Predicting the car accident severity

Rafael Mata M. - September 2020

Business understanding

- Nowadays the car is one of the most used medium to travel
- This phenomenon increases the number of cars in the streets and the possibilities of accidents
- Conditions like location, weather, road staus, speed, light and others can influence the accident
- The question or problem to be answered with this project is knowing certain conditions (weather, location, day, road status, etc) what is the severity when you have a car accident

Data understanding and cleaning

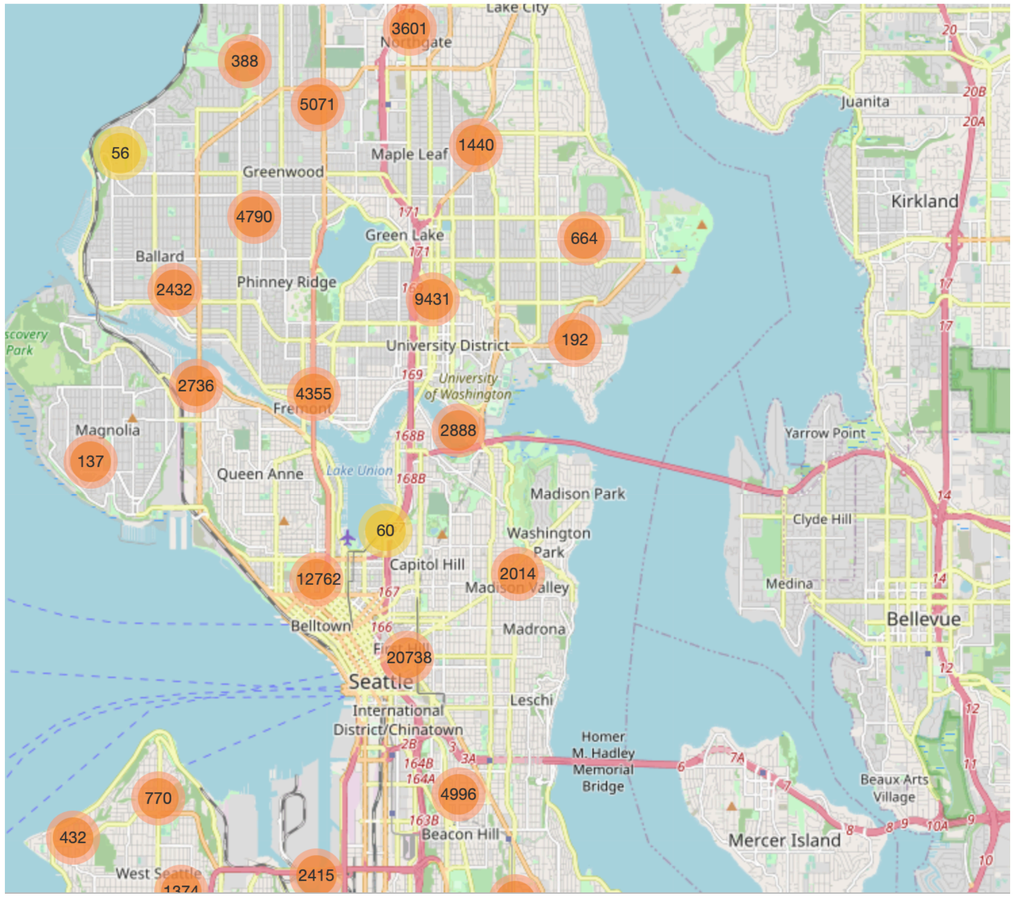
- The dataset is called **Collision -All Years** which has data for the Seattle city car accidents and is in .csv
- Contains **194673** observations and **38** features
- The target feature is the SEVERITYCODE
- The features with more than 50% Nan values were removed and also the features with unique values were removed
- Once cleaned the data, 11 features are used to modeling

Features engineering

- The target feature is unbalanced so the dataset is downsample to have a balance dataset and have a better modeling
- The observations with Nan values are removed
- The outliers are removed
- All features with test values are hot encoded and transformed to numeric values
- All the features are normalized

