Prediction Traffic Accident Severity

Capstone Project
by
Marius Stolz

October 2020

Introduction

- Background
  - Car collisions occur worldwide everyday
  - They lead to human fatality, injuries and property damage
  - Severity can be predicted using machine-learning
  - Input data could be weather, road, light conditions etc.
- Problem
  - Developing a prediction model to predict accident severity
  - Severity outcomes are ‚Injuries' and ‚Property damage'
- Interest
  - Street architecture
  - Navigation- and warning systems

Data acquisition and cleaning

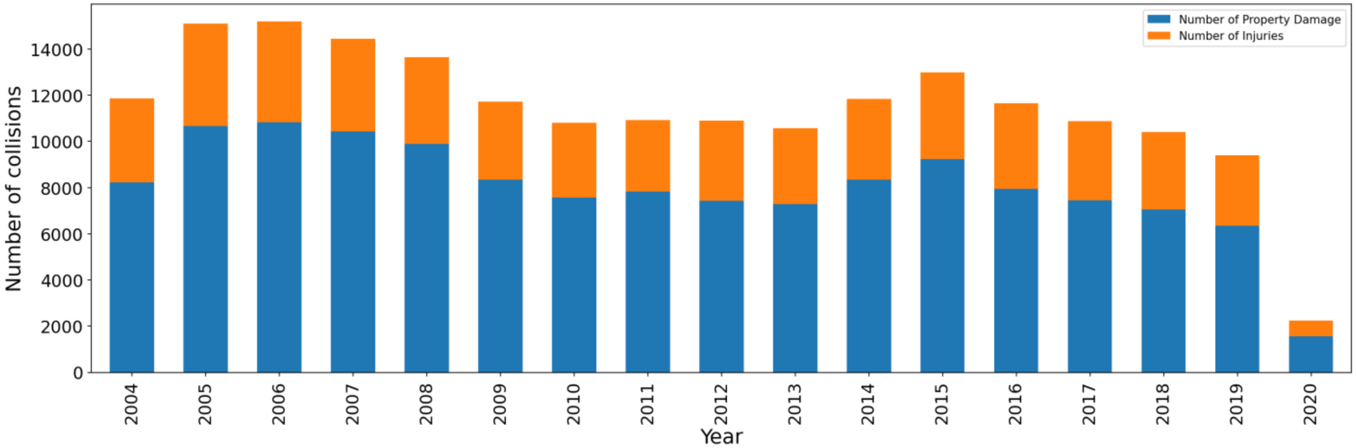
- Data source
  - Seattle Police Department,Traffic Records
  - .csv file
  - ~195000 collisions, described by 38 columns
  - Target label/column ‚SEVERITYCODE‘

- Data cleaning
  - Missing values
  - Redundant information
  - Structural errors

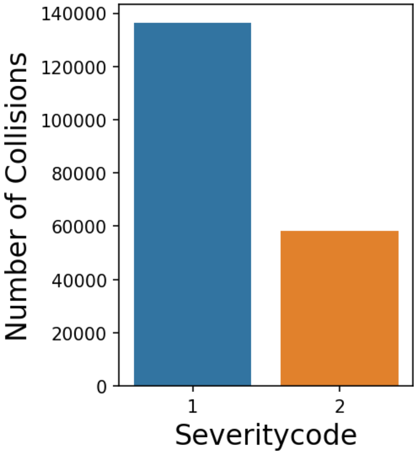
→ 14 columns remaining

Exploratory Data Analysis

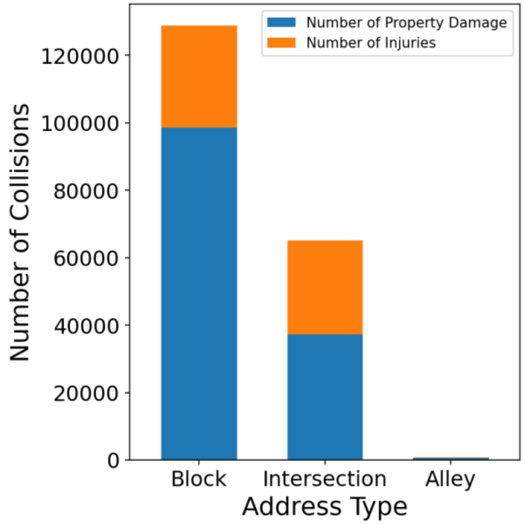
- Annual number of collisions: 2004 - 2020

