Prediction of Road Accident Severity at Los Angeles

Aravindh Siddharth Prabaharan, Abinesh Senthil Kumar,
Venkata Mohan Vamsi Chunduru

December 2019

1 Abstract

This paper presents the prediction of severity of road accidents at Los Angeles with respect to weather data and other traffic events using statistical and machine learning models. Country wide traffic accident dataset by Sobhan Moosavi [24] is used to develop five models in R and as a benchmark for comparison we have considered null model accuracy. We found the factors which has the highest impact on occurrence of accident and predicting the severity of road accidents that are likely to occur will help in taking preventive measures such as setting up amenities.

2 Introduction

Over 37000 people die in road crashes each year in addition to 4.35 million injured or disabled. Los Angeles a popular destination among tourists has faced accidents costing to 20 billion USD per year and has been on a constant rise where steps are taken to predict and halt such incidents. But severity of road accidents has been one of the least studied area though they have a bigger impact on the casualty rate. Recently machine and statistical learning has gone hand in hand predicting solutions to mysterious questions. With advancements in technology, the process of predicting an event based on a given data set after training has led to important findings. This project aims to answer questions related to severity of road accidents in Los Angeles relating to weather and other contributing factors and take appropriate measures to predict the severity of road disasters. The data "A Countrywide Traffic Accident data set" by Sobhan needed for predictive analysis is obtained from smoosavi.org, an online data set repository. Each accident record comprises of a multitude of inherent and contextual characteristics such as place, time, description of the natural language, climate, time and interest points. This project seeks to contribute the knowledge of accidents motorway accidents by creating models such as logistic regression and other classification modules. The expected outcome from

this predictive analysis is to predict the level of severity for an accident namely "Low" and "High" based on previous accidents with predictors of weather and other factors.Here, severity refers to the fatality of accidents with the underlying assumptions that more the traffic delay higher the severity.The coordinates available in the data set can be helpful in predicting further instances of accident at a given locality and various environmental factors influencing them. Predictions based on nearby traffic signal to considerations to enhance the predicted model. The predicted data with "high severity" can be helpful in setting up hospitals or amenity stations near the accident hot spots and thereby significantly reduce the fatality rates.

3 Literature Review

Few of the previous works in this topic include works done by Ren et al [8] using a long-term memory model to predict the frequency of accidents using traffic flow, past traffic accident data, and time data. Yun et al [10] predicted the frequency of traffic accidents using several sources of traffic volume, road condition, rainfall, temperature and satellite images. They evaluated their model using large-scale data of traffic accidents from Iowa state, the performed predictions showed the importance of capturing heterogeneity and temporal trends for effective predictions of traffic accidents. Studies by JD Tamerius, X Zhou showed that precipitation has played a significant role in vehicle accidents. The relative accident risk was low during the summer and rose up for frozen precipitation. These studies showed small changes in temperature and precipitation type can have a severe effect on the crash.
The paper also highlighted weather being an important contributor to accidents. Paper by Ciro, Alessandra has demonstrated a model for the detection of accidents in multilane rural roads under traffic flow, geometric infrastructure characteristics. The study was performed in four-lane median-divided Italian motorway and was monitored for 5 years. The result helped designers in adjusting multilane roads under poor weather conditions.[6]

3.1 Studies Performed in Classification

Prediction on the frequency of accidents by Li-Yen Chang and Wen-Chieh Chen. The scope of this study is to predict the number of expected accidents for a specific geographical segment [3]. Early work in this category is performed by Chang et al. He utilized information such as road geometry, annual average daily traffic and climate data to predict the accident frequencies for a highway using a CART (Classification and Regression Tree). Additionally, Caliendo et al [1] used a set of road-related attributes such as length, curvature, sight distance and presence of junction to predict the frequencies of accidents by a convolution neural network model using large scale accident and imagery data was proposed by Najjar et al [7].
Change compared the performance of an artificial Neural Network with that

of a negative binomial regression model over 1338 accidents.ANN [2] achieved 61.4% and 64% accuracy for testing and training respectively. Chang at el [18] also applied the decision tree model on the same data set to predict high way accidents. Since the testing and training accuracy were less than 55%. Olutayo et al [20] employed decision tree and ANN model on a data set from Nigeria and got precision and recall both around 0.52. Lin et al[5] applied the FP tree to select features that are more likely to contribute to the prediction. Later they used Random Forest, K – Nearest neighbor and Bayesian Network to predict accidents along the same road which resulted in 61% accuracy.
Abella et al [13] used a probabilistic Neural network (PNN) model to predict traffic accidents based on real-time road conditions and achieved an accuracy around 70%. The drawbacks of these analysis performed by them is, most of them only used one model and have not addressed key issues such as class imbalance problem and spatial heterogeneity.
In the study of prediction on the probability of Accident, the prediction is defined as a binary classification task that fits real-time applications approximately. In this analysis, they chose an Interstate Highway as their scope for the project i.e I -64 Virginia (US)[5]. Later, they leveraged a decision tree model to separate pre-cash records from normal ones, using information such as visibility, traffic volume, climatic conditions, and occupancy information. However,their limited scope of data might weaken their resulted analysis. In another study by Chen at al[19] utilized human mobility data in terms of 1.6million GPS records and a set of 300,000 accident records in Tokyo to predict the probability of accident analysis on an hourly basis. In their analysis, they leveraged a stack denoising auto encoder model to get the latest features from human mobility and used a logistic regression model to predict accidents.
In another project, Yuan et al[12] used a heterogeneous set of urban data such as road characteristics, radar-based rainfall data, temperature data, and demographic data to predict the probability of accident for each road segment in the state of Iowa. They leveraged Eigen analysis to capture and represent spatial heterogeneity. Their analyses and results suggest the importance of time, human factors, weather data, and road network characteristics for this data.

