CAR ACCIDENT SEVERITY CAPSTONE PROJECT:

1. INTRODUCTION

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. Car accidents are one of the unexpected and unwanted events that can happen on the roads and drivers can be impacted in different ways as a result of them. Some are fatal, some cause injury while some result in property damage. These crashes post burden on government authorities as well in terms costs associated. According to the data from WSDOT there had been 10,315 total crashes in 2019.

1.1 Business Problem

The goal of this project is to predict severity of an accident and allow the Seattle government to possibly prevent/reduce car accidents that depend on factors like weather and road condition(wet/dry), traffic situation, light condition etc. About 100,000 crashes happen in Seattle every year since 2010 as per WSDOT ten-year summary report(WSDOT 10 Year Summary Report).

1.2 Audience/Stakeholders

The model developed can be used to warn drivers, health authorities, government and police about the possibility of a car accident and its severity so that everyone can be more careful and/or change travel plans (if able to) in such critical situations.

The solution/model developed via this project can be utilized by the end user(car drivers),local government, car insurance companies and hospitals who are considered the target audience for this business problem to use the model and make required decisions to help prevent accidents in the city.

2. DATA

2.1 Data Understanding

The data used in this project is about “accident severity” and was provided by SPD (Seattle Police Department) and recorded by Traffic Records department from 2004 to present. This includes all types of collisions. This problem is seen as supervised machine learning as it has labeled data to train and validate the model.

There are 194,673 observations in the data set. In total, there are 37 attributes(columns) and 1 dependent variable (labelled data) which is “SEVERITYCODE” in the data and 194,673 rows. “SEVERITYCODE” is the code that corresponds to the severity of the collision (as indicated in meta data):

- 1—Accidents resulting in property damage
- 0—Accidents resulting in injuries

2.2 Data Cleaning and Processing

There are numerical and categorical types of data which are needed to be converted into numerical data for machine learning algorithms. Also, some attributes have missing data therefore the data requires some **preprocessing and preparation** to construct the final dataset to be fed into the ML model. Some variables have “unknown/other” values which were treated as missing data.

2.3 Feature Set Selection

Not all attributes are useful/relevant for my model, so I have decided to **select some features** such as SEVERITYCODE (target-variable), SEVERITYDESC, INATTENTIONIND, UNDERINFL, WEATHER, SPEEDINGROADCOND, LIGHTCOND as effects of these variables are considered significant. It was also noticed that data set is unbalanced as distribution of target variable is in favor of property damage so some data balancing will be required. There were only instances of 2 severity types (property damage and injury collisions) in the data set.

Features	Description
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.

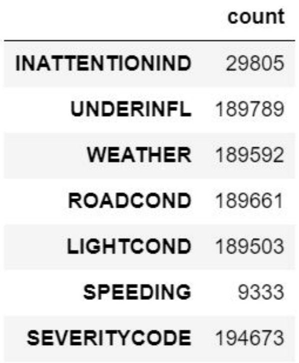
Different machine learning models were used to solve the problem and evaluate the model.

3 METHODOLOGY

3.1 Data Collection

The data was downloaded from the repository and loaded into the Pandas data frame. From the complete dataset, we will choose only the relevant variables which might have an impact in the model training.6 features were selected along with the SEVERITYCODE which was considered the target variable.

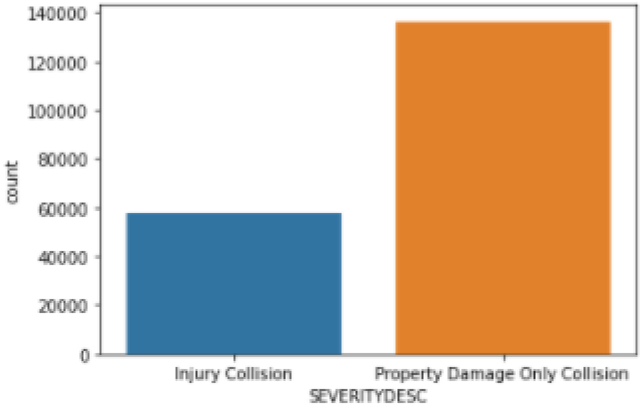
Here is summary of counts for selected variables that have more impact on accident severity.



