1. **Introduction/Business Problem**

For years, road accidents have ravaged lives causing physical, financial and mental effects on individuals involved and families. These accidents can be individually driven (Driving Under Influence (DUI), poor vehicle maintenance, exceeding speed limits etc.) or physically driven (bad roads, unfavorable weather conditions, inadequate traffic signs and symbols). Road accidents are a danger to every society and as such national governments recognize this and put in place necessary infrastructures that can help alleviate this problem and keep the public safe. However, the efficacy of these infrastructures are often times poor and driven by little or limited considerations.

Therefore, the objective of this project is to build a model that takes into consideration both physical and non-physical parameters in order to predict accidents severity and help mitigate this menace. Seeing as the demographic of road users is conventionally limited to age, the recipients of this project is aimed at the general public. In addition, national governments can also benefit from this project as the model could be used to determine areas that are at high risk of severe accidents and allocate the necessary resources to these areas to help combat this issue.

1. **Data acquisition and cleaning**

**2.1 Data Source**

The data to be utilized for this project is the Seattle accident data which has been sourced from Kaggle (an opensource platform for datasets) and made available in csv format. The dataset contains 38 attributes including both physical and non-physical parameters and 194,673 cases of road accidents.

**2.2 Data Cleaning**

The data downloaded was read into a jupyter notebook and displayed as structured data. The data contained quite a lot of missing values and certain columns were missing over 50% of values. These columns that had over 50% of missing values were dropped from the dataset. Other columns had few missing values and would have little effect if they were dropped and as such, those rows were dropped and a new data frame was created.

**2.3 Feature Selection**

Upon completion of data cleaning, the dataset was left with 180,067 samples and 31 attributes. After examining the remaining features, it was clear that certain features were unnecessary as they served as descriptors for other features. For instance, the SEVERITYDESC feature was used to give a detailed description of the SEVERITYCODE feature. This action was performed across the data frame on other features that simply acted as descriptors.

Table 2.1: Features retained during data cleaning

|  |  |  |
| --- | --- | --- |
| **Features Retained** | **Features Discarded** | **Reason for Discarding** |
| SEVERITYCODE, X, Y, OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, ADDRTYPE, LOCATION, SEVERITYCODE.1, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INCDATE, INCDTTM, JUNCTIONTYPE,  SDOT\_COLCODE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, ST\_COLCODE, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR | SEVERITYDESC,  SDOT\_COLDESC, ST\_COLDESC. | Features contained description of other columns and was irrelevant. |

1. **Methodology**

**3.1 Definition of Target Variable**

The objective of this project is to predict road accident severities. The feature SEVERITYCODE contains information regarding this and such was defined as the target variable of the model.

**3.2 Exploratory Data Analysis (Numeric Features)**

In the initial stage, data analysis and visualization techniques were employed to derive initial insights into the data and also determine the level of correlation that exists between the label and predictors. Since this project aims to be predict accidents severity, the SEVERITYCODE attribute was used as the label.

I examined the Pearson correlation and P-value of the numerical independent features and found out that all showed little or no correlation with the dependent variable (correlation values were all less than 0.3). Figure 3 shows that little correlation existed between the numerical columns and the target variable. Some of these relationship between the label and independent variables are explored more in this chapter.

