GLO-4027 - Bike Sharing Demand

Planification

30 janvier 2019

Étudiants du 2ème cycle en programme d'échange:
*Bastien **CHABAL***
*Corentin **GIRAUD***

Projet Kaggle – Repository GitHub

Introduction

Le projet que nous avons choisi est intitulé **Bike Sharing Demand** (plateforme Kaggle). On se propose d'analyser les données du système de partage de vélos de la ville de Washington, D.C.

Le système de partage de vélos au sein d'une ville est très simple : plusieurs kiosques sont répartis dans la ville et des personnes (abonnées ou non) peuvent louer un vélo et faire le trajet qu'elles souhaitent jusqu'à un autre kiosque.

Les données

Les données sont réparties en deux fichiers *.csv* distincts:

- Le **set d'entraînement** contient deux ans de données (année 2011 et 2012), où sont relevées toutes les heures différentes informations, dont le nombre total de locations. Ce set ne contient que les données des **19 premiers jours de chaque mois**.
- Le **set de test** contient des données exactement similaires au set d'entraînement, mais pour **tous les jours après le 20 du mois inclus**.

Après une première visualisation des données proposées par le set d'entraînement, on constate qu'il contient 10 886 entrées, réparties sur **12 attributs**.

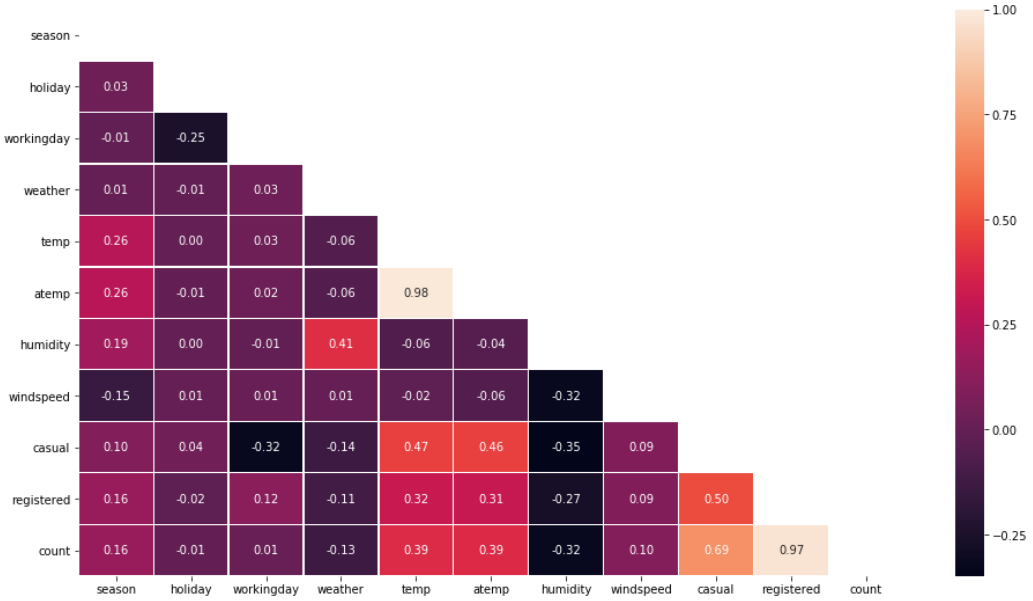
L'objectif

L'objectif ici est **de prédire le nombre de total de vélos loués à chaque heure** pour les périodes couvertes par le set de test, en utilisant seulement les informations du set d'entraînement.

La prédiction sera évaluée automatiquement par Kaggle grâce au **Root Mean Squared Logarithmic Error** (RMSLE) qui mesure le ratio entre les valeurs prédites et réelles. Nous avons créé un script permettant de soumettre automatiquement nos résultats sur la plateforme pour évaluation.

Première approche

Le nombre d'attributs est conséquent (12 attributs), surtout que notre prédiction ne devra se faire que sur le nombre de locations. La première analyse à faire serait donc de classer ces attributs par importance de corrélation avec la donnée finale (attribut *count*). On génère donc la **matrice de corrélation** pour mieux comprendre les relations qui les relient. On s'intéresse pour l'instant principalement à la ligne « count » pour voir quels attributs corrèlent avec le nombre de location.

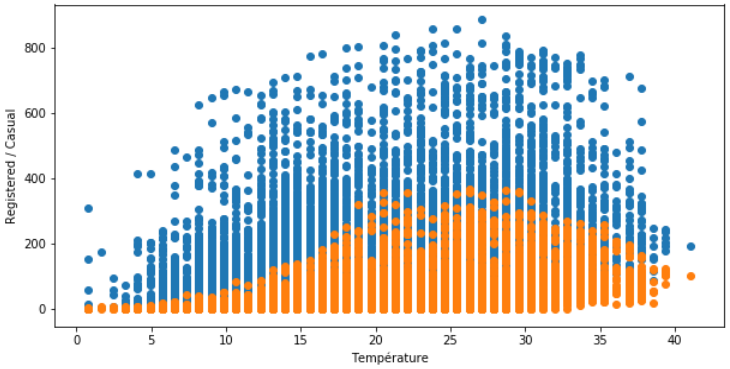


On constate d'abord que les attributs qui corrèlent le plus avec le *bike count* sont des attributs sous-jacents à ce dernier : les locations **occasionnels** (attribut *casual* avec 0.69) et **enregistrés** (attribut *registered* avec 0.97), dont la somme donne le nombre de locations total.

Avec une corrélation de 0.97, l'attribut *registered* va se comporter de la même manière que l'attribut *bike count*. Ainsi, au lieu d'essayer de prédire directement le *bike count*, une première approche intéressante serait d'abord d'analyser les comportements de *registered* et *casual* séparément. Si ceux-ci s'avèrent différents, on découpe notre problème de base en deux sous-problèmes, amenant à deux prédictions plus précises dont on fera la somme.

Observations

On repère déjà une différence entre *registered* et *casual* sur leur corrélation avec *workingday* (négative sur casual). De plus, les deux sont corrélés sur les températures (*temp*) et températures ressenties (*atemp*): on émet donc la conjecture que les journées chaudes sont propices à la location. À l'inverse, les deux sont corrélées négativement avec l'humidité.



Enfin, il faudra s'intéresser l'évolution du nombre de location au sein d'une journée, pour voir si certaines heures sont plus propices à la location.