Python Assignment- Group E

Utkarsh Khemka
Cheer Hung
Amanda Marques Pereira
António Carvalho
Maximiliano Franco
Christoph Stehling
Esther Chaelin Lee

Python Assignment

Bike sharing programs are becoming more and more popular around the world due to environmental issues, pricing and convenience. In an attempt by the city or government to control and understand mobility flows in the city of Washington, bicycle sharing system data can be used as an approximation to explain the commuting of people around the city. Therefore, the main goal of the assignment was to predict the hourly number of bikes rented during the last quarter of 2012.

For this purpose, two datasets with 17 variables were provided with data between the first of January of 2011 until the last day of 2012. One dataset had hourly information while the other one contained daily information regarding the bike sharing system business.

The group decided to take an inclusive approach by allowing each member to contribute with different ideas. In terms of steps, the following ones were crucial phases:

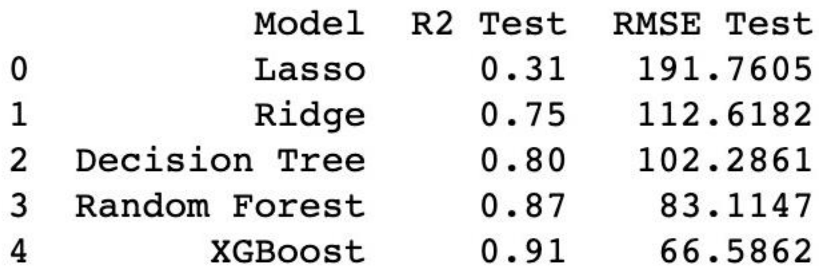
- Understand the data from each dataset
- Explore the data, assuring its quality, correctness and consistency
- Visualize the data to extract relevant insights from it, allowing to a better understanding of the datasets
- Forecast the relevant variables for the final prediction of the number of bikes
- Manipulate the dataset to enable its utilization for the machine learning stage
- Perform feature engineering to improve the quality of the data, anticipating a higher accuracy on the final prediction
- Run and tune the appropriate machine learning models

A step by step strategy was followed by the group where sub-groups worked on specific tasks. At the same time, validation of peer´s work and regular meetings with opened discussions enriched the final output by incorporating each other´s ideas. This was especially important in the EDA stage where visualization and data understanding are subject to people´s interpretations. Members with a higher technical knowledge were able to correctly share their knowledge and contribute to the group´s learning experience.

In terms of a pipeline we followed the below steps in that exact order:

1. Developing an understanding for the data
2. Data cleaning
3. Visualizing important variables and their relationship
4. Feature creating
5. Model selection and training
6. Consolidation and evaluation of results

The results obtained are the following:



```
            Model  R2 Test  RMSE Test
0           Lasso     0.31   191.7605
1           Ridge     0.75   112.6182
2   Decision Tree     0.80   102.2861
3   Random Forest     0.87    83.1147
4         XGBoost     0.91    66.5862
```

As it can be seen the best R-Squared we achieved is very high and the RMSE comes close to the top predictions that could be found online which we attribute to the steps we have undertaken in the feature engineering section. However, there are improvement opportunities which can be tackled in future researches. The recommendations for futures researches can split in broadly 3 categories:

1. Deploying other techniques to predict bike demand.
2. Enriching the data and clarify anomalies
3. Employ more feature engineering techniques

1. Deploying other techniques to predict bike demand

In addition to the models we trained there, are several other models suited to predict the demand of bikes in Washington with the data at hand. Potential candidates would be Neural Networks, Support Vector Machines with various, differing Kernels or linear combination models. To increase the robustness of the results obtained, the models can be stacked, meaning the models themselves would be treated as variables and we train a model on top to predict bike demand.

Another approach would be to treat this problem as a time series problem and fit an ARIMA model or a linear regression of which the error terms would be predicted with a time series model.

As we noticed that the demand on working days and non-working days and holiday differs greatly, it might make sense to train two models, one for working days and one for non-working days. Following the same logic four models could be trained, one for each season or on for daytime and one for nighttime.

2. Enriching the data and clarify anomalies

It could further be investigated which other variables proof to be helpful in predicting the demand of bikes and either enriching the dataset with these variables or starting to gather those variables in the future. Moreover, it is recommended that future researchers get in contact with the owners of these data to clarify the origin (e.g. downtime of the system, measuring error, etc.) of the anomalies of the data in order to make more informed decisions how to handle those outliers.

3. Employ more feature engineering and validation techniques

Future researchers could also opt to employ more advanced feature engineering technique such as PCA, LDA or QDA. In terms of validation, we advise future researchers also to employ rolling cross-validation considering time frames.