Car Accident Severity

Jash Bikash

September 29, 2020

1. Introduction: Business Problem

Reducing traffic accidents is an important public safety challenge. Accident prediction is important for optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure, all in order to make the roads safer. This project is focusing on predicting the accident severity based on different attributes like locations, weather, road and visibility conditions, cause of different vehicles, collision due to inattention, rough driving by influence of drugs or alcohol and so on. In an effort to avoid and reduce the frequency of these type of accidents, we will build a model to predict the severity of an accident given the weather, the road, the light conditions and whether or not a driver involved was under the influence of drugs or alcohol. This way we would be able to bring awareness to the drivers and warn people about the possibility of getting into a car accident and its severity if it happens. This way people would drive more carefully or even change the travel if able to.

We will use our data science powers to be answered that how severe would be the accident if it happens by knowing the weather, road and visibility conditions.

2. Data

We have collected the raw data from SDOT Traffic Management Division, Traffic Records Group and contains data of all types of collisions that happened in Seattle city from 2004 to May/2020.

The data contains 194,673 samples and have 37 features. Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as weather, road condition, light condition, collision type and either driver was involved by influencing drugs or alcohol.

2.1 Missing Values

There are missing values on part of the data; some features have over 40% of missing data for that we will not consider them to our model. Removing the irrelevant data attributes away, the variables we will use to classify the severity of the accidents are:

- COLLISIONTYPE: Collision type
- WEATHER: Weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light condition during the collision
- UNDERINFL: Weather or not a driver involved was under the influence of the drugs or alcohol

These features contain missing values but it’s below 3% of the total amount of samples.

2.2 Target Variable

Our target variable SEVERITYCODE that corresponds to the severity of the collision:

1. Property Damage Only Collision which is the same as not injured collision
2. Injury Collision which is a binary classification problem

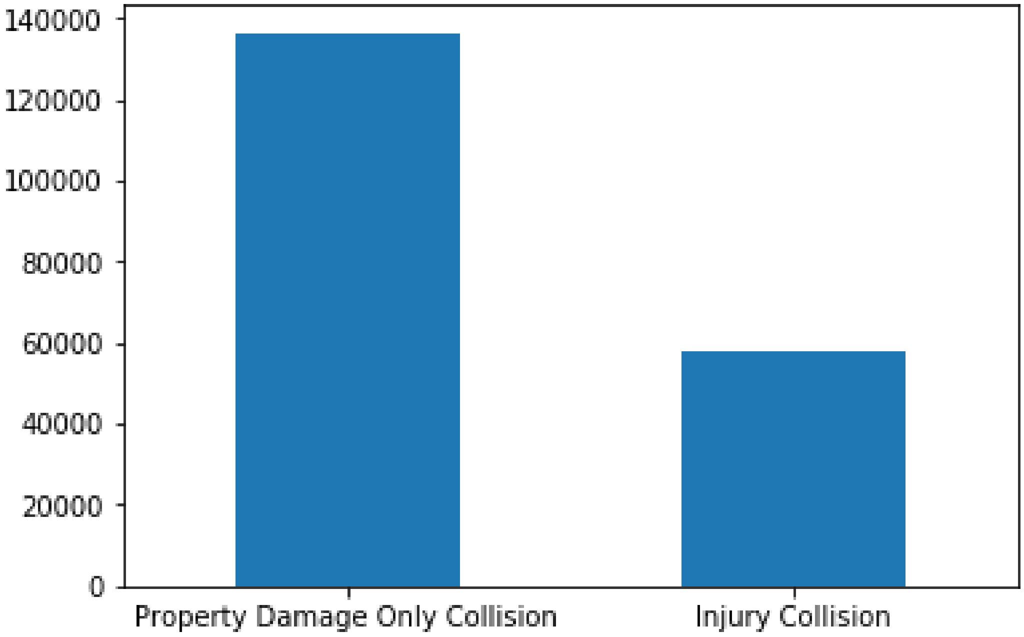


Figure 1: Collision by Severity Type

2.3 Annual Amount of Traffic Incidents in Seattle

We have noticed that there is a considerably high amount of incidents only discrepancy is from 2020 as it was recorded incidents that occurred till May 2020. We can also infer from the plots that no injury collisions are always more likely to happen.

