Prediction of Accident Severity

Section 1: Introduction

Motivation

Road accidents are a serious concern for the majority of nations around the world because accidents can cause severe injuries and fatalities. According to the World Health Organization’s Global Status Report, approximately 1.25 million people deaths happened per year are because of road accident injuries, and most fatality rates were in lower income countries [1]. Our motivation is to predict the accident severity of any road, which will play a crucial factor for traffic control authorities to take proactive precautionary measures.

Objective

The purpose of this project is to predict the severity of an accident by training an efficient machine learning model with the help of existing accidents data from 2005-2015.

Section 2: Design & Implementation

Algorithms, technologies, and tools

Six classification algorithms were used and evaluated to predict the accident severity. Algorithms considered for classification were K-nearest neighbors, Random Forest classifier, Logistic Regression, Gradient Boosting classifier, XGBoost classifier, SVM.

K-Nearest Neighbors Classifier: This model is a non-parametric method used for classification. We tried using different values of neighbors and got the best result for considering three neighbors for each point and weights parameter was set to ‘distance’ which weights the points by an inverse of their distances. This model didn’t perform well on this dataset. F1- score of Accident Severity prediction from this model was 0.57.

Random Forest classifier: A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The F1 scores are .60 for ‘severe’ class and 0.62 overall for the undersampled data.

Logistic Regression: This model is the basic and popular for solving classification problems. Unlike Linear Regression, Logistic regression model uses a sigmoid function to deal with outliers. On this undersampled data, logistic regression model performed better than using ‘balanced’ class weights. As both the classes were not exactly separable, this model gave the f1-score equals 0.62.

Support Vector Machine: SVMs are based on the idea of finding a hyperplane that best divides a data set into two classes. As the features in this dataset were very sparse and nonseparable by any hyperplane, SVM did not work at all for this dataset. For larger dataset, SVM training time is high. When we tried to train the dataset using SVM it ran forever and did not fit the model.

XGBoost : This algorithm, in general, performs faster compared with other algorithms due to the parallel tree boosting method.

Gradient Boosting Machine (GBM): Gradient Boosting is another type in ensemble method which is very popular in decision tree algorithm. GBM performs well in reducing mean squared error (MSE) . This algorithm performed well on undersampled data in terms of both precision and recall values, however, the algorithm is not able to predict even one rarer class when applied on full data set.

Technologies & Tools used:

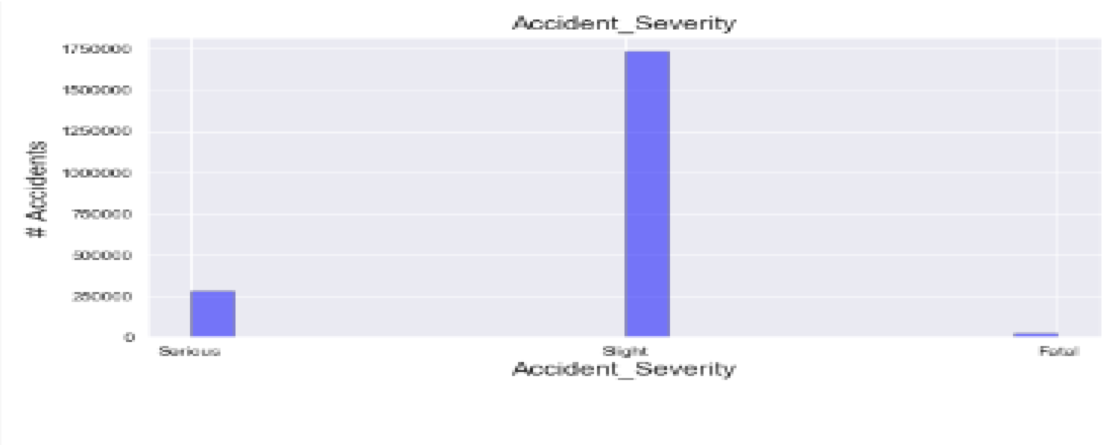
For developing this project, below tools and technologies have been used.

Python: Python is easy to understand language and has a rich set of librarys to use for data pre-processing, modeling, and evaluating the algorithms. Moreover, python has very good community support which is very useful for debugging the code.

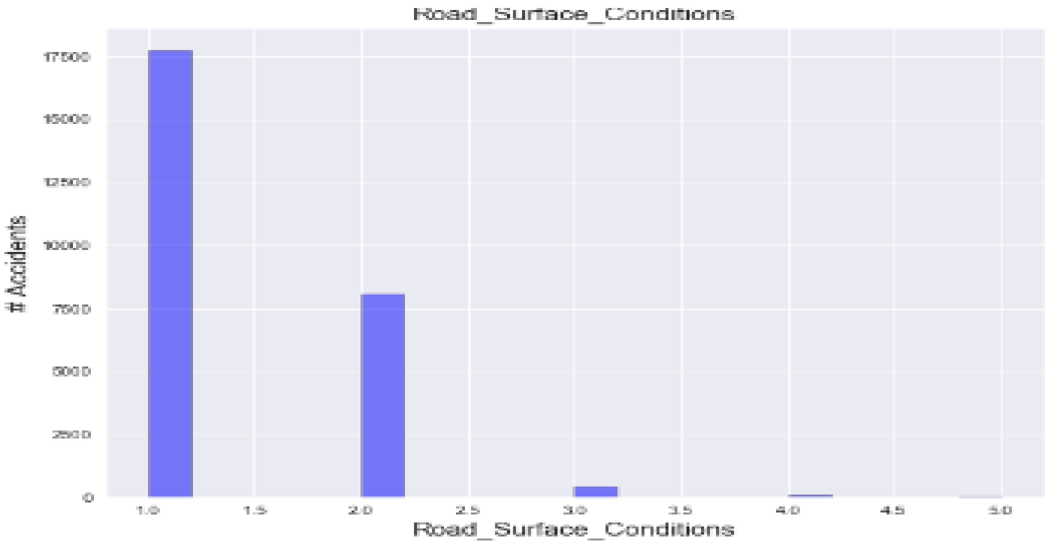
Jupyter Notebook: Jupyter notebook is a simple and interactive tool for running python code. Also, it has many different sets of features such as downloadable to .py, .ipynb, and .html files.

