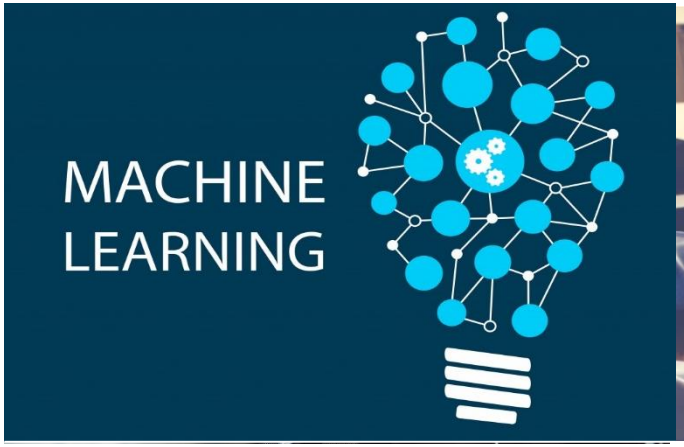
Comparaison de modèles prédictifs pour l'évaluation des coûts matériels automobiles

Maxence de LUSSAC
Bordeaux, le 30 mars 2018



Préambule

- **L'assurance automobile, bien que bénéficiaire, effectue des pertes techniques.**
  - Un ratio combiné d'en moyenne 105% les 10 dernières années
- **Ces pertes techniques sont rattrapées par le résultat financier. D'autres moyens sont également mis en œuvre :**
  - Perfectionnement des modèles de tarification
  - Maîtrise des coûts de gestion
- **Mission : analyse de l'influence d'un réseau d'experts automobiles sur le coût du sinistre**
- **2 problématiques :**
  - Analyse de l'influence d'une variable sur le coût du sinistre (prestataire de services) ;
  - Construction et optimisation d'un modèle prédictif.

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

1. Cadre de travail (1/2)

Présentation de la base de données :

- **Agrégation des portefeuilles de sinistres de 3 assureurs.**
  - Historique : 2015 et 2016.
  - Environ 850 000 observations après retraitement.

- **Dans cette base de données, on retrouve**
  - Les informations concernant le véhicule, le type d'accident, le garage affilié, le lieu, la date d'occurrence du sinistre
  - Le réseau d'experts ayant pris en charge le sinistre (variable cible)
  - Le montant d'expertise (variable à expliquer)

1. Cadre de travail (2/2)

Pour la modélisation :

- **Division de la base de données en 3 échantillons**
  - Echantillon d'apprentissage.
  - Echantillon de validation.
  - Echantillon test.
- **Indicateurs de performance de prédiction**
  - Mesure de corrélation (COR) : A maximiser.
  - Erreur moyenne quadratique (MSE) : A minimiser.
  - Norme L1 : A maximiser.

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

2.a) Modèle linéaire généralisé

- **Le GLM a été choisi pour sa réponse simple à l'évaluation de l'influence de la variable cible**
  - Loi Gamma
  - Fonction de lien log

$$E[Y|X] \approx (1+\beta_1)^{X_1} * \cdots * \left(1+\beta_p\right)^{X_p}.$$

- **Ecrêtage des valeurs extrêmes**
  - Etablissement d'un seuil à droite de 30 000 € et d'un seuil à gauche de 334 €.

- **Sélection ascendante et descendante selon le critère AIC des variables explicatives**
  - Toutes les variables sont conservées.

