PREDICTION OF DAILY BIKE RENTAL COUNT

TEJASH S. POPATE

August 2019

Table of Contents

1. **Introduction3**
   1. Problem Statement3
   2. Data3
2. **Methodology4**

2.1 Data Pre Processing4

2.1.1 Exploratory Data Analysis5

2.1.2 Missing Value Analysis6

2.1.3 Outlier Analysis8

2.1.4 Feature Selection11

2.1.5 Feature Scaling12

2.2 Modelling13

2.2.1 Model Selection13

2.2.2 Linear Regression13

2.2.3 Decision Trees14

2.2.4 Random Forest14

1. **Conclusion15**

3.1 Model Evaluation15

3.2 Model Selection15

1. **Appendices 16**

4.1 Appendix A- Extra Figures17

1. **References**

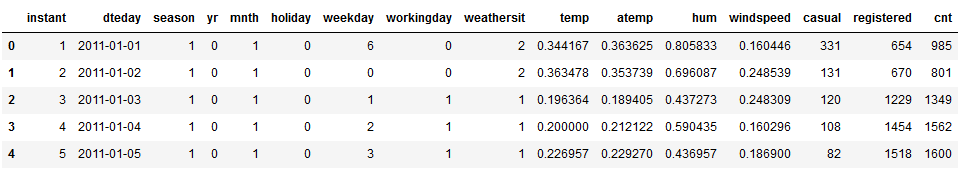
# Chapter 1 **Introduction:**

## **Problem Statement:**

The objective of this project is to predict the daily bike rental count based on the environmental and seasonal settings.

## **Data:**

The aim is to predict daily bike rental count using given dataset and implementation of machine learning models. Given below is a sample of dataset that is for rest of this project.

Table 1.1: Daily Bike Rental Count Data (Columns 1-17)

In the given dataset, there are **16 variables** of data type numeric and total **731 observations** including missing values in some of the rows. After giving a look to the dataset, the variables have been defined in different category:

Table 1.2: Attributes’ Category

|  |  |  |
| --- | --- | --- |
| Type of variable | Data Type | Variable Category |
| Predictor Variables:   1. instant 2. dteday 3. season 4. yr 5. mnth 6. holiday 7. weekday 8. workingday 9. weathersit 10. temp 11. atemp 12. hum 13. windspeed 14. casual 15. registered | **Datetime:**  dteday | **Categorical:**   1. season 2. yr 3. mnth 4. holiday 5. weekday 6. workingday 7. weathersit |
| Target Variable:  cnt | **Numeric:**  All of them except “dteday” | **Continuous:**   1. temp 2. atemp 3. hum 4. windspeed 5. casual 6. registered 7. cnt |

Table 1.3: Attribute Details

|  |  |
| --- | --- |
| Variable | Definition |
| Instant (ID) | Record index |
| Dteday (dateday) | Date |
| season | Type of season (1 = spring, 2 = summer, 3 = fall, 4 = winter) |
| yr (year) | Year (0 = 2011, 1 = 2012) |
| mnth (month) | Month (1 to 12) |
| holiday | whether the day is considered a holiday |
| weekday | Day of the week |
| workingday | If day is neither weekend nor holiday is 1, otherwise is 0. |
| weathersit (weather\_situation) | weather type (extracted from Freemeteo)  1: Clear, Few clouds, Partly cloudy, Partly cloudy  2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| temp | Normalized temperature in Celsius. |
| atemp | Normalized feeling temperature in Celsius. |
| hum (humidity) | Normalized humidity. The values are divided to 100 (max) |
| windspeed | Normalized wind speed. The values are divided to 67 (max) |
| casual | count of casual users |
| registered | count of registered users |
| cnt (count) | count of total rental bikes including both casual and registered |

{Note: Names written in brackets () represent names given in future for easiness of reference}

