**Supply Chain - Satisfaction des clients**

Rapport 1 : rapport d’exploration, de data visualisation et de pre-processing des données

Leonard Heyerdahl

Alexandre PRZYBYLSKI

Alexis Garatti

Huazhen Hou

Objectif principal du projet :

*Extraire de l’information de commentaires clients en ligne*

*Objectifs spécifiques :*

* ***Prédire la satisfaction d’un client*** *: problème de régression (prédire le nombre d'étoiles).*
* ***Identifier les entités importantes*** *d’un message : localisation, nom d’entreprise...*
* *Extraire les propos du commentaire (problème de livraison, article défectueux...) : approche non supervisé*
* *Extraire de la réponse du fournisseur les propos du commentaire dans but d’essayer de les prédire uniquement avec le commentaire.*

# Collecte des données

SurTrustPilot France, nous avons choisi la catégorie Banques pour le volume estimé des données et pour une distribution des modalités (étoiles) relativement plus équilibrée que dans d’autres catégories. En effet lors d’un premier test sur la catégorie immobilier nous nous sommes aperçus qu’il y avait une prédominance des 4 et 5 étoiles.

Trust Pilot ne met pas à disposition d’API, nous avons donc collecté les données par webscrapping. La stratégie de scraping pour Trustpilot a impliqué un processus en deux étapes pour collecter de manière exhaustive les données des entreprises et les avis dans les catégories sélectionnées.

La première phase, **le Scrap 1**, s'est concentrée sur l'extraction d'informations clés des entreprises à partir de la page de listes des banques. Les données ciblées pour le scraping comprenaient le nom des entreprises, la moyenne des étoiles attribuées par les utilisateurs, le nombre total d'avis et le lien direct vers la page de l'entreprise qui contient tous les avis. L'aboutissement de cette phase a été le stockage des données extraites dans un fichier CSV nommé "liste\_entreprises".

Ensuite, le **Scrap 2** a utilisé la liste compilée des entreprises pour se pencher sur les pages individuelles de chaque entreprise. Ici, le scraping était plus détaillé, ciblant les avis des utilisateurs et leurs détails. Le scraper a été conçu pour recueillir le nom d'utilisateur, leur nombre d'avis, la localisation, la date de publication de l'avis, le titre de l'avis, le contenu de l'avis, et la date de l'expérience évaluée. Chaque information a été méticuleusement extraite pour former un ensemble de données détaillé des avis utilisateurs correspondant aux entreprises listées dans la phase de scraping précédente. Les pages d’avis contenaient chacun 20 avis et nous avons adapté le scrapper pour qu’il détermine le nombre de pages au total et qu’il change la clé de page dans l’url d’extraction sur chaque itération du scrapper.

De manière préventive un délais de deux secondes a été appliqué entre chaque page pour éviter de spammer Trustpilot et que le scrapper soit bloqué par le site. Nous avons sauvegardé un csv par entreprise, puis nous avons compilé les différents CSV en une base de données unique.

**Résultat du scrapping :**

Au total, nous avons recueilli 170751 avis sur les banques, en Français.

A screenshot of a computer code

Description automatically generated

# Nettoyage des données

Nous avons enlevé

# Analyse exploratoire des données

Après avoir enlevé les na, converti les datest au format DateTime nous avons fait une exploration des données.

Les 5 étoiles sont prédominantes dans le jeu de données :

A graph with different colored squares

Description automatically generated

Plus en détail, avec 3010 observations, les deux étoiles sont la modalité la plus minoritaire dans le jeu de données :

A black text on a white background

Description automatically generated

Le jeu de données contient des avis sur 56 entreprises, sur l’ensemble, YounitedCredit est l’objet du plus grand nombre d’avis, devançant la deuxième banque, Cofidis, par un facteur de 3 (approximativement 60,000 vs 20,000 avis).

A graph with different colored bars

Description automatically generated

Interprétation :

A graph with different colored lines

Description automatically generated

Nous avons également voulu savoir la distribution du nombre d’avis toutes catégories confondues donnés par les utilisateurs. La majorité n’a publié qu’un seul avis (80,000), suivi de 2 (35,000 utilisateurs) et 3 avis, au-delà de 4 avis, le nombre d’utilisateurs passe sous les 1000 et décroit linéairement.

A graph with different colored bars

Description automatically generated

Nous avons également voulu connaitre la distribution des avis dans le temps. De manière assez surprenante, un pic très important arrive sur le dernier mois, celui de la collecte.

A ce stade, nous n’avons pas d’explication claire sur ce pic. Nous envisageons quelques hypothèses : sur le dernier mois des entreprises listées sous d’autres catégorie (assurance) ont été rattachées à la catégorie banque ;le nombre d’avis est lié au récent durcissement des conditions de crédit ; mois après mois, un nombre potentiellement important de nouveaux avis sont effacés car ils sont contraires à la charte du site (injures, accusations infondées et/ou bots).

