SEVERITY ANALYSIS AND PREDICTION OF US TRAFFIC ACCIDENTS

DS 5110

Mrunmayi Anchawale

Pavan Choudhari

Maanasa Kaza

Zhoucheng Lin

Yashvin Jagarlamudi

**Summary**

Each year in the United States, there are an estimated 1.25 million traffic accidents leading to 2.35 million injuries and 37000 deaths. The United States government estimates that damages caused by accidents account for one percent of the country’s gross domestic product. With such a footprint on society, it is imperative to research the causes of accidents and measures to take to reduce them. This project aims to analyze the factors which contribute to public accidents and use these to predict the severity of accidents.

# The sole dataset used is a Kaggle dataset entitled *“US Accidents (3.0 million records) A Countrywide Traffic Accident Dataset (2016 - 2019)”* by Sobhan Moosavi. The dataset has entries across the last four years and spans 49 states excluding Hawaii. This dataset contains 49 features which explicitly describe the conditions of the accident such as weather conditions, nearby amenities, location, time of accident, etc. Our chief interest, however, is the variable severity. Here severity ranges from 1-4 with 1 being the least severe accident and 4 being the most severe accident with the largest impact on traffic.

To achieve a comprehensive survey of accidents and predictions of severity, several steps must be taken. Initially, the large dataset must be made more useable and condensed to perform data analysis. In addition, several models like logistic regression, sparse logistic regression, decision tree, and random forest must be trained and built. Lastly, a dashboard built from Shiny must be optimized in order to compare the performance metrics of various models to determine the premier one.

**Methods**

Primarily, data preprocessing was done before performing any data analysis of modeling. This consisted of loading the data and packages and tidying the data to remove null values. All variables with a high NA proportion of over fifty percent were dropped and the remaining variables had their NA values replaced with mean values. Afterwards, variables like severity were determined to be specified as factors in order to be used later for modeling. Severity was then further grouped from four levels into two levels. The dataset itself was consolidated to consist of only the top ten states with accidents to avoid overfitting and run the models faster.

In addition, variables deemed redundant and trivial to severity were dropped. A couple examples of this are Airport Code which gives the location of the accident and overlaps with other columns like State, City and County, and Turning Loop which contained all False values. Consequently, all near zero variance predictors were removed as they have less predictive power. The dataset was then portioned into test, training, and validation sets. The preprocessing stage was concluded by utilizing oversampling and under sampling to make the two levels, “Severe” and “Not Severe” more balanced.

Next, during the modeling phase, the models were fitted to predict severity through logistic and sparse logistic regression, decision tree, and random forest. Logistic regression was the base line model used. Using stepwise model selection and Akaike information criterion (AIC), variables were selected to get the best formula and predictions were made on the final dataset. Since there were many variables with several levels, most of them were coefficient zero when the stepwise model was built. Our sparse logistic regression used the “lasso” penalty parameter to improve upon the existing logistic regression model. As the tuning parameter increases, more variables will be forced to have coefficient zero. Upon finding the best lambda value, the prime sparse model containing predictions was made. Here, different cutoff values impacted the final performance.

On the other hand, tree-based algorithms like decision tree and random forest have a built-in feature selection to make selecting predictor variables easier. However, the decision tree model that was constructed had a high accuracy on the training set but a relatively low accuracy on the test set. This is a direct cause of overfitting the data. To remedy this situation, the random forest model used a sampling technique called bootstrapping. The number of variables from the subset were randomly selected as candidate variables using *mtry* and number of trees to be constructed through *ntree.*  The final error rate was plotted against *mtry* to get the optimal value. Furthermore, it was observed that random forest had the superior performance among the four models.

The last addition to this project was the accidents modeling and prediction dashboard built using Shiny. Shiny is a web application framework built on top of R used to build our dashboard. Interactive graphics and tables were incorporated to dynamically visualize changes to the plots made during exploratory data analysis. Toggles were added to easily compare the performance metrics of each of the models.

**Result**

**A close up of a map

Description automatically generated**

Fig. 1

Based on the heatmap above in Figure 1, it is evident that California has the highest number of accidents followed by Texas and Florida. While these are the states with higher population and higher number of registered vehicles, the frequency of accidents between California and Texas is much higher compared to the difference in population

A picture containing drawing

Description automatically generated

Fig. 2

Figure 2 demonstrates that the number of accidents peak at between 7 AM to 8 AM and 4 PM to 5 PM. This can attribute to the fact that this is the time when most people commute to either work or school.

