Abstract

On average, 3,700 people die each day on the road due to traffic accidents equating to 1.35 million annually (Association for Safe International Road Travel, 2021).  By analyzing multiple data sets, this study aims to provide an interface to allow users to identify the likelihood of a motor vehicle fatality in a specific geographical area in real time, potentially decreasing the amount of annual fatalities.  The analysis method for this was to utilize R and develop a predictive model to scale the dependent variables to the independent variable of severity. The severity variable is a measure of how intense a given accident was. To further improve the accuracy of predictions, additional analysis was performed using Python along with sklearn machine learning library to cherry pick a list of independent variables that complements the predictive model in yielding better predictions on large accident volumes. This analysis is to be interpreted as a way to see which variables have an effect into the severity of an accident, rather than the likeliness of a fatality occurring given that all accidents in this study were fatal. This study found that a significant number of fatal accidents occurred in dense population areas, so the user dashboard production was centered around one of these areas, the City of Los Angeles.  Significant factors to accident fatality were found to be the following: Weather Conditions, Time of Day, Manner of Collision, Harmful Event, Speed Limit, Location of accident, and Arrival of Emergency Medical Service.  Selection of these factors are evaluated with multiple predictive models and they consistently provide better prediction accuracies when compared to intuitive feature selection. Throughout this analysis we realized a comparison of fatal accidents to non-fatal accidents may have offered more insight into the causation of fatal accidents. However, we continued with the fatal accident dataset for our analysis due to time constraints.

# 3   Analytics and Algorithms

Data Analysis was performed in a 2 step process, the frequency distribution of individual features are studied using RStudio, and feature scores along with their importance was studied using Python and sklearn library. Data was subjected to multiple iterations of cleansing during the course of analysis to identify and filter out data points that only add noise if considered.

## **3.1 Feature Engineering**

An extensive analysis was performed using Python and sklearn machine learning library to select best features from accidents dataset that would closely predict the dependent variable i.e. severity of an accident. Accident data from years 2016 - 2019 is considered for feature analysis which contains approximately 192,514 fatal accident records.

### **3.1.1 Data Cleansing and Preparation**

Accidents dataset contains numerous attributes each with name and value (numerical) columns. Because analysis and predictions can easily be performed on numerical attributes, the name columns are removed. Dataset was scanned for None, NaN, INF and -INF values and none of the records are found to be matching this criteria. Also, outliers identified in a few key columns such as Vehicles total, Persons, and Fatals are removed as it reduces noise while deriving accident severity variable. Latitude and Longitude variables contain positive and negative values. Values in these variables are normalized to represent a scale between 0 and 100.

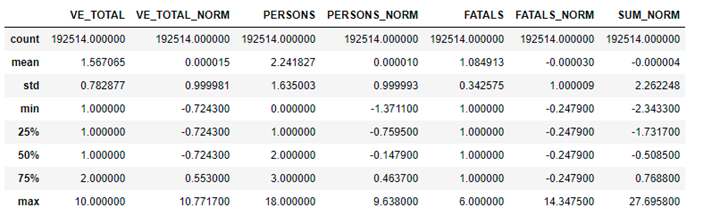
Figure 3.1.1 Frequency distribution of key features after removing outliers

|  |  |  |
| --- | --- | --- |
| Vehicles Total | Persons | Fatals |
| https://lh4.googleusercontent.com/F3tD8YBOTWWgRhIZwZxqAWJ0t9twgIv76-G15NO13LpZS0goRF3qqaJXSe9Atj0kJGC6tLB6qPN-b21Edz6kk7izU_ql66n7j_eukyOclOpP77vUiVNE_V69saNUfSRmlw | https://lh3.googleusercontent.com/gaqE1RiuDZ_jOocO-8i3A9NAFZVrgAcCbzzg4GrsIvs8FGBgj2s3ZfLBROvtoiN43bEwTdA-zQTIJBSEZql-KtTF_JLhJWcDkPdbAdxoJrG4cfN3ByLFe9_JAiel-uGoRg | https://lh4.googleusercontent.com/TIQPpM6V7syYdOaeFYsJLKHv1S7fe6fOutJ5CvS9PQd9LWEZTRO5VrkdeQqgiOp-8Hciq1rTtqsTQTuRTJrK4B4g_7-WvKxyWW8HIdOjuWO4mp6YOdukWOXTWnS-ZuB-Kg |

### **3.1.2 Derivation of Dependent Variable**

Dependent Variable ‘Severity’ is derived from key variables that together may gauge the intensity of crash. Vehicles total, Persons, and Fatals are represented in different scales and the only way to combine them all together is by normalizing the values in these variables to a common scale. This was achieved by calculating the z-scores for individual variables. The resultant values in all 3 columns after the normalization range between -1.37 to +14.34 with mean close to 0 and standard deviation close to +1 for each of the normalized variables.