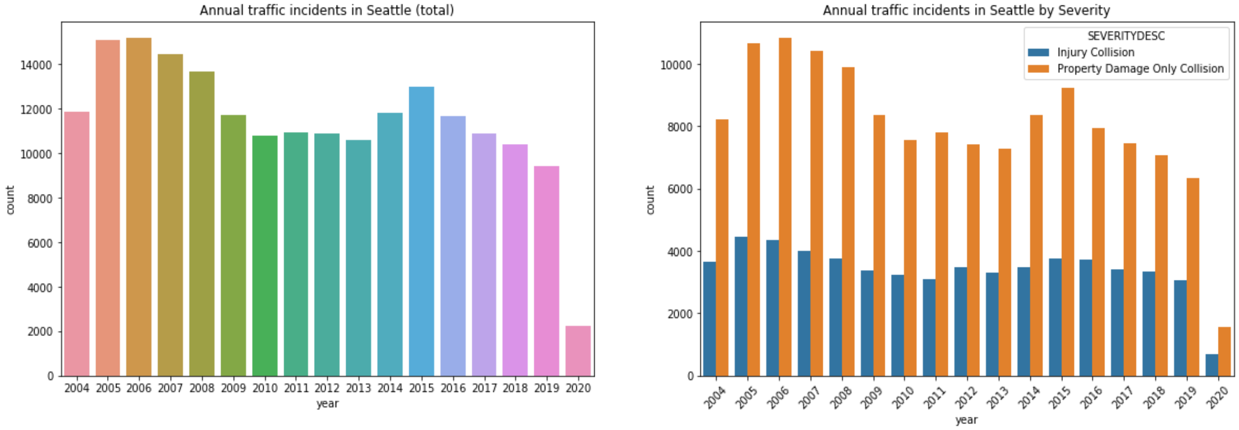
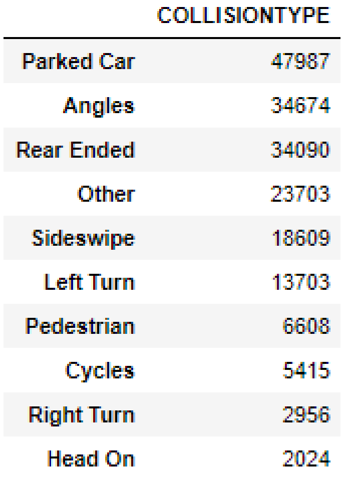


Figure 2: Annual Traffic Incidents in Seattle

2.4 Collision Type

There is a considerable difference on the collision occurrences according to collision type. The most three accidents were with parked cars, angles and rear ended. Other type of collision is also a matter of concern.



	COLLISIONTYPE
Parked Car	47987
Angles	34674
Rear Ended	34090
Other	23703
Sideswipe	18609
Left Turn	13703
Pedestrian	6608
Cycles	5415
Right Turn	2956
Head On	2024

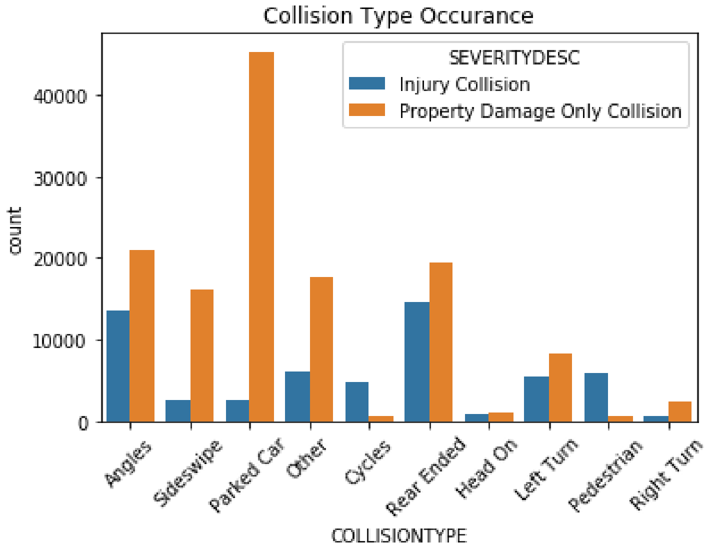
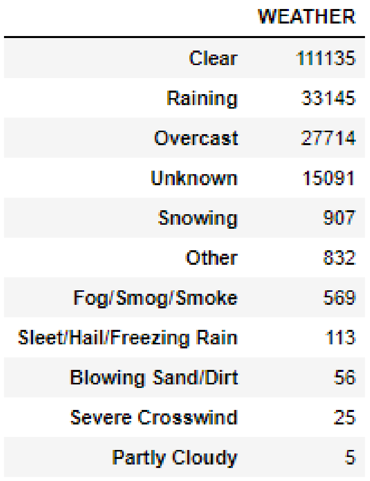


Figure 3: Collision Types by Severity

2.5 Weather Condition

Considering Seattle weather conditions, we notice most incidents happened in a Clear weather. That could be because drivers are less careful when there is no harsh weather condition. It would be interesting to check the correlation between WEATHER and INATTENTIONIND (whether or not collision was due to inattention), but there are too many missing values, 85% of the data is missing.



