Predictive Analysis of Car Accidents' Severity

Project Summary

Names	Merve Dumlu Regina Thahir Tori Wang Xingbo Zhao
Emails	mdumlu@ucla.edu reginathahir@ucla.edu toriluwang2020@gmail.com xingbozhao@g.ucla.edu
Final Rank	10
Final Score	0.9381

Abstract

The goal of this kaggle project is to find a suitable model that could predict the severity of car accidents, and obtain information on the most important and influential predictors in an effort to prevent fatal accidents. This report provides a clear and detailed description on how we build our classification model from start to the end, including introduction, exploratory data analysis, data cleaning, feature selection, model construction, analyzing results, discussion, and limitations.

Our initial approach for the prediction was to clean the data, by removing and replacing missing and unnecessary values or variables in the training dataset. We then tried four models on our data–Knn classification, Logistic Regression, Classification Tree, and Random Forest–and compared their analyses by looking at the number of predictors involved, their Confusion Matrices, and Kaggle accuracy rates.

The final model is based on a random forest using 15 predictors. The most influential variables identified are: date, location, weather, and description keywords “blocked”, “closed” and “caution”. The model has a final Kaggle score of 0.9381 which ranks 10th place.

Introduction

There are an estimated 17,250 automobile accidents per day in the US, with most causing little to no physical harm to the people involved. However, car accidents are annually responsible for more than 46,000 people each year in the US, according to Annual United States Road Crash Statistics (ASIRT). In the last year, the National Highway Traffic Safety Administration (NHTSA) projects there were an estimated 42,915 traffic fatalities, and road traffic injuries crash injuries are now estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for youth and young adults 5–29 years of age, according to the

Association for Safe International Road Travel (ASIRT). Our task of finding a model that could predict the severity of an accident, either "SEVERE" or "MILD", is of great importance and can offer insight into how to detect the attributes of fatal car accidents and avoid them.

We are provided with a dataset "Acctrain.csv", which has 35000 observations and 44 columns (including response variable column Severity) with 32 categorical predictor variables (13 logical and 3 in datetime format) and 11 numerical predictor variables. Our objective is to be able to predict the Severity of 15000 observations in the testing dataset "AcctestnoY.csv" with as much accuracy as possible.

Methodology

1. Data Cleaning

1.1. Dealing with NAs

In our preliminary analysis, we discover that there are 14918 NA values (0.008%) in the training dataset. Since the proportion is less than 5% of the entire dataset, we decide to omit them. However, although the testing dataset has since the resulting prediction has to have the full 15000 observations, omitting the rows of the testing dataset with NAs is not an option. Although some modeling methods are able to work with missing values in the observations (e.g. surrogate splits in tree), in order to avoid having NAs in our resulting prediction, we decide to impute the missing values.

The testing missing values are as follows:

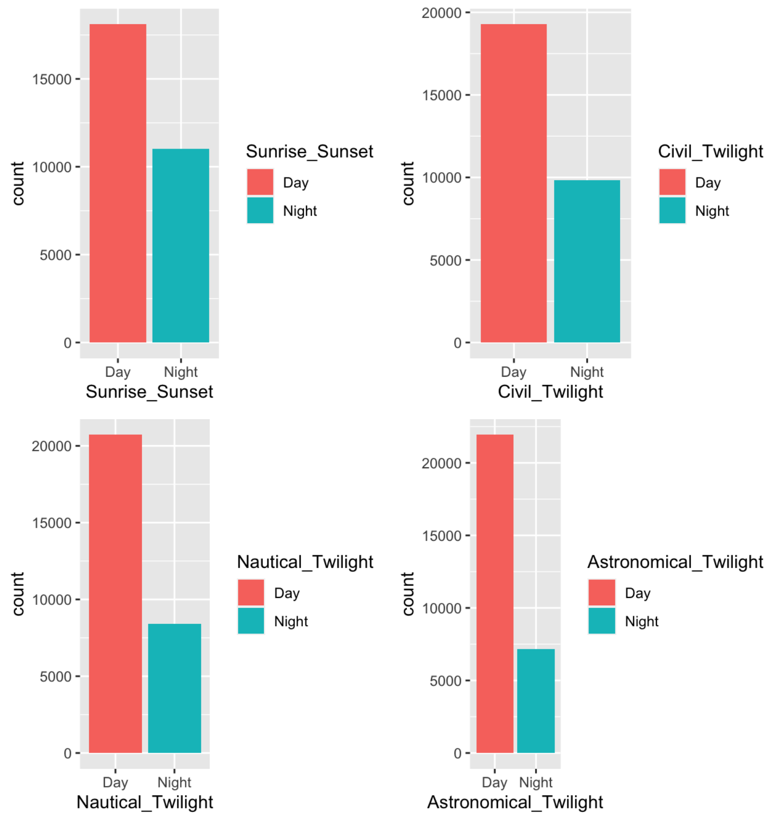
Type	Predictor	Missing Values
Categorical	Zipcode	4

