Washington D.C. Capital Bike Sharing Usage Prediction Model

Introduction

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return has become automatic. Through these systems, the user can easily rent a bike from a particular position and return to another position. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousand bicycles.

Due to the increasing number of sharing bike usage, there is a growing demand to predict the number of bike rental. Therefore, our project aims to examine and find the best prediction model to optimize bike sharing supplement.

However, the bike sharing company or organization have too many inputs to manage such as time, working day or holiday, month, weather condition, and so on so that they don't have a clear idea on what features they should focus.
Therefore, we start our analysis by defining clear research questions to solve:

- What are the key factors influencing the target variable(Count of bike rental)?
- Use different machine learning model, which model shows the best performance under the given circumstance?
- Utilizing feature importance and parameter tuning techniques will increase the model's performance? And how much?
- What actionable insights can we offer to the bike sharing company/organization?

As an analysis task overview, we perform descriptive analysis to have basic summaries of our dataset and graphic analysis to form the basis to virtually analyze the quantitate aspects of data and using three different tree-based machine learning models to find out the key factors influencing the target variable and, also compare their best performance respectively.

Descriptive Analysis and Data Pre-processing

The dataset found at 'Kaggle' which is one of the most famous online platforms for the data science challenges and it contains hourly counts of rental bikes between years 2011 and 2012 from Capital bike sharing system in Washington D.C.
More specifically, it contains 17 variables which are mostly related to season, timestamp, weather condition, holiday, working day and the total rental bike counts per hour with 17397 observations.

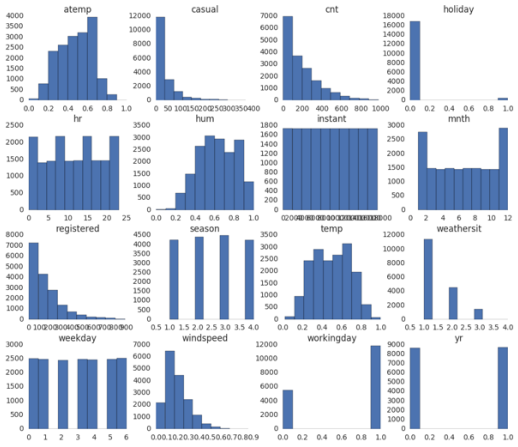


Figure 1: The distribution of numeric variables

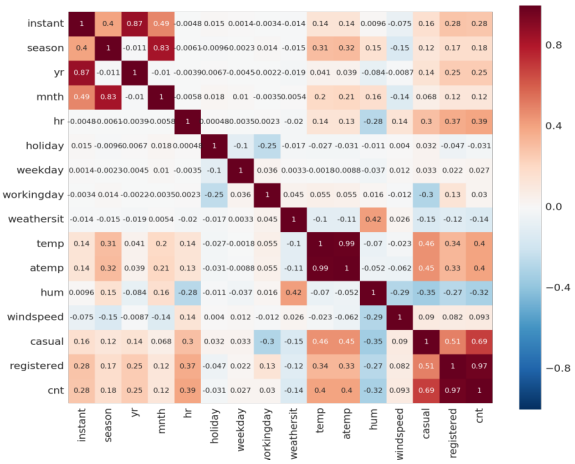


Figure 2: The heat map of correlation

As an Exploratory Data Analysis, we perform initial descriptive analysis to see the numeric variables distribution and correlation corresponding to our target variable(count) by displaying heat map. There are several features like casual, registered, temperature and hour shows relatively high linear relation toward the count.

Next, we plot the season-wise hourly distribution of counts, Weekday wise hourly distribution of counts and Monthly distribution of counts to have more clear abstract.

