Université de Lille

BIG DATA

PRÉDICTION DES TARIFS DE TAXI À NEW YORK

TUTEURS

Mme.Virginie Delsart
M.Maxime Morge

RÉALISÉ PAR

Racky Ka
Komi Adokpe
Reda Belhoti
Amine Abdel Moutaleb

Sommaire

Glossaire

Analyse bivariée	Méthode statistique permettant d'étudier la relation entre deux variables
Analyse univariée	Méthode statistique permettant d'étudier une seule variable
API	Interface logicielle qui permet de connecter un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités
Apprentissage automatique	Branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données historiques
Big Data	Données massives
Distance Matrix	Interface de programmation permettant d'accéder à des services de calcul de distance en ligne
Heure de pointe	Période de la journée où le trafic est à son maximum
mile(s)	Unité de mesure de la distance utilisée dans les pays anglophones
Reporting	Rapports d'activités permettant de suivre les performances d'une entreprise

Table des illustrations, graphiques et tableaux

Table des illustrations

Table des graphiques

Table des tableaux

Introduction

Dans le cadre de notre deuxième année de master Systèmes d'Information et Aide à la Décision (SIAD) à l'Université de Lille, nous sommes amenés à réaliser un projet *Big Data*. Sous forme de challenge, le projet permettra de mettre en pratique les enseignements que nous avons reçus depuis la première année de master. Nous avons été amenés à nous positionner sur une compétition de la plateforme `Kaggle` en équipe de 4. Notre choix s'est alors porté sur la prédiction du tarif des taxis à New York [1].

En effet, les taxis sont un moyen de transport très utilisé dans la ville de New York. La raison est toute simple. New York est une ville très animée avec une population dense. Le trafic est donc intense. Ce qui rend les déplacements compliqués en transports en commun ou avec tout autre moyen de transport. Les taxis restent l'option la plus rapide et la plus accessible pour la population. Mais alors, qu'en est-il de l'apport financier pour les compagnies ?

Dans ce rapport, nous traiterons de cet aspect. Après avoir présenté le projet, nous analyserons les différents facteurs qui peuvent faire varier le tarif d'une course en taxi. À l'issue de cette analyse, nous utiliserons des techniques d'apprentissage automatique pour entraîner des modèles de prédiction à partir de données historiques de trajets en taxi. Enfin, nous proposerons un outil de restitution qui portera sur l'analyse du chiffre d'affaires des compagnies. Cet outil leur permettra d'améliorer leurs stratégies.

I. Le projet

1.1. Contexte

Organisé par `Google Cloud` et `Coursera`, le projet consiste à prédire le coût total d'une course en taxi à New York. Pour réaliser cette prédiction, nous avons à notre disposition des données sur la date de début de la course, les coordonnées géographiques de départ et d'arrivée du taxi. Mais avant de nous intéresser à ces données, nous allons expliquer l'écosystème du transport en taxi à New York.

Les taxis sont un mode de transport prisé par beaucoup de personnes dans la ville de New York. La *New York City Taxi and Limousine Commission* (TLC) est l'organisme qui réglemente l'industrie des taxis à New York [2]. C'est elle qui fixe les tarifs et veille à la qualité du service des taxis. Nous avons deux principaux types de taxi :

- les taxis jaunes : ils sont les plus courants à New York et peuvent circuler n'importe où dans la ville ;
- les taxis verts : moins courants que les taxis jaunes, ils ont un droit de circulation plus restreint et ne peuvent circuler librement qu'en dehors de Manhattan [3].

Dès l'entrée dans un taxi à New York, un tarif forfaitaire de base (2,5 $US) est appliqué. Ce montant augmente de 0,5 $US pour 0,2 mile parcourue. Il peut monter à 1 $US la minute en cas de forts trafics. En cas d'attente, le chauffeur peut facturer 0,40 $US la minute. De plus, un supplément de 1 $US et de 0,5 $US s'appliquent respectivement aux courses qui ont lieu en heures de pointe new-yorkaise (16h-20h) et en heures de nuit (20h-06h). Les courses effectuées le week-end ont un supplément de 1 $US. Les péages restent à la charge du passager et il devra ajouter un pourboire oscillant entre 20 et 30% du prix demandé [4].

1.2. Hypothèses du modèle

Après avoir pris connaissance du contexte, nous avons émis des hypothèses qui peuvent expliquer le tarif des taxis à New York. Ainsi, nous avons :

- la distance parcourue par le taxi : plus la distance sera grande, plus le tarif sera élevé ;
- l'heure du trajet : les heures de pointe (16h-20h) et les heures de nuit (20h-6h) ont un tarif forfaitaire de base plus élevé que les heures normales ;
- la durée du trajet : il peut faire augmenter le montant de la course ;
- le jour où le trajet est effectué : le tarif peut augmenter si le trajet est effectué le week-end ;
- le lieu de départ ou d'arrivée du taxi : les trajets à destination ou en provenance de l'aéroport sont généralement plus chers ;
- les conditions météorologiques : en hiver où il neige par exemple, les taxis sont en nombre plus réduits et donc les tarifs sont plus élevés.

1.3. Gestion du projet

Nous avons divisé le projet en 4 phases :

- la phase Documentation : nous allons rechercher le maximum d'informations sur le sujet dans cette phase ;
- la phase Traitement de données : nous allons épurer notre jeu de données ;
- la phase Statistiques descriptives : nous allons faire des visualisations pour faire ressortir des informations des données ;

- la phase Modèles ML : nous allons créer, entraîner et évaluer nos modèles ;
- la phase Outil de restitution : nous mettrons en place un tableau de bord sur `Power BI`.

Ensuite, nous avons affecté les tâches en fonction des compétences de chacun. Vous retrouverez en annexes (0. Répartition du travail) les tâches effectuées par chaque membre du groupe.

Enfin, nous avons défini des points hebdomadaires lors desquels nous suivons l'avancée du projet. Ces points avaient lieu soit les vendredis soit les samedis.

II. Les données

2.1. Description des données

`Kaggle` a mis à notre disposition deux jeux de données :

- train.csv : il contient les données à partir desquelles nous allons créer et entraîner notre modèle. Sa taille est de 5,7 Go et il comporte huit colonnes et des millions de lignes ;
- test.csv : il contient les données que nous devons prédire. Avec une taille de 983 kB, il comporte sept colonnes et 9914 lignes .

Nous avons sept colonnes qui sont présentes dans les deux jeux de données et une qui n'est présente que dans train.csv :

Colonnes identiques

- key : identifie chaque ligne du jeu de données de façon unique ;
- pickup_datetime : la date de début de la course en taxi ;
- pickup_longitude : la longitude du lieu où la course en taxi a commencé ;
- pickup_latitude : la latitude du lieu où la course en taxi a commencé ;
- dropoff_longitude : la longitude du lieu où la course en taxi s'est terminée ;
- dropoff_latitude : la latitude du lieu où la course en taxi s'est terminée ;
- passenger_count : le nombre de passagers dans le taxi durant la course.

Colonne uniquement présente dans train.csv

- fare_amount : tarif de la course en taxi en dollar américain. Il s'agit de la variable que nous allons expliquer et prédire.

Vous trouverez plus de détails sur les jeux de données notamment le type des colonnes en annexes (1. Structure des données).

2.2. Traitement des données

Nos traitements hormis la création de nouvelles variables vont s'appliquer principalement au jeu de données train.csv car test.csv contient les données réelles (déjà épurées). Pour commencer, nous avons choisi de ne garder que 150 000 lignes dans train.csv. Ensuite, nous avons étudié les données manquantes et les données aberrantes.

2.2.1. Étude des données manquantes

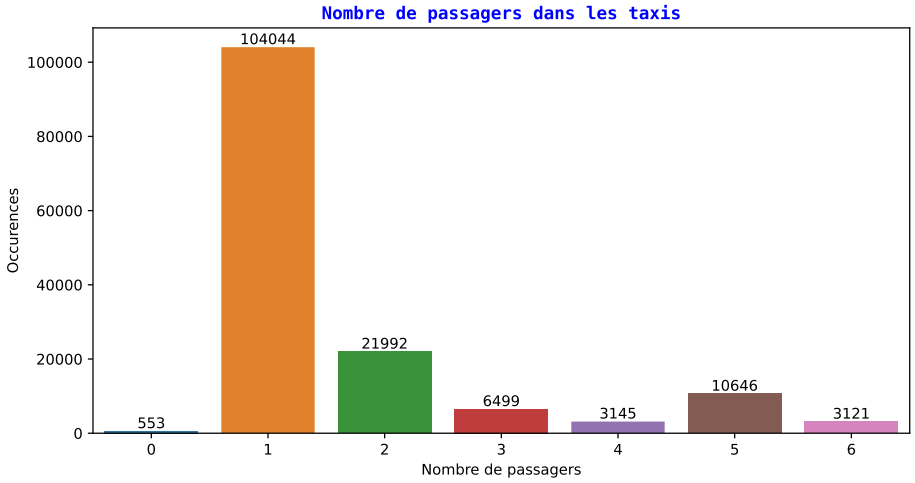
Concernant les données manquantes, nous avons trouvé une observation. Les coordonnées géographiques (longitude et latitude) du lieu où la course en taxi s'est terminée n'étaient pas renseignées. Il n'y a aucun moyen de deviner ces données. De plus, nous n'avons que deux données concernées sur 150 000. Nous les avons donc supprimées.

2.2.2. Étude des données aberrantes

Nous avons détecté un nombre important de données aberrantes. Nous allons donc les présenter variable par variable.

Nombre de passagers

Nous avons relevé 553 courses pour lesquels le nombre de passagers est nul comme vous pouvez le voir sur la figure ci-dessous.



Graphique 1 : Nombre de passagers aberrant
Source : résultats script sur Python

Ce qui n'est pas normal. D'autant plus que pour ces observations, nous avons un tarif appliqué à la course qui n'est pas nul. Ces données sont donc totalement incohérentes. Nous les avons supprimées.

Les coordonnées géographiques

Nous nous sommes intéressés aux coordonnées géographiques qui ne sont pas dans notre champ d'études (New York). Car rappelons-le, notre étude ne porte que sur les courses en taxi à New York. Nous avons détecté 3128 courses qui ne concernent pas la ville de New York. Nous les avons exclues du jeu de données.