Categorical	Timezone	15
Categorical	Airport_Code	36
Categorical (Datetime)	Weather_Timestamp	264
Numerical	Temperature.F.	357
Numerical	Wind_Chill.F.	2485
Numerical	Humidity	373
Numerical	Pressure.in.	312
Numerical	Visibility.mi.	358
Categorical	Wind_Direction	376
Numerical	Wind_Speed.mph.	843
Categorical	Weather_Condition	371
Categorical	Sunrise_Sunset	12
Categorical	Civil_Twilight	12
Categorical	Nautical_Twilight	12
Categorical	Astronomical_Twilight	12

For numerical variables, we decide to use the mean of the rest of the data for imputation of missing values. For categorical variables, we decide to take a closer inspection by converting all of categorical variables to factors. Zipcode, Timezone, Airport_Code, Wind_Direction, and Weather_Condition are all presented as descriptions and have too many levels as factors, hence imputing them seems counterproductive so we decide not to use these variables in our model. Meanwhile, Weather_Timestamp contains the relatively similar information of Year, Date, and Month, as Start_Time and

End_Time, and since these variables do not contain missing values, we can omit Weather_Timestamp and use Start_Time instead.

For Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, and Astronomical_Twilight, since the values are either “Day” or “Night”, we try to improve our model with imputation of missing values using random sampling from a normal distribution standardized to be between 0 and 1.



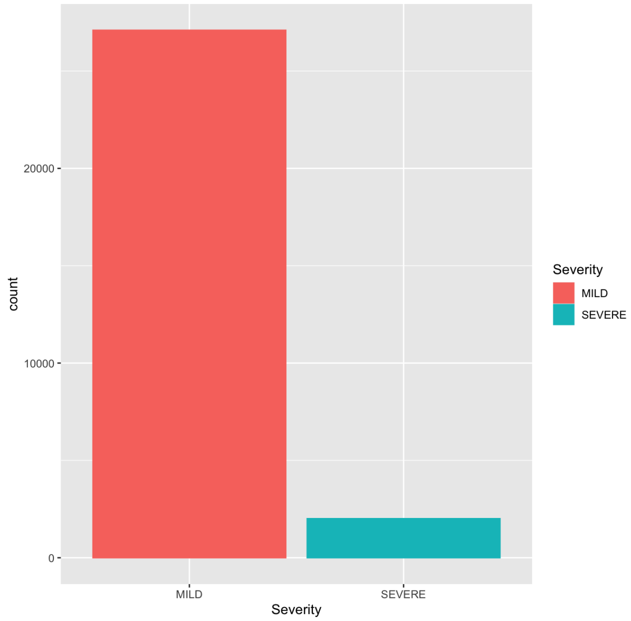
If the value of a random sample is less than or equal to the proportion of its respective variable’s “Night” in the training dataset, then the missing value would be replaced with “Night”. Otherwise, it would be replaced with “Day”. This method of imputation takes into account the fact that values are missing completely at random (MCAR) and determines to keep the distribution of categorical variables as close as possible to the training dataset in hopes of improving the testing accuracy.

2. Exploratory Data Analysis

2.1. Response/Target Variable

The target variable is "SEVERITY", a categorical variable with two categories: "SEVERE" or "MILD". From the training dataset, we obtain the proportion:

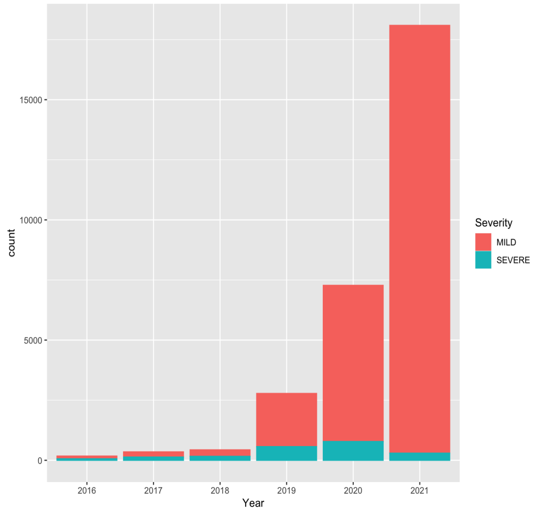
MILD	SEVERE
0.9309788	0.0690212

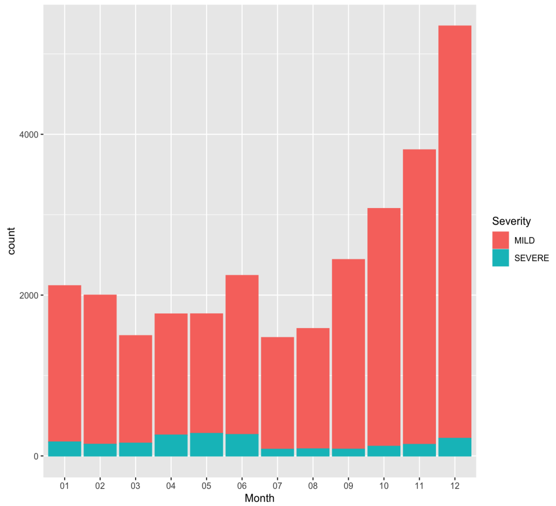


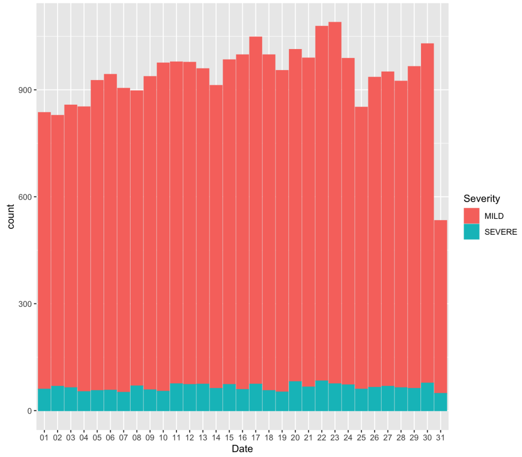
Since the proportion of "SEVERE" accidents is less than 10% while the proportion of "MILD" accidents is more than 90%, we set the misclassification rate threshold of the training dataset to be 10%. Therefore, our model's misclassification rate of predicting the training dataset's Severity has to be less than 10%. We then convert the Severity variable from character class to factor to ensure the result would only have 2 levels (SEVERE and MILD) and each observation belongs to only one or the other.

2.2. Predictor Variables

Since Start_Time, End_Time, and Weather_Timestamp variables are special characters in datetime format, we extracted variables Year, Month, and Date from Start_Time (due to the fact that Weather_Timestamp has NA values in the testing dataset and End_Time is practically analogous) to determine which could be useful in our model.



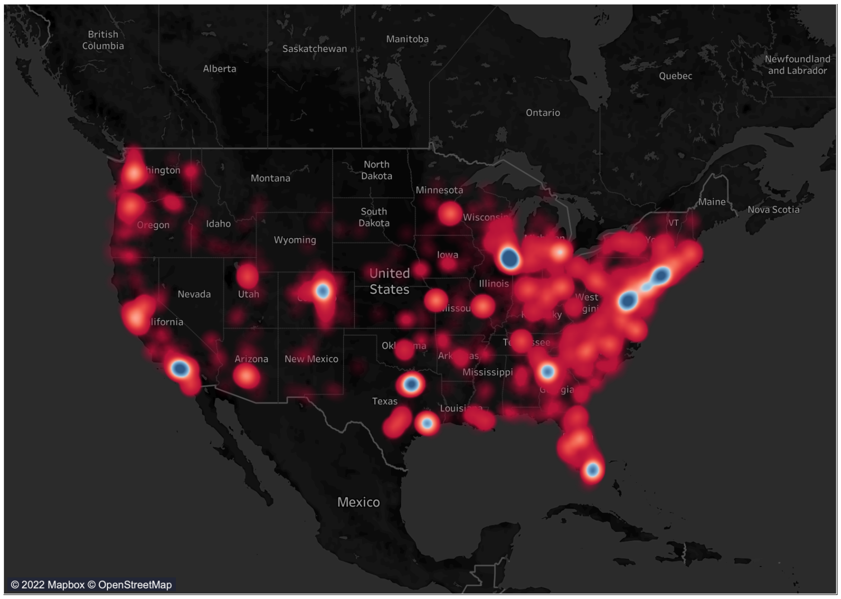




Since the severity of accidents have significant differences by year but are not significant enough to differ by month and date, we create one new variable predictor called Year.

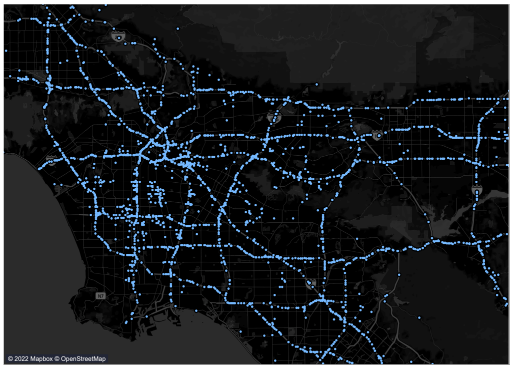
2.3 Exploring the latitude and longitude vs highway and counties

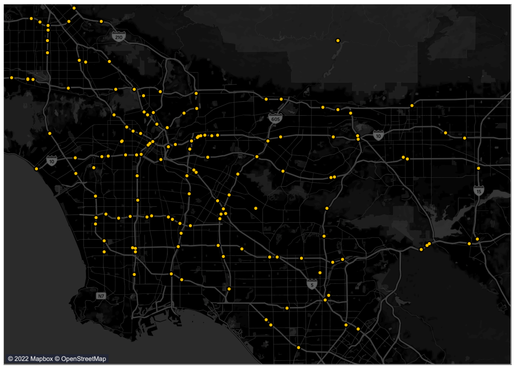
We explored the relationship between counties and longitude and latitude through a heat map. The heat map indicates that some counties have larger ratio between mild and severe car accidents.



Severity Heat Map

Our hypothesis is that severe accidents happen more oftern on highways due to higher speed. We used a Los Angeles map to compare severe cases vs mild cases. We found out that most severe cases happen on highway while mild cases happen on highways and cities.



