Prediction of Car Accident Severity

Mumtaz Murat Arik

Outline

1. Introduction
2. Dataset
3. Feature Selection
4. EDA
5. Predictive Models
6. Conclusion

1. Introduction

Problem statement

In this project we look at the severity of car accidents. Severity is grouped in two categories: 1 Property Damage Only Collision and 2 Injury Collision. We believe that municipalities can use this prediction to improve roads and road signs. Drivers can be more cautious due to some specific conditions.

2. Dataset

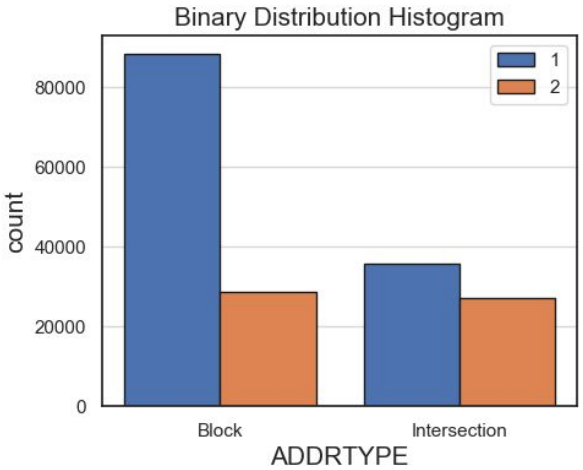
This data set is collected by SDOT Traffic Management Division, Traffic Records Group. It contains collisions that happened in Seattle from 2004 to present. Data set obtained from a csv file and it has 194673 rows and 38 columns including the target column. Our target column is SEVERITYCODE. SEVERITYCODE has two values 1 and 2. 1 is Property Damage Only Collision and 2 is Injury Collision. So this is a binary classification problem.

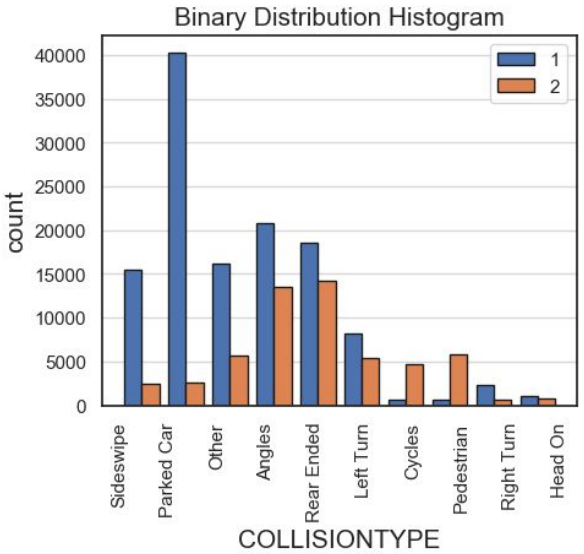
Some features are longitude, latitiuture, location, time, date, collision type, weather, road condition, light condition.

