DataScientest • com

Rapport Technique d'évaluation

ParisPyVelib

Promotion Avril 2021

Participants :

Céline Doussot
Tarik Anouar
Hermine Berthon
Sous la direction de Jérémy Robert

Contexte

Dans ce rapport, nous avons choisi de faire un jeu de question(italique)-réponse pour faciliter la compréhension.

Contexte d'insertion du projet dans votre métier.

Lorsque l'on se promène dans Paris, il est possible d'apercevoir le long des grandes artères de grandes bornes bleues atypiques. Ces totems, comme certains peuvent les appeler, comptabilisent le passage des vélos et ont pour vocation d'analyser le comportement des cyclistes évoluant dans les rues parisiennes.

En effet, la mairie de Paris tend à encourager les habitants à utiliser le vélo comme moyen de transport pour désengorger la ville et réduire la pollution ambiante. Cela se traduit, entre autres, par l'installation de nombreuses pistes cyclables ou encore par des aides financières pour les particuliers. D'où l'essor des compteurs de vélos comme moyen de voir l'impact de ces mesures.

De manière détaillée, ceux-ci comptabilisent chaque jour le nombre de cyclistes, par heure à divers endroits de la ville, le plus souvent dans les deux sens de circulation. Ces différentes données générées donnent lieu à un dataset fourni de 2018 à aujourd'hui. L'analyse des données pourrait non seulement permettre à la mairie de Paris de voir la circulation réelle des cyclistes mais en plus d'aménager l'espace quotidien pour faciliter les déplacements à vélo sur les axes où il y a le plus de passage.

Objectifs

Quels sont les principaux objectifs à atteindre ? Décrivez en quelques lignes.

Les principaux objectifs à atteindre concernent l'analyse de la circulation des vélos dans Paris en fonction de la météo et des événements quotidiens (weekends, jours fériés, confinement, etc.) afin de pouvoir prédire les futurs comportements des utilisateurs en machine learning.

Pour chacun des membres du groupe, préciser le niveau d'expertise autour de la problématique adressée ?

Hermine : Mes compétences se concentrent dans le domaine scientifique et en particulier dans celui de la chimie. J'ai l'habitude de traiter les données que je génère, qui pourraient se traduire en petits datasets. Ce qui fait que mes analyses habituelles se font de manière manuelle, par exemple en traçant les relations entre divers paramètres sur excel et en les présentant sur powerpoint. J'avais au début de la formation peu d'expertise dans la data analyse et dans le machine learning.

Céline : Le niveau d'expertise sur l'étude des compteurs vélib dans Paris était dans mon cas très limité, n'ayant que très peu pratiqué de vélo à Paris ni même eu connaissance de l'existence de ces données en open data. J'ai cependant eu un intérêt particulier à choisir ce sujet d'étude puisqu'à titre personnel, je m'intéresse de plus en plus au vélo comme moyen de transport et à son évolution dans la pratique. Je n'avais au départ avant la formation aucune connaissance dans le traitement de ce genre de données, n'ayant l'habitude de passer que par Excel pour traiter des petites données et affichage graphique usuel. Mes connaissances dans la gestion de projet dans le domaine de la chimie via mon travail m'ont heureusement permis de m'adapter facilement à ce nouveau projet.

Tarik : La problématique qui nous a été adressée par ce projet réside principalement pour ma part autour du machine learning. Une partie de mon activité professionnelle consiste à analyser des jeux de données basées sur des lois connues avant l'analyse. Une des difficultés est d'identifier quel type d'algorithmes de machine learning est le plus adaptée à la problématique. En outre, la diversité et la quantité d'algorithmes de modèle disponible à ce jour ne facilitent pas cette tâche d'identification. Enfin, le temps de traitement, de tests et d'optimisation d'un modèle de machine learning peut être relativement long, ce qui se traduit par une contrainte forte dans un projet d'analyse de données. Le retour d'expérience sur ce projet m'a permis de me confronter avec ces difficultés et de mieux appréhender les enjeux et l'intérêt de la modélisation d'une problématique via un modèle de machine learning.

Êtes vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l'apport de ces intéractions.

Cela ne s'est pas avéré nécessaire pendant le déroulement du projet, le sujet étant plutôt clair et simple.

Avez vous connaissance d'un projet similaire au sein de votre entreprise, ou bien dans votre entourage ? Quel est son état d'avancement ? En quoi vous a-t-il aidé dans la réalisation de votre projet ? En quoi votre projet contribue-t-il à l'améliorer ?

La problématique exploitée dans ce rapport a déjà fait l'objet d'un projet mené par d'autres apprenants dans le cadre d'un bootcamp Data Analyst pour la période du 1er septembre 2019 au 31 décembre 2020 . Leurs résultats étaient orientés autour du trafic des vélos, selon quatre axes : la cartographie, les facteurs d'influence, l'impact des accidents et enfin la prédiction avec des modèles de machine learning. Les axes d'amélioration évoqués sont par exemple, d'élargir la période étudiée et de pousser plus loin les études d'impact en les diversifiant. Ce projet était particulièrement intéressant pour partir d'une base de départ et ainsi mieux comprendre le dataset. En prenant en compte certaines perspectives du projet précédent, nous avons élargi les données pour une période de 2018 à 2021. Et d'autre part, nous nous sommes centrés en particulier sur l'impact de la météo sur la circulation des cyclistes.

Un autre rapport beaucoup plus complet qui étudie l'évolution de la mobilité de tous les modes de transport de 1976 à 2020 dans le Grand Paris (vélo, pieds, voiture, transport en commun…) a permis d'avoir une vision plus globale du projet et de toutes les études pouvant être réalisées dans cette même thématique.

Data

Cadre

Quel(s) jeu(x) de donnée(s) avez vous utilisé pour atteindre les objectifs de votre projet ?
Ces données sont-elles disponibles librement ? Dans le cas contraire, qui est le propriétaire de la donnée ?
Décrivez la volumétrie de votre jeu de données ?

Plusieurs jeux de données ont été utilisés au cours du projet. Tout d'abord, du site open data Paris, nous avons récupéré les données des compteurs de vélos de 2018 à 2021, le data frame résultant compte un peu moins de 2 millions de lignes 9 colonnes au total, après fusion des datas des différentes années.

Un autre jeu de données a également été rajouté concernant les données météos. Un accès à l'historique de Météo France Paris sous format horaire était payant (seul le format mensuel était accessible librement). Nous avons donc opté pour un autre site permettant un accès libre sur toutes les données météos provenant de la ville d'Athis-Mons (91), seul compteur disponible en île-de-France. Cela correspond, après sélection de 2018 à 2021, (même période que pour les compteurs vélib) à un dataframe de 9944 rows × 82 columns. Celui-ci a ensuite été réduit, par la suite, pour ne sélectionner que les données intéressantes.

Enfin, différents jeux de données sur les dates de vacances scolaires, les jours fériés, les dates du confinement aussi en open data ont permis de compléter nos analyses.

Pertinence

Avez vous eu à nettoyer et à traiter les données ? Si oui, décrivez votre processus de traitement.
Quelles variables vous semblent les plus pertinentes au regard de vos objectifs ?
Quelle est la variable cible ?

Pour les données générées par les compteurs entre 2018 et 2021, il a été nécessaire de bien homogénéiser les données. Par exemple, certains compteurs ont été remplacés par des nouveaux, ce qui a pu changer l'identifiant ou les coordonnées géographiques. De même, le format de certaines variables, en particulier les adresses, n'était pas identique entre les différents dataframe (2018, 2019 différent de 2020, 2021).

