Demande de vélos

Constantini

Novembre 2016

1 Statistiques descriptives

1.1 Facteurs influençant la demande en vélos

Le jeu de données porte sur les locations de vélos sur une période. Les relevés dont nous diposons sont: la date et heure du relevé, la saison, la météo, la température, la température ressentie, l'humidité, la vitesse du vent, le nombre de locations d'usagers abonnées et non abonnés ainsi que le nombre total de locations. Les données indiquent aussi s'il s'agit d'un jour de vacances et si le jour est travaillé.

1.1.1 Matrice de corrélation sur les variables continues

Une première approche simple est d'analyser la matrice de corrélation des variables continues.

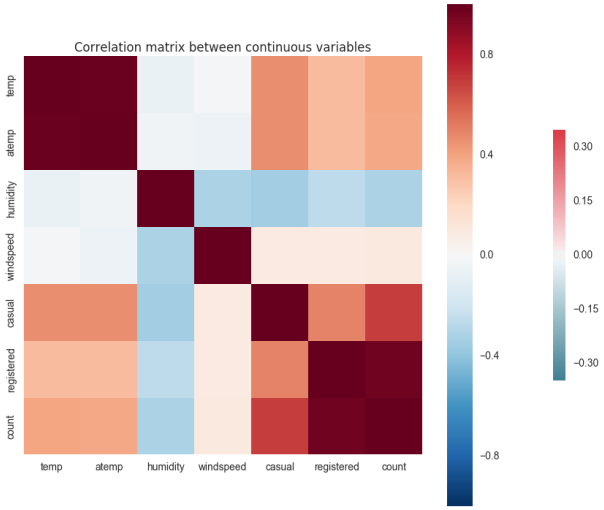

Figure 1: Matrice des corrélations

Ainsi, les températures, absolue et ressentie, sont corrélées avec le nombre de locations. Par ailleurs, les non abonnés sont plus à même de louer un vélo lorsque les températures sont meilleures par rapport aux abonnés qui, eux, sont moins affectés par le changement de température. L'humidité est inversement corrélé avec le nombre de réservations. Plus il fait humide, moins il y a de locations. La vitesse du vent semble très légèrement influencer le nombre de locations.

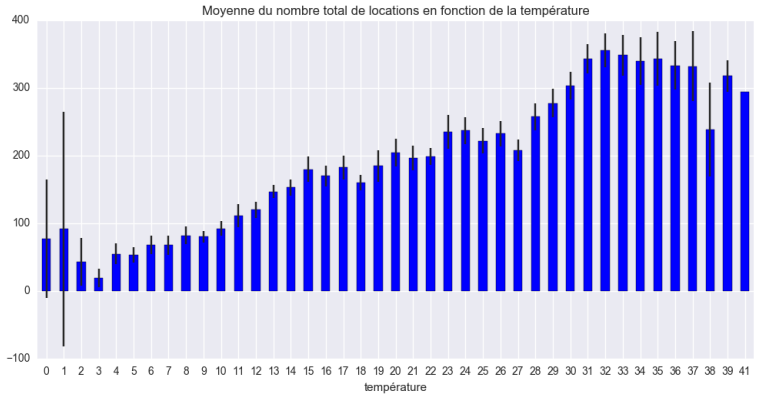

Figure 2: Nombre vélos loués en fonction de la température

Analysons plus en détail l'influence de la température. Hormis les valeurs faibles (0-3 degrès) qui présentent un fort écart-type, le nombre de locations augmentent en fonction de la température jusqu'à un palier. A partir de 30 degrès, la chaleur dissuade les personnes de louer un vélo.

1.1.2 Date et heure des réservations

Analysons ensuite les composantes discrètes du dataset. Pour cela, nous utilisons des histogrammes groupés et des boxplots.

