Big Data Analysis on U.S. car accidents

Akhil Samineni
George Mason University
Fairfax, Virginia
asaminen@gmu.edu

Venkata Sai Ravi Teja Adabala
George Mason University
Fairfax, Virginia
vadabala@gmu.edu

Vaishnavi Putcha
George Mason University
Fairfax, Virginia
vputcha@gmu.edu

Abstract -- Over the years, demand for mode of transportation changed from public to private. Currently, most of the people depend on private transportation especially in the United States. The demand for travel is rising along with the development of transportation in infrastructure and the related traffic safety challenges. Around the world, there are many car accidents. The expense of vehicle crashes and driver injuries has an enormous effect on society. Through our analysis we would like to find out the number of accidents occurring around states, and cities, over the year, and in different months of the year to estimate the severity of accidents.

Keywords – Car accidents, classification, United States, predict, Severity, Astronomical Twilight, Civil Twilight, Sunrise/Sunset, Nautical Twilight

I. INTRODUCTION

A study by the Association for Safe International Road Travel found that each year, 37000 people are killed in car accidents, and 2.35 million more are injured or rendered permanently disabled. 1600 deaths occurred among children under the age of 15, whereas 8000 deaths occurred among those between the ages of 16 and 20 because of car accidents. Vehicle accidents are now one of the four main causes of death for those living in metropolitan areas and have thus become a social risk. Citizens lose hundreds of billions of dollars each year due to road accidents. Several significant incidents were to blame for most of the losses. Major accident prevention aims to prevent potentially dangerous driving conditions. If we can determine the main causes of catastrophic accidents, we may take the required steps and manage our resources more effectively, both financially and personally. The data used in this case study is drawn from a national accident data collection that includes all 49 states in the US. 1.5 million traffic accidents occur between 2016 and 2021.

***Problem Statement*:**

The goal of this paper is to conduct an analysis to be able to predict the severity of the accidents all over the United States using big data tools. This analysis will be done to investigate the occurrence of accidents in different states and cities throughout the year, from 2016 to 2021, in different timings of the day.

Research Questions:

- Which State and City recorded the maximum number of accidents?
- Which month or year had how many accidents and if there is a trend or a pattern?
- Which time of the day has most accidents?
- Predict severity of car accidents.

II. MOTIVATION

Based on a number of data criteria, this study analyzes the current situation with regard to car accidents in the United States and provides suggestions for lowering both accident frequency and severity. We will explore this data considering accidents distributed geographically across the United States. Examine the concentrated accident duration and the recent volatility of total accidents, then look at the data from a temporal viewpoint.

III. LITERATURE SEARCH

To evaluate how bad weather affects accident rates, researchers performed meta-analysis, which showed that weather influenced traffic safety. The average percentage of collision rate on rainy and snowy days is 71 percent and 84 percent, respectively. According to the study, extreme weather in the United States will have an impact on injuries and automobile accidents, particularly on wet and snowy days [2]. K-means clustering was used to estimate and forecast the severity of traffic accidents. To explore different machine learning techniques for forecasting the severity of traffic accidents, this study examined the performance of two machine learning applications, random forest (RF) and Bayesian additive

region trees (Bart). It is found that, when compared to the prediction of the Bayesian additive region trees (Bart) model, the random forest (RF) model with meteorological conditions has a higher prediction probability for determining the seriousness of a traffic collision. The variable importance technique shows that the kind of traffic accidents and the weather conditions are two important components that may provide important information in the modeling and estimation procedures [1].

For US government agencies and the public, their study examines data from 50 states to evaluate traffic accident patterns, causes, and possible prevention measures. Utilizing logistic regression, several variables that were investigated and assessed [6] were associated to the accident's severity.

Another study tries to solve this problem and delve more into the factors that contribute to the increase in the frequency of auto accidents. The data for this study was continuously gathered in the United States from 2016 to 2020 from traffic accident incidences recorded by the Department of Transportation, law enforcement organizations, and traffic cameras. To predict how traffic incidents might affect it, two models were utilized, with a focus on the main factors that contribute to accidents on the road. According to the research, traffic caused by work rush hour and population density are the two main factors influencing car accident rates. [10]

Their analysis includes information on the number of accidents by year, accidents by state, the optimal time to travel by month, day, and hour, accident-prone locations in each state, elements that cause accidents, such as weather, wind flow, temperature, location, and so on, deaths in each state, age groups of fatalities, drivers involved in accidents, drivers' age groups, involved cars, and drivers who have taken alcohol. The platform for analysis was developed using Tableau. [9]

