Sorbonne Data Analytics : Projet 3

Prédire le parti victorieux de chaque Etat US aux élections de 2020

Description du projet

L'objectif de ce projet est de prédire le parti gagnant des élections présidentielles de 2020 aux Etats-Unis à partir de données socio-démographiques. Il s'agit donc d'un projet de classification binaire.

Les étapes du projet

Pour ce projet nous vous recommandons de suivre les étapes suivantes :

1. Constituer les données

Pour ce projet, il vous faudra reconstituer les données. Vous disposez de plusieurs jeux de données qu'il faudra joindre pour obtenir un DataFrame exploitable.

Le fichier des résultats de 2020 vous servira à créer votre target : 1 pour le Parti Républicain, 0 pour les démocrates.

Vous disposez également des résultats de 2008 à 2016. Ce fichier ne vous servira que dans vos analyses pour faire des comparaison avec les élections de 2020.

2. L'analyse exploratoire

Pour ce projet vous devrez mettre en place une analyse exploratoire des données. Vous disposez d'un grand nombre de variables, vous devrez donc choisir un sous ensemble de features sur la base de vos connaissances du sujet et de votre intuition compte tenu de la problématique.

- Vous devrez commencer par **valider la qualité de vos données** : contrôler la présence ou non de doublons, de valeurs manquantes et de valeurs aberrantes.
- **Déterminer les agrégats et statistiques classiques** pour le sous ensemble de variables d'intérêts : moyenne, médiane, écart-type, …
- Tout au long de votre analyse exploratoire **vous produirez des graphiques** permettant de mieux comprendre les données sur lesquelles vous travaillez.
- Vous ferez une **analyse univariée** pour chaque variable qui vous semblera avoir un intérêt compte tenu de votre problématique : distributions, répartitions …

- Vous produirez ensuite une **analyse bivariée** de vos données : Analyse des corrélations, Box-Plot par type de véhicule, scatter plot, pairplot …

Remarques :

- Faites attention à la lisibilité de vos graphiques : netteté, titre, noms des axes, taille, légende

3. La modélisation

Ce projet est une tâche de classification binaire. Pour cette partie vous devrez :

- Vous questionner sur le **features engineering** : création de variables, transformations ou non à appliquer à vos données numériques ex : log transformation, normalisation, standardisation… Cette étape peut être cruciale en fonction du choix de l'algorithme que vous allez retenir (comme vous l'avez vu dans cours de Machine Learning avancé)
- **Encoder vos variables catégorielles** en variables numériques. Il vous faudra choisir entre les techniques que vous avez vu telles que le **label encoding** ou le **one-hot-encoding**
- **Procéder à une sélection de variables**
- **Séparer** vos données en train et en test afin de monitorer la capacité de vos modèles à se généraliser. Attention à la répartition des différents états dans les deux jeux de données.
- **Tester différents modèles de classification :** Vous devrez obligatoirement réaliser une régression logistique comme modèle baseline que vous devrez challenger par des modèles non linéaires.
- **Vous hyper paramétriserez vos modèles avec un GridsearchCV**
- Pour construire vos modèles vous utiliserez des **pipeplines**
- Vous proposerez **une analyse de l'importance des variables globale** (exemples : les coefficients de la régression logistique ou les features importance d'une random forest) **et local** (Shap)
- Dans tout ce projet vous attacherez une grande importance à la **qualité de votre code**

Conseil : Il est recommandé de commencer par des modèles simples et de complexifier votre approche (features engineering, choix des modèles) au fur et à mesure.

4. Evaluation

Pour évaluer la qualité de vos modèles **vous utiliserez le F1-score**. Vous pourrez également regarder des métriques telles que le Recall, la Précision ou encore l'Accuracy.

Vous devrez systématiquement afficher les résultats sur votre jeu de train et de test pour justifier la capacité de votre modèle à se généraliser

Les rendus

- Vous avez jusqu'au dimanche **31 octobre minuit** pour rendre le projet
- Vous devrez nous remettre **le notebook commenté contenant vos analyses exploratoires et vos modélisations** ainsi que la version **html (avec le code exécuté et sans erreur)** du notebook
- Votre projet devra être accessible depuis un dépôt Github, avec au moins 3 versions des notebooks accessibles. Un *readme* décrira votre projet et son fonctionnement et un fichier de *requirements* sera également disponible.
- Votre projet doit être au maximum reproductible
- La semaine suivant le rendu sera consacrée à la préparation de votre soutenance (création de vos slides)
- Vous pouvez réaliser ce projet par **groupe de 4 maximum**