Draft

Ricky Heinrich & Vimaljeet Singh

2023-03-24

Introduction

The dataset we have chosen to analyze is Hadi Fanaee-T's Bike Sharing Dataset, from the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto, accessed in the UCI Machine Learning Repository. This dataset combines the Trip History Data for the years of 2011 and 2012 of 'Capital Bikeshare', which is metro Washington DC's bikeshare service, with weather data and the holiday schedule. We are hypothesizing that we are able to predict the count of bikes rented in a given hour given the predictor variables (weather, calendar date and time, etc). We are also hypothesizing that results will be better if we model casual user bike rentals separately than registered users. Finally, we want to investigate which predictor variables are most important in making the predictions.

Description of Dataset

The data consists of an aggregated count of 'rides' by hour, over the span of the years 2011 and 2012. It contains 17379 rows and 17 columns.

Variable Name	Description	Type
instant	Record index	ordinal
dteday	Date	datetime
season	Season (winter, spring, summer, fall)	categorical
yr	Year (2011, 2012)	ordinal
mth	Month	categorical
hr	Hour	categorical
holiday	Whether day is a holiday or not	boolean
weekday	Day of the week	categorical
workingday	If day is neither weekend nor holiday	boolean
weathersit	Weather conditions	ordinal
temp	Temperature in Celsius	numerical
hum	Humidity	numerical
windspeed	Wind speed	numerical
casual	Count of bikes rented by casual users	numerical
registered	Count of bikes rented registered users	numerical
cnt	Count of total bikes rented	numerical

We've chosen to remove from the original dataset the 'atemp' variable, because it had extremely high collinearity with 'temp', as well as the 'workingday' variable because it was throwing errors. We've also unscaled the temperature, humidity, and windspeed variables and stored those into 'rawtemp', 'rawhum', and 'rawwindspeed'. We are considering our full model to include the following predictors: yr, mth, hr, season, holiday, weekday, weathersit, rawtemp, rawhum, and rawwindspeed. The response variables are cnt, casual, and registered.

Table 2: Statistics of fitted linear model for Total Count

Statistic	Value
R_Squared	0.6875
MSE	9951.4900
F-Stat	447.5300

Table 3: Table of variance and mean for select times

Months	Time	Riders_Mean	Riders_Variance
December, January, February	1am - 4am	12.10479	284.7236
April, May, June	6am - 9am	238.01648	32062.5637

Plan overview

In this report, we will explore different models for predicting the rental bike count like - Linear Regression, Poisson Regression, and Random Forests (RF). We chose these models due to their ability to handle the type of response variable we have and their relative simplicity, which allows for easy interpretation (except RF). We will first fit these models on the total rental count (cnt), and then separately on the 'casual' and 'registered' rental counts. To ensure accurate predictions, we will investigate outliers in the data and use variable selection methods to determine which predictors are most important. Our dataset will be split into a 60/40 training/testing set, and we will use mean squared error (MSE) and R^2 as measures of fit.

Finally, we will present our conclusions and discuss the limitations of our models, as well as possible directions for future research.

Linear Model

In the Bike Sharing dataset, the response variable 'cnt' represents the number of hourly users of a bike. Unlike qualitative or quantitative variables, this response takes on non-negative integer values of counts.

Total Count

The table above shows the statistics of the fitted linear model. The linear model has a R^2 of 0.6875 which means the model is able to explain 68.75% variation in the count data based on the given independent variables. The F-statistic is very high which means one or more of the coefficients is significant. Most of the values that are in the model are significant. Overall, this seems to be a good fit for the model, but there seems to be an issue with the predicted values. 9.84% of the fitted values are negative which means that the linear model predicts a negative number of users during 9.84% of the hours in the data set (check 'Linear Model Fit' chart below too). The negative expected values of bikers in certain situations raises doubts about the reliability of predictions made from the regression model. It also casts doubt on the accuracy of the coefficient estimates and confidence intervals of the model. Moreover, it is plausible to assume that when the expected number of bikers is low, the variance associated with the number of users should also be small. For example, during a heavy December snow at 1 AM, we anticipate that only a few people will use a bike, and there will be less variation in the number of users during such conditions. In contrast, between 6am and 9am in summers, more number of riders are expected and hence the variance should be higher. The table below shows how these statistics vary.