R language demonstrates how data is related by analyzing traffic statistics and graphs. Locations of the accidents and the accident thermal chart were obtained after data preprocessing and data selection using R language Remap package remap and remap functions. In addition, to model the data, they used decision trees, linear regression, and the random forest approach. We can check the model's correctness and obtain the most accurate model based on the actual findings, which will help in predicting the model's accuracy with comparable data in the future. After validating the model and examining the data properties and relationship between the variables, the ultimate purpose of data analysis is to identify the most correct model.[12]

It was the goal of another research project to offer a model that might be utilized to explain why the number of fatal traffic accidents varies so much between nations. National transportation, socioeconomic, and infrastructure factors were investigated as prospective variables from global databases. The model was produced by stepwise regression analyses. [11]

IV. PROPOSED APPROACH

The methodology that will be used to solve the problem is divided into several parts. The stages are as follows: comprehending the problem description; data preprocessing; exploratory data analysis; feature engineering; data modeling; model evaluation; recommendations and insights. Data cleaning, imputation of null values using the mean or median, and transformation of data into meaningful variables or combinations of variables are tasks involved in problem understanding and data preprocessing. The dataset's most crucial attributes were chosen using feature engineering. Data modeling involves dividing data into train and validation sets, using several algorithms like Gradient Boost, Random Forest, and Decision Tree, and modifying hyperparameters to create the optimal model. To choose the optimal model, various metrics are used during model evaluation, including accuracy, precision, recall, F1 score, and AUC/ROC curve. Models created to tackle business challenges with the help of recommendations and insights generated through exploratory investigation.

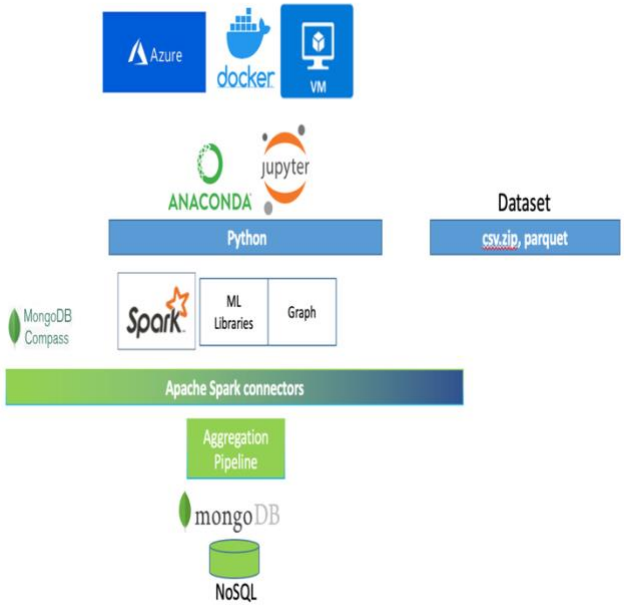
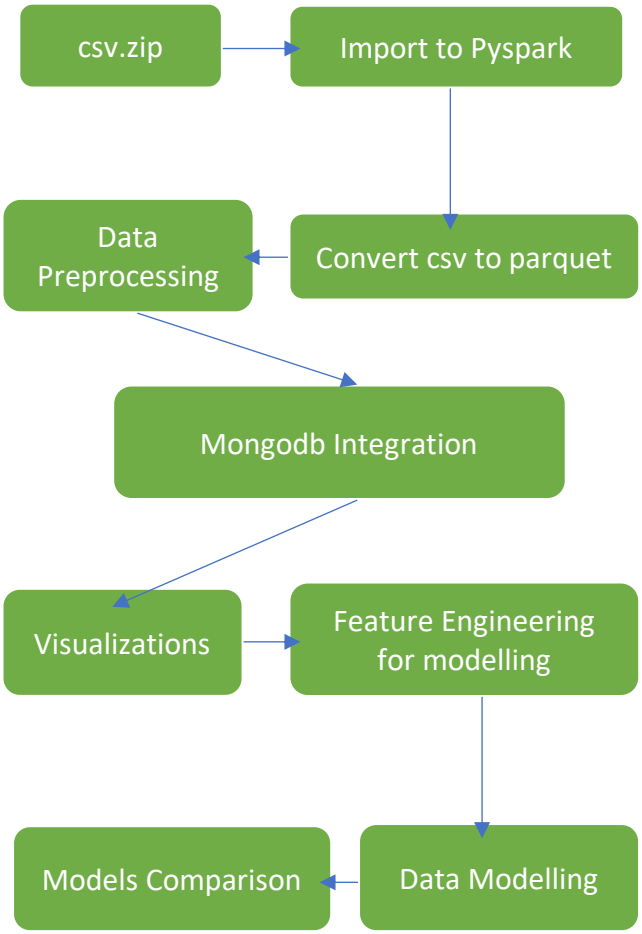


