Capstone Project: Car Accident Severity Prediction

1. Introduction / Business Problem

Road safety is a serious public concern, thus accident analysis and prediction are significant areas of research. Through data analytics, we are able to extract meaningful information and provide insights on key factors that lead to road accidents.

Most road accident data collected by government agencies are public available. This project focuses on data obtained from Seattle City to predict the severity of an accident based on time, traffic conditions and other environmental factors. Target audience would be people in charge of developing traffic rules and accident prevention policies. This analysis would also provide practical information to the general public.

2. Data

2.1 Data Source

The dataset contains collision data collected by the Seattle Department of Transportation from 2004 to 2020. It contains attributes such as location, date, time, weather, light condition, road condition, type of collision, severity of collision, to name a few. The goal is to provide a measure of the severity of accidents at different points in time and space, thus the severity of collision was used as the label to train the model.

2.2 Data Preparation

2.2.1 Missing Values

The dataset has missing data that needed to be dealt with. The junction type, weather, road condition, and light condition columns have observations classified as 'Unknown'. The missing values are practically unknown values, so it made sense to replace the values with the same class. Observations for address type and collision type were dropped as the most frequent values for each column were not high enough in count percentage to replace the missing values. All other attributes not relevant to my analysis were dropped.

2.2.2 Data Formatting

Attributes with incorrect data types were converted to the proper format, namely date and time of incident, and further processed and replaced by dummy variables containing the hour and day of incident.

The dataset mostly contains categorical data. Most machine learning algorithms cannot operate on categorical data directly, therefore these were converted to numerical data using one hot encoding technique.

2.2.3 Class Imbalance

Imbalanced classes have a negative impact on the accuracy of machine learning models. This occurs when there is a disproportionate ratio of observations in each class. The dataset contains more observations for accidents that resulted to property damage (severity code #1) than accidents resulting to injury (severity code #2). To fix the imbalance, I decided to under-sample the data by randomly removing observations from the majority class (severity code #1) to match the number of observations for the minority class (severity code #2).

2.3 Feature Selection

After pre-processing the data, 113,954 samples remained with 9 features (which increased to 52 due to one hot encoding). Some attributes were duplicates or redundant in value, while others were not relevant to my analysis. I'm particularly interested in attributes that relate to time and space, thus attributes influenced by human behaviour were removed. The following features were selected - address type, collision type, junction type, weather, road condition, light condition, vehicle count, hour of day, and day of week.

3. Methodology

3.1 Exploratory Data Analysis

The following plots show the relationship between the target variable and the selected features. It can be observed that most accidents occurred on Fridays and midnight, and that most features have noticeable impact on accident severity.

