Traffic Accident Capstone Report

By: Joshua Clark

Introduction

Background

Traffic accidents, also referred to as traffic collisions, occur daily. An accident can occur when a car collides with: another car, pedestrians, stationary/fixed objects, or animals. Numerous factors contribute to the chances of a traffic accident, some of which include a driver's level of distraction or tiredness, the conditions of the road, the time of day, or even the lighting conditions at the time of the accident. Mitigating factors contributing to and severities of traffic accidents minimizes damage to property and personnel. In the most extreme cases, it prevents fatalities.

Problem Statement

Identify leading attributes that contribute to traffic accidents and how these attributes impact the severity factor of an accident. Ultimately, predict the likelihood and severity of encountering an accident based on live, real-world inputs.

Data

Data Source and Overview

The dataset was obtained from the Coursera Capstone. The data provides information pertaining to traffic accidents derived from the Seattle Department of Transportation (SDOT) and Seattle Police Department (SPD). It is arranged in 194,673 rows and 38 columns.

A complete summary of the data can be found here.

Data Cleaning

The INCDTTM feature provides the Date and Time of the accident. This provides useful insight into times, days of week, and months that accidents are most likely to occur. The column was converted into a datetime, and pertinent information was extracted into separate feature columns.
Categorical columns such as Weather, Road Condition, and Lighting Condition has missing values filled based upon the mode of the column.

Next, comparing the value_counts of the LOCATION feature revealed that accidents frequently occurred at repeat locations. This suggested that there were locations that may be more prone to accidents! An ACCIDENT_FREQ feature was created to reflect the periodicity of accidents at a given location.

Finally, several of the features required conversion from categorical to numerical for the purposes of model training and evaluation. Due to the low cardinality of each column, and in an effort to minimize dimensionality complications created by One Hot Encoding, a Label Encoding approach to data transformation was selected. Each categorical column was mapped to an integer value corresponding to the unique column counts.

Feature Decisions

The final set of features included: Day, Time, Month, Accident Frequency (by Location), Lighting Conditions, Road Conditions, and Weather.

Exploratory Data Analysis

Target Variable

Interestingly, the target variable of Accident Severity was described in the Metadata as containing values of [0, 1, 2, 2b, 3] which described no injury through to fatality. However, the actual dataset contained on values [1, 2]. This was interpreted to mean 1: Minor Injury or No Fatality and 2: Major Injury or Fatality. The target variable column was reformatted to a category dtype containing values: [‘No Accident’, ‘Accident’].

Dependent and Independent Relationships of Interest

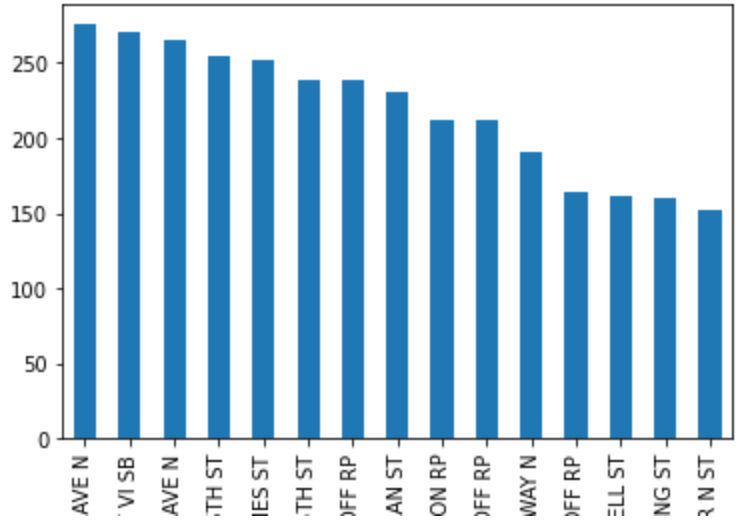
Preliminary data exploration revealed a connection between Friday and Traffic Accidents as well as October and Accidents. Further, this appears to be largely driven by the increase in drivers during these features based on an overall increase in the “No Accident” graph as well.