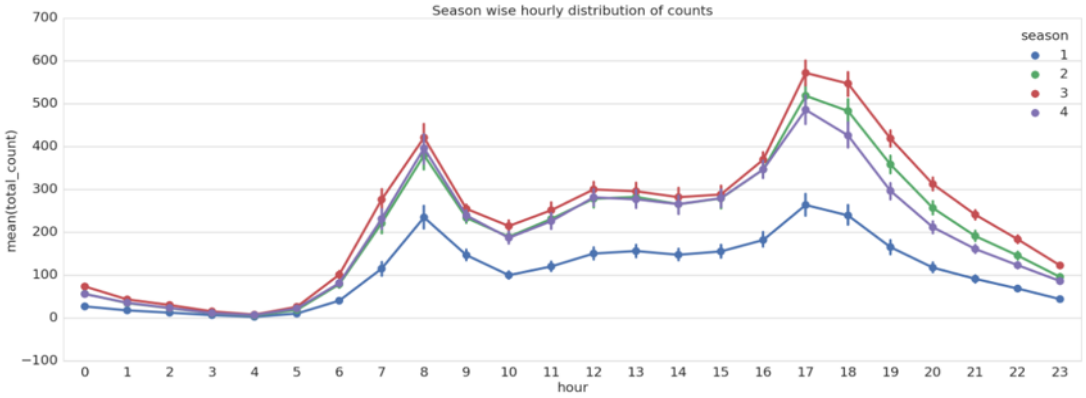


Figure 3: Season wise hourly distribution of counts

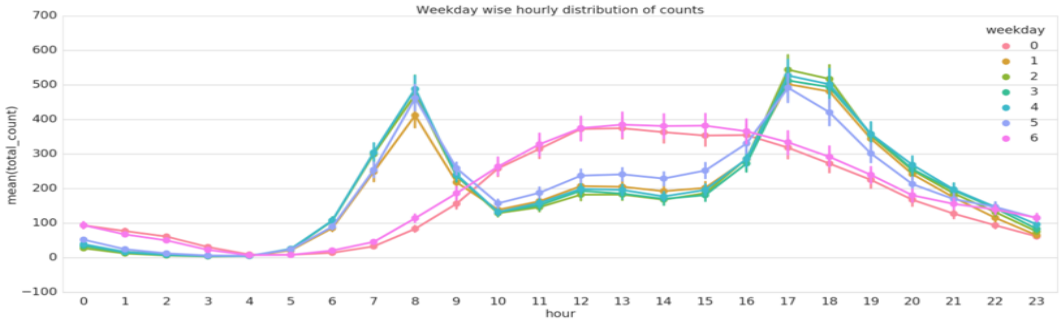


Figure 4: Weekday wise hourly distribution of counts

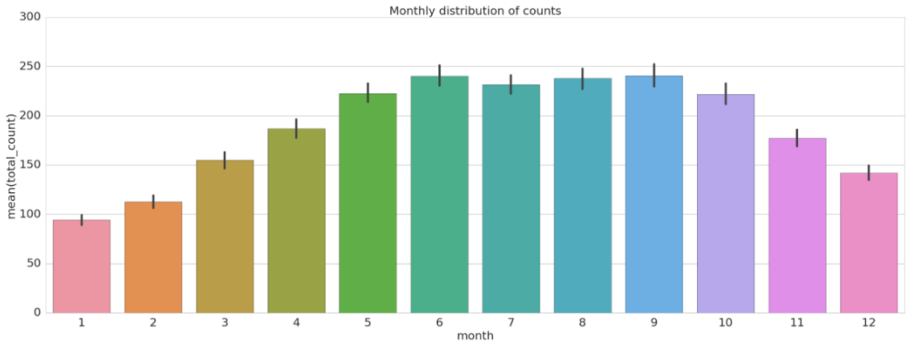


Figure 5: Monthly distribution of counts

We can clearly find usage pattern based on visualization that the commuting hour from 7 to 9 and 16-19 and Summer with Autumn has comparatively high usage than others. Also Weekday and Weekend have a very different usage pattern.

After having a general view of our data, we move on to data pre-processing stage. The main goal is to implement three Tree-based Regression models to identify the relative importance of features and predict and draw conclusions of demands of this bike sharing service at the end. First, we drop the missing values and perform an outlier detection, and there are some extreme outliers in our dataset which are located in upper and lower outer fences. If we include these extreme outliers into our model training, validating and testing, there will be a high chance of an over-fitting problem, so we apply outlier treatment with 'Winsorization' technique to the data above 95% and below 5% of the entire distribution.