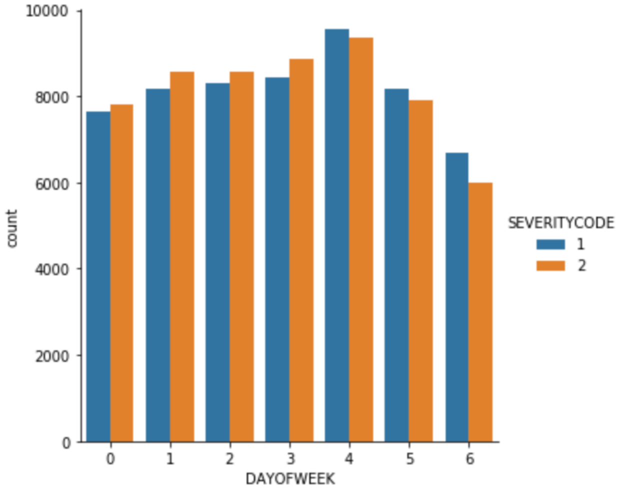


Figure 1. Distribution of severity per day

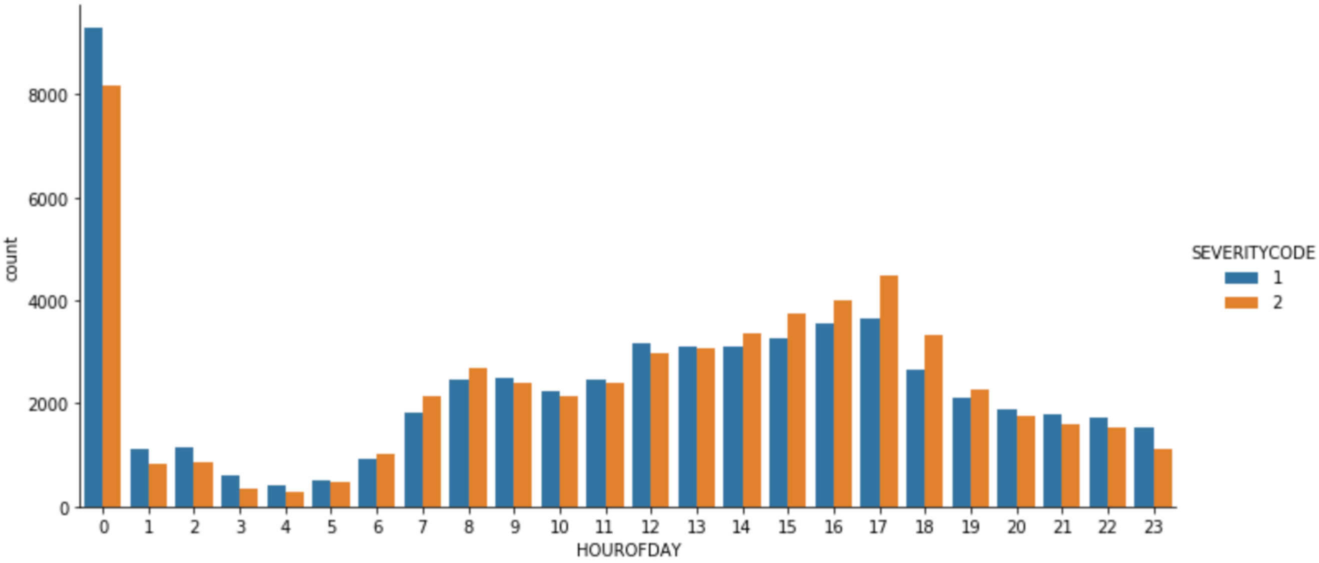


Figure 2. Distribution of severity per hour

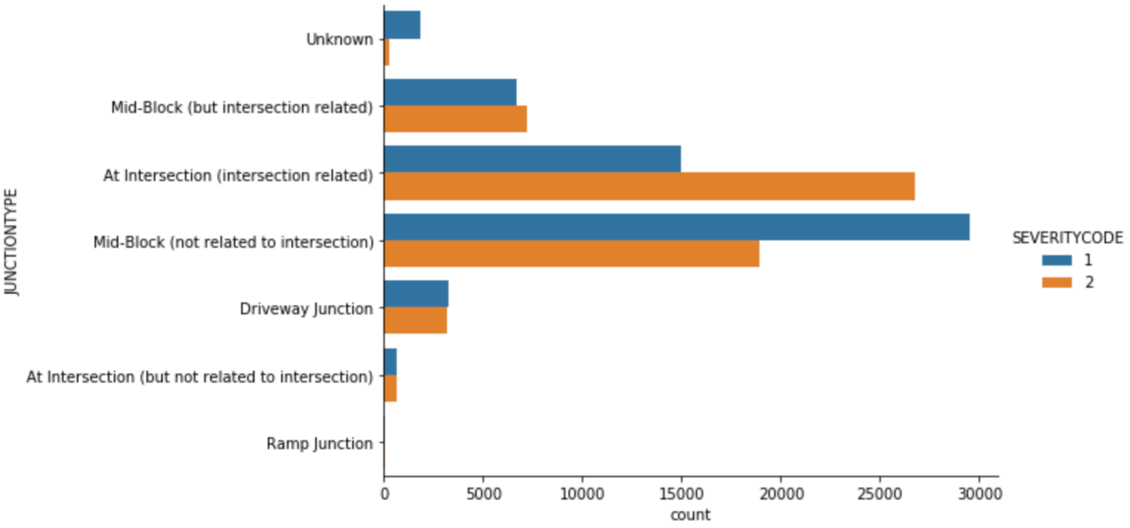
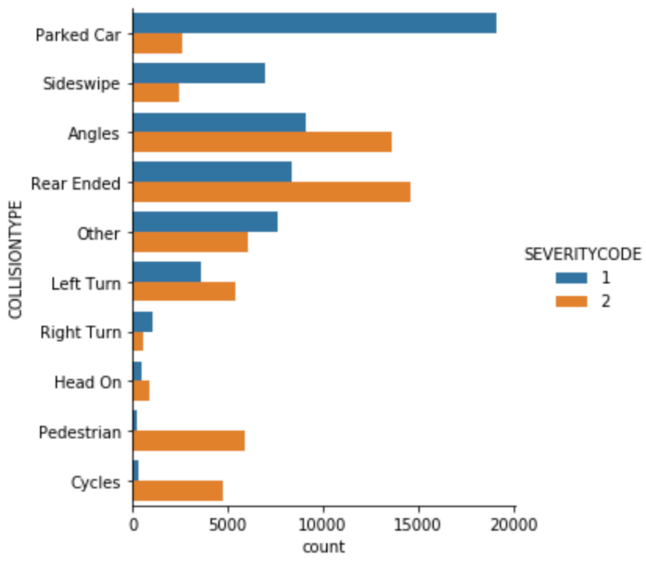


