Introduction

Road accidents lead to a significant number of human fatalities, infrastructure damage and economic losses worldwide. Providing road safety for both pedestrians and drivers is very important to city planning departments and governments. With this problem in mind, this project aims to provide insights into factors that affect severity of car accidents in Seattle. The goal of this project is to develop a model that predicts accident severity based on various external conditions such as current weather, road conditions and visibility can prove to be a valuable tool that provides insights into isolating the factors that are deemed dangerous. This type of model can be used to alert drivers and pedestrians to avoid travelling while these factors are present or take an alternative route for their travel.

Description of the Data

The data used in this analysis is provided by the Seattle police department. It includes 194673 observations of accidents and records the severity of these accidents. The primary variables of interest for this study as the dependant variables are weather conditions(WEATHER), light conditions(LIGHTCOND), road conditions (ROADCOND), whether speeding was a factor (SPEEDING), whether the driver was paying attention(INATTENTIONIND) and whether the driver was under the influence (UNDERINFL). All of these variables of interest are in categorical form.

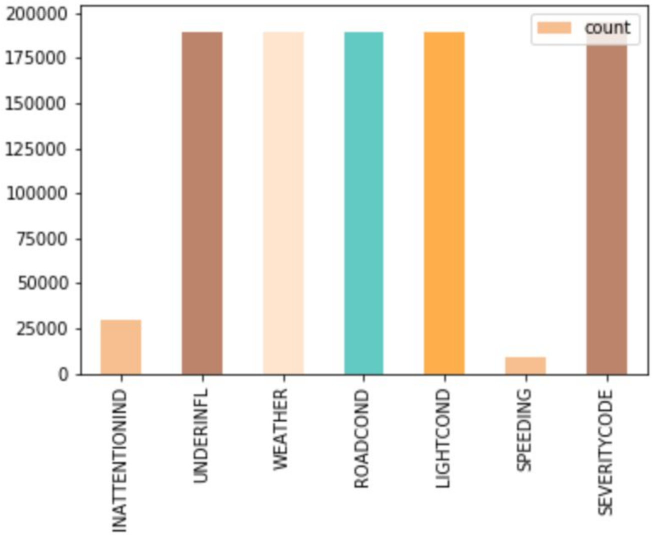
The first step in the analysis of the data was encoding some of the dependent variables in the study. The variable calculating severity of the accidents is recorded as 1 for property damage only and 2 for injury. This is encoded and we use 0 for property damage only and 1 for injury. Dummy variables are provided for whether there was speeding or whether the driver was under the influence. In this analysis, we assign numerical values for the categorical labels of weather condition, light condition and road conditions. Road conditions were encoded in such a way that dry was assigned 0, mushy as 1 and wet as 2. Similarly, weather conditions were assigned 0 if clear, 1 if overcast, 2 if windy and 3 if rain or snow. Lastly, values for light conditions were assigned 0 if bright, 1 if medium and 2 if dark.

There was a lot of data which had values unknown or other.To tackle the problem of missing data, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had ‘Other’ and ‘Unknown’ in them.

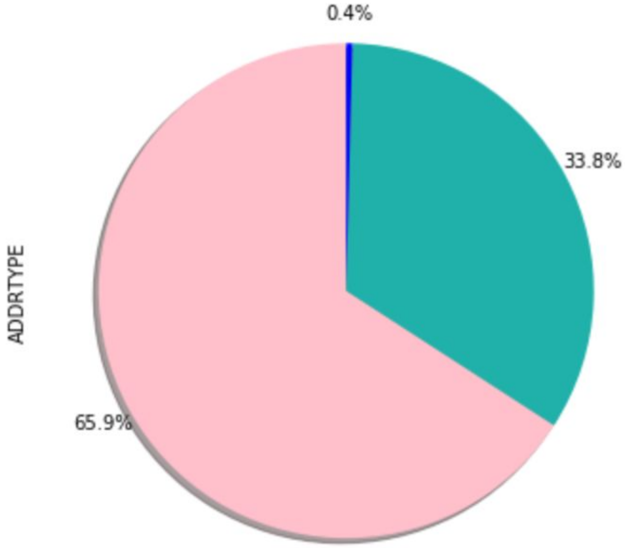
Methodology

Exploratory Analysis

A quick analysis of our dependent variable (severity code) shows that there are significantly more instances of property damage than physical injury. The split is almost in the ratio of 2:1. Due to this unbalanced nature of the dataset, SMOTE was used from the imblearn library in order to balance the target variable in equal proportions in order to have an unbiased classification model.



The graph above is supposed to depict all the nonzero values within each independent variable of the model and can be seen as the frequency of adverse conditions under which accidents took place. The factor which had the most number of accidents under adverse conditions was adverse weather conditions while adverse lighting conditions had the second most number of accidents caused by it. The factors which contributed the least to an instance of an accident are over-speeding and the driver being under the influence.



A quick analysis of the addresses where the most accidents occur, show that almost all of the accidents occur at a block or an intersection. This can be a valuable insight for the city planning office to implement extra security measures in those area such as additional traffic lights, speed bumps or any signs.

Results

Logistic Regression

The model used for analysis of this dataset is Logistic Regression. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The post SMOTE data was used to predict and fit the Logistic Regression Classifier.

Classification Report

	Precision	***Recall***	***F1 score***
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.53	0.54	0.53

Weighted Avg	0.61	0.59	0.60
Log Loss	0.68		

Decision Tree

The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘6’. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

Classification Report

	Precision	***Recall***	***F1 score***
0	0.64	0.72	0.68
1	0.44	0.41	0.38
Accuracy	0.58		
Macro Avg	0.54	0.53	0.53
Weighted Avg	0.56	0.58	0.56
Log Loss	0.68		

Discussion

Algorithm Type	Average f1 score	Property Damage vs Injury	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic	0.60	0	0.72	0.67

Regression				
		1	0.35	0.41

F1 Score:

f1 score is the measure of how accurate a model is. Higher values reflect a higher precision and accuracy. The F1 score calculated in the results is the average of the individual f1-scores of the two elements of the target variable i.e. Property Damage and Injury. The f1 scores are comparable for both of the models used in the analysis and slightly higher for logistic regression.

Precision:

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive. The highest precision for Property Damage is for Logistic Regression, whereas for Injury it is the Decision Tree. The Precision is calculated individually above in order to understand how accurate the model is at predicting Property Damage and Injury individually. For the Decision Tree the precision of 0 is 0.64 and for 1 it is 0.44 which is fairly good. As for the Logistic Regression model, for 0 it is at 0.72 and for 1 it is 0.35

Recall

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative. The recall for both Property Damage and Injury is 0.72 and 0.34 . As for the Logistic Regression, the recall for Property Damage is 0.67 and for Injury it is 0.41. The recall for Property Damage and Injury is the most balanced in terms of being good for both the outputs of the target variable.

Conclusion

When comparing the two models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the three models individually as a whole and how well they perform Car Accident Severity – Seattle, Washington 14 for each output of the target variable. When comparing these scores, we can see that the f1-score is higher for Logistic Regression. Looking at the two models, we can see that the Decision Tree has a more balanced precision for 0 and 1. Whereas, the Logistic Regression is more balanced when it comes to recall of 0 and 1. Furthermore, the average f1-score of the two models are very close but for the Logistic Regression it is higher by 0.04. It can be concluded that both the models can be used side by side for the best performance.