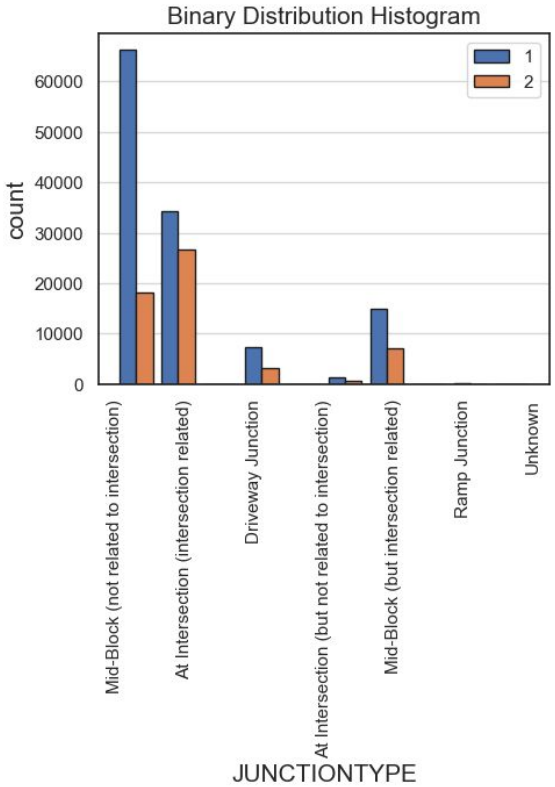
3. Cleaning and Feature Selection

- This is a slightly unbalanced dataset. Percentage of Severitycode categories are: 1 = 70%, 2 = 30%
- Columns with high missing values are dropped(INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'INATTENTIONIND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', and 'SPEEDING
- Rows with missing values are dropped
- Unique values are checked in each column and features represent identification numbers are dropped.(['OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO'].)
- X, Y and LOCATION columns are also dropped because these columns are related to the location of the accident.
- SEVERITYCODE.1 and SEVERITYDESC features are same as SEVERITYCODE so I dropped them.

