Rapport d'Analyse des Données

Table des matières

Projet : BDA-LD

Date de dépôt : 31 déc. 23

Auteurs :

Yahya AARJI

Ayoub ADMESSIEV

Maxime BELLET

Benjamin BERNAUD

Paul ZENAGLIA

Introduction

Ce rapport documente le processus de gestion et d'analyse des données réalisé dans le cadre du projet BDA-LD. L'objectif principal était de créer un modèle de classification pour attribuer aux clients des catégories de véhicules en fonction de leurs caractéristiques.

Gestion des Données

1. Chargement des Librairies

Nous avons utilisé plusieurs librairies R telles que tidyverse, regclass, ggplot, dplyr, rpart.plot et C5.0 pour faciliter la manipulation et l'analyse des données.

2. Analyse Exploratoire des Données

2.1 Chargement des Données

Nous avons commencé par charger les données provenant de fichiers CSV, notamment le catalogue de véhicules, les immatriculations, les données clients et les données marketing.

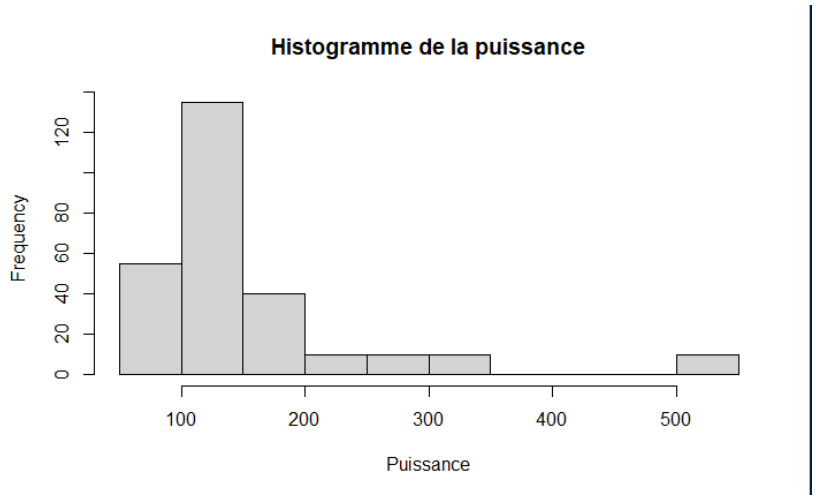
2.2 Statistiques descriptives

Des statistiques descriptives ont été effectuées pour comprendre la distribution des caractéristiques, par exemple, la puissance des véhicules dans le catalogue.

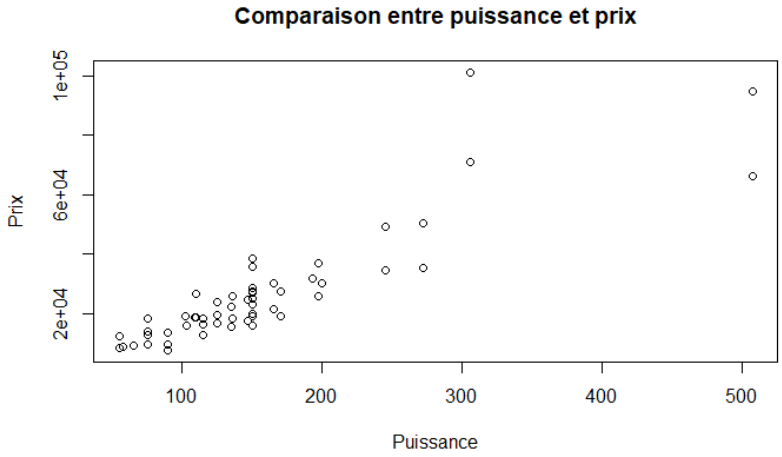
Ainsi on a pu afficher les différentes données, par exemple un résumé de la puissance des véhicules :

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   55.0   109.0   147.0   157.6   170.0   507.0
```

On a aussi fait des graphiques pour montrer de manière plus visuelle ces données, encore une fois la puissance des véhicules :



Ou ici, un graphique montrant la relation entre la puissance et le prix des véhicules :



On peut aussi mettre en évidence les liaisons entre les différentes variables. Par exemple, ici on peut voir que la puissance et le prix d'un véhicule sont très liés !

```
             puissance    nbPlaces  nbPortes       prix
puissance  1.00000000 -0.05708192 0.3109884  0.87545111
nbPlaces  -0.05708192  1.00000000 0.1129385 -0.08189026
nbPortes   0.31098839  0.11293849 1.0000000  0.27147998
prix       0.87545111 -0.08189026 0.2714800  1.00000000
```

2.3 Classification des Véhicules

Nous avons classé les véhicules en fonction de critères tels que la puissance, la taille et le prix, avec des seuils définis pour créer des catégories significatives.

Pour définir ces différents seuils, on s'est basés sur les échelles de valeurs pour les différents critères. Pour la puissance :

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
55.0   109.0   147.0   157.6   170.0   507.0
```

On a donc défini des seuils correspondants aux différentes variables présentes :

```
seuil_puissance <- c(0, 100, 170, Inf)
```

Même procédé avec le prix :