Fig1. Bigdata Architecture



Data Flow

V. ANALYSIS

A. Dataset description

Using information gathered between February 2016 and December 2021, the collection includes car accident data for the USA. Kaggle provided the dataset (US car Accidents Feb16 Dec19). It has 49 qualities with both category and numerical data and 1.1 million observations. The dataset was gathered from numerous APIs, including two APIs that broadcast real-time traffic data, and it covers 50 US states that are next to one another. Geographical features, weather, date-and-time, and POI (point of interest) annotation are some of the different categories that the features fall under.

B. Data Pre-processing

We used a portion of the roughly 1 million observations for our analysis and modeling. Removing unnecessary columns from the data that wouldn't be useful for the research was the first stage in data cleaning. We made the decision to get rid of any features that had a single unique value for the entire column or a single unique value for each row. Additionally, columns with many missing values were deleted.

Handling Outliers

Extreme outliers that are essentially impossible were found in weather-related data. The outlier values in these situations were trimmed to the minimum or maximum thresholds. We conducted a variety of exploratory analyses to identify outliers, and then we managed them accordingly.

Handling Null values

Most of the columns in our sample were null. Missing values for categorical and numerical features were substituted with the mode and median of the corresponding columns. In cases where there was no precipitation, zero was substituted for null values of precipitation, assuming that these quantities were absent. By deleting the entire row, null values were eliminated for some columns. We chose to abandon it since imputing these values would have been deceptive.

Handling duplicate values

There were duplicate values for categorical attributes like Wind Direction, such as "North" and "N." The single character representation of the duplicates was kept, and the other was substituted. The similar strategy was used in the case of Weather Condition, where the terms "Light Rain Shower" and "Light Rain Showers" had equivalent meanings.

Handling data formats

There were data in columns which represent the same information but present in different formats. We handled this type of data by brining all the data to a single format so that processing becomes easy.

Database Integration

For data ingestion and storage, we have chosen Mongodb. Mongodb retrieves the preprocessed data for storage.

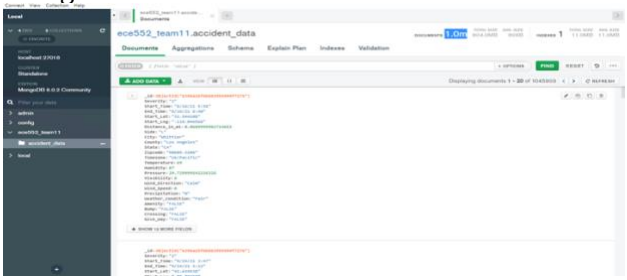


Fig2. Snapshot of Mongodb

C. Exploratory Data Analysis

We performed exploratory data analysis on the data. Firstly, we observed the occurrence of accidents in cities and states around the United States.

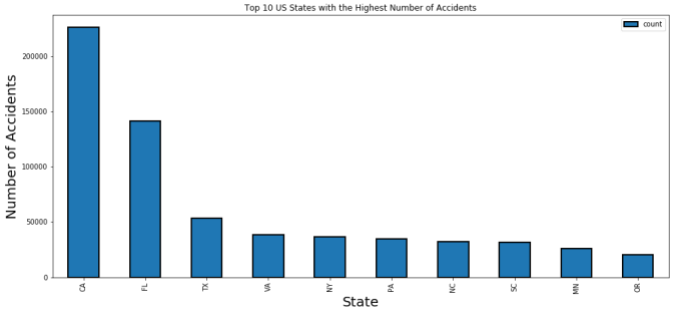


Fig 3. State-wide accidents occurrence

From the bar plot, it can be observed that California has the highest count of accidents amongst other states in the United States. Following CA, Florida (FL) saw the next highest occurrence of accidents.