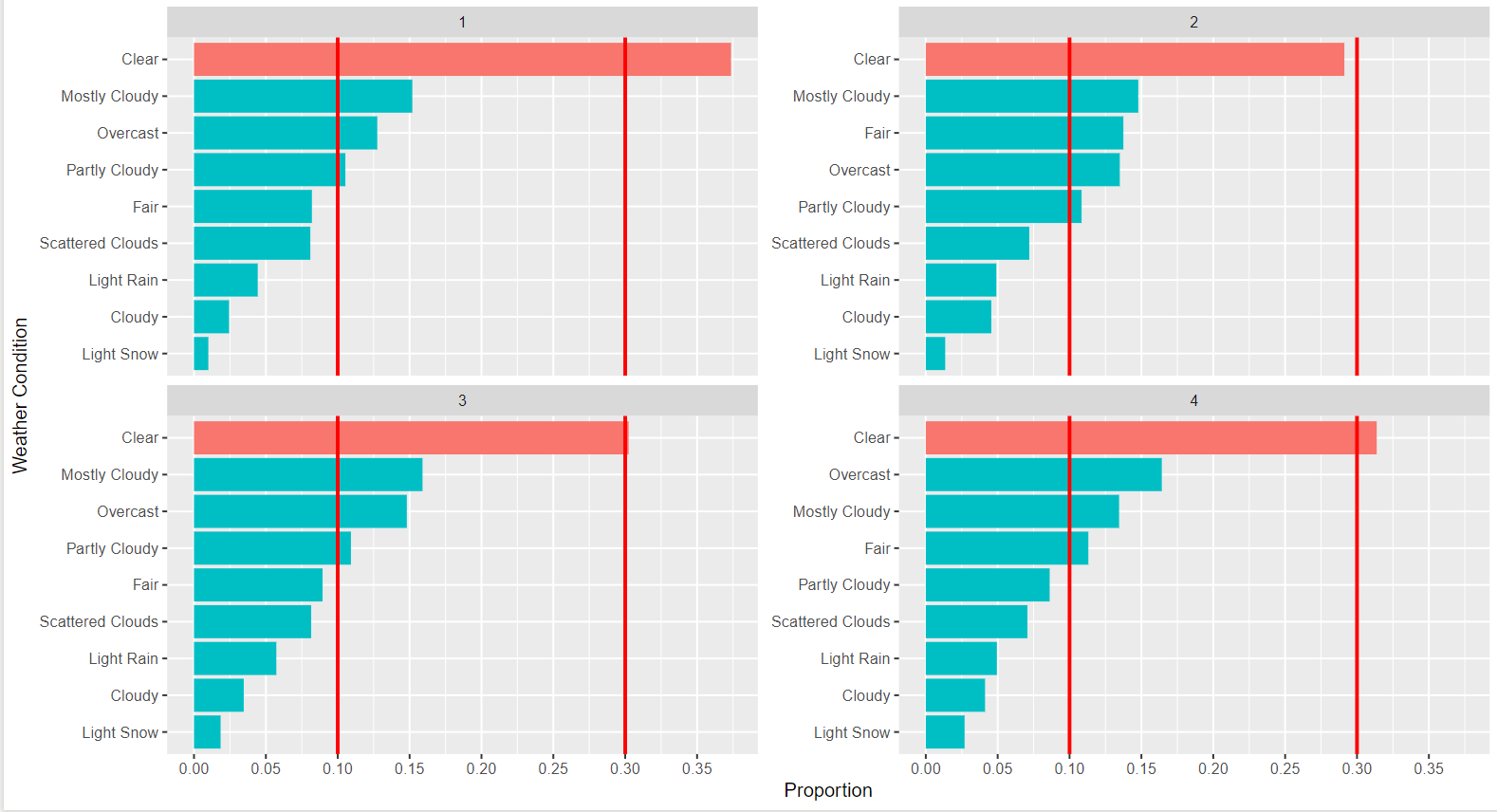


Fig. 3

According to Figure 3, an insight obtained is that most accidents occur during a clear day, and therefore, shows that weather has no discernible relationship with accidents. This is bolstered by the reference line which visualizes a higher proportion on clear days than other additions. Subsequently, other parameters like visibility, wind speed, and precipitation don’t show distinct relationships with severity of accidents when added to the model.

