Accident Severity Prediction

Dhanush P Nayak

November 05, 2020

1. Introduction:

1.1 Business Problem

More than 38,000 people die every year in crashes on U.S. roadways. The U.S. traffic fatality rate is 12.4 deaths per 100,000 inhabitants. An additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of death in the U.S. for people aged 1-54.

Road accidents are increasing day to day all over the world due to several reasons. These accidents may happen due to either avoidable circumstances or un-avoidable circumstances.

- Avoidable scenarios: Speeding, not wearing safety gear such as helmets or seat belts, breaking traffic rules.
- Un avoidable scenarios: Bad roads, weather conditions, lighting conditions, traffic conditions.

While avoidable accident cases could be drastically reduced by enforcing harsh/strict penalties by law. Although, un-avoidable accident situations are difficult to control, chances of fatalities or severe injuries can be reduced by notifying local police authorities, healthcare systems, traffic authorities, safety personnel in advance when there is a high probability of accidents so that they can make better decisions in avoiding the accidents or respond quickly in case of any accidents.

1.2 Target audience/ Benefits of this project:

The target audience of this project are Heath care systems, local government, police/traffic departments.

At the end of this project, we will develop a machine learning model based on several different features which will predict the severity of the accident. This model will aid stake holders to be

proactive and take impactful decisions to reduce chances/severity of the accident.

2. Methodology:

- Data cleaning is the first step by detecting and dealing with null/missing values present in the dataset.
- We will perform EDA (Exploratory data analysis) and feature engineering on most of the data attributes.
- Once the data is ready and in appropriate format, we build a machine learning model using decision tree and logistic regression.
- Finally, we test the accuracy of the model with different metrics.

3. Data gathering and cleaning:

3.1 Data Description:

This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset.

3.2 Acknowledgement:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. “Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International

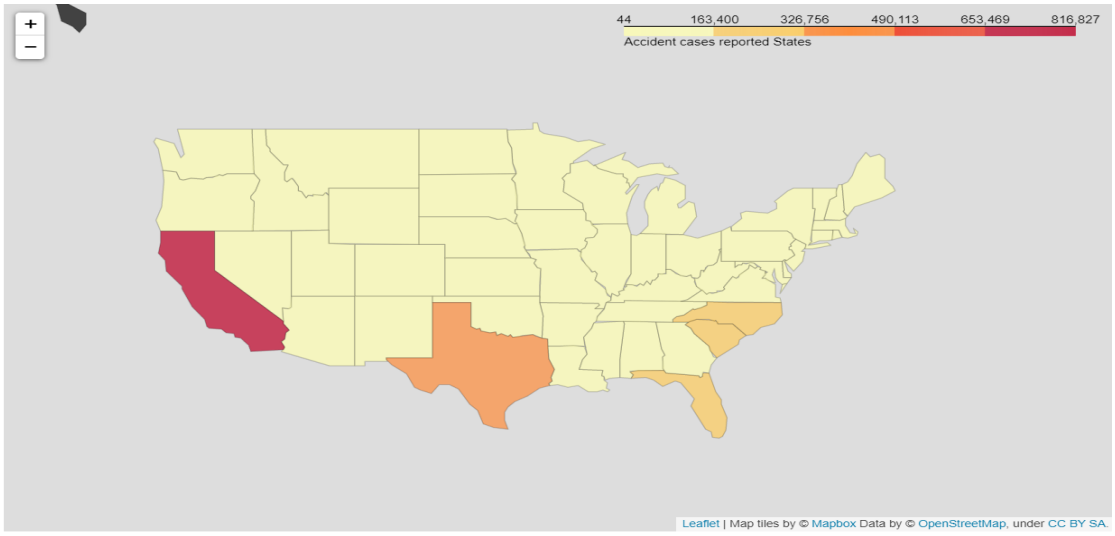
Conference on Advances in Geographic Information Systems, ACM, 2019.