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

2.b) Agrégation de modèles : RGLM (1/2)

- **Objectif : améliorer le pouvoir prédictif du GLM.**

- **Algorithme : boucle de T itérations**
  - Création de l'échantillon *bootstrap*
  - Sélection aléatoire des variables, paramètre q.
  - Construction du GLM
  - Stockage des prédictions

- **Fonction réponse :**

$$E[Y|X] = \frac{1}{T}\sum_{j=1}^{T} exp(\beta_{1j}X_1 + \beta_{2j}X_1 + \cdots + \beta_{pj}X_1)$$

$$\Rightarrow E[Y|X] \approx \frac{1}{T}\sum_{j=1}^{T}\underbrace{\left(\prod_{i=1}^{p}(1+\beta_{ij})^{X_i}\right)}_{\text{Prédictions du j}^{\text{ème}}\text{ GLM}}$$

2.b) Agrégation de modèles : RGLM (2/2)

- **Ajustement de l'algorithme :**
  - Introduction des termes d'interaction.
  - Passage à une régression pénalisée de type LASSO.
    - **Ajout de la contrainte** $\sum_{j=1}^{p} |\beta_j|^{\delta} \leq C,$
  - Méthode d'agrégation « meilleur modèle ».
- **Deux paramètres supplémentaires :**
  - Nombre d'échantillons bootstrap
  - Nombre q de variables sélectionnées aléatoirement
- **Optimisation de q :**

q	3	4	5	6	7	**8**	9
COR	0,62	0,62	0,64	0,65	0,66	**0,66**	0,66
MSE	3 135 499	3 141 007	2 993 409	2 955 974	2 909 052	**2 900 468**	2 902 814
L1	759	758	726	728	721	**716**	717

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

2.c) Gradient boosting model (1/2)

Algorithme de minimisation de la fonction de perte quadratique(1)

$$f(x) = \underset{\varphi \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \varphi(x_i)\right)^2$$

- **Initialisation :** $\forall i \in [1, n], \hat{f}(x_i) = \frac{1}{n} \sum_{j=1}^{n} y_j$
- **Boucle de T itérations :**
  - Calcul du gradient $z_i = \frac{\partial}{\partial \rho} L(y_i, \rho)|_{\rho = \hat{f}(x_i)}$
  - Ajustement d'un arbre de régression g prédisant le gradient à partir des variables x
  - Actualisation $\hat{f}(x_i) \leftarrow \hat{f}(x_i) + \lambda * g(x_i)$

(1) *The State of Boosting*, Greg Ridgeway

2.c) Gradient boosting model (2/2)

Optimisation des paramètres de l'algorithme

- **Deux contraintes :**
  - Nombre importants de paramètres
    - **Nombre d'itérations, le coefficient d'apprentissage, la profondeur maximum des arbres, tirage sans remise de l'échantillon…**
  - Paramètres interdépendants
- **Etape 1 : recherche des valeurs candidates (bornes)**
- **Etape 2 : optimisation par recherche aléatoire dans une grille**
  - +6 millions de combinaisons possibles
  - 149 modèles suffisent à obtenir un excellent modèle avec un probabilité de 99,5 %

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Évaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

3.a) Évaluation de l'influence d'une variable (1/7)

- **GLM : Estimation du coefficient de la variable**
  - Variable X_i : présence du réseau d'experts (binaire : 1 / 0).
  - Coefficient β_i : écart de performance entre les deux concurrents.

$$E[Y|X] \approx (1+\beta_1)^{X_1} * \cdots * (1+\beta_p)^{X_p}$$

- **RGLM : perte de l'information**
  - par la méthode d'agrégation
  - par les variables d'interactions

- **GBM : influence de la variable dans la construction des arbres**

3.a) Évaluation de l'influence d'une variable (2/7)

Exemple d'étude : Ecart de performance entre deux prestataires de services.

