Rapport d'Analyse et Prévision des Réserves d'Assurance

Étude du dataset Porto Seguro's Safe Driver Prediction

Rapport d'analyse

Jules Odje

11 avril 2025

Table des matières

Chapitre 1

Introduction

Ce rapport détaille l'analyse du dataset Porto Seguro's Safe Driver Prediction, utilisé pour développer un modèle de prévision des réserves d'assurance. L'objectif principal est de prédire les risques de sinistres et d'estimer les réserves financières nécessaires pour couvrir ces risques potentiels.

Le dataset contient des informations anonymisées sur les assurés, avec une variable cible binaire indiquant si un client a déposé une réclamation d'assurance l'année suivante (1) ou non (0).

Chapitre 2

Exploration des données

2.1 Structure du dataset

Le jeu de données comprend 595 212 entrées avec 59 colonnes, incluant :

- Un identifiant unique (`id`)
- La variable cible (`target`)
- 57 variables prédictives, dont les noms sont codés pour des raisons de confidentialité

Les variables sont regroupées en catégories :

- Variables individuelles (`ps_ind_XX`)
- Variables de véhicule (`ps_car_XX`)
- Variables régionales (`ps_reg_XX`)
- Variables calculées (`ps_calc_XX`)

Les suffixes indiquent le type de variable :

- `_bin` : variables binaires
- `_cat` : variables catégorielles
- Aucun suffixe : variables continues ou discrètes

2.2 Analyse statistique descriptive

L'analyse statistique a révélé les caractéristiques suivantes :

- Toutes les variables sont de type numérique (int64)
- Aucune valeur NaN ou NULL n'a été détectée dans le dataset
- La distribution de la variable cible est fortement déséquilibrée :
  - Classe 0 (non-sinistres) : 573 518 observations (96,36%)
  - Classe 1 (sinistres) : 21 694 observations (3,64%)

Ce déséquilibre est typique des problèmes d'assurance où les sinistres sont des événements relativement rares.

Chapitre 3

Analyse des valeurs spéciales

3.1 Identification des valeurs -1

Nous avons identifié la présence de valeurs -1 dans plusieurs variables catégorielles :

- `ps_ind_02_cat` : 216 valeurs -1 (0,04% des données)
- `ps_ind_04_cat` : 83 valeurs -1 (0,01% des données)
- `ps_ind_05_cat` : 5809 valeurs -1 (0,98% des données)

3.2 Signification des valeurs -1

L'analyse de la distribution de la variable cible en fonction de ces valeurs -1 a révélé :

Variable	**Condition**	**Distribution (classe 0 / classe 1)**
2*`ps_ind_02_cat`	= -1	81,48% / 18,52%
	≠ -1	96,36% / 3,64%
2*`ps_ind_04_cat`	= -1	60,24% / 39,76%
	≠ -1	96,36% / 3,64%
2*`ps_ind_05_cat`	= -1	91,67% / 8,33%
	≠ -1	96,40% / 3,60%

TABLE 3.1 – Distribution de la variable cible selon les valeurs -1

Ces résultats indiquent que les valeurs -1 ne sont pas des valeurs ordinaires (Je n'ai pas pu trouver le pourquoi de cette notations), mais constituent un signal prédictif fort à voir le tableau ci-dessus. Les observations avec des valeurs -1 dans ces colonnes présentent un risque de sinistre significativement plus élevé que les autres.

Chapitre 4

Analyse des corrélations

4.1 Matrice de corrélation

L'analyse des corrélations entre les variables a révélé :

— Des corrélations généralement faibles avec la variable cible (toutes inférieures à 0,06)
— Quelques clusters de corrélation entre certaines variables explicatives

4.2 Variables les plus corrélées avec la cible

Variables	Corrélation avec la cible
ps_car_13	0,054
ps_car_12	0,039
ps_ind_17_bin	0,037
ps_reg_02	0,034
ps_ind_07_bin	0,034
ps_car_07_cat	-0,036
ps_ind_06_bin	-0,034
ps_car_02_cat	-0,032
ps_ind_16_bin	-0,028
ps_ind_15	-0,022

TABLE 4.1 – Variables les plus corrélées (positivement et négativement) avec la cible

Les faibles corrélations suggèrent que des modèles linéaires simples ne seraient probablement pas optimaux pour ce problème.

Chapitre 5

Modélisation prédictive

Pour prédire les risques de sinistres, nous avons testé trois algorithmes d'apprentissage automatique :

5.1 Préparation des données

— Séparation des features et de la variable cible
— Utilisation de SMOTE (Synthetic Minority Over-sampling Technique) pour gérer le déséquilibre des classes

5.2 Modèles testés

Nous avons implémenté et comparé trois algorithmes :

5.2.1 Random Forest

— Modèle basé sur un ensemble d'arbres de décision
— Hyperparamètres : n_estimators=100, random_state=42

5.2.2 XGBoost

— Modèle de boosting basé sur les arbres de décision
— Hyperparamètres : n_estimators=100, learning_rate=0.1, max_depth=5, etc.

5.2.3 Gradient Boosting

— Algorithme de boosting qui construit des arbres de façon séquentielle
— Hyperparamètres : n_estimators=100, learning_rate=0.1, max_depth=5, etc.