A graph with a line

Description automatically generated

Nous avone également voulu explorer la répartition géographique des utilisateurs, prenant avantage du tage « location » extrait dans el scrapping, cependant l’écrasante majorité renseignent la France, cete variable sera sans doute « droppée » pour la suite des analyses et modélisations.

A screen shot of a graph

Description automatically generated

Longeur des avis

De manière très intérressante pour la modelisation, la longueur des avis et la note donnée par les utilisateurs semblent être en interaction :

A graph of different colored squares

Description automatically generated

# Preprocessing et premiers résultats modèle

**Renforcement du jeu de données**

Pour la modélisation nous avons prévu de renforcer le jeu de données en rajoutant certaines variables

Premièrement, nous avons ajouté la longueur de l’avis, étant donnée la relation forte entre celle-ci et le nombre d’étoiles observée dans la phase d’exploration.

Deuxièmement, nous avons utiliser le modèle de langage de Deep Learning (Transformers) Camembert pour établir le sentiment général de l’avis. Dans son analyse de sentiment, Camembert retourne un label (soit positif, soit négatif) et un score de confiance dans sa labélisation. Pour le renforcement de notre jeu de données, nous combinons les deux, en réétiquetant positif en 1 et négatif en -1, et en multipliant ce résultat par le score de confiance, de manière à avoir des scores entre -1 (très probablement négatif) et 1 (très probablement positif.

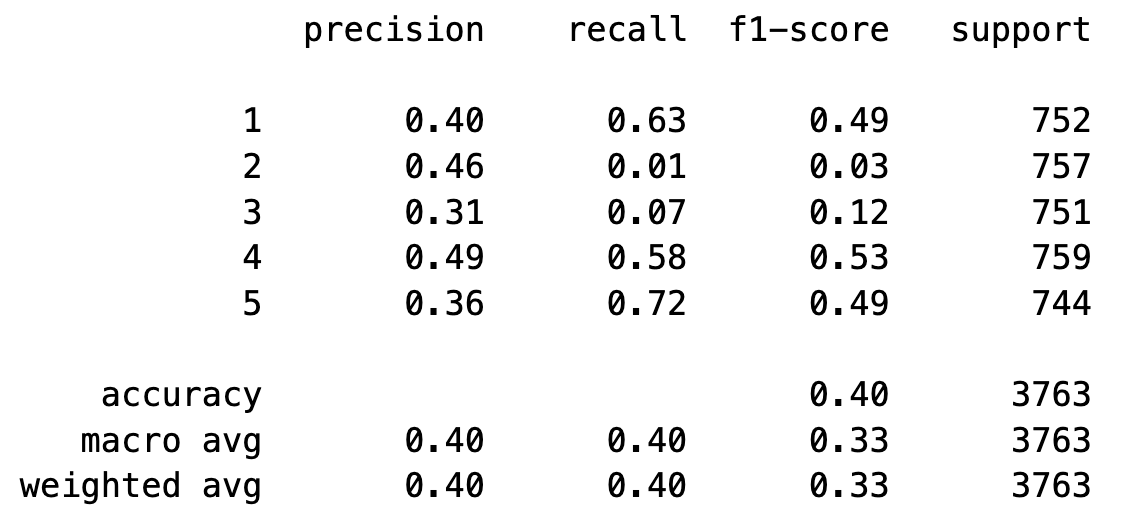
# Essais de modélisation

Nous avons prévu de faire un benchmark de différents modèles dont SVM, KNN, Random Forest et Camembert pour la classification des étoiles. Baseline : pour 5 modélités, le score d’étiquetage aléatoire serait en moyenne de 0,20.

**SVM simple**

La prédiction de SVM simple donne un résultat deux fois meilleur qu’un étiquetage aléatoire à 0,40, mais qui reste en deça d’une classification performante.

| **col\_0** | **1** | **2** | **3** | **4** | **5** |
| --- | --- | --- | --- | --- | --- |
| **etoiles** |  |  |  |  |  |
| 1 | 475 | 10 | 41 | 61 | 165 |
| 2 | 354 | 11 | 61 | 109 | 222 |
| 3 | 208 | 2 | 55 | 188 | 298 |
| 4 | 43 | 1 | 20 | 441 | 254 |
| 5 | 100 | 0 | 1 | 104 | 539 |



Par la suite nous tenterons de faire une grille pour essayer différents paramètres.

**Tests sur Camembert**

Le test de Camembert sur un échantillon réduit (15500 observations équilibrées entre les modalités – 3100 par étoiles) donne d’emblée de bons résultats, avec un score de 0,7.

Nous pourrons améliorer le score en rajoutant des poids et en faisant tourner le modèle sur l’ensemble du jeu, étant donné que les modèles Transformers sont relativement moins sensibles aux déséquilibres de modalités que les modèles plus classiques de machine Learning.

Plans pour la suite :