REPORT

Capstone project-final

Introduction/Business Problem:

There are several factors associated with the risk of car accidents including traffic, weather conditions, lighting conditions on streets, drunk driving, roads situation, etc. My project report is focused on analyzing all these attributes and developing a model in order to predict whether the outcome of a road accident is influenced by these outside factors, and the degree of severity of these collisions, taking in consideration the weather situation, traffic or road problems.
To interpret the regular mishaps of car accidents, we'll be deploying a model in order to predict the degree of severity of an accident given the current weather, road and visibility situation. This model would help and assist the drivers to gauge the sort of conditions that can help them avoid situations that could lead to such accidents.

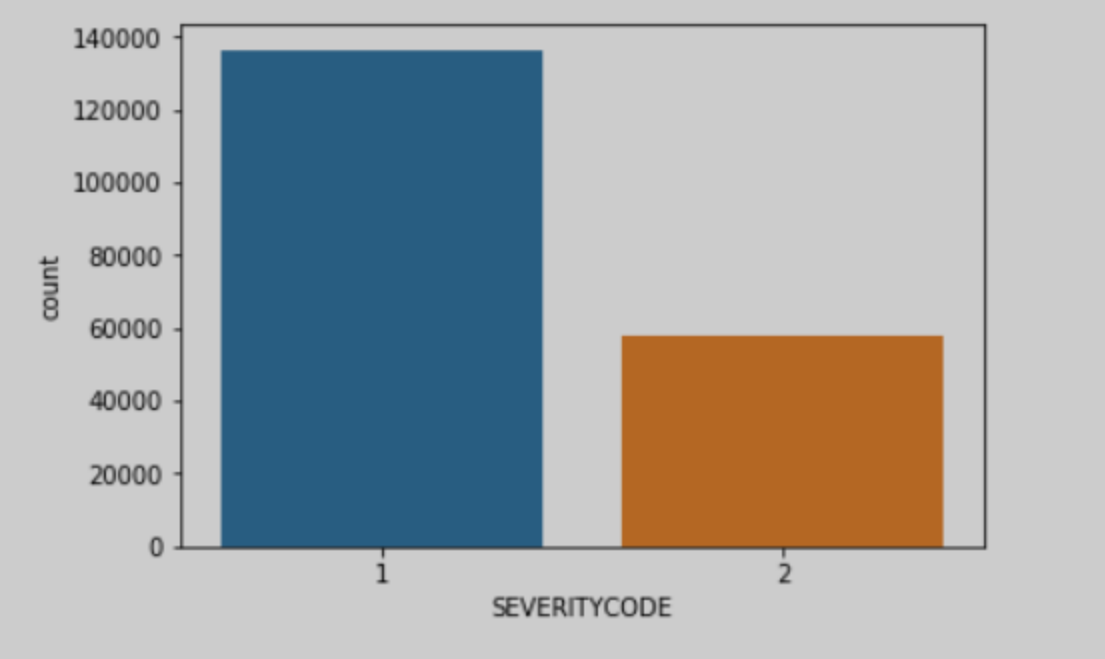
Data understanding:

This is the initial phase where we come to understand the project's objective from the business or application perspective. Then, we need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.

In this phase, we'll need to collect or extract the dataset from various sources(here, csv file). Then, we'll need to determine the attributes (columns) that we're going to use to train the machine learning model. Also, we will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

The data we are using is recorded by Traffic Records division and the information comes from Seattle Police Department that contains data of accidents that happened in the city from the period of 2004 to May,2020. The data that is contained within a csv file is constituted of 37 attributes or features and 194673 rows, that contains various samples, the information is labeled and covers road conditions, weather, fatality records etc.

We have the dependent or target variable as 'SEVERITYCODE' in the dataset which describes the degree of severity of the accident within the data. SEVERITYCODE takes values 1 (for non-injury accidents) and 2 (for accidents involving injury)



Data preparation and cleaning:

The data preparation stage involves all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

It is the process of cleaning and transforming raw data prior to further processing and analysis. This is an important step that precedes further steps of processing and it usually involves modifying, removing missing values, reformatting data, making corrections and adjustments to the data or even combining of different data sources to enrich the dataset.