# Chapter 2 **Methodology:**

## **2.1 Data Pre-processing:**

### 2.1.1 Exploratory Data Analysis: (EDA)

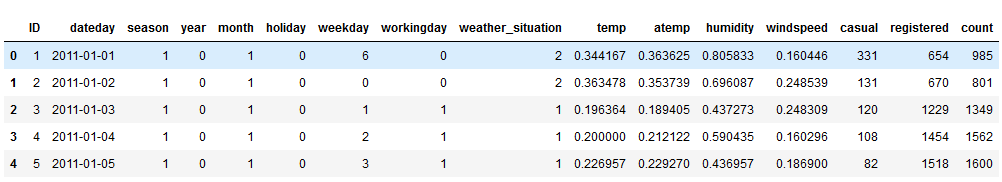
It is used to get the initial insights of data. It helps in understanding all the statistical parameters of both categorical and numeric variables/features from the data and also it helps in understanding presence of missing values and unique values in each feature. The imputation of these missing values and impact of imputation on the overall data is explained in the next step. For simplicity, some column names were given proper names using following code:

bike\_data.rename(columns={'instant':'ID','dteday':'dateday','mnth':'month',

'weathersit':'weather\_situation','yr':'year',

'hum': 'humidity','cnt':'count'},inplace=True)

[Table 2.1: Dataset with proper column names]



EDA includes two parts:

1. Univariate Analysis b) Bivariate Analysis

Univariate Analysis: This type of analysis helps in detecting anomaly in the data. Exploration depends on type of variable. If it is a continuous variable, the parameters such as central tendency, dispersion and distribution of variable (symmetric /right skewed/ left skewed)) are considered. If it is categorical variable, then frequency table, histogram and bar-plot are checked. Following tables and figures shows univariate analysis of each variable.