Pour cela, nous avons dans un premier temps identifié les problèmes tels que la gestion des doublons. Ensuite, nous avons harmonisé tous les fichiers afin d'attribuer chacune des coordonnées à une adresse puis créé un nouvel identifiant unique pour chacun des compteurs.

Dans un second temps, nous avons transformé, supprimé et ajouté certaines colonnes qui nous semblaient pertinentes pour la suite tel que l'ajout de colonnes séparant l'heure, le jour, le mois et l'année de comptage.

Enfin, nous avons renommé certaines colonnes pour faciliter la compréhension. Sur le dataframe final, la colonne cible est le comptage horaire mais les autres colonnes importantes pour son étude et prédiction, sont les colonnes temporelles ainsi que les coordonnées géographiques.

Lorsque nous avons voulu introduire les données météos à notre dataset initial, nous nous sommes aperçus que de nombreuses colonnes étaient inutiles ou comportaient beaucoup de NANs. De plus, les données météo n'étaient mises à jour que toutes les 3h et non toutes les heures comme notre dataset standard. Pour pallier ce problème, nous avons sélectionné certaines colonnes pertinentes pour l'étude de la fréquentation vélo comme la pluie, le vent ou encore la neige, puis nous avons utilisé un backfill pour compléter les données manquantes sur certaines heures.

Enfin les données vacances, jour fériés et confinement ont été ajoutées afin d'apporter plus de paramètres influant dans l'étude du comptage horaire.

Décrivez la distribution de ses valeurs ?

Concentrons maintenant sur la distribution des variables qui nous intéressent, nous pouvons trouver en Annexe I les boxplot de chacune des variables du dataframe.

Les latitudes et longitudes sont globalement similaires selon les années, ce qui est logique. Les différences proviennent des compteurs qui se sont ajoutés, qui ont été changés voire supprimés. Les dates et heures de comptage et d'installation correspondent bien au calendrier et aux aspects logiques du quotidien (pas de 13ème mois ou 13ème heure). Les valeurs de 2021 ne sont pas exactement les mêmes mais cela est normal car le jeu de données ne couvre que de janvier à mai.

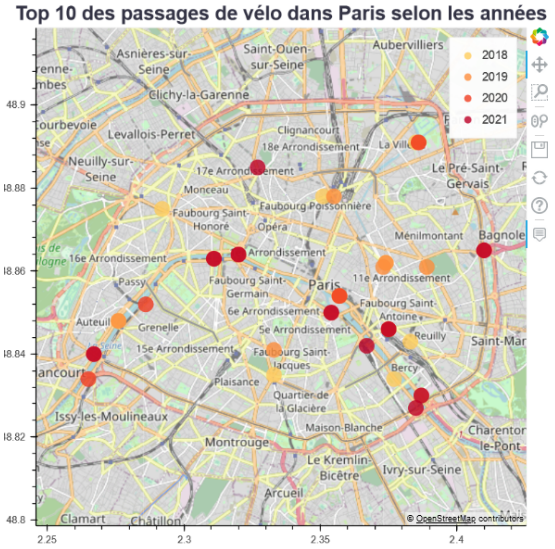
Les températures enregistrées vont de -10°C jusqu'à plus de 40°C selon les années. Comme expliqué précédemment, 2021 diffère car il couvre moins de mois et donc il n'y a pas comptabilisation des températures élevées des mois d'été. Les autres variables sur l'humidité et la vitesse du vent sont semblables selon les années et cohérentes avec la réalité.

Si nous regardons la variable High_ice, les valeurs sont entre janvier et février. Cependant, il est intéressant de voir qu'en 2020 il n'y a pas eu de neige à Paris contrairement aux autres années. La variable précipitation a des valeurs sur toute l'année, avec de fortes précipitations de mai à août. Cette variable, ainsi que la variable High_ice ont des valeurs négatives aberrantes qui sont enlevées dans le dataframe final.

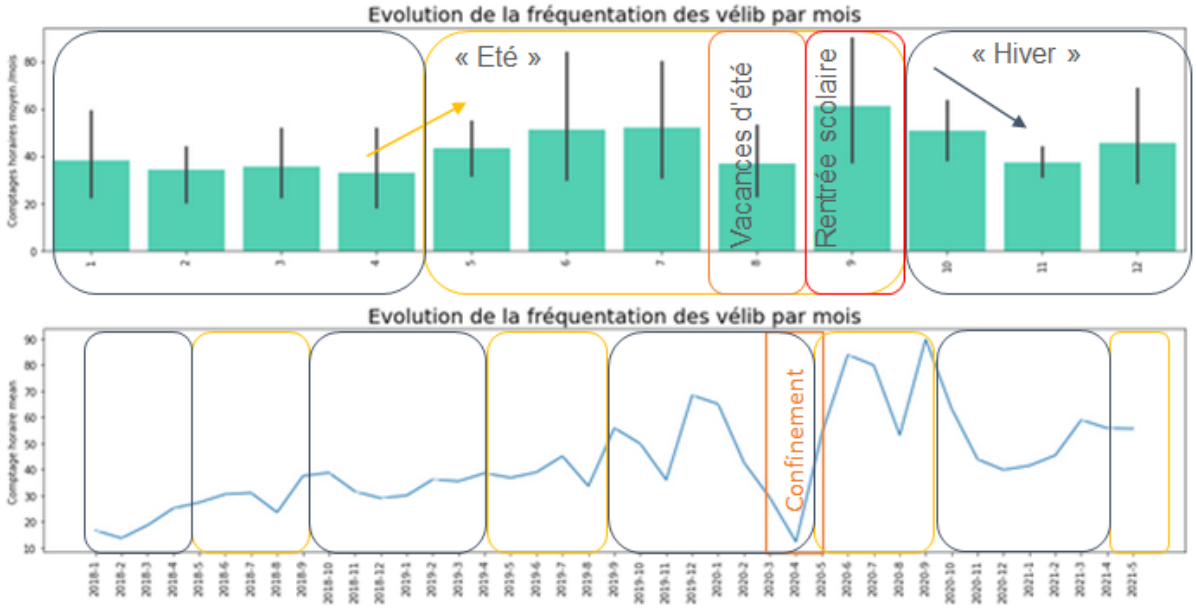
Analyse de la variable cible

Avez-vous identifié des relations entre différentes variables ? Entre variables explicatives ? et entre vos variables explicatives et la/les cible(s) ?
Quelles particularités de votre jeu de données pouvez-vous mettre en avant ?

Notre variable cible, Count_by_hour, comptabilise le nombre de vélo qui passe chaque heure devant un compteur entre 2018 et 2021. En Annexe II, les différentes visualisations de la variable cible sont visibles. De manière générale, la distribution des compteurs est dans tout Paris, le long des axes les plus empruntés, comme nous pouvons l'observer sur la représentation graphique (illustration ci-dessous). Les appareils comptabilisant le plus de passage par an (= top 10) évoluent au fil des années, sûrement grâce à l'ajout de nouveaux compteurs. Néanmoins, nous pouvons remarquer que la circulation est importante le long de la Seine et au niveau des axes ferroviers.