Figure 3.1.2 Statistics of key features after being normalized



Adding the normalized values of each of 3 variables resulted in summation column (SUM\_NORM) that was used to derive the ‘Severity’ dependent variable

|  |  |
| --- | --- |
| **SUM\_NORM value** | **Severity** |
| Sum\_norm value < 25% of distribution | Low |
| 25% <= sum\_norm value < 75% of distribution | Medium |
| Sum\_norm value >= 75% of distribution | High |

### **3.1.3 Feature Analysis**

Sklearn machine learning library is utilized to identify the scores and importance of all 49 features in the dataset. SelectKBest model of sklearn is leveraged to identify feature scores and ExtraTreesClassifier model of sklearn is leveraged to identify feature importance. Based on the statistics gathered, top 20 features are selected based on their scores and importance in the dataset to help predict the severity of an accident.

Figure 3.1.3 Scores and Importance of top 20 features selected by sklearn models

|  |  |
| --- | --- |
| Feature Scores | Feature Importance |
| https://lh5.googleusercontent.com/LeYZw0XXWivGyxAhu_61z8e5lMAAl_JKAWaGhfq-Nj7HDmVKrF66NAxxXbuAWy2_pILCGBZYNszniSzoKpsvcuoJLXOJw4XI4W9-9NPbq8PT674_Tg6FfAFvQMObDw6Jbg | https://lh6.googleusercontent.com/-_zNNFoFl5S-65qVpq9nvJLZ4pi-zbkDPS1fADBEEH5SesmGbRMNgig0BwBC4t4fQ1pWyoiyw4FRqNo57ZoUOt7eHJJPYqevC_3z390MXb-Asn3sEG5_Pptgs7Fwx-hR7A |

### **3.1.4 Feature Evaluation**

Based on the set of features selected intuitively before performing feature engineering process; and set of features selected from analysis after performing feature engineering process; different predictive models were built and prediction accuracies identified from model executions.

Scenario 1 – Features selected intuitively by looking at the dataset.

'DAY\_WEEK', 'MONTH', 'HOUR', 'LGT\_COND', 'WEATHER'

Scenario 2 – Features selected after performing feature analysis as described in previous section.

'MILEPT', 'COUNTY', 'STATE',  'CITY', 'ROUTE', 'LATITUDE', 'LONGITUD',                  'MAN\_COLL', 'HARM\_EV', 'RD\_OWNER', 'REL\_ROAD', 'CF1', 'CF3', 'PVH\_INVL', 'WEATHER',  'NOT\_HOUR', 'ARR\_HOUR', 'ARR\_MIN', 'DAY\_WEEK', 'DAY', 'NOT\_MIN', 'MONTH', 'HOUR', ‘PEDS', 'PERNOTMVIT', 'HOSP\_HR', 'HOSP\_MN', 'FUNC\_SYS', 'YEAR'

Figure 3.1.4 Evaluation of selected features

|  |  |
| --- | --- |
| Performance of Linear Regression Model | Performance of other Models |
| https://lh3.googleusercontent.com/D8nCb8430W21EQSyvY_UizJz0Bse8BDYzPda1-IvyNUPKN8k5FOBAJ63QCYog9uV7mOFv_1FnBZmrdmOCGWiwalEC9VOBDtL1RiRDwswyqGD7OosFVjLiFNtC6Llsx6kUA | https://lh5.googleusercontent.com/fulHVyZrPO3MpE7BNKhTuTxOLz_R9vCESHIXWzee4z0ic5X8EYwPe9MlUAkVFNWtgwDmaP-eDfQ_diaFLlfU_hDUTxbZyQFh3HjlbGlvuctpqUnWs3dOC0LJTXgqweuLMQ |

MAE - Mean Absolute Error

MSE - Mean Squared error

RMSE - Root Mean Squared Error

Noticing the performance evaluation comparison among different predictive models in the figure above, it can be concluded that features from Scenario 2 i.e. features selected from the feature engineering process are consistently more responsive and yield better predictive accuracies.