Projet de Machine Learning

Doulson Mathis

I. Introduction

A. Présentation de la base de donnée

Nous avons choisi d'effectuer notre projet de Machine Learning sur une base de données en libre accès qui nous tient à cœur depuis longtemps, il s'agit de « London bike sharing dataset ». Cette base est une concaténation des bases mensuelles publiées régulièrement sur le site anglais https://cycling.data.tfl.gov.uk/. La table que l'on a traitée (*london_merged.csv*) est constituée de toutes les données mensuelles de l'année 2015 et 2016 qui répertorient le nombre de vélos en libre services empruntés en fonction de la date et de la météo londonienne au moment de l'emprunt. La partition de la date consiste en une ligne pour chaque heure du jour considéré. On dispose donc sur 2 ans d'un total de 17414 observations (ou dates différentes), pour un total de 10 variables :

- timestamp : Date sous forme (AA-MM-JJ HH:MM:SS), il s'agit de la clé d'identification de la table
- cnt : variable qui comptabilise le nombre de nouveaux emprunts
- t1 – Température réelle en degrés Celsius
- t2 – Température ressentie en degrés Celsius
- hum – Taux d'humidité (en %)
- windspeed : Vitesse du vent en km/h
- weathercode – Catégorie météo définie comme tel :
  1 = Clear ; mostly clear but have some values with haze/fog/patches of fog/ fog in vicinity
  2 = scattered clouds / few clouds
  3 = Broken clouds
  4 = Cloudy
  7 = Rain/ light Rain shower/ Light rain
  10 = rain with thunderstorm
  26 = snowfall
  94 = Freezing Fog
- "isholiday" – Booléen qui caractérise un jour férié (1)/ Non férié (0)
- "isweekend" - Booléen qui caractérise un jour de weekend (1)/ de semaine (0)
- "season" – Catégorie de la saison : 0-Printemps ; 1-Eté; 2-Autumn; 3-Hiver.

B. Objectifs

En appliquant les techniques de machine learning à cette base de données, nous souhaiterions prédire l'évolution du nombre d'emprunts de vélos par rapport aux informations météorologiques ainsi qu'aux périodes (annuelles, mensuelles, journalières) recensées dans la base. Nous avons la prétention que nos travaux pourront aider les autorités britanniques à anticiper le nombre idéal de vélos devant être mis à disposition à Londres, et de ce fait, à les mieux répartir sur le territoire en fonction des périodes de l'année et des conditions météorologiques présentes. Nous aspirons à une meilleure accessibilité des territoires provinciaux de l'Angleterre aux vélos en libre accès.

II. Analyse de données

A. Transformations préalables des variables

Après avoir importé la base .csv sur Python, on effectue une rapide analyse du nombre de lignes, de variables et du type de variables auxquelles on a affaire. Afin de simplifier la manipulation des données, on transforme la variable *TimeStamp* (qui est sous format *object*) en une variable *datetime64*. Cela pour nous permettre de créer les trois variables *dayofweek*, *hour* et *month*. Puis on convertit les variables dont le format *float* est inutile (que des 0 en décimales) en *integer* et en type *Category*, et les variables *is_hoiday* et *is_weekend* en booléens.

B. Analyse du nombre de locations par période

1. Par mois

On commence notre analyse en affichant le plot du nombre de locations (*cnt* en ordonnées) en fonction de la date (*timestamp* en abscisses). On remarque une fréquence assez faible en hiver (décembre/janvier) et une fréquence bien plus importante pendant l'été (juillet/août) avec même des valeurs extrêmes (aberrantes ?) qui culminent à près de 8000 en juillet 2015.
En filtrant par nombre de locations, on observe que toutes les observations (8 observations) dont la variable *cnt* est supérieure à 5500 ont lieu en juillet et août 2015. En élargissant à celles supérieures à 5000 on obtient la même prépondérance des mois de juillet et août pour l'année 2016 également. Si l'on se concentre sur les observations pour lesquelles on observe moins de 50 emprunts, elles se situent en grandes majorités pendant les mois hivernaux (janvier, mars, février). Il est intéressant de noter qu'il existe une observation, datant du 29 mars à 1h du matin, qui ne comptabilise aucun emprunt de vélo.
Par agrégation du nombre d'emprunts par mois, le mois qui cumule sur les deux ans le moins d'emprunts est celui de février (1 136 504) et celui qui en cumule le plus est juillet (2 286 214). Entre ces deux mois, on observe un rapport du simple au double sur le nombre de vélos empruntés.

