PROJET BIG DATA ANALYTICS

Analyse de la Clientèle d'un Concessionnaire Automobile pour la Recommandation de Modèles de Véhicules

2e Partie : Data visualisation

Groupe DEEP BRIDGE 1 : Chaptoukaev Hava, El Yakoubi Elidrissi Mohamed, Khchiou Mohamed, Malainine Cheikh El Wali, Trujillo Alexandre

I- PRESENTATION DU PROJET

CONTEXTE

Nous avons été contacté par un concessionnaire automobile afin de l'aider à mieux cibler les véhicules susceptibles d'intéresser ses clients. Le client sera satisfait si nous lui proposons un moyen afin :

- Qu'un vendeur puisse en quelques secondes évaluer le type de véhicule le plus susceptible d'intéresser des clients qui se présentent dans la concession
- Qu'il puisse envoyer une documentation précise sur le véhicule le plus adéquat pour des clients sélectionnés par son service marketing

Ce projet se découpera en trois parties distinctes :

- Mise en place d'une architecture Big Data.
- Data Visualisation.
- Data Mining et Machine Learning.

Nous allons développer la deuxième partie dans ce document.

DATA VISUALISATION

Pour ce projet plusieurs fichiers csv nous ont été fournis :

- Un fichier *Clients_2.csv* contenant les achats de l'année en cours.
- Un fichier *Clients_10.csv* contenant les achats de l'année en cours.
- Un fichier *Immatriculation.csv* contenant les informations sur les immatriculations effectuées cette année.
- Un fichier *Catalogue.csv* contenant le catalogue des véhicules du concessionnaire.
- Un fichier *Marketing.csv* contenant l'ensemble des clients sélectionnés par le marketing.

Le but de cette partie est d'analyser les données du domaine d'application et de les catégoriser, on va pour cela développer des techniques de visualisation avec l'API D3JS selon les contraintes suivantes :

- Utiliser une base de données multidimensionnelle.
- Prévoir de l'interaction.
- Prévoir deux niveaux de visualisation ; c'est à dire une vision globale et une vision plus précise.
- Donner la possibilité de charger des ensembles de données indépendants de l'application.

Nous allons faire un prototype exécutable avec D3JS qui démontre que l'utilisateur peut explorer l'ensemble des données pour atteindre des buts fixés. Avant nous allons définir dans ce document :

- La problématique, le but, et les utilisateurs visés.
- Les visualisations à créer pour répondre à notre problématique, leurs objectifs et les taches utilisateur.
- La chaine de traitement qui permet, à partir des données brutes, de créer une représentation graphique et interactive avec l'ensemble des données.

II- PROBLEMATIQUE ET APPROCHE ADOPTEE

PROBLEMATIQUE ET UTILISATEURS VISES

Le souhait du concessionnaire est d'attribuer une voiture à un client. Pour aider à répondre à cette problématique, une approche possible est de se concentrer sur l'historique des transactions passées. Autrement dit nous nous intéresserons aux liens entre un client et la voiture qu'il a achetée.
Cette application ne sera pas destinée au concessionnaire ou aux vendeurs, mais à nous, data Scientists, qui devrons trouver une solution pour aider le concessionnaire à répondre à ce besoin.

APPROCHE ADOPTEE

Pour cette problématique, nous allons visualiser l'ensemble véhicules de la base de données, c'est à dire explorer l'hétérogénéité des véhicules, leurs points communs, etc.
Mais nous allons également nous intéresser aux transactions, et plus précisément aux corrélations entre les attributs du client et les attributs des véhicules qu'ils ont achetés.

On peut supposer que l'élément principal qui détermine quel véhicule un client va acheter, est son prix. C'est pourquoi on propose de chercher plus particulièrement quels attributs d'un véhicule influent sur son prix, ou encore quels attributs d'un client déterminent à quel prix il va acheter. Mais bien évidemment, on ne va pas se contenter d'analyser uniquement cet aspect.

Avant de commencer la visualisation, il faut cependant d'abord s'assurer de la qualité des données, et si besoin les nettoyer.

III- PRE TRAITEMENT DES DONNEES

CHOIX DES DONNEES

Pour cette problématique particulière, nous allons oublier les fichiers *Catalogue.csv* et *Marketing.csv*, et prendre seulement les fichier *Immatriculations.csv* et Clients.csv :

- Pour cela nous allons d'abord réunir les fichiers *Clients_2.csv* et *Clients_10.csv* en un seul.
- Ensuite on joindra ce nouveau fichier *Client* au fichier *Immatriculations.csv* sur la colonne Immatriculation.