Ensuite, nous avons constaté que 68 lignes du jeu de données avaient des coordonnées de latitude à la place des coordonnées de longitude et inversement. Nous avons donc interchangé ces données.

Enfin, nous avons eu 4327 observations où les coordonnées d'origine de la course en taxi sont les mêmes que les coordonnées de destination. Nous pensons qu'il peut s'agir d'une course où le taxi a fait un aller-retour. Malheureusement, nous n'avons pas la durée du trajet ni l'heure d'arrivée du taxi pour la calculer. Nous avons choisi de garder ces observations.

Le tarif de la course en taxi

En présentant le contexte, nous avons parlé de la tarification des courses en taxi à New York. En effet, le tarif forfaitaire de base en heures normales est fixé à 2,5 $US. Nous ne devons donc logiquement pas avoir des courses avec un tarif inférieur à ce montant. Cependant, nous en avons 14 dans le jeu de données. Nous les avons supprimées.

Aussi, nous aimerons savoir si nous avons des courses qui se sont déroulées en heures de pointe par exemple mais qui ont un tarif inférieur à 3,5 $US. Pour pouvoir le faire, nous avons besoin de l'heure où le trajet s'est déroulé. Or, nous n'avons pas encore cette variable. Nous allons donc la créer à partir de la variable pickup_datetime (date de début de la course). Nous allons en profiter pour créer d'autres variables également.

2.2.3. Création de variables

Les variables de temps

Nous avons créé à partir de la variable pickup_datetime plusieurs variables de temps :

- pickup_heure : l'heure où la course en taxi a débuté ;
- pickup_jour : le jour de la semaine de la course en taxi ;
- pickup_dateJour : le jour du mois de la course en taxi ;
- pickup_mois : le mois de la course en taxi ;
- pickup_annee : l'année de la course en taxi ;
- pickup_saison : la saison de la course en taxi (1 : hiver, 2 : printemps, 3 : été, 4 : automne).

Ces variables nous ont permis de créer deux autres variables assez intéressantes :

- pickup_weekEnd : pour indiquer si la course en taxi s'est effectuée un week-end ou non. Elle prend la valeur 1 s'il s'agit d'un week-end et 0 sinon ;
- heureDeNuitEtDePointe : pour indiquer si nous sommes en horaires de nuit, en horaires de pointe ou en horaires normaux. Elle prend la valeur 0 en heures normales, 1 en heures de nuit et 2 en heures de pointe.

Autres variables

Une autre variable très importante que nous avons créée est la distance. Elle indique le nombre de miles parcourus par le taxi entre le départ et l'arrivée. Pour la calculer, nous avons fait appel à l'API (*Application Programming Interface*) `Distance Matrix` de `Google Maps`. Il s'agit d'un service web qui permet de calculer le temps de trajet entre un point de départ et une destination. Son avantage est qu'il offre la possibilité de spécifier le mode de transport ici le taxi. Ainsi, nous avons une distance basée sur l'itinéraire emprunté.

Nous avons aussi créé la variable courseAeroport qui indique si une course est à destination ou en provenance d'un des aéroports de New York (LaGuardia, Newark, JFK). Elle prend la valeur 1 si le trajet répond à ce critère et 0 sinon.

Vous retrouverez la structure complète de notre jeu de données dans Annexes (1. Structure du jeu de données).

Nous allons continuer le traitement de notre jeu de données notamment la détection des valeurs aberrantes sur le tarif de la course en taxi.

Nous avons relevé des courses qui se sont déroulées le week-end mais qui affichent un tarif inférieur à 3 $US. Ce qui est incohérent par rapport à la tarification en vigueur à New York. De même, nous avons des courses où le montant appliqué est nettement inférieur au montant minimum que nous sommes censés avoir au vu de la distance parcourue par le taxi. Nous avons supprimé ces données.

À l'issue de tous ces traitements, nous nous retrouvons avec 122 847 observations sur les 150 000 que nous avons au départ.

III. Statistiques descriptives

À présent, nous allons réaliser une analyse descriptive de nos données. Cette étape est nécessaire pour décrire les caractéristiques de notre jeu de données et tirer des conclusions importantes. Nous allons vous présenter deux types d'analyses : une analyse univariée, ensuite une analyse bivariée et pour terminer, nous étudierons les liens entre les variables.

3.1. Analyse univariée

Nous allons commencer par quelques statistiques descriptives sur les variables distance et tarif de la course en taxi comme le montre le tableau ci-dessous :

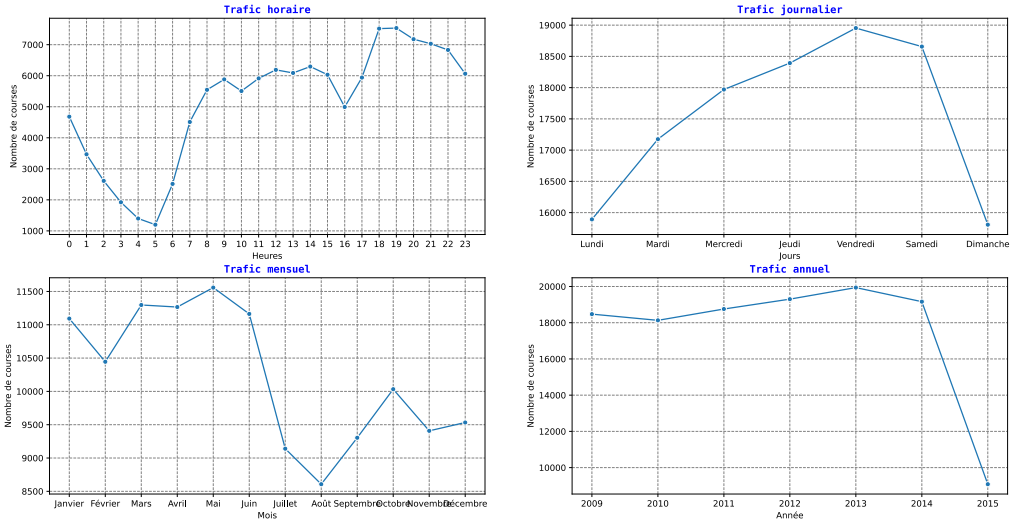
	Moyenne	Ecart type	Minimum	Médiane	Maximum
Tarif de la course (en $US)	11,82	10,2	2,5	8,5	500
Distance (en miles)	2,22	2,44	0,000189	1,5	67,7

Tableau 1 : Statistiques sur le tarif et la distance
Source : résultats script sur Python

Ce qui interpelle à première vue est la distance minimale parcourue par les taxis (0,000189 mile). Elle correspond aux courses de taxi pour lesquelles les coordonnées d'origine de la course sont égales aux coordonnées d'arrivée. Il est donc normal que nous ayons cette valeur. Nous avons une distance moyenne de 2,22 miles, une médiane de 1,5 mile et un écart type de 2,44. Ce qui signifie que la moitié des courses en taxi se font à moins de 1,5 mile. De même l'écart type montre une grande variabilité des données autour de la moyenne. Nous déduisons donc que les distances parcourues par les taxis à New York varient énormément avec une concentration importante sur les trajets courts. Les facteurs tels que la zone géographique et l'heure de la course peuvent l'expliquer.

Nous observons ensuite un tarif moyen de 11,82 $US, une médiane de 8,5 $ et un écart type de 10,2. Cette médiane indique que la moitié des tarifs est inférieure à 11,82 $US et l'autre moitié est supérieure. Les tarifs sont donc assez dispersés autour de la moyenne. L'écart type observé nous le confirme. Ce qui peut s'expliquer par les variations de distances, les lieux de course ou encore les horaires où la course a lieu.

Pour avoir des informations plus détaillées, nous allons nous intéresser à l'évolution temporelle du trafic des taxis à New York. Il s'agit du nombre de courses réalisées par les taxis en fonction des heures, des jours, des mois et des années. Le graphique ci-dessous nous permet de l'observer.



Graphique 2 : Évolution temporelle du tarif
Source : résultats script sur Python

Nous observons au niveau du trafic horaire qu'entre 8h et 17h, nous avons une activité d'environ 6000 courses. Le trafic atteint son pic entre 18h et 19h avec plus de 7000 courses réalisées par les taxis. Ce qui est assez normal car nous sommes en heures de pointe. Ce sont des horaires où les travailleurs ont tendance à rentrer chez eux après leur journée de travail. Le trafic tend à baisser à partir de 20h pour atteindre son plus bas niveau à 5h du matin. Nous sommes dans la plage des horaires de nuit. Il est commun d'avoir moins de trajets en ce moment de la journée.

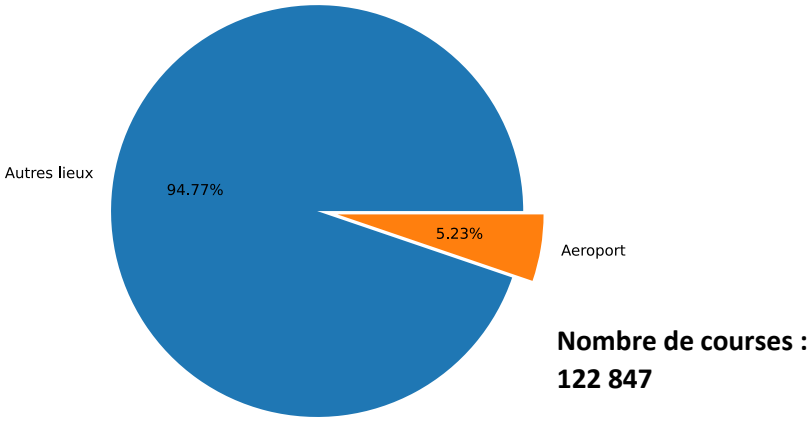
Au niveau du trafic journalier, nous constatons que le trafic démarre timidement en début de semaine pour atteindre son pic le vendredi avec près de 19 000 courses. Le vendredi marque le début du week-end et il arrive que des événements spéciaux soient programmés en ce jour. Il peut s'agir de concerts, de festivals ou encore d'activités sportives. La baisse du trafic constatée à partir du vendredi peut s'expliquer par le fait qu'il y a moins de travailleurs en déplacement le week-end. De plus, les transports en commun peuvent être privilégiés les week-ends.