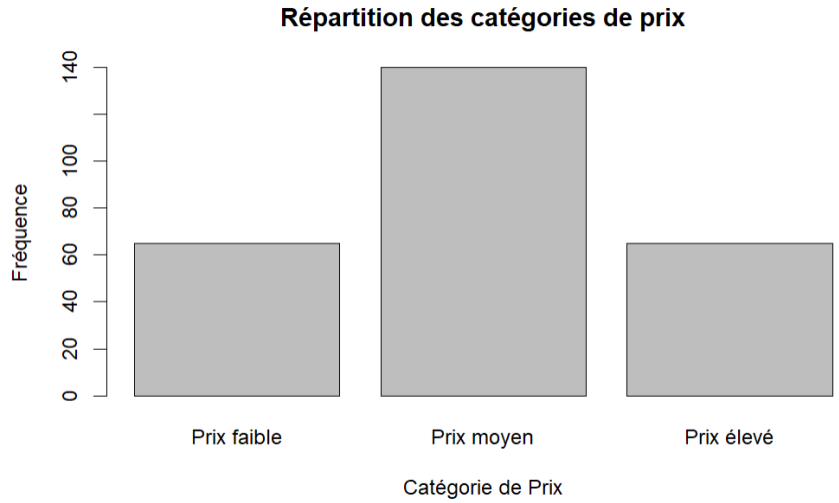
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
7500   16029   20598   26668   30000  101300
```

```
seuil_prix <- c(0, 16000, 30000, Inf)
```

Ensuite on a défini des classes pour séparer en différentes classes les prix et les puissances (les tailles des véhicules étant déjà qualitatives, il n'est pas nécessaire de les séparer)

couleur <fctr>	occasion <fctr>	prix <int>	puissance_class <fctr>	prix_class <fctr>
blanc	false	50500	Puissance élevée	Prix élevé
noir	false	50500	Puissance élevée	Prix élevé
rouge	false	50500	Puissance élevée	Prix élevé
gris	true	35350	Puissance élevée	Prix élevé
bleu	true	35350	Puissance élevée	Prix élevé
gris	false	50500	Puissance élevée	Prix élevé
bleu	false	50500	Puissance élevée	Prix élevé
rouge	true	35350	Puissance élevée	Prix élevé
blanc	true	35350	Puissance élevée	Prix élevé
noir	true	35350	Puissance élevée	Prix élevé

On peut maintenant visualiser les catégories de prix.



Ensuite on a défini des critères de classifications pour définir les catégories de véhicule :

```
catalogue <- catalogue %>%
  mutate(classe = case_when(
    puissance_class == "Puissance élevée" & longueur != "Courte" ~ "Routière",
    puissance_class == "Puissance élevée" & longueur == "Courte" & prix_class == "Prix élevé" ~ "Sportive",
    puissance_class == "Puissance moyenne" & longueur == "Courte" & prix_class == "Prix élevé" ~ "Sportive",
    puissance_class == "Puissance moyenne" & longueur == "Courte" & prix_class != "Prix élevé" ~ "Citadine",
    puissance_class == "Puissance moyenne" & longueur != "Courte" ~ "Routière",
    puissance_class == "Puissance faible" ~ "Citadine",
    TRUE ~ "?"
  ))
```

Modèle de Classification

3. Application des Catégories aux Données des Immatriculations

Nous avons appliqué les catégories définies dans le catalogue aux données d'immatriculations, en utilisant les critères de puissance, taille et prix.

4. Fusion des Données Clients et Immatriculations

Les données clients et immatriculations ont été fusionnées pour créer un ensemble de données complet associant les informations des clients aux véhicules achetés.

Nous avons tout d'abord réuni les 2 fichiers clients en une seule variable clients. Dans un second temps nous avons effectuer une jointure entre les immatriculations et les clients, en se basant sur le champ immatriculation.

prix <int>	puissance_class <fctr>	prix_class <fctr>	classe <chr>	age <fctr>	sexe <fctr>	taux <fctr>
50500	Puissance élevée	Prix élevé	Routière	24	F	497
13750	Puissance faible	Prix faible	Citadine	77	M	520
94800	Puissance élevée	Prix élevé	Routière	59	M	1114
13750	Puissance faible	Prix faible	Citadine	58	F	547
25900	Puissance moyenne	Prix moyen	Routière	71	F	1320
18200	Puissance moyenne	Prix moyen	Routière	24	F	1381
37100	Puissance élevée	Prix élevé	Routière	37	M	1252

Après cette étape nous avons donc une donnée qui comprend les immatriculations et les clients concernés.

5. Création du Modèle de Classification

Nous avons choisi d'utiliser un modèle d'arbre de décision pour la classification en fonction des caractéristiques des clients, telles que l'âge, le sexe, le taux, la situation familiale, le nombre d'enfants à charge et la présence d'une deuxième voiture.

Afin de pouvoir tester notre arbre de décision, nous avons séparé nos données pour avoir d'un côté, un dataset d'apprentissage, et de l'autre un dataset de test. C'est ainsi que nous avons crée notre arbre de décision rpart, qui a obtenu un taux de succès de 86%.

Résultats

6. Application du Modèle aux Données Marketing

Le modèle créé a été appliqué aux données marketing pour prédire les catégories de véhicules les plus appropriées pour chaque client.

Après avoir crée notre arbre, l'avoir testé sur notre dataset de test, il ne restait plus qu'à appliquer l'arbre sur les nouvelles données à prédire : le fichier marketing.csv.

Problèmes Rencontrés

Malheureusement, en raison de difficultés rencontrées dans d'autres aspects du projet, notamment des problèmes sur l'architecture globale, nous avons été limités par des contraintes temporelles. Cela a réduit le temps disponible pour explorer et tester diverses méthodes de prédiction. C'est en partie

pour cela que nous n'avons pu tester qu'une seule méthode de prédiction, à savoir les arbres de décision.

Conclusion

Ce rapport a mis en lumière le raisonnement derrière chaque étape du projet, du choix des librairies à l'application du modèle aux données marketing. Les décisions ont été prises en fonction des caractéristiques des données et des objectifs du projet.

Nous sommes conscients que notre cheminement n’a peut-être pas été optimal, et qu’avec plus de temps disponible, nous aurions pu pousser les recherches plus loin pour trouver un algorithme de classification plus efficace.