Analyse de données de location de vélos partagés - ancien concours Kaggle Bike Sharing Demand

Morgane Flauder

Partie I - Statistiques descriptives

Nous nous intéressons à un jeu de données de location de vélos partagés. Ces données ont été collectées en deux ans, et concernent la première moitié de chaque mois (du 1er au 19). Nous avons à notre disposition des variables temporelles (date et heure, jour travaillé ou non...) et météorologiques (température, force du vent, humidité...), ou les deux (saison). Nous cherchons à prédire le nombre de locations de vélos à partir de ces variables.

Valeurs manquantes

Nous constatons tout d'abord que ce jeu de données est incomplet : il ne concerne que les 19 premiers jours de chaque mois. Il s'agit d'une caractéristique particulière de ces données, nous pouvons donc les analyser sans s'en préoccuper.

Ces donnéees présentent un autre type de valeur manquante. En effet, nous pouvons remarquer que la valeur minimale de nombre de location de vélo est 1. Les heures correspondant à "aucune location" ne sont tout simplement pas présentes. On peut rapidement calculer le nombre d'entrées manquantes en considérant qu'on a ici 2 années de 12 mois chacunes, chaque mois faisant 19 jours de 24h : $2 \times 12 \times 19 \times 24 = 10944$. Or le jeu de données contient 10886 lignes. On conclut que 58 entrées sont manquantes, ce qui correspond à 0.53% du jeu de données s'il était complet. La proportions de valeurs manquantes étant très faible, j'ai fait le choix de les ignorer.

Dans la suite de cette analyse, nous allons travailler sur un sous-ensemble aléatoire de 60% de ces données, qui représentera l'échantillon d'apprentissage. Les 40% restant sont séparés aléatoirement et de manière égale en un échantillon de validation et un échantillon de test. Nous les utiliserons par la suite, lors de la construction du modèle prédictif.

Variables temporelles

Tendances annuelles et saisonnières

Nous allons tout d'abord observer le nombre total de locations de vélos par jour sur l'ensemble de ces deux années (figure 1).

Nous constatons deux tendances : l'augmentation du nombre de locations de vélos en 2012 par rapport à 2011, ajouté à une périodicité saisonnière. En effet, les locations de vélos semblent augmenter durant les mois les plus chaud (de mai à septembre), et sont à leur minimum en hiver (de décembre à février).

Nous allons à présent vérifier statistiquement ces deux hypothèses. La variable d'intérêt, `count`, étant une valeur de comptage (discrète, toujours positive, et avec une forte fréquence de valeurs faibles), elle ne suit pas une loi normale. Nous allons donc utiliser des tests non-paramétriques. Pour comparer les moyennes du nombre de locations en 2011 par rapport à 2012, j'ai tout d'abord créé une nouvelle variable appelée `year`. J'ai ensuite choisi d'utiliser un test de Mann-Whitney (dans le cas d'un jeu de données complet, un test de Wilcoxon aurait été plus indiqué car les deux échantillons peuvent être considérés appariés). Nous obtenons une p-value de 2.2e-16 significative à $\alpha = 0.05$, nous pouvons donc rejeter H_0 (égalité des moyennes), et nous pouvons conclure qu'il existe une différence significative entre le nombre de locations en 2011 et en 2012.