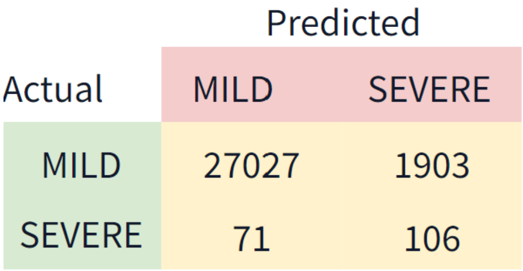


3. Modeling

3.1. Model 1 : KNN Classifier

Our first Model is a KNN Classifier. We use the Predictors: Start_Lat, Start_Lng , Junction, Temperature.F. , Wind_Chill.F., Humidity…, Pressure.in., Visibility.mi., Civil_Twilight, Caution, Blocked, and Closed with K = 40. We choose K = 40 since the testing dataset contains 15,000 observations such that the square root of n is close to K = 40, making it likely an optimal value to use.

Testing several K values shows that indeed, our K = 40 KNN classifier is the most effective. Our training confusion matrix is shown below



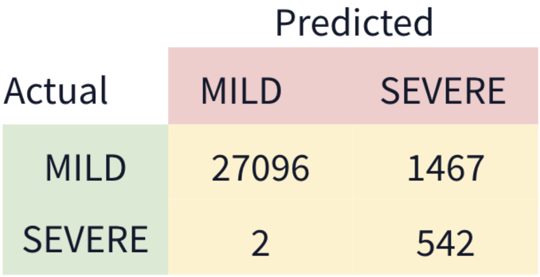
	Predicted	
Actual	MILD	SEVERE
MILD	27027	1903
SEVERE	71	106

Training Confusion matrix for KNN classifier
Error Rate: 7.27%

The KNN classifier model is a very flexible model, as it does not assume our data follows any distribution or patterns. The K value of 40 is relatively moderate, therefore we decrease the risk of overfitting to our training data, while also minimizing the risk of data generalization. Overall, our KNN model was able to pass the 9% threshold of naturally occurring SEVERE accidents, however the KNN classifier does not seem to be our best option. We move on to try another simple model.

3.2. Model 2: Logistic Regression

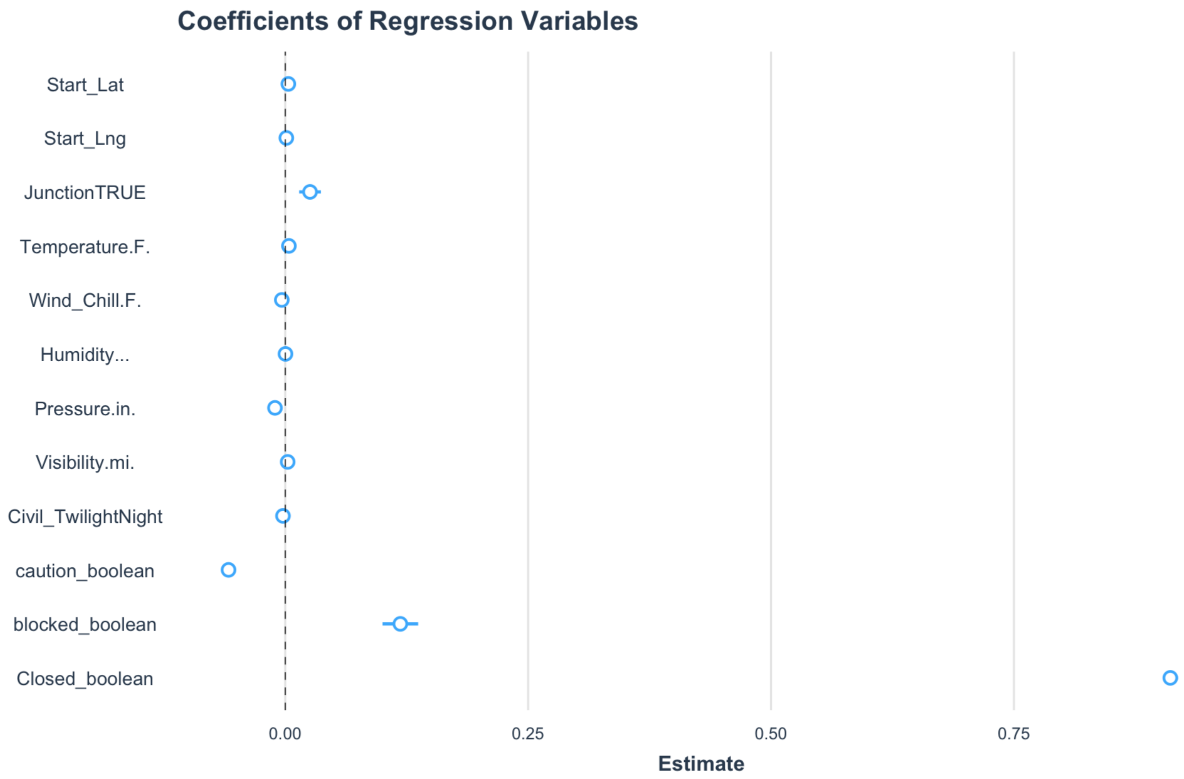
Next, we attempt a logistic regression using the same variables as our KNN classifier model used. Since the logistic regression predicts values between [0,1], we turn the Severity variable into a factor with levels 0 and 1, with 0 being MILD and 1 being SEVERE. We set out classifier conditions to label all predictions >0.5 to be SEVERE. The confusion matrix of this model is shown below.



