**Rapport**

Objectif 1

* Etablir le plan du rapport décrivant la méthodologie

1. Intro

* Identifier les besoins et objectifs du projet

1. Partie 1

* Explorer le fichier (découvrir les variables, type de donnée, corrélation entre les variables)
* Traité les données (retirer valeurs manquantes, aberrantes, doublons, égale à 0, égale NULL, erreur de type)

1. Partie 2

* Séparer entre données d’entrainement et données de test
* Choix des modèles
* Coder les modèles (optimiser les hyper-paramètres)

1. Partie 3

* Comparer les modèles
* Choisir le modèle le plus performant

1. Conclusion

* Interprétation de la performance du modèle final

**Etape 2 : Dictionnaire de données**

VendorID : variable qualitative (float) indiquant quelle société de taxis a fourni la donnée. Vaut 1 pour "Creative Mobile Technologies LLC", 2 pour "VeriFone Inc."

tpep\_pickup\_datetime : variable qualitative (string) indiquant à quelle heure la course a commencé

tpep\_dropoff\_datetime : variable qualitative (string) indiquant à quelle heure la course s'est terminée

Passenger\_count : variable quantitative (float) indiquant combien de passagers sont inclus dans la course

Trip\_distance : variable quantitative (float) indiquant la longueur de la course en miles

PULocationID : variable qualitative (int) indiquant dans quelle zone géographique TLC le taxi a été demandé / la course a commencé

DOLocationID : variable qualitative (int) indiquant dans quelle zone géographique TLC le taxi a été laissé / la course s'est terminée

RateCodeID : variable qualitative (float) indiquant le code tarifaire final de la course. (1= Standard rate, 2=JFK, 3=Newark, 4=Nassau or Westchester, 5=Negotiated fare, 6=Group ride)

Store\_and\_fwd\_flag : variable qualitative (string) indiquant si la course avait été stockée en mémoire avant d'être transmis au chauffeur. Cela arrive en cas de perte de connexion au serveur (permettant la transmission de course en temps réel je suppose).

Payment\_type : variable qualitative (float) indiquant le moyen de paiement utilisé pour payer la course (1= Credit card, 2= Cash, 3= No charge, 4= Dispute, 5= Unknown, 6= Voided trip )

Fare\_amount : variable quantitative (float) Tarif calculé en fonction du temps et de la distance.

Extra : variable quantitative (float) Suppléments et suppléments divers. Actuellement, cela ne comprend que les frais d'heure de pointe et de nuit de 0,50 $ et 1 $.

MTA\_tax : variable quantitative (float) Taxe MTA de 0,50 $ déclenchée automatiquement en fonction du tarif mesuré utilisé.

Improvement\_surcharge : variable quantitative (float) $0.30 de taxe au déclenchement de la course. La surtaxe d'amélioration a commencé à être perçue en 2015.

Tip\_amount : variable qualitative (float) Ce champ est automatiquement rempli pour les cartes de crédit. Les pourboires en espèces ne sont pas inclus

Tolls\_amount : variable qualitative (float) Montant total de tous les péages payés en voyage.

Total\_amount : variable quantitative (float) Le montant total facturé aux passagers. Qui n'inclue pas le pourboire.

congestion\_surcharge : variable quantitative (float) Taxe pour tous les véhicule qui transporte des personnes.

**Etape 3 : Nettoyage des données.**

VendorsID : 58 891 lignes vide ce qui représente plus de 15% des données totals.

tpep\_pickup\_datetime : 4 valeurs au mois d’avril. Le dataset provient des données de mai initialement.

tpep\_dropoff\_datetime : 2 valeurs au mois d’avril. Le dataset provient des données de mai initialement. 85 valeurs au mois de Juin.

passenger\_count : 8 et 9 sont des valeurs aberrantes.

trip\_distance : 21 valeurs aberrantes.

RatecodeID : la valeurs 99 est une valeur aberrante, elle est représentée au nombre de 25. 58 891 lignes vide ce qui représente plus de 15% des données totals.

store\_and\_fwd\_flag : 58 891 lignes vide ce qui représente plus de 15% des données totals.

PULocationID :

DOLocationID :

payment\_type : 58 891 lignes vide ce qui représente plus de 15% des données totals.

fare\_amount : 1529 valeurs négatives.

extra : 576 valeurs négatives.

mta\_tax : 1498 valeurs négatives. 3.3 est une valeur aberrante elle est représenté 2 fois dans le fichier.

tip\_amount : 20 valeurs négatives.

tolls\_amount : 31 valeurs négatives.

improvement\_surcharge : 1518 valeurs négatives.

total\_amount : 1528 valeurs négatives. 170 valeurs égales à 0.

congestion\_surcharge : 1045 valeurs négatives. 1 valeur aberrante égale à 1.

**Objectif du projet :**

Optimiser puis comparer trois modèles permettant de prédire la variable TotalAmount.