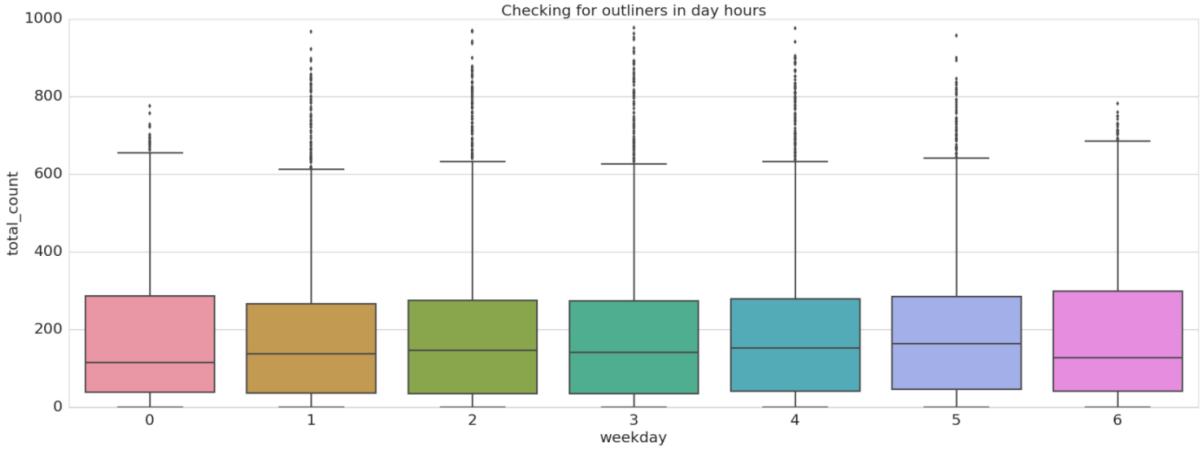


Figure 6: Weekday wise box plot of total counts

Also, we create dummy for the categorical variables including 'season', 'year', 'month', 'hour', 'weekday', 'weathers' to prepare for our model input. Once a categorical variable has recorded as a dummy variable, the dummy variable can used in regression analysis just like any other quantitative variable. Through data pre-processing we clean the noise and inconsistency problems and prepare the values to input properly to our regression models.

Model Description

In order to derive useful business insights and understand which of the variables in the dataset would affect the demand for sharing bikes, we have utilized three tree-based regression modeling techniques for our data analysis.

- **Decision Tree Regression Model**

The decision tree regression model built in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

- **Random Forest Regression Model**

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models.

- **Gradient Boosting Decision Tree Model**

The Gradient Boosted Decision Trees (GBDT) is a machine learning algorithm that iteratively constructs an ensemble of weak decision tree learners through boosting.

Feature Selection

After we divided the dataset into training, validation and testing dataset (60%, 30% and, 10%), we build our models with the training dataset. At first, we intentionally include all features in the models to detect the impact of a specific feature on the model's performance which is feature importance. Then, we ranked the features based on the feature importance in each model.



Figure 7: Ranked features in decision tree model

For the decision tree regression model, we noticed that the following 14 variables have an impact on model performance:

- atemp, hour_17, temp, hour_18, hour_8, humidity, is_workingday, year_0, hour_1, hour_19, hour_2, hour_9, weekday_5, month_8

