Capstone Project Report:

*Prediction of Emergency Roadside Assistance Call Volume.*

# Introduction

The goal of the project is to predictively model call volume of emergency roadside assistance calls for car breakdown. In particular, this paper will describe the following:

1. Wrangling / merging of data sets
2. Exploratory analysis of the data.
3. A detailed approach of building predictive models using the ‘caret’ R package that covers evaluation of metrics for each algorithm used.
4. Finally, predictions and accuracy of different models tried are compared and finalized the model.

# Data set

The data set used in this study is the “Call volume count” from two datasets.

1). New Jersey department of transportation (NJ DOT) <http://www.state.nj.us/transportation/refdata/> 2). “Weather data” from [www.weatherunderground.com](http://www.weatherunderground.com) .

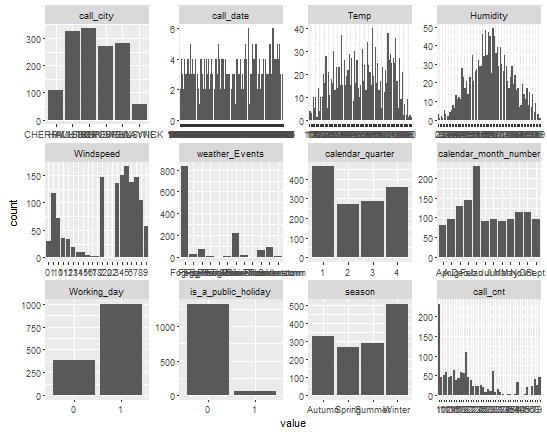
There are total 6 data sets in this study. One data set with actual call volume for each city with breakdown location, and the call date. And 5 datasets from weather data for each weather station in NJ which are located to that nearest city. Weather data has the weather information for each day with 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 using TSQL. The call volume data is merged with the weather data using the variables call city and call date using SQL.

The final data set was then cleaned to handle the missing data using R.

* Call date was converted to data date using R.
* The variables month number, working day, weekend, is public holiday variables are converted to factor using R.
* The variables date\_key, wevents, Month\_name, week\_day\_name, week\_day\_number, is\_weekend\_day are removed due to duplicate data.
* Temperature variable is missing in some observations and so imputed with median value using R.
* And then the data set is sampled into train and test data with 80% train data and 20% test data using R. 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.

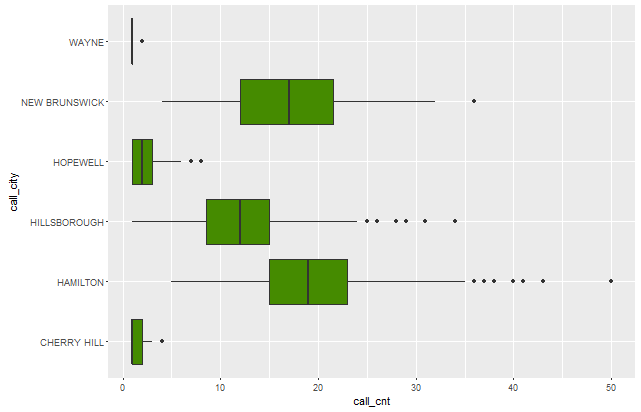


**Figure 1**

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

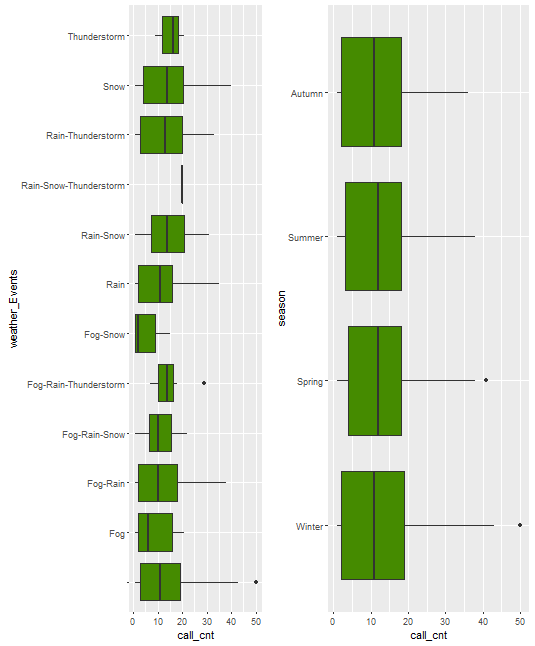
* There are two cities that have more observations than the rest of the cities.
* Humidity, Temp seem to have somewhat ‘balanced’ distributions.
* Wind speed seems to be varying throughout the observations.
* Weather events has maximum number of calls during one weather event compared to others.
* The first quarter seems to have a higher number of observations compared to the other quarters. Similarly, the month of XXXX seems to have a higher number of observations compared to the other months.
* Working days seem to have a higher number of observations.
* Public holidays have a lesser number of calls than regular days.
* Winter season has more number of calls
* The outcome variable (call\_cnt) seems to be right skewed

Boxplots were plotted to understand data distributions of two variables (bivariate analysis) in the data set.



