  - ➢ *L'Aire sous le courbe ROC ou AUC offre ainsi une vue globale du modèle sans se soucier du seuil de décision*

❑ Appréciation du modèle :

- ✓ *0.90 - 1.00* = excellent (A)
- ✓ *0.80 - 0.90 = good (B)*
- ✓ *0.70 - 0.80 = fair (C)*
- ✓ *0.60 - 0.70 = poor (D)*
- ✓ *0.50 - 0.60 = fail (F)*



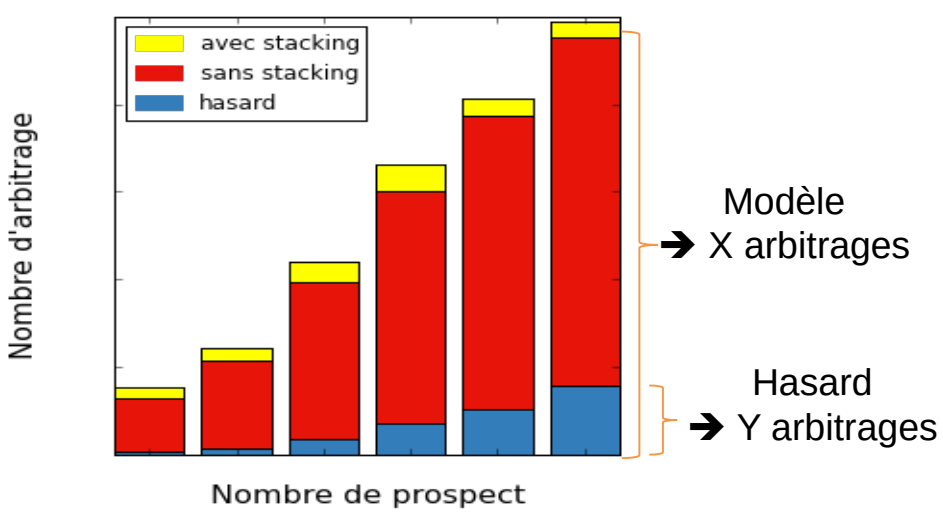
De la théorie à la pratique, quelques repères…:
Focus sur les modèles de prédiction

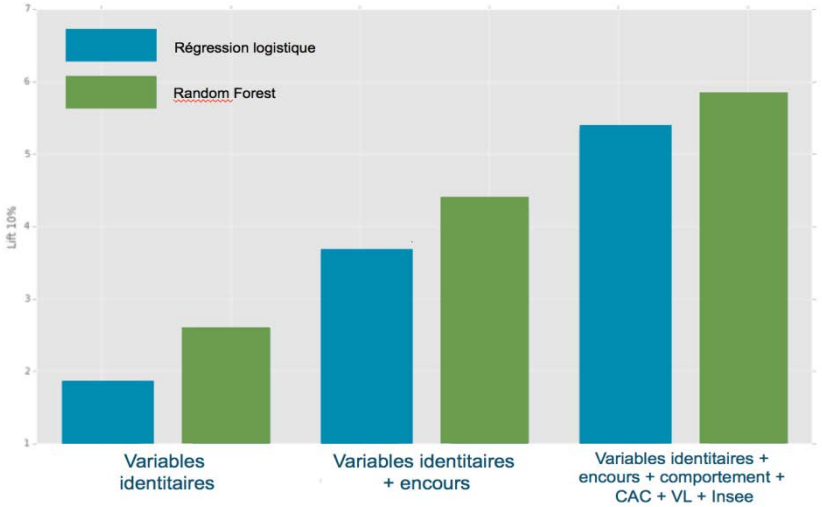
❑ Evaluation et performance (les métriques utilisées)

✓ Le Lift :

- ➢ *Le Lift permet d'apprécier à quel point le modèle est meilleur que l'aléatoire*
- ➢ *Le Lift est plus important si la cible retenue diminue*
- ➢ *Le lift est très utilisé en marketing (et permet de faire un ciblage des campagnes efficaces)*

Lift 10%[1] = X / Y





AGENDA

1	Contexte et enjeux pour CNP Assurances
2	De la théorie à la pratique, quelques repères…
3	Quelques exemples de Proof of Concept
4	Conclusion