	WEATHER
Clear	111135
Raining	33145
Overcast	27714
Unknown	15091
Snowing	907
Other	832
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rain	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5

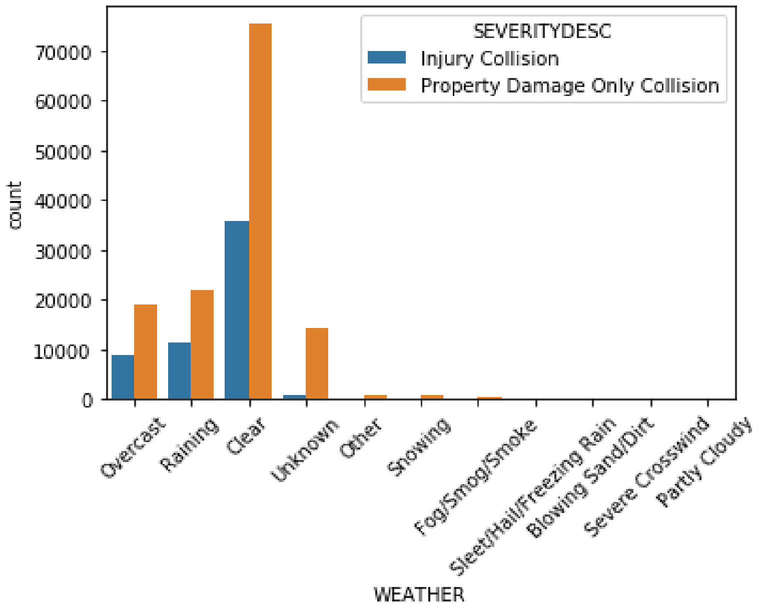
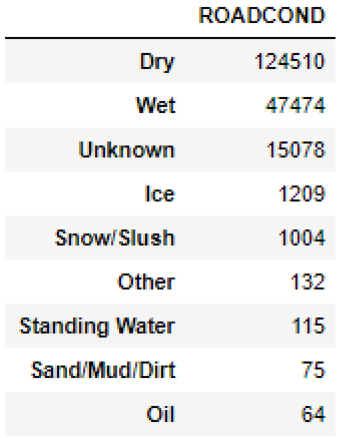


Figure 4: Weather by Severity

2.6 Road Condition

Considering the road condition we have found that there were more occurrences happened in dry road condition.



	ROADCOND
Dry	124510
Wet	47474
Unknown	15078
Ice	1209
Snow/Slush	1004
Other	132
Standing Water	115
Sand/Mud/Dirt	75
Oil	64

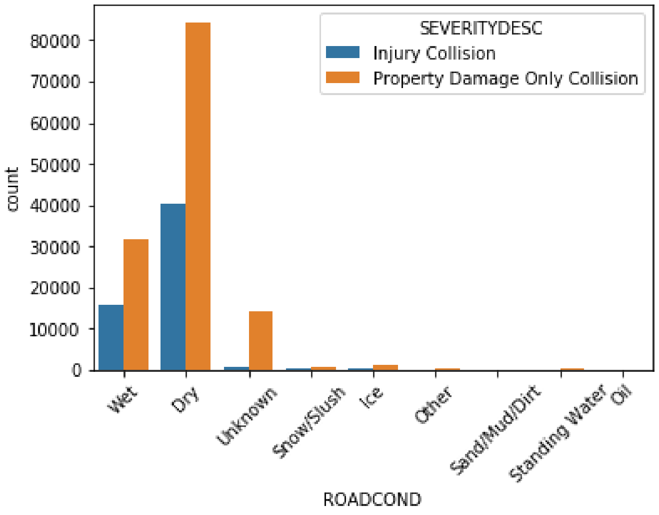
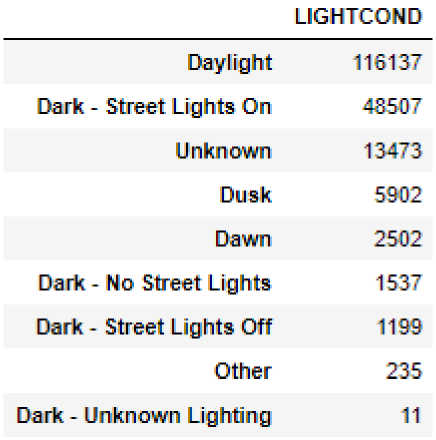


Figure 4: Road Condition by Severity

2.7 Light Condition

Most of the accidents happened in Daylight whereas Dark-Streer Lights On is also considerable for accident.