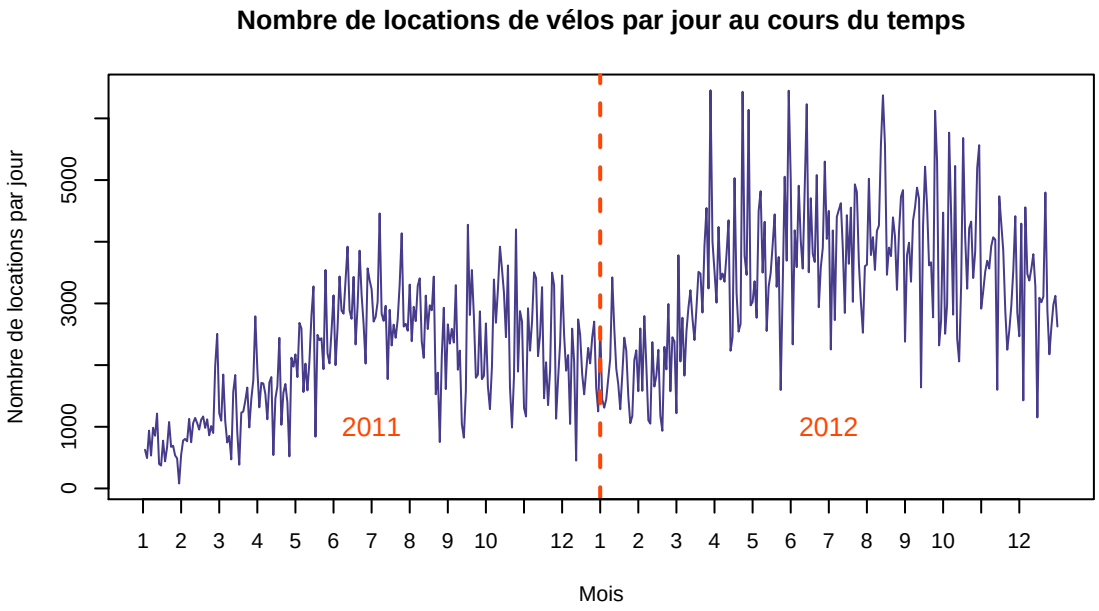

Figure 1: Mise en évidence de l'augmentation du nombre de locations de vélos entre 2011 et 2012 et de la périodicité saisonnière

De la même manière, nous pouvons vérifier qu'il existe une différence selon les saisons. Pour cela, j'ai choisi d'utiliser un test de Kruskal-Wallis. La p-value (2.2e-16) est également significative à $\alpha = 0.05$, nous pouvons donc conclure qu'il existe une différence selon les saisons. Ce test ne permet pas de déterminer plus précisément les saisons présentant une différence, mais ce n'est pas important ici, notre but étant de confirmer ou non l'importance de cette variable.

Nous pouvons finalement conclure que la variable `season` ainsi que la nouvelle variable `year` ont une influence sur le nombre de locations. Nous allons pouvoir utiliser ces variables dans notre modèle prédictif : `season` pour représenter la périodicité saisonnière, et `year` pour représenter l'augmentation annuelle constatée.

Tendances journalières et hebdomadaires

Afin de mieux comprendre les différentes périodicités de ces données, nous allons nous intéresser à l'influence de deux variables combinées : l'heure (variable `time`, que nous avons créée) et le jour de la semaine, plus particulièrement, s'il s'agit d'un jour travaillé ou non (variable `workingday`) (Figure 2).

Nous pouvons constater que le nombre de locations de vélos semble être fortement influencé par l'heure. Il est très faible durant la nuit et augmente durant la journée. Nous pouvons également remarquer une différence de tendance selon le jour de la semaine : les pics de fréquentation semblent être différents en semaine par rapport au week-end.

Il semble effectivement que le nombre de locations est réparti différemment au cours de la journée selon si il s'agit d'un jour travaillé ou non. En effet, on remarque deux pics de fréquentation à 8h et 17-18h lors des jours travaillés, ce qui pourrait correspondre aux usagers louant un vélo pour se rendre sur leur lieu de travail et revenir. Nous pouvons également noter une légère augmentation de fréquentation lors de la pause déjeuner, entre midi et 13h. En revanche, les jours de week-end ou de vacances, les usagers louent des vélos tout au long de la journée, principalement entre 10h et 18h, ce qui pourrait correspondre à une utilisation de type loisir.

La variable `workingday` semble donc être une variable importante pour prédire le nombre de locations de vélos, quand elle est associée à l'heure. Nous allons donc ajouter une nouvelle variable, `time`.

