# Coursera Capstone:-

# *Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Seattle city*

## Table of contents

* Introduction
* Data
* Data Pre-Processing
* Methodology
* Classification
* Map
* Results and Discussion
* Conclusion

# 1.Introduction

In recent years, the road accident has become a global problem and marked as the ninth prominent cause of death in the world. Due to the enormous number of road accidents every year. It is entirely inadmissible and saddening to allow its citizen to kill by road accidents. Consequently, to handle this overwhelmed situation, a precise analysis is required. Road Accident is the most undesirable and unexpected thing to occur to a road user, though they happen quite often. It has a massive impact on society as well as in the economy of our country as there is an immense cost of fatalities and injuries. Besides this, according to WHO, the economic cost of road accidents to a developing country like us is 2-3% of GDP, which is a significant loss for a country.

"This notebook aims to use historical dataset to explore car accident severity and to build a model that can predict its future occurring severity level. The sole objective of the process is to minimize future road accident occurrence. The identification of potential safety hazards on new road project at the appropriate Type, so that they can be eliminated or otherwise treated to mitigate their adverse effect at minimum cost. To decrease the rate of accidents at a particular location. The identification of potential safety hazards features of an existing road so that they can be eliminated or otherwise treated before they become accident prone location."

# Data description

Our goal in this project is to build a machine learning solution to predict the severity of accidents. Data that will be used for analysis was obtained from the Seattle government open data portal website. Dataset was downloaded as CSV file and includes records of traffic accident that took place in Seattle City from 2004 to Present. All collisions provided by SPD and recorded by Traffic Records. The raw sample has a total of 194673 accidents with 38 columns. The main target variable is “Severity” which is a categorical variable that supports 2 classes: “Injury Collision”, “Property Damage Only Collision”. The attributes include the time and date of accident, road type where the accident occurred, the state of road surface, weather conditions, as well as number of people and bicycles and vehicles involved etc.

# Methodology

## Data pre-processing

Once the raw data is obtained, a pre-processing step is required to clean data and finalize the

feature that will be used for modelling. For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the approaches. We split our entire dataset into two parts Training Dataset and Test Dataset. 70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset.

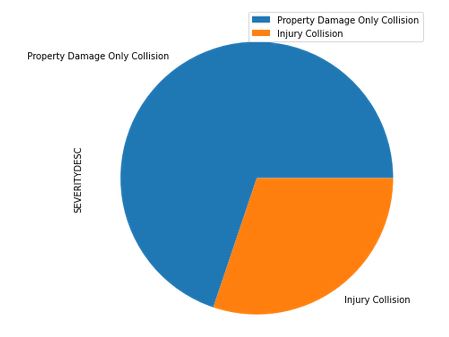
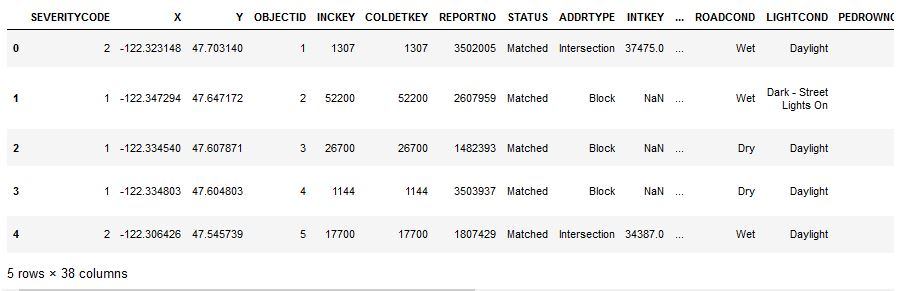


Figure 1 Severity Proportion for each class

I try to use dataset for Seattle city and learn how to deal with the accidents data. The first column it labelled data. The remaining columns have different types of attributes. Some or all can be used to train the model. The label for the data set is severity, which describes the fatality of an accident. You will notice that the shared data has unbalanced labels. We should balance the data; I recommend you go through them carefully.



In this dataset, all the accident records were written with formal words. We properly organize this total dataset based on the feature. In total, we have found 37 factors that affect previous accidents in some way. Firstly, we methodize all accident records by using these features. After that, for many accident records, we have found missing values in the total dataset as these missing values can affect the performance, on account of this, we have applied a method by using the mean value of that feature column to provide an amount where it is required. We use this method as there presents no extreme value which can affect the mean. Working with a large number of features may affect the performance because training time increases exponentially with the number of features. Even, it has also the risk of overfitting with the increasing number of features. So, for getting a more accurate prediction, feature selection is a critical factor here.