When filtered with cities, the highest number of accidents occurred in Miami.

City	count
Miami	38929
Orlando	21476
Los Angeles	19595
Dallas	15442
Houston	14839
Charlotte	13044
Sacramento	8953
San Diego	8489
Raleigh	7608
Nashville	7084

only showing top 10 rows

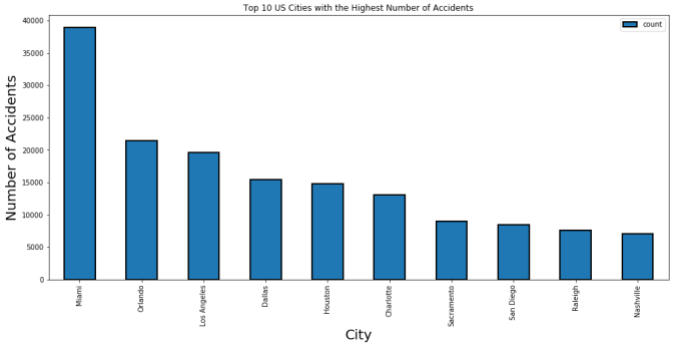


Fig 4. City wise occurrence of accidents

	Severity	month_of_year	count
0	2	December	147875
1	2	November	105865
2	2	October	84983
3	2	September	77075
4	2	June	75395
5	2	July	72567
6	2	August	71309
7	2	May	56200
8	2	April	49188
9	2	March	45995

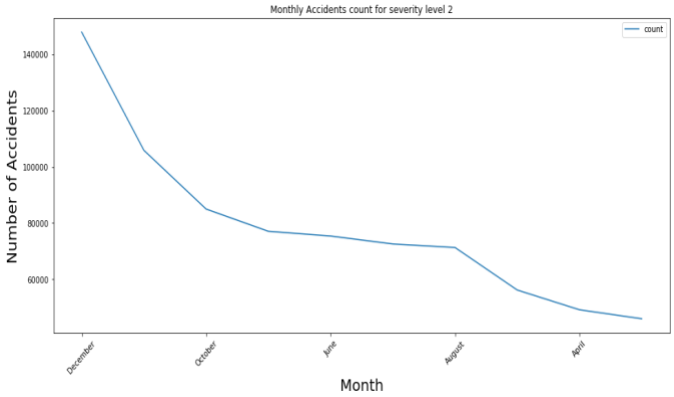


Fig 5. Monthly accidents count

When considering the occurrence of accidents per month, most of the accidents were observed with severity 2.

December had the highest count of accidents when compared to other months. When filtered with years, 2021 has the highest count of accidents. The below figure shows the top 3 years: 2016, 2017 and 2021with occurrence of accidents.

```
In [71]: pc4=parquetaccidents.groupBy('Year').count()
In [72]: pc4=pc4.sort(col('count').desc())
In [73]: pc4.show(3)
```

```
+----+------+
|Year| count|
+----+------+
|2021|637546|
|2016|121919|
|2017|102842|
+----+------+
only showing top 3 rows
```

Fig 6. Yearly accidents count

Severity 2 has the highest count of accidents. Least distribution of severity was observed with severity 3 & 4.

```
In [74]: pc5=parquetaccidents.groupBy('Severity').count()
In [75]: pc5=pc5.sort(col('count').desc())
In [76]: pc5.show()
```

```
+--------+------+
|Severity| count|
+--------+------+
|       2|809360|
|       3| 39963|
|       4| 39138|
+--------+------+
```

Fig 7. Severity distribution

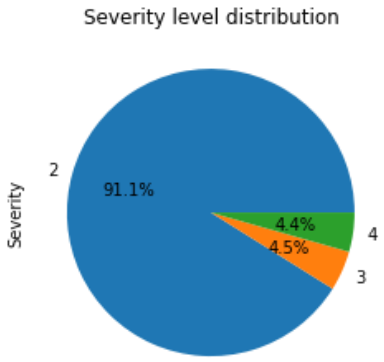


Fig 8. Percentage severity distribution

Friday recorded the highest number of accidents from the years 2016 to 2021. We can see from the bar graph that the accident count is reducing drastically on the weekends. This might be due to low traffic and less flow on the roads.

```
In [108]: pc6=parquetaccidents.groupBy('day_of_week').count()
In [111]: pc6=pc6.sort(col('count').desc())
In [112]: pc6.show()
```

```
+-----------+------+
|day_of_week| count|
+-----------+------+
|     Friday|157983|
|   Thursday|146045|
|  Wednesday|145424|
|    Tuesday|141861|
|     Monday|130827|
|   Saturday| 91551|
|     Sunday| 74770|
+-----------+------+
```

