Coursera Data Science Project

Predict severity of vehicles accident in Seattle by building Machine Learning Model

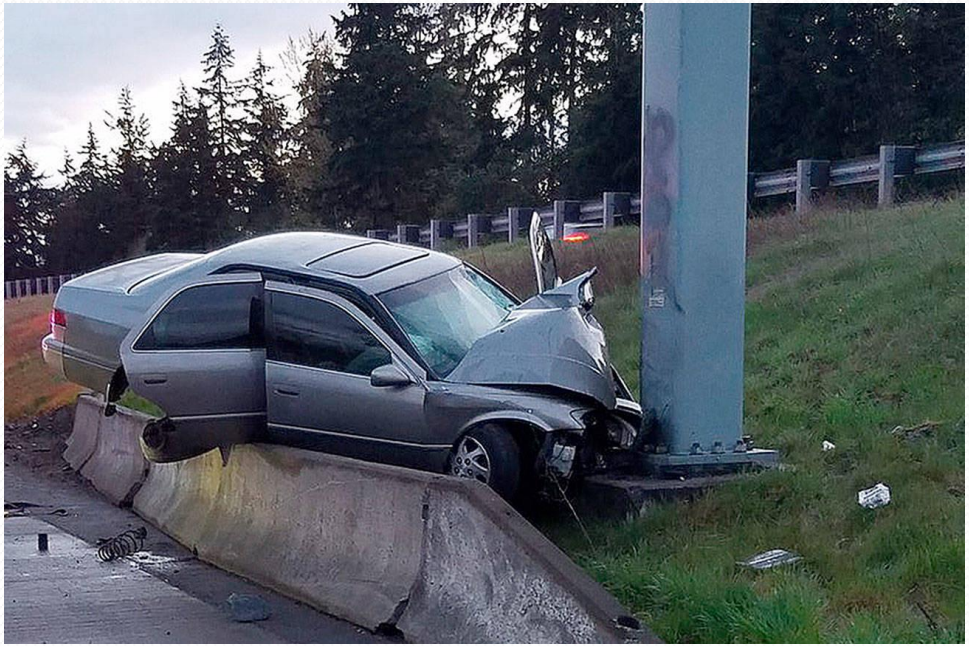
Irina Dzyu

Sep 2020

Introduction

Accident factors:

- vehicle condition
- road design
- environment
- weather
- human behavior
- other destruction objects



Problem description

- Seattle's status as the country's fastest-growing city — adding 19,901 people since the 2016 report — is a backdrop to the traffic rundown
- not a well-established mass-transit system in Seattle, most of the residents are drivers
- city is surrounded on three sides by water and is a fairly mountainous city. With all of its rain, hills, bridges, tunnels and narrow roads, this city is the perfect storm for car accidents

Data Analysis

Data source

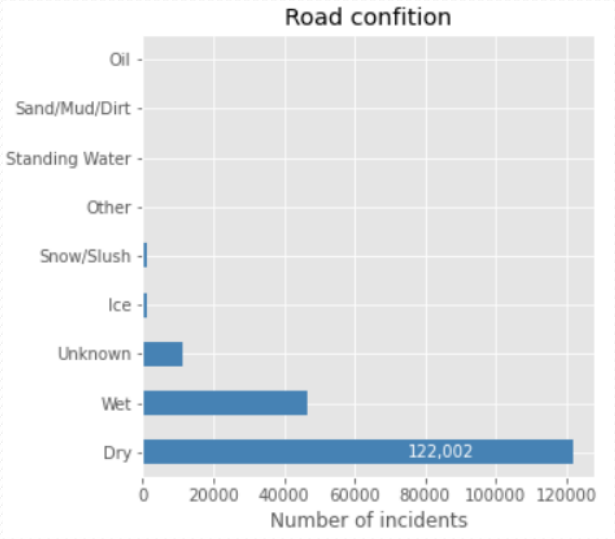
- CVS file (Data-Collisions.csv), SDOT Traffic Management Division, Traffic Records Group. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present
- Metadata form (Collisions All Year.pdf) contains a description of all features.