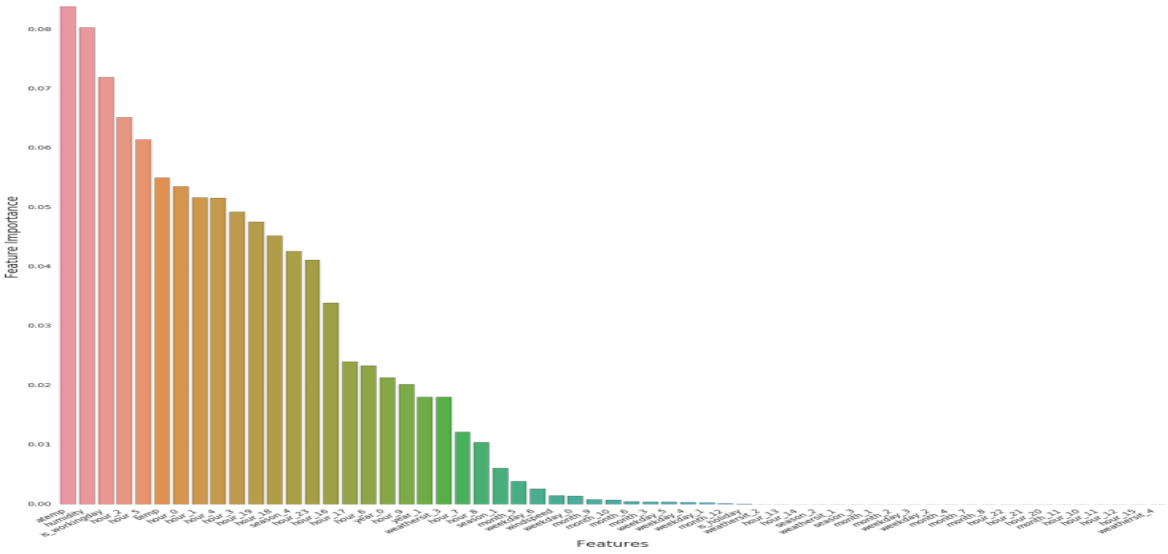


Figure 8: Ranked features in GBDT model

For the GBDT model, we noticed that the following 37 variables have impact on model performance:

- atemp, humidity, is_workingday, hour_2, hour_5, temp, hour_0, hour_1, hour_4, hour_3, hour_19, hour_18, season_4, hour_23, hour_16, hour_17, hour_6, year_0, hour_9, year_1, weathersit_3, hour_7, hour_8, season_1, month_5, weekday_6, month_9, month_10, month_6, month_3, weekday_5, weekday_4, weekday_1, month_12, is_holiday, weathersit_2
-

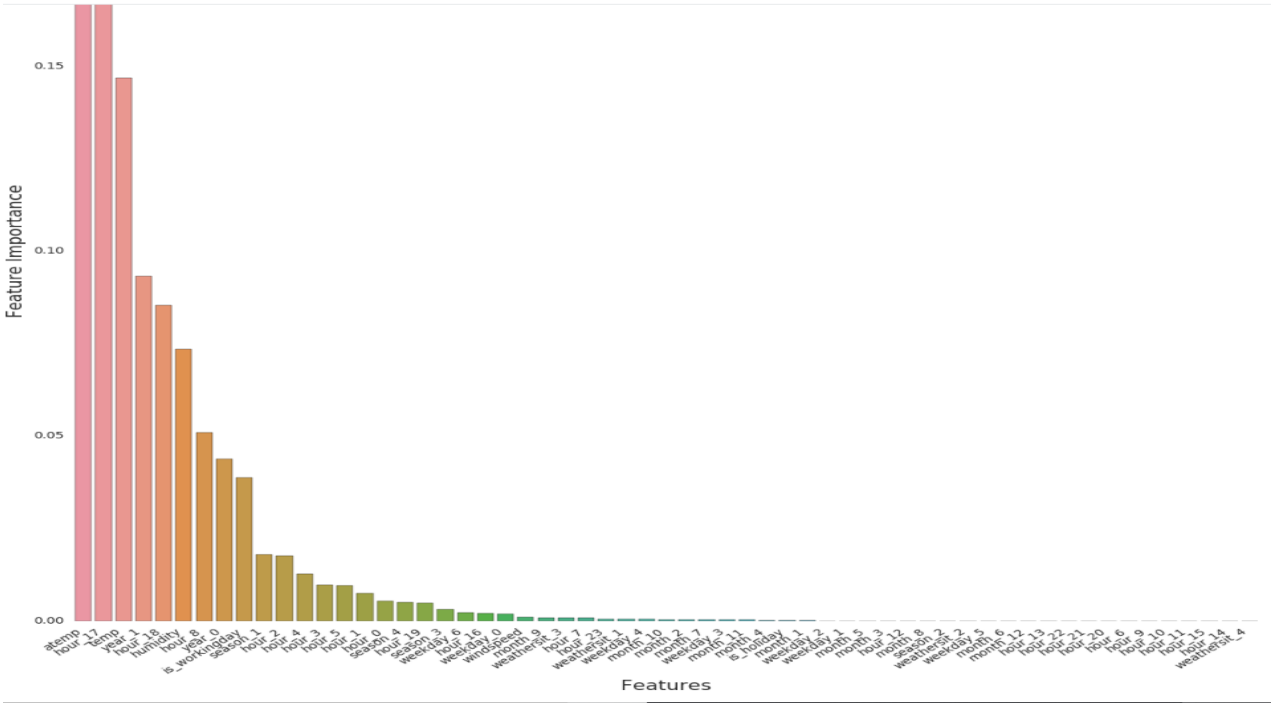


Figure 9: Ranked features in random forest model

For the random forest regression model, we noticed that the following 14 variables have impact on model performance:

- atemp, hour_17, temp, year_1, hour_18, humidity, hour_8, year_0, is_workingday, season_1, hour_2, hour_4, hour_3, hour_5, hour_1, hour_0, season_4, hour_19, season_3, weekday_6, hour_16

Since including all variables that have impact the model performance would increase the computation time, we attempted to reduce the dimensionality based on the model's mean square error (MSE). We applied the step-wise method to calculate the model's MSE and found the best combination of features for each model.