Quelques exemples de PoC :
Etude réalisée sur des portefeuilles d'assurance vie

PoC N°1

❑ **Use case** : Ciblage des assurés arbitreurs au 1er semestre 2014 (euros vers UC)

❑ **Contexte** :

- Contrat d'épargne multi support grand public
- Taux de cible d'arbitrage au S1 : inférieur à 1%
- Volume de données disponibles : plus de 40 GB
- Historique : 01/01/2009 -31/12/2014

❑ **Principaux enseignements** :

- Mesure du Lift à 10% (au moins 7 fois mieux que le hasard!)

- Près de 80% des arbitreurs conservent l'intégralité de leur UC après 1 an vs 60% après 3 ans

- Près de 60% des arbitreurs sont des primo arbitreurs (assuré n'ayant jamais arbitré avant)

➔ le modèle prédit au moins 40% des primo-arbitreurs

- Le modèle prédit également :
  - ✓ la quasi-totalité des assurés ayant arbitré au moins une fois
  - ✓ des arbitreurs S1 et des arbitreur S2
  - ✓ des assurés ayant réalisé des Versements Libres au cours de l'année

Quelques exemples de PoC :
Etude réalisée sur des portefeuilles d'assurance vie

PoC N°2

- **Use case** : Identification des assurés susceptibles de réaliser un rachat total ou partiel (Euro ou UC) au cours des 6 prochains mois

- **Contexte** :
  - Contrat d'épargne multi support grand public
  - Taux de cible d'arbitrage au S1 : inférieur à 1%
  - Volume de données disponibles : plus de 40 GB
  - Historique : 01/01/2009 -31/12/2014

- **Principaux enseignements** :
  - Mesure du Lift à 10% (au moins 6 fois mieux que le hasard)
  - La tendance à racheter dépend des activités passées de l'assuré (nombre de rachats passés)
  - Près de 50% des racheteurs sont des primo racheteurs

  ➔ le modèle prédit au plus de 40% des primo-racheteurs

AGENDA

1	Contexte et enjeux pour CNP Assurances
2	De la théorie à la pratique, quelques repères…
3	Quelques exemples de Proof of Concept
4	Conclusion

Conclusion

- **Facteurs clés de succès un processus en 5 étapes : AIMER**
  - *ANIMER*
  - *IMAGINER*
  - *MATERIALISER*
  - *EXPLOITER*
  - *RELAYER*
- **La valeur majeure se trouve à l'intérieur de vos bases de données. Regardez et exploitez les d'abord avant de regarder les données externes**
- **Approche « Business Driven » est efficace :**
  - *Facilite la communication et la mobilisation de l'organisation jusqu'au plus haut niveau*
  - *Assure le succès de la démarche car prônant la recherche de use case sous un angle business*
- **Nécessité d'avoir le soutien des dirigeants et favoriser une démarche AGILE**
- **Plusieurs projets souhaités suite aux premières études**

REVUE BIBLIOGRAPHIQUE

❑ Big Data & Data science :

- ✓ *Anani Olympio, Romain Méridoux, travaux de R&D - Data'Lab CNP Assurances : Big Data et data science*
- ✓ *Eric Biernat (Auteur), Michel Lutz (Auteur), Yann LeCun (Préface) : Data science : fondamentaux et études de cas : Machine learning avec Python et R Broché – 1 octobre 2015*
- ✓ *de Pirmin Lemberger (Auteur), Marc Batty (Auteur), Médéric Morel (Auteur), Jean-Luc Raffaëlli (Auteur) : Big Data et Machine Learning - Manuel du data scientist Broché – 18 février 2015,*
- ✓ *Hastie Trevor, Tibshirani Robert, Friedman Jerome : The Elements of statistical learning, Data mining, Inference and Prediction, 2nd edition, Springer*
- ✓ *Publication de Xebia IT Architects : Imaginer, Matérialiser, Exploiter, TechTrends n° 6, mars 2015 – (xebia.fr)*
- ✓ *Dries De Dauw : Customer Intelligence and Analytics, From data to insight to value, 21/12/2015, AG Insurance*
- ✓ *Sources INSEE*