After examining the attributes of the dataset, we can see that some columns are irrelevant to our problem. So, we simply drop them from the dataset. A total of 17 attributes were excluded from further analysis such as 'INCKEY’,'COLDETKEY’,'REPORTNO', 'STATUS', 'INTKEY', 'LOCATION', 'SEVERITYCODE', 'JUNCTIONTYPE', 'UNDERINFL', 'PEDROWNOTGRNT’, 'EXCEPTRSNCODE’, 'SDOTCOLNUM’, 'SPEEDING’, 'ST\_COLCODE', 'SEGLANEKEY', 'CROSSWALKKEY’, 'HITPARKEDCAR' and so on. The first 5 rows of our dataset is shown in Table 1.

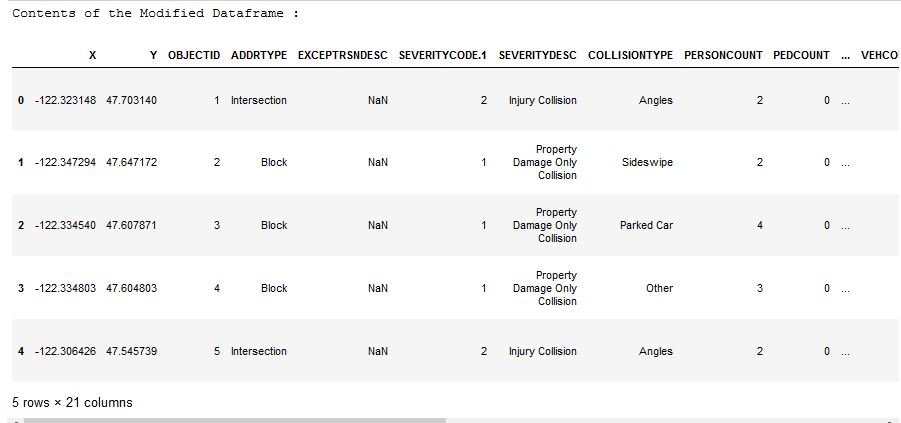
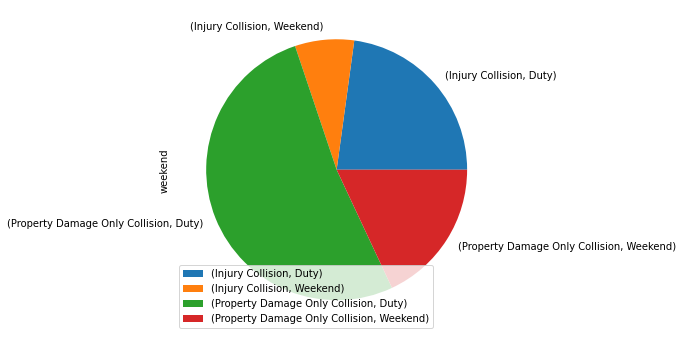


Table 1 First 5 rows of accident dataset after cleaning

Before moving to model building, some feature engineering work is thought to be useful. For the time, I think it is better to convert it categorical variable.

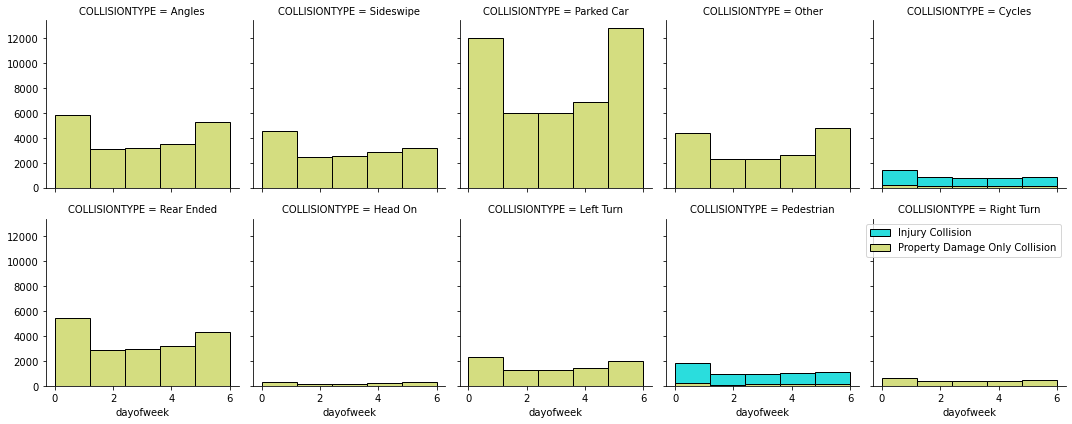
Lets analysis some relationships between Severities and some attribute column:-