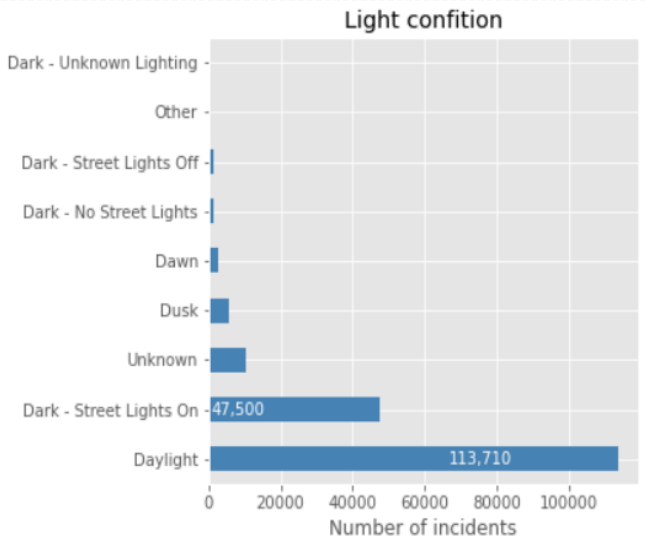
Data cleaning

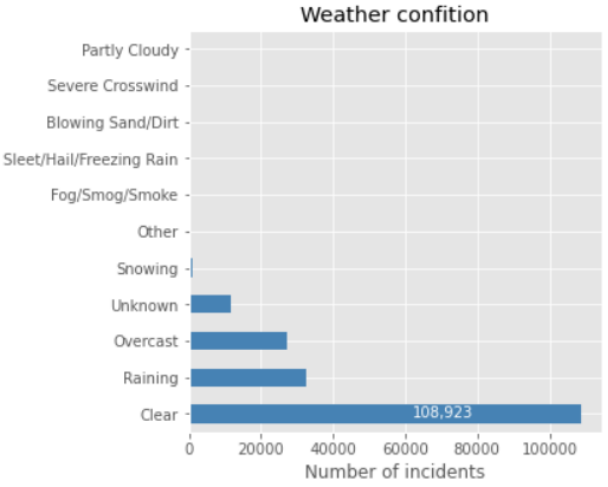
- Removal of Non-Relevant Features (initial dataset 34 features, dropped to 13 features)
- Removal N/A values (initial dataset 194,674 samples, dropped to 182,660)