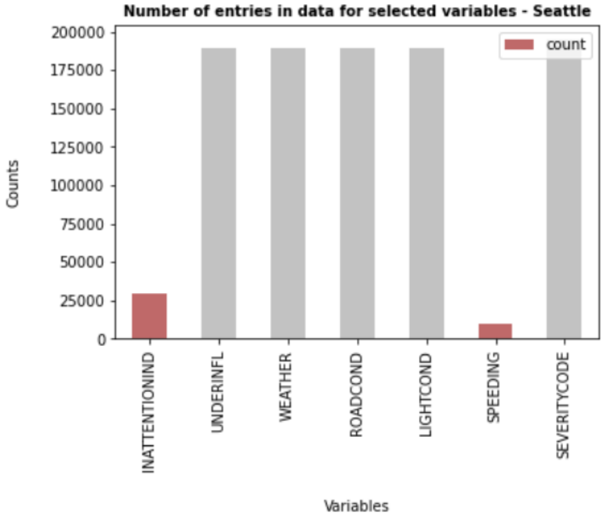
	count
INATTENTIONIND	29805
UNDERINFL	189789
WEATHER	189592
ROADCOND	189661
LIGHTCOND	189503
SPEEDING	9333
SEVERITYCODE	194673

3.2 Exploratory Analysis

It was observed that there are more accidents resulting in property damage than leading to injury based on the data set as seen below. The distribution ratio of this unbalanced data set is almost 2:1. SMOTE package was used to balance the data in equal proportion before feeding into the machine learning models for training.

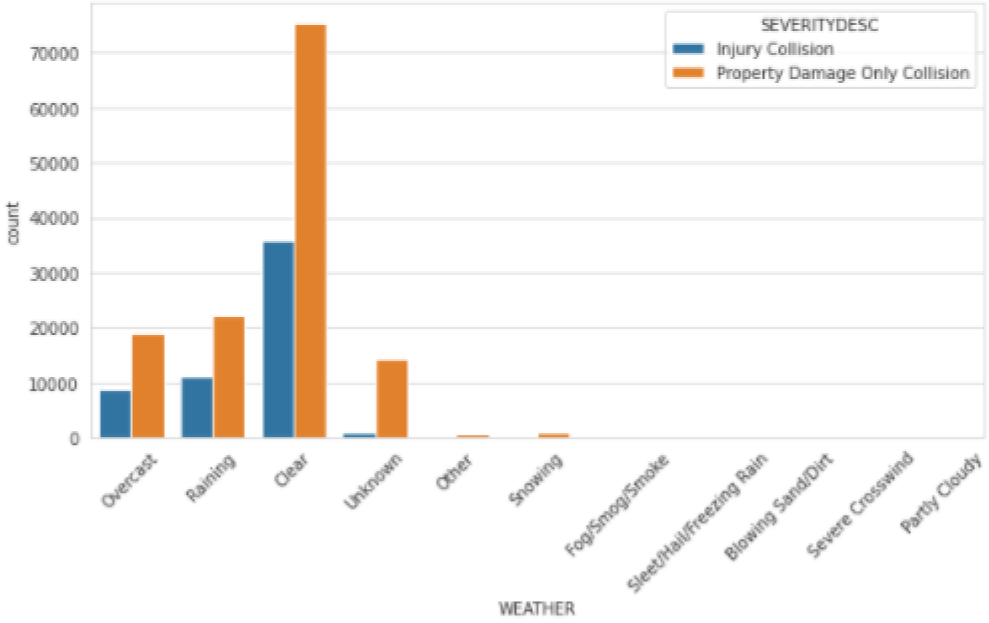


The effects of selected variables like weather, light condition, road condition, whether if under influence or not correspond to higher number of accidents hence these variables were selected. Graphical representation of these can be seen below:



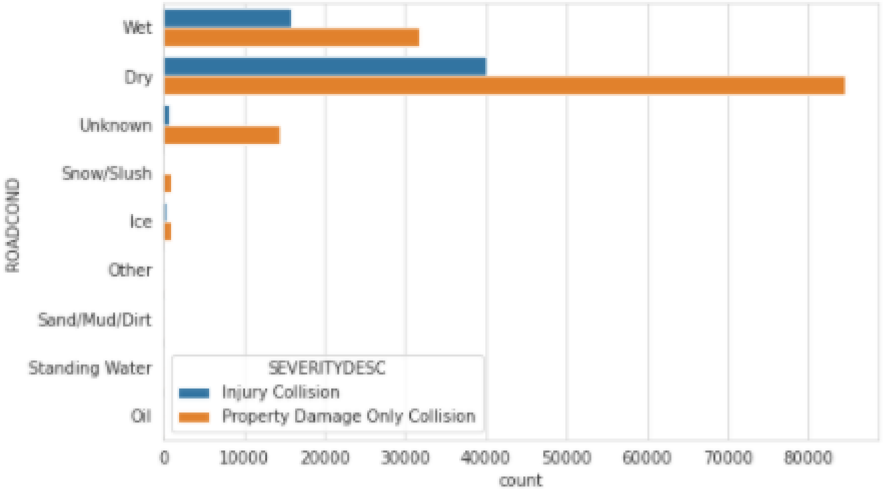
3.2.1 WEATHER

“WEATHER” has significant impact on accident severity. As seen in the figure below, we can see that most of weather conditions lead to higher number of property damage collisions as compared to injury incidents.



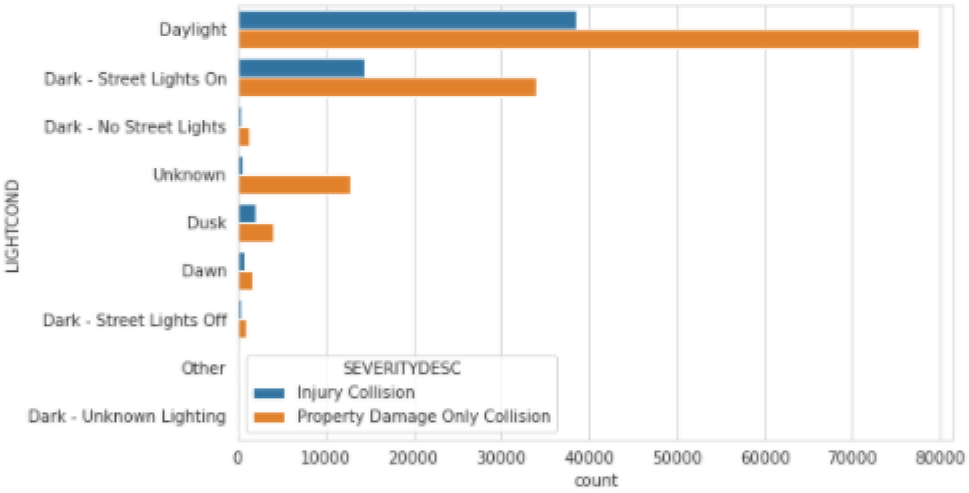
3.2.2 ROADCOND

“ROADCOND” attribute impact can be seen visually as in the below figure. It can be inferred that wet and dry road conditions have significant impact on the number of collisions.



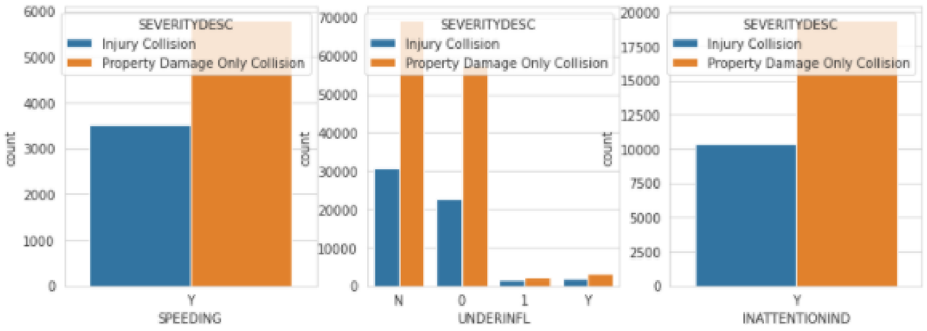
3.2.3 LIGHTCOND

“LIGHTCOND” also has inference on the severity of the accident. Below is the pictorial representation of the same. It is interesting to note that many accidents happen in day light and streetlights on also although majority are property damage type. “Other” and “Unknown” lighting types could possibly be the factors influencing the severity type, but we don’t have visibility to their values in this data set.