2. Par jour de la semaine

Sur le graphique « Nombre utilisateurs pour chaque jour de la semaine » on observe que l'effet saisonnier été/hiver se retrouve pour tous les jours de la semaine. En revanche, la fréquence est différente entre les jours. Les samedis et dimanche font figure de jours creux en terme d'emprunts, comparés aux jours de semaine. A l'inverse, le mercredi et jeudis sont les jours qui affichent le plus d'emprunts. Cette première dichotomie entre jours de semaine et jours de semaine nous laisse penser

que ces vélos sont empruntés par un certain nombre de travailleurs qui s'en servent de manière quotidienne pour se rendre sur leur lieu de travail.
Une agrégation de la variable de comptage par jour de la semaine nous montre que les mardi, mercredi et jeudi sont bel et bien les jours qui comptabilisent sur les deux ans le plus d'emprunts (3 136 956 au total pour jeudi). Le samedi et dimanche sont ceux qui en comptent le moins (2 403 716 pour le dimanche).
En prenant en compte la saison, on n'observe pas sur les graphiques « Nombre utilisateurs pour chaque jour de la semaine » d'impact particulier de la saison en fonction des jours de semaines.
Enfin, on analyse un potentiel effet de début de mois. Excepté pour le mois de décembre, on n'observe pas de pics réguliers en débuts de mois.

3. Par heure du jour

Si l'on regarde tout d'abord le nombre moyen d'emprunt en fonction des heures de la journée sur les deux ans, on remarque une très forte occurrence pour les intervalles 7h- 9h et 17h-19h. On retrouve ici l'idée d'une utilisation pour se rendre sur le lieu de travail. Dans le graphe « Utilisation moyenne par heure », on remarque que la courbe des utilisations moyennes pour les jours de semaine affiche clairement deux pics aux heures de pointes des trajets professionnels. Les jours fériés ou de weekend affichent quant à eux des courbes relativement similaires avec des densités plus fortes entre 12h et 16h (intervalle le moins investi de la journée pour les utilisations en semaine).
Les graphiques suivants nous montrent l'utilisation moyenne par jour et par heure. On remarque que la tendance aux heures de pointe est sensiblement la même pour tous les jours de la semaine (comme attendu). Les samedis et dimanches affichent des tendances similaires également.
Ces mêmes graphes journaliers appliqués pour chacune des trois saisons nous montrent que l'effet saisonnier sur l'utilisation en semaine est quasi nul. On remarque en revanche que pour les samedis et dimanches, les sorties hivernales à vélo sont bien moins fréquentes que pour les autres saisons. C'est moins le cas pour les jours de semaine. L'impératif professionnel dépasse donc l'aversion à prendre le vélo en hiver.

4. En fonction des conditions météo

Nous créons une variable *hour_cat* définit en fonction de la variable *hour* et qui prendra 4 modalités : soir/nuit ; matin ; journée ; Après-midi.

Les statistiques descriptives des variables météorologiques nous indiquent une température ressentie en moyenne d'un degré plus faible que la température réelle. Une humidité très marquée, comme c'est souvent le cas dans la capitale britannique, avec une moyenne de 72.32 %. Et une vitesse de vent en moyenne de 16 Km/h pouvant attendre 56 Km/h. On sait que la température ressentie est fonction de la température réelle, du taux d'humidité, et de la vitesse du vent. Nous vérifions cela à l'aide d'une matrice de corrélation. Si l'on se concentre seulement sur les variables continues *t1, t2, hum* et *wind_speed* on remarque une forte corrélation positive entre *t1* et *t2*, ainsi qu'une corrélation négative importante entre *t2* et *hum*. On remarque par ailleurs une très faible corrélation (positive) entre t2 et *wind_speed*.
Les quatre graphiques suivants nous montrent par ailleurs une température ressentie, qui est fonction des saisons ainsi que des heures de la journée (plus forte densité dans l'après-midi, plus faible au milieu de la nuit).

a) Température (réelle et ressentie)

Nous nous intéressons aux « Emprunts (moyens) par saison et température ». On observe que pour les basses températures, il n'y a pas d'effet de saisonnalité. En revanche, les hautes températures de printemps (18-25°) affichent de plus nombreux emprunts qu'aux mêmes températures en automne et en été. Le début des fortes chaleurs au printemps serait donc un argument pour emprunter des vélos. Les pics d'emprunts moyens sont atteints en été et en automne pour des températures supérieures à 30°.
En filtrant en fonction des différentes parties de la journée, on remarque que le printemps affiche des pics de moyenne importants dès que les températures se réchauffent. C'est notamment vrai en soirée et en journée. L'arrivée des beaux jours pourrait impliquer donc une véritable recrudescence pour l'utilisation des vélos en libre-service. Le constat est le même pour les températures ressenties.