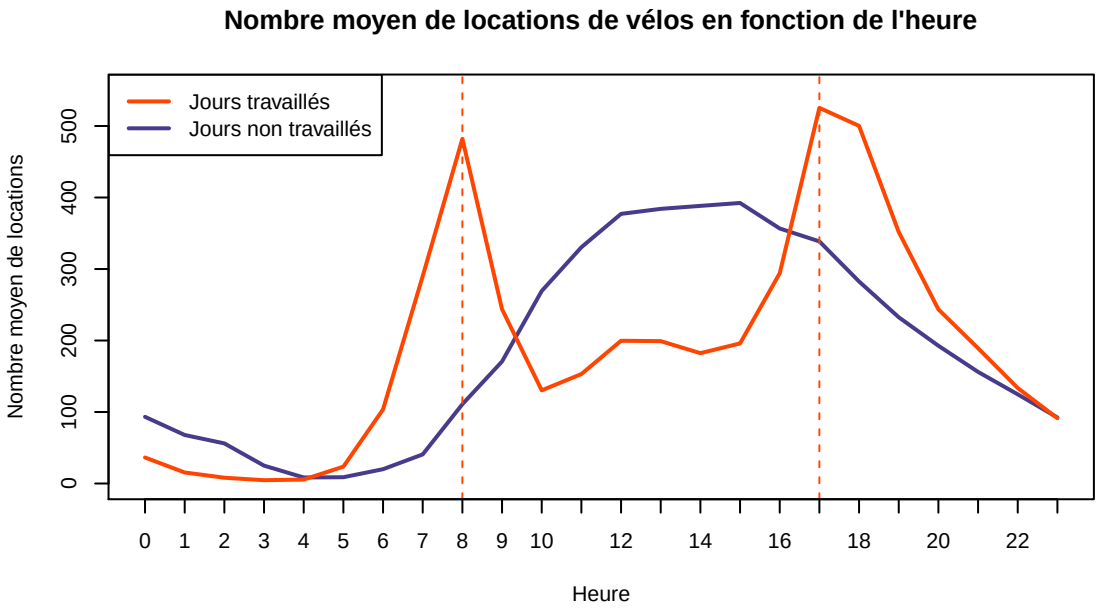

Figure 2: Mise en évidence des différences de fréquentation entre les jours travaillés et non travaillés

Variables météorologiques

Nous pouvons remarquer que la température et le taux d'humidité semblent fortement corrélés avec le nombre de locations (figure 3).

On remarque que le nombre de locations de vélos semble augmenter linéairement avec la température, jusqu'à atteindre une valeur maximale à environ 36°C. Dans les cas plus rares où la chaleur est plus forte, les vélos sont délaissés par les usagers. On constate une tendance similaire avec l'humidité : 20% d'humidité semble être la valeur idéale. La fréquentation diminue quand l'humidité augmente, et dans le cas de temps très secs, la fréquentation diminue également. On peut supposer qu'un temps très sec est également très chaud, d'où la baisse de fréquentation.

Partie II - Machine Learning

Nous allons maintenant construire un modèle prédictif à partir de ces différentes variables.

Critère de performance

Nous cherchons à prédire la variable `count`, qui représente le nombre de location de vélos en une heure. Bien que discrète et toujours positive, elle prend des valeurs très élevées et variées. Nous pouvons donc la considérer comme une variable continue.