3.2 Studies Performed in Numeric Prediction and Correlation Analysis

These groups of works aim to fit regression or other models to predict the number of traffic accidents on specific roads or in certain regions. Many studies focus on identifying the correlation between attributes like weather, accident risk, and road accidents. Chen et al [19] developed a learning model to predict the traffic accident risk level using human mobility data. Caliendo et al[15] developed Poisson, Negative binomial, and negative multinomial regression models to predict the number of accidents on given roads. Oh et al[23] employed a zero-inflated Poisson regression model to predict the number of crashes at railway highway intersections and identified the correlation between several factors and crash rates. Begel et al [14] employed an autoregression regressive model to study the

correlations between weather attributes and injury incidents.
Eisenberg et al[21]used a negative binomial regression model to study the relationship between monthly precipitation and fatal crashes. In the case study of Predicting traffic accidents through heterogeneous urban data by Zhuoning Yuan, James Tamerius, Xun Zhou, Ricardo Mantilla, Tianbo Yang, they investigated the problem of traffic accident prediction using heterogeneous urban data, an important issue to transportation and public safety. They formulated the problem as a binary classification problem. They obtained and map matched fine-grained datasets such as all the motor vehicle crashes in Iowa from 2006 – 2013 detailed road network, and hourly weather data, and evaluated the performance of several classification models. They incorporated road network connectivity relationships into the classification process using eigenanalysis. From their model, they improved DNN accuracy and AUC to 0.9512 and 0.9612, respectively.
In the paper Traffic Accident analysis using machine learning paradigms by Miachong, Ajith Abraham,Marcin Paprzycki,[17] analyzed the GES automobile accident data from 1995 to 2000 and investigated the performance of the neural network, decision tree, support vector machines, and a hybrid decision tree. The classification accuracy obtained in their experiments revealed a hybrid approach performed better than a neural network, decision trees and support vector machines. According to their analysis for no injury classes, the hybrid approach performed better than the neural network. The no injury and the possible injury classes could be best modeled directly by decision trees. The ability to predict fatal and nonfatal injury is very important since driver fatality has the highest cost to society economically and socially and their experiments showed that model for fatal and nonfatal injury performed better than other classes.
In the Review of accident prediction models for road intersections by BB Nambussi, T Brijis, and E Hermans, their main objective is to derive the most appropriate technique and significant explanatory variables in assessing the safety of road intersections. In their study, they analyzed that multiple regression model has some limitations to describe adequately the random, discrete and non-negative accident events. This limitation includes the presence of undesired statistical properties such as the possibility of negative accident counts and the lack of distributional properties such as the condition of normally distributed accident occurrence. In their analysis, they proposed that Poisson regression models are more suitable for this kind of analysis than multiple linear regression models as the accident occurred is unavoidably discrete and more likely random events. But these types of models have potential problems like the mean must be equal to the variance. If that assumption is not satisfied that is the accident data are significantly over dispersed.
To solve the problem of overdispersion the negative binomial distribution is selected instead of the Poisson model. But the negative binomial model ignores the correlation and treat within the intersection accidents the same as between intersection accidents, thereby producing biased results. So, it is suggested to consider techniques which adjust for variation in accident due to locations. Besides, multiple logistic regression, multiple linear regression, negative binomial

and Poisson regression models assume independent residuals across units. So, for a single solution to all these problems random effects model is selected which accounts for correlation within clusters. So, the Random-effects model is a plausible choice if data are serially correlated. Finally, after performing all necessary analysis, they would prefer a model for intersections to be of this form

$$\mu_i = \beta_o * Q_{MI}^{\beta_1} * Q_{MA}^{\beta_2} * e^{\sum \beta_j x_{ij}}$$

where μ_i = expected number of accidents at intersection type i
Q_{MA}= number of vehicles entering an intersection from the major road
Q_{MI} = number of vehicles entering an intersection from the minor road
x_{ij} = vector of explanatory variables, j, other than traffic flow on intersection i
β_o = intercept
$\beta 1, \beta 2$= effect on traffic volume on the expected number of accidents and is modeled as elasticity.
Since there is sufficient studies progressed in the field of accidents, those are focusing on predicting the number of accidents, frequency of the accidents, factors that can cause accidents and probabilities of accidents, but the prediction for the rate of severity have been a question. This work will address the prediction of the severity of an accident. In this study, we analyzed that Random forest model and The regression tree model will best suit our analysis of our data set.

4 Data Cleaning and Transformation

4.1 Data Collection and Preprocessing

The data was collected from Kaggle and Soban Moosavi [24]. The table below shows the original list of predictors used in the obtained data set. Predictors were filtered since most were redundant to the response "Severity". Moreover, 243 rows of NA values were removed as they contributed to less than 0.5% of the final data. The highlighted cells in the table shows the filtered list of final predictors used in modelling.
Red-Response Variable
Purple- Predictors used along with response
Green-Used to plot accident point in Exploratory Data Analysis

Table 1: List of Original Predictors.

