Capstone Project Milestone Report:

*Prediction of Emergency Roadside Assistance car breakdown Call Volume.*

# Introduction

The goal of the project is to build the predictive model to predict the call volume of the emergency roadside assistance calls for car breakdown. This paper will describe the data wrangling on how the data sets are merged together and exploratory analysis of the data. Outlined the detailed approach of building the predictive model using R caret package and covers evaluation of metrics for each algorithm used in the model. And finally compared the predictions and accuracy metric between the best predicted model of the training data and test data. Used the models lm, glmnet, svmRadial, knn, CART, rf, gbm, cubist and treebag in *caret* for predicting the call volume.

# Data set

The data set used in this study is the “Call volume count” from the NJ DOT <http://www.state.nj.us/transportation/refdata/> and “weather data” from [www.weatherunderground.com](http://www.weatherunderground.com) .

There are total 6 data sets in this study. One data set with actual call volume for each city with breakdown location, and the call date. And 5 datasets from weather data for each weather station in NJ which are located to that nearest city. Weather data has the weather information for each day with details of weather events, season, humidity, wind speed and temperature.

Call date is split into additional variables working day, is\_a\_public\_holiday, is\_weekend, Month and quarter. Converted the data type for call date to date. And converted the other variables month number, working day, weekend, is public holiday variables to factor.

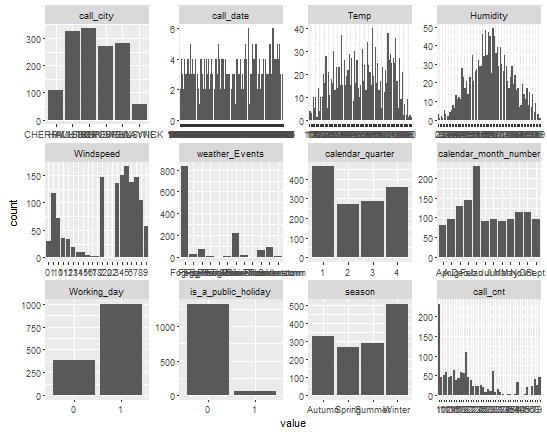
Removed the duplicate variables date\_key, wevents, Month\_name, week\_day\_name, week\_day\_number, is\_weekend\_day

The call volume data is merged with the weather data using the variables call city and call date. The final data set was then cleaned to handle the missing data. Temp is missing in some observations and so imputed with median value.

And then the data set is sampled into train and test data with 80% train data and 20% test data.The predictive model is built on the train data set. And Test data is used after the model is built to test the accuracy of the model.

# Exploratory analysis of data

Exploratory analysis is done to understand the distribution of data on each variable using histograms and the distribution of data between the response variable and independent variables is done using boxplots.

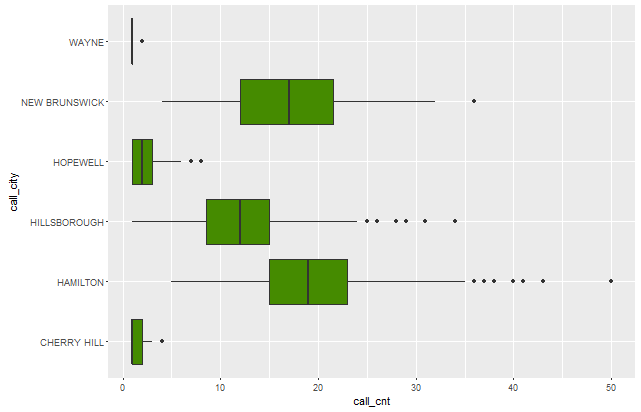


**Figure 1**

Here are some quick inferences drawn from Figure1 for the variables in the train data set.