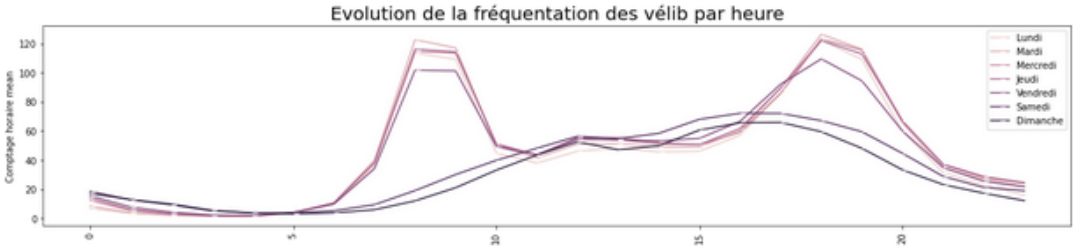


Lorsque nous analysons la variable plus en détail, selon les mois de l'année, il semble y avoir une différence entre les mois que l'on pourrait qualifier "d'hiver", en bleu et "d'été" en jaune. Dans les premiers, la température est plus fraîche avec de gros écarts de température entre la matinée et l'après-midi et les seconds ont une température plus douce et homogène. Nous pouvons noter que le mois d'août enregistre particulièrement peu de vélos. Cela est synonyme que les parisiens partent souvent en vacances durant ce mois, avec une recrudescence à la rentrée. De 2018 à 2021, nous voyons une tendance en augmentation, du nombre de vélos (à mettre néanmoins en parallèle avec le nombre croissant de compteurs), avec une coupure lors du confinement stricte et une grosse augmentation juste après à la fin de celui-ci.



La dernière observation très caractéristique de la variable est sur la différence entre la semaine et les weekends. En semaine, deux pics de circulation sont présents le matin et l'après-midi, correspondant aux heures de pointe du travail. Les deux pics ayant une allure quasi identique, cela pourrait montrer que ce sont les mêmes cyclistes à l'aller et au retour de leur trajet domicile-travail. Le weekends, la distribution est plus

dispersée tout au long de la journée, avec une gaussienne plus aplatie. Nous avons remarqué également que les jours fériés se comportent comme les weekends.



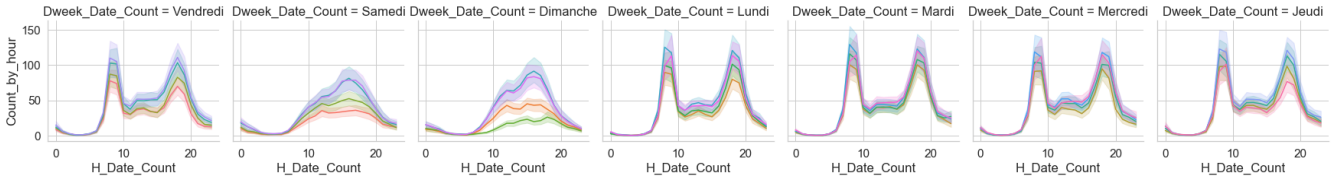
Notre jeu de données est particulier car il se prête facilement aux analyses graphiques. Ainsi, nous avons voulu laisser une place importante aux analyses visuelles des données.

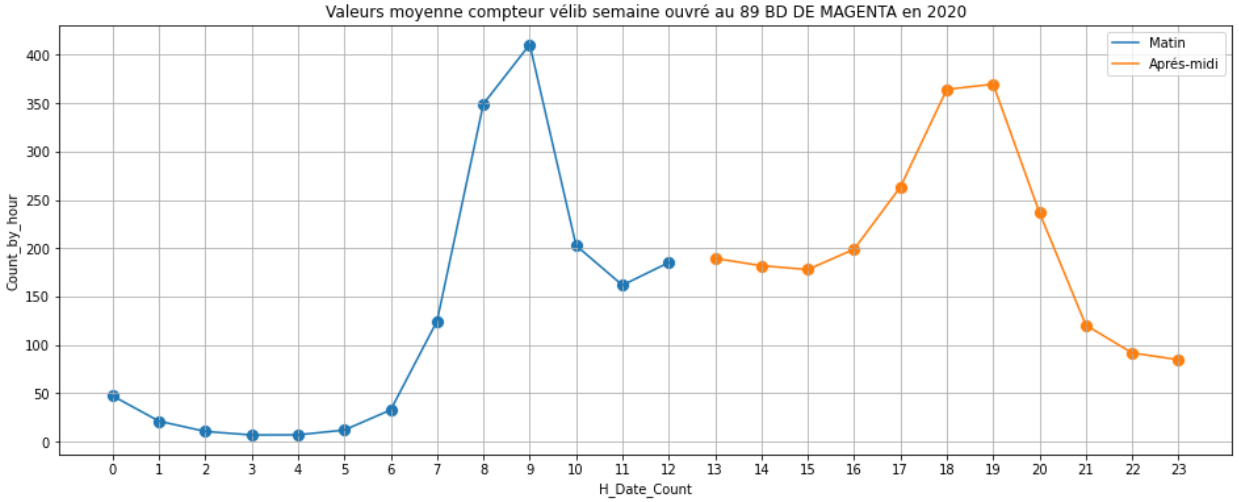
Projet

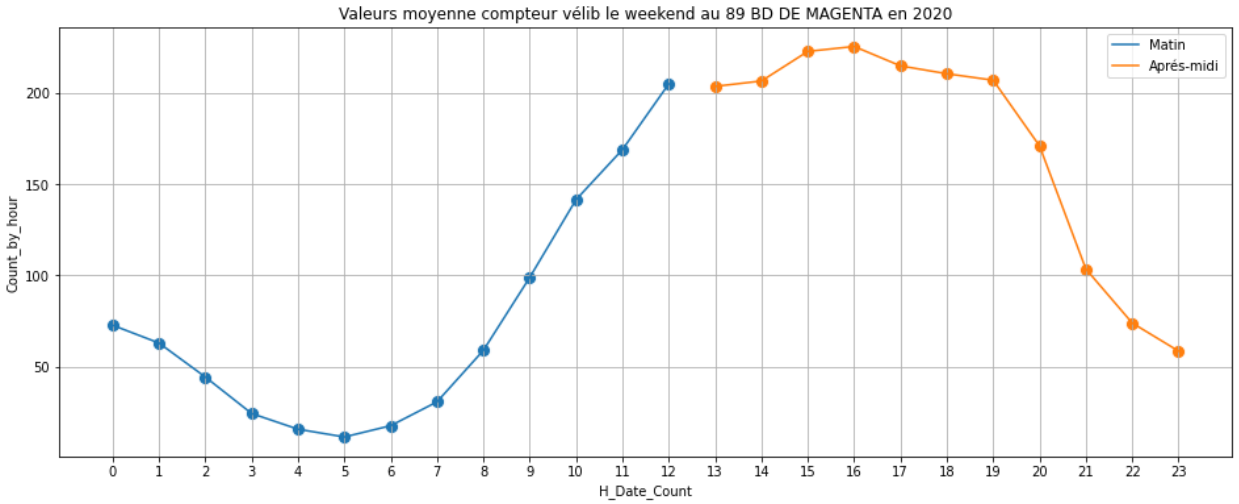
Classification du problème

À quel <u>type</u> de problème de machine learning votre projet s'apparente-t-il ? (Classification, régression, clustering ….)

L'objectif réside dans la prédiction du nombre de vélos passant devant chaque compteur à Paris. Ainsi, la visualisation des données a permis de mettre en lumière une influence du trafic vélo en fonction du jour de la semaine (voir illustration ci-dessous), du moment de la journée : matin ou après-midi (voir illustration ci-dessous). Il est aussi tout à fait probable que la météo joue un rôle significatif. Ces éléments s'apparentent donc à un problème de machine learning de type classification.