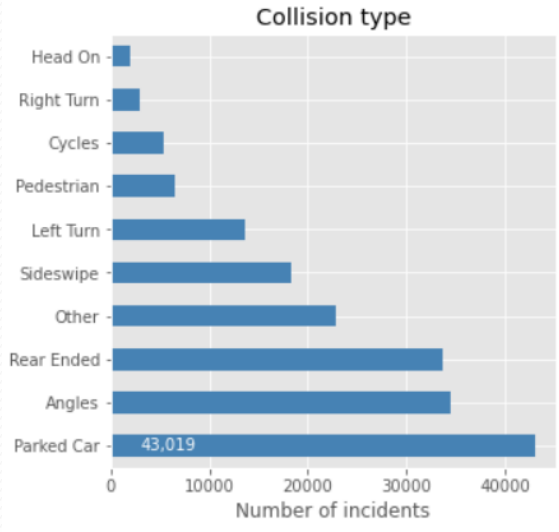
Data Analysis

Independent variables



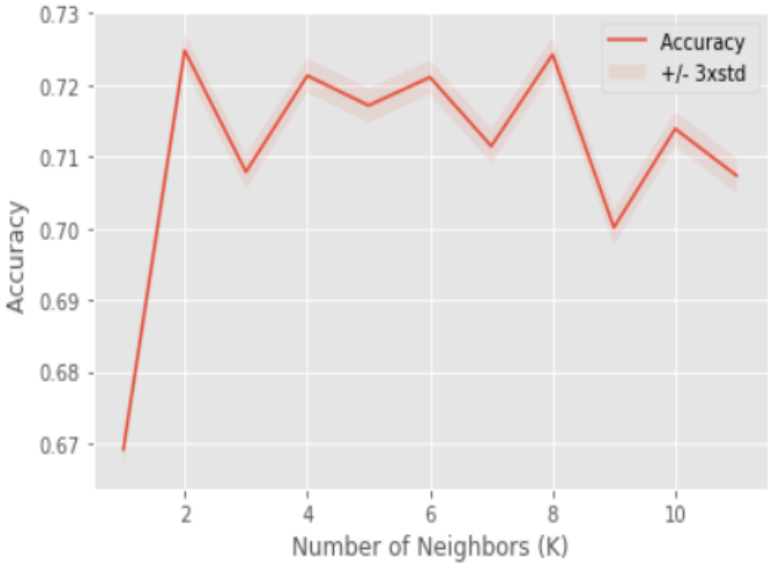






Predictive modeling (1/3)

- K Nearest Neighbor(KNN) - KNN will help predict the severity code of an outcome by finding the most similar to data point within k distance



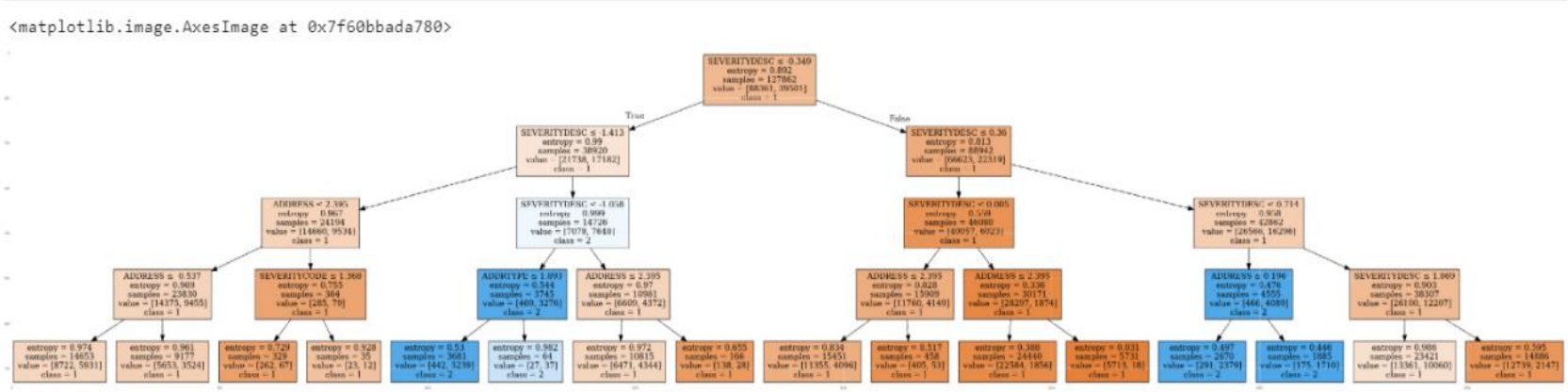
The best accuracy was with 0.7248165991459542 with k= 2

Model Evaluation

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.728939	0.665036	NA

Predictive modeling (2/3)

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



DecisionTrees's Accuracy: 0.7388955801306617

Model Evaluation

Algorithm	Jaccard	F1-score	LogLoss
Decision Tree	0.740616	0.675393	NA

Predictive modeling (3/3)

- Because our dataset only provides us with two severity code outcomes (SEVERITYCODE), our model will only predict one of those two classes (1-proper damage, 2-injury / 2b – serious injury and 3-fatality, not presented in our dataset population). This makes our data binary, which is perfect to use with logistic regression

[45]:
```
#Predict
yhat = LR.predict(X_test)
yhat_prob = LR.predict_proba(X_test)

#Evaluation
from sklearn.metrics import jaccard_similarity_score
jaccard_similarity_score(y_test,yhat)
```

[45]: 0.6873791014270594

Model Evaluation

Algorithm	Jaccard	F1-score	LogLoss
Logistic Regression	0.689352	0.566360	0.605631

Conclusion

- particular road, light, weather conditions during the collision have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).
- to predict possibility of car accident, KNN and Decision Tree classification models are ideal for this project
- logistic regression made most sense because of its binary nature.
- Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and log loss for logistic regression.
- Choosing different k, max depth and hypermeter C values helped to improve our accuracy to be the best possible.