A picture containing screenshot

Description automatically generated

Fig. 4

Through Figure 4, distance, which is the length of the road affected by the accident, is clearly affected. Accidents having a more severe level stretch for longer distances of about 25 to 100 miles.

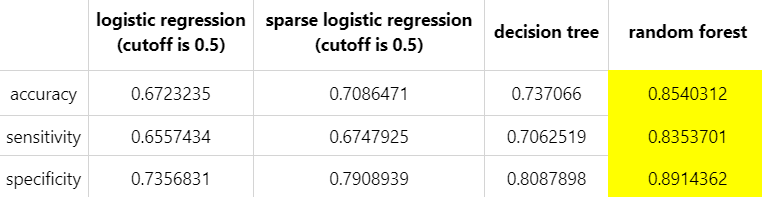


Fig. 5

Figure 5 demonstrates that random forest had the foremost performance with the greatest accuracy, sensitivity, and specificity.

**Discussion**

In conclusion, random forest was chosen as our final classification model as it gave the best accuracy of 85%. Its existing feature selection can make extending the model easier. Since the scope of the project was limited to the top ten states, a future model could include data from an additional number of states and perhaps include the missing state: Hawaii. Predicting the severity of the accident could potentially help with retrieving the estimate of overall traffic delay and assisting with traffic rerouting. Similarly, the project may be expanded to a real time accident prediction system to warn the users of possible accidents in the vicinity.

**Statement of Contributions**

**Yashvin Jagarlamudi** was responsible for data tidying and data visualizations.

**Maanasa Kaza** was responsible for data preprocessing and data visualizations.

**Mrunmayi Anchawale** was responsible for data preprocessing and data visualizations.

**Pavan Choudhari** was responsible for modeling and construction of Shiny dashboard.

**Zhoucheng Lin** was responsible for the modeling and construction of Shiny dashboard.

**References**

[1] Moosavi, Sobhan. “US Accidents (3.0 Million Records).” *Kaggle*, 17 Jan. 2020, www.kaggle.com/sobhanmoosavi/us-accidents

[2] Sarver, Cory. “Introduction to Regression and Classification in Machine Learning.” *Springboard Blog*, 18 July 2019, [www.springboard.com/blog/introduction-regression-classification-machine-learning](http://www.springboard.com/blog/introduction-regression-classification-machine-learning)

[3] Rawat, Shubhankar. “USA Accidents Data Analysis.” *Medium*, Towards Data Science, 21 Feb. 2020, towardsdatascience.com/usa-accidents-data-analysis-d130843cde02.

[4] Dyke, Ben Van. “Exploring U.S. Traffic Fatality Data.” *Oracle Data Science*, blogs.oracle.com/datascience/exploring-us-traffic-fatality-data.

[5]Santanam, Hari. “Real World Data Science Project: Traffic Accident Analysis.” *FreeCodeCamp.org*, FreeCodeCamp.org, 7 Dec. 2018, www.freecodecamp.org/news/real-world-data-science-project-traffic-accident-analysis-e5a36775ee11/.