Data preparation steps usually involve-

- Collecting the data
- Discovering and assessing the data
- Cleansing and validating data
- Transform and enrich data
- Store data

The columns ‘EXCEPTRSNCODE’, ‘PEDROWNOTGRNT’, ‘EXCEPTRSNDESC’, ‘INATTENTIONIND’ ,‘INTKEY’ have more than 50% of missing or null values hence we’re going to have to drop them from the DataFrame. We’ll then deal with the null or the missing values and clean the data.

```
null = df1.isnull().sum()
null[0:15]
```

```
SEVERITYCODE         0
OBJECTID             0
INCKEY               0
STATUS               0
ADDRTYPE             0
SEVERITYCODE.1       0
SEVERITYDESC         0
COLLISIONTYPE        0
PERSONCOUNT          0
PEDCOUNT             0
PEDCYLCOUNT          0
VEHCOUNT             0
INCDATE              0
JUNCTIONTYPE         0
UNDERINFL         4884
dtype: int64
```

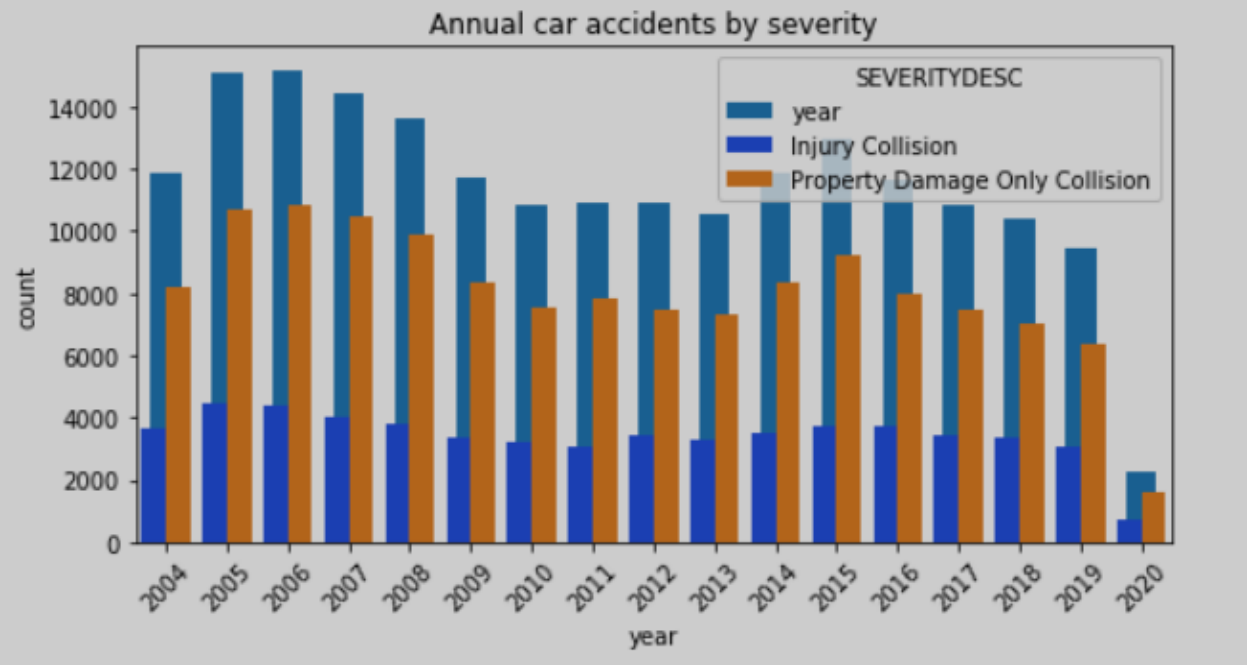
We will extract only the attributes we are using for applying the algorithms to calculate the severity, like 'WEATHER', 'ROADCOND' and 'LIGHTCOND' and COLLISIONTYPE and build our model based on those attributes. There are many columns with type *object*, we're going to convert them to numerical data type in order to be able to fit the machine learning algorithm.

We'll employ label encoding in order to covert the columns into our required data type to make them usable for the ML modeling.