Figure 3. Impact of junction type on severity



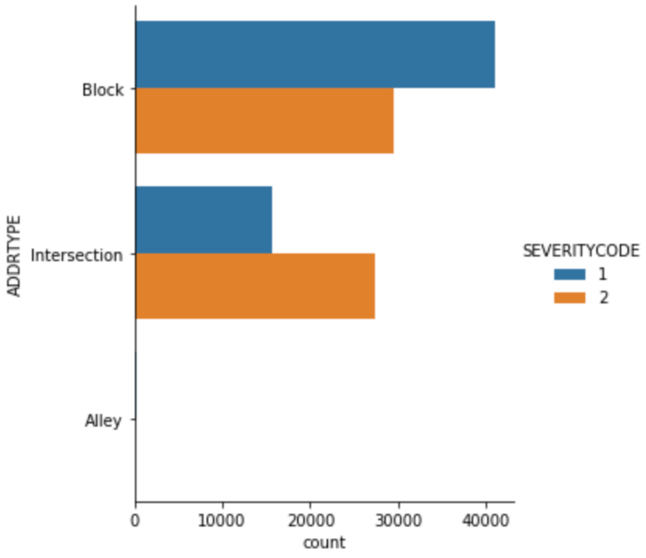
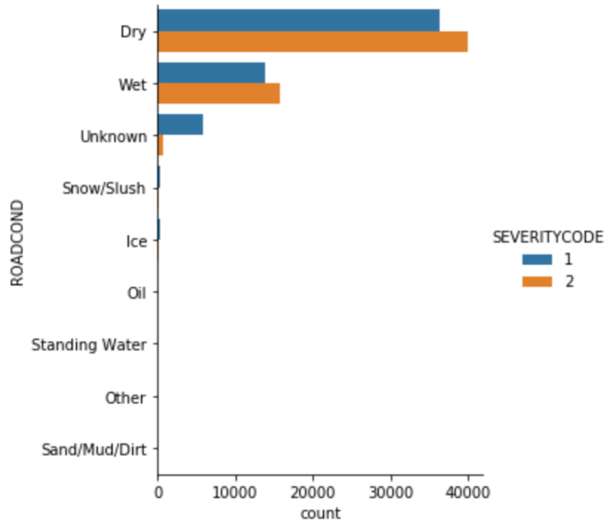


