# Car Accident Severity

Abah Jemimah

October 2020

**Introduction: Business Problem**

Traffic accidents cause serious threat to the human life worldwide, the economy and reduces efficiency in transportation. Some factors that contribute to the risk of collisions are; vehicle design, speed of operation, road design, road environment, weather conditions, lighting conditions, driving skills, impairment due to alcohol or drugs, and behaviour, notably distracted driving, speeding, and street racing.

It is therefore important to build a model for accident severity prediction for effective performance of road traffic systems for improved safety and minimized collisions.

***Stakeholders***

Stakeholder involved in this includes:

1. Car owners
2. Healthcare workers
3. Government
4. Commuters
5. Logistics
6. Professional Drivers

**Data Understanding**

The data to be used for this project is raw data from the SDOT Traffic Management Division, containing all types of collisions that occurred Seattle city from 2004 to 2020

The data contains 194,673 samples and have 37 features. we will use SEVERITYCODE as our dependent variable Y. Since the observations are quite large, we will have to work on 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, either driver was involved by influencing drugs or alcohol, junction type, speeding or not speeding.

***Data Description***

1. COLLISIONTYPE: Collision type
2. WEATHER: Weather conditions during the time of the collision
3. ROADCOND: The condition of the road during the collision
4. LIGHTCOND: The light condition during the collision

5. UNDERINFL: Whether a driver involved was under the influence of the drugs or alcohol
6. SPEEDING: speeding or not speeding
7. ADDRTYPE: Type of area