J'ai choisi comme critère de performance le RMSE (Root Mean Square Error), qui est la racine carrée de la moyenne du carré des erreurs de prédiction. C'est une des métrique les plus utilisées dans le cas de données continues. Il pénalise de la même manière les prédictions trop élevées ou trop basses par rapport à leurs valeurs réelles. Il pénalise très fortement les erreurs de prédiction de grande amplitude, ce qui oriente le modèle à faire de nombreuses petites erreurs plutôt que peu d'erreurs importantes. C'est important dans le cadre de la location de vélos car on cherche à avoir un ordre de grandeur du nombre de locations. En revanche, une erreur de grande amplitude serait beaucoup plus grave : cela voudrait dire beaucoup sous-estimer ou sur-estimer la fréquentation à une heure donnée, ce qui amène soit à des usagers mécontents car n'ayant pas

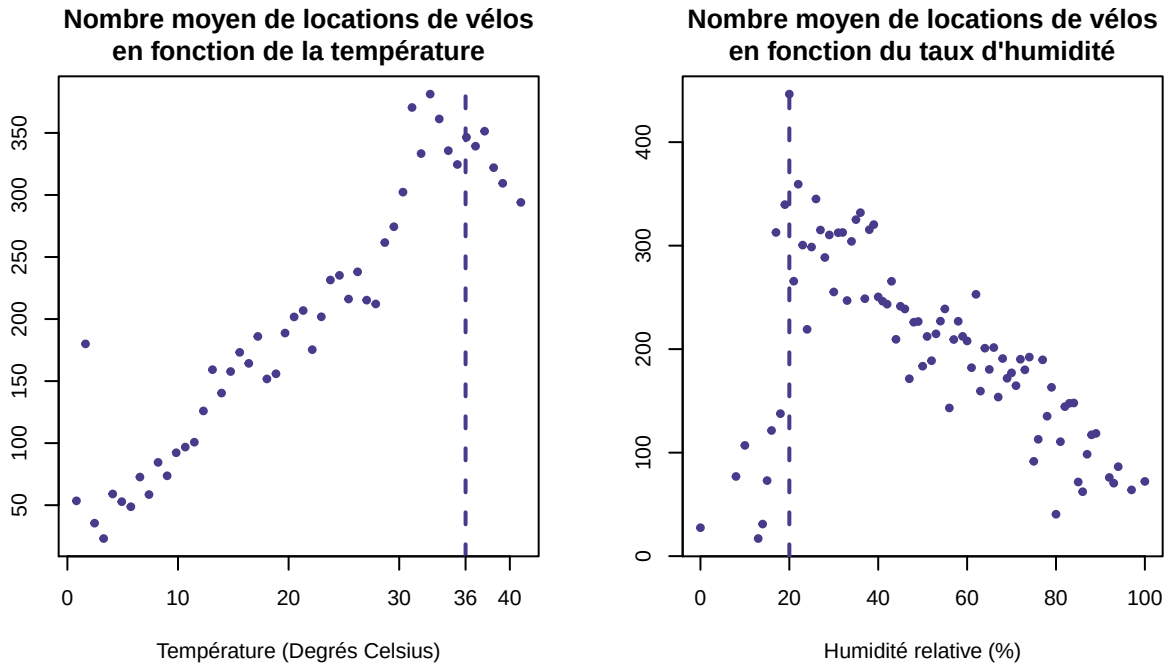

Figure 3: Influence de l'humidité et de la température sur la location de vélos

de vélo disponible, soit à un investissement de matériel non utilisé, donc dans les deux cas, une perte d'argent importante.

Construction du modèle prédictif

Random Forest

La première partie de cet exercice a été consacrée à l'analyse exploratoire du jeu de données. Elle a permis de sélectionner les variables qui semblent les plus informatives afin de prédire le nombre de locations de vélos à une heure donnée, et également de créer deux nouvelles variables.

J'ai choisi de créer et comparer trois modèles identiques, mais utilisant différentes variables prédictives. J'ai choisi d'utiliser l'algorithme des Random Forests appliqué à la régression (Regression Forests), connu pour ses bonnes performances prédictives. J'ai utilisé pour cela les paramètres par défaut de la fonction `randomForest` de la librairie `randomForest` de R. Pour comparer ces modèles, j'ai calculé le RMSE à partir de l'échantillon de validation précédemment créé (Figure 4).