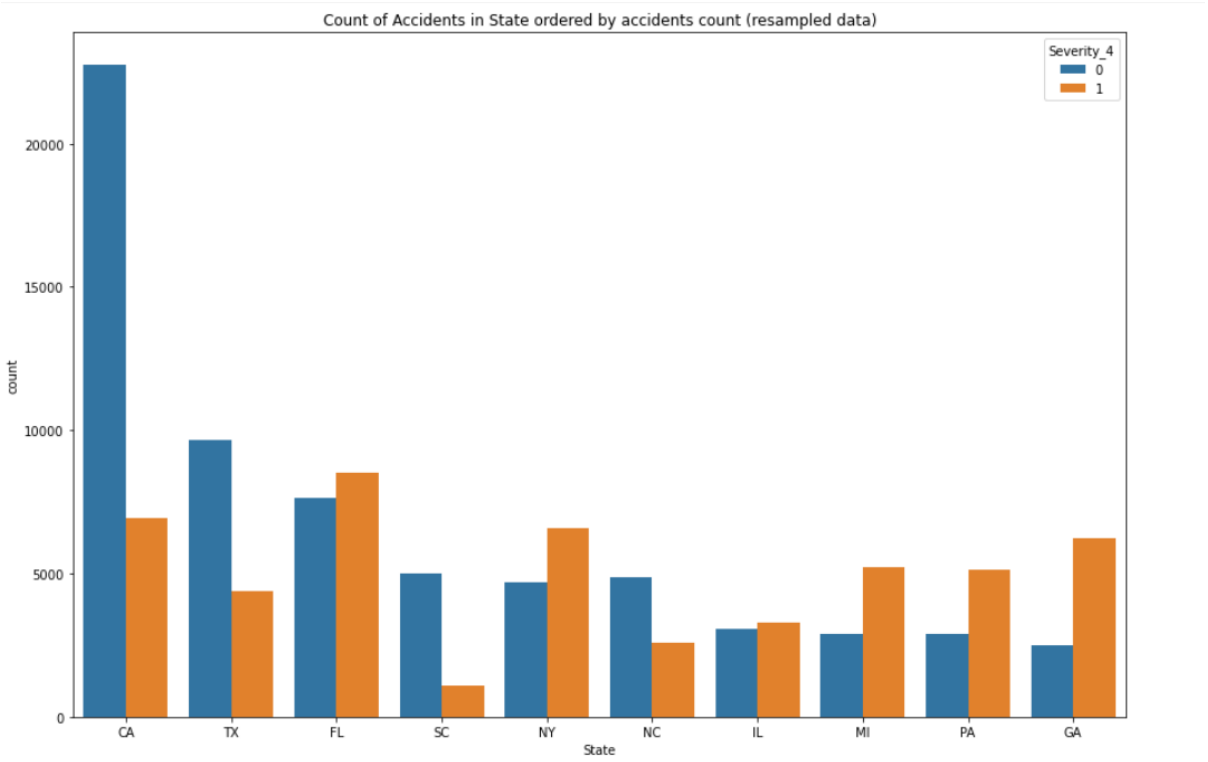
3.3 Data Cleaning:

Dataset was downloaded from Kaggle had 49 columns. There were a lot of missing values in the dataset. Missing values had to be dropped and features having >50% missing values was dropped entirely. Target variable had four values depicting different levels of accident severity. During analysis, dataset was observed to be unbalanced. Therefore, to address this issue, the combination of over- and under-sampling will be used since the dataset is large enough. Level 4 will be randomly oversampled to 100000 and other levels will be randomly undersampled to 100000. Through analysis, many different features were dropped for simpler modelling and the dropped features did not have any strong impact on accident prediction.

4. Exploratory Data Analysis:

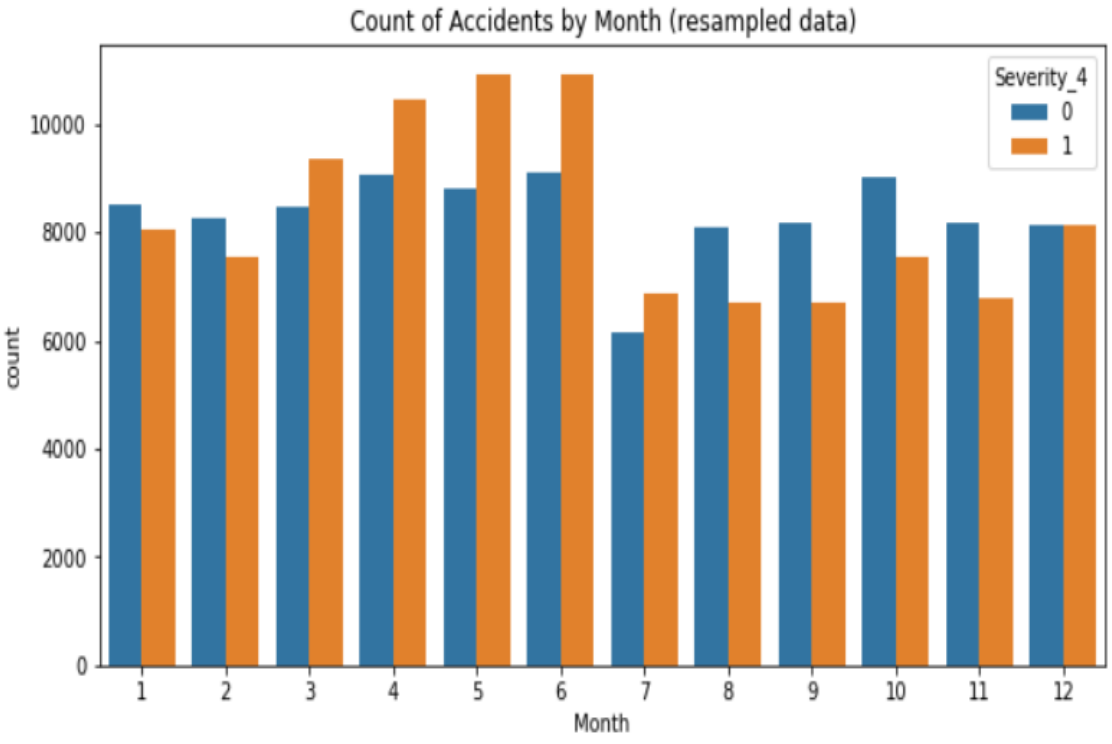
4.1 Count of cases reported across US states:





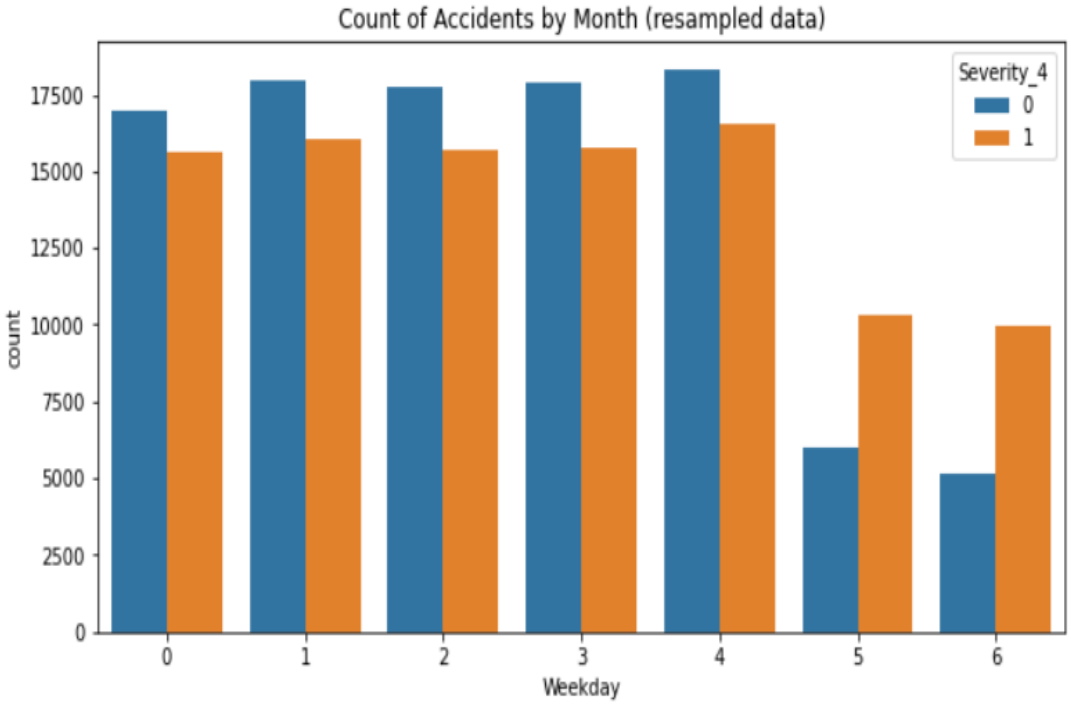
Graph shows higher accident cases in states CA, TX, and FL. The population and amount of traffic in the states may also affect the accident count. But we will not be considering population in this study.