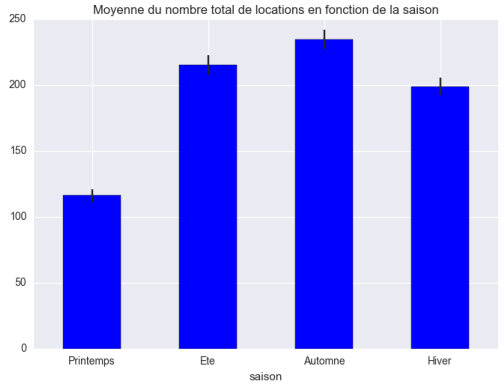

Figure 3: Graphique du nombre total de locations en fonction de la saison

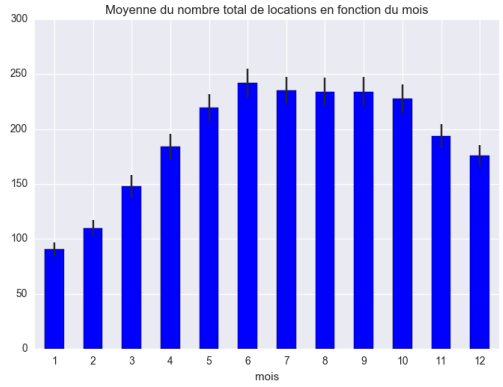

Figure 4: Graphique du nombre total de locations en fonction du mois

Ces deux graphiques montrent l'importance de la période dans l'année. Au printemps (et un peu en hiver), le nombre de locations diminue à cause des faibles températures: la demande diminue d'environ 50% en moyenne.

Les quatre graphiques suivants montrent la différence de comportement entre les abonnés et non abonnés. Les abonnés louent des vélos les jours de travail, du lundi au vendredi, tandis que les non-abonnés louent des vélos le samedi et dimanche pour profiter du week end. De plus, les abonnés effectuent des locations le matin pour aller au travail et en fin d'après-midi pour rentrer du travail tandis que la répartition des locations faites par des non-abonnés au cours de la journée s'approche d'une gaussienne.

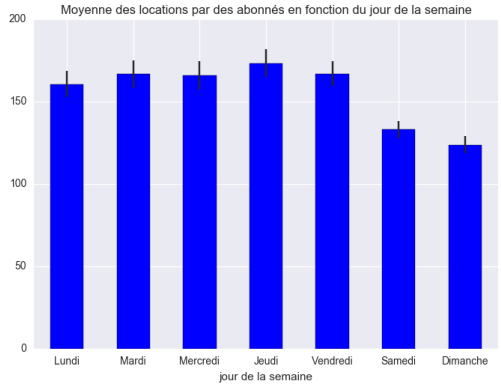

Figure 5: Graphique du nombre de locations par des abonnés en fonction du jour de la semaine

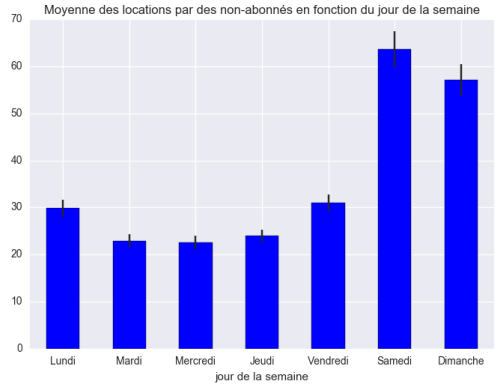

Figure 6: Graphique du nombre de locations par des non-abonnés en fonction du jour de la semaine

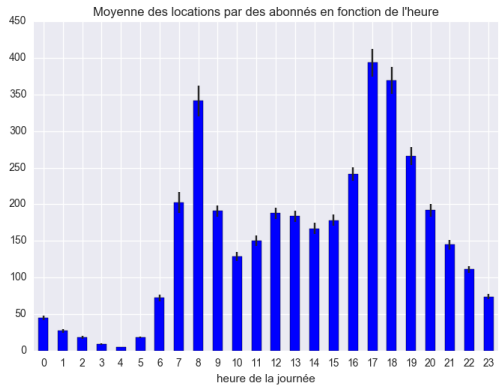

Figure 7: Graphique du nombre de locations par des abonnés en fonction de l'heure