Table 2.2 EDA of categorical variables

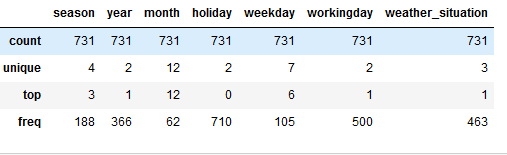
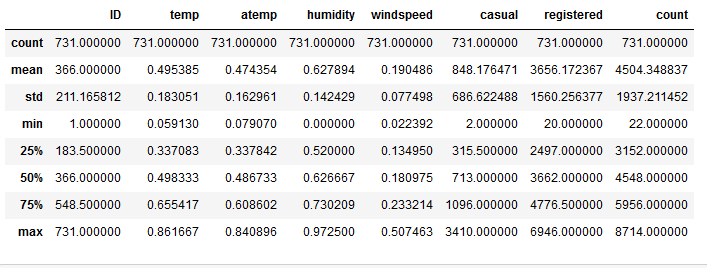
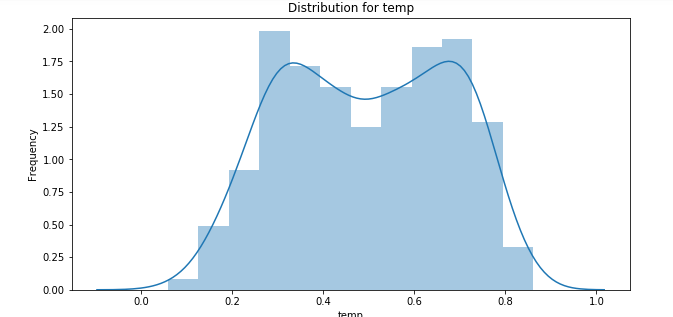
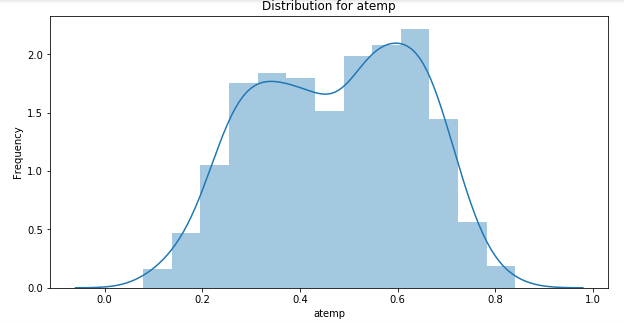
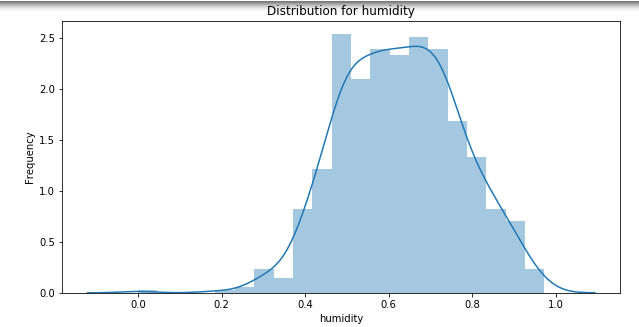
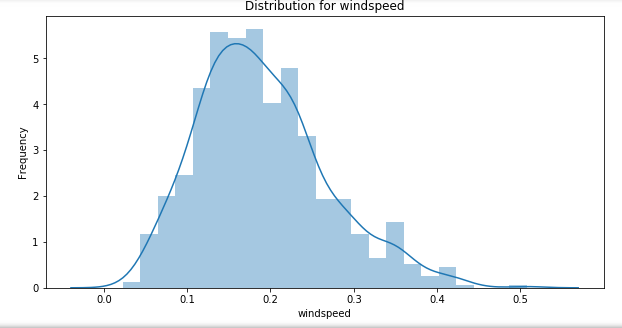
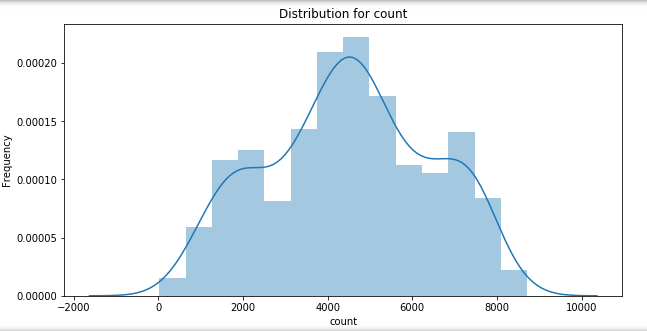
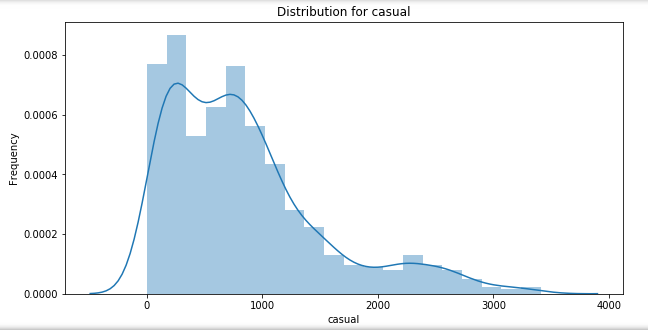
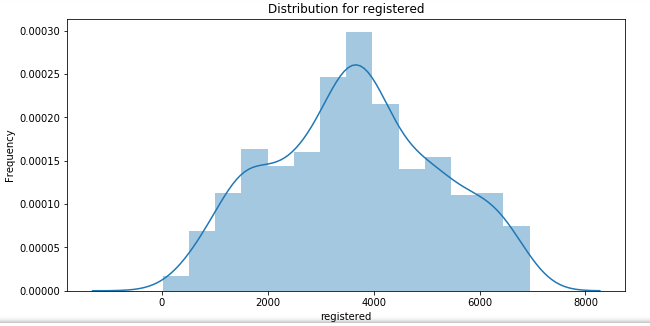