À quelle tâche de machine learning votre projet s'apparente-t-il ? (détection de fraude, reconnaissance faciale, analyse de sentiment …) ?

La tâche de machine learning utilisée est le classement. Une étiquette est déterminée pour chaque classe {0, 1, 2, 3} de la cible, en l'occurrence le nombre de vélo compté par heure (variable Count_by_hour).

Quelle est la métrique de performance principale utilisée pour comparer vos modèles ?

La métrique utilisée afin de comparer les différents modèles est : mean_test_score. La matrice de confusion est aussi un outil visuel qui à permis d'évaluer la performance de classification.

Avez vous utilisé d'autres métriques de performances qualitative ou quantitative) ? Si oui, détaillez.

Pour le modèle KNN, la précision du modèle en fonction de la valeur de K à permis d'évaluer visuellement la meilleure performance en fonction de l'hyper paramètres K. Toutefois, ce modèle n'a pas été retenue

Choix du modèle & Optimisation

Quels algorithmes avez vous essayés ?

Les différents algorithme testé sont :

- KNN : modèle non retenue
- SVM : modèle non retenue
- Random Forest : modèle retenue
- Moyenne année n-1 : modèle se basant sur la moyenne du nombre de vélos compté à l'année n-1. Ce modèle à l'avantage d'avoir un niveau d'interprétabilité élevée comparativement au modèle Random Forest. De plus, les performances de ce modèle reste équivalente à une classification de type Random Forest

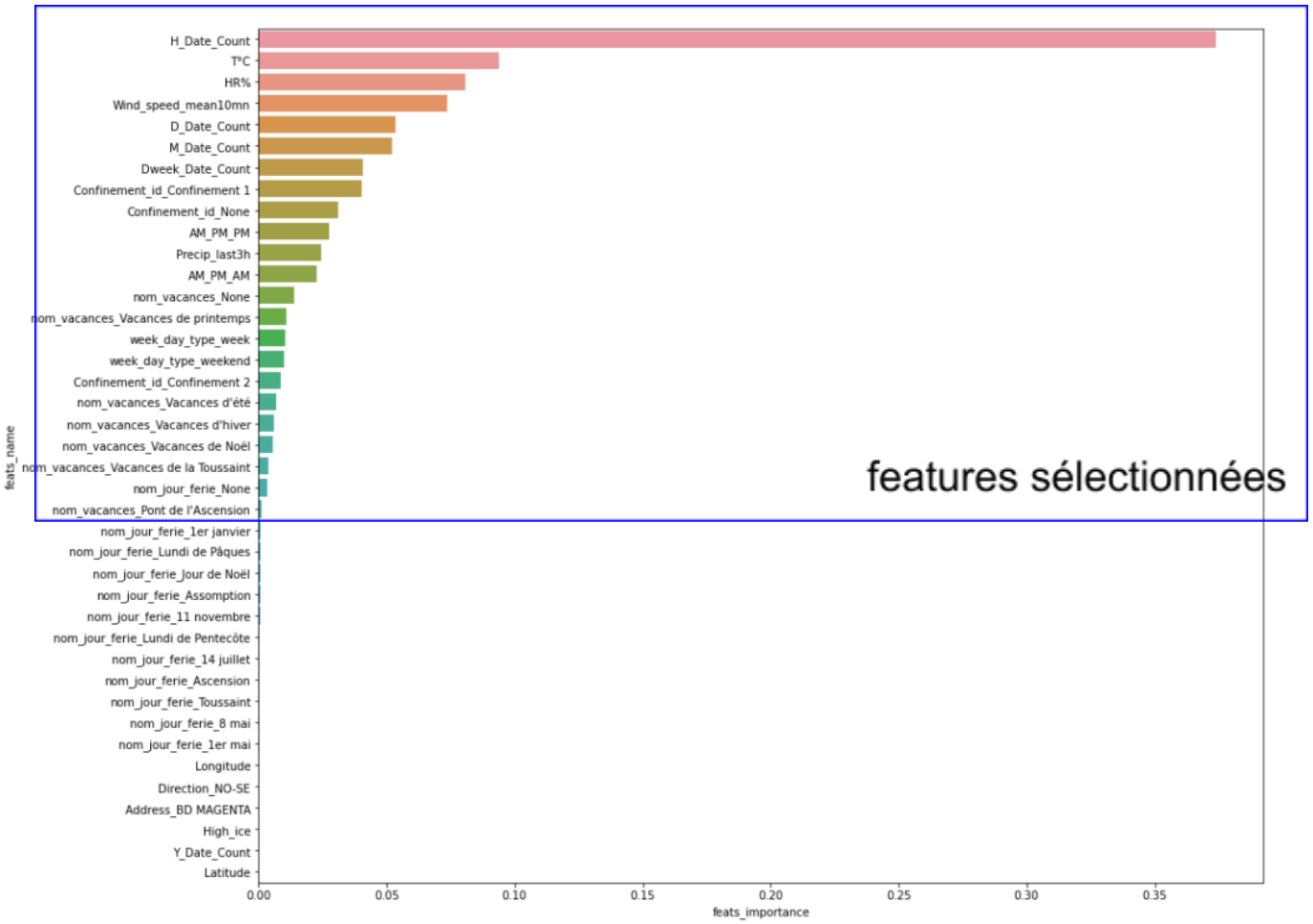
Décrivez celui / ceux que vous avez retenu et pourquoi ?

Nous avons retenu l'algorithme Random Forest, car ce modèle nous donnait un score supérieur à 80% mais aussi parce que ce modèle est le plus adapté pour une problématique de classification. En outre, un score supérieur à 80% reste suffisant pour prédire un nombre de vélos avec une incertitude acceptable.

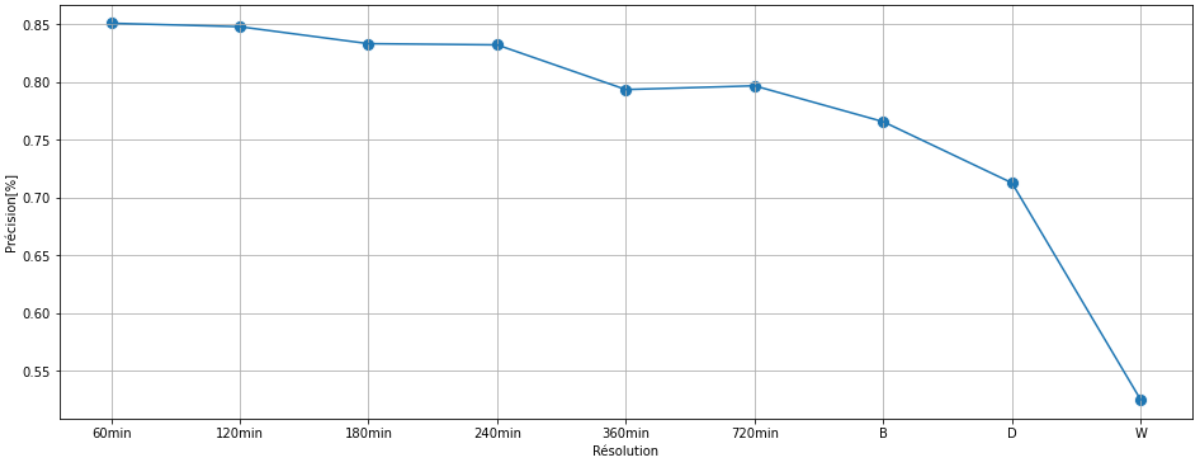
Qu'est ce qui a engendré une amélioration significative de vos performances ?