***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.

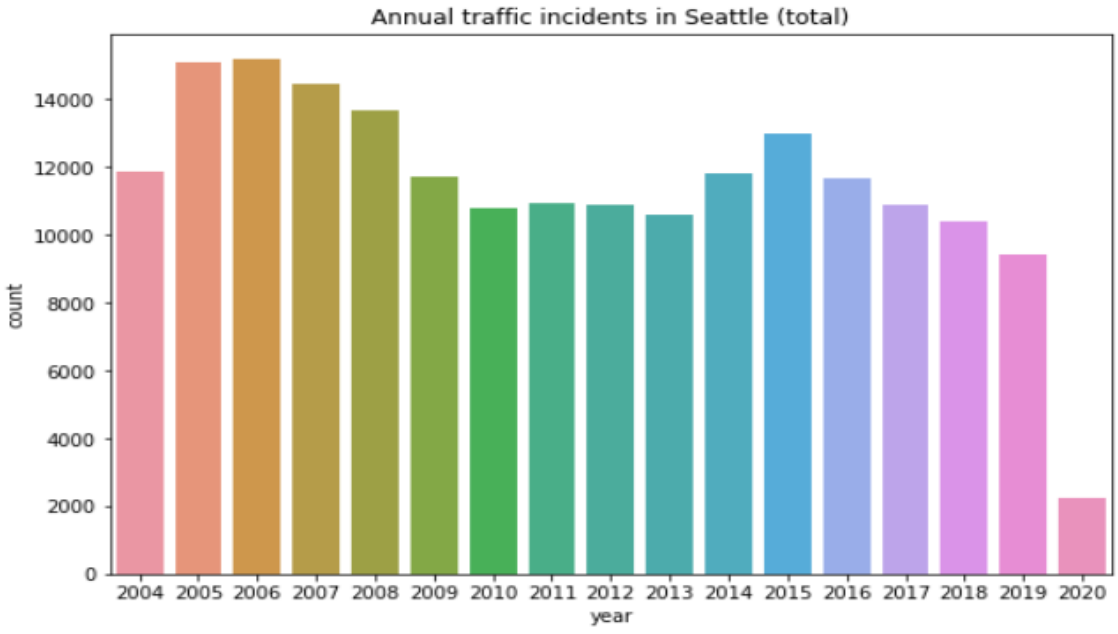
***Target Variable***

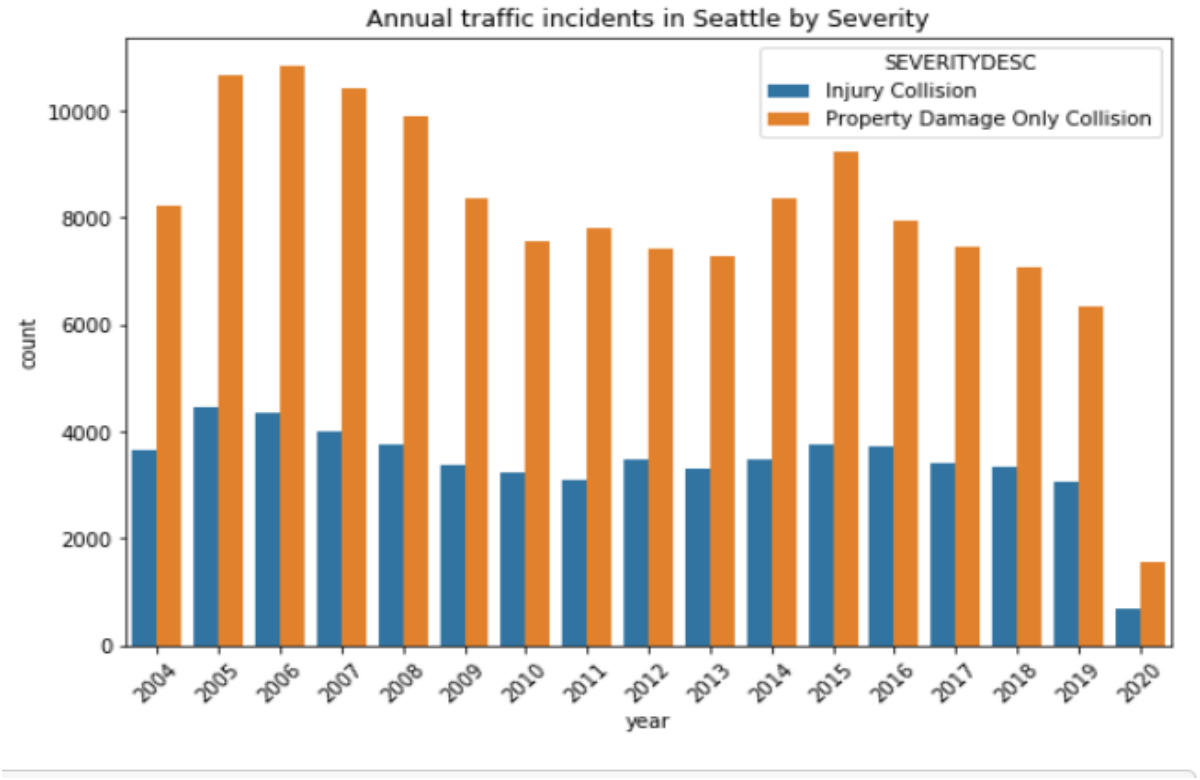
The target variable for this project is 'SEVERITYCODE' that corresponds to the severity of the collision and is represented as follows:

1. Property Damage Only Collision
2. Injury Collision

| | SEVERITYDESC |
|---|---|
| Property Damage Only Collision | 136485 |
| Injury Collision | 58188 |

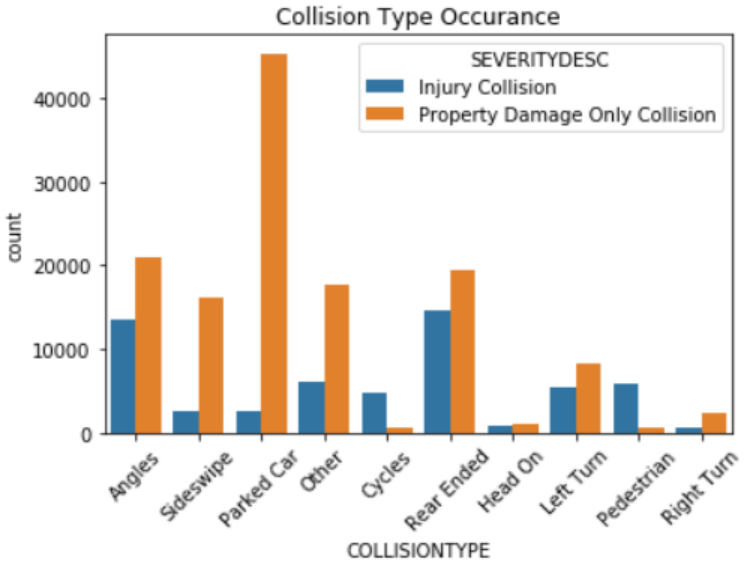
***Annual Amount of Traffic Incidents in Seattle***