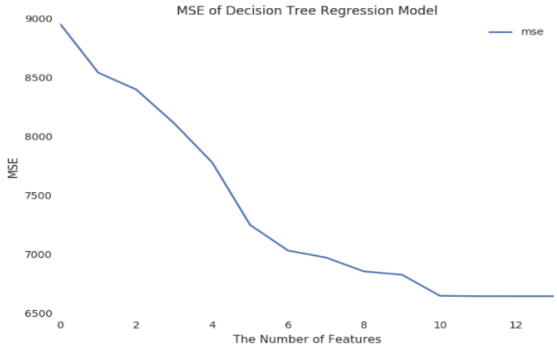


Figure 10: MSE of Decision Tree Regression Model

Based on the distribution of MSE, we would include Top 10 most important features in the Decision Tree Regression model which includes:

- atemp, hour_17, temp, hour_18, hour_8, humidity, is_workingday, year_0, hour_1, hour_19

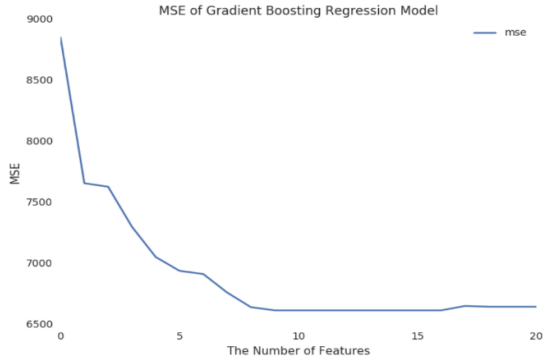


Figure 11: MSE of GBDT Model

For the GBDT model, we would include Top 9 most important features which are:

- atemp, humidity, is_workingday, hour_2, hour_5, temp, hour_0, hour_1, hour_4

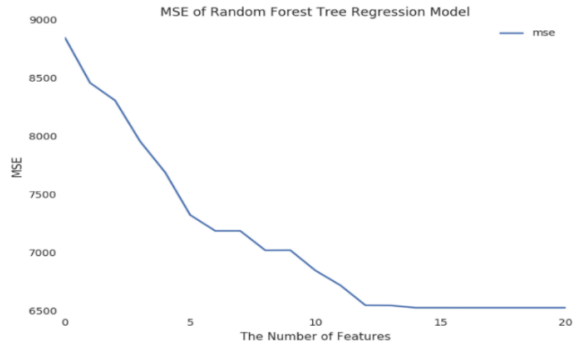


Figure 12: MSE of Random Forest Regression Model

For Random Forest model, we would include Top 12 most important features which are:

- atemp, hour_17, temp, year_1, hour_18, humidity, hour_8, year_0, is_workingday, season_1, hour_2, hour_4, hour_3, hour_5, hour_1, hour_0, season_4, hour_19, season_3

Parameter Tuning

After we figure out the best combination of features, we need to do the parameter tuning to improve the model's predictive ability. We focused on the parameters which have the most impact on the model performance. For the Decision Tree model, we set MaxBins from 2 to 20 and MaxDepth from 1 to 20. For the Gradient Boosting Decision Tree model, we set MaxIter from 2 to 10 and MaxDepth from 2 to 15. For the Random Forest model, we set NumOfTrees from 1 to 200 and MaxDepth from 2 to 10. Furthermore, we wrote an iteration function in python to go through every combination of parameters and calculate the MSE automatically. Then, we applied a similar approach to find the best parameter.