	LIGHTCOND
Daylight	116137
Dark - Street Lights On	48507
Unknown	13473
Dusk	5902
Dawn	2502
Dark - No Street Lights	1537
Dark - Street Lights Off	1199
Other	235
Dark - Unknown Lighting	11

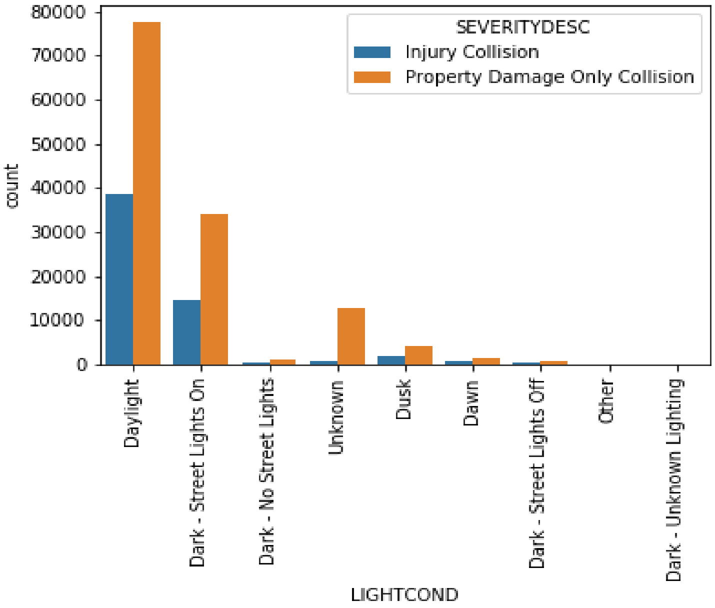
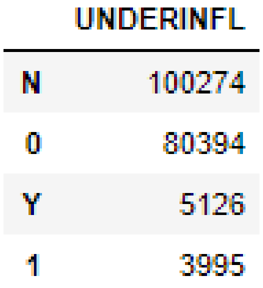


Figure 5: Light Condition by Severity

2.8 Diver Under Influence of Drugs or Alcohol

It has been seen that drivers were not under any influence in most the incidents.



	UNDERINFL
N	100274
0	80394
Y	5126
1	3995

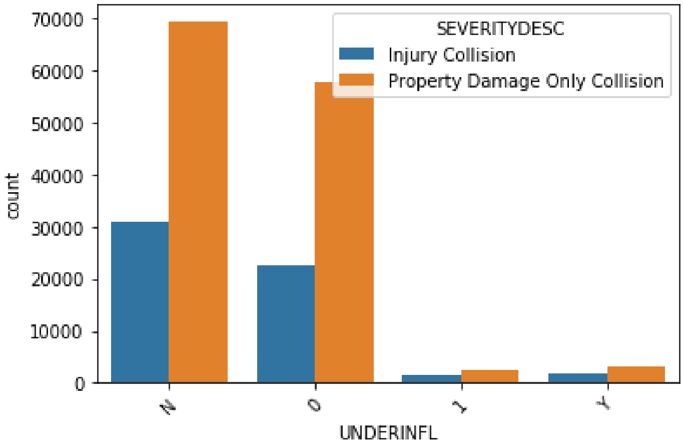


Figure 5: Driver Influences Chart

Each feature have a different weight of influence on the severity of the collision. Overall, all of them are consistently infering that no-injury accidents in normal driving conditions are more recurrent.

We will use COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND and UNDERINFL as attributes to classify SEVERITYCODE. For that we will need to prepare this features so it is suitable for a binary classification model. We will use some popular machine learning algorithms like SVM, Logistic Regression, Naive Bayes and KNN for build up models to analyze their performance and predict the collision severity.

3. Methodology

In this project we will use COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND and UNDERINFL as attributes to classify SEVERITYCODE**.** We will limit our analysis to this four independent variables. For that we will need to prepare these features so it is suitable for a binary classification model.

In the first step we will prepare and clean the dataset to make it readable and suitable for the machine learning algorithms. There are 37 attributes whereas we will consider the features of COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND and UNDERINFL to classify SEVERITYCODE. Those attributes has 3% missing data which we will drop them. We will split this dataset as train and test split whereas 70% to train the model and 30% to test the model.

Second step in our analysis will be calculation and exploration of different models to find out the main problem for severity. We will use 3 classification models which are Logistic Regression, Decision Tree an KNN. After obtaining each model's predictions we will evaluate their accuracy, precision, f1-score, log-loss and compare and discuss the results.