Les principales améliorations des performances du modèle concernent le temps de traitement. Pour cela deux axes ont été exploré :

- Réduction du nombre de features : l'importance des features à révélé que seulement 20 features sur 40 au total sont pertinente pour le modèle de classification. Ainsi la diminution des features par deux à été une amélioration significative afin de réduire le nombre de variables à traiter par le modèle.
  Ci-dessous l'illustration de l'importance de chaque features :



- Évaluation de la résolution temporelle : nativement le nombre de vélos est comptabilisé toutes les heures. Toutefois, nous avons évalué d'autres résolutions temporelles afin de mesurer si la précision restait équivalente. Ainsi les résolutions suivantes ont été évaluées : 60 mn, 120 mn, 180 mn, 240 mn, 360 mn, 720mn, B (Business Day), D (Day), W (Week).
  Enfin la résolution 60 mn et 120 mn donne une précision équivalente (voir ci-dessous), il est donc possible de réduire la dimension de la matrice des features en utilisant une résolution de 120 mn, ce qui contribue à l'amélioration des performances du modèle.



NB : l'analyse de la résolution temporelle à été réalisée uniquement sur un seul compteur.

Avez-vous analysé les erreurs de votre modèle ? Cela a-t-il contribué à son amélioration ? Si oui, décrivez.

Non, les erreurs du modèle non pas été analysées, nous n'avons pas connaissance de ce type de pratique ainsi que la démarche à réaliser pour ce type d'analyse.

Détaillez quelle a été votre contribution principale dans l'atteinte des objectifs du projet.

cf répartition des tâches

Description des travaux réalisés

Répartition de l'effort sur la durée et dans l'équipe

Morceler votre projet en un maximum de tâches unitaires. Produisez le diagramme de Gantt a posteriori en spécifiant qui s'est occupé de quelle tâche et à quelle moment. (joindre le diagramme en annexe du rapport)

En Annexe III, nous pouvons trouver le diagramme de Gantt fabriqué a posteriori. De manière générale, la distribution des tâches s'est faite au fil des idées et des hypothèses de chacun. En particulier, Céline a fait une grosse partie du preprocessing et de la description visuelle du dataframe, Hermine s'est concentrée sur la visualisation cartographique du dataframe et Tarik s'est attelé au machine learning.

Les différentes tâches se sont faites en collaboration grâce à l'utilisation de GitHub, google collab et drive. Des points réguliers entre nous ou avec Jérémy ont permis de vérifier nos progrès et de voir les points bloquants ensemble.

Bibliographie

Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc...) vous êtes vous appuyé pour réaliser votre projet ?

Partie dataframe et preprocessing :

1. Données sources

Datasets des compteurs de vélo de 2018 à 2021 :
https://opendata.paris.fr/explore/dataset/comptage-velo-donnees-compteurs/information/?disjunctive.id_compteur&disjunctive.nom_compteur&disjunctive.id&disjunctive.name
https://opendata.paris.fr/explore/dataset/comptage-velo-historique-donnees-compteurs/information/

Dataset météo :
https://public.opendatasoft.com/explore/dataset/donnees-synop-essentielles-omm/table/?flg=fr&sort=date

Dataset vacances scolaires :
https://www.data.gouv.fr/fr/datasets/vacances-scolaires-par-zones/

Dataset jours fériés :
https://www.data.gouv.fr/fr/datasets/jours-feries-en-france/

Dataset confinement : Données créent par Tarik

2. Bibliographie

Projet PyCycle : https://studio.datascientest.com/project/pycycle/

Dossier sur l'évolution des mobilités dans le Grand Paris :
https://www.apur.org/fr/nos-travaux/evolution-mobilites-grand-paris-tendances-historiques-evolutions-cours-emergentes
Article du Monde :
https://www.lemonde.fr/les-decodeurs/article/2021/09/19/a-paris-aux-heures-de-pointe-les-velos-sont-plus-nombreux-que-les-voitures-sur-certains-axes_6095203_4355770.html

Partie Visualisation :
Compteurs vélo à Paris:
https://compteurs.parisenselle.fr/

L'utilisation de Bokeh avec a été réalisée avec le support du module de cours optionnel 113 et du site bokeh : https://docs.bokeh.org/en/latest/docs/user_guide/geo.html

Partie ML :
https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

Difficultés rencontrées lors du projet

Quel a été le principal verrou scientifique rencontré lors de ce projet ? Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elle vous ont ralenti dans la mise en place de votre projet :

Prévisionnel : (tâches qui ont pris plus de temps que prévu etc ….)

La réalisation du Notebook 1, qui sert à importer et traiter les données, a pris beaucoup plus de temps que prévu. En particulier, à cause des besoins d'harmonisation du dataframe après fusion des différents compteurs vélib entre 2018 et 2021. Mettre l'écriture des adresses sous le même format a été la tâche prenant le plus de temps. D'ailleurs, nous n'avons pas réussi à totalement automatiser cette partie par un code et il a été nécessaire de la finaliser manuellement, faute de temps. Une autre tâche sous-estimée a été d'obtenir des coordonnées uniques par adresse, quel que soit le sens de direction. Les arrondis ou valeurs étant légèrement différentes en fonction de la source de données utilisée, il a fallu trouver un compromis d'arrondir à 3 chiffres puis de modifier manuellement les doublons.

Jeux de données : (Acquisition, volumétrie, traitement, agrégation etc….)

Le jeu de données total faisant ~2 millions de lignes, il y a eu des moments difficiles pour trouver le code le plus optimisé, lors de la modification de certaines colonnes par exemple. En effet, modifier avec une simple boucle pouvait parfois être beaucoup trop long et il était nécessaire d'optimiser le code afin de faire le changement de la colonne entière et non pas ligne par ligne.

Compétences technique / théoriques : (Timing d'acquisition des compétences, compétence non proposée en formation etc…)

Les compétences qui ont pu manquer par rapport au projet étaient celles en rapport avec le text mining ou du moins savoir comment harmoniser une écriture. Les modules y faisant rapport interviennent tard dans la formation. Notre jeu de données s'y prêtant bien, le module

Power BI a aussi manqué. Il aurait été intéressant de l'utiliser plus rapidement dans le projet afin de visualiser rapidement le jeu de données. De façon générale, le projet étant une application de la formation, il aurait peut être mieux valu le faire démarrer un peu plus tard afin d'avoir suffisamment de notions sur différents modules proposés.

Pertinence : (de l'approche, du modèle, des données etc ...)

L'identification du modèle de machine learning à utiliser ainsi que la visualisation des prédictions ont été cruciaux pour comprendre quels leviers utiliser pour développer des axes d'étude et d'amélioration. Les différents modules d'apprentissage abordés lors de la formation ne nous ont pas apporté l'expérience suffisante pour avoir une approche pertinente et rapide d'un projet basé majoritairement sur le machine learning. Cela a occasionné un gros ralentissement du projet.

IT : (puissance de stockage, puissance computationnelle, etc….)

La plupart des datasets manipulés ne posait pas de problème concernant l'espace de stockage requis. Toutefois, la fusion et la multiplication de plusieurs sources de données peuvent être une source d'erreur. Ainsi, un service de stockage en ligne (Google Drive) a permis à chaque membre du projet d'utiliser une source de données unique pour exploiter les mêmes fichiers.
Enfin, les temps de traitement des modèles de machine learning utilisent beaucoup de mémoire. Les nombreux tests et optimisation d'hyper-paramètres, d'influence des features, etc. ont nécessité beaucoup de temps. Afin de palier à ce problème, deux stratégie ont été adopté :

- Modéliser un seul compteur pour réaliser l'étude du modèle de machine learning.
- Utiliser un service cloud pour exécuter plusieurs notebooks en même temps
- Modéliser les compteurs par années afin de diminuer la dimension du nombres de données à entraîner

Une fois le modèle, ainsi que les paramètres optimum déterminés, l'évaluation sur tous les compteurs requiert entre 1h30 à 2h.