Column Name with Description	
ID	Unique identifier of the accident record
Source	Indicates source of the accident report
TMC	Traffic accident may have a Traffic Message Channel
Severity	Severity number between 1 and 4
Start_Time	Start time of the accident in local time zone
End_Time	End time of the accident in local time zone
Start_Lat	Latitude in GPS coordinate of the start point

Table 1: List of Original Predictors.

Column Name with Description	
Start_Lng	Longitude in GPS coordinate of the start point
End_Lat	Latitude in GPS coordinate of the end point
End_Lng	Longitude in GPS coordinate of the end point
Distance(mi)	The length of the road affected
Description	natural language description of the accident
Number	Street number in address field
Street	Street name in address field
Side	Relative side of the street
City	City in address field
County	County in address field
State	State in address field
Zipcode	Zipcode in address field
Country	Country in address field
TimeZone	Timezone based on the location
Airport_Code	Denotes an airport-based weather station
Weather_Timestamp	Time-stamp of weather observation
Temperature(F)	Temperature (in Fahrenheit)
Wind_Chill(F)	Wind chill (in Fahrenheit)
Humidity	Humidity (in percentage)
Pressure(in)	Pressure (in inches)
Visibility(mi)	Visibility (in miles)
Wind_Direction	Wind direction
Wind_Speed(mph)	Wind Speed (in miles per hour)
Precipitation(in)	precipitation (in inches)
Weather_Condition	Weather condition (rain, snow, thunderstorm, fog, etc.)
Amenity	Indicates presence of amenity
Bump	Indicates presence of speed bump
Crossing	Indicates presence of Crossing
Give_Way	Indicates presence of Give Way
Junction	Indicates presence of Junction
No_Exit	Indicates presence of No Exit
Railway	Indicates presence of Railway
Roundabout	Indicates presence of Roundabout
Station	Indicates presence of Station
Stop	Indicates presence of Stop
Traffic_Calming	Indicates presence of Traffic Calming

Table 1: List of Original Predictors.

Column Name with Description	
Traffic_Signal	Indicates presence of Traffic Signal
Turning_Loop	Indicates presence of Turning Loop
Sunrise Sunset	Period of day based on sunrise/sunset
Civil_Twilight	Period of day based on Civil Twilight
Nautical_Twilight	Period of day based on Nautical Twilight
Astronomical_Twilight	Period of day based on Astronomical Twilight

4.2 Grouping similar Weather Conditions

The initial list of weather condition had 17 levels ranging from clear to thunderstorms and rain.

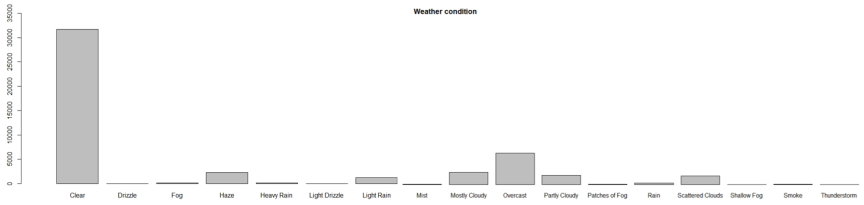

Figure 1: Spread of initial Weather Conditions

Similar weather conditions were grouped together to provide easy interpretability with 4 standard levels.
The grouping is as follows:
1.Clear
2.Rain - Drizzle,Heavy Rain,Light Drizzle,Light Rain,Thunderstorms and Rain
3.Cloudy - Mostly Cloudy,Overcast,Partly Cloudy,Scattered Clouds
4.Fog - Smoke,Fog,Haze,Mist,Patches of Fog,Shallow Fog

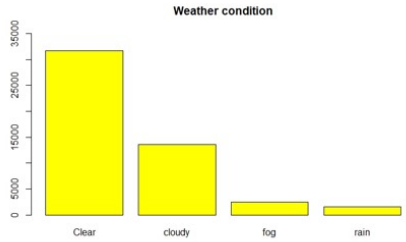

Figure 2: Grouped Weather Condition

4.3 Severity Classification

Severity is classified into "High Severity" and "Low Severity" based on the initial level from the data set. High Severity category include levels from 3 and above with duration more than 10 minute traffic delay. Whereas low severity include levels 1 and 2 with duration less than 10 minute traffic delay. The classification was made for easy understanding of severity zones and set up amenities. The count of each severity category are
Low Severity-26705
High Severity-22579

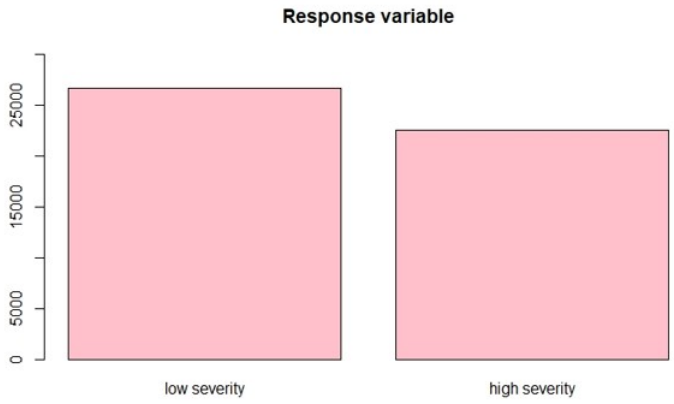

Figure 3: Severity Count

5 Exploratory Data Analysis

5.1 Severity Distribution in Los Angeles

The map below [Figure 4] shows the accident points in the city of Los Angeles. The map was created using google maps API, ggmap in R and made use of latitude, longitude co-ordinates from the data set.