	Predicted	
Actual	MILD	SEVERE
MILD	27096	1467
SEVERE	2	542

Confusion matrix for Logistic Regression

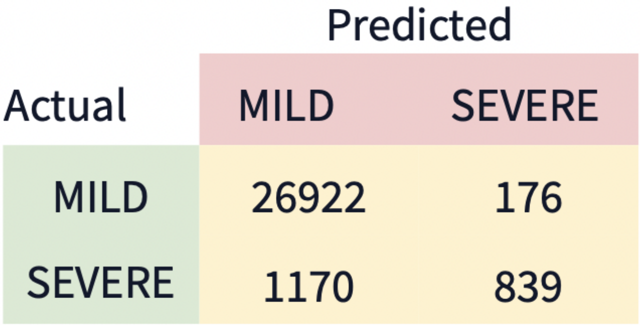
Next, we take a look at the importance of the variables used. We can do this by looking at the standard estimate values for each variable.



Based on the plot, we can observe that the most important predictors that stand out as having robust estimates are the Closed and blocked variables. Junction also has a notable effect on our model. However, we can observe that many of our estimators are not influential. This makes our Logistic model more complex without a significant difference in the increase in accuracy. Overall, our logistic model achieves a training misclassification rate of 5.05%, passing our 9% threshold. Next, we decide to look into tree classification methods.

3.3. Model 3 : Classification Tree

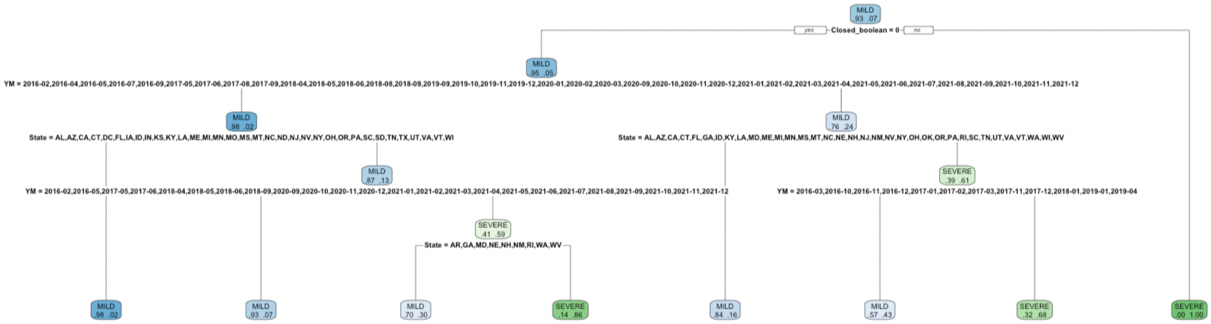
In addition, we tried the classification tree model and used all predictors. We changed character variables into factors and added in the descriptions containing “blocked”, “caution”, and “closed”. Below is the aonfusion Matrix for the classification tree model.



	Predicted	
Actual	MILD	SEVERE
MILD	26922	176
SEVERE	1170	839

Confusion Matrix for the classification tree model

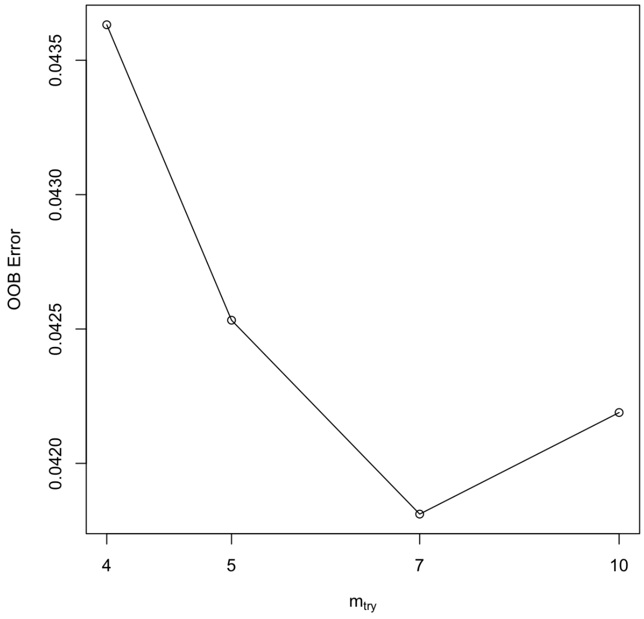
When we take a look at the result, the most significant predictors are “description” with the word “blocked”, “month”, and “states”. This matched with our hypothesis from the beginning when we explored data using Tableau.