EDA - Accident locations

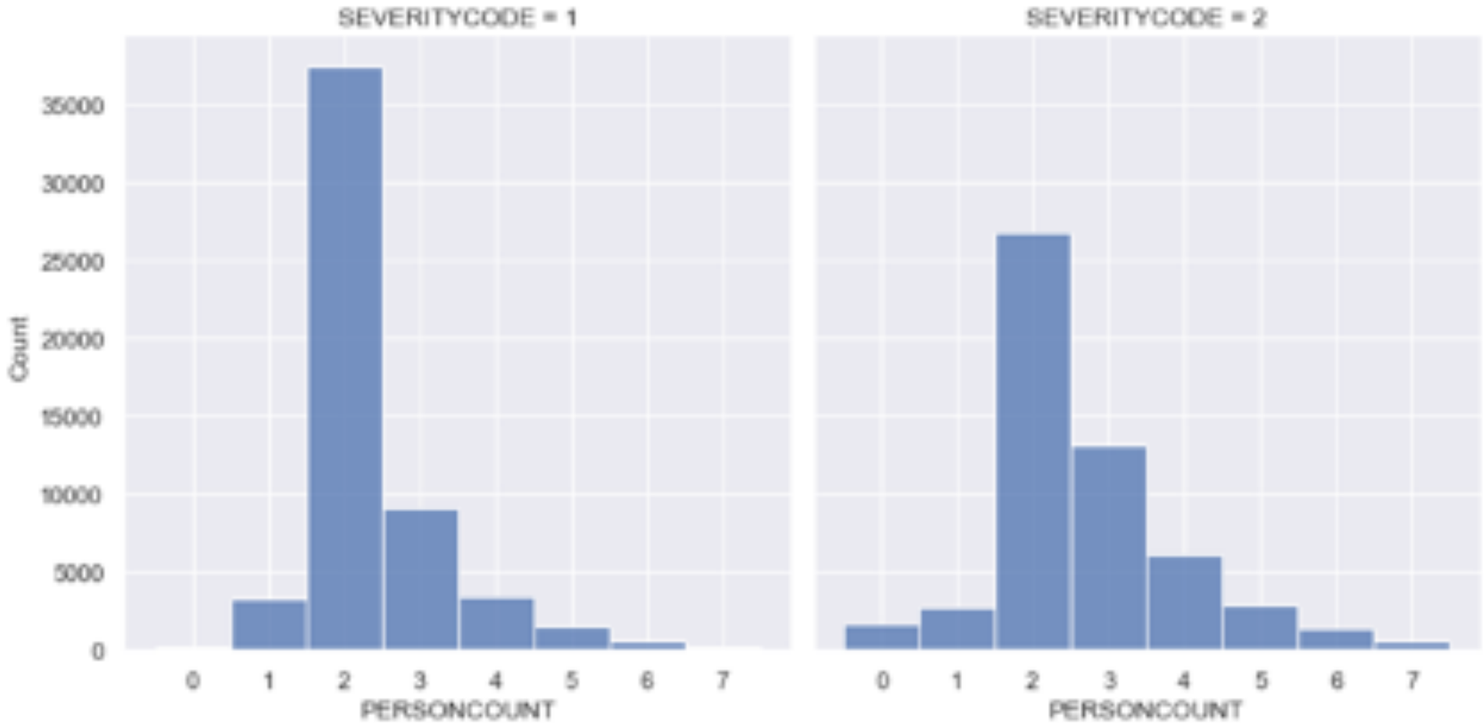
- The accidents are distributed in difference Seattle areas
- The severity is balanced between the different locations
- Only a few areas does not have accidents



