Introduction

For this Coursera capstone project, we will be trying to predict the severity of an accident using the dataset provided by Coursera. There are several potential business applications for a prediction model of this nature. By using certain attributes, it can help insurance companies better determine accident severity and provide competitive pricing. The dataset also has geometric and positional information so this can be used by city planners to determine if a certain type of intersection or speed threshold leads to an increased severity of accidents.

The scope of this project is to determine if accident severity can be predicted based on the data set provided by Coursera and the Seattle Police Department. We will be taking the role of insurance company trying to determine the possibility of an accident severity and we will also be assuming that certain probabilities are known to us (explained in more detail in the Data Understanding section report).

Data Understanding

This dataset is provided by the Seattle Police Department (SPD) and contains information about collision data that was recorded by the SPD. The following table shows all the features provided by the dataset and the ones used as well as to the reasoning if it was discarded:

Column Name	Non Null Elements	Description	Used
SEVERITYCODE	194673	Target	Y
X	189339	Not used as it is not defined	N
Y	189339	Not used as it is not defined	N
OBJECTID	194673	SPD identifier key	N
INCKEY	194673	Incident key. Not giving much value	N
COLDETKEY	194673	Secondary key. Not much value	N
REPORTNO	194673	Not used as it is not defined	N
STATUS	194673	Not used as it is not defined	N
ADDRTYPE	192747	Collision type, like alley, block, intersection	Y
INTKEY	65070	Not used as it is too specific	N
LOCATION	191996	Not used as it is too general and not expandable	N
EXCEPTRSNCODE	84811	Not defined	N
EXCEPTRSNDESC	5638	Not defined	N
SEVERITYCODE.1	194673	Not defined	N

SEVERITYDESC	194673	Not used too specific	N
COLLISIONTYPE	189769	Type of the collision	Y
PERSONCOUNT	194673	Number of people	Y
PEDCOUNT	194673	Number of pedestrians	Y
PEDCYLCOUNT	194673	Number of cyclists	Y
VEHCOUNT	194673	Number of vehicles	Y
INCDATE	194673	Date of the incident. Not used as feature	N
INCDTTM	194673	Date and time of the incident	N
JUNCTIONTYPE	188344	Not used as ti is not well defined	N
SDOT_COLCODE	194673	SOPD code not used	N
SDOT_COLDESC	194673	SOPD code not used	N
INATTENTIONIND	29805	Was the driver not paying attention	Y
UNDERINFL	189789	Weather the driver was under influence	Y
WEATHER	189592	Weather condition	Y
ROADCOND	189661	Road condition	Y
LIGHTCOND	189503	Lighting condition	Y
PEDROWNOTGRNT	4667	Did pedestrian have right of way	Y
SDOTCOLNUM	114936	SDOT id, not used	N
SPEEDING	9333	Was driver speeding	Y
ST_COLCODE	194655	SDOT code, not used	N
ST_COLDESC	189769	SDOT code, not used	N
SEGLANEKEY	194673	SDOT code, not used	N
CROSSWALKKEY	194673	Key for cross walk, not used as too specific	N
HITPARKEDCAR	194673	Hit a parked car, not used	N

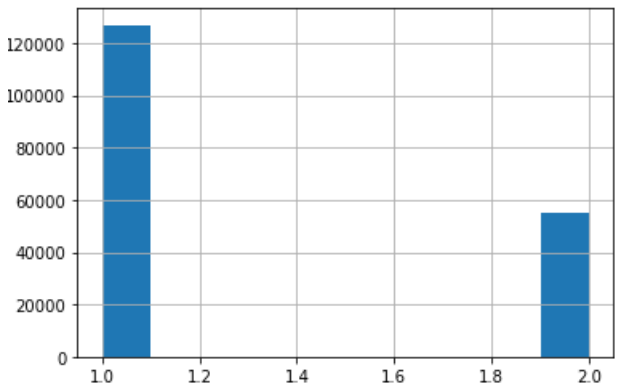
Data Preparation

With the features selected, we had to deal with a quite a few missing values. The below table summarizes how these are dealt with:

Column Name	Non Null Objects	How NA values are dealt with
SEVERITYCODE	194673	No NA values
ADDRTYPE	192747	Dropped
COLLISIONTYPE	189769	Dropped
PERSONCOUNT	194673	No NA values
PEDCOUNT	194673	No NA values
PEDCYLCOUNT	194673	No NA values
VEHCOUNT	194673	No NA values
INATTENTIONIND	29805	NA values assumed to be N as the only unique is Y
UNDERINFL	189789	NA values assumed to be N
WEATHER	189592	NA values defaulted to be clear
ROADCOND	189661	NA mapped to unknown road
LIGHTCOND	189503	NA values mapped to Daylight, as I will assume that is when most of the driving is doen
PEDROWNOTGRNT	4667	NA mapped to N as the only other is Y
SPEEDING	9333	NA mapped to N, as the only other is Y

For a few of the columns, such as light condition and road condition, I have made the default assumptions on these to be clear or given the best daylight. These are reasonable as it can be assumed that individuals do not want to drive in adverse conditions. Additionally, the insurance company has probabilistic models of what each weather condition is.

The balance of our target variable is also viewed:



As we can see, most of the target variable are 1 (property damage), with a limited set of injury. This is dealt with by lowering the weights of the training set on the property damage class when we do our modelling.

For modelling purposes, 75% of the data will be used for training while 25% is used for testing.

Modelling

The following models are used to try and predict the severity of the accident:

- K Nearest Neighbour
- Decision Tree classifier
- SVM
- Logistic regression

K Nearest Neighbours

In this model, the data point is assigned a class based on the k nearest neighbour from the training data. To train this model, a brute force approach is used to determine the ideal value of k in the training set.

Decision Tree Classifier

This model tries to separate the classes based on the features that provide the best separation first and work its way down. At each end point (called a leaf) the purity of the remaining classes are measured to determine the importance of the feature.

SVM

This model aims to construct a hyperplane between the classes based on the feature set provided.

Logistic Regression

This model uses the sigmoid function to determine a probability and differentiate the classes.

Model Evaluations

The models were trained on the 75% of the original data provided. Each of the models are trained using the sklearn libraries and the accuracies are determined on the test set to evaluate the models. Since the class set is unbalanced it will be judged based on the F1 score. The below table summarizes the models results on the test set:

Model	Jaccard Index	F1 Score	Log Loss
KNN	0.67	0.68	NA
Decision Tree	0.73	0.69	NA
SVM	0.74	0.70	NA
Logistic Regression	0.73	0.70	0.55

Based on the above table, it can be seen that the SVM and Logistic regression work the best.

Conclusion

Based on the results, we have shown that the accident severity can be predicted based on certain attributes that have been described in the Data Understanding section. It is important to note that there is an important assumption made:

- Probabilities of certain features are known prior and can be inputted into the pricing model.
    - These values will become important once we try to expand this model into different geographic areas.

Since the dataset is unbalanced, the models were judged based on the F1 Score. From the table in the Model Evaluation section, it is shown that support vector machine and logistic regression provide the best results. It is recommended to use a logistic regression model as it has a faster run time when training.