MILD .93 .07
yes
Closed_boolean = 0
no
MILD .95 .05
YM = 2016-02,2016-04,2016-05,2016-07,2016-09,2017-05,2017-06,2017-08,2017-09,2018-04,2018-05,2018-06,2018-08,2018-09,2019-09,2019-10,2019-11,2019-12,2020-01,2020-02,2020-03,2020-09,2020-10,2020-11,2020-12,2021-01,2021-02,2021-03,2021-04,2021-05,2021-06,2021-07,2021-08,2021-09,2021-10,2021-11,2021-12
MILD .98 .02
State = AL,AZ,CA,CT,DC,FL,IA,ID,IN,KS,KY,LA,ME,MI,MN,MO,MS,MT,NC,ND,NJ,NV,NY,OH,OR,PA,SC,SD,TN,TX,UT,VA,VT,WI
MILD .76 .24
State = AL,AZ,CA,CT,FL,GA,ID,KY,LA,MD,ME,MI,MN,MS,MT,NC,NE,NH,NJ,NM,NV,NY,OH,OK,OR,PA,RI,SC,TN,UT,VA,VT,WA,WI,WV
MILD .87 .13
YM = 2016-02,2016-05,2017-05,2017-06,2018-04,2018-05,2018-06,2018-09,2020-09,2020-10,2020-11,2020-12,2021-01,2021-02,2021-03,2021-04,2021-05,2021-06,2021-07,2021-08,2021-09,2021-10,2021-11,2021-12
SEVERE .39 .61
YM = 2016-03,2016-10,2016-11,2016-12,2017-01,2017-02,2017-03,2017-11,2017-12,2018-01,2019-01,2019-04
SEVERE .41 .59
State = AR,GA,MD,NE,NH,NM,RI,WA,WV
MILD .98 .02
MILD .93 .07
MILD .70 .30
SEVERE .14 .86
MILD .84 .16
MILD .57 .43
SEVERE .32 .68
SEVERE .00 1.00

3.4. Final Model : Random Forest

Since the classification tree does better than all of the other models, using the same idea, we seek to improve the model with Random Forest. First we try using the full train model without "Start_Time", "End_Time", "Description", "Street", "City", "County", "Zipcode", "Airport_Code", "Weather_Timestamp", "Weather_Condition", "Starting_Date", and “Country”.

In terms of the tuning parameter, we use ntree = 128[1] instead of the default 500 due to the fact that using less ntree lessens model complexity and time complexity of the function run. For subset of predictors tried at each split, we use the function tuneRF and obtain the following result:

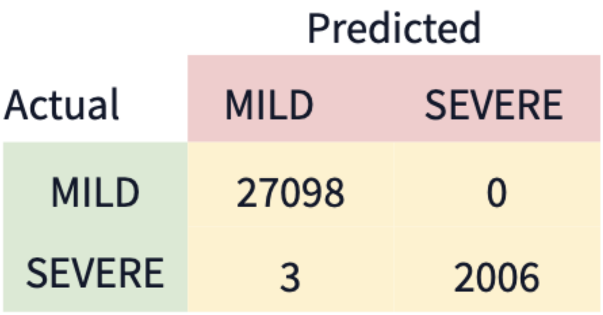


We set mtry = 7 as our tuning parameter since OOB (Out-Of-Bag error) is estimated to be lowest at that point, which is consistent with the fact that the value of mtry is approximately equal to the square root of the number of predictors.

OOB estimate of error-rate : 4.26%

Confusion matrix for training :

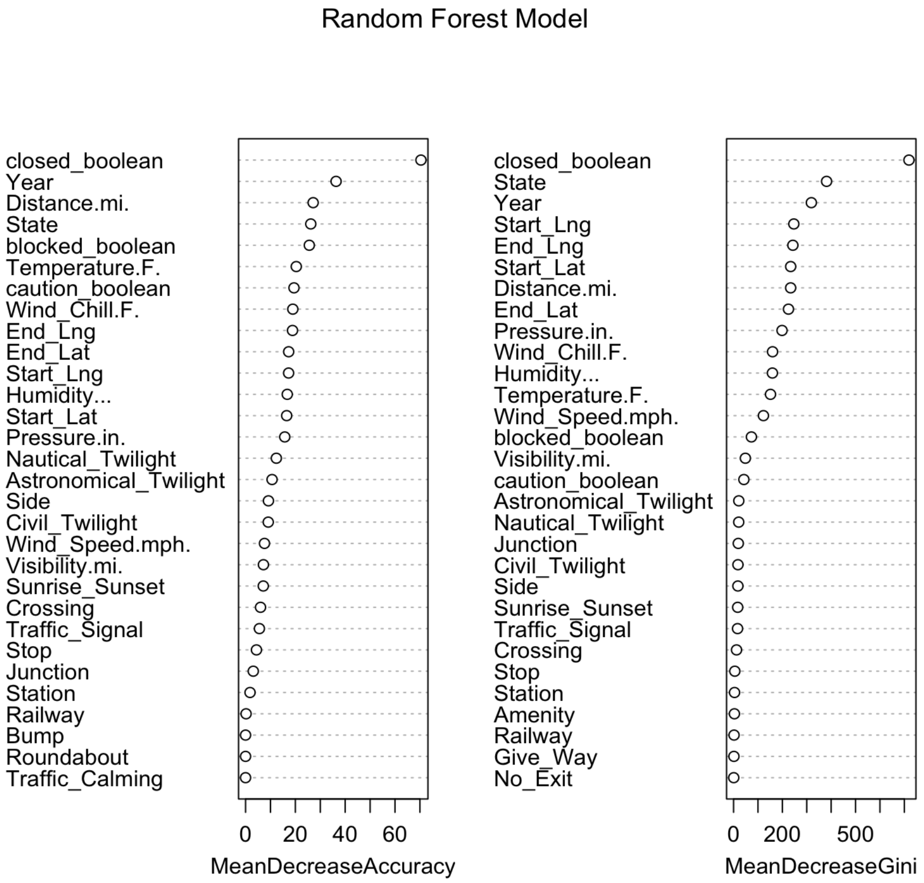
[1] Oshiro, Thais Mayumi et al. "How Many Trees in a Random Forest?" *IAPR International Conference on Machine Learning and Data Mining in Pattern Recognition* (2012)