EDA - Relation between people involved and the severity

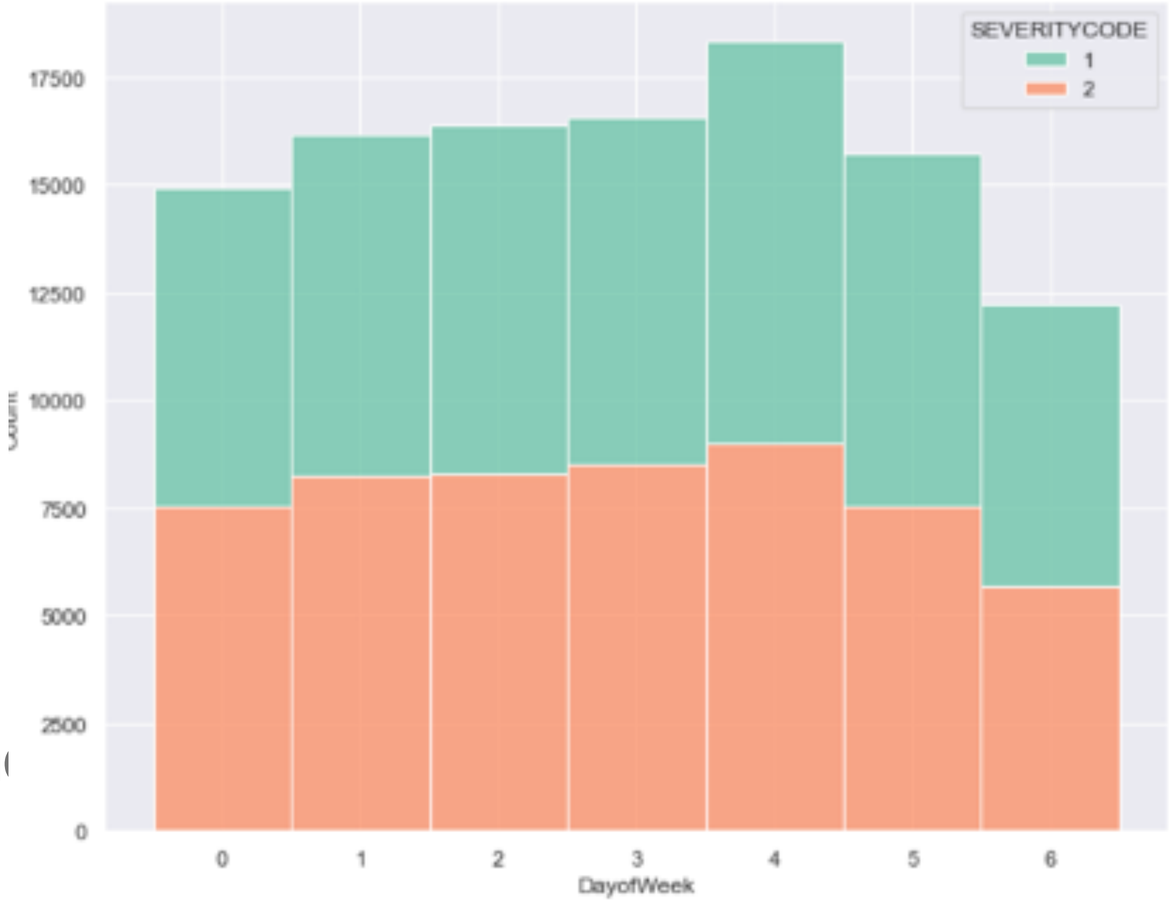
- Most of the accidents have 2 and 3 people involved in the accidents

- The severity is balanced between the different persons that participated in the accidents.