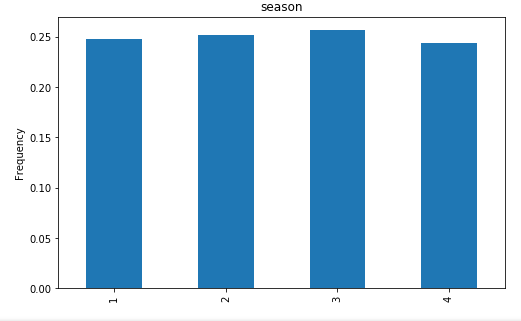
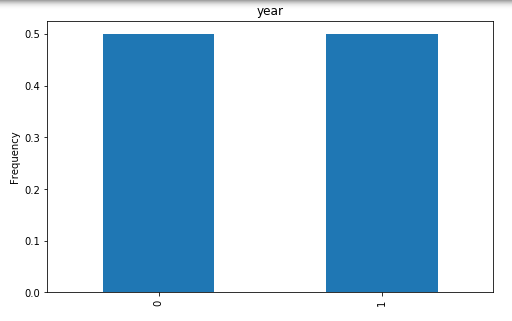
Table 2.3 EDA of continuous variables

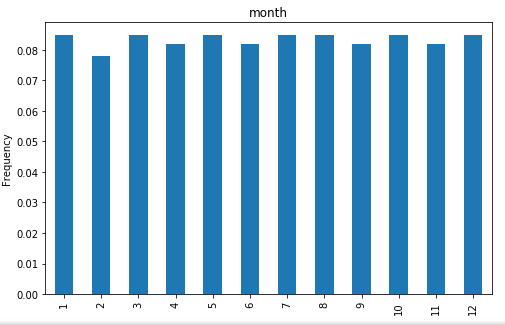
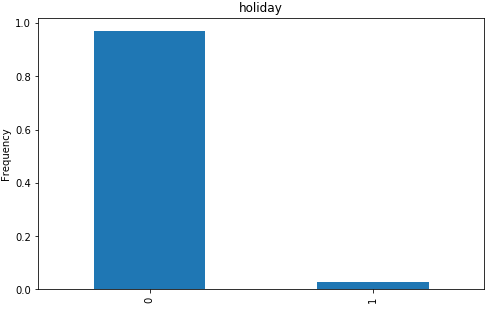


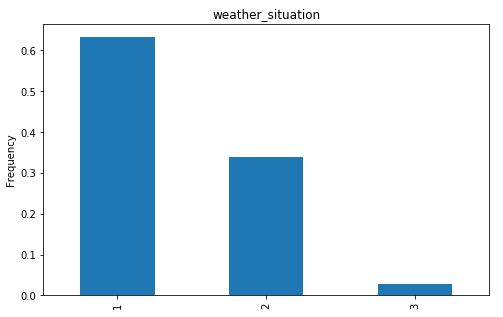
Figure 2.1.a Histogram plots of continuous variables

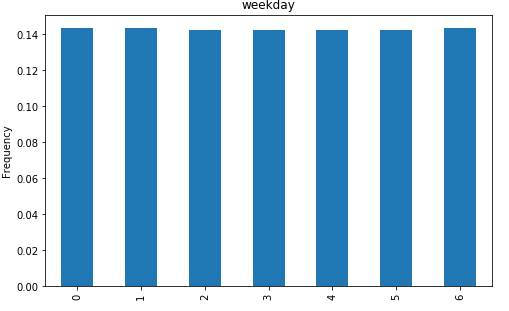
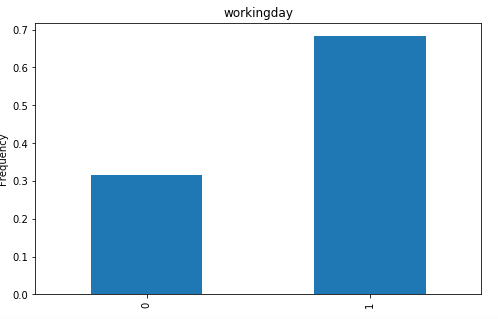




Figure 2.1.b Bar-plots of Categorical Variables

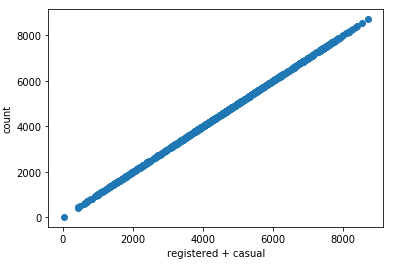




Inferences:

Bivariate Analysis: Here, two variables are studied together for their empirical relationship and association of two variables is studied. It helps in feature selection and predictions and also helps in detecting anomalies in the data. This type of analysis can be categorized based on relation between variables as follows:

* Continuous -continuous variable: Both considered variables for analysis are of continuous type. Scatter plot and correlation plot are used for this kind of analysis. In case of given data, addition of two variables “registered” and “casual” gives equivalent result with predictor “count”
* Categorical-Continuous variable: Here one variable is categorical in nature and another is continuous in nature. Bar plot and two sample t-test are used for such analysis
* Categorical-Categorical variable: Both variables are of categorical in nature. Two way table and chi-squared test are used for such type of analysis. This also helps in feature selection part.

### 2.1.2 Missing Value Analysis:

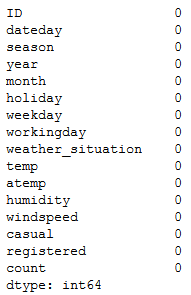
None of the variables in the dataset have any missing values.

Fig 2.2 Missing values percentages in each variable

### 2.1.3 Outlier Analysis:

Mean is most affected by outliers and outliers will not affect median, mode.

Graphical method of detection: Boxplot

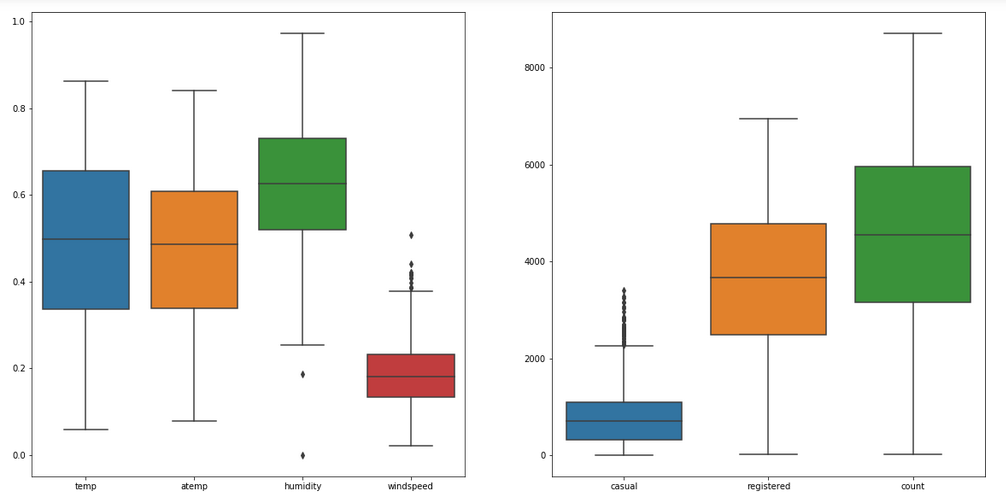
Formula: If a value satisfies following condition, then it’s an outlier

Where IQR=Q3-Q1 (Inter Quartile Range)

Treating Outliers: There are multiple ways for treatment of outliers

1. Deleting these outliers
2. Transforming and Binning these outliers (like taking log() of value)
3. Imputing outliers similar to missing value
4. Treat them separately. (Perform diff operations on these sets of variable and diff operations on remaining sets of variable.
5. Replace the outliers with cut-off values of the variables.

Fig 2.3 Boxplots of continuous variables



In case of given problem statement, out of **7** continuous predictors, only “humidity”, “windspeed”, ”casual” have any outliers. Out of them, outliers from predictors “humidity” and “windspeed” were treated as outliers from “casual” held important information. Outliers were visualized using boxplots of each continuous predictors.