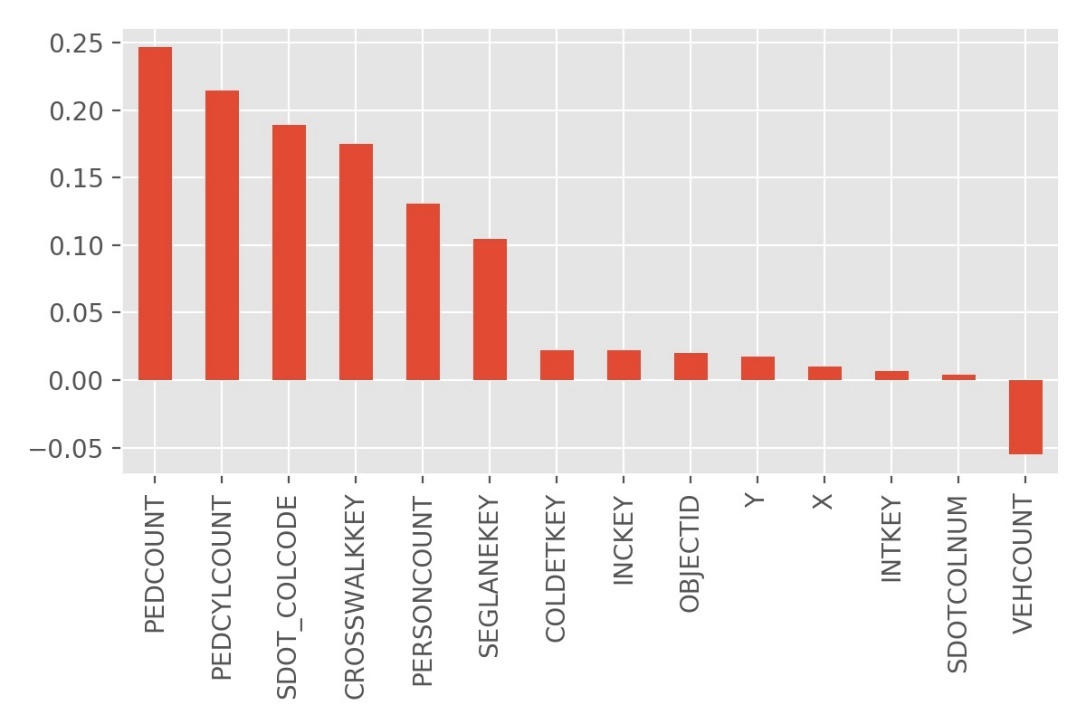


Figure 3.1: Bar plot of Pearson correlation between target variable and numerical independent variables

**3.2.1 Relationship between severity and pedestrian count**

To explore the relationship between severity and pedestrian count a regression plot was generated to visualize if any relationship occurs. However, the points were not randomly distributed and as such the peasron correlation coefficient had a value of 0.25 (indicating little correlation), while the p-value was given at 0.0 indicating a weak evidence that the correlation is significant. This feature was dropped from the dataframe.