Heteroscedasticity refers to a violation of the assumption in the linear model:

$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$

where the variance of the response variable (cnt) is not constant across the range of predictor variables. The most common form of heteroscedasticity in the response variable is that the variance of the response variable may change as the mean of the response variable changes. The estimate for the variance of the slope and variance will be inaccurate. Heteroscedasticity can be detected by examining the scatter plot of the data before performing the regression. We will plot a graph of mean vs variance for the 'cnt' values for both the years to inspect this.

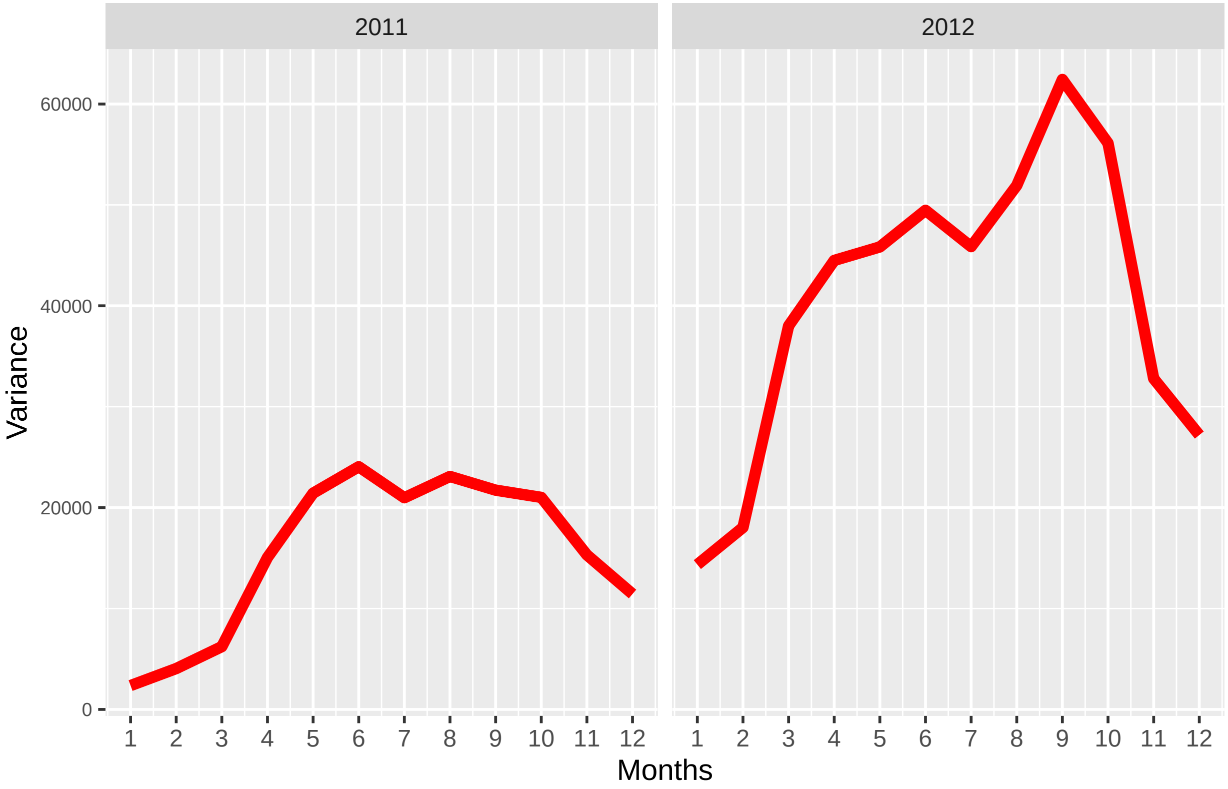

Figure 1: Variance of Total Count over 2011 and 2012

The plot above clearly shows that the variance varies throughout the year and the assumption of a linear relationship between the predictor variable and the response variable is severely violated due to unequal variance of the response variable. In fact, the variance in 2012 is visibly more too on the whole. As a result, the assumption of homoscedasticity is not met, which raises concerns about the appropriateness of using a linear regression model to analyze the data.

The following plot shows the observed values vs predicted values using linear model. Our concern here is visualzied when we see the red dots falling below the 0 on the x-axis.