b) Humidité

On constate une décroissance régulière du nombre d'utilisateurs en fonction du taux d'humidité. Cela se remarque pour toutes les saisons. Néanmoins on observe une décroissance plus importante pour l'été, l'automne et le printemps par rapport à l'hiver qui affiche une décroissance très faible voir une pente moyenne quasi nulle en fonction du taux d'humidité.

c) Vent

Le nombre d'utilisateur en fonction de la vitesse du vent semble afficher une décroissance à partir de 30 Km/H. Cela se vérifie pour toutes les saisons.

d) WEATHER

En moyenne :
On effectue tout d'abord une analyse moyenne sur toute la période. En moyenne, ce sont les catégories *Clear*, *Few Clouds* et *Broken Clouds* qui affichent la plus grande affluence. Néanmoins, les jours de chute de neige, qui sont un cas extrême de météo affichent, en moyenne toujours, une affluence assez élevée de plus de 300 utilisateurs par heure. C'est la catégorie *Rain with thunderstorm* qui comptabilise la plus faible moyenne d'utilisateurs.

En cumulé :
En revanche lorsqu'on regarde sur le graphique suivant la fréquence cumulée, on obtient une affluence très faible pour *Snowfall* dû à sa très faible occurrence : 14 occurrences sur les plus de 17 000 observations que compte la table. On remarque que pour *Rain with thunderstorm,* si sa fréquence d'apparition est plus forte (60), le nombre d'utilisateurs qui se rapporte à cette catégorie est très faible : 8168 utilisations sur les deux ans.

Par saison :
En printemps et en été c'est la modalité *Mostly Clear* qui affiche la plus grande affluence. Au printemps, c'est la modalité *Snowfall* qui en affiche le moins tandis qu'en été c'est *Rain with thunderstorm* (aucune chute de neige n'est recensée en été).

En automne et hiver, c'est la modalité *Few Clouds* qui affiche la plus grande affluence et dans les deux cas c'est *Rain with thunderstorm* qui affiche la plus faible affluence (avec moins de 250 utilisations pour ces deux saisons au total).
A noter que cette catégorie (*Rain with thunderstorm)* est beaucoup plus investie au printemps et en été avec près de 8000 utilisations totales pour ces deux saisons. Si l'on regarde le tableau en moyenne juste en dessous, on remarque qu'en moyenne les vélos sont plus investies pour *Rain with thunderstorm* au printemps et en été qu'en hiver et en automne. On peut donc supposer que les pluies avec tempêtes sont plus décourageantes en hiver et en automne qu'en printemps/été.

III. Prédictions

La régression Ridge

La régression ridge est basée sur le principe de régularisation des contraintes à la base de la régression linéaire classique sur une matrice d'input mal conditionnée. La contrainte de la RL est : min ($||bX - y||^2$) où ||.|| est la norme euclidienne. Si la matrice A est mal conditionnée (en général trop grande colinéarité entre les variables), on peut utiliser une régression ridge qui ajoute un terme de régularisation, d'où la nouvelle contrainte : min $||bX - y||^2 + ||k.b||$.
La contrainte quadratique qui s'ajoute à la régression linéaire permet d'ajuster la matrice X^TX en une matrice $(X^TX)^{-1} + k.I$ qui minimise la variance des estimateurs en fonction de *k* (laquelle est égale à : $\sigma^2(X^TX)^{-1} + k.I$).

Avec la régression Ridge on obtient un R^2 sensiblement similaire à celui de la Régression linéaire, que ce soit pour la base test ou train. La MAE est un peu plus faible (545.84) mais la MSE est un peu plus élevée d'où une réduction des faibles erreurs avec la régression ridge mais pas des fortes erreurs. Si l'on observe le ratio des coefficients de la régression linéaire et la régression ridge on observe que les ratios sont en général proche de 1 (le cas extrême pique à 1.06). La régression ridge n'a donc pas modifié beaucoup la matrice de variance-covariance. On peut penser que la régresion Ridge n'était donc pas d'un grand secours pour cette table. On ne répète pas les plots des erreurs qui n'apportent pas beaucoup plus d'information que ceux de la partie Régression Linéaire.

La régression Lasso

La régression Lasso est basée sur le même principe de régularisation que la régression Ridge et s'opère dans les mêmes conditions, mais avec une contrainte linéaire sur les coefficients différente, à savoir : min $||bX - y||^2 + k||b||$.
Comme l'on pouvait s'y attendre étant donnée la conclusion tirée de la régression Ridge dans la partie précédente, la régression Lasso n'apporte pas une meilleure qualité dans les prédictions.

