Onglet 1

Projet Data & Économétrie

Titre du Projet :

Analyse et Modélisation des Données d'Assurance

Réalisé par :

Alan-Long Le Menelec

Clément Duran

Formation :

Diplôme Universitaire Big Data,

Data Science et analyse des risques sous Python

Etablissement :

Université de Montpellier - Faculté d'économie

Année universitaire :

2024-2025

Contexte

Dans le cadre d'un projet de data science appliquée au secteur de l'assurance, nous avons étudié différentes bases de données (contrats, sinistres, météo, données communales) en utilisant exclusivement Python. L'objectif était de nettoyer et enrichir les données, de créer une base unique cohérente puis de mener une analyse descriptive et économétrique afin de mieux comprendre les déterminants de la sinistralité.

a) Traitement et retraitement des données

Nous avons commencé par une exploration structurelle et statistique des différents jeux de données disponibles.

- **Valeurs manquantes** : les variables continues ont été imputées par la médiane (robuste aux valeurs extrêmes) et les qualitatives par la modalité dominante ou "NA" selon le contexte.
- **Valeurs aberrantes** : des traitements ont été appliqués sur des variables comme pol_bonus, vh_weight, ou vh_din avec exclusion ou correction.
- **Transformation des données** : typage, conversion de dates, suppression de doublons (par exemple sur id_policy, id_year) et nettoyage de variables superflues (comme Unnamed: 31).

Certaines valeurs de claim_amount étant négatives (≈ 1,2 %, probablement des régularisations ou remboursements), nous les avons remplacées par zéro afin de ne pas biaiser l'apprentissage du modèle tout en conservant les observations concernées

Des variables dérivées ont été créées pour renforcer l'analyse :

- id_policy pour croiser contrats et sinistres,
- has_claim (variable cible binaire),
- DEP (code département),
- geo_cluster via clustering KMeans sur les coordonnées GPS,
- moyennes climatiques agrégées (températures, précipitations, etc.) par département.

b) Jointures entre bases de données

Les jeux de données ont été fusionnés comme suit :

- **Contrats + sinistres** via l'identifiant id_policy,
- **Contrats + données climatiques** via la variable DEP (appliquant des moyennes mensuelles par département),
- **Contrats + géolocalisation** via retraitement des données geometry des communes.

La base finale obtenue, df_assurance, réunit ainsi les caractéristiques assurantielles, climatiques et spatiales, prête pour les analyses.

c) Analyse descriptive

Une première analyse univariée et multivariée a permis d'observer la répartition des variables clés, leur dispersion et les interactions potentielles avec les sinistres. Les outils utilisés incluent :

- Statistiques descriptives (moyenne, écart-type, etc.),
- Visualisations (histogrammes, boxplots, heatmaps),
- Groupements et moyennes conditionnelles (ex : sinistralité moyenne par type de carburant ou de couverture).

Deux approches de clustering ont été menées :

- geo_cluster, fondé sur les coordonnées GPS (clustering KMeans),
- Cluster, fondé sur un clustering hiérarchique sur des variables assurantielles et techniques.

Bien que non utilisé pour les modèles supervisés (la variable étant construite à partir de nb_sinistres), le cluster métier a montré une forte hétérogénéité en termes de sinistralité moyenne par segment.

d) Modélisation économétrique

Régression linéaire (OLS)

Nous avons modélisé le montant des sinistres (claim_amount) parmi les contrats sinistrés.

- **Variables explicatives** : caractéristiques techniques (valeur, puissance, ancienneté du véhicule), de contrat, du conducteur, et geo_cluster.
- **Tests réalisés** :
    - VIF pour identifier la multicolinéarité,
    - Histogramme et QQ-plot pour vérifier la normalité des résidus,
    - Test de Breusch-Pagan pour vérifier l'homoscédasticité (non rejetée, $p > 0.3$).

Le modèle OLS montre que des variables comme vh_din (puissance moteur), exposure (durée d'exposition) ont un effet significatif. Le modèle est globalement significatif.

Régression Logit

Nous avons modélisé la probabilité qu'un contrat ait au moins un sinistre (has_claim) via un modèle Logit. Les variables explicatives incluaient les caractéristiques véhicule, contrat et conducteur. L'efficacité a été évaluée par :

- Calcul de l'AUC via la courbe ROC : **AUC = 0.63**, indiquant une capacité de discrimination modérée,
- Rapport de classification (précision, rappel).

Conclusion

Cette étude a permis de :

- Nettoyer, transformer et enrichir une base complexe multi-sources,
- Explorer finement la structure des données d'assurance,
- Identifier les variables les plus influentes sur le risque ou le coût des sinistres,
- Construire deux modèles économétriques valides (OLS et Logit) à des fins explicatives.

En complément, l'analyse non supervisée a mis en évidence des segments clients à sinistralité différenciée. La prochaine étape sera l'application de modèles de machine learning pour mieux capter la non-linéarité et les interactions complexes.