Red dots represent high severity [Figure 5]
Yellow dots represent low severity [Figure 5]
Closer inspection from the maps show, high severity occurs along the state freeway which forms the backbone for most accidents in the city of Los Angeles.

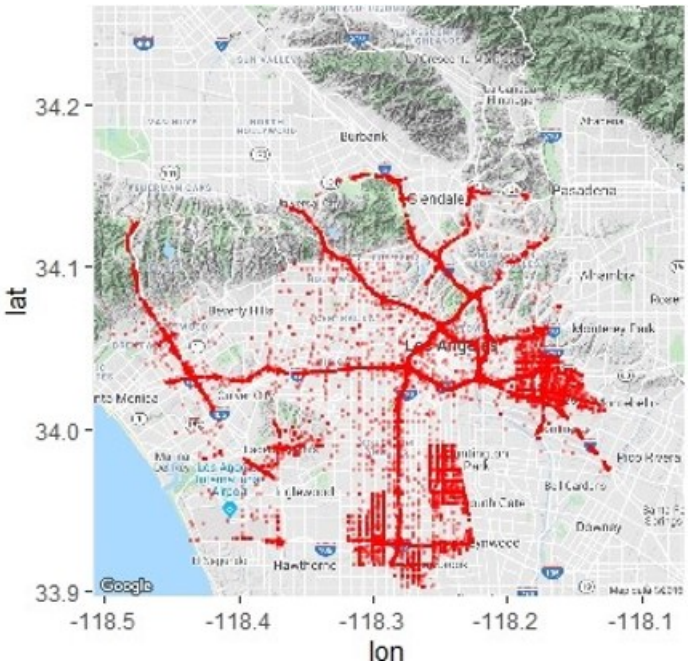

Figure 4: Severity Location in Los Angeles

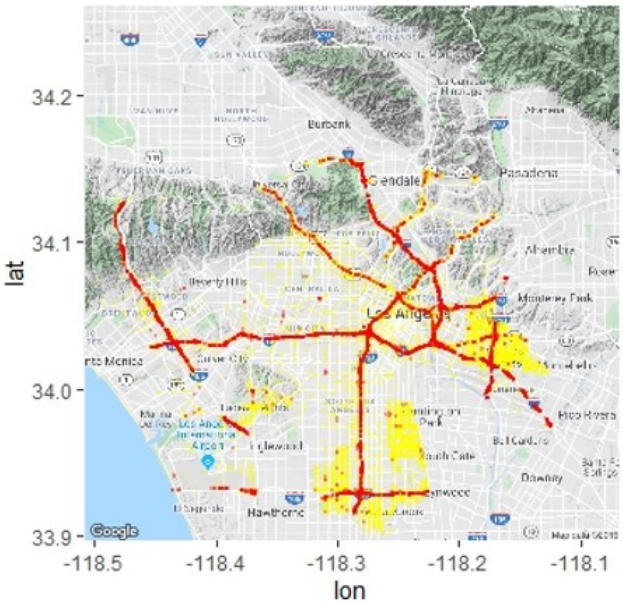

Figure 5: Comparison of High and Low Severity

Low severity accidents are observed as small clusters within the city.

5.2 Predictor Distribution and Interpretation

Predictors such as temperature, humidity and pressure are continuous data. From the plots temperature and pressure are normally distributed but humidity is skewed to the left indicating more accidents occur at higher humidity.

Plots on weather condition and time of day show most accidents occur in clear

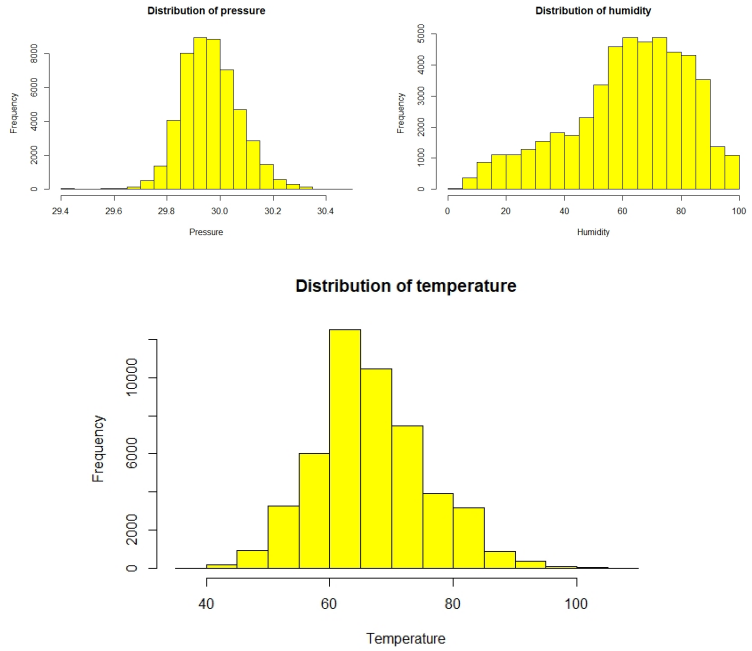

Figure 6: Distribution of Pressure, Humidity, Temperature

weather when the visibility is high and this is aided with the visibility plot.