Le trafic mensuel connaît énormément de fluctuations. Le mois de mai apparaît clairement comme celui où les taxis enregistrent le plus de demandes environ 11 500. Les mois de janvier et mars également ont une activité similaire au mois de mai. Nous voyons cependant que de juin à août, nous avons une baisse drastique du trafic qui atteint son plus bas niveau en août avec seulement 8600 courses. Le trafic tend à remonter ensuite de septembre à octobre mais sans jamais dépasser la barre de 10 000 courses. Ces fluctuations peuvent s'expliquer par les événements saisonniers, les périodes de vacances notamment entre juillet et août. Le faible trafic observé sur la période de décembre à février peut s'expliquer par l'hiver qui rend les conditions météorologiques moins favorables aux sorties. Le fort pic observé en mai peut être lié à la présence de touristes à cette période [5].

Enfin sur le trafic annuel, nous voyons qu'il y a une évolution plus ou moins homogène de 2009 à 2014. Nous avons une hausse du trafic de 2010 à 2013 qui est suivie par une légère baisse en 2014. Des facteurs économiques peuvent expliquer les variations annuelles mais il convient de noter aussi l'arrivée de concurrents. Nous avons notamment les services de covoiturage. La baisse observée sur l'année 2015 s'explique par le fait que nous n'avons que les données du premier semestre de cette année.

Maintenant que nous avons mieux compris le trafic à New York, analysons à présent la destination des courses. Nous allons plus précisément étudier le nombre de courses qui sont à destination ou

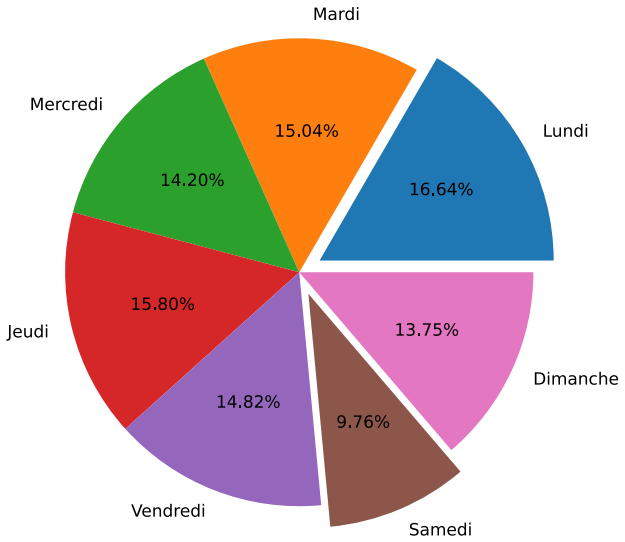
en provenance d'un des trois aéroports de New York. Le graphique suivant nous montre la répartition des courses entre les aéroports et les autres destinations :



Graphique 3 : Proportion des courses vers les aéroports locaux
Source : résultats script sur Python

Nous voyons que 5 % des courses sur 122 847 ont pour origine ou destination les aéroports locaux. Ce qui représente environ 6000 courses et reflète donc un marché important pour les taxis new-yorkais.

Après cette observation, nous trouvons judicieux d'analyser les tendances de voyage des New-Yorkais. Nous souhaitons connaître les jours où il y a plus de courses en taxi à destination ou en provenance de l'aéroport. Nous avons donc représenté la répartition des courses en direction ou en provenance des aéroports locaux en fonction des jours via le graphique ci-dessous :



Graphique 4 : Proportion des courses vers les aéroports locaux selon les jours
Source : résultats script sur Python

D'après le graphique, le lundi se révèle être le jour le plus prolifique pour les courses en taxi qui sont en provenance ou à destination des aéroports. Par contre, le samedi est le jour où moins de courses sont enregistrées. Les autres jours ont une tendance similaire au lundi. Nous déduisons

que les voyages d'affaires en semaine notamment le lundi et le jeudi expliquent en grande partie les courses en provenance ou en direction de l'aéroport.

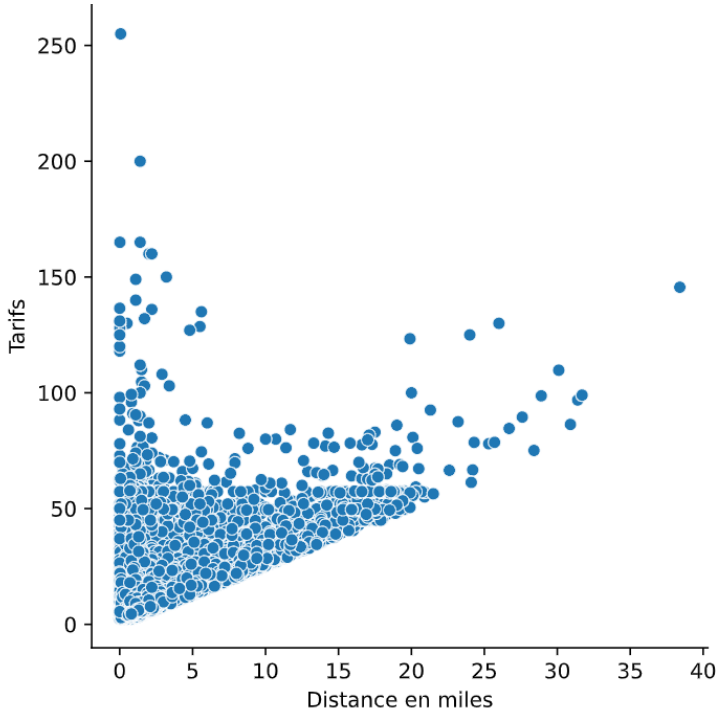
Nous allons désormais étudier la relation entre les variables. Nous commencerons par une analyse bivariée.

3.2. Analyse bivariée

Cette analyse nous permettra d'apercevoir l'effet qu'une variable a sur une autre, comme par exemple comment la distance affecte le tarif de la course. Nous allons pouvoir étudier également les différences entre groupes notamment les différences de chiffre d'affaires entre les saisons.

Tarif de la course en taxi et la distance

Le graphique ci-dessous montre la relation entre la distance parcourue par les taxis et le tarif de la course. Afin de mieux observer cette relation, nous avons exclu les courses aller-retour de l'analyse. C'est-à-dire les courses où les coordonnées d'origine du trajet sont égales à celles d'arrivée. Nous avons également exclu les courses vers les aéroports car elles ont des tarifs particuliers.

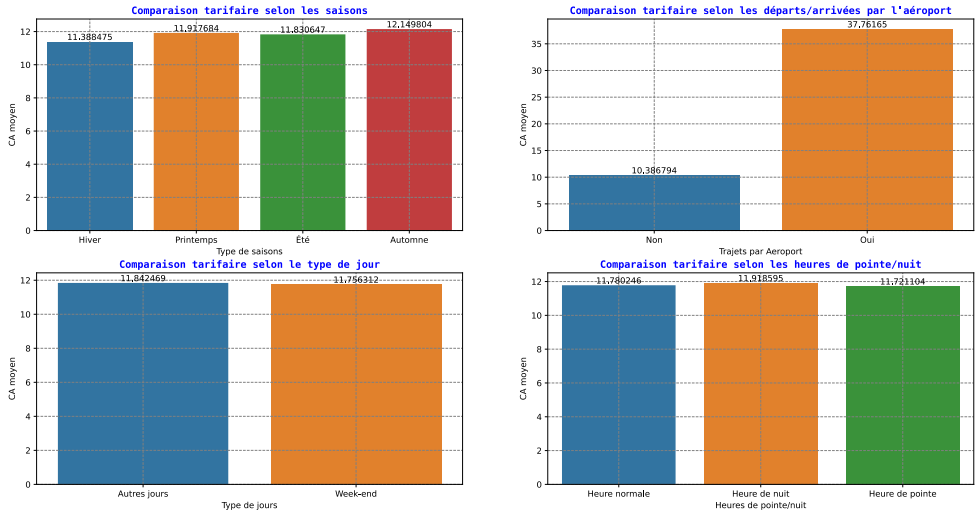


Graphique 5 : Évolution du tarif en fonction de la distance
Source : résultats script sur Python

Nous apercevons une évolution croissante du tarif en fonction de la distance. Plus la distance est longue, plus le tarif appliqué à la course est élevé. Néanmoins, nous avons quelques courses où la distance est assez courte mais le tarif est élevé. C'est assez logique de notre point de vue. En fonction de la zone géographique ou encore d'autres facteurs tels que les heures d'affluence, les temps d'attente, le tarif peut être revu à la hausse même si la distance parcourue est courte.

Comparaison tarifaire

Étudions maintenant les différences de chiffre d'affaires entre les saisons, les heures de nuit et les heures de pointe, les types de jours et les départs ou arrivées par l'aéroport. Le graphique ci-dessous nous le montre :

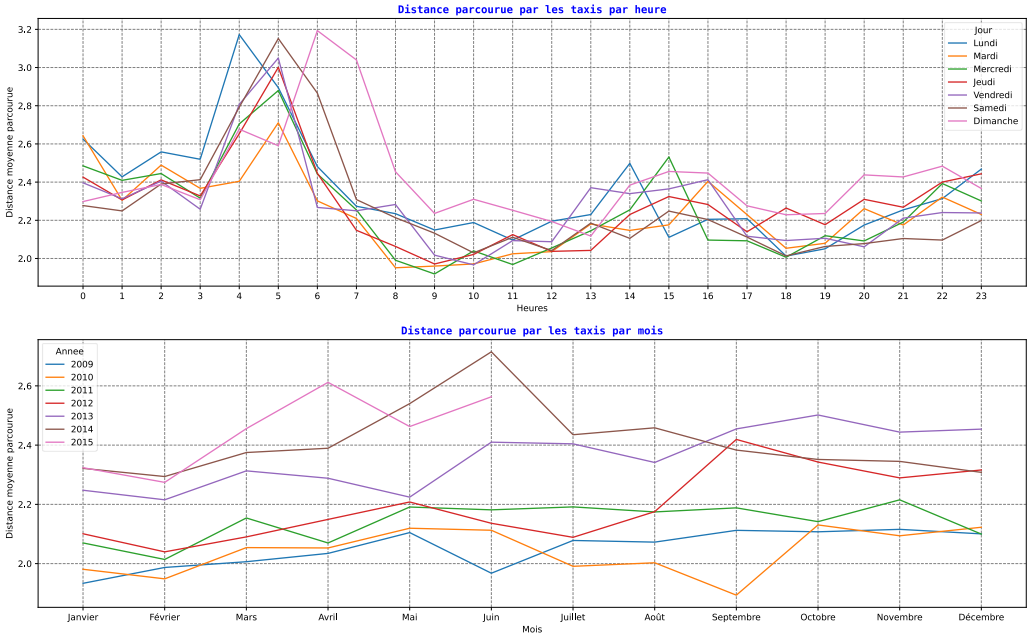


Graphique 6 : Comparaison tarifaire
Source : résultats script sur Python

Nous constatons que le coût moyen appliqué aux courses de taxi est relativement équilibré par saison avec une légère hausse en automne par rapport aux autres saisons de l'année. Nous observons aussi que les trajets en provenance ou en direction de l'aéroport ont un coût moyen nettement plus élevé (d'environ 25 $US) que ceux qui ne le sont pas. Ce qui confirme notre hypothèse de départ. Les coûts au niveau des week-ends et des autres jours de la semaine s'équilibrent assez bien en moyenne. Au niveau des heures, nous faisons la même remarque. Il y a un équilibre entre les coûts en moyenne. Mais rappelons que le trafic en heures de nuit est beaucoup plus faible que le trafic en heures normales ou en heures de pointe. Cela signifie donc que si les coûts s'équilibrent en moyenne, c'est parce que les trajets effectués la nuit sont beaucoup plus coûteux.