- **Evaluation du coefficient β_i : - 2,4%.**

$$Coût\ du\ sinistre \approx (1 + \beta_i)^{X_i} * \mathrm{C}$$

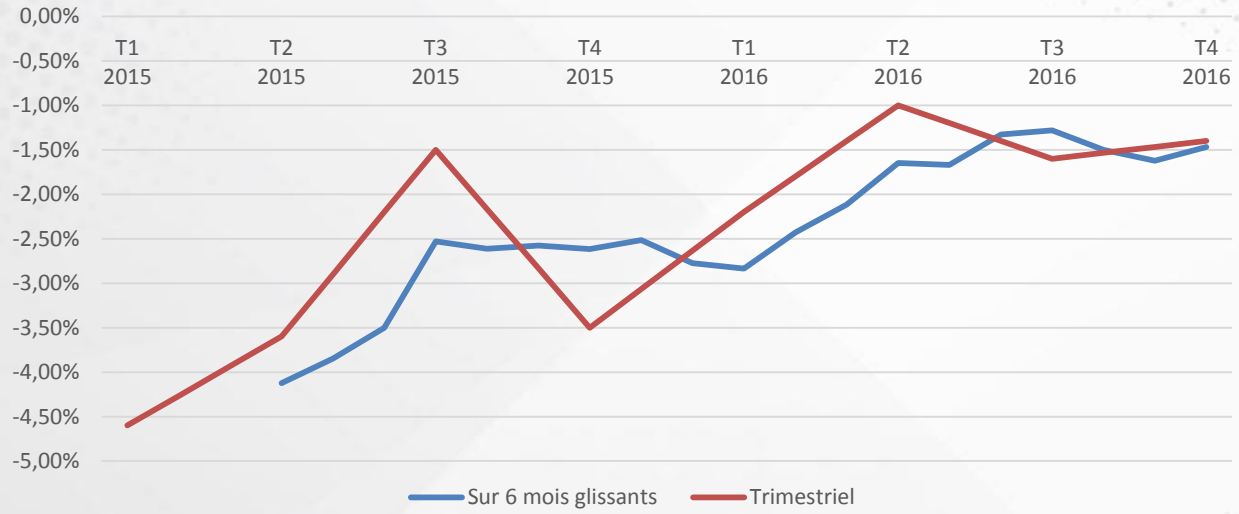
  - C constante qui ne dépend pas de X_i.
  - $X_i = 1$: le coût du sinistre baisse de 2,4%.

- **Le β_i représente l'écart de performance entre les deux réseaux d'experts.**

- **Intérêt de la méthode : évaluation « isolée » de l'influence du réseau d'experts**
  - Il ne suffit pas de comparer les coûts moyens
  - Méthode analytique similaire : comparé les coûts moyens de chaque sous-segment.

3.a) Évaluation de l'influence d'une variable (3/7)