We cannot have predicted count values that are negative (see the red dots below 0 in the graph below) as the number of bikes rented in an hour can never be negative. This is another reason we should not use linear model for this data. Furthermore, the response in this dataset 'cnt' is in the form of integers, while a linear model assumes that the error term is continuous. This means that the response variable in a linear model must be continuous as well. Therefore, the integer nature of 'cnt' response implies that a linear regression model may not be entirely suitable for this dataset.

Transforming the response into `log` could help us eradicate some of the problems that we are facing with linear model. We could fit something like:

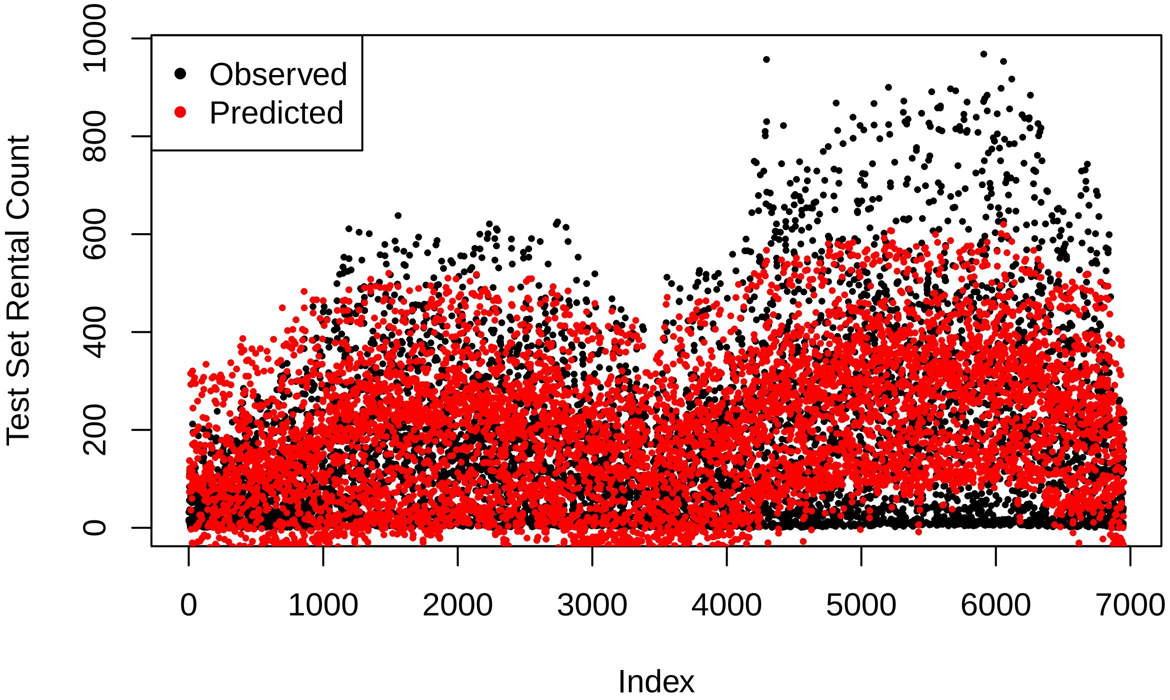

Figure 2: Linear Model Fit

Table 4: Statistics of fitted linear model for Casual Count

Statistic	Value
R_Squared	0.5920
MSE	984.6553
F-Stat	295.2100

Table 5: Statistics of fitted linear model for Registered Count

Statistic	Value
R_Squared	0.6848
MSE	7245.4982
F-Stat	442.0000

$$log(cnt) = \sum_{j=1}^{p} X_j \beta_j + \epsilon$$

Transforming the response variable in the Bikeshare data can be helpful in addressing two main issues associated with fitting a linear regression model: the occurrence of negative predictions and the presence of heteroscedasticity in the original data. By transforming the response, we can avoid negative predicted values and reduce heteroscedasticity, resulting in a more accurate and reliable model. While transforming the response variable can address some issues in fitting a linear regression model, it is not entirely satisfactory. This is because the predictions and interpretations are made in terms of the logarithm of the response rather than the response itself, which can be challenging for interpretation. Moreover, this transformation cannot be applied to data sets where the response can take on a value of 0. Therefore, although using a transformation of the response can be a reasonable approach for some count-valued data sets, it may not always be the optimal solution.