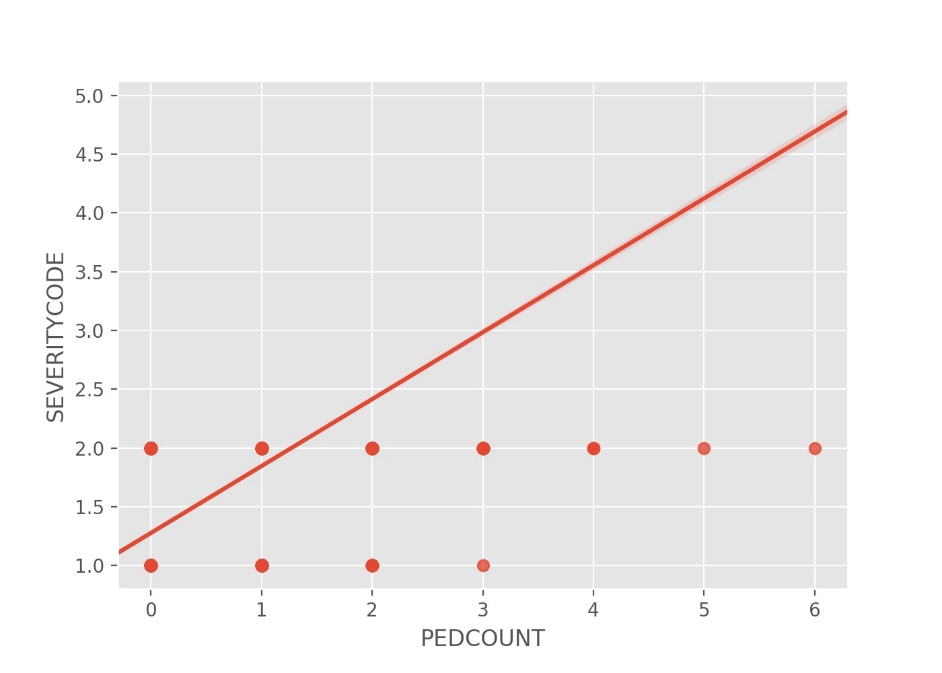


Figure 3.2.1: Regression plot of label (SEVERITYCODE) and PEDCOUNT (pedestrian count)

**3.2.2 Relationship between severity and number of bicycles involved in collision**

The relationship between severity and the number of bicycles was visualized through a regression. As in the case of pedestrian count, the points were not randomly distributed and as such the peasron correlation coefficient had a value of 0.21 (indicating little correlation), while the p-value was given at 0.0 indicating a weak evidence that the correlation is significant. This feature was dropped from the dataframe.

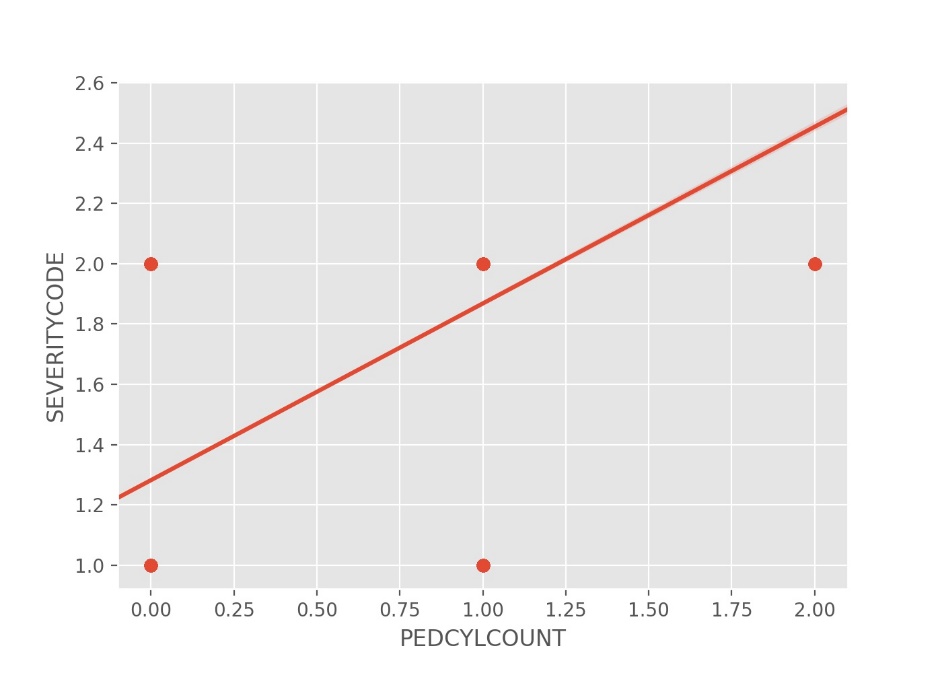


Figure 3.2.2: Regression plot of label (SEVERITYCODE) and number of bicycles involved in collision

**3.3 Exploratory Data Analysis (Categorical Features)**

**3.3.1 Hit parked cars**

The HITPARKEDCAR feature represented a categorical column showing if a parked car was hit. The number of cars that weren’t hit overwhelmed the numbers of cars that were hit, and thus created a bias. This bias could have an effect of the model and was therefore dropped from the dataset.