Bilan & Suite du projet

Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non. Dans le cas contraire, quelles pistes d'amélioration suggérez-vous pour améliorer les performances de votre modèle ?

L'objectif de se familiariser avec les données, de les nettoyer et les traiter dans le but de pouvoir s'en servir pour de la visualisation ou de la prédiction a été accompli. De même, on peut considérer que la partie analyse de données par visualisation a également été un succès. En effet, nous avons réussi à tirer des informations et des tendances sur l'utilisation des vélos dans Paris (fréquentation majoritairement en heure de pointe en semaine pour trajet domicile-travail, augmentation au fil des années, effet du 1er confinement, quartier les plus fréquentés…). Cependant, l'analyse n'a pas pu être poussée comme souhaitée. Initialement, nous voulions étudier l'augmentation de la part de vélo par rapport aux autres modes transports (voiture, métro…). Les datasets source ont été trouvés sur ce sujet mais non exploités, faute de temps. Une autre piste d'amélioration pourrait concerner l'analyse en temps réel de la circulation des cyclistes.

Concernant la partie modélisation, l'objectif a été partiellement atteint. La comparaison entre un modèle "simple" basé sur une moyenne sur l'année n-1 et un modèle Random Forest donne une précision équivalente. Il est envisageable d'améliorer les scores de prédictions en explorant plusieurs pistes :

- Simplification des features actuelles
- Analyse des erreurs du modèle et biais potentiel du dataset
- Ajouts de features (jours de grèves, événements sportifs, manifestation, etc)
- Test d'autre algorithme de classification

Cependant, la précision de modèle atteint une moyenne d'environ 80% ce qui reste suffisant tout en s'accordant une incertitude raisonnable. L'optimisation du modèle demanderait beaucoup plus de temps d'étude qu'initialement envisagé en début de projet.

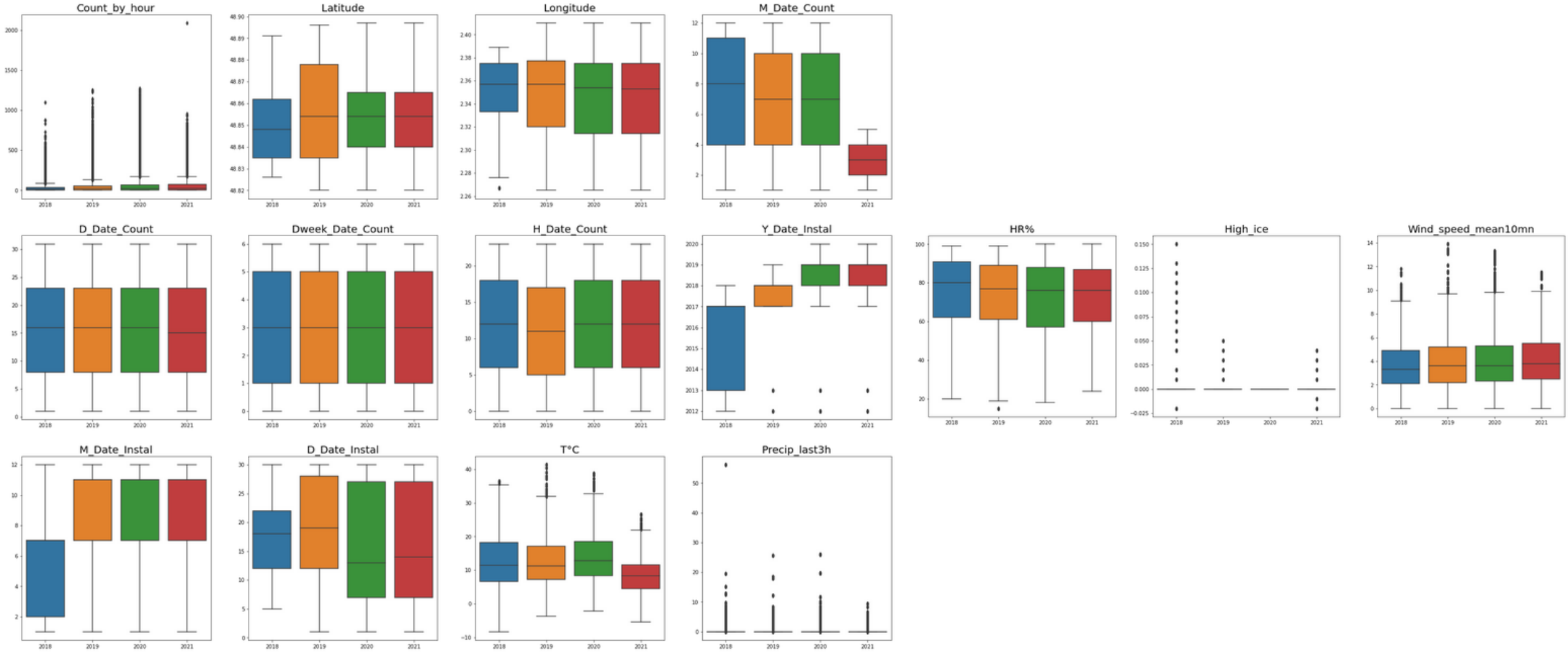
S'ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s'inscrire ? Détaillez.

Tout d'abord, l'analyse et la prédiction du trafic cycliste peut permettre à la mairie de Paris d'étudier de manière précise l'impact de certains facteurs (météo, accidents, etc.). Ceux-ci influent sur le comportement des vélos et ainsi la mairie peut optimiser leur circulation en les protégeant grâce à l'aménagement de pistes cyclables efficaces.

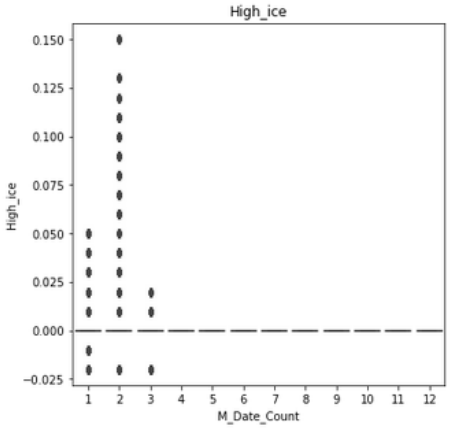
De plus, en combinant l'analyse et le machine learning, il est envisageable de pouvoir créer à l'image d'autres applications pour voiture (Waze), une application pour améliorer les trajets des vélos au quotidien à Paris en fluidifiant le trafic, par exemple. Celui-ci pouvant être étendu aux autres grandes agglomérations comme Bordeaux.

