Lien GIT : https://github.com/Nessnessrine/Projet\_TaxiDriver

**TAXI DRIVER**

1. Note de Cadrage (Contexte/ Rappel du besoin/ Enjeux/ Objectifs)
2. Introduction

Le secteur des services de taxi à New York est depuis longtemps ancré dans le paysage urbain animé de la ville. Des milliers de taxis jaunes sillonnent les rues, offrant aux passagers un moyen pratique et fiable de se déplacer dans cette métropole en constante effervescence. Dans ce contexte dynamique, une nouvelle entreprise de taxi, nommée Taxi Driver, fait son entrée sur le marché avec une ambition claire : offrir une expérience de transport optimale à ses clients tout en maintenant une tarification compétitive. Pour y parvenir, Taxi Driver souhaite exploiter les données disponibles provenant du site de la Taxi and Limousine Commission (TLC) de New York pour prédire avec précision le prix des trajets.

1. Contexte

Dans le cadre de ses activités, la TLC recueille et conserve une quantité considérable de données sur les trajets effectués par les taxis jaunes, y compris des informations détaillées sur les itinéraires, les temps de trajet, les distances parcourues et bien d'autres variables pertinentes. Ces données sont accessibles au public et offrent une opportunité précieuse pour les entreprises du secteur des taxis d'exploiter ces informations afin de mieux comprendre les facteurs qui influent sur les tarifs des trajets.

Taxi Driver a reconnu la valeur de ces données en tant que ressource essentielle pour améliorer sa planification et son efficacité opérationnelle. En utilisant des techniques d'analyse de données avancées, Taxi Driver souhaite développer des modèles prédictifs fiables capables d'estimer le prix d'un trajet en fonction des variables disponibles dans les fichiers TLC. Une telle prédiction permettrait à Taxi Driver de mieux gérer ses tarifs, de proposer des offres compétitives à ses clients et de garantir une expérience de transport satisfaisante tout en optimisant ses revenus.

1. Objectifs

L’objectif est d’avoir une application qui nous donnerait le prix d’une course de taxi new-yorkais après avoir renseigné les informations nécessaires liés au trajets, tels que :

* Le point de départ
* Le point d’arrivée
* La date et l’heure

Nous allons exploiter les données provenant du site de la TLC pour développer 3 modèles prédictifs capables d'estimer le prix des trajets effectués par Taxi Driver. En utilisant une combinaison de statistiques descriptives, de préparation de données et de modélisation, nous chercherons à identifier les variables les plus significatives pour prédire le prix des trajets. Nous évaluerons ensuite plusieurs modèles de prédiction et comparerons leurs performances afin de déterminer celui qui offre les résultats les plus fiables et les plus précis.

En mettant en œuvre ce modèle prédictif, Taxi Driver espère améliorer sa capacité à fixer des tarifs compétitifs, à optimiser ses revenus et à offrir une expérience de transport satisfaisante à ses clients.

1. Présentation des données

La première étape consiste à prendre connaissance du contenu du fichier.

Voici un tableau décrivant les champs du fichier :

|  |  |
| --- | --- |
| Nom variable | Description |
| VendorID | Un code indiquant le fournisseur TPEP qui a fourni l'enregistrement |
| tpep\_pickup\_datetime | La date et l'heure auxquelles le compteur a été engagé |
| tpep\_dropoff\_datetime | La date et l'heure auxquelles le compteur a été désactivé |
| Passenger\_count | Le nombre de passagers dans le véhicule.  Il s'agit d'une valeur saisie par le conducteur |
| Trip\_distance | La distance parcourue en miles rapportée par le taximètre. |
| PULocationID | TLC Taxi Zone dans laquelle le taximètre a été engagé |
| DOLocationID | TLC Taxi Zone dans laquelle le taximètre a été désengagé |
| RateCodeID | Le code tarifaire définitif en vigueur à la fin du voyage. |
| Store\_and\_fwd\_flag | Cet indicateur indique si l'enregistrement de voyage a été détenu dans le véhicule  mémoire avant l'envoi au fournisseur, alias "stocker et transmettre",  car le véhicule n'avait pas de connexion au serveur. |
| Payment\_type | Un code numérique indiquant comment le passager a payé le voyage. |
| Fare\_amount | Le tarif temps-distance calculé au compteur. |
| Extra | Suppléments et suppléments divers. Actuellement, cela ne comprend que  les frais de 0,50 $ et 1 $ aux heures de pointe et de nuit. |
| MTA\_tax | Taxe MTA de 0,50 $ déclenchée automatiquement en fonction du compteur  taux d'utilisation. |
| Improvement\_surcharge | Supplément d'amélioration de 0,30 $ pour les voyages évalués à la chute du drapeau. Le  la surtaxe d'amélioration a commencé à être perçue en 2015 |
| Tip\_amount | Montant du pourboire – Ce champ est automatiquement rempli pour les cartes de crédit  Les pourboires en espèces ne sont pas inclus |
| Tolls\_amount | Montant total de tous les péages payés en voyage |
| Total\_amount | Le montant total facturé aux passagers. Ne comprend pas les pourboires en espèces. |
| Congestion\_Surcharge | Montant total collecté lors du voyage pour le supplément de congestion NYS. |
| Airport\_fee | 1,25 $ pour la prise en charge uniquement aux aéroports LaGuardia et John F. Kennedy |

1. Modèles de prédiction

Voici les trois modèles de Machine Learning que nous allons utiliser pour prédire le prix des trajets de Taxi Driver :

1. Réseaux de neurones

Les réseaux de neurones sont des modèles d'apprentissage profond capables de modéliser des relations complexes entre les variables. En utilisant un réseau de neurones, vous pouvez créer une architecture avec plusieurs couches cachées pour capturer des modèles non linéaires dans les données. On entrainera le modèle en utilisant des caractéristiques variées, telles que la distance, la durée, l'heure, le jour de la semaine, ainsi que d'autres données supplémentaires spécifiques aux trajets.

1. Régression linéaire

La régression linéaire est un modèle classique utilisé pour prédire une variable continue en fonction d'un ensemble de variables explicatives. Dans ce contexte, on entrainera un modèle de régression linéaire en utilisant des caractéristiques telles que la distance du trajet, la durée du trajet, l'heure de la journée, le jour de la semaine, etc. Ce modèle fournira une estimation linéaire du prix du trajet en fonction de ces variables.

1. Forêt aléatoire (Random Forest)

Les forêts aléatoires sont des modèles d'ensemble qui combinent plusieurs arbres de décision pour effectuer des prédictions. Ils sont flexibles, capables de gérer des données complexes et peuvent capturer des relations non linéaires. En utilisant une forêt aléatoire, vous pouvez inclure diverses variables explicatives, telles que la distance, la durée, l'heure, le jour de la semaine, ainsi que d'autres informations spécifiques aux trajets, pour prédire le prix du trajet.