Seattle’s Car Accidents Severity Project

Applied Data Science Capstone Project

IBM Data Science Professional Certificate Specialization

Jordan López

https://github.com/lopezjordan12

Introduction

- This is the final project of my IBM Data Science Certified Professional offered by Coursera.

- It is the 9th course of the program that I started early 2020.

- This project will focus on exploratory and statistical analysis of data from 2004 to august 2020 about car accidents in the city of Seattle, USA. This data give an indicator of severity of the accidents, and this is the goal of the project. Predict this indicator.

- With several machine learning supervised algorithm the indicator is estimated with a successful accuracy percentage.

Data

- One main data that is used to solve this problem is the Seattle data of car accidents. This data is an open data, and everyone can access to it.

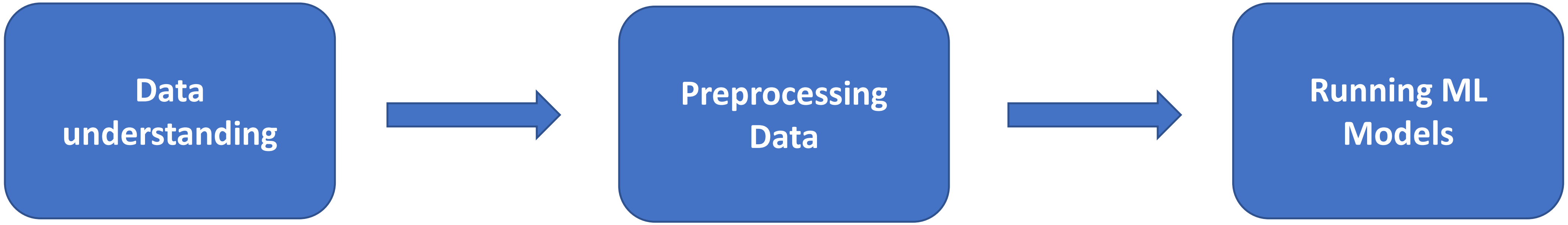
- The link to the portal is the following: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

- Our predictor or target variable will be 'SEVERITYCODE' because it measure the severity of an accident from 0 to 3 within the dataset.The variable can take the following values:

- * 3—Fatality
- * 2b—Serious Injury,
- * 2—Injury,
- * 1—Property Damage,
- * 0—unknown

Methodology

- This picture shows the resume of methodology involved in approaching this project:



Data understanding

Preprocessing Data

Running ML Models

- The upcoming slides will talk more in details about each of the steps involved

Data Understanding

- The Dataset contains 221.266 records and 40 columns.

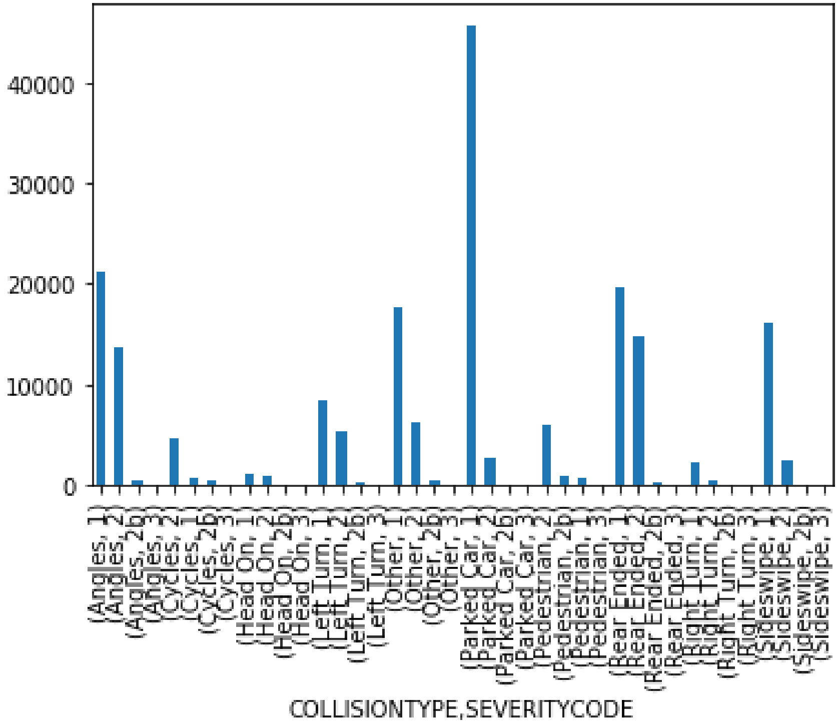
- The distribution of the records is the following:

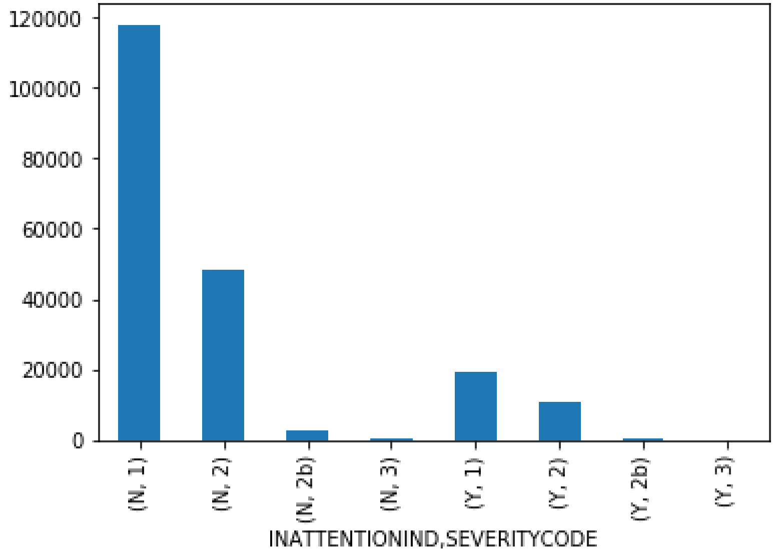
- Property Damage Only Collision 137485
- Injury Collision 58698,
- Unknown 21636,
- Serious Injury Collision 3098,
- Fatality Collision 349

- Finally, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

Data Preprocessing

After analyzing the data set, I have decided to focus in only five features: PERSONCOUNT (Number of people involved in the accident, VEHCOUNT (number of vehicles in the accident , SPEEDING (Whether or not speeding was a factor in the collision), INATTENTIONIND (Whether or not collision was due to inattention) and COLLISIONTYPE (Type of Collision). Here are some graphics of the data.





Data Preprocessing

- After performing the Data Understanding, and the preprocessing, the final set of data have the following records.

Severity Description	Number of Records
Property Damage Only Collision	58.698
Injury Collision	58.698
Serious Injury Collision	3.098
Fatality Collision	349

Running ML Algorithm

- With this in mind, the variables are used to create the machine learning models:
 Decision Tree, K-Neighbors, SVM and Logistic Regression K-Nearest Neighbor (KNN)

- KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

- Decision Tree A decision tree model gives us a layout of all possible outcomes so we can fully analyze the concequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

- Also I will implement SVM (Support Vector Machine) and Logistic Regression.

Results

- In the following table are the results of the evaluation tests to the models.

Algorithm	Jaccard	F1- score	LogLoss
KNN	0.649	0.638	N/A
Decision Tree	0.663	0.73	N/A
SVM	0.687	0.675	N/A
Log Regression	0.682	0.670	0.657

Discussion and Conclusion

- The Preprocessing of data was important, specially because some data was empty, or with an incorrect format. Also, with most categorical data that was of type 'object'. This is not a data type that we could have fed through an algorithm, so Transforming some of them with dummies (1 and 0) was used to created new classes that were of type int8; a numerical data type.

- Once we analyzed and cleaned the data, it was then fed through four Machine Learning models; K-Nearest Neighbors, Decision Tree Support Vector Machine (SVM) and Logistic Regression. Although the first two are ideal for this project, logistic regression and SVM helps because it was a classification model.

Discussion and Conclusion

- Based on the dataset provided for this capstone from type of Collision, number of people and cars involved into the accident, and some crucial factors like over speed and inattention of drivers we can conclude that particular conditions have a somewhat impact on whether or not travel could result in property damage (class 1), injury (class 2), or Class 2b and 3.

- It's an important thing check the accuracy of the models, the numbers are near of 70%, this means that the model can be improved, analyzing new variables, changing train and test dataset, etc.