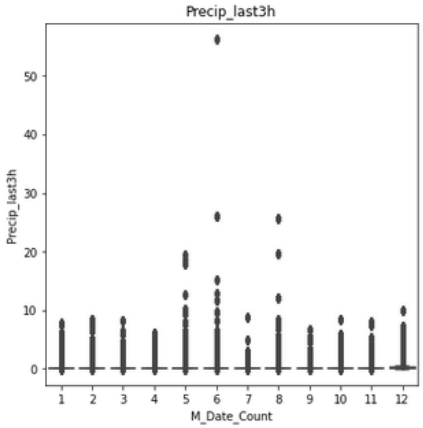
Annexe I : Distribution des valeurs du dataframe

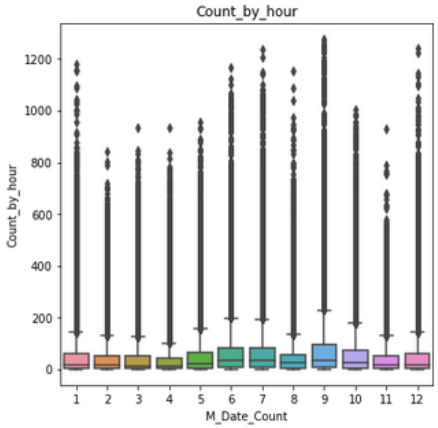
Sous forme de Boxplot selon les années :

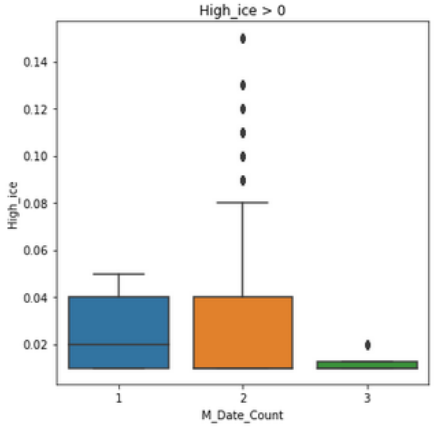


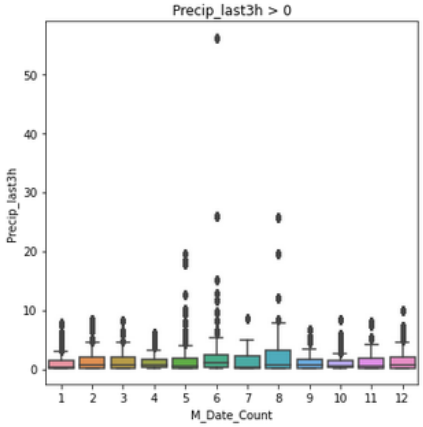
Sous forme de Boxplot selon les mois :