4.2 Count of accidents by month:



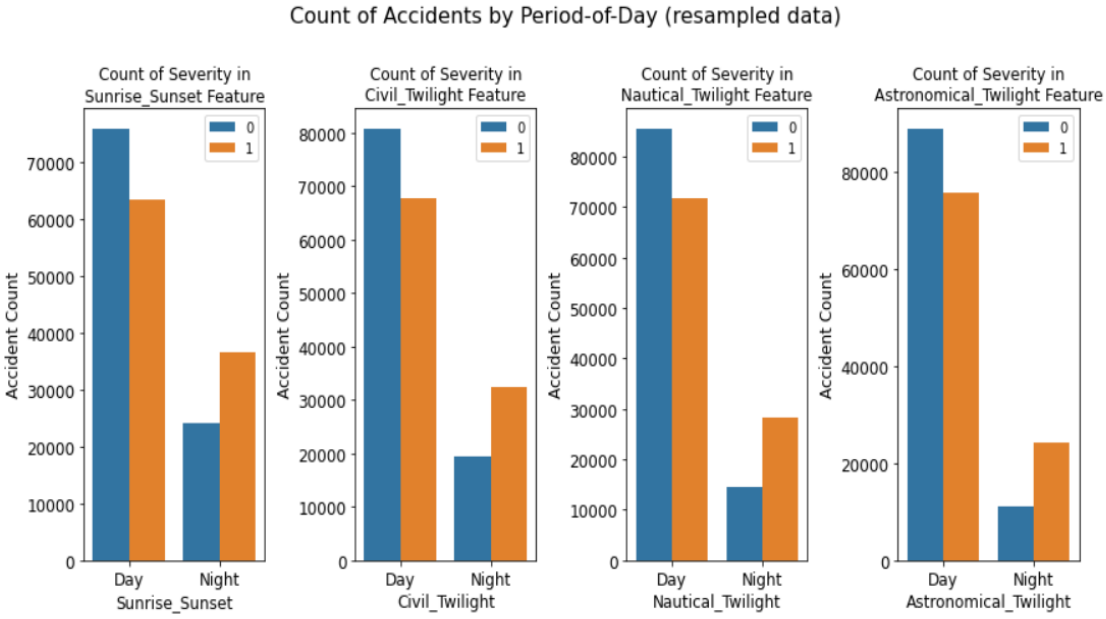
We see that accidents gradually increases from January to June and then rapidly falls down during July and again increases until December.

4.3 Accident count based on month:



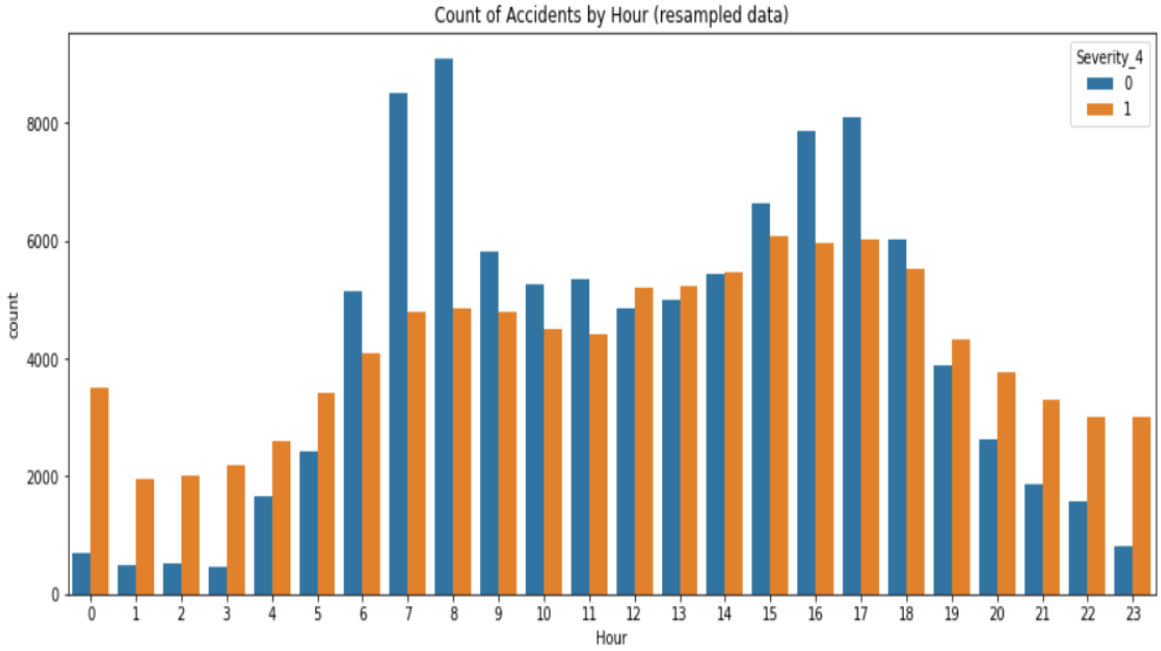
We see that number of accidents are less during weekends although the severity of accidents is high on weekends than weekdays.