Fig 9. Weekly count of accidents

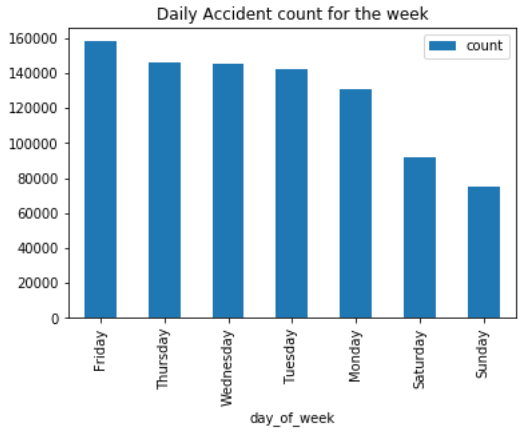


Fig 10. Weekly distribution of accidents

The below results are to know which specific times of the day have more severity in accidents. We have four different times in a day namely - Sunset/Sunrise, Civil Twilight, Nautical Twilight, Astronomical Twilight. They are further segregated as day and night. From the analysis below we can see that Astronomical twilight has more accidents with severity 2 and 3 and most severe accidents (Severity=4) occur during the day given any time of the day from the results.

```
In [134]: pc7=parquetaccidents.groupBy('Sunrise_Sunset','Severity').count()
          pc8=parquetaccidents.groupBy('Civil_Twilight','Severity').count()
          pc9=parquetaccidents.groupBy('Nautical_Twilight','Severity').count()
          pc10=parquetaccidents.groupBy('Astronomical_Twilight','Severity').count()

In [135]: pc7=pc7.sort(col('Count').desc())
          pc8=pc8.sort(col('Count').desc())
          pc9=pc9.sort(col('Count').desc())
          pc10=pc10.sort(col('Count').desc())
```

Sunrise_Sunset	Severity	count
Day	2	574820
Night	2	234540
Day	3	26353
Day	4	22926
Night	4	16212
Night	3	13610

Figure11

Civil_Twilight	Severity	count
Day	2	610202
Night	2	199158
Day	3	28097
Day	4	24499
Night	4	14639
Night	3	11866

Figure12

Nautical_Twilight	Severity	count
Day	2	650235
Night	2	159125
Day	3	29971
Day	4	26228
Night	4	12910
Night	3	9992

Figure13

Astronomical_Twilight	Severity	count
Day	2	682487
Night	2	126873
Day	3	31755
Day	4	27951
Night	4	11187
Night	3	8208

Figure14

Fig 11-14. Accidents distribution based on times of the day

VI. FEATURE ENGINEERING

A crucial step in creating any machine learning model is feature engineering. Numerous algorithms that we want to use are effective when used with numerical data. One Hot Encoding was used to convert category attributes to numerical attributes using dummies.

Our analysis doesn't really benefit much from time-related variables that tell us when the accident began and ended. These already-existing features were transformed into new ones that will be more useful for our study, such as the day of the week and the hour of the day.

Severity, our primary variable, has a range from 1 to 4. The dataset is unbalanced since Severity 2 data points make up many of the observations while Severity-1 observations are rarer. To solve this issue, we consolidate Severity 1 and 2 into 2.

We divided the complete dataset into training and validation data, with training data comprising 70% of the observations and validation data the remaining 30%. The models are developed using the training data, and the validation data are used to assess the models' performance. Both multiclass and binary classifications are possible using the data that has been handled in this way.

Handling Unbalanced data

We applied oversampling and under sampling techniques to address the imbalance in our data. We oversampled the data with severity 4 and under sampled the data with severity 2 for the multiclass target to match the number of data points in class 3. The data we collected was balanced since each class had the same number of data points. For binary classification, we under sampled class 0 until the number of data points matched class 1.

VII. DATA MODELING

Logistic Regression

Logistic Regression was the first model we tested. Assigning observations to a discrete set of classes is done using the classification process known as logistic regression. It applies the logistic sigmoid function to convert its output and returns a probability value that can be transferred to two or more discrete classifications. To normalize the input for the logistic regression, we utilized a standard scaler to center the data to its mean.