	Modèle 1	**Modèle 2**	**Modèle 3**
Variables	season, workingday, temp, humidity	season, workingday, temp, humidity	season, workingday, temp, humidity
	atemp, holiday, weather, windspeed	atemp, holiday, weather, windspeed	
		time, year	time, year
RMSE (validation)	~ 145	~ 57	~ 57

Figure 4: Comparaison entre trois modèles de Random Forests utilisant différents variables prédictives

Le modèle 1 correspond à un modèle naïf n'utilisant que les variables disponibles à l'origine dans le jeu de données. Le modèle 2 correspond à un modèle plus complet avec deux variables de plus : `year` et `time`. Nous pouvons constater que l'ajout de ces deux variables améliore grandement le modèle, le RMSE de l'échantillon de validation étant plus de 2 fois plus petit. Pour créer le troisième modèle, j'ai choisi de ne garder que les variables les plus informatives, analysées dans la partie I. Nous remarquons un RMSE sur l'échantillon de

validation similaire entre ce modèle et le modèle 2. Retirer ces quatre variables ne résulte donc pas en une perte de performance du modèle, nous pouvons donc conclure qu'elles ne sont pas indispensables pour prédire le nombre de locations de vélos par heure.

SVR

Les SVM (Support Vector Machine) sont également des modèles très puissant dans le cadre de la prédiction. Habituellement utilisés dans des cas de classifications, ils sont également utilisés pour prédire des valeurs continues, et sont alors appelé SVR (Support Vector Regression).

J'ai choisi de construire un modèle utilisant les même variables que le modèle 3 de la partie précédente. Pour cela, j'ai utlisé la fonction `svm` de la librairie `e1071` de R, avec ses paramètres par défaut. J'ai obtenu un RMSE sur l'échantillon de validation d'environ 83, ce qui est clairement moins bon que le modèle de Random Forest, qui avait un RMSE de 57. J'ai ensuite cherché à ajuster les paramètres afin de créer un modèle plus performant (Figure 5).

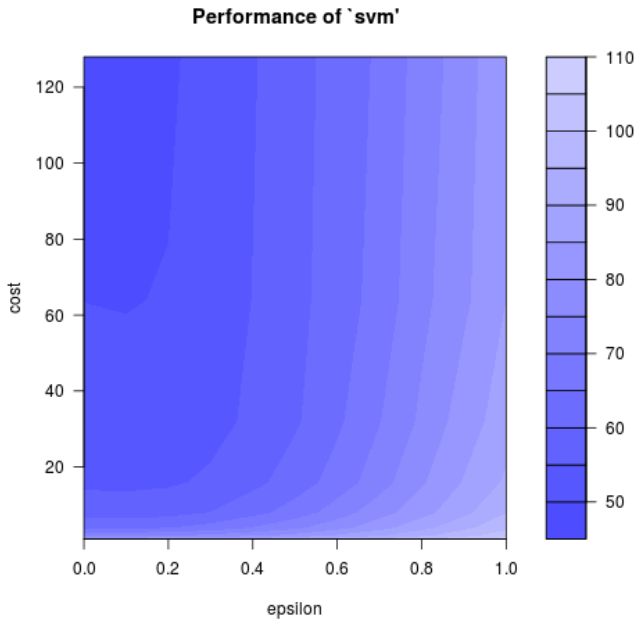

Figure 5: Ajustement des paramètres d'un modèle de SVR en fonction du RMSE de l'échantillon de validation

J'ai choisi d'optimiser les paramètres `cost`, qui représente la pénalité attribuée aux erreurs de prédiction, et `epsilon`, qui représente la marge de tolérance des prédictions. Par défaut, ces valeurs sont égales à 1 pour `cost`, et 0.1 pour `epsilon`. Après avoir testé différents couples de paramètres, leurs valeurs optimales sont déterminées : 128 pour `cost` et 0.1 pour `epsilon`, qui garde donc sa valeur par défaut. Ces valeurs correspondent à un RMSE de l'échantillon de validation d'environ 49, ce qui est non seulement plus performant que le modèle ayant les paramètres par défaut, mais également plus performant que le modèle de Random Forest créé dans la partie précédente.