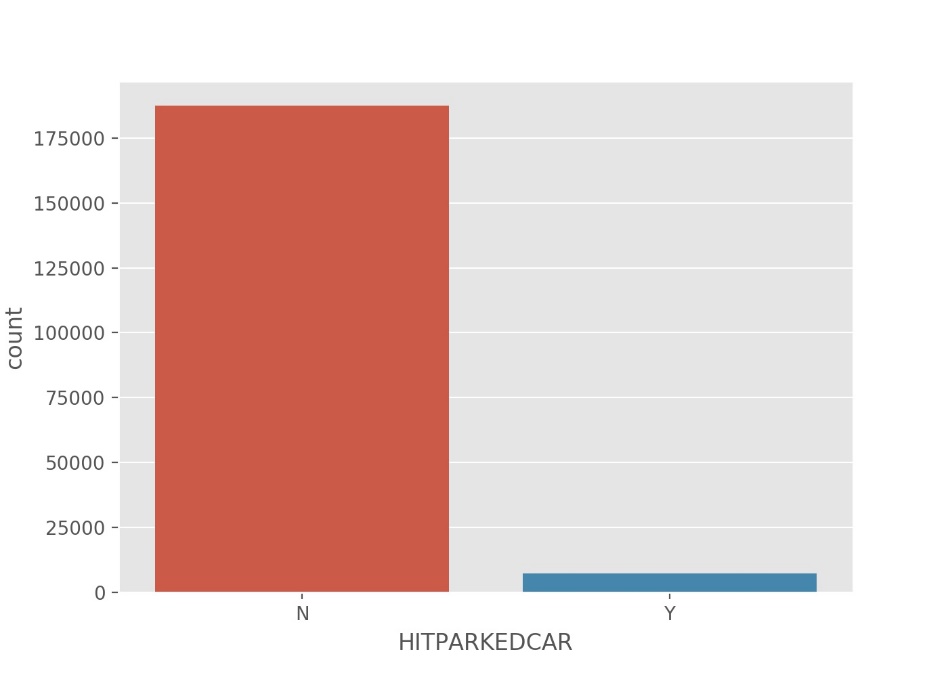


Figure 3.3.1: Count plot of hit parked cars

**3.3.2 Weather Conditions**

The categories within the WEATHER column showed some variation with most of the accidents occurring while the weather was clear. This variation would be useful in building the model, so this column was retained.