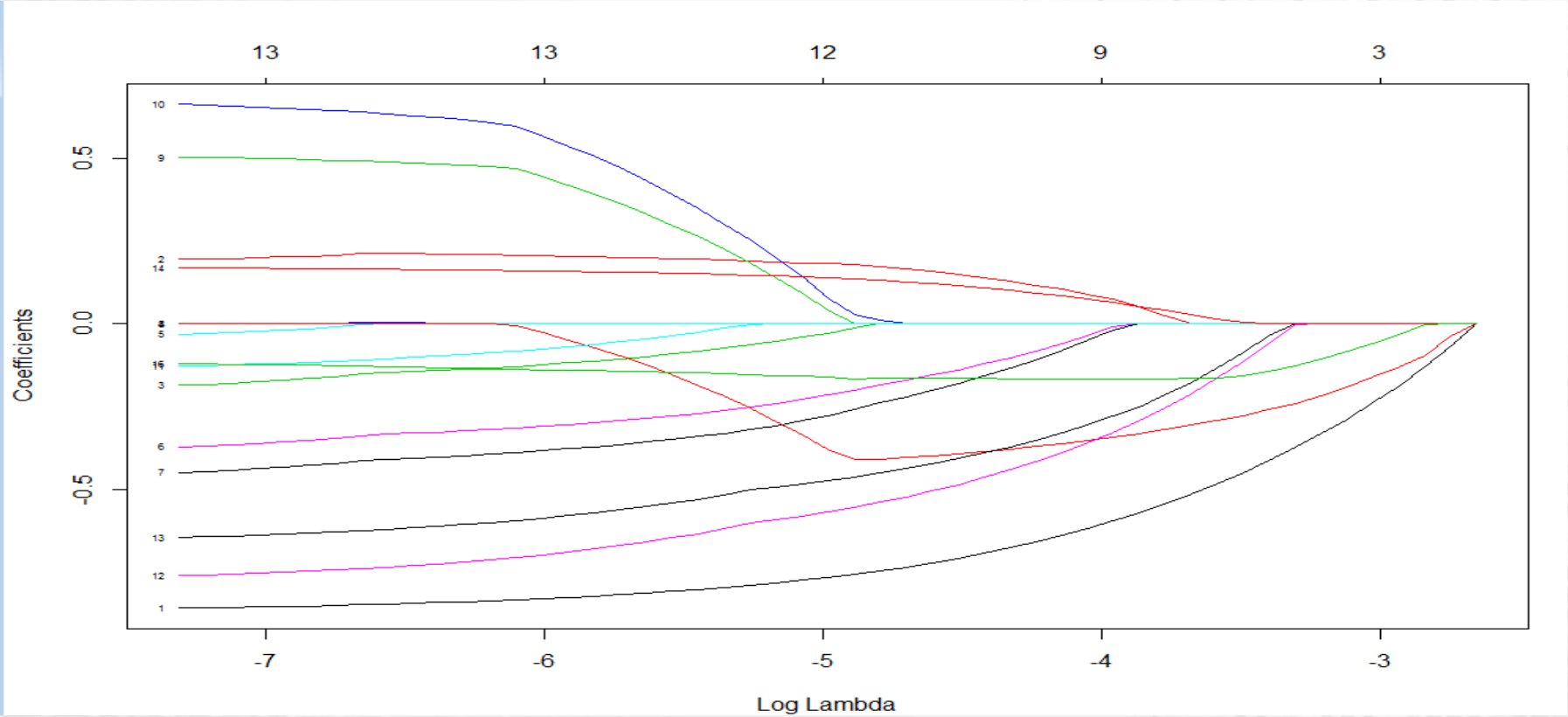
- **Suivi trimestrielle de l'écart de performance (Courbe rouge).**
  - Calcul du coefficient β_i chaque trimestre.
  - Vision instantanée.



- **Etude sur 6 mois glissants (Courbe bleue).**
  - Vision tendancielle avec une historique plus important.

3.a) Evaluation de l'influence d'une variable (4/7)

- **RGLM : perte de l'information avec les ajustements**
  - Retrait des interactions
  - Retrait de la pénalisation



3.a) Évaluation de l'influence d'une variable (5/7)

- **Amélioration de l'agrégation**
  - Nécessité de comparer la qualité des GLM entre eux
  - OOB error : erreur de prédiction sur les observations écartées par la méthode bootstrap
    - **Attribution d'un poids :**

$$p_i = \frac{1/OOBe_i}{\sum_{i=1}^{T}(\frac{1}{OOBe_i})}$$

- **Résultat : le coefficient oscille entre -2,2 % et -2,6 %.**

- **Agrégation : -2,38 % (contre -2,44 % avec la méthode classique).**
  - Résultats a priori plus précis (expérience)
  - Prédictions équivalentes au GLM

3.a) Évaluation de l'influence d'une variable (6/7)