Boxplot for the continuous predictors were inspected. Outliers were considered as they carry important information about severity and is usedfor capturing the variation in the data. Outliers are generally 1.5 times of the interquartile range in the boxplot.

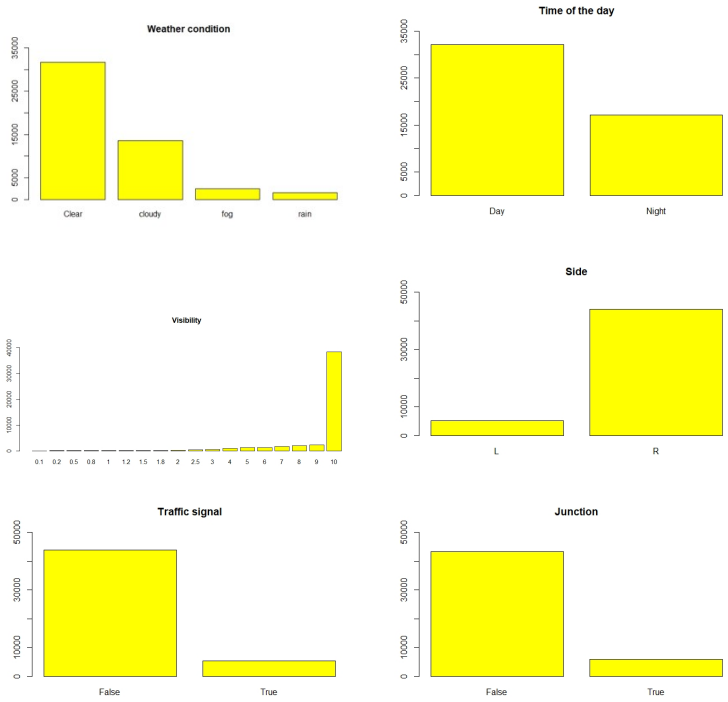

Figure 7: Categorical Predictor Distribution

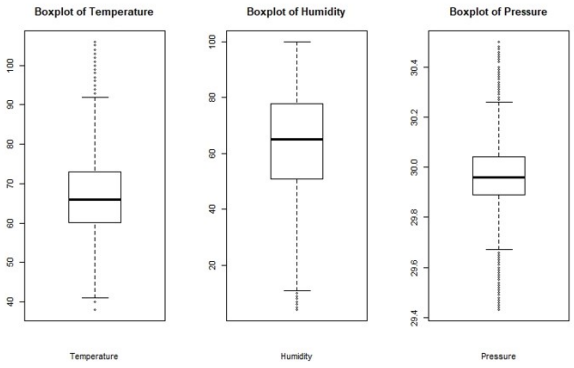

Figure 8: Boxplot of continuous predictors

6 Modeling Methodologies

6.1 Terminology

6.1.1 Bias

If the model has a high bias it is considered to over fit the data. In broader terms, the inability for a model to capture the true relationship is called to Bias.

6.1.2 Variance

Once a model is fit to the training data set, the model is run through the testing data set to make predictions. The difference in fits between data sets is called Variance. A model is chosen if they have a low bias and low variance as they are considered as a good fit to the data. Methods like regularization, boosting and bagging are used to achieve low bias and variance.

6.1.3 Confusion matrix

Trees make use of confusion matrix were the actual and predicted values are compared together in the matrix. The rows in the confusion matrix correspond to the predicted values and the columns are the actual values. Votes are added to the corresponding grid of the confusion matrix to assess the tree model. The top left and bottom right grids of the confusion matrix show the True positives and True negatives respectively, these correspond to the accuracy of the model as they make correct positive and negatives predictions. In summary, the diagonal of the matrix has the correct predictions. The top right and bottom left of the confusion matrix show the False positive and False negatives respectively, these correspond to false predictions about response The percentage of correctly identified positives is called sensitivity.

$$Sensitivity = \frac{True\,Positives}{True\,Positives + False\,Negatives}$$

The percentage of correctly identified negatives is called specificity.

$$Specificity = \frac{True\,Negatives}{True\,Negatives + False\,Positives}$$

The sensitivity and specificity values are compared with different models and a model according to our importance of positivity/negativity is chosen.

6.2 Modeling Methods

Since the response variable is categorical, we use classification methods. Few of the modelling methodologies are discussed as follows.