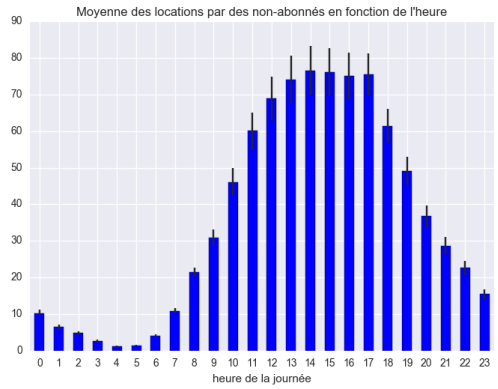

Figure 8: Graphique du nombre de locations par des non-abonnés en fonction de l'heure

1.1.3 Locations et météo

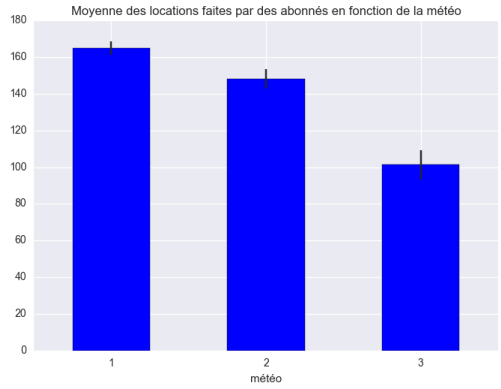

Figure 9: Moyenne des locations faites par des abonnés en fonction de la météo

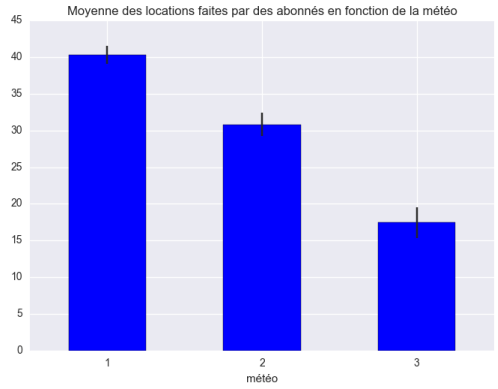

Figure 10: Moyenne des locations faites par des non-abonnés en fonction de la météo

La météo est ainsi un facteur clé influencant la demande en vélos. Plus le temps se dégrade donc plus les conditions sont mauvaises, moins la demande en vélos est importante.

1.1.4 Locations et jour de travail ou jour de vacances scolaires

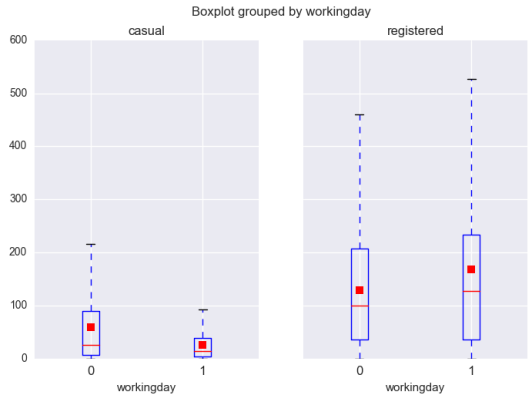

Figure 11: Nombre vélos loués en fonction que ce soit un jour de travail ou non

Ici, la différence d'utilisation est encore visible. La moyenne (carré rouge) des locations faites par des non-abonnés augmente de 50% lorsque le jour n'est pas travaillé. Cela est probablement due au fait que les personnes profitent de ce jour pour se ballader ou effectuer des courses en vélo. Nous retrouvons le résultat obtenu en étudiant les locations en fonction du jour de la semaine. Les conclusions sont similaires pour les jours de vacances scolaires.

Par ailleurs, remarquons que le nombre de vélos réservés augmente d'une année sur l'autre. Cela est probablement dû à l'augmentation du nombre de vélos en circulation entre 2011 et 2012.