A re-grouping by Weather and Severity Code provided insight into some key Weather Conditions leading to accidents. Based upon most driving occurring during clear days, there are far more accidents on clear days than any other Weather Condition day. However, the percentage of clear accidents to clear non-accidents is not the largest percentage contributor. Rather, Rain is the largest statistical contributor to accidents than any other Weather feature.

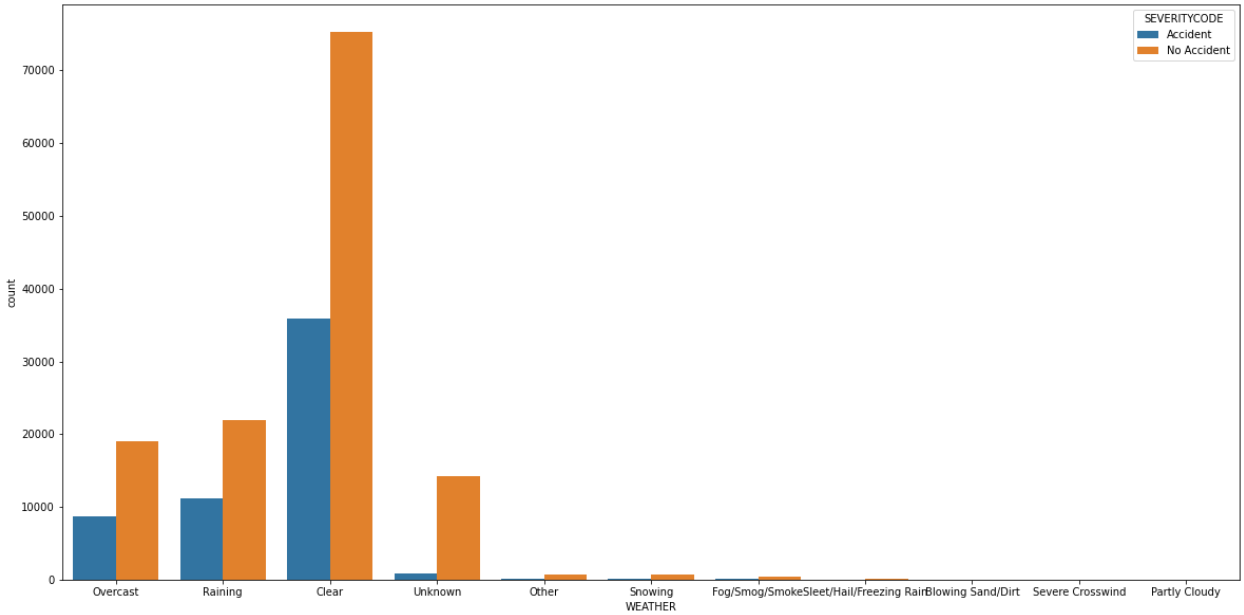
Similarly, a Wet Road Condition is the largest percentage contributor to accidents than any other Road Condition.

Finally, the most unsuspected feature value was the largest percentage contributor from Lighting Condition. The original hypothesis of Night Driving, both with and without adequate street lighting, was not the largest percentage contributor to Lighting Condition features and accidents. Rather, Dusk conditions were most significant. Nearly 66% of Dusk Driving resulted in an accident as compared to 30% of Night Driving.