6.2.1 Logistic Regression Model

Logistic Regression is a simple linear model where the response is categorical i.e the response falls into binary classification. It is calculated based on the sum of log likelihood an event is about to happen.

$$Z_i = ln(P_i/1 - P_i)$$

The log of events are converted back to probability

$$P_i = 1 - (1/1 + e_i)$$

where, $P_i = probability\, of\, an\, event$

6.2.2 Decision trees

Decision trees are generally constructed top-down with several internal nodes followed by leaf nodes. Tress are used for classification data and they make use of *gini* impurity to generate a list of questions. Each row of a predictor is compared to the result of the response and votes are given in the leaf node. Generally, none of the leaf nodes have all the votes under one split and so they are all considered "impure". For example, the relative side of the street is compared to the level of severity and the same is followed with other predictors. The *gini* impurity for each leaf node is calculated by

$$Gini\, Impurity = 1 - (Probability\, of\, high\, severity)^2 - (Probability\, of\, low\, severity)^2$$

The gini impurity for the internal node or the column of predictor,

$$\frac{Total\, in\, the\, Leaf\, Node}{Total\, in\, Leaves\, Node} * Gini\, Impurity + \frac{Total\, in\, the\, Leaf\, Node}{Total\, in Leaves\, Node} * Gini\, Impurity$$

The tree with the lowest *gini* impurity value is chosen and the process is repeated for the next internal node with the remaining columns. If the *gini* impurity value is higher in the leaf node than the internal node the internal node becomes a leaf node. Decision trees can also be used with numeric values. Here the numeric predictor column is sorted in ascending order and the average value for adjacent rows are calculated. For example, average temperature value is calculated for adjacent rows and are matched with the response variable. The lowest impurity value is chosen as the internal node as the tree grows downwards. To handle missing values in decision trees, the most common value in the column is used to fill the missing value or a high correlation with another column is used. In case of numeric values the mean/median is used or a high correlation with another column is used in linear regression and by using least squares the missing value can be predicted. Decision tress are easy to understand and can handle both categorical, numerical data but they are prone to over fitting leading to a biased tree. But by requiring a large reduction in the impurity score to make a split the tree can be avoided from over fit i.e feature selection. Methods such as pruning, the branches of the tree with low important features are removed to avoid over fitting and can increase the predictive accuracy. Rpart function has been used in this analysis.

6.2.3 Bagging

The process of sampling the existing data into new data, but the rows can be repeated. Generally, one-third of the data won't enter the new data set, these are called "Out of bag samples". Using the bootstrapped data and aggregating them to make decisions is called bagging.

6.2.4 Random Forest

Random forests make use of the simplicity in decision trees and adds flexibility thereby increasing the accuracy. The data is first bootstrapped, random forest method randomly selects variables to create internal nodes in the tree and chooses the option with the most votes. The accuracy of the random forest is checked on every iteration using a new bootstrapped data and the variables used are changed. The testing data is run through the different trees in the random forest and choose the option with the most votes. The accuracy of the random forest is calculated with the help of "out of bag samples" correctly identified. The proportion of out of bag samples that were incorrectly classified is called "Out of Bag error". We choose the random forest which is the most accurate by changing the number of variables in the data. Missing data in random forest are handled by using the most common value in the original data for categorical variables and by median value in case of numeric.

6.2.5 Boosting

Boosting is a ensemble learning technique used is supervised learning which combines multiple weak learners to create strong one. The concept of boosting is to train weak learners sequentially, each one trying to correct the predecessor. Two types of boosting methods are used in this project namely, Adaptive boosting(Adaboost) and Gradient boosting(gbm).
In Adaboost, the shortcomings are identified by high weight data points by calculating weights for the previously mis classified points. Gradient boosting uses gradients to improve the model by directly considering the residuals (misclassified points).

7 Modeling and Results

7.1 Model Framework

The diagram shows the process flow followed for fitting the model. The final data set with required predictors is used for modeling and the model evaluated using randomized holdout method, were the data set is split into a holdout and train data corresponding to inputs in the holdout function. Later models are built on the train data and prediction is done on the holdout data. The prediction accuracy of each model is compared to select the best model. A null model

is used as a benchmark to evaluate whether the developed models perform better.The training models used for prediction include Logistic Regression, CART, Random Forest, Boosting.

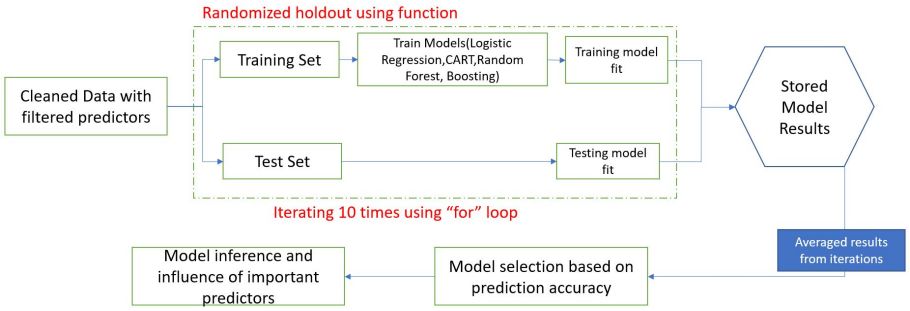

Figure 9: Model Framework

The process of model fitting is repeated for ten iterations to check prediction consistency.