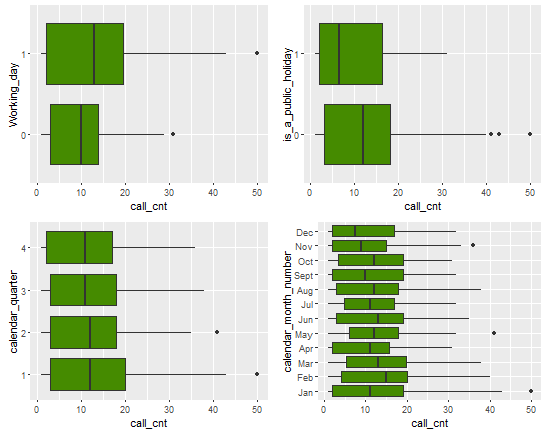
**Figure 2**

From Figure2, we can see that Hamilton and New Brunswick cities in NJ have more call volume.



**Figure 3**

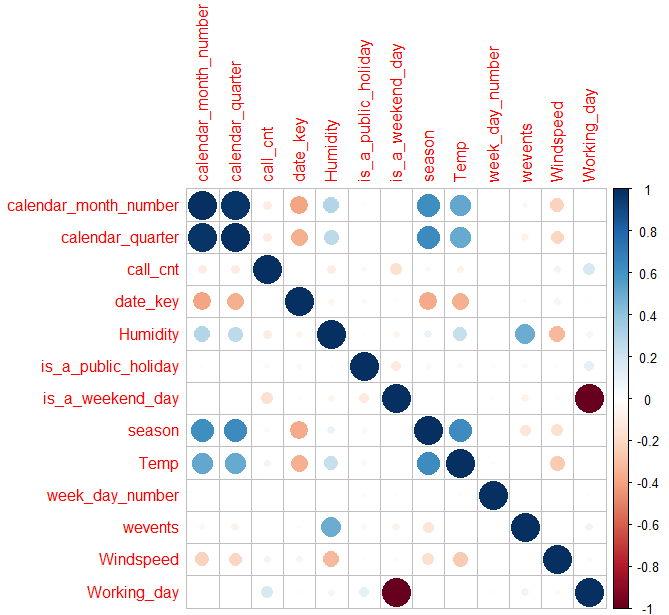
From Figure 3, the average call volume for each season seems to be approximately same but the spread of calls during the winter seems to be slightly high compared to the other seasons. But each weather event has different distributions and when the weather event is clear which means no fog, no rain, 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**

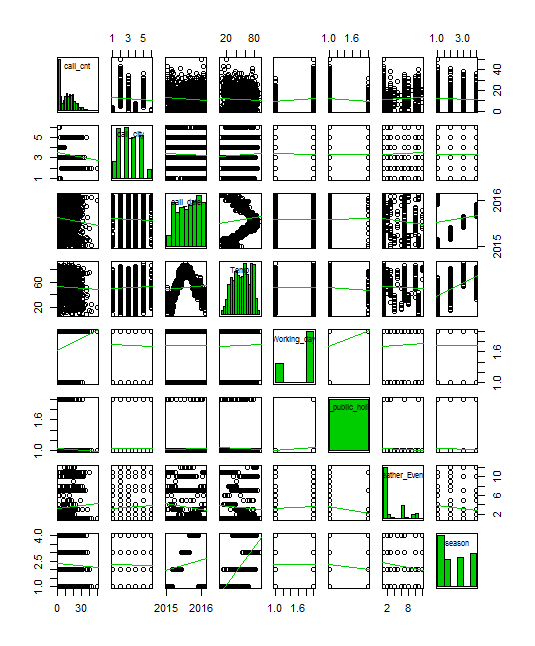
From Figure 4, it seems call volume is high during regular days (working and non-holidays). Also, January

Seems to have a wider range of calls compared to other months. February seems to have highest average call volume counts compared to other months.



**Figure 5**

From Figure 5, it seems that weather events (wevents) and Humidity are positively correlated. Weather events 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. And there is a very high correlation between the two created variables Working\_day and is\_a\_weekend.



**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 we used 9 different machine learning algorithms (lm, glmnet, svmRadial, knn, CART, rf, gbm, cubist and treebag) to train the model using the ‘caret’ package in R for estimating the accuracy and evaluated the metrics using cross validation with 10 folds.

After the models are trained, the results are compared using RMSE scores.

Table 1 contains a summary of results with RMSE score and R2 metrics for each algorithm.

|  |  |  |
| --- | --- | --- |
| **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 results of Table1, RMSE (Root Mean Square Error) of the glmnet model seems to be the lowest. In the ensemble methods tried, ‘random forest’ has the lowest RMSE.

|  |  |  |
| --- | --- | --- |
| After Box-Cox 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 Neighbor (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**

Table 2 is the result of the evaluation of metrics after using model using Box-Cox transform in the preprocessing. But it seems there was not much improvement compared to Table1.

# Improving Accuracy

The models mentioned in the previous section were tuned to improve their accuracy using ‘tune Length’ parameter in ‘caret’. Following 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, the following 4 models from Table 3 were finalized to make predictions on the test data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Training results | | Test results | |
| Model | RMSE | Rsquared | RMSE | Rsquared |
| gradient 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 poor. And random forest seems to be good with improved accuracy on the test data set with a value of 4.48. ‘Cubist’ and ‘lm’ perform approximately similarly on the test dataset as on the training dataset.

# Conclusion

In conclusion, I merged the two data sets, one call Volume data set and the other weather data set. Analyzed the data sets by using different graphs Histograms, boxplots, correlation plot, scatterplot matrix. Understood the frequency distributions of each variable and also between the two variables. And build the predictive model using 9 different algorithms to predict the call volume and finalized the best model out of these 9 models with least RMSE value.

For next steps, I would like to add additional variables to improve the model prediction

1). Traffic Volume

2). Construction Data

3). Handle the outliers

And then predict the model with additional variables to improve the Accuracy of the model.