**Appendices**

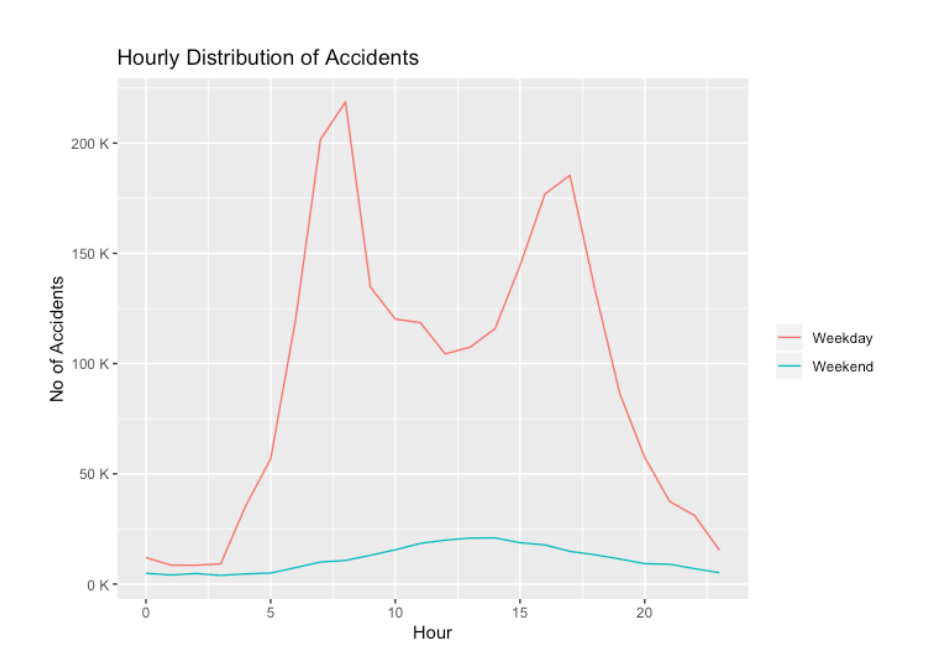


Fig. 6

A screenshot of a cell phone

Description automatically generated

Fig. 7

A screenshot of a cell phone

Description automatically generated