4. Analysis

4.1 Data Preparation and Cleaning

Data preparation and cleaning is required to make the dataset readable and suitable to the machine learning algorithms.

As we have choose 5 attributes to classify SEVERITYCODE so that we have dropped other attributes out of 37 attributes and also dropped the 3% missing data from those attributes. We have converted the categorical values to numerical values by using label encoding technique.

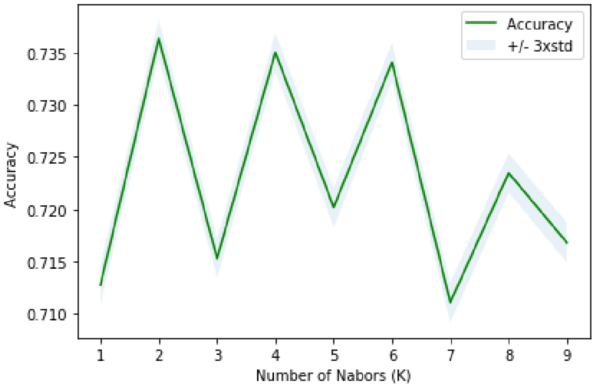
Then we separated the independent variables to a dataset and dependent variable 'SEVERITYCODE' to another dataset. After that use this data to randomly pick samples and split in below ratio:

- 70% to train model
- 30% to test model

4.2 Classification: Modeling and Evaluation

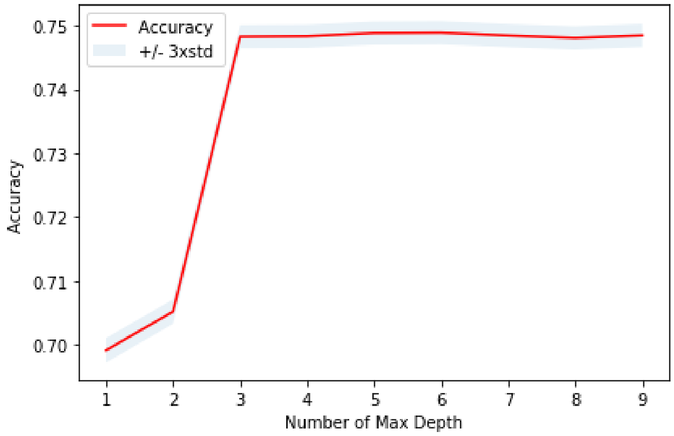
The prepared was used to model three classification models.

KNN: Classifies unseen data through the majority of its ‘neighbours’. In this case we already know K=2 (2 classes of SEVERITY CODES). After obtaining each model’s predictions we will evaluate their accuracy, precison, f1-score, log-loss and compare and discuss the results.



```
The best accuracy was with 0.7363676379963024 with k= 2
```

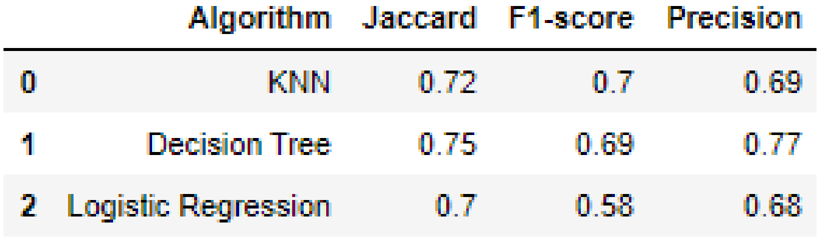
Decision Tree: Classifies by breaking down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.



Logistic Regression: Classifies data by estimating the probability of classes.

4.3 Model Evaluation Using Test Set

Among all three models, Jaccard score's measures accuracy is above 70%. The highest accuracy model is the Decision Tree Classifier. The same model also presents the best F1_score and Recall(True positive rate).



	Algorithm	Jaccard	F1-score	Precision
0	KNN	0.72	0.7	0.69
1	Decision Tree	0.75	0.69	0.77
2	Logistic Regression	0.7	0.58	0.68

Figure 6: Model Evaluation

5. Results and Discussion

In this analysis we evaluated the performance of 3 machine learning algorithms on the Seattle Collision dataset to predict the severity of an accident knowing the weather and road conditions.

The three models performed very similary, but Decision Tree stood out with a difference from KNN and Logistic Regression during the evaluation with the model's accuracy.

6. Conclusion

Purpose of this project was to analyze the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. We picked 5 features out of 37 where it showed to be a reasonable choice to find the answer that were searcing for. It was able to achieve 77% accuracy however there were still significant variances that could not be predicted by the models in this study. Those 5 features have somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).