Casual Count

Registered Count

Poisson Model

The Poisson distribution is commonly employed to model counts due to several reasons, such as the fact that counts, like the Poisson distribution, are restricted to non-negative integer values. This makes it a suitable and natural choice for modeling count data.

Total Count

When using a Poisson regression to model bike usage, we make an implicit assumption that the mean bike usage in an hour is equal to the variance of bike usage during that hour. This is because the Poisson distribution is typically used to model counts, and counts, like the Poisson distribution, take on non-negative integer values. In contrast, a linear regression model assumes that the variance of bike usage always takes on a constant value. Therefore, the Poisson regression model is better suited to handle the mean-variance relationship observed in the Bike sharing data compared to the linear regression model. In fact from the table below, we can see that the variance in ‘cnt’ appears to be much higher than the mean, a situation referred to as “overdispersion” which can seemingly be handled by quasi-poisson model. We checked the results from a quasi-poisson model as well and the result is exactly the same as that of a Poisson model.

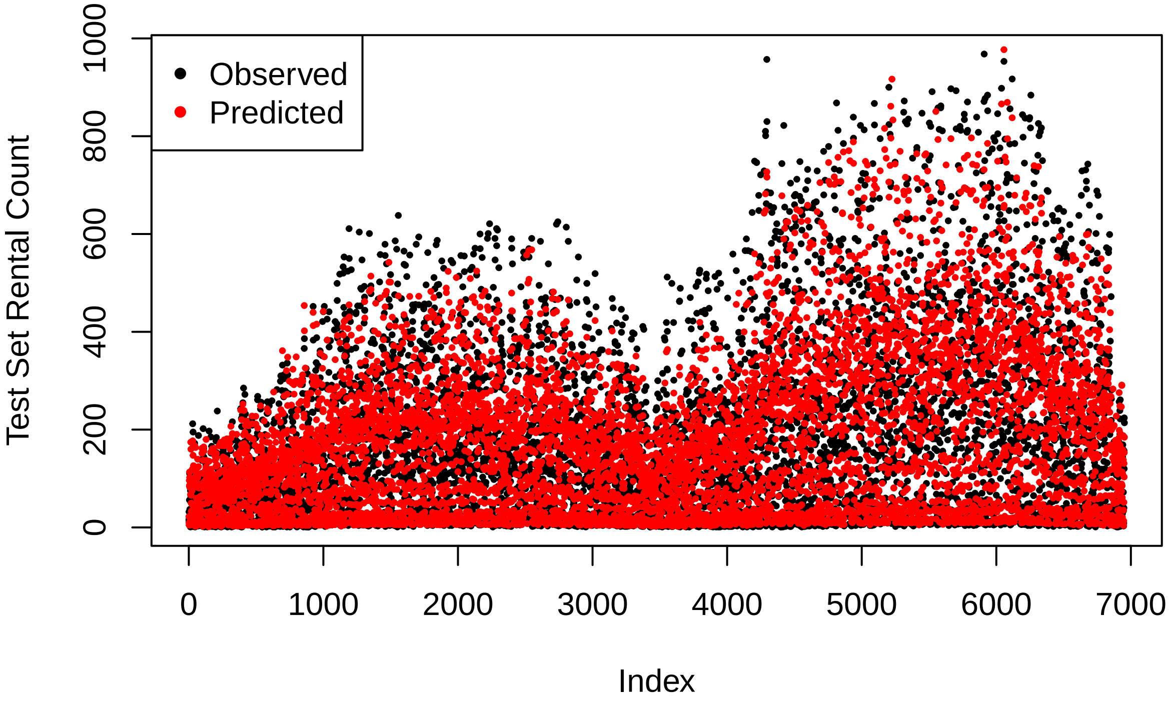

Figure 3: Actual vs Predicted Values of Test Set with GLM-Poisson model

Table 6: Table Mean vs Variance for Poisson

Year	Month	Mean	Var
2011	1	55.51	2363.97
2011	2	74.29	4048.27
2011	3	87.73	6221.97
2011	4	131.95	15056.42
2011	5	182.56	21431.03
2011	6	199.32	24050.94
2011	7	189.97	20977.03
2011	8	186.99	23089.84
2011	9	177.71	21731.45
2011	10	166.23	21014.77
2011	11	142.10	15313.78
2011	12	117.84	11436.88
2012	1	130.56	14351.25
2012	2	149.04	18032.86
2012	3	221.90	38013.55
2012	4	242.65	44495.40
2012	5	263.26	45834.14
2012	6	281.71	49466.60
2012	7	273.67	45863.84
2012	8	288.31	51927.01
2012	9	303.57	62430.32
2012	10	280.85	56122.56
2012	11	212.62	32788.45
2012	12	166.73	27190.42