Distance parcourue par les taxis en fonction du temps

Nous allons observer l'évolution de la distance parcourue par les taxis selon les heures et les mois à travers les graphiques suivants :



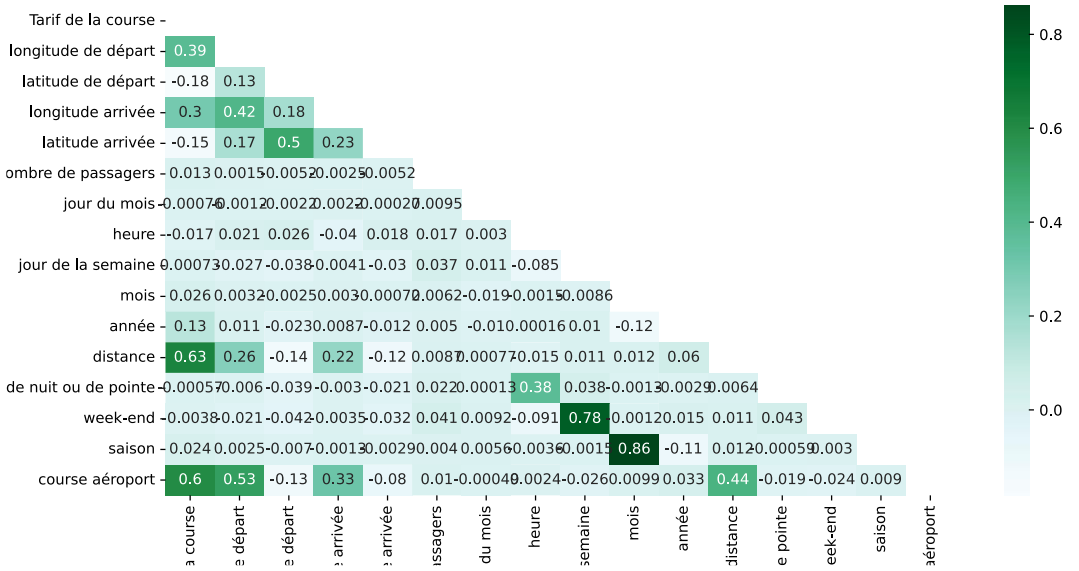
Graphique 7 : Distance parcourue en fonction du temps
Source : résultats script sur Python

À partir du premier graphique, nous pouvons voir l'évolution de la distance en fonction des heures et des jours. Nous pouvons déduire du graphique que les taxis parcourent généralement une plus grande distance entre 4h et 6h du matin quel que soit le jour. Le lundi, le samedi et le dimanche enregistrent des distances plus élevées par rapport aux autres jours dans cette tranche horaire. Nous remarquons aussi qu'entre 8h et 12h, la distance parcourue en moyenne par les taxis est faible.

Le deuxième graphique nous montre quant à lui l'évolution de la distance en fonction des mois et des années. Nous observons qu'en fonction des années, les pics ne sont pas atteints sur les mêmes mois. Ainsi 2014, atteint son pic en juin alors que l'année 2013 atteint son pic en octobre. Nous observons néanmoins que les taxis enregistrent généralement de faibles distances en février.

3.3. Analyse des corrélations

Nous allons réaliser la matrice de corrélation de Pearson entre nos variables dans le but de déterminer lesquelles nous garderons pour construire notre modèle. Cette matrice se base sur le test de corrélation de Pearson [6]. Plus la corrélation est forte, plus le coefficient est proche de 1 (corrélation positive) ou -1 (corrélation négative). Vous trouverez des détails sur ce test en annexes (2. Test de corrélation de Pearson).



Graphique 8 : Matrice de corrélation
Source : résultats script sur Python

À partir de cette matrice, nous pouvons voir l'intensité du lien entre les variables. Commençons donc par nous intéresser à notre variable d'intérêt, le tarif de la course en taxi.

Nous voyons que le tarif de la course est positivement corrélé avec la distance avec un coefficient de 0,63. Il est moyennement corrélé avec les longitudes d'origine et d'arrivée de la course en taxi. Nous constatons aussi une corrélation avec la variable Course aéroport qui rappelons-le, indique si la course a pour origine ou destination un aéroport local. Ces observations ne nous surprennent guère. Nous avons déjà émis comme hypothèse que le tarif pouvait dépendre du lieu de départ ou d'arrivée du taxi, de la distance parcourue par le taxi. Nous avons des corrélations beaucoup plus faibles pour les autres variables.

Intéressons-nous désormais aux liens entre les autres variables qui constituent ici les variables explicatives du modèle que nous allons construire.

Nous voyons de fortes corrélations entre certaines variables :

- heure et heure de nuit ou de pointe ;
- week-end et jour de la semaine ;
- saison et mois ;
- distance et course aéroport, distance et longitude départ, distance et longitude d'arrivée ;
- course aéroport et longitude de départ, course aéroport et longitude d'arrivée ;
- longitude de départ et longitude d'arrivée ;
- latitude de départ et latitude d'arrivée.

Ces corrélations ne sont pas surprenantes car elles concernent pour la plupart des variables qui sont dérivées d'autres. Par exemple, si nous prenons le cas de la corrélation entre les variables saison et mois, cela est dû au fait que la variable saison a été créée à partir de la variable mois. Il en est ainsi pour les variables week-end et jour de la semaine, distance et longitude de départ, course aéroport et longitude de départ…

Les corrélations que nous avons observées entre les variables longitude de départ et la longitude d'arrivée, latitude de départ et latitude d'arrivée ne sont pas étonnantes non plus. Ces variables

indiquent les lieux de départ et d'arrivée du taxi. Nous pouvons donc avoir les mêmes lieux qui se retrouvent aussi en départ qu'en arrivée. Un client peut avoir comme provenance l'aéroport JFK alors qu'un autre peut l'avoir comme destination.

À l'issue de l'étude des liens entre les variables, nous déduisons que certaines variables ne peuvent être intégrées simultanément dans notre modèle. Cela rendrait notre modèle inefficace en raison de la redondance de l'information. Les variables concernées sont :

- heure et heure de nuit ou de pointe
- jour de la semaine et week-end ;
- mois et saison ;
- course aéroport et les coordonnées de longitude.

Nous pouvons dès lors passer à l'étape de la modélisation.

IV. Modélisation

Il s'agit de la partie la plus importante de notre rapport. Toutes les étapes réalisées en amont avaient pour objectif de préparer la modélisation. Dans cette partie, nous allons construire différents modèles et les entraîner. Nous retiendrons ensuite le meilleur modèle qui nous permettra de faire nos prédictions.

Nous allons cependant commencer par le choix des variables à inclure dans notre modèle.

4.1. Choix des variables

Ce choix sera basé sur les corrélations et les conclusions que nous avons tirées précédemment. De ce fait, nous utiliserons les variables suivantes dans notre modèle :

- le tarif de la course en taxi : notre variable à expliquer ;
- les coordonnées géographiques : longitude de départ, latitude de départ, longitude d'arrivée, latitude d'arrivée ;
- la distance parcourue par le taxi ;
- heure de nuit ou de pointe (la variable prend la valeur 2 si la course a eu lieu en heure de pointe, 1 si c'est en heure de nuit et 0 sinon) ;
- week-end (la variable prend la valeur 1 si la course a eu lieu le week-end et 0 sinon) ;
- la saison où a eu lieu la course en taxi.

4.2. Échantillonnage des données

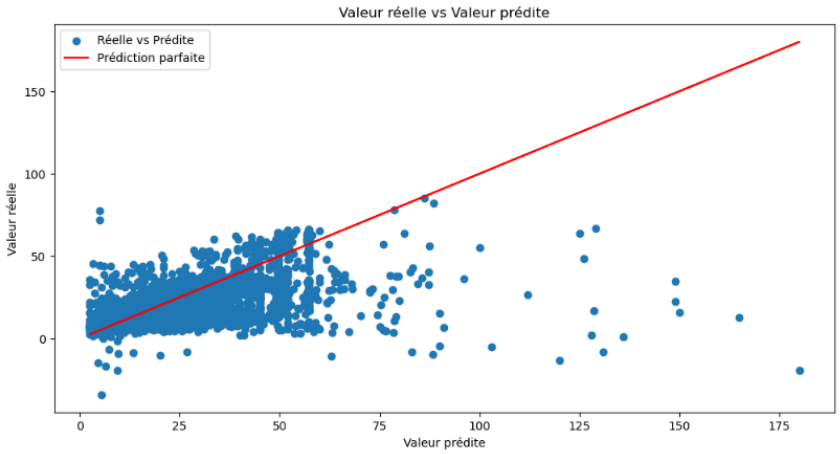
Nous allons à présent réaliser un échantillonnage de nos données [7]. Cela consiste à diviser notre jeu de données en deux sous-ensembles : un pour l'entraînement et l'autre pour la validation du modèle. Cette étape a l'avantage de nous permettre également de réduire la taille de nos données et donc d'accélérer le processus de développement de nos modèles. Nous avons choisi de prendre 80 % des données pour l'entraînement et 20 % pour la validation. Vous trouverez des détails en annexes (3. Échantillonnage des données).