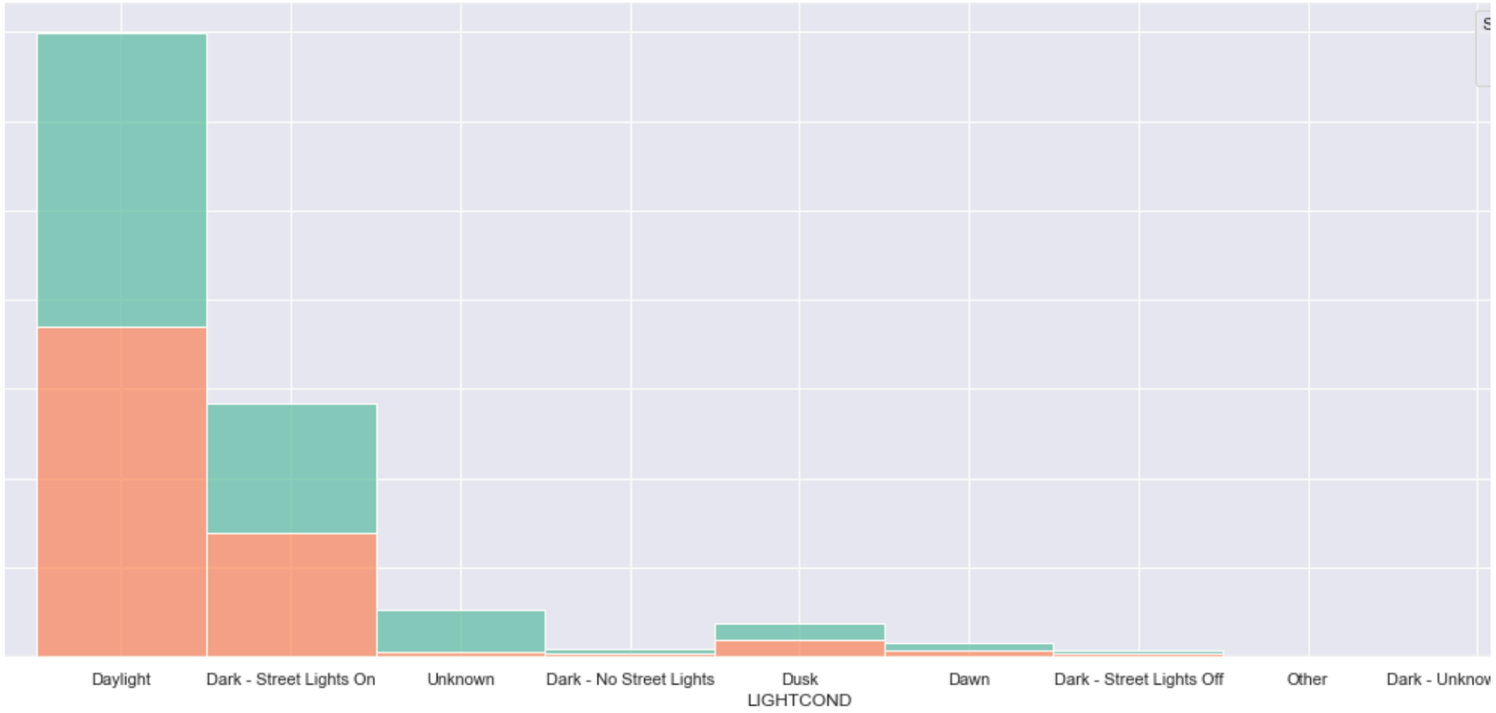
EDA - Relation between day of the week and accidents

- There is not big difference between monday thru thursday in the number of accidents
- Friday is the peak day of the week
- During weekends the number of accidents decreased
- The accident severity is balanced in th different days