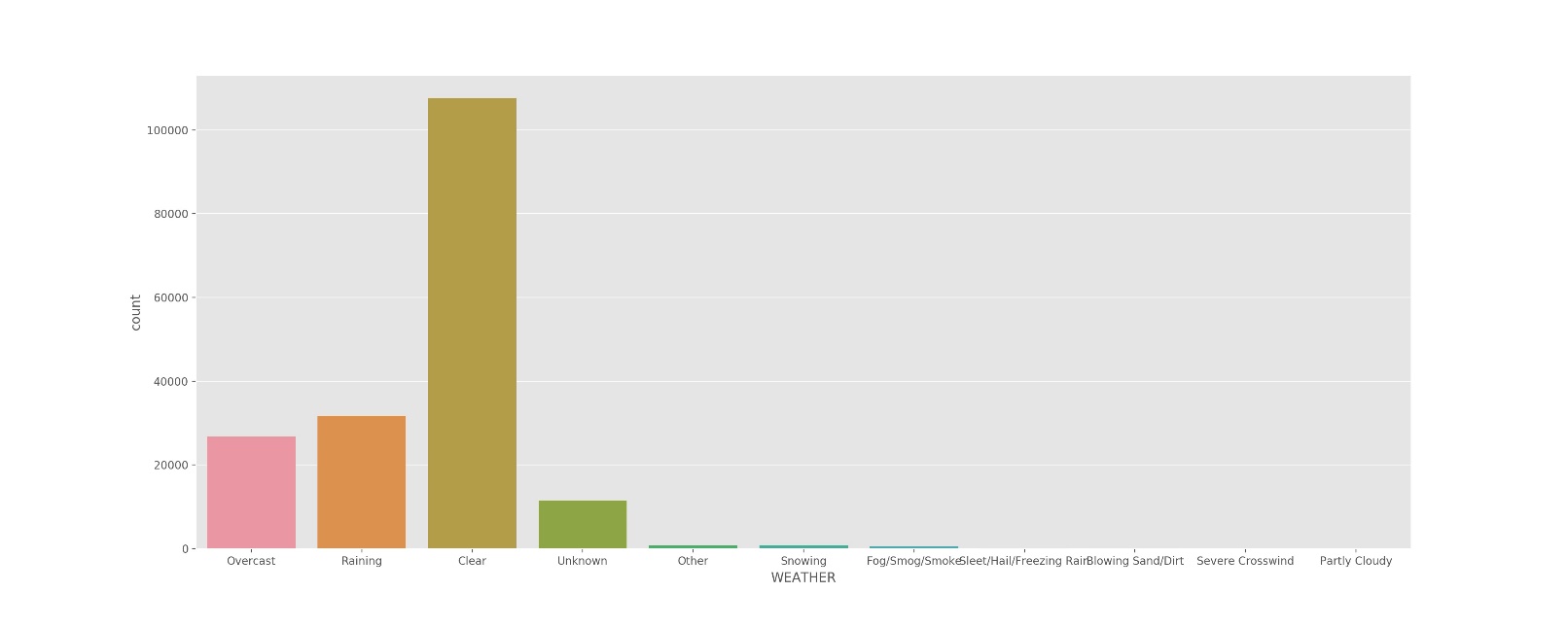


Figure 3.3.2: Count plot of weather conditions

**3.3.3 Collision Types**

The collision variable shows variation within the categories, were must collision occurred with parked cars. The angles and rear-ended collisions also showed high counts. The least collisions made were head-on collisions.