Fig. 8

A screenshot of a social media post

Description automatically generated

Fig. 9

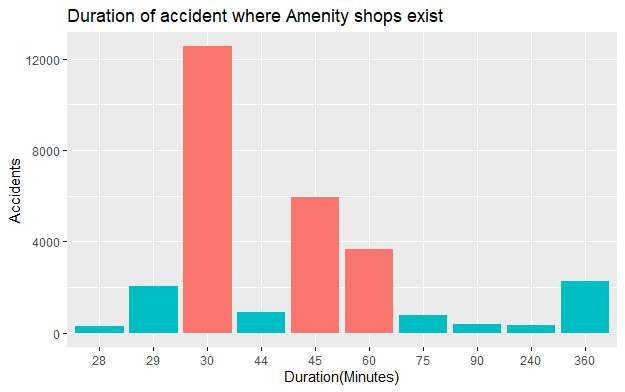


Fig. 10

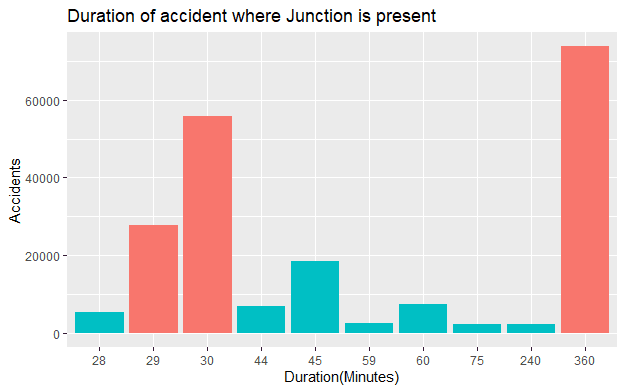