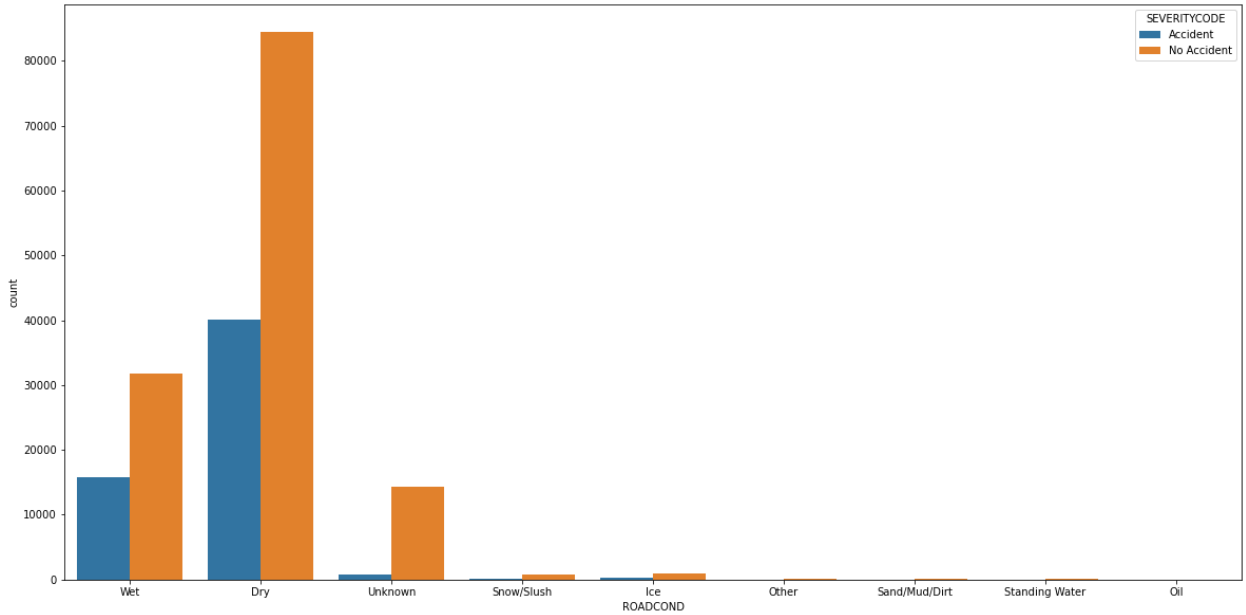
Frequency of Accident by Location (threshold > 150)



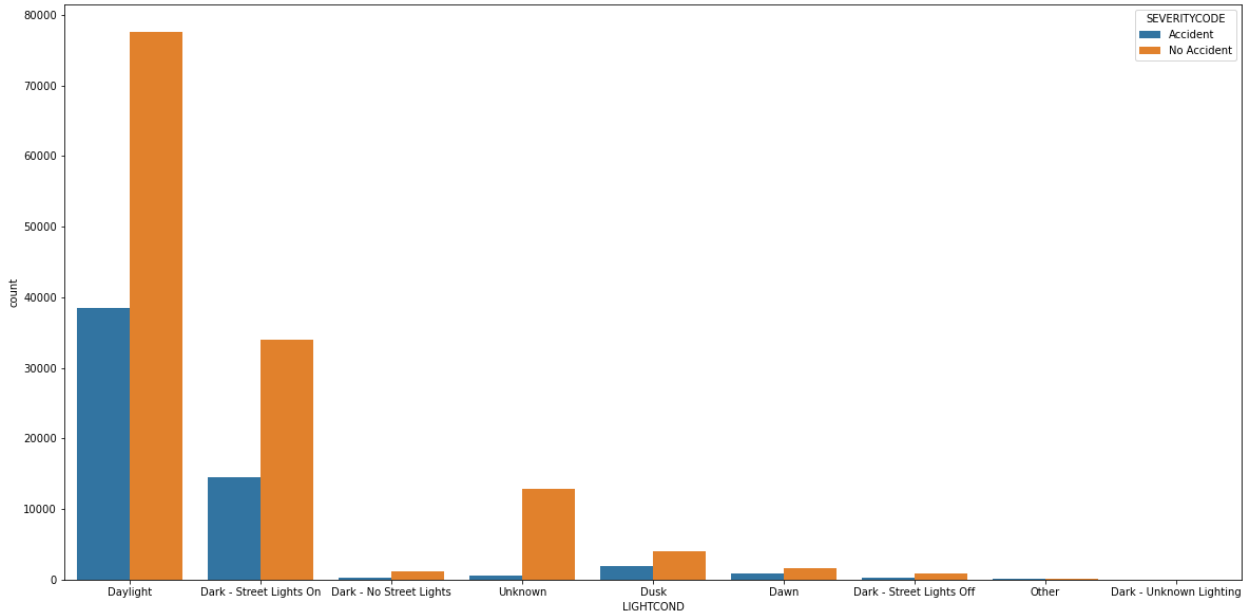
Accident/No Accident Frequency by Weather Conditions