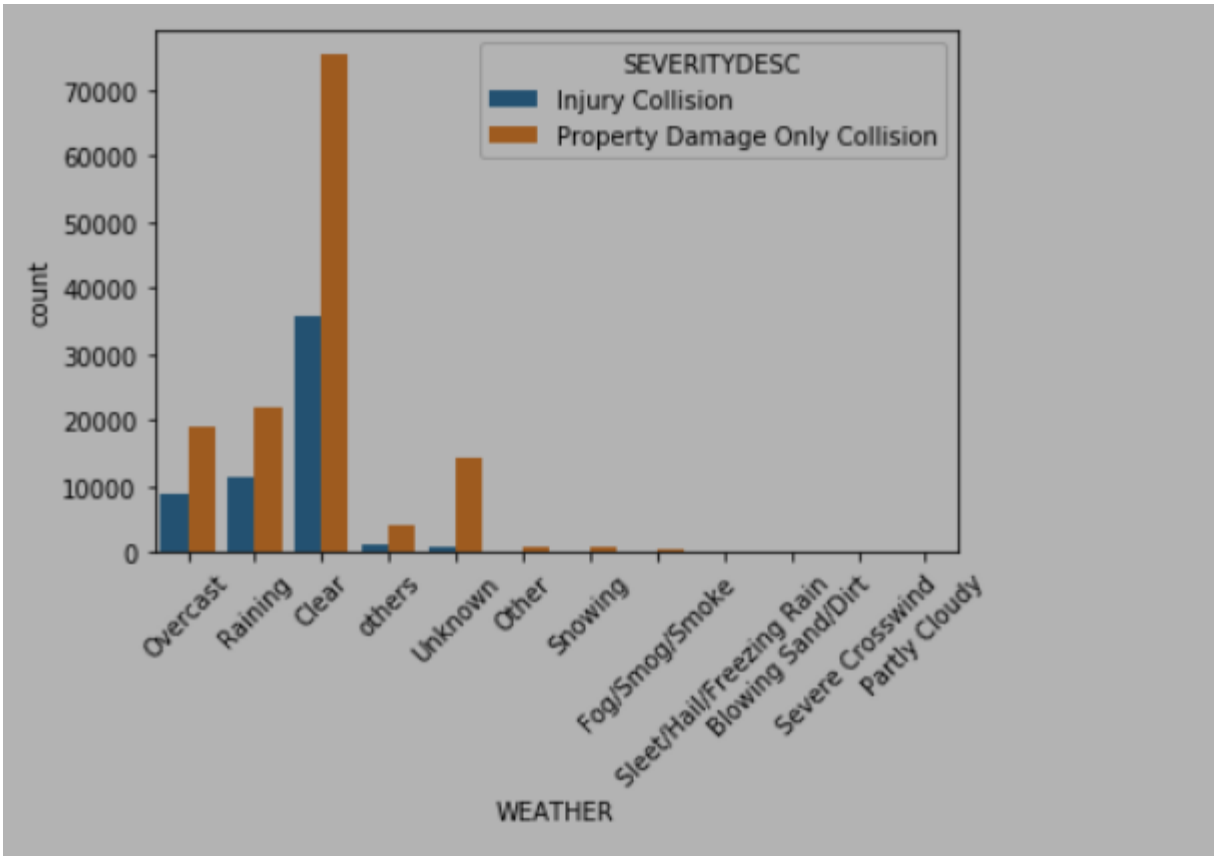
Data visualization:

Now we'll plot the annual accident data that's occurred from 2004 till May 2020, since the data for 2020 is incomplete hence we have less number of accidents recorded for the year.

From the plot, we observe there's higher possibility of property damage caused due to accidents rather than them leading to injury.