	Predicted	
Actual	MILD	SEVERE
MILD	27098	0
SEVERE	3	2006

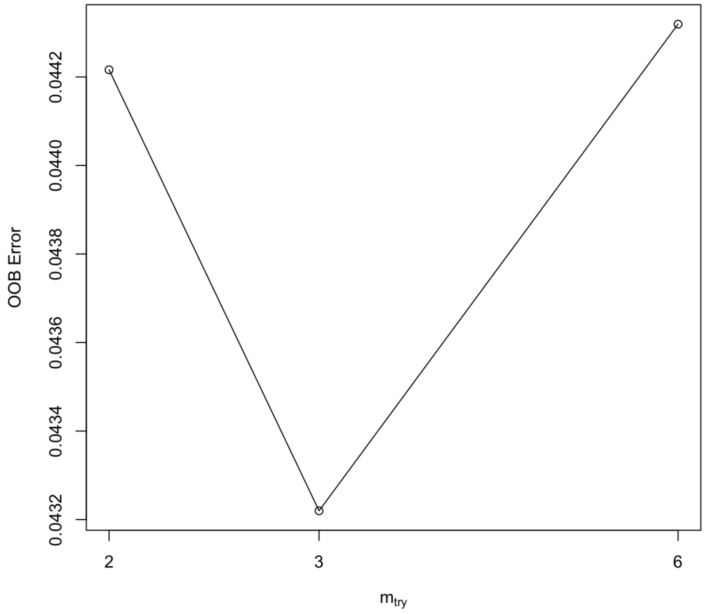
Training Misclassification Rate : 0.000103

Variable Importance Plot



From the Variable Importance Plot, we can see that a lot of the variables from the full training dataset are not significant to our model, as seen from the fact that their Mean Decrease Accuracy are very close to 0. Thus, we decide to improve our model by subsetting only 15 of the most important predictors in terms of Mean Decrease Accuracy, i.e. “closed_boolean”, “Year”, “Distance.mi.”, “State”, “blocked_boolean”, “Temperature.F.”, “caution_boolean”, “Wind_Chill.F.”, “End_Lng”, “End_Lat”, “Start_Lng”, “Humidity…”, “Start_Lat”, and “Pressure.in..”.

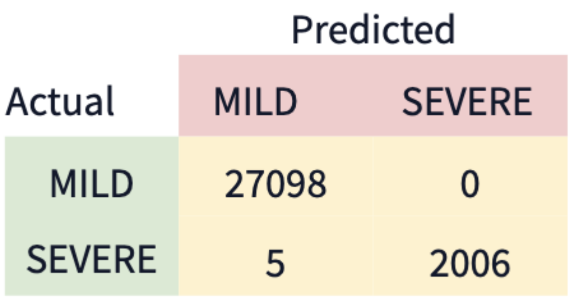
We keep ntree to be 128 but tune our mtry using the same function.



Therefore we use mtry = 3 as our tuning parameter.