Accident/No Accident Frequency by Road Conditions



Accident/No Accident Frequency by Lighting Conditions



Modeling

Overview

A K-Nearest Neighbor, Decision Tree, Support Vector Machine, Logistic Regression, and XGBoost Model were trained and evaluated based upon standard metrics within the sklearn library. Of note, the SVM Model was abandoned due to repeated computational complexity and failure to converge. This is hypothesized to be due to the large dataset.

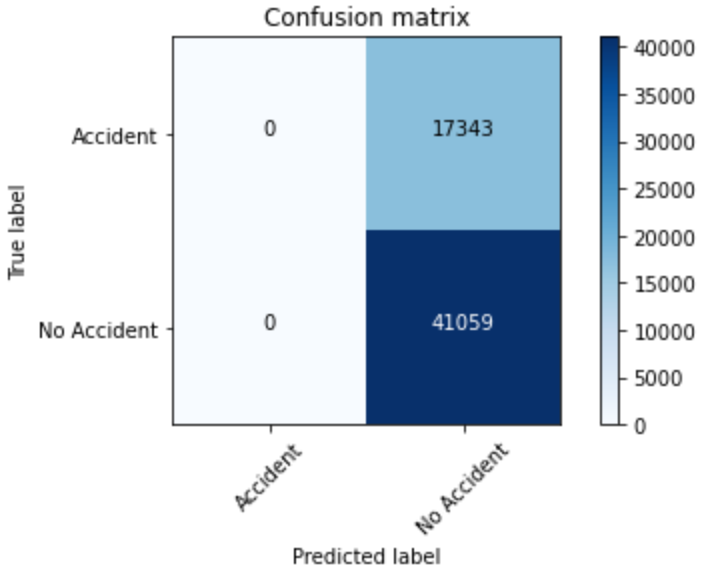
Classification Models

Describe some of the models to be used and any issues encountered with modeling.

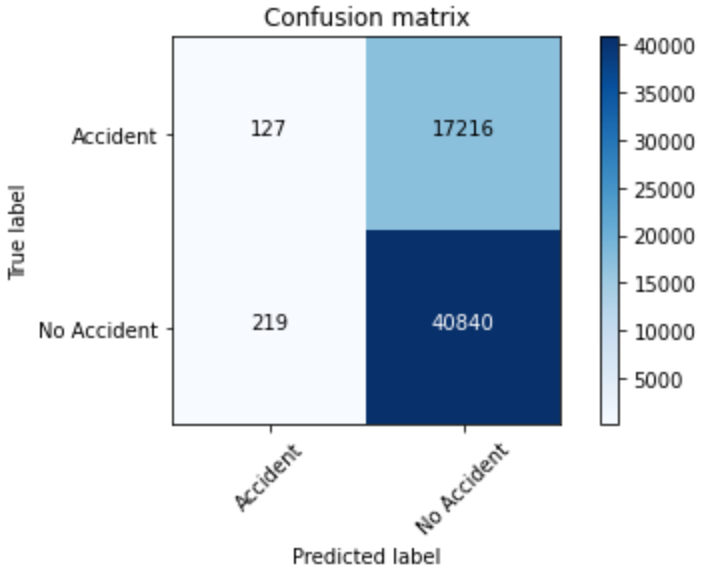
Modeling Summary

Below are the Confusion Matrix Plots of each of the trained and evaluated models. Overall summary statistics for each model is provided below. Based upon a holistic ability to predict success in actual accident scenarios, the KNN Model is assessed to have performed best. Though there was an increase in False Positive outcomes, this degrade came with the tradeoff of a decrease in False Negatives. The KNN had the best performance with regard to True Positive Categorization as shown below.