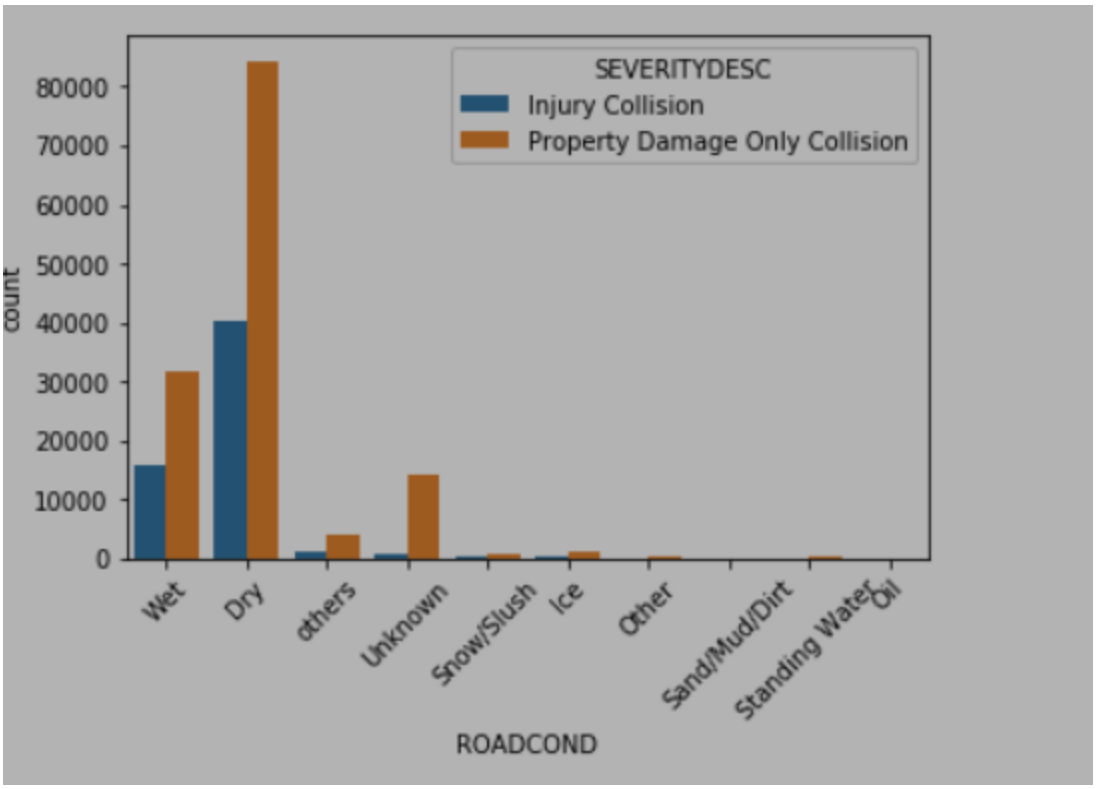


On plotting against weather, it can be observed that the most accident occurrences have happened in 'Clear' weather situations unlike we originally expected. This tells us weather is not as much responsible and indicative of having accidents as we thought it was, there's likely possible that drivers are less likely to pay attention on roads/drive rashly while its clear weather than say, rainy.



We also observed that the most number of accidents happened in ‘Dry’ road conditions hence no big correlation b/w road conditions and accidents.

Similar outcomes were found when we plotted for road conditions, lightning and driving under influences.



All in all, it can be concluded that a majority of these accidents were non-injury based accidents happened under fairly normal driving conditions with little to no outside influences.

Modeling and Evaluation results: As our data is cleaned and extracted, we're ready to build our models. We've used the following main algorithms for our modeling.

- K Nearest Neighbor(KNN)
- Decision Tree
- Logistic Regression

	Algorithm	Jaccard	Precision
0	KNN	0.73	0.7
1	Decision Tree	0.75	0.77
2	Logistic Regression	0.7	0.68

In conclusion out of all the classification algorithms, the best classifier algorithm seems to be **Decision Tree** with the highest positive accuracy score.

We find the confusion matrix for the given algorithms, reveals the umber of samples that were rightly classified. *'A **confusion matrix** is a table that is used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.'* We note variation that

occurs while comparing false positives & true positives while true negatives & false negatives seem stable.

Discussion and Conclusion:

In this notebook we employed a number of libraries(mainly pandas, NumPy) to interpret, modify and understand the dataset in order to identify the severity of accidents or type of accidents happening during road collisions.

The major objective of our project was to predict whether these accidents are caused due to changes in weather conditions, roads, drugs influence, lighting etc. and to find out the severity of these accidents so as to help the drivers avoid such mishappenings.

We find out the missing values, clean the data and apply various data visualization tools to interpret the data, plot various findings, used label encoding to create new class to get the new required data type for modeling purpose. Finally we run our cleaned data through 3 main machine learning algorithms, namely; K-Nearest Neighbor, Decision Tree and Logistic Regression. We further employ these algorithms to build a model and get optimal values to predict the outcome of the analysis. The three models we used were almost similar in performance, but Decision Tree provided the best possible accuracy score for the dataset.