Casual Count

Registered Count

Random Forest model

Because we are interested in prediction, we are ready to lose some interpretability in our model in exchange for better predictive power. Using random forest over a simple decision tree decreases the variance as well as the bias, usually resulting in closer predictions.

Total Count

As seen in Figure ##, the percent increase in MSE shows that when either hr or yr are permuted, then the MSE increases by over 150%: the values of these variables are really important in prediction. Weekday and temperature values are also important, as MSE can increase by 75 or 100% as well when they get jumbled up. All variables except for the windspeed seem to sit at %IncMSE values of over 25%, suggesting that they all carry some importance in the model. [I don't understand the plot vs the raw values, why raw temps has greater %IncMSE than yr but in plot yr comes up first?]

"mean decrease of accuracy in predictions on the out of bag samples when a given variable is permuted" "measure of the total decrease in node impurity that results from splits over that variable, averaged over all trees"

The IncNodePurity shows that the hr variable is by far the most important variable in regards to decreasing the node impurity. [What's the scale on these values idk . . .]

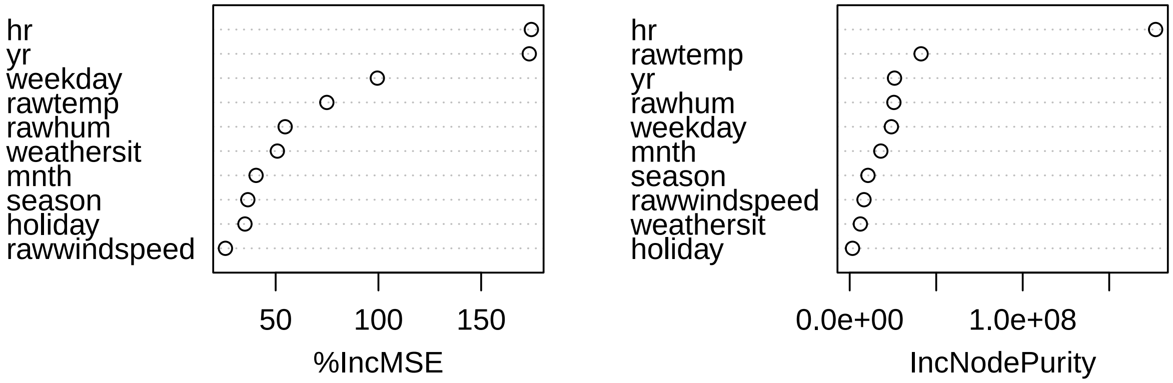

Figure 4: Variable Importance for Total Count

Casual Count

Modeling for just the Casual users, we see in Figure ## that the most important variable in terms of the mean decrease of accuracy in predictions on the out of bag samples when the variable is permuted is now

Table 7: MSEs for models

Model	MSE.Total.Count	MSE.Casual.Count	MSE.Registered.Count
Linear Model	9951.490	984.6553	7245.498
Poisson Model	8635.064	498.8537	4890.431
Random Forest	3361.434	247.0523	1983.619

weekday, at over 200%, followed by hr just below 150%, and then a year just under 100%. Still, there's a slew of variables that change over 50% in MSE when they are permutated.

In terms of IncNodePurity, the hr variable still ranks the highest, but the value is not as high as in the total cnt, and the weekday and rawtemp values follow closer behind than other models.