5.3 Importance des variables selon les modèles

Random Forest		XGBoost		Gradient Boosting	
Variable	**Importance**	**Variable**	**Importance**	**Variable**	**Importance**
ps_reg_01	0,062	ps_ind_17_bin	0,080	ps_ind_08_bin	0,131
ps_reg_03	0,056	ps_calc_15_bin	0,073	ps_calc_03	0,079
ps_reg_02	0,051	ps_calc_20_bin	0,072	ps_ind_07_bin	0,066
ps_ind_06_bin	0,046	ps_ind_08_bin	0,069	ps_calc_02	0,060
ps_ind_09_bin	0,042	ps_calc_19_bin	0,064	ps_ind_09_bin	0,060

TABLE 5.1 – Top 5 des variables les plus importantes selon chaque modèle

Chapitre 6

Sélection du meilleur modèle

6.1 Comparaison des performances

Les modèles ont été évalués à l'aide de la validation croisée (5-fold) et de la métrique AUC-ROC :

Modèle	**AUC moyen**	**Écart-type**
Gradient Boosting	0,6389	0,0027
XGBoost	0,6388	0,0023
Random Forest	0,5863	0,0025

TABLE 6.1 – Performances des modèles en validation croisée

6.2 Analyse détaillée du meilleur modèle

Le Gradient Boosting a obtenu les meilleures performances, bien que très proches de XGBoost. L'analyse détaillée de ses prédictions a révélé :

	Prédiction	
Réalité	**Classe 0**	**Classe 1**
Classe 0	573 284	234
Classe 1	21 671	23

TABLE 6.2 – Matrice de confusion du modèle Gradient Boosting

Classe	**Précision**	**Recall**	**F1-score**
0 (non-sinistres)	0,96	1,00	0,98
1 (sinistres)	0,09	0,00	0,00
macro avg	0,53	0,50	0,49
weighted avg	0,93	0,96	0,95

TABLE 6.3 – Rapport de classification pour le Gradient Boosting

Cette analyse révèle une limitation importante : bien que le modèle classe relativement bien les risques (AUC de 0,64), il détecte très peu de sinistres réels (23 sur 21 694).

6.3 Approche ensemble

Pour atténuer les limitations des modèles individuels, nous avons adopté une approche d'ensemble en combinant les prédictions des trois modèles avec une pondération reflétant leurs performances :

- Random Forest : poids de 0,2
- XGBoost : poids de 0,4
- Gradient Boosting : poids de 0,4

Chapitre 7

Calcul des réserves estimées

7.1 Méthodologie

Pour calculer les réserves, nous avons :

1. Utilisé les probabilités pondérées issues de l'approche ensemble
2. Appliqué un coût moyen de sinistre de 10 000 unités monétaires
3. Segmenté les polices en 5 catégories de risque

7.2 Segmentation des risques

Les polices ont été classées en catégories selon leur probabilité de sinistre :

— Très faible : 0,004 à 0,070
— Faible : 0,070 à 0,112
— Moyen : 0,112 à 0,162
— Élevé : 0,162 à 0,205
— Très élevé : 0,205 à 0,686

Chapitre 8

Analyse des résultats

8.1 Répartition des réserves

Catégorie	Montant des réserves	Pourcentage	Nombre de polices	Probabilité moyen
Très faible	145 877 100	31,46%	357 127	0,041
Faible	105 108 700	22,67%	119 042	0,088
Moyen	79 795 810	17,21%	59 521	0,134
Élevé	53 833 320	11,61%	29 761	0,181
Très élevé	79 113 000	17,06%	29 761	0,266

Table 8.1 – Répartition des réserves par catégorie de risque

8.2 Statistiques globales

— Réserve totale estimée : 463 727 891,61 unités monétaires
— Nombre total de polices : 595 212
— Réserve moyenne par police : 779,10 unités monétaires

8.3 Distribution des probabilités de sinistre

La distribution des probabilités de sinistre est fortement asymétrique, avec une concentration importante vers les faibles probabilités (majorité des polices avec une probabilité inférieure à 0,1).

Chapitre 9

Conclusion et recommandations

9.1 Synthèse des résultats

Cette étude a permis de développer un modèle de prévision des réserves d'assurance basé sur l'analyse du dataset Porto Seguro. Le modèle d'ensemble final offre une capacité discriminante modérée (AUC ∼0,64) pour identifier les risques relatifs des différentes polices.

Les résultats montrent que :

— La majorité des polices (60%) présente un risque très faible
— Une minorité de polices (10%) concentre près de 30% des réserves estimées
— La réserve totale estimée s'élève à 463,7 millions d'unités monétaires

9.2 Recommandations

Pour la gestion des réserves d'assurance, nous recommandons :

9.2.1 Stratification des réserves

— Allouer les réserves en fonction des catégories de risque identifiées
— Porter une attention particulière au segment "Très élevé" qui représente un risque concentré

9.2.2 Utilisations pratiques

— Mettre en place un suivi régulier des polices à haut risque
— Utiliser les scores de risque pour ajuster la tarification
— Intégrer les estimations de réserves dans la stratégie financière globale

Ce dashboard de prévision des réserves constitue un outil d'aide à la décision précieux pour les actuaires et gestionnaires de risques, permettant une allocation plus précise des ressources financières et une meilleure anticipation des risques d'assurance.