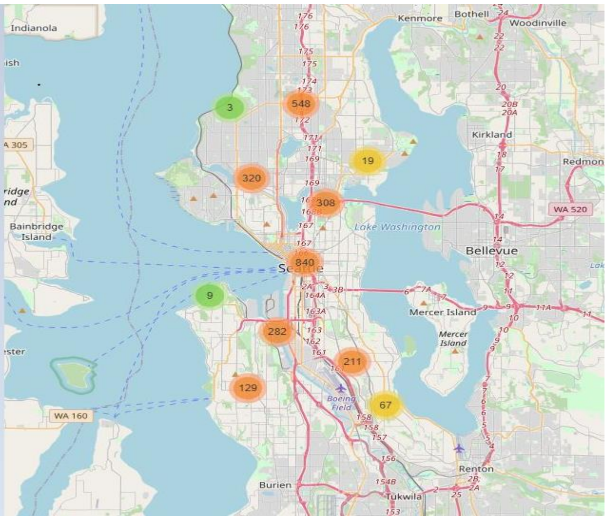
3.2.4 SPEEDING & INATTENTIONID

At the same time, factors like “SPEEDING” and “INATTENTIONID” contributed to low number of incidents as per the data set used. Following figure depicts number of instances related to speeding, influence and attention factors. These attributes are mostly associated with property damage only.



3.2.5 GEOGRAPHICAL ANALYSIS

Majority of injury collisions happen around downtown Seattle and busy intersections near it. High traffic is expected in busy hours. Areas like Pioneer square, First Hill, International district show significant instances for injury collisions. Random selection of records is shown in the map below:



3.3 Machine Learning Model Selection

Decision Tree Classifier, Logistic Regression and k-nearest neighbors' algorithms were used for machine learning model in this project. Decision Trees (DTs) predict the value of a target variable by learning simple decision rules inferred from the data features. t uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).Logistic Regression uses "Sigmoid" as cost function for the classification problems to predict the target variable. K nearest neighbors' algorithm used similarity measure (based on distance) to make prediction. SVM model wasn't used in this analysis as it is generally considered inaccurate for large data sets.

4 RESULTS

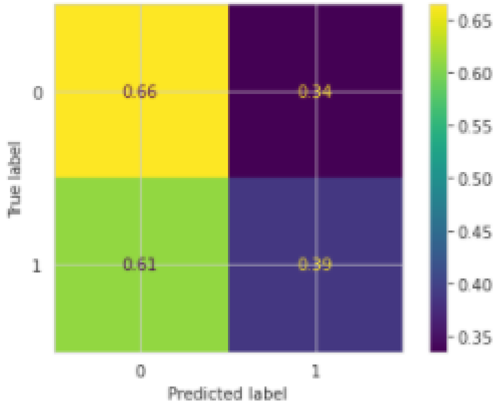
4.1 Decision Tree Classifier

Sci-kit library's "DecisionTreeClassifier" algorithm was used to fit and predict the results. Model was trained on the balanced data. Parameters used were "criterion" and "max_depth". For the purpose of this report "criterion" was chosen as "entropy" while "max_depth" was set to 9.

4.1.1 Classification Report

	Precision	Recall	F1-score
0	0.69	0.66	0.68
1	0.36	0.39	0.38
Accuracy	0.57		
Macro avg	0.53	0.53	0.53
Weighted avg	0.58	0.57	0.58

4.1.2 Confusion Matrix (Normalized)



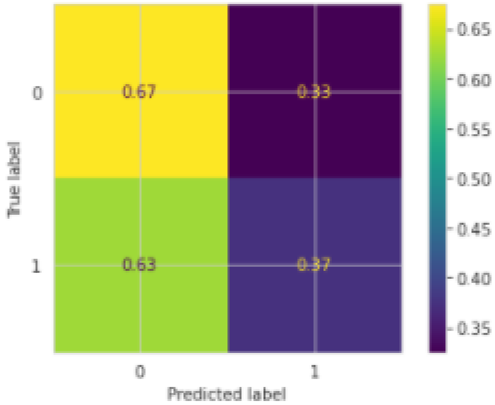
4.2 Logistic Regression

Sci-kit library's "Logistic Regression" algorithm was used to run prediction based on Logistic Regression machine learning model for accident severity. Regularization strength value of 6 and 'saga' solver was used which is considered more efficient for larger data sets. Balanced data was used to fit the model.