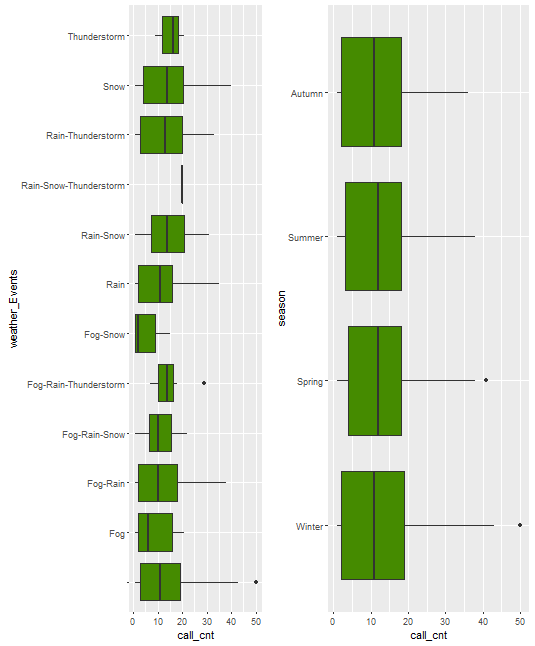
* There are two cities that have more observations than the rest of the cities.
* Humidity, Temp seems to be normally distributed through the number of observations.
* Wind speed seems to be varying throughout the observations.
* Weather events has maximum number of calls during one weather event compared to others.
* During the first quarter seems to be the number of observations are high and one month has peak values.
* Working day has more number of observations.
* During public holidays there are less number of calls compared to regular days.
* Season 1- which means winter season has more number of calls
* At the end last graph call\_cnt seems to be right skewed instead of normally distributed.

Plotted the Boxplot to understand the data distributions of two variables (bivariate analysis) in the data set.



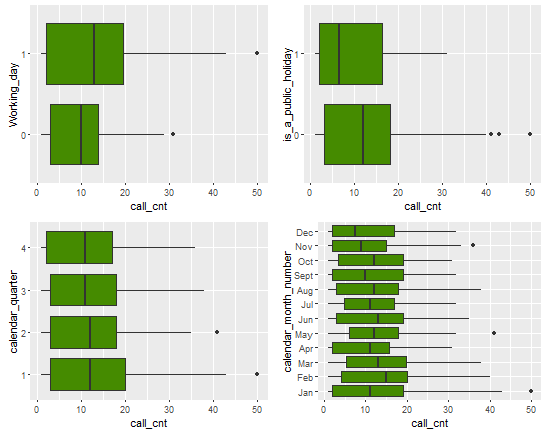
**Figure 2**

From the Figure2, we can see that Hamilton and New Brunswick cities in NJ has more call volume. And the dotted lines on the Hamilton boxplot shows there are few calls with peak values.



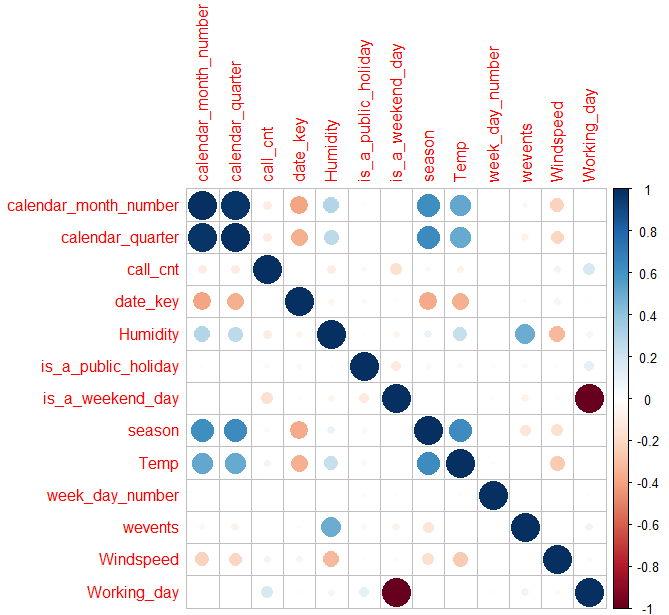
**Figure 3**

* From Figure 3, the average call volume for each season seems to be approximately same but the spread of calls during the winter seems to be pretty high compared to the other seasons. But each weather event has different distributions and when the weather event is clear which means no fog, no rain, no thunderstorm and no snow then the call volume is high, because the data from the weather underground has not provided the description for weather events when there is no rain, fog, thunderstorm, or snow.