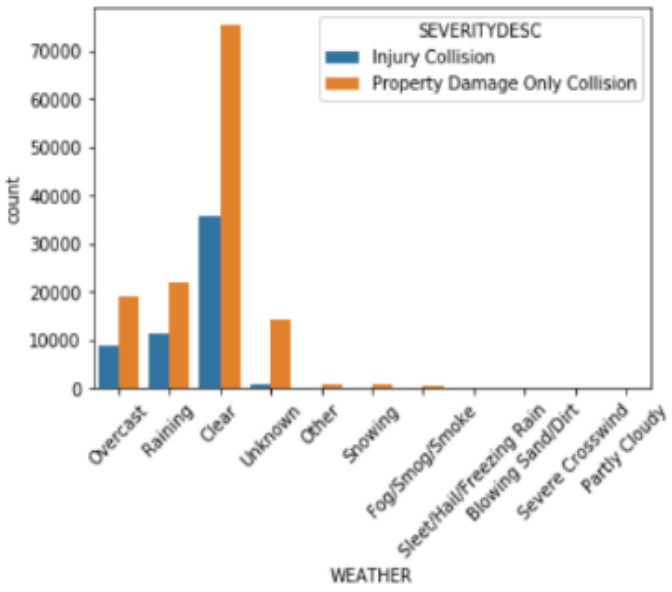



***Collision Type***

| 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 |



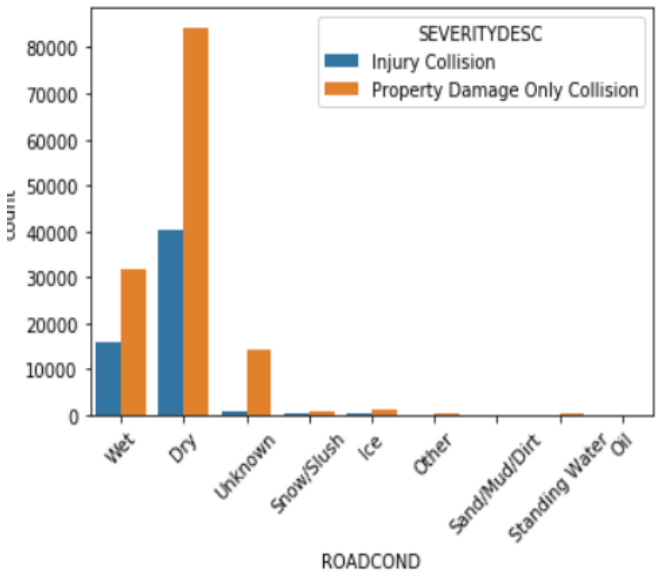
***Weather Condition***




| | 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 |

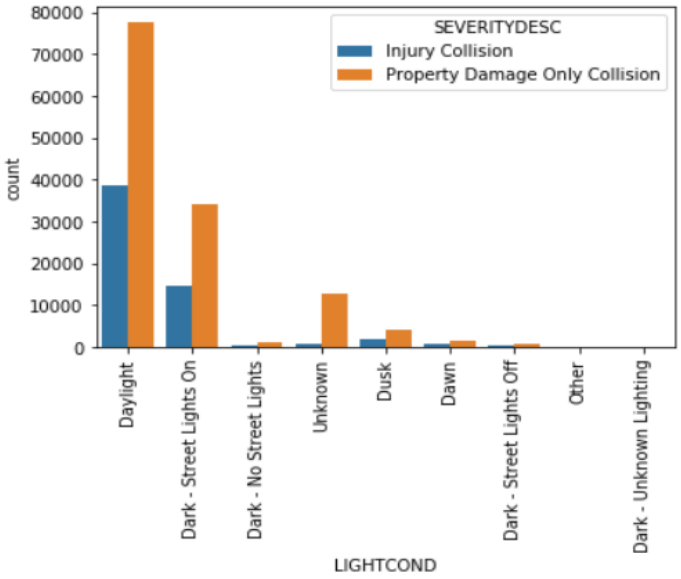
***Road Condition***

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



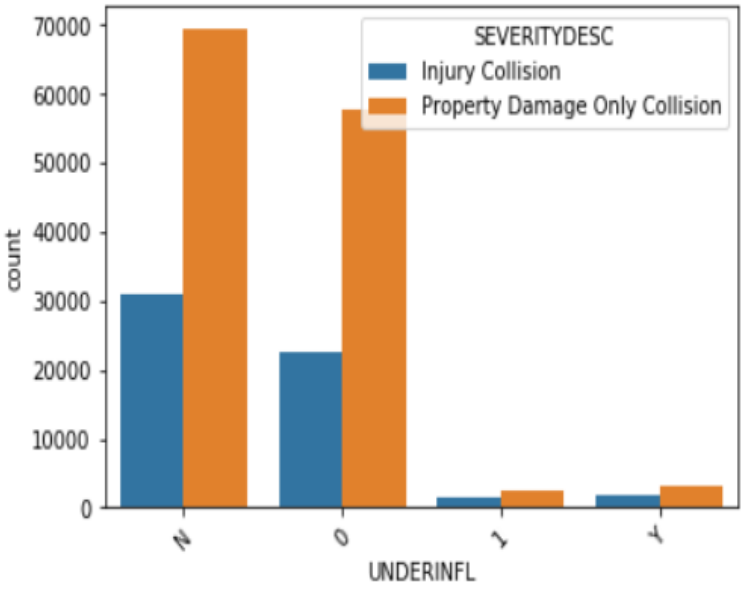
***Light Condition***

| | 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 |




***Under Influence of Alcohol***

| | UNDERINFL |
|---|---|
| N | 100274 |
| 0 | 80394 |
| Y | 5126 |
| 1 | 3995 |



**METHODOLOGY**

In this project we will use ADDRTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, COLLISIONTYPE as features (Independent variables) to classify SEVERITYCODE. 