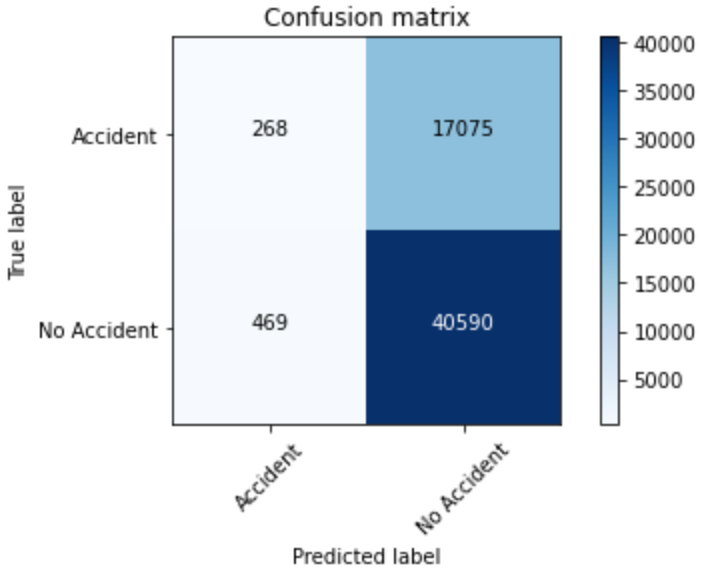
Decision Tree Confusion Matrix:



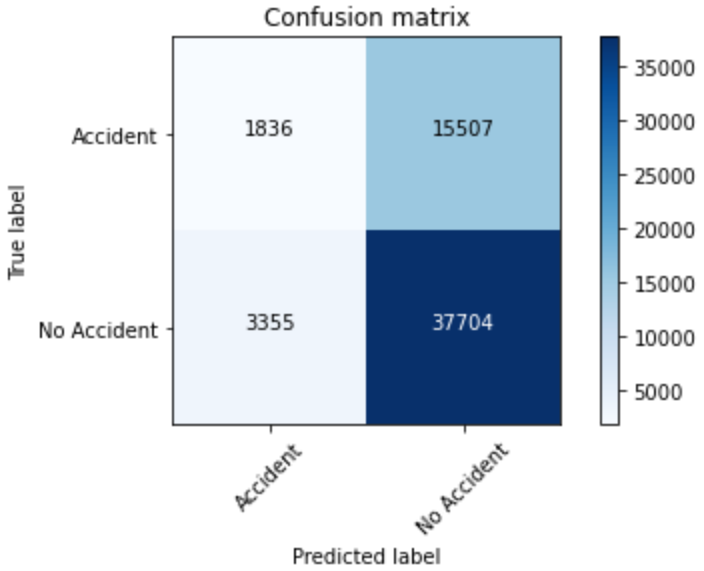
Logistic Regression Confusion Matrix:



XGBoost Confusion Matrix:



KNN Confusion Matrix:



Summary Table:

	Algorithm	Accuracy	F1-Score	LogLoss
0	KNN	0.68	0.61	NA
1	Decision Tree	0.70	0.60	NA
2	XGBoost	0.70	0.59	NA
3	Log Regression	0.70	0.58	0.6

Conclusion

In conclusion, a series of models was developed to predict an accident, or lack thereof, based upon Weather, Lighting Conditions, Road Conditions, Location, Day of the Week, and Month. The models all achieved approximately 70% accuracy. The largest drawback of the models was the high False Negative rate. This can largely be attributed to the unbalanced classes, and is addressed under the next section of Future Contributions.

Future Contributions

The large amount of False Negatives is primarily driven by the Unbalanced Classes. Future workarounds would include distributing the Test and Training sets to include a larger percentage of No Accident data. Increasing overall test_size to 50% maintains approximately 70% Accuracy while reducing False Negative outcome. However, because the train_test_split was equally and randomly split at 50%, it still favored No Accident outcomes. Future model development would split the data into Accident and Non Accident subsets, perform any standardization, and concatenate a percentage of each dataset after standardization.