Projet SD701: Exploration de grands volumes de données

Sujet : Recommandation des stations VELIB dans le premier arrondissement de Paris

1- Introduction

Ce projet a pour but, la mise en place d'un système de recommandation de velib dans le premier arrondissement de paris en se basant sur 2 critères principaux:

- La distance d'une habitation à une station velib
- La distance d'une habitation à une station velib et la probabilité d'avoir un vélo

Par la suite, nous essayerons de mettre en place un système de prédiction de la disponibilité des vélos en se basant sur des variables telles que:

- Le jour de la semaine
- L'heure de la journée
- La météo

2 - Récupération et nettoyage des données

2-1 Récupération des données

Dans cette phase, notre réflexion s'est portée sur le type de données à utiliser pour atteindre les objectifs fixés.

Ainsi, Pour la mise en place de ce système de recommandation, nous avons eu besoin de 2 types de données:

- Les données de disponibilité des vélos dans chaque station velib en temps réel, que nous avons pu récupérer via l'API « Open data paris » en effectuant la requête suivante https://opendata.paris.fr/api/records/1.0/search/?dataset=velib-disponibilite-en-temps-reel&facet=records&rows=1391"
- La base de données adresses des arrondissements de Paris qui est disponible sur le site data.gouv.fr

2-2 Nettoyage des données

Pour nos données cette phase a essentiellement consisté à:

- Renommer les colonnes afin d'avoir des noms plus courts et plus parlant
- Supprimer certaines colonnes pour ne garder que celles importantes pour notre étude
- Sélectionner les adresses du premier arrondissement parisien pour ce qui est de la base de donnée adresses

3- Recommandation des stations velib se basant sur la distance

Ce principe de recommandation consiste à recommander la station la plus proche d'une adresse en se basant uniquement sur le calcul de la distance euclidienne entre la station et l'adresse :

$$distance = \sqrt{((x_{station} - x_{velib})^2 + (y_{station} - y_{velib})^2)}$$

Avant de réaliser le calcul de notre distance, nous avons eu besoin de réaliser une jointure croisée entre nos données Velib et notre base de données adresse:

```
val crossJoin = adresseClean.crossJoin(velibClean)
```

Nos deux jeux de données possédant chacune des positions en coordonnées planes Lambert 93, nous avons pu calculer la distance euclidienne, entre chaque adresse et chaque station. Par la suite, nous avons récupérer la plus petite distance entre une adresse et une station grâce aux lignes suivantes:

```
val inter: DataFrame = crossJoinDist.groupBy("id").agg(min($"dist_km").as("dist_min"))
```

qui permet d'agréger nos résultats par adresse en utilisant la plus petite distance.

4- Recommandation des stations velib se basant sur la distance et la probabilité d'avoir un vélo

Pour cette recommandation, nous avons juste combiner la distance d'une adresse à une station et la probabilité d'avoir un vélo:

$$Esperance = \sqrt{((x_{station} - x_{velib})^2 + (y_{station} - y_{velib})^2)} * \left(\frac{nbreVelo_{Dispo}}{nbreBornes_{Station}}\right)$$

La station recommandée pour une adresse sera celle présentant la plus forte espérance.

5- Essai de mise en place d'un modèle de prédiction

Pour essayer de prédire le nombre de vélos disponibles dans une station, nous avons d'utiliser principalement 2 algorithmes de classification:

- La Régression logistique
- L'arbre de décision

Nous avons dans une première approche, décidé d'utiliser les variables ***«jour de la semaine et heure***» comme variables explicatives. En utilisant ces variables explicatives pour prédire le nombre de vélos disponible dans les stations, les précisions obtenues sur nos modèles sont très médiocres, respectivement 22% et 17% pour la régression logistique et l'arbre de décision.

Afin d'essayer d'améliorer notre modèle, nous avons introduit les variables météorologiques telles que la température.

L'ajout de cette variable n'a pas eu l'effet escompté car la précision de notre modèle est toujours aussi médiocre (21%).

6- Perspectives et conclusion

En définitive, nous avons pu mener à bien la partie recommandation des stations vélib en fonction de la distance et de l'espérance.

Concernant la partie prédictive, le constat est un peu plus mitigé car la très faible quantité de données et les variables explicatives utilisées pour prédire le nombre de vélos disponible en station ne nous permettent pas d'obtenir un modèle d'une grande précision.

Pour améliorer nos modèles, il faudrait d'une part essayer de travailler sur des données sur une plus grande période (plusieurs années) et d'autre part trouver d'autres variables explicatives permettant d'expliquer le nombre de vélos disponible dans une station (proximité, d'un monument touristique, de bureaux, fonctionnement des autres transports...). Si ces nouvelles variables explicatives n'améliorent toujours pas nos modèles, il faudrait peut être songer aux modèles non linéaires.