**Accident Severity Prediction**

**Introduction/ Business Problem**

The Washington State Department of Transportation Crash Data Portal provides crash information for accidents that occurred state-wide. According to the 2019 data, there were 45,524 accidents on all roads. Of those:

* 235 were fatal crashes
* 973 were suspected of serious injury accidents
* 2,798 were suspected of minor injury accidents
* 9,412 were possible injury crashes
* 32,106 were no apparent injury collisions

Our motivation is to use the weather, location and road condition data provided in the dataset, made available by the Seattle Department Of Transportation Traffic Management Division, to arrive at a correlation to predict the severity of road accidents. This tool/data can then be made available to the public and the Seattle traffic authorities to possibly prevent/reduce severe or fatal accidents in the future by taking precautionary measures.

**Data Understanding**

We chose the unbalanced dataset provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. It covers accidents from January 2004 to May 2020. Some of the features in this dataset include and are not limited to Severity code, Location/Address of accident, Weather condition at the incident site, Driver state (whether under influence or not), collision type. Hence we think its a good generalized dataset which will help us in creating an accurate predictive model.

The unbalance with respect to the severity code in the dataset is as follows.

|  |  |
| --- | --- |
| SEVERITY CODE | Count |
| 1 | 136485 |
| 2 | 58188 |

**Data Pre-processing**

An unbalanced dataset is used, provided by the Seattle Department of Transportation Traffic Management Division with 194673 rows (accidents) and 37 columns (features) where each accident is given a severity code. The steps taken in pre-processing the dataset are as follows.

1. **Removal of irrelevant columns or features**
2. **Identification and handling missing values**
3. **Balancing the dataset**
4. **Encoding of data**
5. **Splitting into training and testing datasets**
6. **Normalizing/ Feature scaling of data**

**Understanding Correlation in Dataset**

**Machine Learning Algorithms**

* **Logistic Regression Classifier**

It’s a classifier that estimates discrete values **(binary values like 0/1, yes/no, true/false)** based on a given set of an independent variables. It basically predicts the probability of occurrence of an event by fitting data to a logisticfunction**.** Hence it is also known as**logistic regression**. The values obtained would always lie within 0 and 1 since it predicts the probability.

* **K Nearest Neighbours Classifier**

K nearest neighbours algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbours. The case assigned to the class is most common amongst its K nearest neighbours measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming).

* **Naïve Bayes Classifier**

Naive Bayes classifies objects based on Bayes' Theorem with an assumption that the predictors (features) are independent of each other. Bayes theorem is a way to calculate posterior probability P(c|x) from the P(c), P(x), P(x|c). Naive Bayes is naive because it assumes the presence of a particular feature is completely unrelated to the presence of another, and each of them contributes to the posterior probability independently.

* **Decision Tree Classifier**

Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets (leaves) based on the most significant differentiators in the input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

* **Random Forest Tree Classifier**

Random Forest Classifier is an ensemble (algorithms which combines more than one algorithms of same or different kind for classifying objects) tree-based learning algorithm. RFC is a set of decision trees from randomly selected subset of training set. It **aggregates the votes from different decision trees** to decide the final class of the test object. Used for both classification and regression.

* **Support Vector Machine Classifier**

Support Vector Machine is an algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes.

**Results**

**Conclusion**

**Future Work**

Cross Validation