Fig. 11

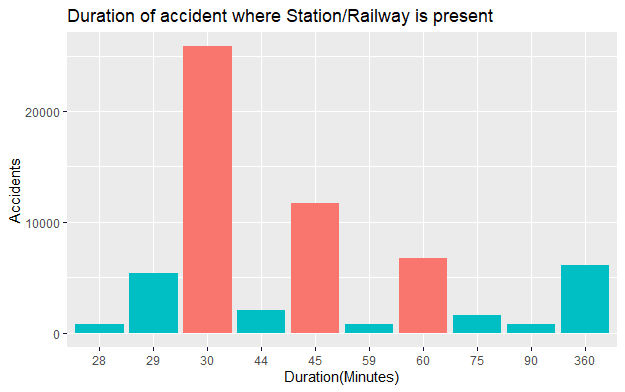


Fig. 12

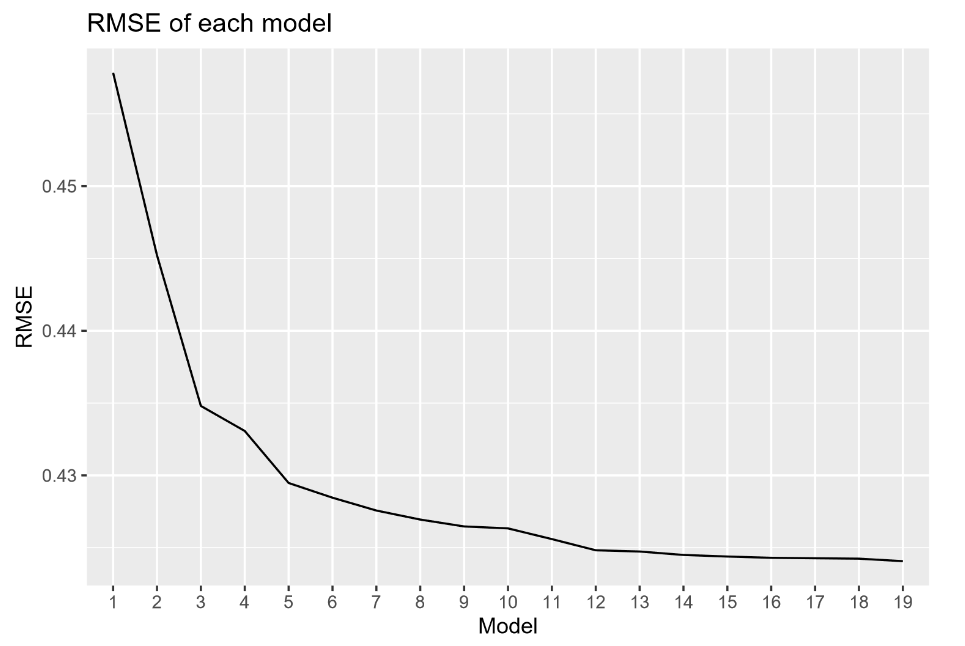


Fig. 13

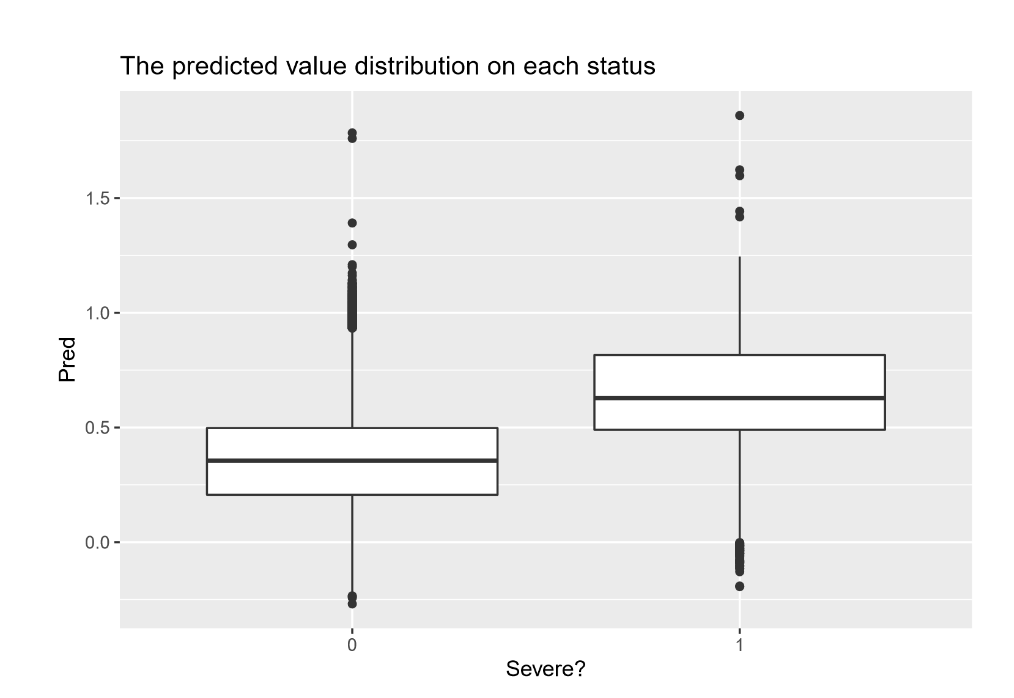


Fig. 14