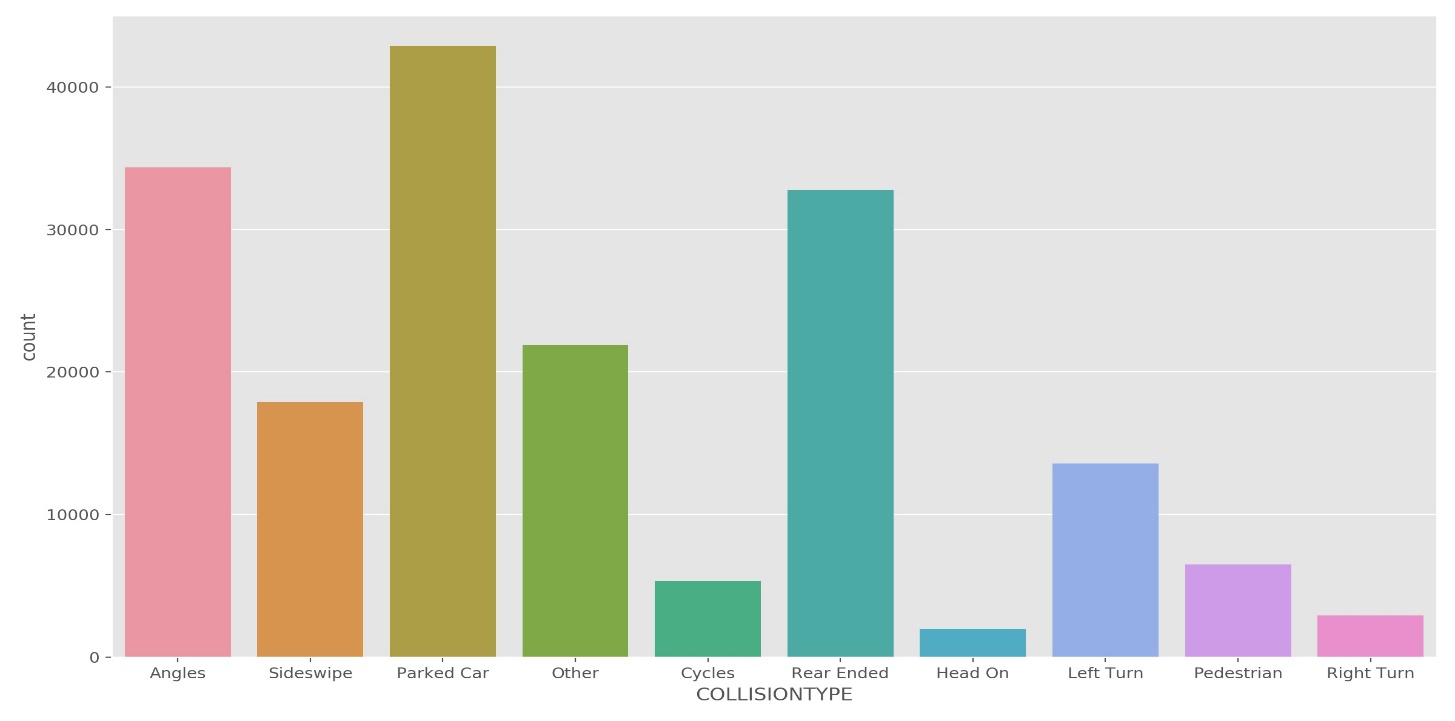


Figure 3.3.3: Count plot of collision types

**3.3.4 Light Conditions**

The light conditions within the data frame showed a bit of variation with daylight accounting for the overwhelming value within the feature. This tells us that majority of the collisions occurred during the day.