### 2.1.4 Feature Selection:

Variable Importance is crucial in ML modelling where a subset of relevant features/variables are selected for use of model construction. Feature selection is done to avoid over-fitting, to make fast predictions and training, to decrease storage required for model and the dataset. Two techniques for feature selection:

1. Domain Knowledge
2. ML algorithm’s usage.

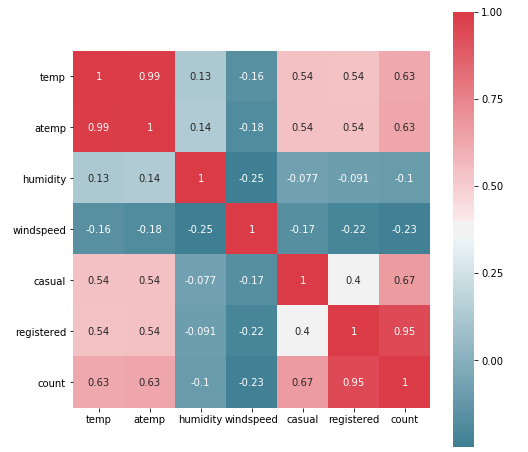
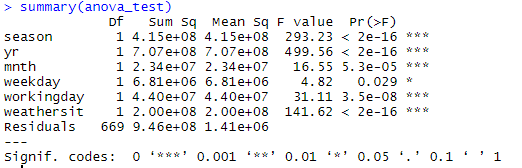


Fig 2.4 Correlation plot of continuous variables

Here correlation analysis helped in dropping the highly correlated continuous variables (i.e. “Body mass index” column).Also ANOVA test was used to check dependencies of categorical variable with the continuous target variable. ANOVA uses one categorical and one numerical variable to calculate the relevancy of that particular variable. Using the probability value generated by ANOVA test, those variables which were having p value less than 0.05 used as features for prediction.   


Further “atemp” has collinearity issue after doing VIF analysis

---------- VIFs of the remained variables --------

Variables VIF

1 temp 1.0

2 hum 1.1

3 windspeed 1.1

Following is the list of variables (features) selected for model construction:

1. Temp
2. Humidity
3. Windspeed
4. Season
5. Year
6. Month
7. Weekday
8. Workingday
9. Holiday
10. Weather\_situation

### 2.1.5 Feature Scaling:

All of continuous predictors in the data were already normalized. So there’s no need of scaling of predictors here.

## **2.2 Modelling:**

### 2.2.1 Model Selection:

After pre-processing of the data, ML models are used for making predictions. Given problem is a regression problem. Thus regression based models such as linear regression, decision trees, random forest and gradient boosting were selected to predict the target variable.

### 2.2.2 Linear Regression:

This algorithm is used to predict one variable using another variable when both of them are continuous in nature. It is a part of supervised learning algorithm. Linear regression is used for regression problems. R squared metric and RMSE (root mean squared error) metric will help in evaluating regression models.

Table 2.4 Evaluation metrics for Linear Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Linear Regression | R | | Python | |
| Train data | Test data | Train data | Test data |
| RMSE | 731 | 693 | 764.94 | 744.40 |
| MAPE | 46 | 19 | 44.41 | 18.03 |
| R squared | 0.85 | 0.85 | 0.842 | 0.850 |
| MAE | 537 | 513 | 559.97 | 522.18 |
| MSE | 534746 | 480128 | 585135.87 | 554140.08 |

### 2.2.3 Decision Trees:

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. A decision tree is a structure that describes a basic process to follow to reach a conclusion.

Table 2.5 Evaluation metrics for Decision Trees

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decision Tree | R | | Python | |
| Train data | Test data | Train data | Test data |
| RMSE | 833 | 800 | 690.13 | 744.40 |
| MAPE | 55 | 24 | 16.73 | 18.03 |
| R squared | 0.8 | 0.8 | 0.871 | 0.85 |
| MAE | 630 | 590 | 529.17 | 522.18 |
| MSE | 693437 | 639283 | 476279 | 554140 |

### 2.2.3 Random Forest:

In random forest algorithm, number of decision trees created internally is decided by the error rate. It will build the trees until the error no longer decreases. Thus it is not possible to predict number of tree created in random forest algorithm if number of trees are not defined. Random Forest is an ensemble technique that consists of many decision trees. The idea behind Random Forest is to build n number of trees to have more accuracy in dataset.