4.4 Accident count during period of day:



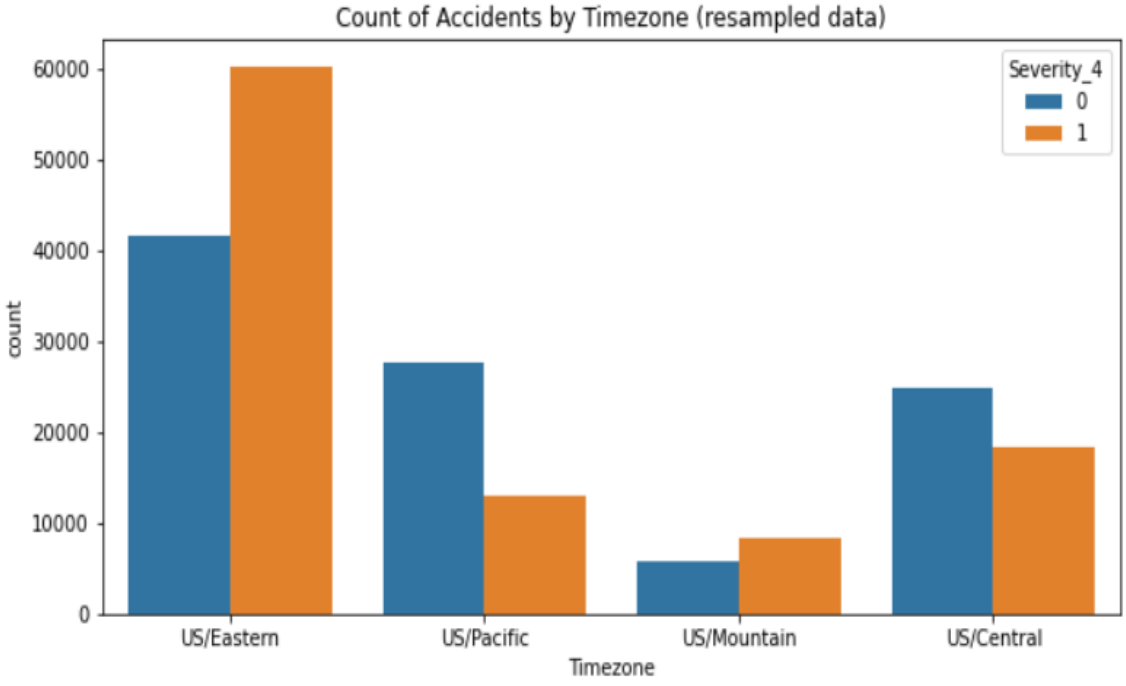
We see that although, the number of accidents was higher during day time severity of accidents was higher during night.

4.5 Accident count by hour (through a day):



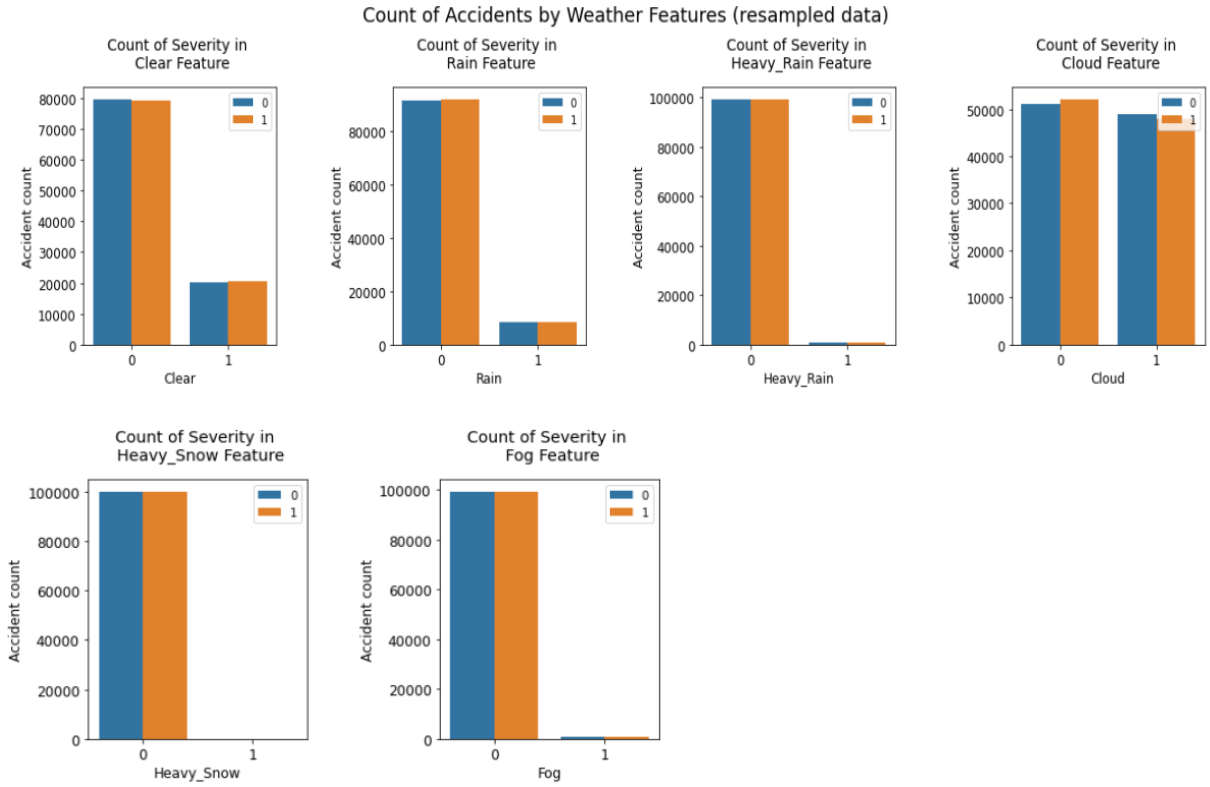
Although accidents are higher during early morning may be due to huge population travelling to work, accidents during night tend to be more severe.