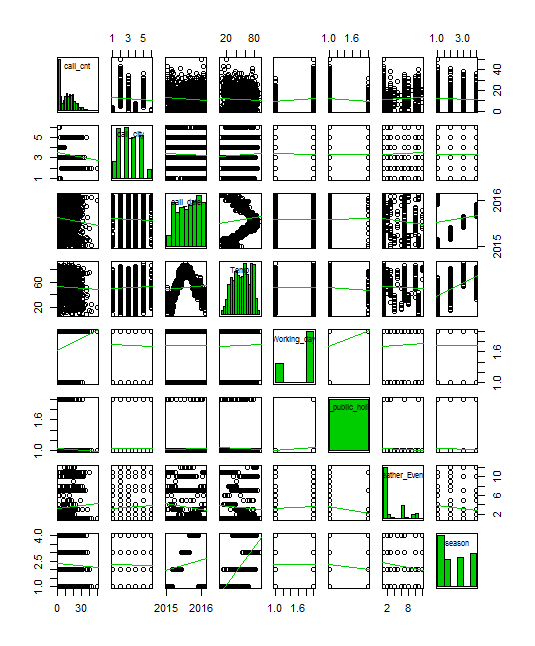
**Figure 4**

From Figure4, It seems the call volume is high during regular days and in that January and February seems to be high compared to other months.



**Figure 5**

From the Figure 5, it seems that weather events (wevents) and Humidity are positively correlated. Wevents and season are negatively correlated. Wind speed, Temp and Humidity are highly correlated to each other. Season and wind speed are highly correlated to each other.



**Figure 6**

Figure 6 is a scatter plot matrix to understand the correlation between all the variables. Temp and season seems to have linear relationship and highly correlated to each other. But the response variable Call count seems to have a nonlinear relationship with independent variables.

# Building and evaluating predictive model

In this study I used 9 different machine learning algorithms (lm, glmnet, svmRadial, knn, CART, rf, gbm, cubist and treebag) to train the model for estimating the accuracy and evaluated the metrics using cross validation with 10 folds.

After the models are trained. Added the results to a list using resamples () for comparing the RMSE scores for each model. And then called the summary function to see the results.

Here is the summary of results in a table with each algorithm for each row of RMSE score and R2 metrics for each column.

|  |  |  |
| --- | --- | --- |
| **Algorithm Models** | **RMSE** | **Rsquared** |
| glmnet | 4.73 | 0.72 |
| CART | 6.48 | 0.47 |
| SVM | 4.87 | 0.71 |
| Linear regression (lm) | 4.74 | 0.72 |
| K-Nearest Neighbour (knn) | 5.95 | 0.56 |
| gradiant boosting(gbm) | 4.57 | 0.74 |
| random forest(rf) | 4.54 | 0.74 |
| cubist | 4.61 | 0.73 |
| treebag | 4.66 | 0.73 |

**Table 1**

From the above results of Table1, the glmnet Model, the RMSE (Root Mean Square Error) the average deviation of the predictions from the observations seems to be lowest compared to the other models and from the ensemble methods random forest has the lowest RMSE value.

|  |  |  |
| --- | --- | --- |
| After BoxCox transformation | | |
| **Model** | **RMSE** | **Rsquared** |
| glmnet | 4.73 | 0.72 |
| CART | 6.48 | 0.47 |
| SVM | 4.86 | 0.71 |
| Linear regression (lm) | 4.74 | 0.72 |
| K-Nearest Neighbour (knn) | 5.94 | 0.56 |
| gradiant boosting(gbm) | 4.57 | 0.74 |
| random forest(rf) | 4.54 | 0.74 |
| cubist | 4.61 | 0.73 |
| treebag | 4.66 | 0.73 |

**Table 2**

Table2 is the result of the evaluation of metrics after using model using BoxCox transform in the preprocess. But it seems there was not much improvement compared to Table1.