7.2 Randomized Holdout Method

A matrix is created to store the prediction accuracy of each model used and a function called random string is designed which takes percentage and length as input. The function randomstring assigns zeros for training and ones for the holdout testing data randomly to a dataframe corresponding to the the length(input) and percentage(input). In our case we have used length as the length of the dataset and percent as 20 percentage(holdout percent). A initialization for loop is used to iterate for 10 times. First step in for loop is to initialize the function so that it generates zeros and ones for the length of the dataset with 20 percent zeros and remaining to ones. Later, a dataframe is created to combine the full data with the created dataset from the function. Another dataframe is created to store the response variable corresponding to the function.A Dataframe holdout is built using the subset function to subset the data that contains only ones(test data), the same is followed for the response. Dataframe train is created for zeros corresponding to the data. Then, models are built inside the for loop with train data to predict the responses using holdout data. Table function is used to create a confusion matrix and Prediction accuracy is calculated for the respective developed model. For each iteration the randomstring function generates zeros, ones randomly. The process is repeated for 10 iterations inside for loop. The prediction accuracy is calculated and stored inside the matrix created initially. The mean for each column (models) is calculated and the best model is selected based on the mean prediction accuracy.

7.3 Model Results

Results of both train and test set are stored in a table. Average of each column is taken and the model with the highest average prediction accuracy is chosen as the best model.

	Logistic regression	Cart using rpart	Randomforest	Gbm boost	Ada boost	Null model
1	0.5976268	0.6001820	0.7162217	0.6413769	0.5978940	0.5404967
2	0.5953270	0.6002086	0.7182956	0.6380204	0.5996099	0.5403518
3	0.5967979	0.5985307	0.7136457	0.6382401	0.5990353	0.5409044
4	0.5970523	0.5975403	0.7160321	0.6397315	0.5993002	0.5410432
5	0.5973904	0.6016958	0.7103807	0.6413560	0.5992710	0.5428124
6	0.5979335	0.5989511	0.7112285	0.6388685	0.6058773	0.5413877
7	0.5966963	0.5972549	0.7218032	0.6364654	0.6004423	0.5428050
8	0.5953134	0.5979074	0.7140720	0.6407064	0.5977522	0.5420563
9	0.5969523	0.5983386	0.7124239	0.6387701	0.6053528	0.5411890
10	0.5992242	0.6001874	0.7174757	0.6395136	0.5988552	0.5434368

	Logistic regression	Cart using rpart	Randomforest	Gbm boost	Ada boost	Null model
1	0.5944326	0.5929395	0.6133603	0.6046727	0.6027411	0.5473076
2	0.6066202	0.5930233	0.6142091	0.6074231	0.5964322	0.5477422
3	0.5965357	0.5996380	0.6147844	0.6049919	0.5991510	0.5457245
4	0.6041324	0.6036146	0.6179259	0.6041185	0.6004267	0.5450972
5	0.5985327	0.5870400	0.6085593	0.5997181	0.5973813	0.5379610
6	0.5974642	0.5979623	0.6154956	0.6024997	0.6053601	0.5437475
7	0.6044896	0.6047984	0.6164075	0.6081244	0.5985123	0.5380793
8	0.6021260	0.6020796	0.6117432	0.6057160	0.5962708	0.5410687
9	0.5956295	0.6004032	0.6100163	0.6086435	0.6044266	0.5445414
10	0.5956517	0.5929782	0.6203544	0.6079879	0.5981432	0.5356246

Figure 10: Accuracy of Train and Test data sets

7.4 Random Forest- Variable Importance Plot

A variable importance plot is taken from the random forest model to study the predictors which has the significant impact.

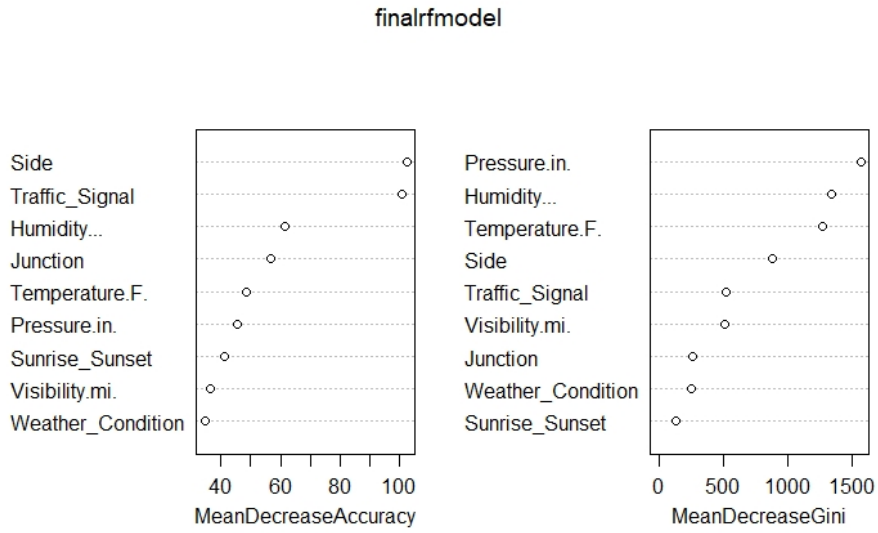

Figure 11: Variable Importance plot- Random Forest

From the variable importance plot we can infer that the two most influential predictors are presence of traffic signal and the side of the road. It is obvious

that the presence of traffic signal can largely affect the likelihood of accident to be occurred.It is interesting to note that Weather condition and Visibility have low significance. So, from our results we can suggest to have more number of traffic signals and for the places where traffic signals can't be placed we can suggest to has appropriate signboards and roundabouts to control and reduce the speed of the moving traffic.
Also, another most influential predictor is side of the road while traveling and we can infer from the exploratory analysis that most of the high severe accidents have happened on the right side of the road and since most of the high severe accidents have happened on freeways we could say the accidents have happened when a vehicle is trying to make an exit from the right lane. So, restricting the speed of the vehicles and providing roundabouts for the traffic that are taking a exit will help reduce the accident fatality.