4.3. Régression linéaire

La régression linéaire est une méthode statistique qui permet d'établir une relation linéaire entre une variable dépendante (ici le tarif de la course) et une ou plusieurs variables indépendantes (les autres variables que nous avons gardées dans notre modèle) [8]. Ce modèle permet de prédire la

valeur de la variable dépendante à partir des variables indépendantes via une équation linéaire. Les détails techniques de ce modèle sont à retrouver dans les annexes (4. Régression linéaire).

Suite à l'entraînement de ce modèle, nous avons construit le graphique suivant qui compare les valeurs réelles et les valeurs prédites par le modèle.



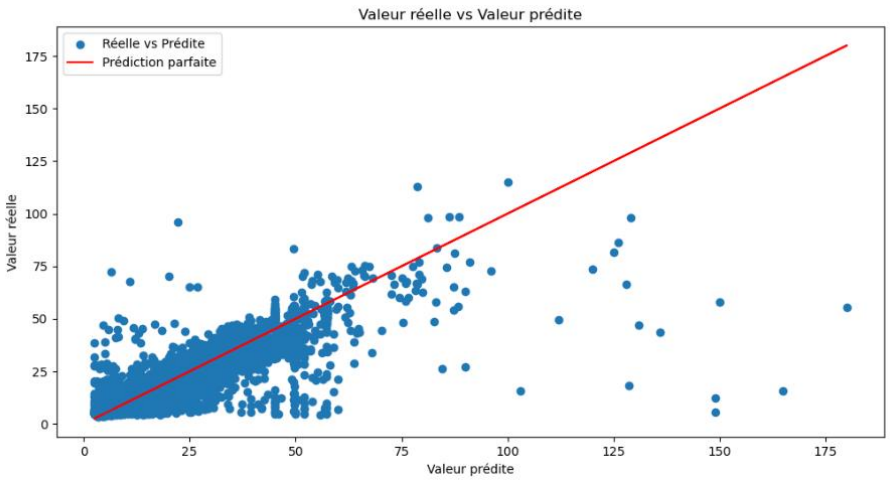
Graphique 9 : Valeurs réelles vs valeurs prédites Régression Linéaire
Source : résultats script sur Python

Nous pouvons voir que nous avons une grande dispersion des données. Nous avons beaucoup de valeurs qui sont surestimées ou sous estimées. De plus, les résultats du modèle ne sont pas satisfaisants (voir Annexes 4. Régression linéaire/Résultats du modèle). D'après ces résultats, le modèle ne prédit correctement que 50 % des observations. De plus, l'écart entre les valeurs réelles et les valeurs prédites est d'environ 7,32 $US en moyenne. Ce qui est assez grand.

4.4. Random Forest Regression

Le `Random Forest` (ou encore forêt aléatoire) est algorithme d'apprentissage supervisé qui construit des arbres de décision basés différents échantillons. Il prend ensuite la moyenne des prédictions fournies par ces arbres de décision et la retourne comme résultat final [9]. Vous trouverez plus de détails en annexes (5. Random Forest).

Les résultats de ce modèle (voir Annexes 5. Random Forest/Résultats du modèle) ont été beaucoup plus satisfaisants que le précédent comme l'atteste ce graphique.



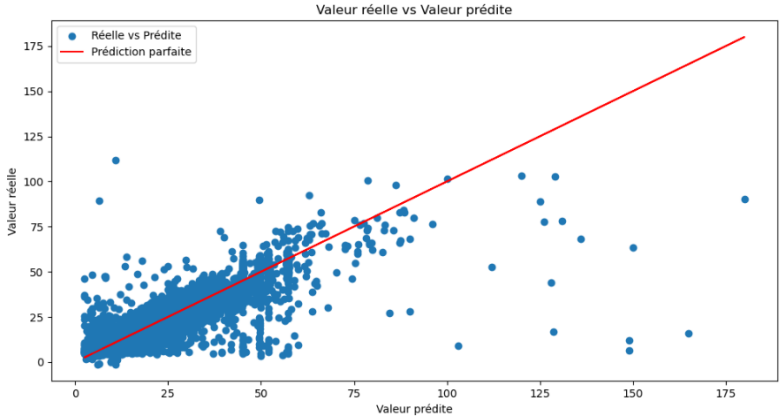
Graphique 10 : Valeurs réelles vs valeurs prédites Random Forest
Source : résultats script sur Python

Nous voyons que cette fois-ci, nous avons moins de dispersions. Il y a beaucoup moins de valeurs surestimées ou sous estimées. Et surtout la plupart des valeurs sont autour de la droite de régression. Le modèle a un pouvoir prédictif de 80 % et l'écart entre les valeurs prédites et les valeurs réelles est en moyenne d'environ 4,59 $US.

4.5. XGBoost Regression

`XGBoost` (*eXtreme Gradient Boosting*) est également un algorithme d'apprentissage supervisé comme le `Random Forest`. C'est une méthode qui repose sur les arbres de décision et qui s'améliore à partir d'autres méthodes comme le `Random Forest`. Elle est efficace sur des jeux de données volumineux et complexes. Elle permet d'avoir des prédictions précises [10]. Vous trouverez plus de détails en annexes (6. XGBoost)

Ce dernier modèle est légèrement plus performant que le `Random Forest`. La comparaison entre les valeurs réelles et les valeurs prédites nous permet de représenter un graphique quasi similaire à celui du `Random Forest`.



Graphique 11 : Valeurs réelles vs valeurs prédites XGBoost
Source : résultats script sur Python

Sur le graphique, nous ne voyons pas vraiment de différences. C'est assez normal car les performances de ce modèle ne sont pas si élevées par rapport à l'ancien. Les résultats obtenus

(voir Annexes 6. XGBoost/Résultats du modèle) montrent que le modèle explique le tarif des courses en taxi à New York de 81 %. L'écart entre les valeurs prédites et les valeurs réelles est en moyenne de 4,56 $US.

Les résultats obtenus pour l'ensemble de ces modèles vont nous permettre à présent de faire le choix de notre meilleur modèle. Celui qui nous permettra de faire les prédictions.

4.6. Choix du meilleur modèle

Dans le choix de notre meilleur modèle, nous allons regarder principalement le coefficient de détermination (R^2), l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE) [11]. Le coefficient de détermination sert à connaître le pourcentage de précision de notre modèle. L'erreur absolue moyenne permet d'exprimer l'erreur de la prédiction en unités de la variable cible. L'erreur quadratique moyenne est un indicateur pour mesurer la qualité de la prédiction. Un modèle qui est de meilleure qualité doit avoir son coefficient de détermination élevé, son erreur absolue moyenne et son erreur quadratique moyenne faibles. Nous allons également tenir compte du temps d'exécution de nos modèles dans notre choix. Vous trouverez plus de détails sur ces indicateurs en annexes (7. Qualité d'un modèle).

Nous avons donc regroupé dans le tableau ci-dessous les indicateurs de performance de chacun des modèles.

	R2	RMSE	MAE	Temps d'exécution
Régression linéaire	0,5	7,32 $US	4,01 $US	1,23 s
Random Forest	0,8	4,64 $US	2,20 $US	1 min 49 s
XGBoost	0,81	4,56 $US	2,19 $US	48,2 s

Tableau 2 : Comparaison entre les modèles
Source : résultats script sur Python

Comme vous pouvez le voir, il se dégage clairement que de tous les modèles réunis, le `XGBoost` est le plus performant. Nous avons un pouvoir prédictif de 0,81. Ce qui signifie que les prédictions du modèle sont précises à 81 %. Nous avons une erreur absolue moyenne de 2,19 $US. Donc en moyenne, nos prédictions sont fausses à hauteur de 2,19 $US. De plus, le temps d'exécution du `XGBoost` reste plus faible que celui du Random Forest qui est assez performant également.

À partir de ces observations, nous choisissons de garder le modèle `XGBoost`.

4.7. Classement sur Kaggle

Nous avons réalisé nos prédictions avec le modèle `XGBoost`. Nous avons extrait les données prédites dans un fichier que nous avons soumis à `Kaggle`. À la suite de cette soumission, nous avons obtenu un score de 4,69132 qui nous place à la 1128ième place au classement sur 1485 participants.

V. Outil de restitution

Afin de permettre aux compagnies propriétaires des taxis à New York d'améliorer leur stratégie, nous avons conçu un outil de restitution. Cet outil vise à analyser le chiffre d'affaires obtenu sur les courses de taxi et à faire des recommandations aux compagnies. L'outil a été conçu sur `Power BI` qui est un logiciel de *reporting* de `Microsoft`.

Pour le réaliser, nous avons d'abord conçu un schéma factuel à partir du jeu de données que vous retrouverez en annexes (8. Schéma factuel).

5.1. Présentation de l'outil

L'outil se présente comme suit :

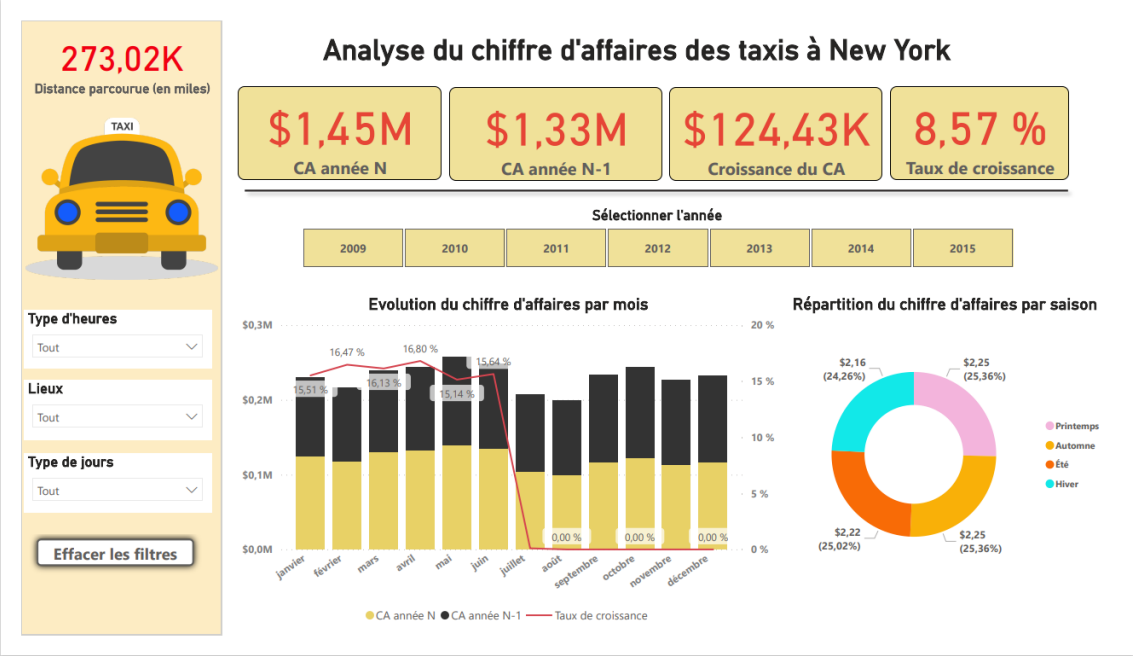


Illustration 1 : Outil de restitution
Source : capture sur Power BI

Dans le coin supérieur gauche, nous avons la distance totale parcourue par les taxis. Tout juste en dessous, nous avons différents filtres : un filtre permettant de choisir le type d'heures (heures de nuit, heures de pointe, heures normales), un filtre permettant de choisir le type de lieux (aéroport ou autres lieux) et un filtre permettant de choisir le type de jours (week-end ou autres jours). L'outil offre la possibilité de faire une sélection combinée sur ces différents filtres. Juste après les filtres, nous avons le bouton « Effacer les filtres » qui permet d'effacer tous les filtres qui ont été sélectionnés.