1.2 Procédure statistique

Mettons en place un test statistique pour affirmer ou réfuter l'hypothèse selon laquelle les distributions en âge des deux populations (femme et homme) sont identiques. En supposant que les modèles paramètriques sont différents, nous devons utiliser un modèle non paramètrique pour des populations non appariées. Il s'agit du test de Kolmogorov-Smirnov qui permet d'affirmer qu'une distribution est comparable à une distribution semblable.

1. Notons $\underline{X} = (X_i, 1 \leq i \leq m)$, i.i.d. de fonction de répartition F continue, l'échantillon d'âges issue de la population femme. Notons $\underline{Y} = (Y_j, 1 \leq j \leq n)$, i.i.d. de fonction de répartition G continue, l'échantillon d'âges issue de la population homme.

2. Les hypothèses sont H_0 : "les distributions en âge des deux populations (femme et homme) sont identiques i.e. F=G". Et H_1 : "F≠G".

3. Alors, la statistique du test de KS est: $\zeta_{m,n} = \sqrt{\frac{mn}{m+n}} D_{X,Y}$, avec $D_{X,Y}$ la statistique de Kolmogorov-Smirnov. $D_{X,Y} = sup_{t \in R} |F_X(t) - G_Y(t)|$. Ce test est ainsi basé sur la distance entre deux fonctions de répartitions.

4. Sous H_0, lorsque min(m,n) tend vers l'infini, $\zeta_{m,n}$ converge en loi vers la loi de fonction de répartition définie. Sous H_1, $\zeta_{m,n}$ tend vers $+\infty$.

5. Ainsi, en posant $a > 0$, la région critique du test sera $[a, +\infty]$. Et le test est convergent pour $min(m, n) \to +\infty$. Le test est de niveau asymptotique α si a est égal au quantile d'ordre $1 - \alpha$ de la loi précédente. En comparant la valeur de la statistique observée au quantile d'ordre 1- α, l'hypothèse H_0 est acceptée ou réfutée.

6. Il reste ensuite à calculer la p-valeur. Si $p \geq \alpha$, alors H_0 est accepté au seuil α% et H_1 est rejetée. Sinon, l'hypothèse H_0 est rejetée.

Ainsi, en calculant la valeur de la statistique observée puis la p-valeur, il suffit de comparer la p-valeur au seuil pour affirmer ou rejeter H_0. Nous aurions aussi pu traiter le problème en utilisant le test de Mann-Whitney.

Si nous disposons du modèle paramètrique des deux populations, à priori identique et gaussien. Nous pouvons effectuer un test de comparaison de moyennes de deux échantillons gaussiens puis un test de Fisher pour comparer la variance des deux échantillons. Notons que les procédures de test restent les mêmes.

Ce test permettrait d'unifier les deux populations de manière à réduire le nombre de variables du modèle.

2 Machine Learning

La prédiction de la variable **count** est un problème de régression.

2.1 Modèles

Parmi les algorithmes populaires de regression, la regression linéaire, les machines à vecteurs de support, les arbres de décisions (Gradient Boosting ou Random Forest) peuvent être utilisés. Pour cette étude, nous entrainons chacun de ces modèles pour obtenir le modèle qui a la meilleure précision. Comme le modèle contient des variables discrètes, les algorithmes utilisant des arbres semblent à priori les plus adaptés.

Concernant l'utilisation des données, nous disposons de données qui constituent count: registered, casual. La variable count provient de la somme de ces deux variables. D'après la partie statistique, ces deux variables se comportent différemment en particulier en fonction de l'heure et du jour. D'abord, les modèles seront entrainés de manière à trouver la variable 'count'. Une fois le meilleur modèle trouvé, nous entrainerons deux mêmes modèles séparés (sur les abonnés et non-abonnés) et aditionnerons les résultats. Il s'agit d'une piste d'amélioration de l'algorithme.

2.2 Critère de performance

Par ailleurs, nous utilisons comme critère de performance l'erreur quadratique moyenne. Sur scikit-learn, la méthode **score** donne le coefficient de corrélation. C'est celui-ci que nous utiliserons.