Figure 4. Impact of collision type on severity

Figure 5. Impact of address type on severity



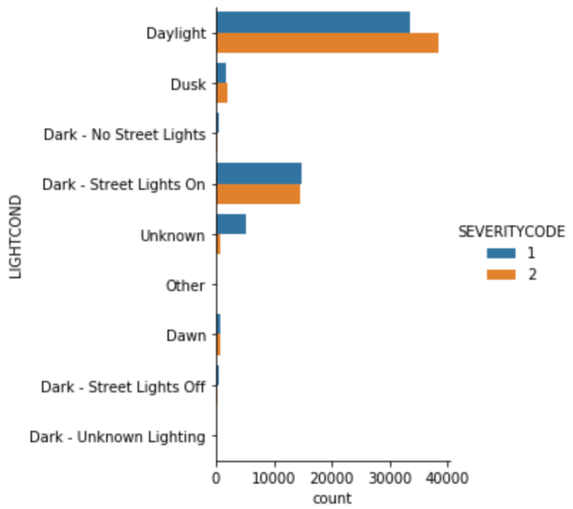


Figure 6. Impact of road condition on severity

Figure 7. Impact of light condition on severity

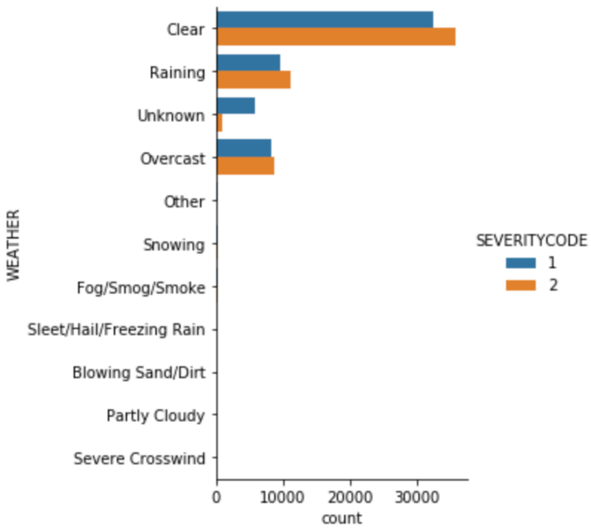


Figure 8. Impact of weather on severity

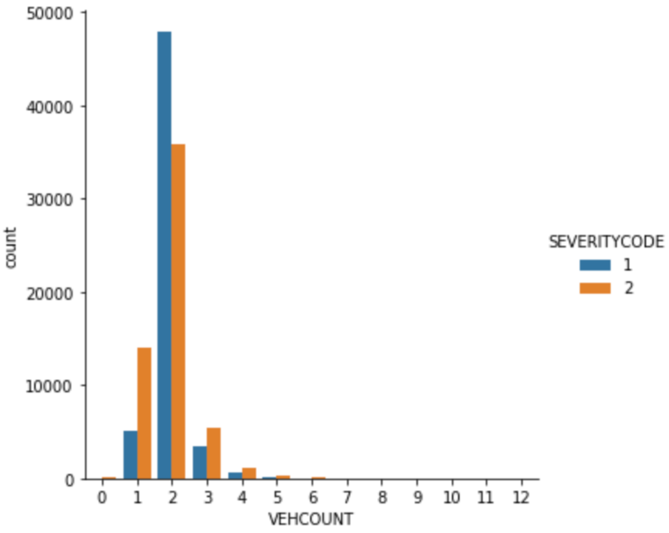


Figure 9. Impact of vehicle count on severity

3.2 Modeling

Classification models were used since the goal was to predict the class of a categorical label The dataset was split into 80% training set and 20% test set for a more accurate evaluation of out-of-sample accuracy.

3.2.1 Random Forest Classifier

A random forest merges multiple decision trees together to obtain a more accurate prediction. It also allows easy measurement of the importance of each feature on the prediction. I opted for this algorithm as tree ensembles usually outperform single decision trees.

Random Forest usually performs better when less important features are removed. However, this was not the case for the model as removing irrelevant features did not necessarily improve its accuracy. For the parameters, I kept the number of trees to 100 and the maximum depth to 10 as increasing the value decreased the accuracy.

3.2.2 Gradient Boosting Classifier

Gradient boosting classifiers are also popular due to their effectiveness. I used a regular boosting algorithm and tested with different parameter values. The best accuracy is with a learning rate of 0.75 and maximum depth of 4. Any more than that only decreased the accuracy.

4. Results

The models were applied on the test set to predict the class of the label. Both models performed quite decently, each with an accuracy of 70%.

	Accuracy	**Precision**	**Recall**	**F1-score**
Random Forest	0.70	0.71	0.70	0.70
Gradient Boosting	0.70	0.71	0.70	0.70

Confusion Matrix

Random Forest predicted severity 1 with higher accuracy whereas Gradient Boosting predicted severity 2 with higher accuracy. Overall, both models have the same balance between precision and recall, as indicated by the F1 score.

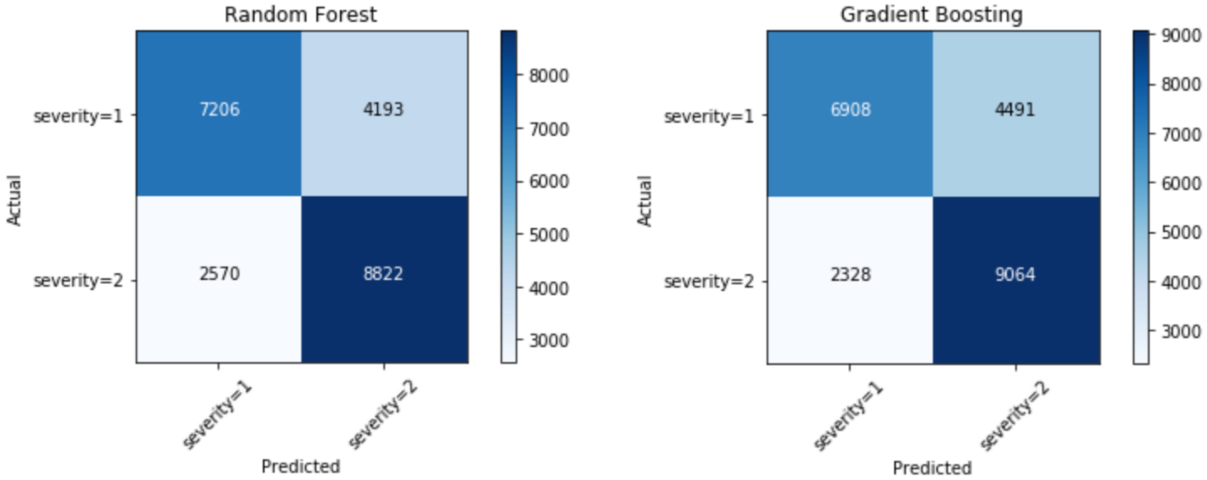


Figure 10. Confusion Matrix - Random Forest vs Gradient Boosting

Feature Importance

The feature importance were ranked differently by each model. However, both models ranked the collision type and vehicle count as most important features.