GBM

- **Dans l'arbre A, l'importance de la variable j est déterminé par(2)**

$$\hat{I}_j^2(A) = \sum_{t=1}^{J} i_t^2 1_{v_t=j}$$

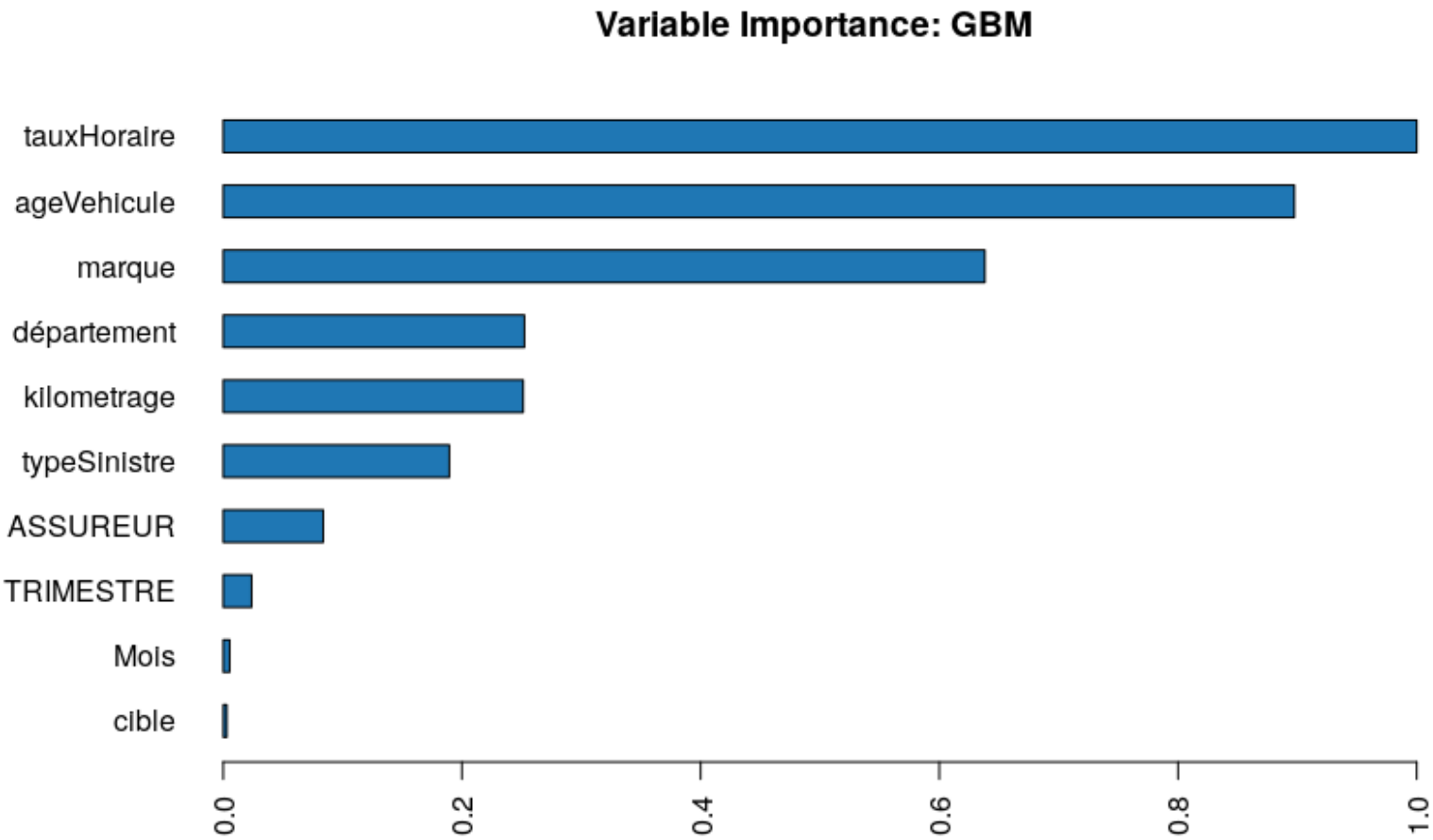
  - **J le nombre de nœuds non terminaux de l'arbre A,**
  - **v_t la variable de décision servant à la division du nœud t,**
  - **i_t^2 est le critère d'amélioration des moindres carrés résultant de cette division.**

- **Importance de la variable j dans l'algorithme**

$$\tilde{I}_j^2 = \frac{1}{T} \sum_{m=1}^{T} \hat{I}_j^2(A_m)$$

(2) **Breiman, Friedman, Olshen and Stone. 1983.** *Classification and Regression Trees.* s.l. : Wadsworth, 1983.