1. Weekends' and Severity Cases:
2. Different types of Collision and their respective counts day wise:
3. Light Condition and Severity Cases:
4. Road Condition and Severity Cases:
5. Weather and Severity Cases :



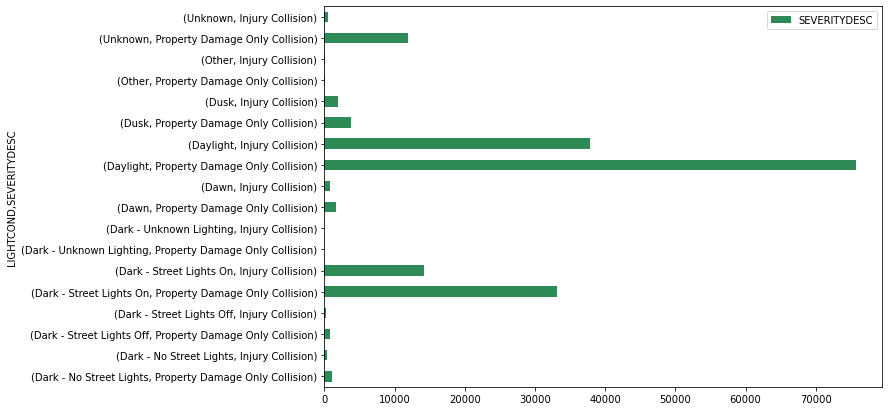
Weekends' and Severity Cases:

In this we can see that most of the collision takes place during rush hour, when people are hurry in their daily jobs and accidents cause most of the property damage.



Different types of Collision and their respective counts day wise:

From this plot it is evident that most of the collision incidents takes place during Car Parking, at angles, slide swipe and rear ended and severe injuries occur during right turn and beside pedestrian.

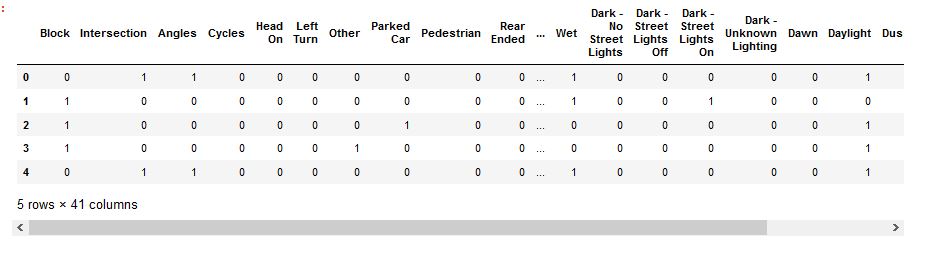


Light Condition and Severity Cases:

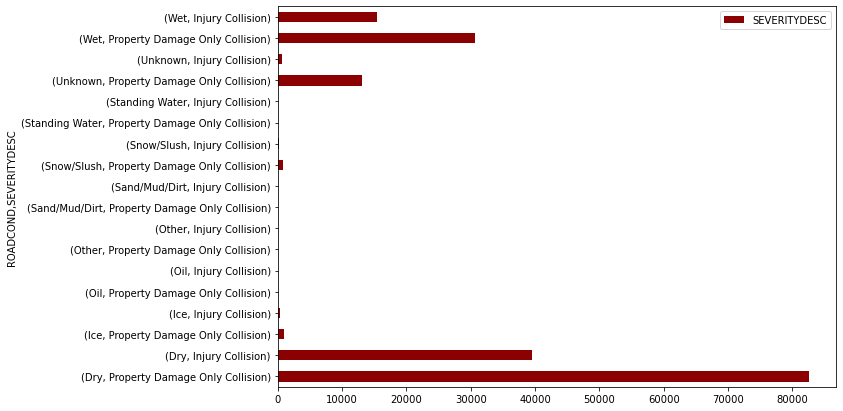
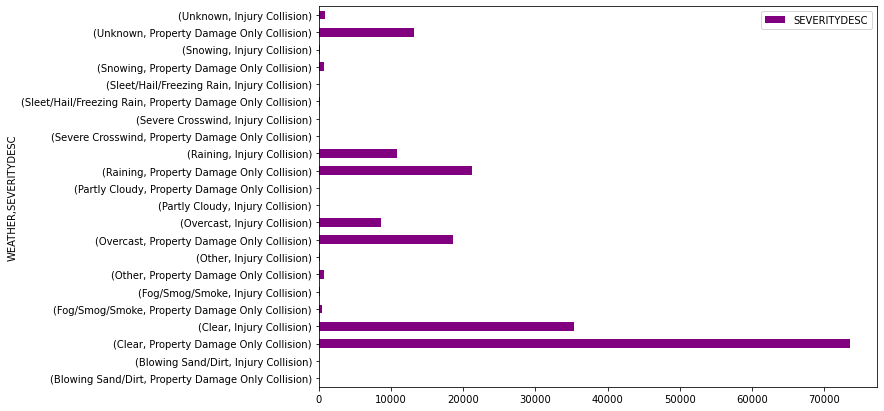
Since the data is obtained from Seattle, a quick internet search indicates that there are 2-time intervals that are considered as peak hour: 6:30 -> 9:30 and 16:00 -> 19:00. So, “Time” column will be converted to 2 values: Daylight and Dark.

From this we can infer that most accident takes place when road is dry, clean, broad and after second cases occur during rain when some accident occur due to poor visibility and skidding of tires.

In the final stage of the data pre-processing, we'll need to convert categorical features to dummy variables using pandas. Otherwise, our machine learning algorithm won't be able to directly take in those features as inputs. The final table of input variables can be seen in the Jupyter notebook.



Converted Categorical Features to Dummy Variables

Road Condition and Severity Cases Weather and Severity Cases

## Model Building

In order to build the optimal model, 4 Machine learning algorithms will be used, evaluated, and compared to choose the classifier that shows the best performance.

The classifiers are:

1. Logistic Regression (LR)
2. K-Nearest Neighbour (KNN)
3. Decision Tree (DT)
4. Support Vector Machine (SVM)

For the best model, further evaluation will be performed using precision, recall, F1-score and support to evaluate the models.

# Results

All the code implementation was performed in python using a Jupyter notebook. Various packages were used for data pre-processing as well as build the machine learning models. These packages include: Numpy, Pandas, Matplotlib, Seabron, and sklearn.

Data were splitted into 80% training data and 20% validation data. We can clearly see that Logistic Regression model and Support Vector Machine model are the best. The results are shown in table 2.

Table 2 Accuracy of candidate models.

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score** |
| LR | 0.7325 | 0.8254 |
| KNN | 0.727 | 0.79 |
| DT | 0.742 | 0.82 |
| SVM | 0.734 | 0.8254 |

SVM slightly outperforms, so we choose it for further validation. So we choose it for further analysis. This time we will use other evaluation metrics such as precision, recall, F1-score, and support. Please see Table 3.

Table 3 Evaluation of classification performance of SVM model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Injury Collision | 0.69 | 0.19 | 0.30 | 597 |
| Property Damage Only Collision | 0.74 | 0.96 | 0.83 | 1403 |

# Discussion

In previous section, we found that SVM, LR, DT model performs the best, however, with further analysis, SVM does not seem to show optimal performance. One reason could be the critical imbalance between the 2 severity classes. Other data treatment techniques can be used to improve the input data by using oversampling or under sampling techniques to create balance between the different classes.

Another problem can be the hyper parameter tuning which can improve the classification performance of the models and provide better results through optimization of parameters of each model.

# Conclusion

In this project, we propose a machine learning model to predict the severity of traffic accidents in the city of Seattle. A comparative approach was followed to identify the best performing model among 4 models. The best model was support vector machine and Decision Tree. Further improvements are required to optimize the model performance.

Note:

Some Points are inspired by studying other research paper and the way they write their project report. So,

**Thanks** to them.

Regards

Suraj