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 thirty-seven (37) attributes and six (6) were used in this project. Missing data was replaced and then the data was SMOTE(undersampled). We will split this dataset as train and test split whereas 70% to train the model and 30% to test the model.

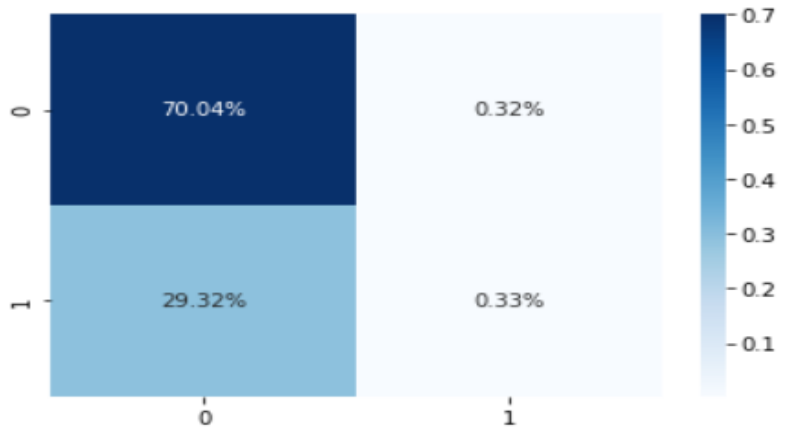
Second step in our analysis will be calculation and exploration of different models to find out the main problem for severity. We will use 3 classification models which are Decision Tree, Logistic Regression an SVM. After obtaining each model's predictions we will evaluate their accuracy, precision,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