3.a) Évaluation de l'influence d'une variable (7/7)



Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

3.b) Prédiction des modèles et tarification (1/3)

- **Valeurs des indicateurs sur l'échantillon test :**

	COR	MSE	L1
GLM	0,55	3 639 942	770
RGLM	0,66	2 900 468	716
GBM	**0,73**	**2 411 535**	**651**

- **Le GBM est supérieur aux deux autres modèles pour les trois indicateurs.**

3.b) Prédiction des modèles et tarification (2/3)

- **Balance entre qualité de prédiction, d'exécution et d'interprétation.**

	Prédiction	Exécution	Interprétation	Implémentation
GLM	-	-	++	++
RGLM	+	+	+	--
GBM	++	++	--	-

- **GBM : difficile de retracer le détail des calculs, paramétrisation parfois confuse.**

- **Actuaire tarificateur : utiliser le modèle adapté à la situation.**
  - Délais de livraison
  - Degré de vulgarisation
  - Contrainte de performance

3.b) Prédiction des modèles et tarification (3/3)

Application des modèles en pratique :

- **Projet de courte durée pour la mise en place d'un modèle de tarification :**

→ ***GLM***

- **Projet de longue durée avec contrainte de vulgarisation :**

→ ***RGLM***

- **Construction d'un modèle interne avec contrainte de performance :**

→ ***GBM***

Sommaire

1. Cadre de travail

2. Modélisation

2.a) Modèle linéaire généralisé (GLM)

2.b) Agrégation de modèles (RGLM)

2.c) Gradient boosting model (GBM)

3. Analyse des résultats

3.a) Evaluation de l'influence d'une variable

3.b) Prédictions des modèles et tarification

4. Conclusion

Conclusion (1/2)

- **2 problématiques :**
  - Analyse de l'influence d'une variable sur le coût du sinistre
  - Optimiser un modèle de prédiction dans une optique de tarification
- **3 modèles : GLM, RGLM et GBM**
  - Fonctionnement, paramétrisation et optimisation
- **Le GLM apparait comme le plus pertinent pour la première problématique.**
  - Estimation d'un coefficient qui renseigne directement l'influence de la variable sur le coût du sinistre.
- **La réponse de la deuxième problématique dépend des besoins de l'utilisateur.**
  - Besoin de vulgarisation
  - Contrainte de temps
  - Contrainte de performance

Conclusion (2/2)

Limite de l'étude : faible nombre de variables explicatives

Mis en évidence par une segmentation insuffisante

montExp	marque	kilometrage	cible	typeSinistre	ageVehicule	TRIMESTRE	ASSUREUR	Mois	département	tauxHoraire	densite
870,29	RENAULT	15710	O	ACT	0,79945243	T3 2015	Assureur_1	201507	62	TH fort	221,470394
1137,29	RENAULT	15845	O	ACT	0,89801506	T2 2016	Assureur_1	201606	62	TH fort	221,470394
863	RENAULT	26055	O	ACT	1,60985626	T3 2015	Assureur_7	201509	62	TH fort	221,470394
1439	RENAULT	26848	O	ACT	1,83436003	T1 2015	Assureur_7	201503	62	TH fort	221,470394
3273,13	RENAULT	36774	O	ACT	2,23408624	T1 2016	Assureur_3	201602	62	TH fort	221,470394
4578	RENAULT	36776	O	ACT	2,11362081	T1 2016	Assureur_7	201602	62	TH fort	221,470394
413,43	RENAULT	48734	O	ACT	2,9596167	T3 2016	Assureur_1	201609	62	TH fort	221,470394
1753	RENAULT	49480	O	ACT	3,02806297	T1 2015	Assureur_7	201501	62	TH fort	221,470394

Impact sur l'étude

- Erreur de prédiction « absolu » important
- Pas d'impact sur l'analyse d'influence d'une variable
- La comparaison des modèles reste inchangée

Prim' Act

Merci de votre attention.

DAUPHINE
UNIVERSITÉ PARIS