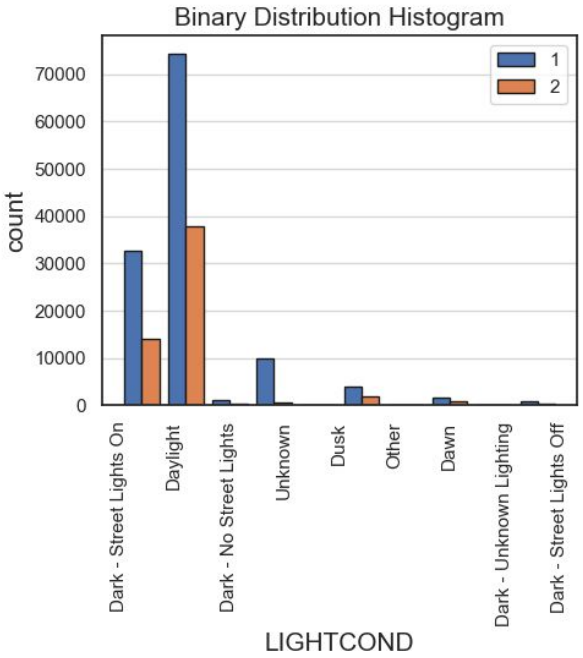
4. Exploratory Data Analysis Plots-1

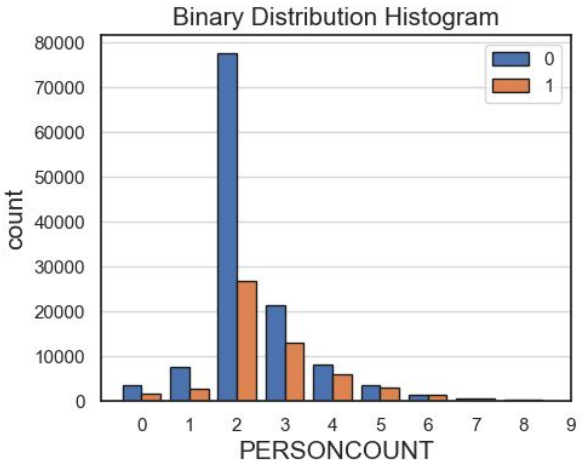


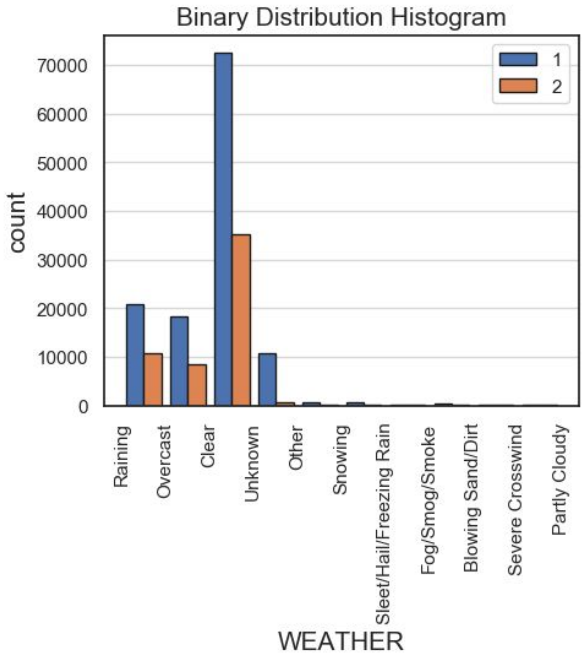




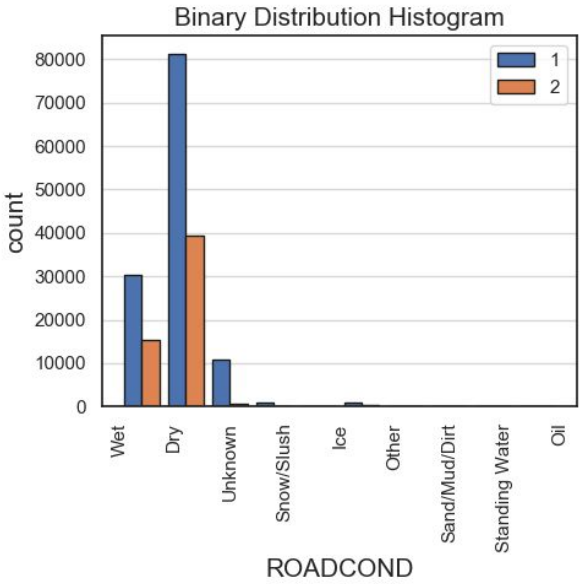
4. Exploratory Data Analysis Plots-2

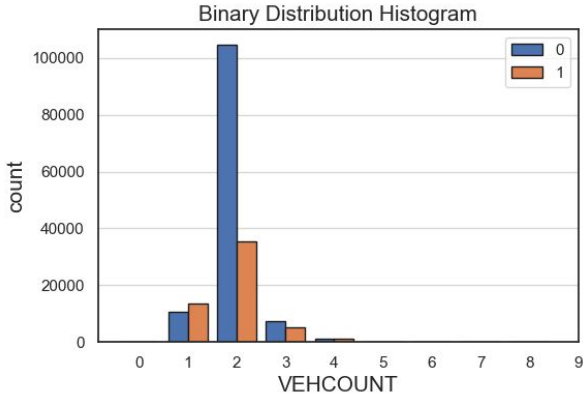


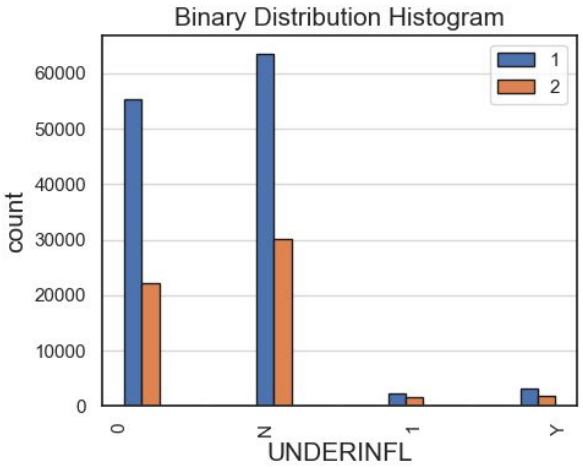




4. Exploratory Data Analysis Plots-3







5. Predictive Models

- Using train_test_split function data set is divided into train(X_train, y_train) and test sets (y_train, y_test) with ratio 80:20.
- Train sets are trained with 3 different machine learning models. Logistic Regression, Random Forest Classifier, and K Neighbours Classifier.
- The data set is unbalanced, because of this sampling is applied. For each model I also did random under sampling model(RUS). I reported confusion matrices and classification reports for all base and and their RUS models.
- Weighted Average F1-Score:

Model Name	Base Model	RUS Model
Logistic Regression	0.67	0.63
Random Forest Classifier	0.72	0.68
K Neighbours Classifier	0.72	0.70

AUC Score:

Model Name	Base Model	RUS Model
Logistic Regression	0.57	0.64
Random Forest Classifier	0.64	0.69
K Neighbours Classifier	0.65	0.61

6. Conclusion

- Random under sampling was not helpful to my models. But I employed 3 models and compared results. Random forest and KNN works better than logistic regression.

- Since this data set is unbalanced, I employed random under sampling method. Sampling method improved AUC scores but did not improve weighted average F1-scores.

- I have not included time and day. The next step will be using accident time and day in the machine learning process.