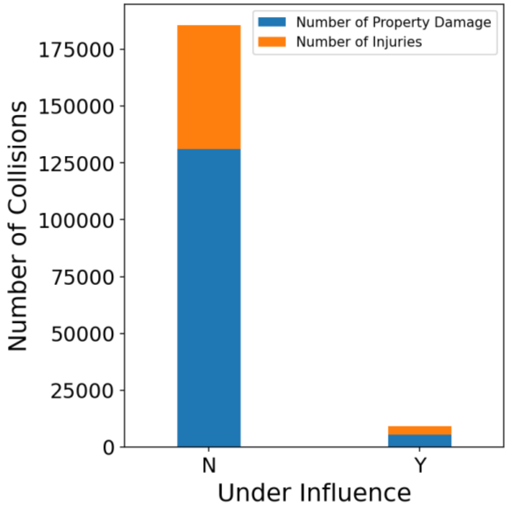


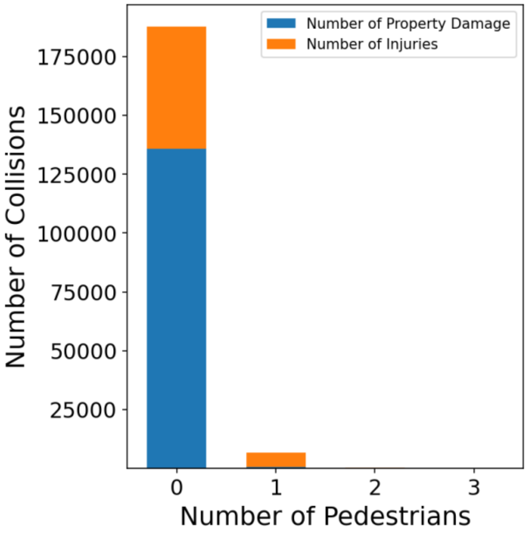
- Distribution of accident severity
  - 1 – Property damage
  - 2 – Injury