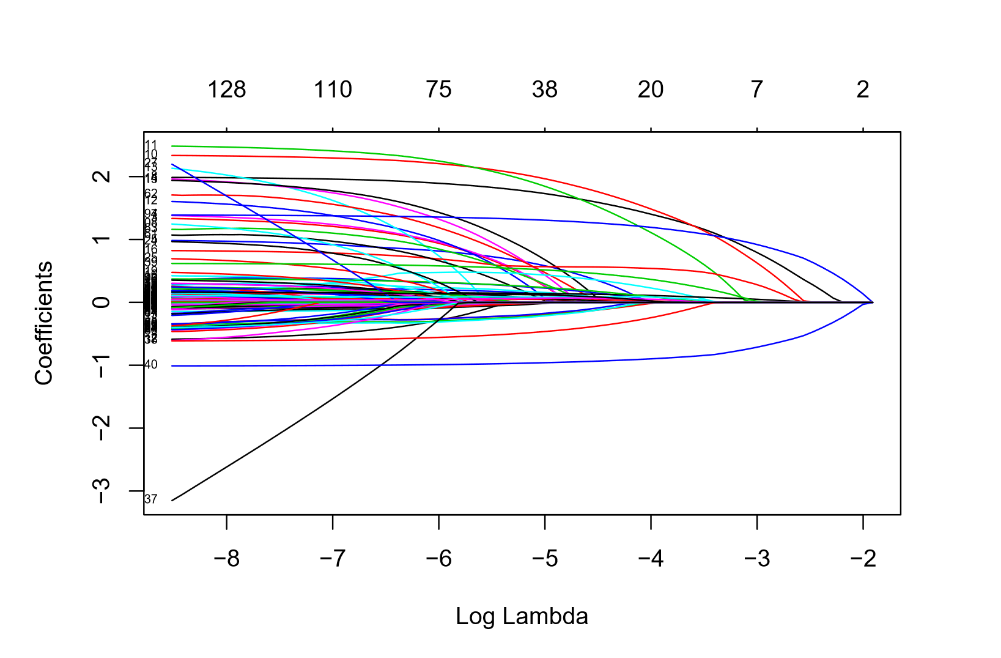


Fig. 15

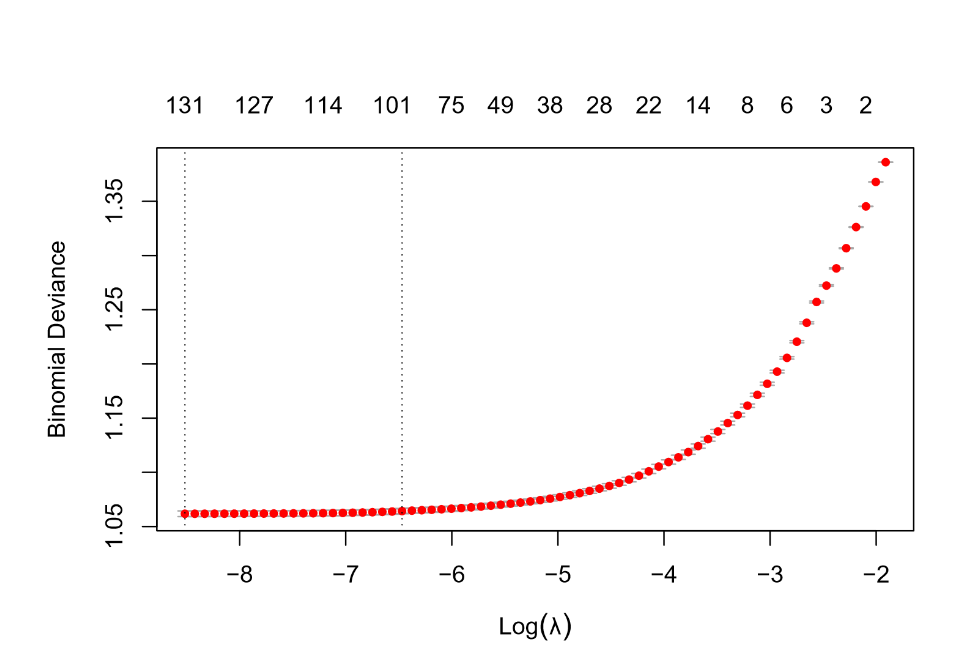


Fig. 16

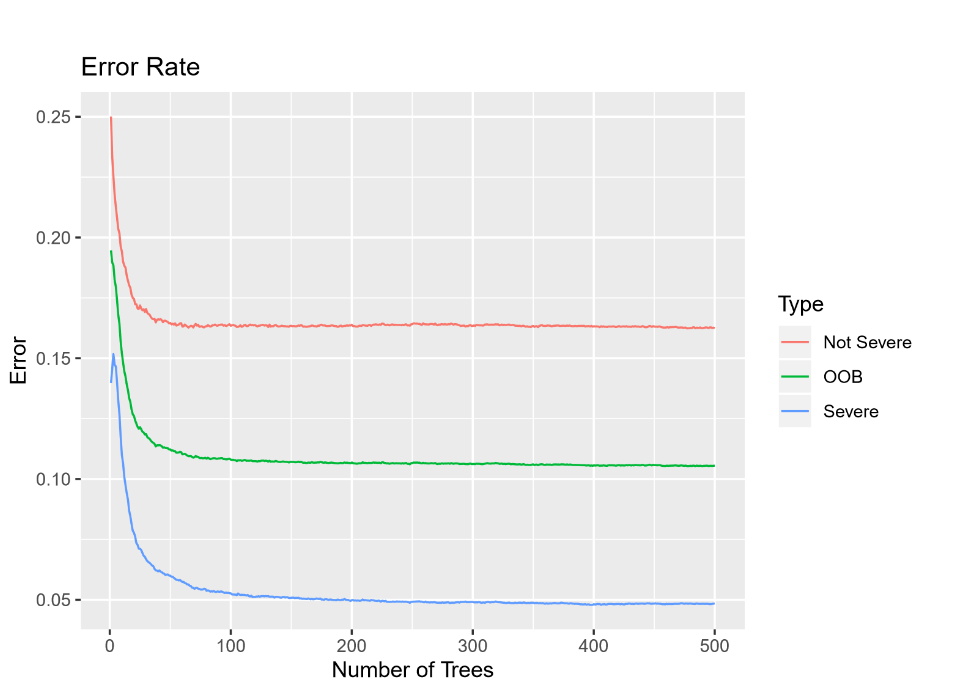


Fig. 17

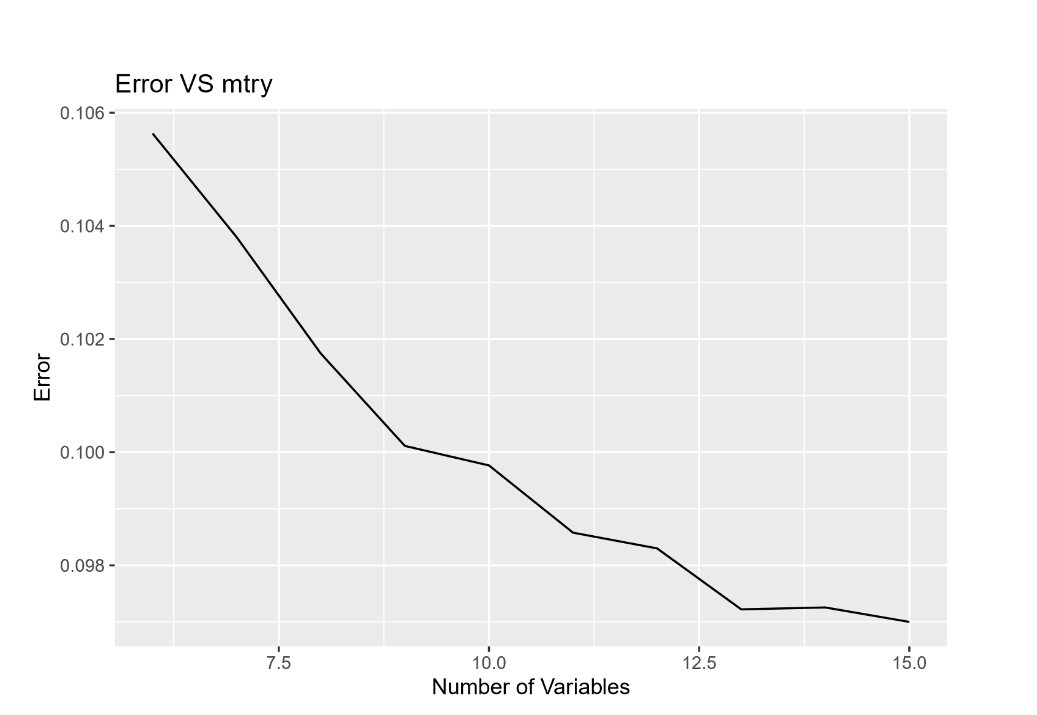


Fig. 18