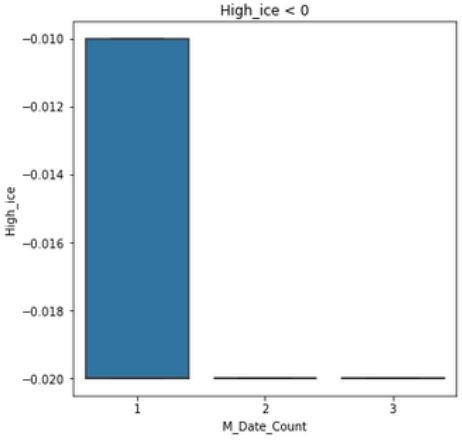


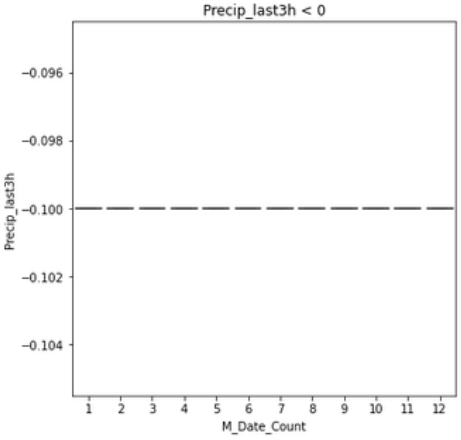




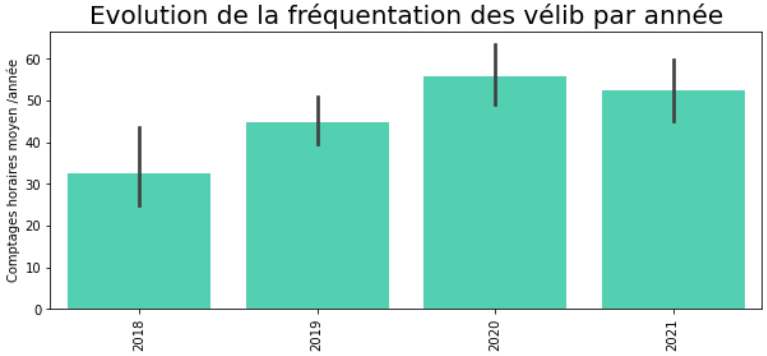


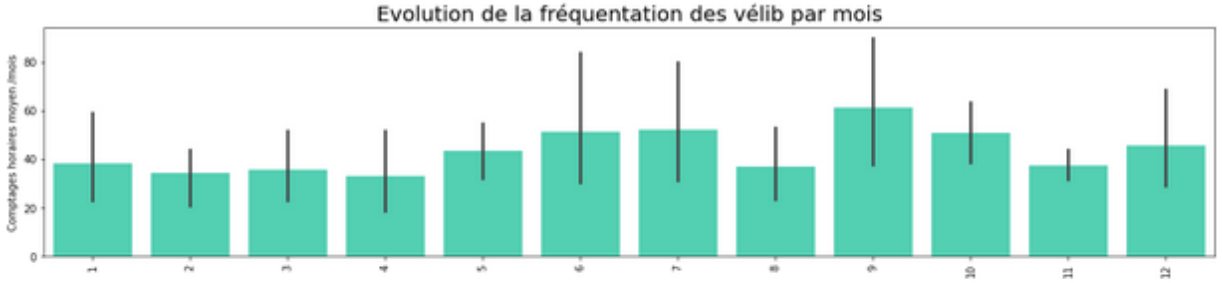


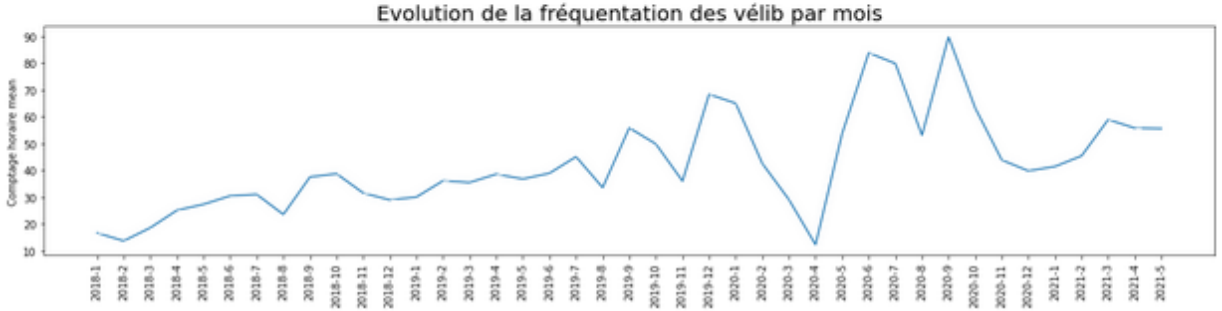


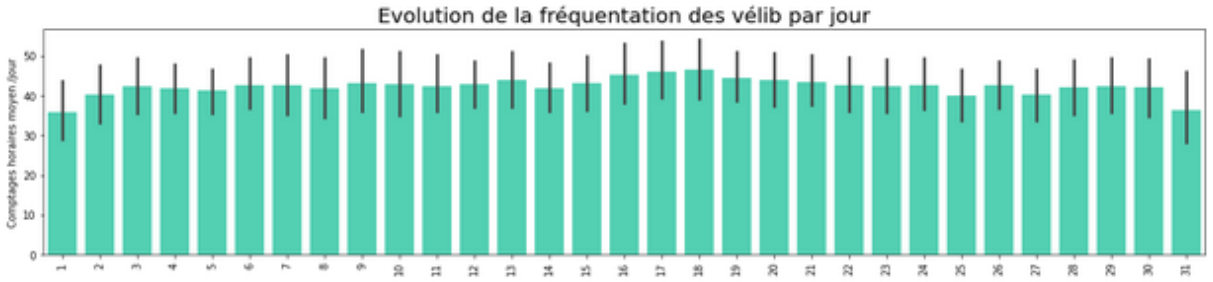


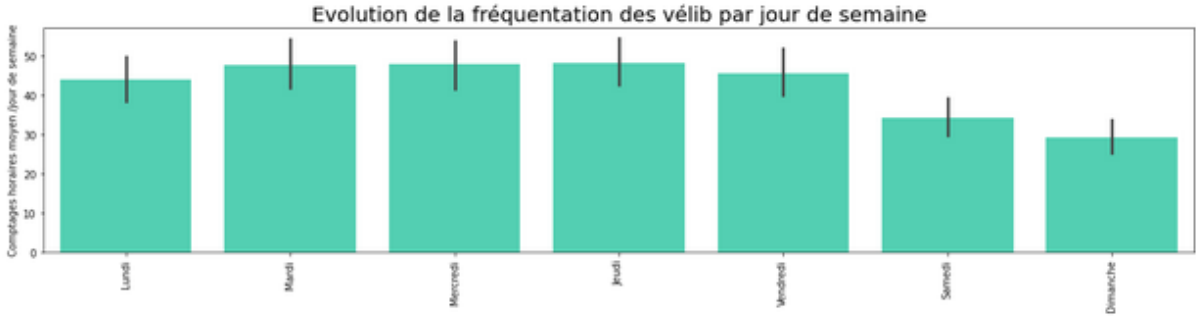
Annexe II : Distribution des valeurs de la variable cible

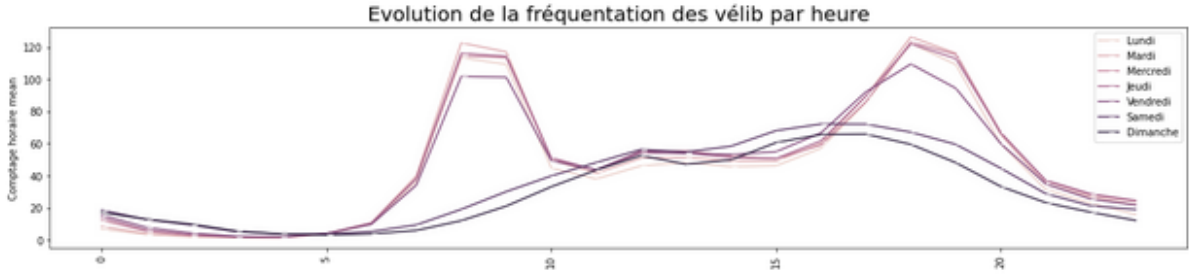






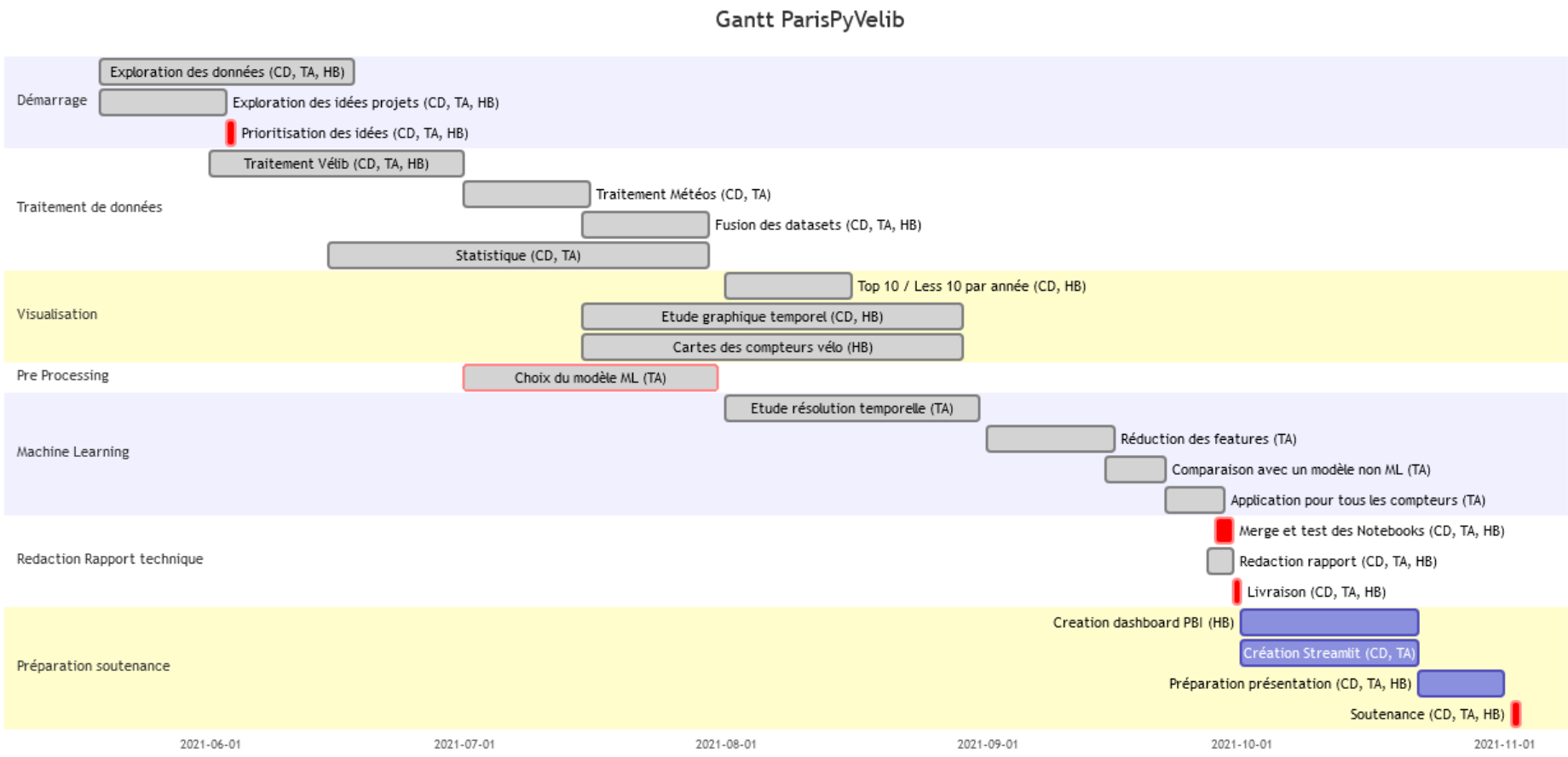

Evolution de la fréquentation des vélib par jour
Comptages horaires moyen /jour


Evolution de la fréquentation des vélib par jour de semaine
Comptages horaires moyen /jour de semaine
Lundi
Mardi
Mercredi
Jeudi
Vendredi
Samedi
Dimanche


Evolution de la fréquentation des vélib par heure
Comptage horaire mean
Lundi
Mardi
Mercredi
Jeudi
Vendredi
Samedi
Dimanche

Annexes III :

Diagramme de gantt



Description des fichiers de code