Dans la partie droite, nous avons tout en haut 4 étiquettes qui représentent respectivement :

- le chiffre d'affaires de l'année de l'année sélectionnée ;
- le chiffre d'affaires de l'année qui précède l'année sélectionnée ;
- la croissance enregistrée sur ces deux années sur le chiffre d'affaires ;
- et le taux de croissance du chiffre d'affaires.

Juste en dessous, nous avons un filtre sur les années. Ce filtre va permettre de sélectionner l'année à étudier. Il donne la possibilité de sélectionner plusieurs années.

Après ce filtre, nous avons nos deux visualisations. La première est un graphique à barre qui représente l'évolution du chiffre d'affaires par mois. Nous avons en jaune le chiffre d'affaires de l'année sélectionnée, en noir le chiffre d'affaires de l'année qui précède l'année sélectionnée et nous avons le taux de croissance représenté en ligne rouge.

Juste à côté, nous avons la seconde visualisation qui porte cette fois-ci sur la répartition du chiffre d'affaires par saison. Nous avons opté pour le chiffre d'affaires moyen dans ce cas comme indicateur car il nous permet de faire une comparaison plus juste.

Remarque : Dans le cas où aucune année n'est sélectionnée, l'outil fait le calcul du chiffre d'affaires sur toutes les années confondues.

5.2. Recommandations

L'outil de restitution nous a montré qu'en hiver, le chiffre d'affaires des taxis tend à baisser. Ce qui est logique car les gens ont tendance à rester chez eux. Face à cette situation, nous pensons qu'il peut être intéressant pour les compagnies de taxi de proposer des services de livraison. Ces services peuvent concerner n'importe quel type de produits.

Nous avons également constaté que le chiffre d'affaires sur les trajets en provenance ou en direction de l'aéroport croît tous les ans. Il serait intéressant de proposer des services de navette aux passagers qui se rendent ou qui reviennent de l'aéroport ou même de la gare. Ces services seront facturés à un prix fixe.

Nous avons constaté également que la distance parcourue par les taxis augmente d'année en année. Il serait intéressant de mettre en place un système de suivi de flotte pour optimiser l'utilisation des taxis.

Enfin, ils peuvent offrir des réductions aux clients qui effectuent des trajets réguliers.

Conclusion

Lors de ce projet, nous étions amenés à prédire le tarif des courses en taxi à New York. Pour répondre à cette problématique, nous avons mobilisé toutes les connaissances techniques et fonctionnelles acquises depuis la première année du master. Nous avons notamment mis en pratique les méthodologies de gestion de projet, les techniques de traitement et d'analyse de données. La partie la plus passionnante du projet réside dans le fait d'avoir travaillé sur une problématique réelle. Même si nous n'avons pas été en tête au classement, nous savons néanmoins que nous avons les compétences pour y arriver. Le modèle que nous avons développé peut toujours être amélioré en incluant la durée du trajet, les adresses précises des lieux de départ et d'arrivée du taxi. Par exemple le nom des villes ou quartiers. Nous avons aussi mis en place un outil de restitution qui devrait aider les compagnies de taxi à New York à améliorer leur chiffre d'affaires s'ils suivent nos recommandations.

Bibliographie

[1] Kaggle. *New York City Taxi Fare Prediction*. Dernière consultation 19/03/2023.

url : https://www.kaggle.com/c/new-york-city-taxi-fare-prediction.

[2] Wikipédia. *Taxis de New York*. Dernière consultation 19/03/2023.
url : https://fr.wikipedia.org/wiki/Taxis_de_New_York.

[3] Partir à New York. *Les taxis verts de New York.* Dernière consultation 19/03/2023.
url : https://www.partir-a-new-york.com/les-taxis-verts-de-new-york.

[4] New York Mon Amour. *Taxi New York : prix et infos pratiques, tout ce qu'il faut savoir*. Dernière consultation 19/02/2023.
url : https://newyorkmonamour.fr/taxi-new-york/.

[5] New York welcome. *Le printemps à New York : que faire et que visiter*. Dernière consultation 19/03/2023.
url : https://www.newyorkwelcome.net/fr/explorez/ce-quil-faut-savoir/.

[6] Nagwa. *Fiche explicative de la leçon : Coefficient de corrélation de Pearson.* Dernière consultation 19/03/2023.
url : https://www.nagwa.com/fr/explainers/143190760373/.

[7] Éditions eni. *L'échantillonnage*. Dernière consultation 19/03/2023.
url : https://www.editions-eni.fr.

[8] Wikipédia. *Régression linéaire*. Dernière consultation 19/03/2023.
url : https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire.

[9] DataScientest. *Random Forest : Forêt d'arbre de décision – Définition et fonctionnement*. Dernière consultation 19/03/2023.
url : https://datascientest.com/random-forest-definition.

[10] ArcGIS Pro. *Fonctionnement de l'algorithme XGBoost*. Dernière consultation 19/03/2023.
url : https://pro.arcgis.com/fr/pro-app/latest/.

[11] XLSTAT by Lumivero. *Indicateurs de performance de modèles.* Dernière consultation 19/03/2023.
url : https://www.xlstat.com/fr/solutions/fonctionnalites/indicateurs-de-performance.

[12] Jybaudot. *Le coefficient de détermination.* Dernière consultation 19/03/2023.
url : http://www.jybaudot.fr/Correl_regress/coeffdeterm.html.

[13] IBM. *Erreur absolue moyenne*. Dernière consultation 19/03/2023.
url : https://www.ibm.com/docs/fr/cloud-paks/cp-data.

[14] DataScientest. *Qu'est-ce que l'erreur quadratique moyenne ?*. Dernière consultation 19/03/2023.
url : https://datascientest.com/erreur-quadratique-moyenne.

Annexes

0. Répartition des tâches

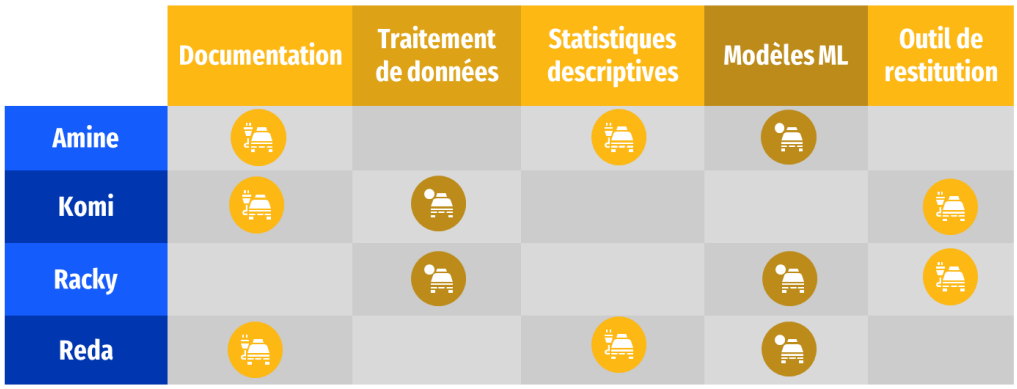


Illustration 2 : Répartition des tâches
Source : réalisée sur PowerPoint

		Tâches	Responsable	Etat
2	09/01/2023	Choix du sujet/Constitution des groupes	AA	Fait
3	16/01/2023	Prise en main de la problématique	Tous	Fait
4		Spécifications	KA&RK	Fait
5		Dictionnaire de données	KA&RK	Fait
6		Importation de la base de données sur Github	AA&RB	Fait
7	21/01/2023	Traitement de données	Tous	Fait
8		Créer les variables temps	AA	Fait
9		Création d'une fonction pour calculer la distance	KA	Fait
10		Exclure longitudes et latitudes aberrantes	RB	Fait
11		Exclure les données aberrantes de passenger_count	RK	Fait
12		Traiter la variable cible	KA&RK	Fait
13		Création du schéma factuel	KA	Fait
14	23/01/2023	Statistiques descriptives	AA&RB	Fait
15		Sur la variable cible	KA&RK	Fait
16		Sur les variables explicatives	AA&RB	Fait
17		Créer un support de présentation pour la soutenance blanche	Tous	Fait
18	30/01/2023	Soutenance blanche	Tous	Fait
19		Distance de Manhattan	KA	Fait
20		Hypothèses sur le prix des taxis	KA&RK	Fait
21		Détailler le contexte métier	RK	Fait
22		Compléter le fichier README	KA	Fait
23		Discussion sur les données aberrantes	KA	Fait
24	06/02/2023	Création d'un google docs rapport	Tous	Fait
25		Définition du ou des modèles à entraîner	AA&RB	Fait
26		Préparation des données (création de nouvelles variables)	KA&RK	Fait
27		Définition des modèles (description et étapes nécessaires à la modélisation)	AA&RB	Fait
28	11/02/2023	Corrélation avec les variables	AA&RB	Fait
29	13/02/2023	Entraînement des modèles	AA&RB	Fait
30		Optimisation du code de traitement de données	KA	Fait
31		Amélioration des modèles	KA&RK	Fait
32	20/02/2023	Amélioration des modèles	Tous	Fait
33	27/02/2023	Amélioration des modèles	Tous	Fait
34	06/03/2023	Rédaction du rapport	Tous	Fait
35	13/03/2023	Conception de l'outil de restitution	KA&RK	Fait
36		Amélioration du modèle	KA	Fait
37		Rédaction du rapport	Tous	Fait
38	16/03/2023	Finalisation de l'outil de restitution	KA	Fait
39	20/03/2023	Préparation du support de présentation de la soutenance	Tous	En cours
40	27/03/2023	Soutenance		