OOB estimate of error-rate : 4.34%

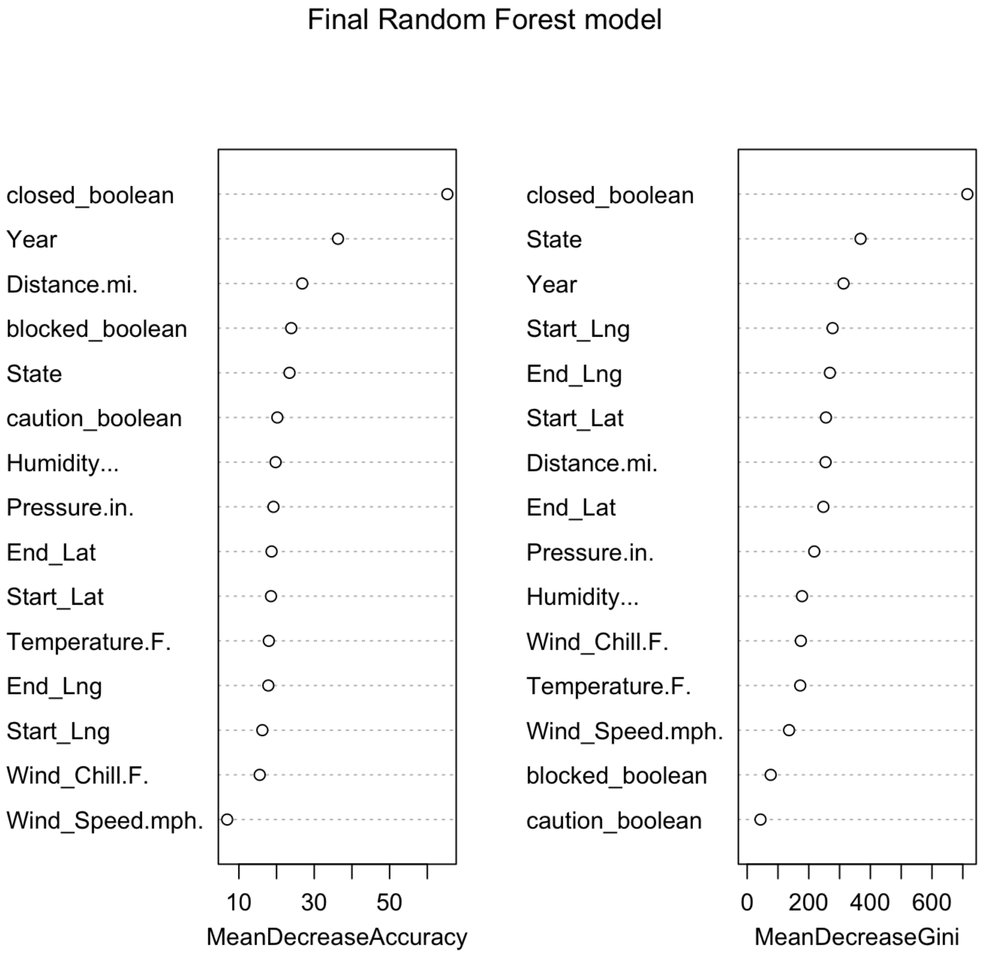
Confusion matrix for training :



	Predicted	
Actual	MILD	SEVERE
MILD	27098	0
SEVERE	5	2006

Misclassification Rate : 0.0001718

Variable Importance Plot



Since the performance of this subsetted Random Forest model with 15 predictors is much better than all the other models, we conclude that this is our best and final model.

Results & Discussion

In studying the US car accident data, we found out that the severity level of car accidents may be influenced by a number of predictors. According to our research and analysis, location, time, and scene description are some of the most important predictors. Through our initial data exploration, we strongly suspect that the location of accidents may be one of the primary causes of severe car accidents. The fact that some counties have more severe cases on the heat map and that severe cases mostly happen on highways suggests that severe cases happen more frequently when high speed and large population occur. Research shows that greater severity reaches 10% at an impact speed of 17.1 miles per hour (mph), 25% at 24.9 mph, 50% at 33.0 mph, 75% at 40.8 mph, and 90% at 48.1 mph. (Tefft 2013). Through data mining, we found that key words like “caution”, “blocked”, and “closed” play a significant role in predicting the severity level of car accidents.

Limitations and Conclusion

We assume that imputing the missing values within the numerical variables with their column means could have skewed and changed our final prediction. Using the column median for imputation could have given a slightly better and more accurate prediction than the mean since the median would be a value more likely observed in the dataset. The logistic regression and KNN classification models were both not very effective because most predictors were not as influential to the models. A reason behind this could be that the training dataset is large and models like KNN do not perform well with higher dimensions. Moreover, feature scaling is needed for the KNN model and the selection of the appropriate k-value has significant effects on the prediction. Our chosen random forest model is highly complex and also computationally intensive since the training time is more compared to other models. Additionally, another limitation with the random forest model is that, by definition, it bootstraps random samples of the dataset to choose a random sample of m predictors, which leads to the result being slightly different for each iteration.

Analysis of the confusion matrix of our chosen random forest model's prediction depicts that all mild cases were predicted accurately, while there exists severe cases that were falsely predicted as mild. However, the severe cases that were falsely predicted are very low, making only 0.017% of all predictions, meaning that the random forest model gives a perfect prediction for mild car accident cases and a close to perfect prediction of severe car accidents.

Our model's final Kaggle score of 0.93857 passes the 0.9 threshold of cases that are severe by nature. Through our analysis we were able to gain valuable insight into the predictors of classifying cases including interesting finds within Description and Timestamp variables. Further work can be done to analyze some of the more complex variables such as county, street, zip code, and city, as these may provide further insights, but for now it is not considered in the scope of this project due to the time restriction.

References

Brian C. Tefft, Impact speed and a pedestrian's risk of severe injury or death, Accident Analysis & Prevention, Volume 50, 2013, Pages 871-878, ISSN 0001-4575, https://doi.org/10.1016/j.aap.2012.07.022.

Almohalwas, Akram.