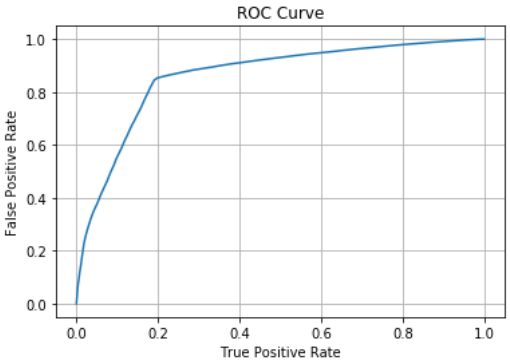


Fig 15. Logistic Regression ROC curve

Decision Tree

A straightforward visual aid for categorizing examples is the decision tree. It is a type of supervised machine learning in which the data is continually divided based on a particular parameter. Starting at the tree's base, we divide the data according to the characteristic that produces the greatest information gain using the decision method (IG). The splitting technique can then be repeated at each child node until only pure leaves remain. This indicates that all the samples at each leaf node belong to the same class. In actuality, we might impose a restriction on the depth of the tree to avoid overfitting. As the finished leaves may still contain some impurity, we make a slight compromise in terms of purity here.

Random Forest

An approach for learning with several trees is called Random Forest. A group of decision trees from a subset of the training data that was randomly chosen to make up the Random Forest Classifier. The final class of the test object is chosen by averaging the votes from various decision trees.

One of the best learning algorithms out there is random forest. It creates a classifier that is incredibly accurate for many different data sets. Large datasets can be used with efficiency. Without deleting any variables, it can manage a huge number of input variables. It provides estimates of the key classification-relevant variables. The generalization error is estimated objectively inside as the forest is being built. It retains accuracy even when a sizable amount of the data is missing thanks to an efficient mechanism for guessing missing data.

The top 10 features and their weights for utilizing a binary random classifier to determine severity are displayed below. ‘Year’ variable has the highest importance.

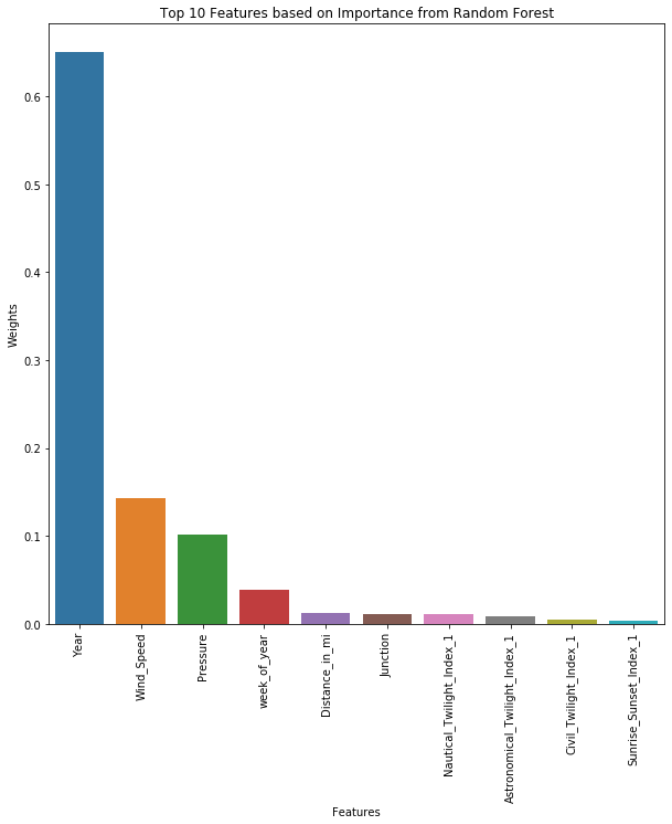


Fig 16. Feature importance from Random Forest

Model	Accuracy	AUC ROC
Logistic Regression	80.73%	86.1%
Decision Tree	80.88%	72.37%
Random Forest	81.67%	84.26%

Table 1. Models Comparison

The train and test accuracy and AUC values were practically identical, indicating that the models did not overfit. As seen in the table of model comparison, Random Forest performs best in terms of accuracy and AUC ROC Score. From all the models for the balanced data, the maximum AUC ROC score was 0.84, and the accuracy was 0.81. Using balanced data instead of imbalanced data resulted in lower accuracy for the models, while the balanced dataset had a better AUC ROC score. Due to the models' tendency to forecast the class with the most instances, accuracy was higher for the dataset that was unbalanced.