Section 3: Data Analysis

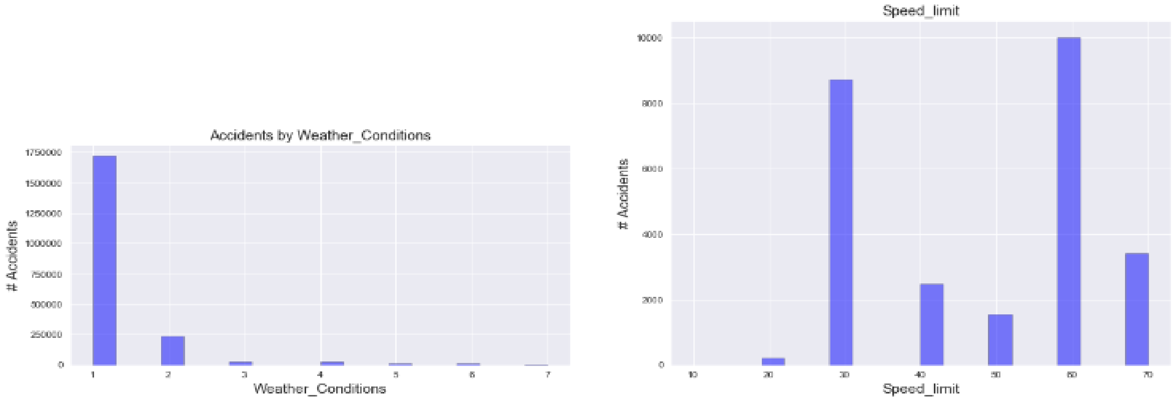
Dataset used: We have used UK road traffic accidents dataset for our project. The dataset contains 1,920,000 records and 34 columns including weather conditions, Road class, road type, junction details, road surface conditions, light conditions, etc. Our dataset is very imbalanced with data corresponding to slight severity is 84.84%, serious severity is 13.86% and for fatal severity is 1.30%. For validation of data, we have separated the dataset based on the accident severity and choose data for train and test dataset in equal ratios. We have used k-fold validation with 5 folds from each subset



The above figure shows the Total accident counts with accident severity as Slight, Serious and Fatal.



The above figure shows the Total accident counts with Road Surface Conditions.



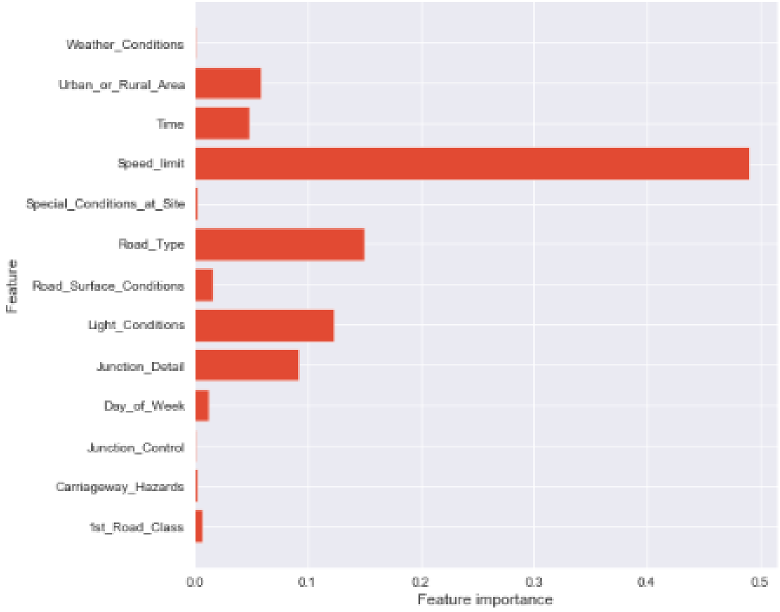
The above figure shows the Total accident counts with Weather_conditions and Speed limit.

Feature Selection:

The dataset has 34 attributes describing the incident of an accident. There are mixed types of data such as continuous and categorical. Manually dropped few columns due to its inconsistency in values such as Accident ID, and Location ID. For selecting the best features, below functions are used from sklearn library.

1. SelectKBest: SelectKBest is a sci-kit learn library provides the k best features by performing statistical tests i.e., chi squared computation between two non-negative features. Using chi squared function filters out the features which are independent of target attribute.

2. Recursive Feature Elimination (RFE): RFE runs the defined model by trying out different possible combinations of features, and it removes the features recursively which are not impacting the class label. Logistic regression algorithm is used as a parameter for RFE to decide on features.



The above figure shows the feature importance for all the features in decision classifier model.

Section 4: Discussion & Conclusions

Decisions, difficulties and discussions:

Our main aim was to predict the severity of the accident when it is “serious” and “fatal”. It was very difficult to handle this large-sized data. Using HPC we were able to run most of our algorithms. Data is highly imbalanced so even though most of our algorithms were giving > 89% accuracies, it was of no use. It was predicting all the accidents as slight accidents. After checking on all these algorithms, the team even tried dimensionality reduction techniques and but the results were not improved. Then the team decided to use the undersampled dataset as it was giving better results in predicting the severe/fatal accidents. This decision was made on trying out oversampling, undersampling, test and train data with an equal ratio of classification classes.