KA	Komi ADOKPE
RK	Racky KA
AA	Amine Abdel Moutaleb
RB	Reda Belhoti

Illustration 3 : Suivi du projet
Source : réalisée sur Google Sheet

1. Structure du jeu de données

		Libellé	Type
train.csv test.csv	key	Identifiant unique du jeu de données	String
	pickup_datetime	Date de début de la course en taxi	Date/Heure
	pickup_longitude	Longitude du lieu où la course en taxi a débuté	Float
	pickup_latitude	Latitude du lieu où la course en taxi a débuté	Float
	dropoff_longitude	Longitude du lieu où la course en taxi s'est terminée	Float
	dropoff_latitude	Latitude du lieu où la course en taxi s'est terminée	Float
	passenger_count	Nombre de passagers dans le taxi	Integer
	distance	Nombre de miles parcourus par le taxi	Float
	pickup_heure	Heure où la course en taxi a débuté	Integer
	pickup_jour	Jour de la semaine de la course en taxi	Integer
	pickup_dateJour	Jour du mois de la course en taxi	Integer
	pickup_mois	Mois de la course en taxi	Integer
	pickup_annee	Année de la course en taxi	Integer
	pickup_saison	Saison de la course en taxi	Integer
	pickup_weekEnd	Indique si la course en taxi s'est déroulée un week-end (1) ou non (0)	Integer
	heureDeNuitEtDePointe	Indique si la course s'est déroulée en heures de nuit (1), heures de pointe (2) et 0 sinon	Integer
	courseAeroport	Indique si la course s'effectue vers ou à destination d'un aéroport (1) et 0 sinon	Integer
train.csv	fare_amount	Tarif de la course en taxi en $US	Float

Tableau 3 : Structure du jeu de données
Source : jeu de données

2. Test de corrélation de Pearson

Le test de corrélation de Pearson permet de calculer la dépendance entre deux variables quantitatives. Les deux échantillons sont supposés suivre une loi de distribution normale. Le coefficient prend une valeur comprise entre [-1, 1]. -1 indique une corrélation négative parfaite, 0 indique une absence de corrélation et 1 une corrélation positive parfaite [6].
Le test évalue l'hypothèse nulle selon laquelle il n'y a pas de corrélation entre les deux variables. La statistique de test est calculée selon la formule suivante :

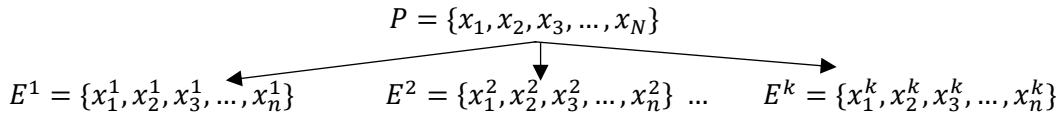
$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

où r est le coefficient de Pearson, $Cov(X,Y)$ est la covariance de X et Y et $\sigma_X \sigma_Y$ est le produit de l'écart type de X et de l'écart type de Y.

Si la valeur calculée de la statistique de test est grande, cela signifie qu'il y a une forte corrélation entre les deux variables. Ce qui conduit au rejet de l'hypothèse nulle.

3. Échantillonnage des données

L'échantillonnage consiste à réaliser des tirages d'individus à partir d'une population [7]. Soit la population $P = \{x_1, x_2, x_3, \dots, x_N\}$ avec x_i, $i \in \{1, 2, 3, \dots N\}$ les individus de cette population et N le nombre d'individus de l'ensemble de la population. Sur cette population P, des échantillons peuvent être sélectionnés comme suit :



Avec E^q l'échantillon numéro q réalisé sur la population P, x_n^q le $n^{ième}$ individu de l'échantillon E^q et $\forall j, \forall i : (x_i^j \in P)$.

4. Régression linéaire

La régression linéaire est une technique d'analyse de données qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives. Elle s'utilise pour des variables quantitatives [8].
Le modèle se présente généralement sous la forme suivante :

$$y = \alpha_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + \varepsilon_i \, , i \in \{1, 2, \ldots, N\}$$

où y est la variable expliquée (dépendante), $x_1, x_2, \ldots, x_n$ sont les variables explicatives (indépendantes), $b_1, b_2, \ldots, b_n$ sont les coefficients de régression qui représentent l'effet de chaque variable indépendante sur la variable dépendante, α_0 est l'ordonnée à l'origine et ε_i est l'erreur résiduelle de l'individu i.

Résultats du modèle

```
Erreur absolue moyenne (MAE) de Régression linéaire  :  4.007270714022742
Écart quadratique moyen (RMSE) de Régression linéaire  :  53.60213533750311
Erreur quadratique moyenne (RMSE) de Régression linéaire  :  7.321347918075134
Erreur absolue moyenne en pourcentage (MAPE) de Régression linéaire  :  0.3893022895375035
Variation expliquée (EVS) de Régression linéaire  :  0.49840964781908503
R2 de Régression linéaire  :  0.5
```

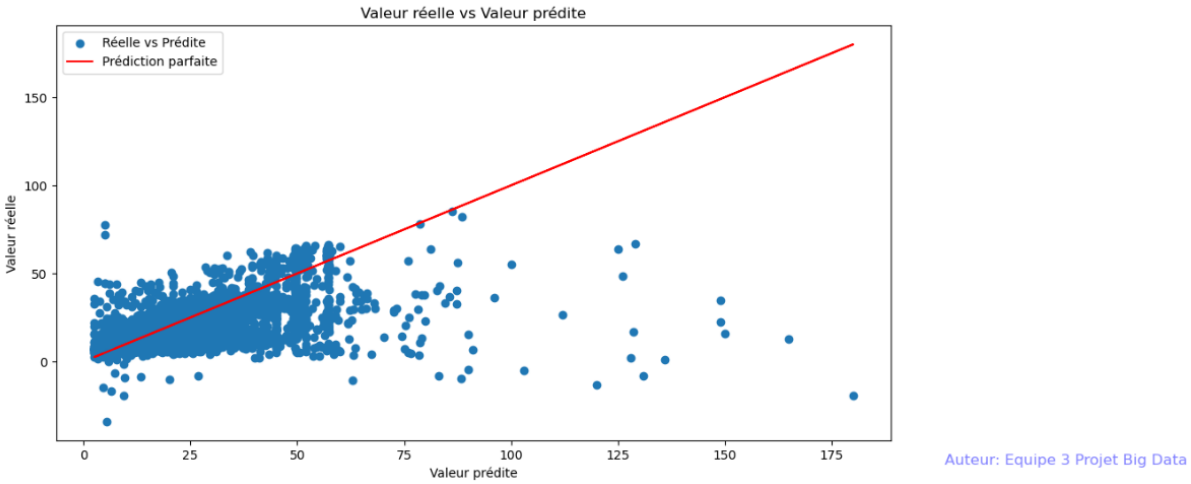


Illustration 4 : Résultat Régression Linéaire
Source : résultats script sur Python

5. Random Forest

Le `Random Forest` est un algorithme d'apprentissage automatique supervisé qui peut être utilisé pour la classification ou la régression. Il est basé sur un ensemble d'arbres de décision aléatoires. Chaque arbre est construit en utilisant un sous-ensemble aléatoire des données d'entraînement et un sous-ensemble aléatoire des variables d'entrée. Le résultat final est obtenu en agrégeant les prédictions de chaque arbre, en utilisant soit la moyenne (dans le cas de la régression), soit la majorité (dans le cas de la classification) [9].

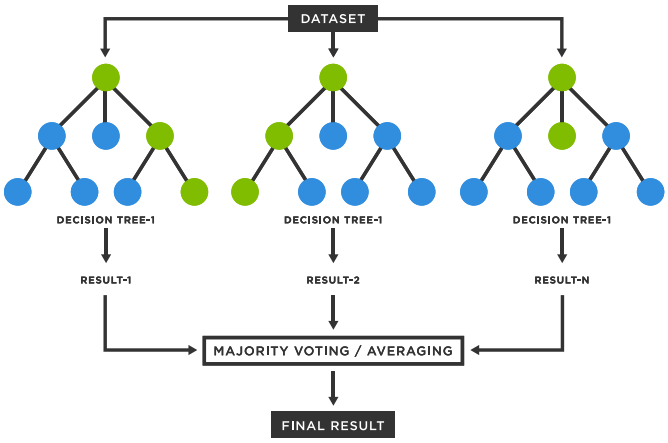
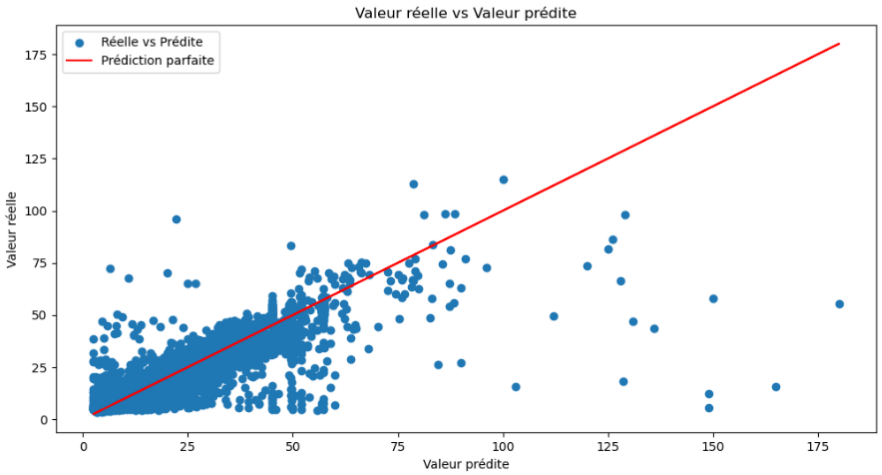


Illustration 5 : Mécanisme du Random Forest
Source : Google image

Résultats du modèle

```
Erreur absolue moyenne (MAE) de Régression Random Forest  :  2.2014009383211954
Écart quadratique moyen (RMSE) de Régression Random Forest  :  21.037608366693554
Erreur quadratique moyenne (RMSE) de Régression Random Forest  :  4.586677268643779
Erreur absolue moyenne en pourcentage (MAPE) de Régression Random Forest  :  0.20531724749159
Variation expliquée (EVS) de Régression Random Forest  :  0.8031451731300021
R2 de Régression Random Forest  :  0.8
```



Auteur: Equipe 3 Projet Big Data

Illustration 6 : Résultat Random Forest
Source : résultats script sur Python

6. XGBoost

`XGBoost` est un algorithme d'apprentissage automatique supervisé utilisé pour la classification et la régression. Il est basé sur la méthode de *gradient boosting*, qui consiste à combiner plusieurs modèles de prédiction plus simples pour créer un modèle prédictif plus puissant. Il utilise les arbres de décision comme modèles de prédiction. Il les ajoute de manière itérative à un modèle existant, en cherchant à minimiser l'erreur de prédiction du modèle sur les données d'entraînement [10].