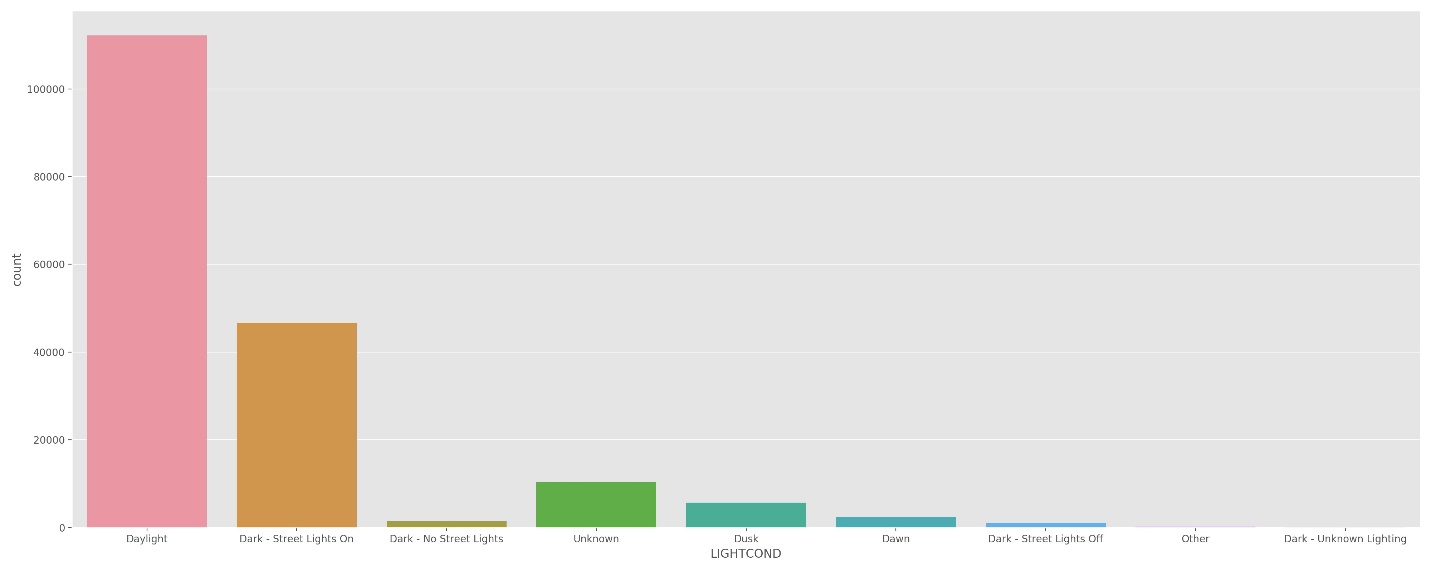


Figure 3.3.4: Count plot of light conditions

Table 3.1: Features retained during exploratory data analysis

|  |  |  |
| --- | --- | --- |
| **Features Retained** | **Features Discarded** | **Reason for Discarding** |
| SEVERITYCODE, REPORTNO, STATUS, ADDRTYPE, LOCATION, COLLISIONTYPE, INCDATE, INCDTTM, JUNCTIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, ST\_COLCODE, HITPARKEDCAR | OBJECTID, INCKEY, COLDETKEY, SEVERITYCODE.1, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT\_COLCODE, SEGLANEKEY, CROSSWALKKEY, X, Y | Features showed no correlation to target variable. |
| SEVERITYCODE, REPORTNO, STATUS, ADDRTYPE, LOCATION, COLLISIONTYPE, INCDATE, INCDTTM, JUNCTIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, ST\_COLCODE | HITPARKEDCAR | Feature showed too much bias. |
| SEVERITYCODE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND | REPORTNO, STATUS, ADDRTYPE, LOCATION, INCDATE, INCDTTM, UNDERINFL, ST\_COLCODE | Irrelevant for prediction. |

**3.4 Construction of final data set**

After dropping the irrelevant features, we were left with 5 independent variables (COLLISIONTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND) and the dependent variable (SEVERITYCODE). The independent features were categorical features and needed to be one-hot encoded for it to be acceptable by the model. A list containing the independent features was created and one-hot encoded using the get\_dummies pandas’ method to get a new data frame. This new data frame was called final\_data and consisted of 180,067 samples and 42 columns.

**3.4 Splitting data into train and test sets**

Upon completion of the construction of the final data set, the next step was to get our train and test data. To split our data into train and test sets the train\_test\_split function from the sklearn model was imported and 70% of the data was assigned to training, while 30% was assigned to testing. A random state of 101 was also chosen.

1. **Predictive Modeling**

**4.1 Classification Model**

The models chosen were classification models as the target variable was a binary classification. Three classification models were chosen for this project, which included Decision Trees, Logistic Regression and Support Vector Machines. The train and test data were used for all three models and were tuned and built.

1. **Results and Discussions**

The results showed that all three models had equal performance, however, logistic regression accounted for the highest number of false positives with 405, while both decision trees accounted for highest true positives with 36,934. In this instance, lower false positive rate is more important than higher true positive rate. In other words, fewer false positive rates tell us that severe accidents are less likely to be falsely classified. The models all had similar AUROC values and as such the plots were stacked up on each other.

Table 5.1: Performance of classification models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Decision Tree** | **Logistic Regression** | **SVM** |
| **Accuracy** | 0.742 | 0.742 | 0.742 |
| **No. of True Positives** | 36,934 | 36,932 | 36,934 |
| **No. of False Positives** | 403 | 405 | 403 |
| **No. of False Negatives** | 13,530 | 13,526 | 13,529 |
| **No. of True Negatives** | 3,154 | 3,158 | 3,155 |
| **AUROC** | 0.589 | 0.589 | 0.589 |

1. **Conclusion**

In this study, I analyzed the relationship between road accidents severity and physical and non-physical data. I identified collision type, junction type, weather, road condition, light conditions as the most important features that can determine severity of road accidents. I built classification models to predict how much an accident could be severe by using three different models. These models could be of grave importance in helping national governments in predicting how severe accidents could be and allocate necessary resources to save lives.