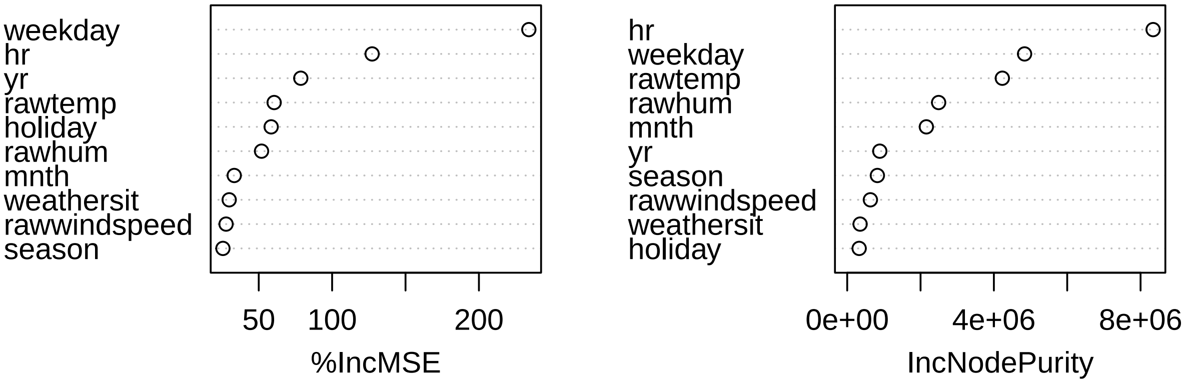

Figure 5: Variable Importance for Casual Count

Registered Count

Modeling for just the registered users follows more closely to the total count, which makes sense given that registered users take up a greater proportion of total rides than casual users. We see that hr, year, and weekday are again top variables when it comes to %IncMSE, and that hr is vastly more important when it comes to IncNodePurity.

Perfomance of all models on test set

The final draw: comparing the model performances on the test set. We've calcualted the MSE for each model, and here are the results:

We see that 'yr' is an important variable all around: it makes sense, as Capital Bikeshare began operations in 2010, and so usage growth would grow a lot in the first few years. To predict count of bike rentals per hour in future years, we would have to take into account the growth in the business and popularity. Therefore, a limitation in our data is that is only covers two years and so it is hard to forecast usage in future years.

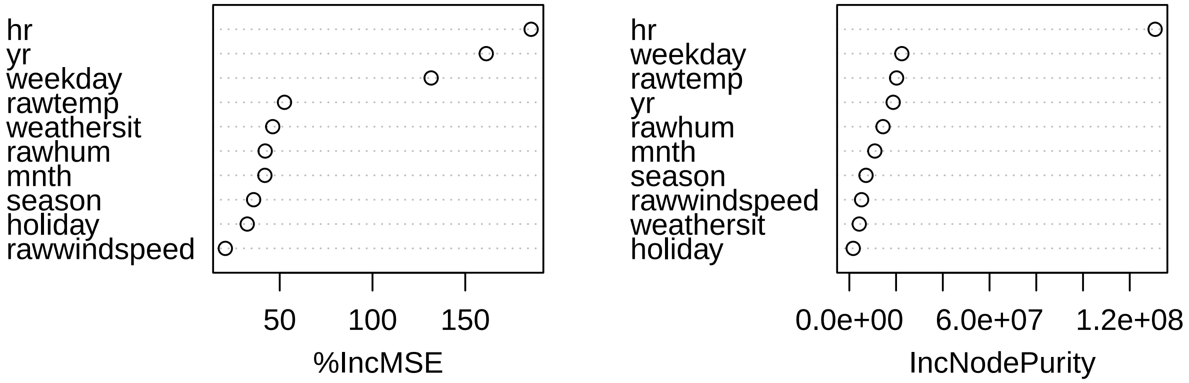

Figure 6: Variable Importance for Registered Count

Limitations

The data set that was aggregated on the UCI Machine Learning Repository excluded all the data where the count of bikes rented was zero. This is a limitation of the dataset from a data collection standpoint. The dataset which has zeros for 'cnt' variable would not fit well with log transformation of the response and hence it was excluded in case kaggle updates the dataset to include zeros for the response variable.

Random Forests models provide excellent prediction capability but they can be difficult to interpret, especially if there are a large number of trees. Even though the MSE is lowest for RF, the model is hard to interpret in the context of how increase in a unit of a specific independent variable will affect the response. Sometimes, overfitting could also be a problem with this model. Also, in some cases Random forest models tend to perform better when there are categorical variables in the data. This can be a problem if the data is mostly continuous.

The Linear Model (LM) is a good fit but the negative predictions for the count of rented bikes make us question the reliability of the model. This model would be better for interpretability as compared to RF. Furthermore, the data violates the normality assumptions of the model which limits its capacity to provide meaningful results.