4.6 Accident cases across different time zones:



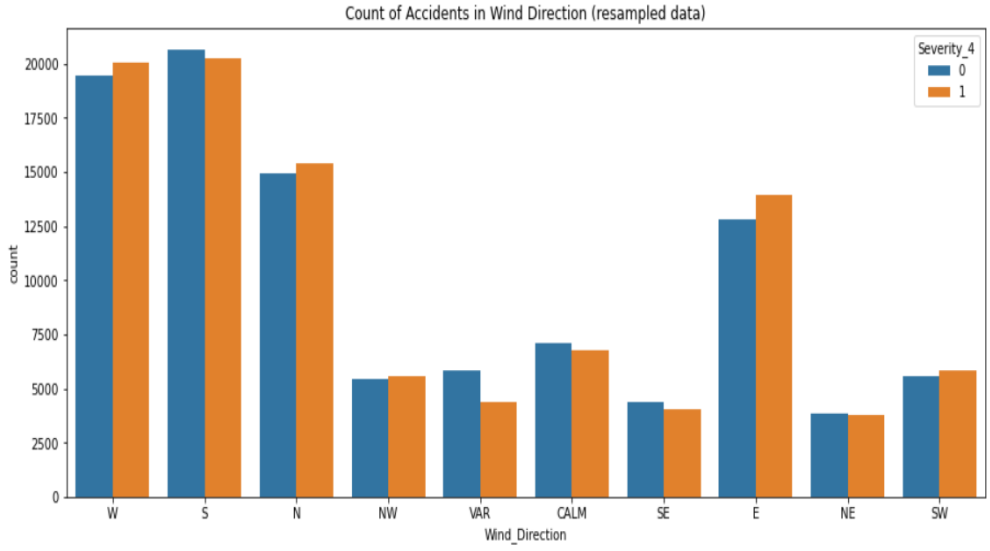
We can see that highest number of accident cases occur in **US/Eastern** Time zone with higher severity.

4.7 Accident count during different weather conditions:



We can see that accidents occur to be more serious during Heavy Rain/Snow rather than a Cloudy weather.

4.8 Accident count during different weather conditions:



Accidents were higher and more severe with Wind directions from West, South and North.

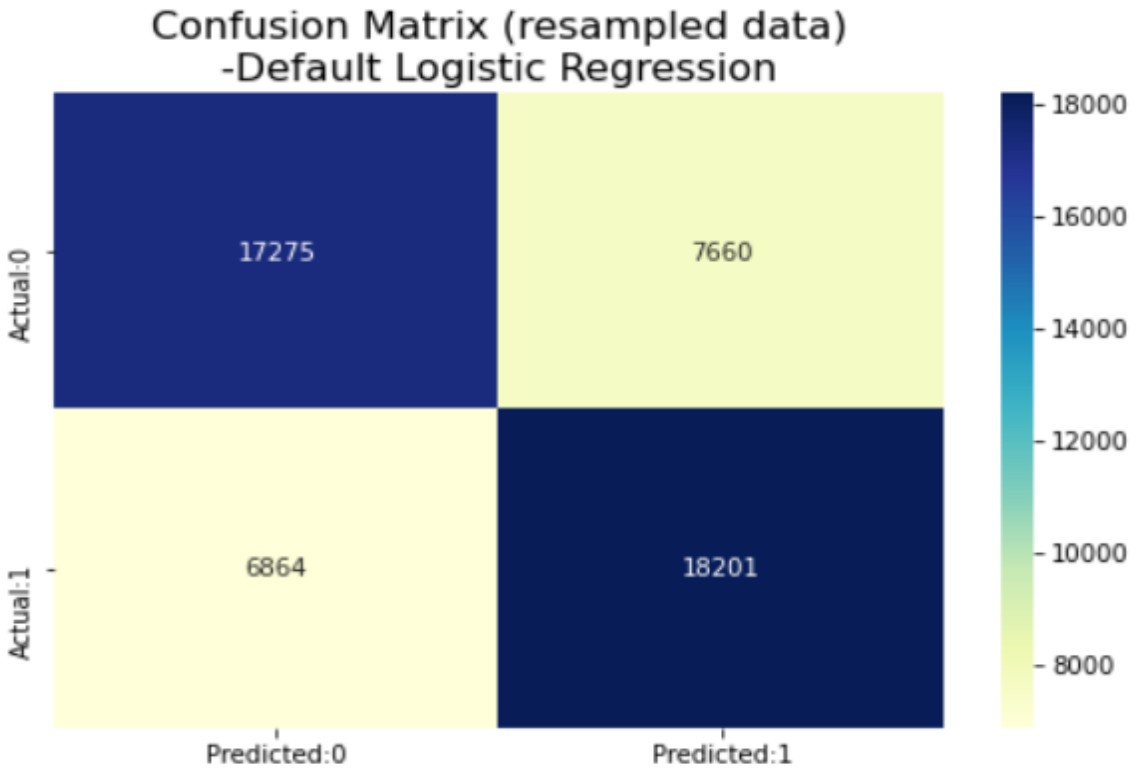
5. Modelling and Prediction:

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can be used to predict the accident numbers, while classification models focus on the probabilities of accident occurrence. I have carried out only classification modelling.

5.1 Logistic Regression Classification:

I divided the samples into two classes 0 (less severe) and 1 (more severe) both of same size.

5.1.1 Confusion Matrix from Logistic regression classifier:

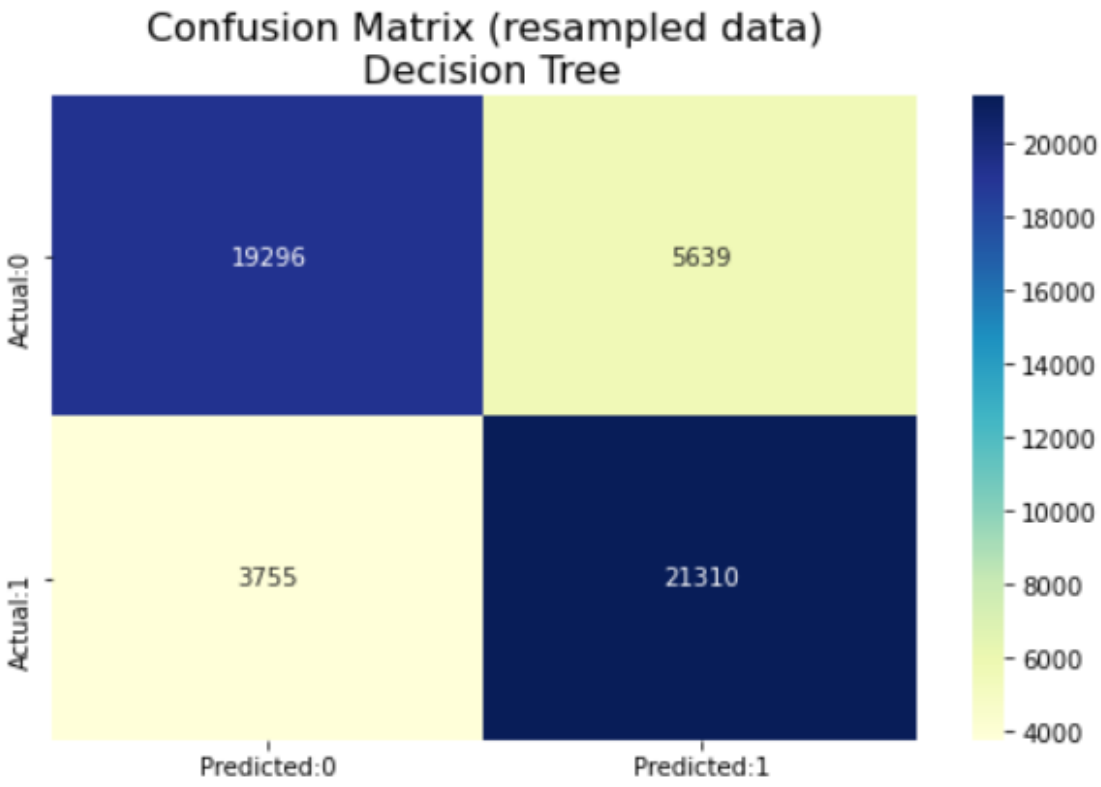


- Accuracy of the train set was found to be 70.9%
- Accuracy of the test set was found to be 71.1%.
- Jaccardian Similarity Index was found to be 0.55