Model Performance

After finished the feature selection and parameter tuning, we fitted each model with the testing dataset and compared the model performance based on the MSE. The following table and graph show each model's detailed information and final performance:

Model	Features	Parameter	MSE
Decision Tree	'atemp','hour_17','temp','hour_18','hour_8','humidity','is_workingday','year_0','hour_1','hour_19'	MaxDepth(10), MaxBins(5)	5734.32877
Gradient Boosting Tree	'atemp','humidity','is_workingday','hour_17','hour_18','temp','hour_8','year_0','season_1'	MaxDepth(8), MaxIter(8)	6010.0221
Random Forest	'atemp','hour_17','temp','year_1','hour_18','humidity','hour_8','year_0','is_workingday','season_1','hour_2','hour_4'	NumTrees(121), MaxDepth(8)	5515.0596

Figure 13: Detailed information for each model

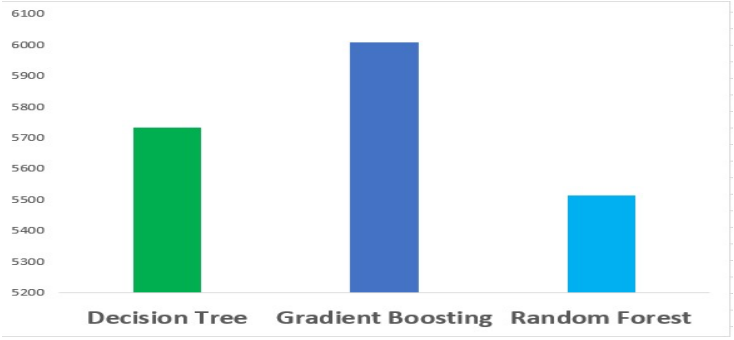


Figure 14: Model performance comparison by MSE

As seen in the graph above, the Random Forest regression model has the best performance with the lowest MSE (5515.0), followed with Decision Tree regression model and Gradient Boosting Decision Tree Model. The best Model includes 12 features, with the parameter of Number of Trees 121, and Max Depth of tree 8.

Conclusion and Actionable Insights

The table below shows the most important features in the Random Forest Regression Model

1	atemp
2	Hour_17
3	temp
4	Hour_18
5	Hour_8
6	Is_workingday
7	Season_1

Figure 15: Top important features in RF

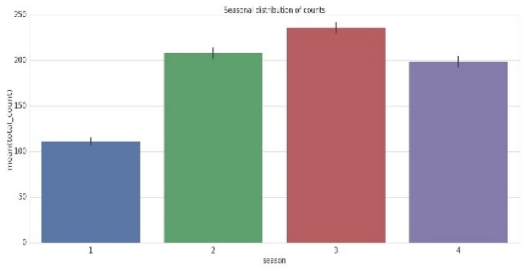


Figure 16: Season wise distribution of counts

Based on the feature importance and descriptive analysis we did before, we can provide actionable insights as follow:

- Features like Hour_17, Hour_18, Hour_8 have significant impact on model performance. There are two peaks of renting bikes in the commuting period. Therefore, we can conclude that there are a great number of people who live in Washington, D.C would like to use sharing bike as commuting transportation. The company should be well-prepared for the rapidly increased demand for the number of sharing bikes in the 8 am, 5 pm and 6 pm.

- Based on the model, whether it is a working day is essential in predicting the renting number of bikes. The difference in renting number between workdays and weekends is that the demand at the weekends during the daytime (especially in the middle of the day) is more than it in the workdays. Therefore, the company should increase the supplement of sharing bikes in the daytime during the weekends.

- According to the distribution of renting numbers in each season, we can conclude that the demand for sharing bikes is relatively low during spring. The company should attempt to renovate or repair damaged bikes in this time.

- The temperature is an essential driver for predicting sharing bike renting numbers. We can see that there is a rising number of demands for sharing bikes in warmer weather. Therefore, the company should increase the supplement of sharing bikes as temperature increases.

References

Kaggle, Bike Sharing in Wahington D.C. *Kaggle* [Online] Available at: <https://www.kaggle.com/marklvl/bike-sharing-dataset> [Accessed 1 March, 2019]
Turi Machine Learning Platform User Guide. *Turi* [online] Available at: <https://turi.com/learn/userguide/supervised-learning/random_forest_regression.html> [Accessed 30 April, 2019]
Gradient boosting *WikiPedia*. [online] Available at: <https://en.wikipedia.org/wiki/Gradient_boosting> [Accessed 30 April, 2019]
Decision Tree - Regression. *Data mining Map* [online] Available at: <https://www.saedsayad.com/decision_tree_reg.htm> [Accessed 30 April, 2019].