Nous étudierons ce coefficient sur les scores des données d'entrainement (70%) et de test (30%). Pour mieux appréhender la généralisation des modèles, nous calculerons aussi le score de cross validation. En découpant en 5 parties les données, le modèle est entrainé successivement sur 4 parties, puis testé sur la dernière. Ainsi, l'erreur due à la variance est mieux appréhendée.

2.3 Choix des paramètres

Le choix des paramètres pour chaque modèle s'effectue grâce à la fonction GridSearchCV. Elle permet d'entrainer l'algorithme sur une grille de paramètres et de sélectionner ceux qui performent le mieux sur les données test. La fonction GridSearchCV prenant un temps de calcul considérable, les paramètres sont trouvés en les modifiant successivement et en selectionant les meilleurs.

Notons que l'algorithme de Gradient Boosting est robuste. Un fort nombre d'estimateurs donnera de meilleures performances sans overfitting.

2.4 Résultats

Modèle	Train score	Test score	Cross validation
Regression linéaire	0.39	0.39	0.28
Kernel ridge	0.99	0.45	0.005
SVR	0.99	0.60	0.12
Gradient Boosting	0.97	0.93	0.72
Random Forest	0.99	0.94	0.71
2-M Random Forest	0.99	0.94	n

Le kernel ridge et le regresseur à vecteurs de support (à kernel gaussien) ont un mauvais score de cross validation. Ces modèles ne se généralisent pas aux nouvelles données et overfittent.

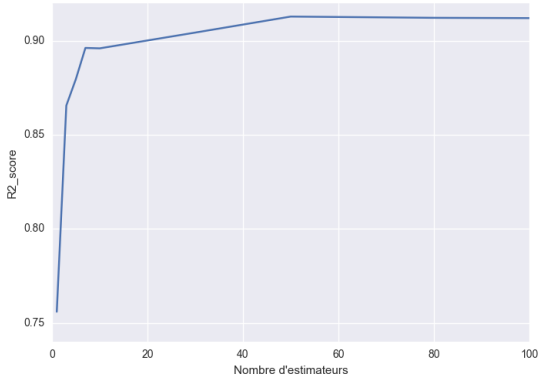

Figure 12: Coefficient de corrélation en fonction du nombre d'estimateurs

Les modèles de type arbre sont les plus performants car prennent en compte les variables discrètes et réduisent de manière significative la variance. Nous constatons qu'un nombre d'estimateurs de 40 convient pour obtenir la meilleure performance. L'algorithme de Random Forest sera donc choisi.

Nous constatons aussi que l'entrainement séparé (2-M modèle) sur les données des abonnés et des non-abonnés n'améliore pas les performances. Cela est probablement dû au fait que les algorithmes de Random Forest ont un biais faible. La complexité de ce modèle permet la modélisation intrinsèque des deux populations.

2.5 Amélioration du modèle

2.5.1 Amélioration essayée: deux populations

Nous avons essayé, dans le modèle 2-M Random Forest, d'entrainer deux modèles: l'un sur les abonnés, l'autre sur les non-abonnés. Les performances du meilleur modèles ne se sont pas améliorés. Si l'on avait choisi un modèle avec plus de biais, une amélioration aurait été visible.

2.5.2 Créer des dummy variables pour les variables discrètes

Pour augmenter la précision de la prédiction, il est commun de créer des "dummy variables". Ces variables sont des booléens des valeurs que peuvent prendre les variables discrètes. Par exemple, pour les saisons, quatre nouvelles variables booléennes seraient ajoutées: Ete, Automne, Hiver, Printemps.

2.5.3 Modéliser l'évolution d'une année sur l'autre

En modélisant l'augmentation du nombre de stations dans la ville, il devient possible alors de le prendre en compte dans l'algorithme. Cette solution parait plus appropriée que la variable discrète.

2.5.4 Utiliser les données des transports en commun

Si les transports en commun subissent une grève ou si quelques stations sont en rénovation ou momentanément indisponible, les personnes auront tendance à louer des vélos comme alternative de transport.