La Random Forest

On exécute une *Random Forest* sur une forêt de 500 arbres. Après l'avoir exécutée sur 10 000, puis sur 1000 avec un R^2 et des erreurs assez similaires dans les deux cas, nous jugeons que 500 arbres est un bon compromis pour une forêt comme la nôtre. On décide également d'un minimum de 10 feuilles, comme on l'a paramétré pour l'arbre de décision.

La moyenne des résultats pour ces 500 arbres de décision nous donne finalement une MAE de 120.66 pour la base train et de 141 pour la base Test, soit des résultats un peu meilleurs pour ce qui est de la base *train* et bien meilleurs dans le cas de notre base *test*, lorsqu'on les compare à ceux du seul arbre de décision à 10 feuilles. Les R^2 de la base *train* sont les mêmes dans les deux cas, celui de la base *test* pour la forêt est un peu meilleur. Ainsi la *Random Forest* est un modèle prévisionnel plus adapté pour une table de petite taille dont la distribution des variables est sûrement peu équilibrée.
Nous définissons la fonction *rf_feat_importance* qui nous donne pour chaque variable le *score d'importance* de cette dernière, grâce à l'attribut *rf.feature_importances_*. L'importance de chaque variable est calculée par la réduction de l'index de Gini, aussi appelé indice d'impureté. Le concept de pureté (ou impureté) fait référence au caractère discriminant de la séparation effectuée par un noeud. Une séparation est dite pure quand chaque partie (après la séparation) contient des éléments d'une même classe. À l'inverse, le maximum d'impureté est atteint lorsque chaque séparation contient la même probabilité d'éléments de chaque classe. Soit un data-set contenant AABB — la segmentation la plus pure classe en deux groupes AA et BB, et la plus impure en AB et AB. Dans le cas de la random forest, il s'agit donc de trouver la segmentation qui donne des résultats le plus purs possibles.
On obtient donc que les deux variables les plus significatives sont l'*heure* et la période de la journée (*hour_cat*) avec un indice de 0.28 pour *hour* et 0.46 pour *hour_cat*. Ces deux variables sont suivies par la variable *weekofday* ainsi que la température et *is_weekend.* Parmi ces 6 variables que l'on vient de listées, 4 d'entre-elles caractérisent directement ou indirectement la ségrégation entre les trajets pour se rendre au travail ou non. En effet, on l'a vu précédemment, on a une grande différence de moyenne de la variable *cnt* pour les heures de pointe d'aller ou retour du travail. Les variables *weekofday* et *is_weekend* séparent bien quant à elles les jours où l'on se rend au travail de ceux où l'on ne s'y rend pas. En se fiant à l'indice de Gini fournit par la *random forest* on peut donc sans risque valider notre hypothèse d'une dichotomie trajet professionnel et trajet non-professionnel. Outre cette dichotomie, c'est la variable continue de température et de température ressentie (les deux sont très corrélées on l'a vu) qui va déterminer le plus les différents groupes à chaque nœud.

Comparaisons de modèles :

Si l'on observe le plot des erreurs pour la *Random Forest* maintenant, on peut voir que les points se situent beaucoup plus proches de la bissectrice que ce n'était le cas pour les modèles linéaires. On observe que le segment *best fit* se situe légèrement en dessous de la bissectrice, d'où une présence un peu plus forte d'erreurs positives, comme c'était le cas pour les modèles linéaires. On remarque par ailleurs plus de cas extrêmes en dessous de la bissectrice avec des erreurs qui culminent à 3000 unités d'écart.

On effectue juste en dessous une représentation des R^2 pour l'arbre de décision, la *random forest* et la régression linéaire (on s'épargne une représentation avec Lasso et Ridge qui nous fournissent un

R^2 sensiblement égal à celui de la régression linéaire sans régularisation), en fonction du minimum de feuilles. On remarque que pour un nombre de feuille (trop) faible la random forest nous fournit un très bon R^2 mais avec un modèle sans doute trop *over-fitté*. Pour un *min leaves* entre 10 et 20 on a toujours une meilleure goodness of fit pour la random forest, et à partir de 20 les R^2 des deux modèles se rejoignent.
Etant donné la taille de notre table et le nombre de variables, un *minl* inférieur à 20 nous semble faible. On peut donc être amené à conclure à une qualité de prédiction très proche entre un seul arbre de décision et une *random forest* dans notre cas.

RESTE XGBOOST ….