Le modèle conservé comme étant le meilleur est donc ce dernier modèle de SVR.

Estimation de la performance réelle du modèle

Nous allons à présent estimer la performance réelle de ce dernier modèle avec l'échantillon de test créé précédemment. Avec l'échantillon d'apprentissage, le RMSE obtenu est d'environ 47. Avec les échantillons

de validation mais également de test, le RMSE est de 49. Ce modèle génère donc autant d'erreurs sur les différents échantillons, le risque de sur-apprentissage est donc faible. Nous constatons également que les erreurs sont de faible amplitude, et le modèle sous-estime et sur-estime le nombre de locations de vélos de manière équivalente (Figure 6).

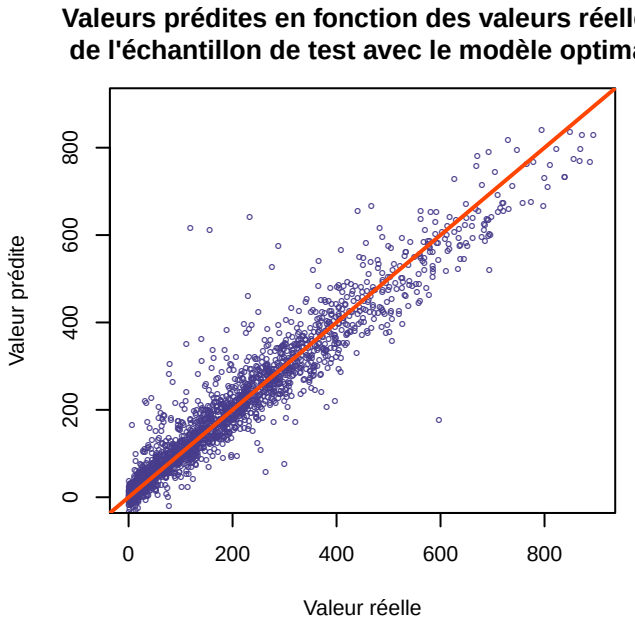

Figure 6: Illustration de la performance du modèle sur l'échantillon de test. La ligne rouge représente une prédiction parfaite

Nous pouvons noter un léger défaut de ce modèle : il prédit des valeurs négatives, ce qui n'a pas de sens ici. Il suffit donc d'attribuer les valeurs négatives à zéro.

Pistes d'amélioration

Ce modèle, bien que satisfaisant par rapport aux autres modèles étudiés, peut encore être largement amélioré.

- Le RMSE pénalise de la même manière les prédictions trop grandes et trop petites par rapport à la valeur réelle. Dans le contexte de la location de vélos, on pourrait imaginer donner la priorité à la satisfaction de l'usager, et donc réduire au maximum les pénuries de vélos. Dans ce cas, il faudrait utiliser un autre critère de performance qui pénaliserait plus les sous-estimations par rapport aux sur-estimations.
- Ces deux modèles peuvent être combinés par une méthode dite de *stacking*. Les prédictions des deux modèles seraient utilisées comme entrée d'un troisième modèle, qui prédirait le résultat final.
- Le jeu de données original nous donne une informations supplémentaire : le nombre de locations par des abonnés et des non-abonnés, représentés par les variables `casual` et `registered`. Il serait donc possible de construire deux modèles : un modèle prédisant le nombre de locations par les abonnés, et un deuxième modèle prédisant le nombre de locations des non-abonnés. Il suffirait d'additionner ces deux prédictions pour obtenir le nombre total. Cette méthode se base sur le fait que les abonnés et les non-abonnés ont un comportement différent. En effet, les abonnés utilisent principalement les vélos en semaine, pour se rendre et revenir de leur lieu de travail, alors que les non-abonnés louent des vélos plutôt le week-end en tant que loisir.