PREDICTING TRAFFIC ACCIDENT SEVERITY

Applied Data Science Capstone

github.com/Benider

Introduction

Traffic accidents are…

- Cause of 1.35 million deaths globally in 2016.
- Main cause of death among those aged 15–29 years.
- Predicted to become the 7th leading cause of death by 2030.

Predicting the accident severity in advance could save lives each year.

Road safety should be a prior interest for governments, local authorities and private companies

Data

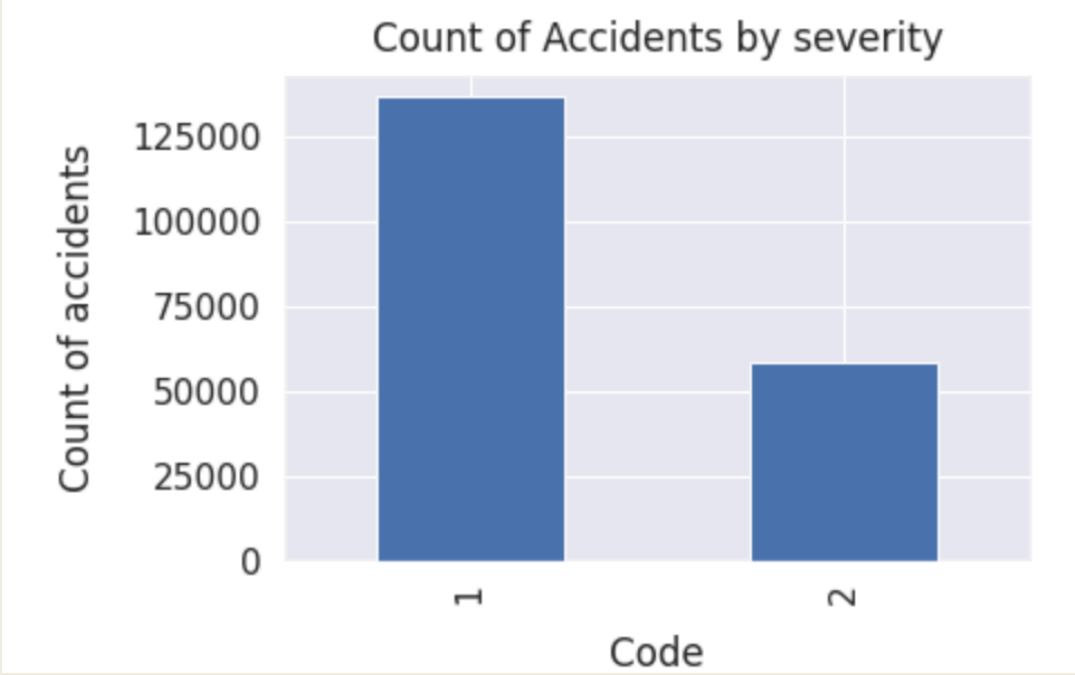
All the recorded accidents in SeatIle from 2013 to 2018, both years included.

- *Initial dataset from the CognitiveClass.*
- *Pre-selcted features on my GitHub report and notebook.*
- *In total 37 features, 194,672 rows*
- *Redundant and not relevant features were dropped*
- *5 features pre-selected*
- *On the data cleaning missing values and outliers were replaced.*

EDA

The target feature a binary classifier, describing the accident severity.

- *1: low severity.*
- *2: high severity, from hospitalized wounded injuries to death.*

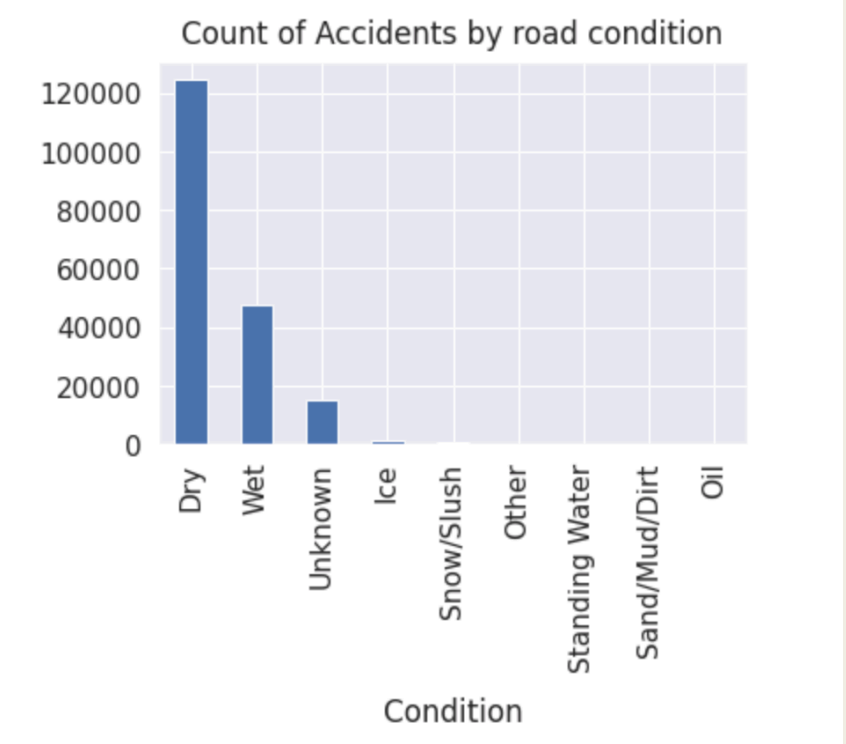


EDA

Road condition are either :

- *Dry*
- *Wet*
- *Unknown*
- *Ice*

Other conditions are negligeable.



Classification Models

- Random Forest
- Logistic Regression
- K-Nearest Neighbor
- XGBoost

Results

This table reports the results of the evaluation of each model.

	K Neighbors Classifier	Random Forest Classifier	logistic regression	XGBRegressor
f1 score	0.8198	0.8246	0.8243	0.8243
jaccard similarity score	0.6946	0.7015	0.7011	0.7011
Accuracy	0.6970	0.7019	0.7011	0.7011

Logistic regression is the better model since it is the fastest and more accurate!

Conclusion

- Built models to predict the severity of a traffic accident
- Still room for improvement for the Accuracy of the models