5.2 Decision Tree Classifier:

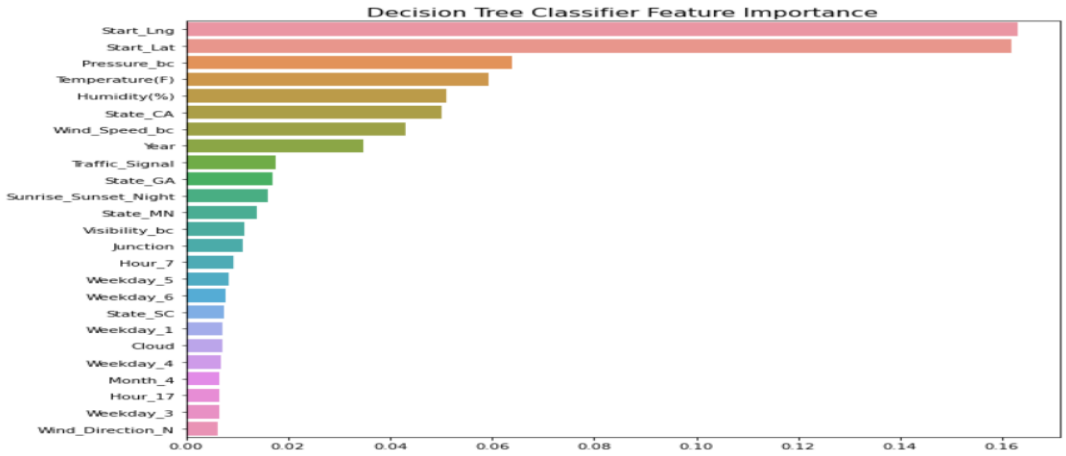
When Decision Tree classifier was used for modelling, model accuracy was found to be higher.

5.2.1 Confusion Matrix for Decision Tree Classifier:



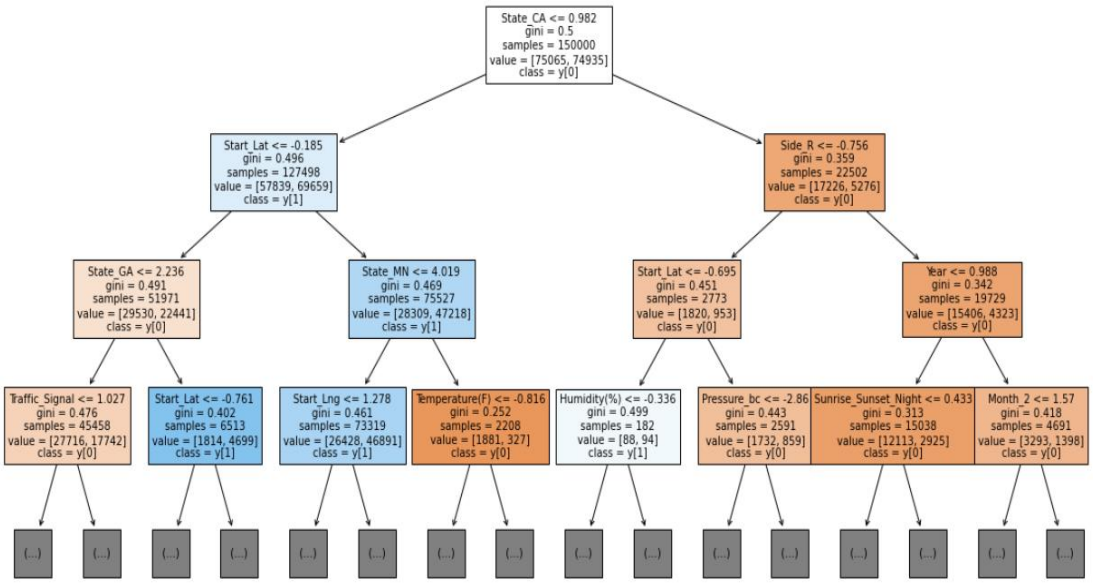
- Train Accuracy was found to be 99.5%.
- Test Accuracy was found to be 81.2%.

5.2.2 Important features for Decision Tree classifier:



Plot shows us the important features for Decision Tree Classifier such as spatial data like longitudes and latitudes at the top spot. Pressure, Temperature and Humidity also shows higher precedence.

5.2.3 Decision Tree Diagram:



6. Conclusion:

- Accident Severity can be predicted with few attributes such as time, period of day, location and weather.
- Accidents during night time are much sever than day time accidents.
- Accidents are much likely to occur on the right side of the road.
- Wind Directions also play an important role in accident severity prediction.
- Number of accidents are higher during early hours of the day.

7. Future Scope:

- ML models created in this project can be incorporated in real time to predict accidents.
- Detailed relations between different important factors can be further studied.