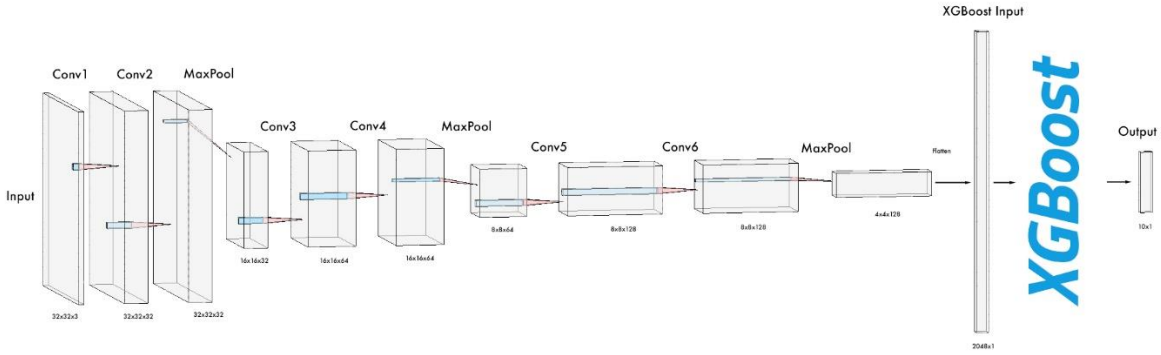


Illustration 7 : Mécanisme XGBoost
Source : Google image

Résultats du modèle

```
Rapport d'évaluation du modèle :
Erreur absolue moyenne (MAE) de Régression XGBoost  :  2.1895769
Écart quadratique moyen (RMSE) de Régression XGBoost  :  20.793148
Erreur quadratique moyenne (RMSE) de Régression XGBoost  :  4.5599504
Erreur absolue moyenne en pourcentage (MAPE) de Régression XGBoost  :  0.20138909
Variation expliquée (EVS) de Régression XGBoost  :  0.8054208159446716
R2 de Régression XGBoost  :  0.81
```

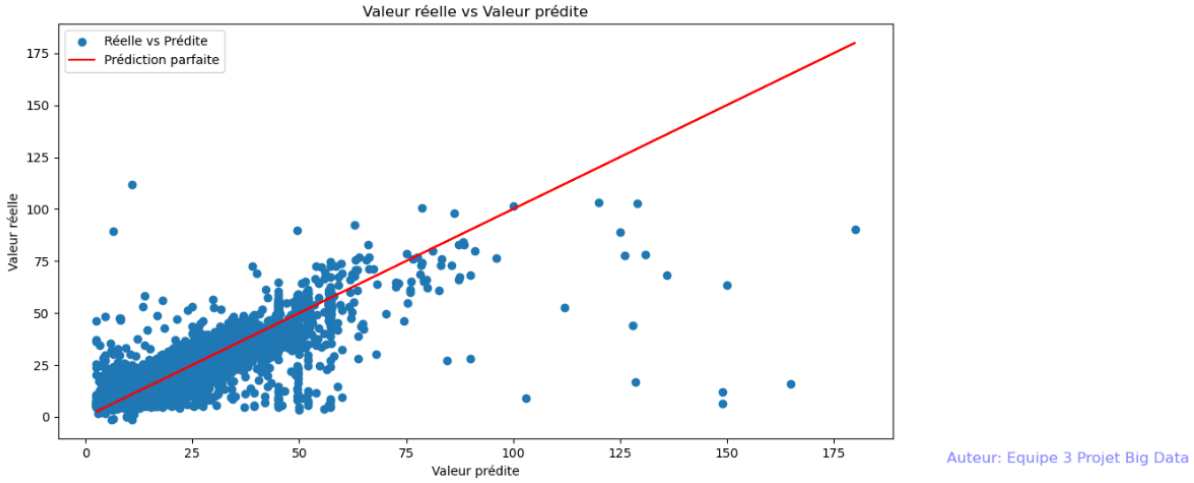


Illustration 8 : Résultat XGBoost
Source : résultats script sur Python

7. Qualité d'un modèle

Coefficient de détermination

Le coefficient de détermination mesure la proportion de variance expliquée par le modèle par rapport à la variance totale des données. Il est compris entre 0 et 1. Plus il est proche de 1, plus le modèle est performant. S'il est proche de 0, cela indique que le modèle n'est pas en mesure d'expliquer les données. Il est calculé à partir de l'erreur quadratique moyenne du modèle (EQM).

$$R^2 = 1 - (\frac{EQM\ du\ modèle}{EQM\ de\ la\ moyenne})$$

où $EQM\ du\ modèle$ est l'erreur quadratique moyenne du modèle de régression et $EQM\ de\ la\ moyenne$ est l'erreur quadratique moyenne si nous avions simplement prédit la moyenne de toutes les valeurs réelles [12].

Erreur absolue moyenne

L'erreur absolue moyenne (MAE pour *Mean Absolute Error* en anglais) est une mesure de la qualité d'un modèle de prédiction. Elle sert à évaluer la précision du modèle dans la prédiction des valeurs de valeurs numériques. Il est calculé grâce à la formule :

$$MAE = \frac{\sum_i^n Y_i - X_i)}{n}$$

où X_i est la valeur réelle, Y_i est la valeur prédite et n est le nombre d'erreurs.
Elle est facilement interprétable car elle permet d'exprimer l'erreur en unités de la variable cible [13].

Erreur quadratique moyenne

L'erreur quadratique moyenne est une mesure de l'écart entre un ensemble de valeurs réelles et les valeurs prédites par un modèle. Elle permet d'évaluer la précision des prédictions d'un modèle. Sa formule de calcul est :

$$RMSE = \sqrt{(\frac{1}{n} * \sum_i^n (Y_i - X_i)^2)}$$

où Y_i est la valeur réelle de l'observation i, X_i est la valeur prédite de l'observation i et n est le nombre total d'observations.
Elle est également très facile à interpréter car elle s'exprime en unités de la variable cible [14].

8. Schéma factuel

L'illustration 9 représente un schéma factuel. Le fait correspond à un tarif appliqué à une course en taxi effectuée à une date donnée d'une adresse de départ vers une adresse d'arrivée.

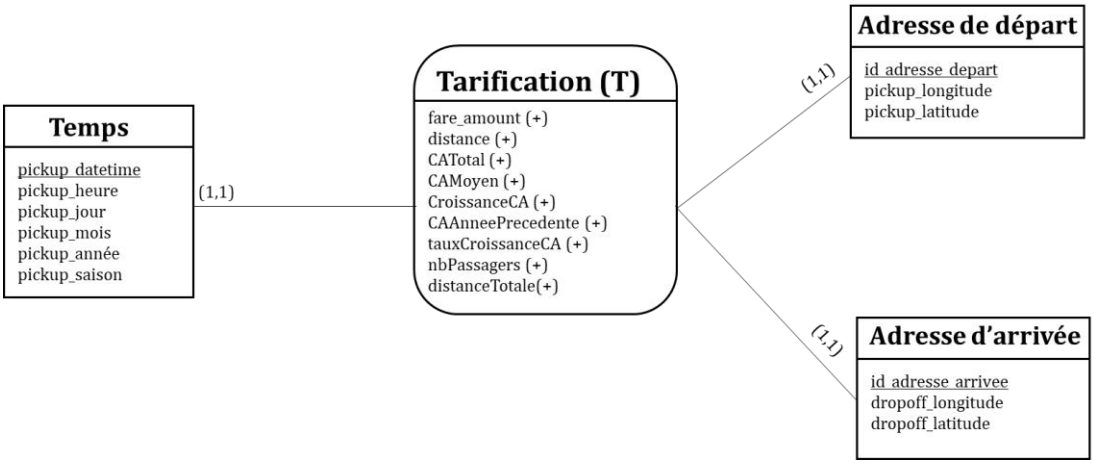


Illustration 9 : Schéma factuel
Source : réalisée sur Powerpoint

Dans notre schéma factuel, nous avons 3 dimensions qui sont liées à la table de fait « Tarification » :

- la dimension « Temps » qui contient des attributs qui indiquent le temps où la course en taxi a débuté ;
- la dimension « Adresse de départ » qui contient des attributs qui indiquent les coordonnées géographiques du lieu de départ de la course ;
- la dimension « Adresse d'arrivée » qui contient des attributs qui indiquent les coordonnées géographiques du lieu d'arrivée.

La table de fait « Tarification » contient 9 indicateurs :

- fare_amount : le tarif de la course de taxi
- distance : la distance parcourue lors de la course de taxi ;
- CATotal : le chiffre d'affaires total ;
- CAMoyen : le chiffre d'affaires moyen ;
- CroissanceCA : la croissance du chiffre d'affaires ;
- CAAnneePrecedente : le chiffre d'affaires de l'année précédente ;
- tauxCroissanceCA : le taux de croissance du chiffre d'affaires ;
- nbPassagers : le nombre de passagers dans le taxi lors de la course ;
- distanceTotale : la distance totale parcourue par le taxi.