Examen final

Professeur Olaf Kouamo
PDS - Python Pour La Data Science
Année académique 2020-2021

November 2, 2020

Exercice 1. Téléchargez le dataset qui se situe dans *data/measurement.csv*. Ce dataset contient les informations sur les consommation de carburant. L'idée etant de comparer les prix deux types de carburant SP 98 et E10. Le but de cet exercice est de prédire la consommation de carburant en fonction du type de carburant la distance parcourue etc ....

1. faire l'inventaire du nombre de lignes et de colonnes du data set.
2. Calculer la moyenne et l'ecart type des variables quantitative en fonction du type de carburant .
3. Y'a t il une différence significative entre la consommation dans les deux types de carburant?
4.
5. Quelles sont les variables qui contiennent des valeurs manquantes et quelle est la proportion des valeurs manquantes pour chacune de ces variables?
6. Calculer ainsi la matrice X constituée des variables explicatives et la cible Y. Ensuite découper les observations en X_{train}, Y_{train}, X_{test}, Y_{test}, ou la première partie servira de base d'apprentissage et la seconde de base de test.
7. Mettre en place un modele d'apprentissage pour prédire la consommation.
8. tester plusieurs modèles et plusieurs combinaisons de paramètres afin de fournir le modèle avec la meilleure prédiction. On définira de facon claire et précise les méthodes d'évaluation de modèles mis en place.

Exercice 2. Il est primordial pour un constructeur automobile de fidéliser ses clients une fois qu'ils ont acheté un véhicule. Un client fidèle est un client qui effectue ses opérations d'après-vente (entretien, réparation du véhicule) au sein d'une concession du groupe. Un client est considéré comme *churner* s'il ne s'est pas rendu dans une concession groupe depuis plus de 18 mois. Afin d'inciter le client à effectuer ses opérations d'après-vente au sein du

réseau, la Direction Marketing peut effectuer un certain nombre d'actions. L'objectif de cette étude est donc d'utiliser les modèles prédictifs, afin de mieux identifier les clients susceptibles d'être *churneurs*. Pour ce faire, nous disposons de deux fichiers:

- Un fichier véhicule contenant les colonnes suivantes:
  - id_veh : id du véhicule
  - brand : marque du véhicule
  - model : modèle du véhicule
  - segment : segment du véhicule (ex : 208 -> segment B)
  - registration_date : date de mise en circulation.
- Un fichier visite contenant les colonnes suivantes
  - id_veh : id du véhicule
  - date : date de la visite
  - country : pays de la concession
  - brand : marque du véhicule
  - dealer_id : id de la concession
  - mileage_km : km du véhicule à la date de la visite
  - client_amount : montant payé par le client lors de la visite

NB : Un client_amount nul signifie que la facture a été prise en charge par le constructeur (véhicule sous garantie) ou par l'assurance (dans le cas d'une visite suite à un accident par exemple).

1. Analyse descriptive et visualisation des données.
   Après avoir chargé les deux datasets, le candidat devra répondre aux questions suivantes :
   (a) Combien de véhicules sont présents dans les deux datasets ? Combien de véhicules distincts ?
   (b) Combien de véhicules ont été mis en circulation chaque année ?
   (c) Combien de marques différentes de véhicules contiennent les deux datasets ?
   (d) Top 5 id_veh ayant effectué le plus de visites
   (e) Top 5 dealer_id ayant reçu le plus de visites
   (f) Quelle est la proportion de véhicules ayant effectué une visite lors des 18 derniers mois (entre le 01/03/2018 et le 01/09/2019) ?
   (g) N'hésitez pas à ajouter toute analyse pertinente pour mieux comprendre les données.

2. Construction d'un modèle de classification binaire.
   L'objectif est de construire un modèle de classification binaire afin de prédire un client *churner*, c'est-à-dire un client qui ne s'est pas rendu dans une concession entre le 01/03/2018 et le 01/09/2019.

   L'objectif n'est pas de construire le modèle optimal (qui n'existe pas), mais de construire un pipeline complet de modélisation : preprocessing des données, feature engineering, construction de la target et entraînement du modèle.

   (a) Data preprocessing
   - identifier les variables categorielles et les variables continues.
   - analyser les valeurs manquantes dans les données. remplacer les valeurs manquantes pour les features categorielles par le mode et les features continues par la médiane.

   (b) Feature engineering: Exemples de features possibles :
   - Age du véhicule
   - Nombre de visites réalisées par le client
   - Informations de la dernière visite

   (c) Construction de la target: Un *churner* est un client qui n'a pas réalisé de visite en concession entre le 01/03/2018 et le 01/09/2019.

   (d) Mettre en place un modèle de machine learning afin de predire la probabilité pour un client d'être *churneur*

3. Critique du modèle et prochaines étapes

   (a) Expliquer succinctement le feature engineering réalisé et le modèle choisi.

   (b) Quelle métrique avez-vous adopté pour évaluer les résultats du modèle ? Pourquoi ?

   (c) Que feriez-vous pour améliorer les résultats de votre modèle ?

   (d) Quelles données pourrait-on ajouter afin de mieux expliquer le *churn* d'un client ?

   (e) Quelles seraient les étapes nécessaires afin de déployer votre modèle en production ?