Table 2.6 Evaluation metrics for Random Forest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | R | | Python | |
| Train data | Test data | Train data | Test data |
| RMSE | 325 | 6500 | 261.71 | 602.37 |
| MAPE | 25 | 15 | 18.05 | 15.05 |
| R squared | 0.97 | 0.90 | 0.981 | 0. 901 |
| MAE | 230 | 53000 | 175.38 | 394.51 |
| MSE | 105327 | 4.3e+07 | 68495.63 | 362852 |

### 2.2.4 Gradient Boosting:

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function

Table 2.6 Evaluation metrics for Gradient Boosting

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gradient Boosting | R | | Python | |
| Train data | Test data | Train data | Test data |
| RMSE | 10000 | 7600 | 422.79 | 610.0 |
| MAPE | 23 | 16 | 11.98 | 14.09 |
| R squared | 0.95 | 0.87 | 0. 951 | 0. 899 |
| MAE | 1.8e+05 | 6.2e+04 | 307.41 | 426.93 |
| MSE | 1.0e+08 | 5.7e+07 | 178758 | 372109 |

# Chapter 3 **Conclusion**

This chapter deals with evaluation of models and selection of best model for the given problem statement as mentioned in chapter 1.1 and also patterns were analysed from visualizations made throughout the project.

## **3.1 Model Evaluation:**

Model evaluation is done based the values of metrics such as RMSE, MSE, MAE, R-squared value. **RMSE** (Root Mean Squared Error) is the most popular evaluation metric used in regression problems. It follows an assumption that error are unbiased and follow a normal distribution. The power of ‘square root’ empowers this metric to show large number deviations. It represents the sample standard deviation of the differences between predicted values and observed values (called residuals). **MAE** is the average of the absolute difference between the predicted values and observed value. The MAE is a linear score which means that all the individual differences are weighted equally in the average. The MAE is also the most intuitive of the metrics since we’re just looking at the absolute difference between the data and the model’s predictions. RMSE penalizes the higher difference more than MAE. Generally, RMSE will be higher than or equal to MAE. The only case where it equals MAE is when all the differences are equal or zero. It is important to note that the units of both RMSE & MAE are same. The range of RMSE & MAE is from 0 to infinity. MAE is robust to outliers whereas RMSE is not. **The coefficient of determination, or R²** (sometimes read as R-two), is another metric we may use to evaluate a model and it is closely related to MSE, but has the advantage of being scale-free — it doesn’t matter if the output values are very large or very small, the R² is always going to be between negative infinity and 1. The absolute value of RMSE does not actually tell how bad a model is. It can only be used to compare across two models whereas R² easily does that.  **R**-**squared** is a relative measure of fit, **RMSE** is an absolute measure of fit. Theoretically, if a model has adjusted R² equal to 0.05 then it is definitely poor. The maximum value of R² is 1 but minimum can be negative infinity. **MSE** basically measures average squared error of our predictions. For each point, it calculates square difference between the predictions and the target and then average those values. The higher this value, the worse the model is. It is never negative, since we’re squaring the individual prediction-wise errors before summing them, but would be zero for a perfect model.

## **3.2 Model Selection:**

The “**Random Forest**” model has best set of evaluation metric when compared with other models, and so it was chosen for modelling. Also, over-fitting will be less as evaluation parameters for both test and train do not differ much when compared.

# Chapter 4 **Appendices** **(Extra Figures):**

# 

# Chapter 5 **References**

1. <http://thestatsgeek.com/2013/10/28/r-squared-and-adjusted-r-squared/?source=post_page>
2. Martiniano, A., Ferreira, R. P., Sassi, R. J., & Affonso, C. (2012). Application of a neuro fuzzy network in prediction of absenteeism at work. In Information Systems and Technologies (CISTI), 7th Iberian Conference on (pp. 1-4). IEEE.
3. Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.
4. Grigorios Papageorgiou, Stuart W Grant, Johanna J M Takkenberg, Mostafa M Mokhles, Statistical primer: how to deal with missing data in scientific research?, *Interactive CardioVascular and Thoracic Surgery*, Volume 27, Issue 2, August 2018, Pages 153–158,
5. <http://www.cs.columbia.edu/~amueller/comsw4995s19/schedule/>
6. [1] Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software