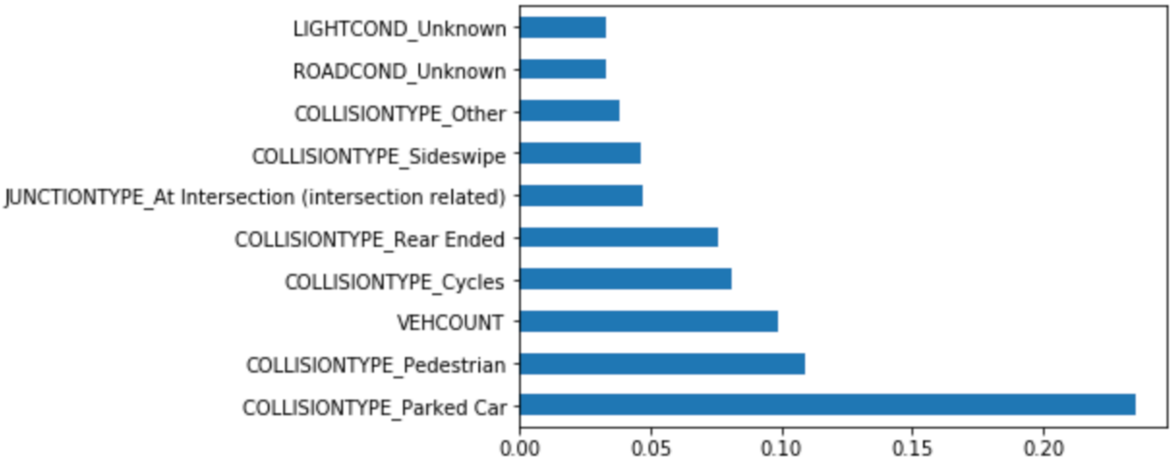


Figure 11. Feature Importance - Random Forest

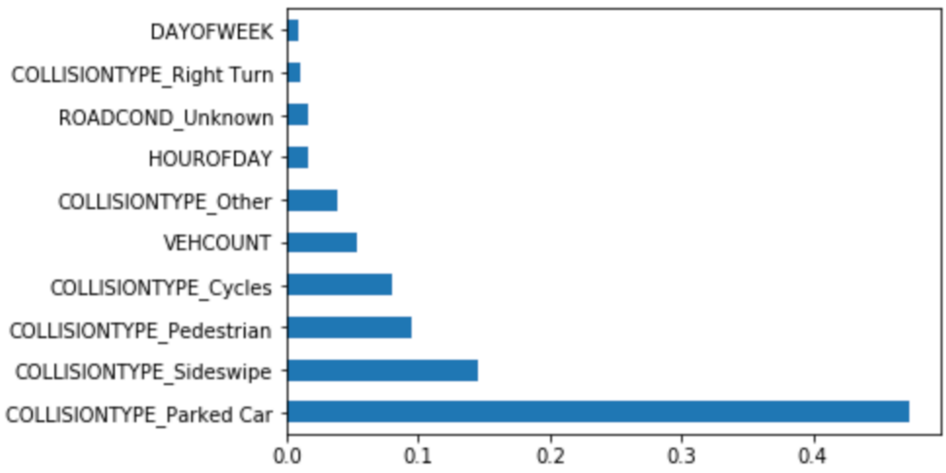


Figure 12. Feature importance - Gradient Boosting

5. Discussion

The built models performed similarly, thus it is not fair to say that one is better than the other. The only difference is that the Gradient Boosting classifier has a slightly longer runtime than Random Forest. The performance could also be better. To obtain better results, perhaps it is worth using a more advanced machine learning algorithm.

6. Conclusion

In this project, I analyzed the impact of time, traffic conditions and environmental factors on accident severity. I used Random Forest and Gradient Boosting classifiers to predict the class of the label. Vehicle count and collision type were identified as the most important features in predicting accident severity. While both models performed decently, accuracy can still be improved.