# Improving Accuracy

Tried to tune the model and improve the accuracy using tune length. And here are the results for each model after tuning the parameters.

|  |  |  |
| --- | --- | --- |
| After tuning parameters | | |
| **Model** | **RMSE** | **Rsquared** |
| gradient boosting(gbm) | 4.42 | 0.76 |
| random forest(rf) | 4.54 | 0.74 |
| cubist | 4.55 | 0.74 |
| glmnet | 4.73 | 0.72 |

**Table 3**

Table3 is the results after further tuning the model using the tune parameter tune length=10. And it seems there is quite an improvement in the RMSE and Rsquared value.

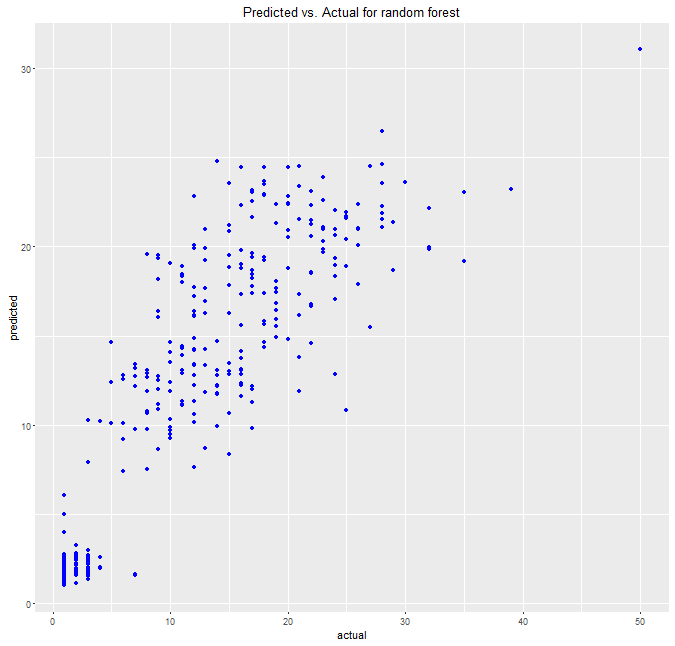
# Finalizing the model

After evaluating the algorithms and tuning the parameters to improve the accuracy, finalized the 4 models from Table3 to test with predictions on the test data and see the comparison of RMSE with training data and test data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Trained Model | | After Predicting with test data | |
| Model | RMSE | Rsquared | RMSE | Rsquared |
| gradiant boosting(gbm) | 4.42 | 0.76 | 7.9 | 0.77 |
| random forest(rf) | 4.54 | 0.74 | 4.48 | 0.75 |
| cubist | 4.55 | 0.74 | 4.53 | 0.75 |
| glmnet | 4.73 | 0.72 | 4.74 | 0.72 |

**Table 4**

From Table 4, after making the predictions though gbm did pretty good with training data with RMSE of 4.42 value, with test data set RMSE value is very bad. And random forest seems to be good with improved accuracy on the test data set with a value of 4.48.



**Figure 7**

Figure7 is the visualization of Predicted Vs Actual values of the data set using random forest data set.

|  |  |
| --- | --- |
| Actual | Predicted |
| 1 | 6 |
| 1 | 5 |
| 1 | 3 |
| 1 | 1.8 |
| 1 | 1.5 |
| 1 | 1.3 |

**Table 5**

Table 5 is the head of 5 rows actual and predicted data using random forest model.