VIII. RESULTS

- California has the highest count of accidents amongst other states in the United States.
- When filtered with cities, highest number of accidents were occurred in Miami.
- When considered occurrence of accidents per month, majority of the accidents were observed with severity 2 and December month has highest number of accidents. Year 2021 has seen major number of accidents.
- Severity 2 has the highest count of accidents. Least distribution of severity was observed with severity 3 & 4.
- Severe accidents (Severity=4) occur during the day given any time of the day from the results.
- Friday recorded the highest number of accidents from the years 2016 to 2021.Accidnet rate is greatly reduced during the weekends
- Random forest model is effective comparatively to other models in predicting severity of accidents with an accuracy of 81.67% and AUC ROC of 844.26%.

IX. CONCLUSION

We discovered from our exploratory data analysis that there are some values for several factors, such as the day of the week, the hour of the day, and the month of the year, where the number of accidents is very high. As a result, the authorities may be better able to plan ahead for certain times and enforce stronger traffic laws to reduce the number. Additionally, we discovered that, intuitively, accident severity increases during the day. This can mean that daytime safety precautions need to be increased. We discovered from the feature importance plots that Year, Wind speed, and Hour were some of the most prevalent significant elements from our models that can be utilized to forecast the severity of an accident. In the case of a balanced dataset, Random Forest was the best model in terms of metrics, accuracy and AUC ROC score.

REFERENCES

[1] A. R. Mondal, M. A. E. Bhuiyan, and F. Yang, "Advancement of weather-related crash prediction model using nonparametric machine learning algorithms," SN Applied Sciences, vol. 2, no. 8, Jul. 2020, DOI:10.1007/s42452-020-03196-x.

[2] L. Qiu and W. A. Nixon, "Effects of Adverse Weather on Traffic Crashes," Transportation Research Record: Journal of the Transportation Research Board, vol. 2055, no. 1, pp. 139–146, Jan. 2008, DOI: 10.3141/2055-16.

[3] Moosavi, S. (2022, March 12). US accidents (2016 - 2021). Kaggle. Retrieved December 12, 2022, from https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[4] T. Sridharan, "How Many Car Accidents Occur in the U.S. Each Year?" 1-800-THE-LAW2, 2020. https://www.1800thelaw2.com/resources/vehicle-accident/how-many-accidents-us/

[5] S. Saha, P. Schramm, A. Nolan, and J. Hess, "Adverse weather conditions and fatal motor vehicle crashes in the United States, 1994-2012," Environmental Health, vol. 15, no. 1, Nov. 2016, DOI: 10.1186/s12940-016-0189-x.

[6] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," Accident Analysis & Prevention, vol. 34, no. 6, pp. 729–741, Nov. 2002, DOI: 10.1016/s0001- 4575(01)00073-2.

[7] M. Feng, J. Zheng, J. Ren, and Y. Liu, "Towards Big Data Analytics and Mining for UK Traffic Accident Analysis, Visualization & Prediction," Proceedings of the 2020 12th International Conference on Machine Learning and Computing, pp. 225-229, Feb. 2020, DOI: 10.1145/3383972.3384034.

[8] Motor Vehicle Safety Data. Motor Vehicle Safety Data | Bureau of Transportation Statistics. (n.d.). Retrieved December 12, 2022, from https://www.bts.gov/content/motor-vehicle-safety-data

[9] LI Gang,HUANG Tong-yuan, YAN He,et al.Grey residual error model of highway traffic accident forecast[J].Journal of Traffic and Transportation Engineering,2009,09(05):88-93.

[10] Accident analysis & prevention - journal - elsevier.Journal. (n.d.). Retrieved March 20, 2022, from https://www.journals.elsevier.com/accident-analysis-and-prevention

[11] M. Aljaban, "Analysis of Car Accidents Causes in the USA Analysis of Car Accidents Causes in the USA," 2021. [Online]. Available: https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12218&context=theses

[12] C. J. Bester, "Explaining national roadfatalities," Accident Analysis & Prevention, vol. 33, no. 5, pp. 663–672, Sep. 2001, doi: 10.1016/s0001- 4575(00)