4.2.1 Classification Report

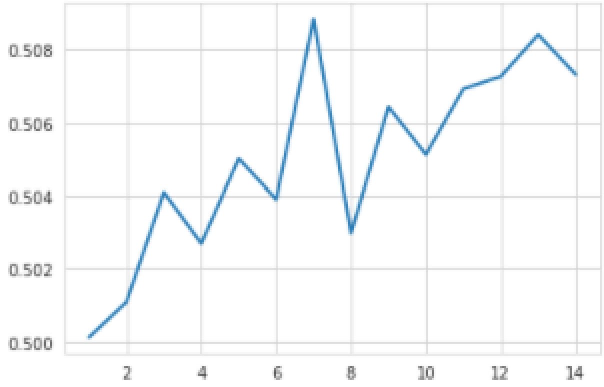
	Precision	Recall	F1-score
0	0.69	0.67	0.68
1	0.36	0.37	0.37
Accuracy	0.58		
Macro avg	0.52	0.52	0.52
Weighted avg	0.58	0.58	0.58
Log loss	0.69		

4.2.2 Confusion Matrix (Normalized)



4.3 KNN

K-nearest neighbor algorithm from sci-kit library was used for classification on accident severity data set. Best value of k was selected to be 7 as determines from the graph (accuracy vs value of k) below:

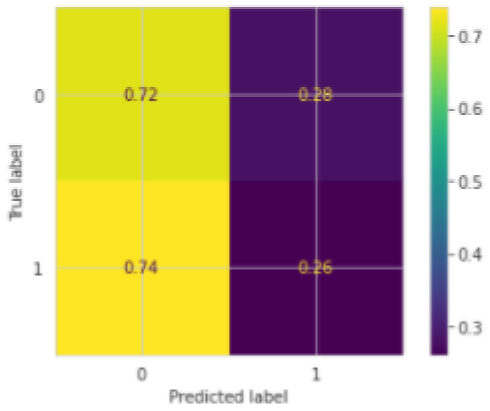


Then the balanced data was used to train the model and predict the severity using value of k as 7.

4.3.1 Classification Report

	Precision	Recall	F1-score
0	0.67	0.72	0.69
1	0.31	0.26	0.28
Accuracy	0.57		
Macro avg	0.49	0.49	0.49
Weighted avg	0.55	0.57	0.56

4.3.2 Confusion Matrix (Normalized)



5. DISCUSSION

Different training models used in this project were compared and summary table is presented below:

Model	Avg f1-score	Avg Precision	Avg Recall
Decision Tree	0.58	0.58	0.57
LR	0.58	0.58	0.58
Knn	0.56	0.55	0.57

5.1 f1-score(average)

F1-score is indicator of accuracy of the model. DT classifier and LR models have similar performance and have higher average f1 score as compared to knn. However, average f1-score doesn't represent clear picture as precision and recall are different for both the classes of the target variable and more biased towards the class "0-property damage".

5.2 Precision

Precision is represented as ratio of TP (True Positives) to sum of TP+FP(True Positives+False Positives).DT classifier and LR models have similar average precision values but again are predict class "0" with better accuracy. Both DT and LR exhibit similar precision of 0.69 for class 0 and 0.36 for class 1 which is reasonable. KNN model on the other hand is slightly lower in precision for both the classes with 0.67 value for "0" and "0.31" for "1".

5.3 Recall

Recall refers to the percentage of total relevant results correctly classified by the ML model. It is obtained by dividing TP (true positives) by TP+FN (True Positives+False Negatives).Recall for knn is highest for property damage while it is highest for injury collision in DT classifier.

6. CONCLUSION

It can be concluded all the learning models used behave similar for this data set and have reasonable scores but not great when compared with industry benchmarks. Few reasons accounting towards this could be the unbalanced data set, missing and unclear data values. In addition, other factors and more instances of the incidents could have helped in better predicting models.