8 Conclusion and Future Works

The aim of the present study was to predict the severity of road accidents at Los Angeles utilizing the traffic accident data set. A total of six models were developed including Null model in R and Randomforest seemed to be a promising model for prediction since it has the highest prediction accuracy averaging around 62% out of all the models developed. Also the variable importance plot from random forest show traffic signals and side of the road are most influencing factors while deciding the severity of the accidents that are likely to occur. Predicting high severity and low severity accident zones can help us take appropriate countermeasures by setting up variable speed limits and appropriate signboards. Also, this prediction will help in setting up amenities such as hospitals and other emergency facilities near high severe accident zones.
This paper studies the prediction of accidents at Los Angeles. However, in future the same methodology can be replicated for a state or a country while including more predictors, which will improve the model performance and help us predict more accurately.

References

[1] Caliendo, Ciro and Guida, Maurizio and Parisi, Alessandra *A crash-prediction model for multilane roads* 2007 - Elsevier.

[2] Chang, Li-Yen. *Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network* 2005 - Elsevier.

[3] Chang, Li-Yen and Chen, Wen-Chieh *Data mining of tree-based models to analyze freeway accident frequency* 2005 - Elsevier.

[4] Chen, Quanjun and Song, Xuan and Yamada, Harutoshi and Shibasaki, Ryosuke *Learning deep representation from big and heterogeneous data for*

traffic accident inference 2016 - Thirtieth AAAI Conference on Artificial Intelligence.

[5] Lin, Lei and Wang, Qian and Sadek, Adel W *A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction* 2015 - Elsevier.

[6] Mannering, Fred L and Shankar, Venky and Bhat, Chandra R *Unobserved heterogeneity and the statistical analysis of highway accident data* 2015 - Elsevier.

[7] Najjar, Alameen and Kaneko, Shun'ichi and Miyanaga, Yoshikazu *Combining satellite imagery and open data to map road safety* 2017 - Thirty-First AAAI Conference on Artificial Intelligencer.

[8] Ren, Honglei and Song, You and Wang, Jingwen and Hu, Yucheng and Lei, Jinzhi *A deep learning approach to the citywide traffic accident risk prediction* 2018 - IEEE

[9] Briceño-Navarro, Pablo and Sánchez-Squella, A and Orellana, Álvaro and Shah, Dhruv *2017 2nd IEEE International Conference on Intelligent Transportation Engineering* 2017 - ICITE

[10] Yuan, Zhuoning and Zhou, Xun and Yang, Tianbao *Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data* 2018 - ACM

[11] Yuan, Zhuoning and Zhou, Xun and Yang, Tianbao and Tamerius, James and Mantilla, Ricardo *Predicting traffic accidents through heterogeneous urban data: A case study* 2017 - UrbComp

[12] Yuan, Zhuoning and Zhou, Xun and Yang, Tianbao and Tamerius, James and Mantilla, Ricardo *Predicting traffic accidents through heterogeneous urban data: A case study* 2017 - UrbComp

[13] Abellán, Joaquín and López, Griselda and De OñA, Juan *Analysis of traffic accident severity using decision rules via decision trees* 2013 - Elsevier

[14] Bergel-Hayat, Ruth and Debbarh, Mohammed and Antoniou, Constantinos and Yannis, George *Explaining the road accident risk: weather effects* 2013 - Elsevier

[15] Caliendo, Ciro and Guida, Maurizio and Parisi, Alessandra *A crash-prediction model for multilane roads* 2007 - Elsevier

[16] Chang, Li-Yen *Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network* 2005 - Elsevier

[17] Chong, Miao and Abraham, Ajith and Paprzycki, Marcin *Traffic accident analysis using machine learning paradigms* 2005 - Informatica

[18] Chang, Li-Yen and Chen, Wen-Chieh *Data mining of tree-based models to analyze freeway accident frequency* 2005 - Elsevier

[19] Chen, Quanjun and Song, Xuan and Yamada, Harutoshi and Shibasaki, Ryosuke *Learning deep representation from big and heterogeneous data for traffic accident inference* 2016 - Thirtieth AAAI Conference on Artificial Intelligence

[20] Chong, Miao M and Abraham, Ajith and Paprzycki, Marcin *Traffic accident analysis using decision trees and neural networks* 2004 - arXiv preprint cs/0405050

[21] Eisenberg, Daniel *The mixed effects of precipitation on traffic crashes* 2004 - Elsevier

[22] Lin, Lei and Wang, Qian and Sadek, Adel W *A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction* 2015 - Elsevier

[23] Oh, Jutaek and Washington, Simon P and Nam, Doohee *A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction* 2006 - Elsevier

[24] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Rajiv Ramnath *Countrywide Traffic Accident Dataset* 2019 - arXiv:1906.05409v1