De telle sorte, on aura un unique fichier où chaque ligne représentera un client et ses attributs, ainsi que l'immatriculation de la voiture qu'il possède et les attributs de son véhicule.

Ce fichier est composé de *183 306 instances*, et *17 attributs* : immatriculation, âge, sexe, taux d'endettement, situation familiale, nombre d'enfants à charge, 2eme voiture, marque, nom de modèle, puissance, longueur, nombre de places, nombre de portes, couleur, occasion, et prix.

NETTOYAGE DES DONNEES

Une fois que le jeu de données final est construit, en explorant les données on réalise rapidement que les données comportent des anomalies et des problèmes de qualité :

- Les immatriculations ne sont pas uniques, et pour deux mêmes numéros il ne s'agit pas du même modèle de voiture. Il faut les supprimer.
- Il y a des variables pour lesquelles on rencontre des valeurs hors de leur domaine de définition, il faut gérer ces valeurs : pour les valeurs catégorielles, on peut remplacer par la valeur correcte : par exemple on peut rencontrer « False » alors que les valeurs doivent être comprises dans {F, T}. Pour les valeurs numériques hors de leur domaine, il faut probablement supprimer, ou remplacer par la valeur médiane des instances similaires.
- On rencontre également des NA: pour certains champs la seule solution est de supprimer la ligne. Pour les autres, remplacer par la valeur médiane des instances similaires.

TRANSFORMATIONS

On va catégoriser des variables numériques :

- Age
- Taux
- Prix

Le fichier *Clean_data.csv* joint à ce document correspond au jeu de données décrit. Cependant pour faciliter l'exécution de l'application, nous avons travaillé avec un sous échantillon aléatoire *Small_clean_data.csv.*

IV- MISE EN ŒUVRE

VISUALISATIONS

- Fréquence de vente des véhicules : On commence par réaliser une visualisation globale pour représenter les taux de ventes des différents modèles de voitures. On espère détecter les marques les plus vendues. Pour cela on choisit des *Histogrammes* prenant les différentes marques en dimension, et la fréquence d'apparition dans le jeu de données en mesure. Autrement dit la fréquence d'achat en mesure. On voudrait également avoir un histogramme des différents modèles de la marque, pour chaque marque.

- Prix selon les marques de véhicules : On va ensuite entrer dans un contexte plus précis, et analyser les différents liens entre les attributs d'un véhicule, plus précisément on voudrait voir s'il y a des différences notables de prix d'une marque à l'autre, ou encore s'il y a un lien entre la puissance d'un véhicule et son prix. On choisit de faire un *Parallel Coordinates* qui prendrait comme dimensions les marques des véhicules, leur prix, et leur puissance.

- Prix selon la puissance et l'état du véhicule : Toujours dans l'exploration des attributs qui influent sur les prix des véhicules, on veut observer s'il y a une corrélation entre la puissance d'un véhicule et son prix, mais on va également voir si celui ci varie selon s'il s'agit d'un véhicule neuf ou d'occasion. On va pour cela réaliser un *Scatter Plot*, et placer les véhicules en fonction du prix en abscisses et de la puissance en ordonnées. On va

également distinguer les voitures d'occasion des voitures neuves en les représentant de deux couleurs différentes.

- Prix des véhicules achetés selon les moyens des clients : Dans le même contexte, on va chercher une corrélation entre le prix d'une voiture et le taux d'endettement d'un client. On veut en effet observer si le revenu d'un client influe sur le prix du véhicule qu'il achète. Pour cela, on va réaliser une *2D Density* Map, sous forme de *Contour Map*, du taux d'endettement d'un client en fonction du prix de la voiture qu'il possède.

- Situation du client et son véhicule : Pour finir, on veut explorer les différentes tailles des véhicules en cherchant en même temps des liens avec la situation familiale des clients. On va représenter, dans un *Parallel Sets*, si un véhicule a été acquis en tant que 2e voiture, son nombre de portes, sa longueur, la situation familiale des clients, et le nombre d'enfants à charge des clients. Cette représentation permettra également d'avoir une visibilité sur la proportion des véhicules/clients qui entrent dans ces catégories, puisque la taille d'un « set » dépend de la fréquence d'apparition dans le jeu de données.

PIPELINE