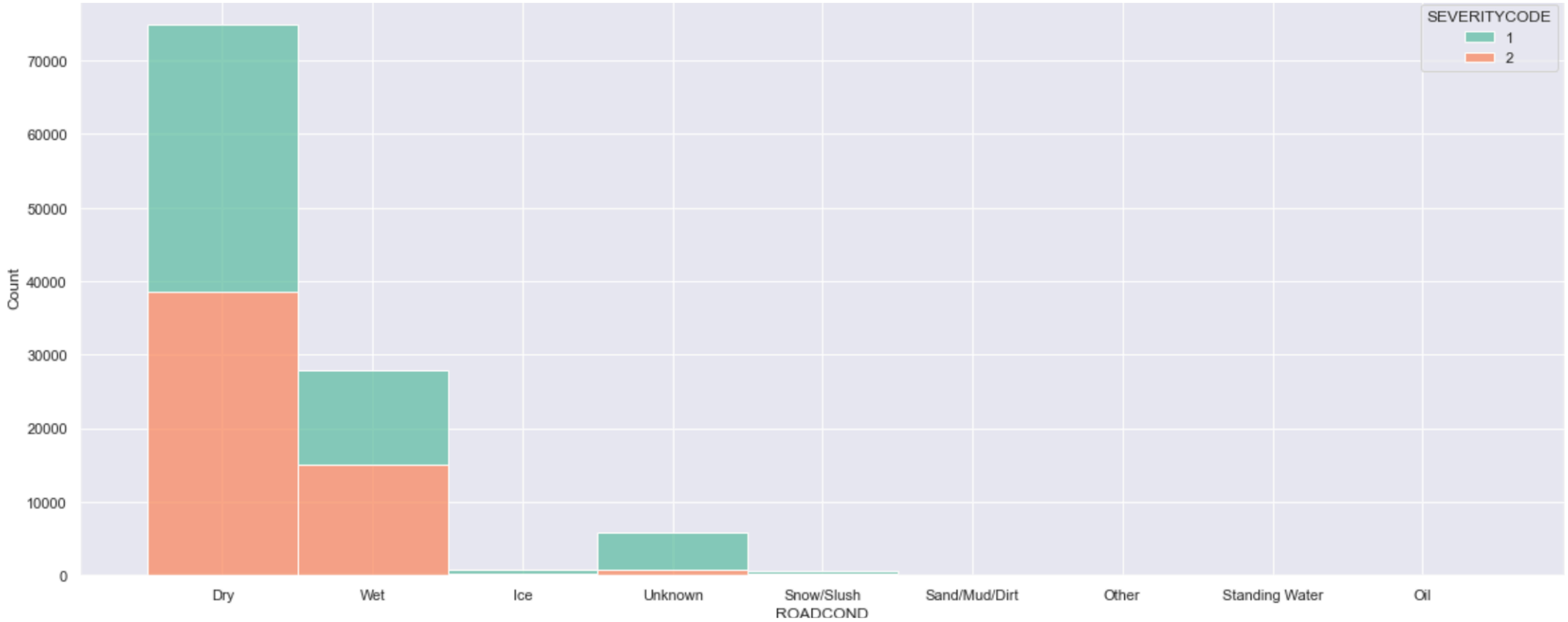
EDA- Relation between accidents and road condition

- Most of the accidents occurred during daylight or street lights on



EDA - Relation between road condition and accidents

- Most of the accidents occurred in dry conditions and the severity is balanced
- During wet conditions also occurred accidents but in minor quantity



Modeling

- Three classification machine learning algorithms are used to resolve the problem:
  - KNN - K nearest neighbor
  - SVM - Support Vector Machine
  - Logistic regression
- The dataset is split in training and testing to avoid overfitting
- Grid search is used for the three models to adjust the hyper parameters and get the best results

Modeling Performance

- Different metrics are used to evaluate the models performance and select the best model

Algorithm	MCC	RECALL	Precision	F1-score	Accuracy	LogLoss
KNN	42.9%	70.3%	72.0%	71.5%	71.5%	NA
SVM	43.9%	63.7%	75.9%	71.5%	71.7%	NA
LogisticRegression	29.4%	68.1%	63.7%	64.6%	64.6%	63.5%

- The model with the best prediction accuracy is the SVM

CONCLUSSIONS

- Due to the nature of the problem three classification machine learning methods were used to classify the car accident severity

- The SVM model has the best results, however a 75% accuracy could be considered low for this kind of problem

- As starting point this model can be used to analyze the accidents in the Seattle area and predict the severity to take preventive measurements

- Using another features that are not in the dataset like cellular phone usage during the accident can be used and improve the model accuracy