**MODELING AND EVALUATION**

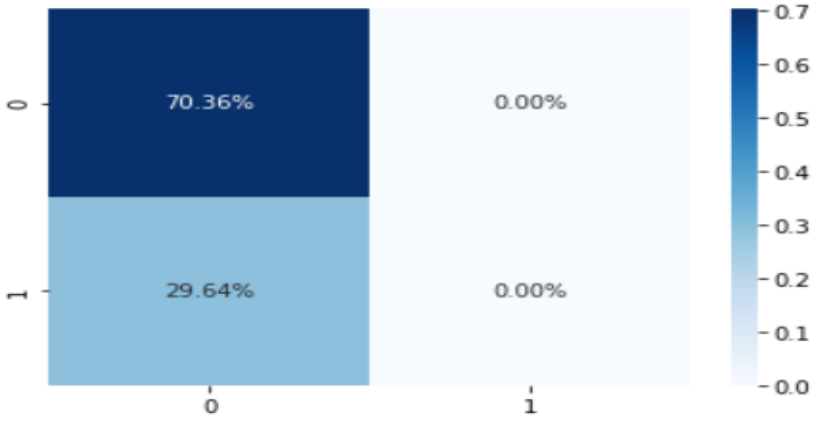
Dataset was smote and divided into training and testing dataset in a ratio of 7:3.

***Decision Tree***

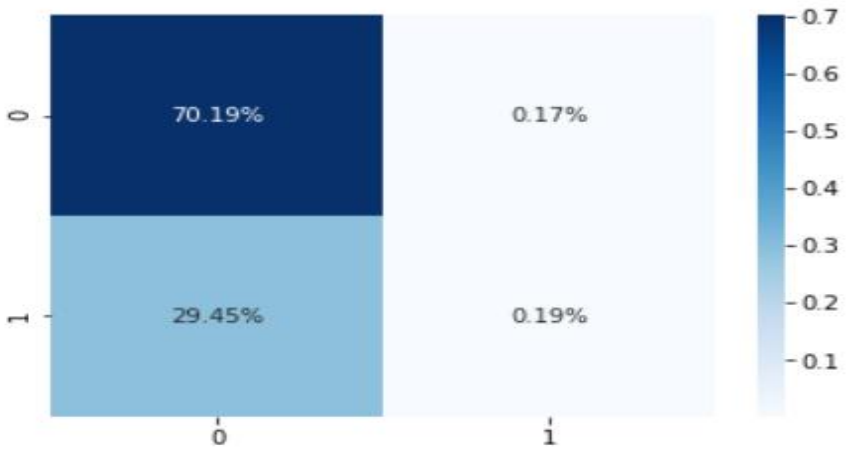
Classifies by breaking down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. 0.7037772679017842 accuracy was gotten from this model




***Logistic Regression***




***SVM***




***Evaluation***

Among all three models, accuracy score's measures accuracy is above 70%. The highest accuracy model is the SVM Classifier. The same model also presents the best F1_score , jaccard and precision.

| | Algorithm | Accuracy | Jaccard | F1-score | Precision |
|---|---|---|---|---|---|
| **0** | Decision Tree | 0.7 | 0.0 | 0.6 | 0.6 |
| **1** | Logistic Regression | 0.7 | 0.0 | 0.58 | 0.5 |
| **2** | SVM | 0.704 | 0.006 | 0.585 | 0.652 |

**RESULT 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 similar, but SVM had better accuracy than Decision Tree and Logistic Regression during the evaluation with the model's accuracy.

**CONCLUSION**

Purpose of this project was to analyse the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. We picked 7 features out of 37 where it showed to be a reasonable choice to find the answer that were searching for. It was able to achieve 70.4% accuracy however there were still significant variances that could not be predicted by the models in this study.