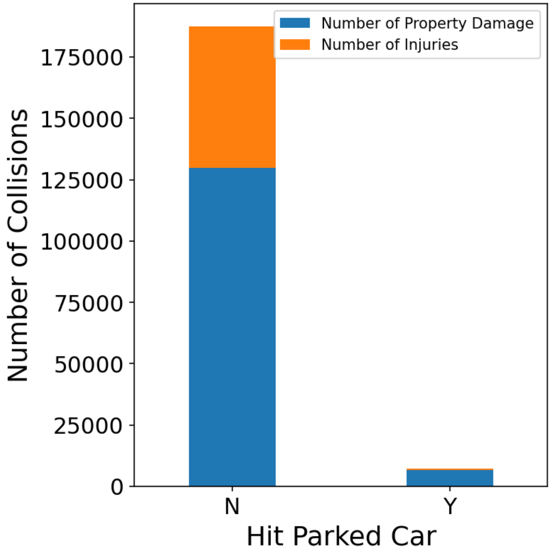


Exploratory Data Analysis

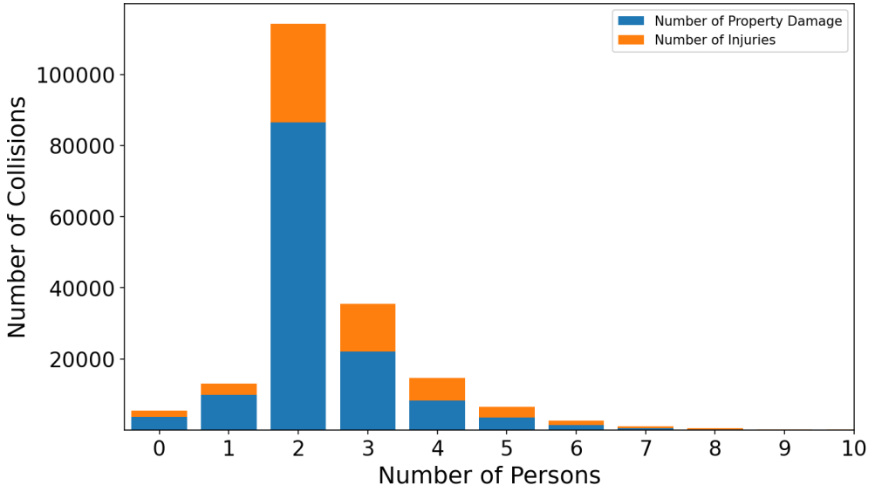


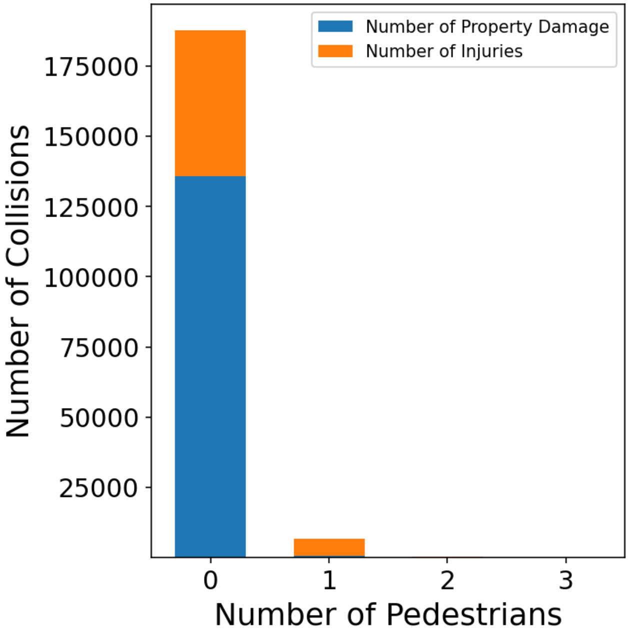


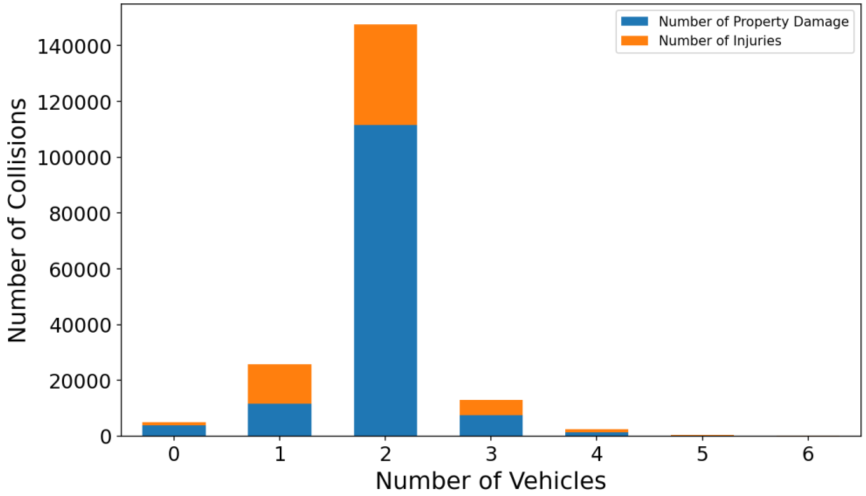




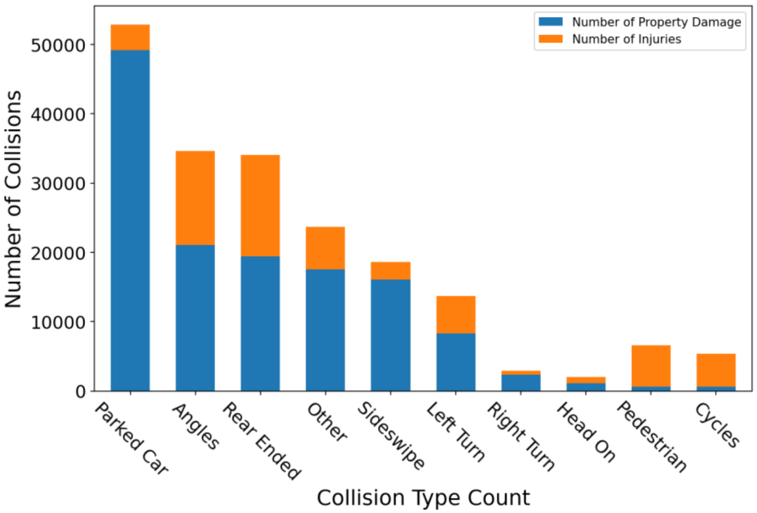
Exploratory Data Analysis

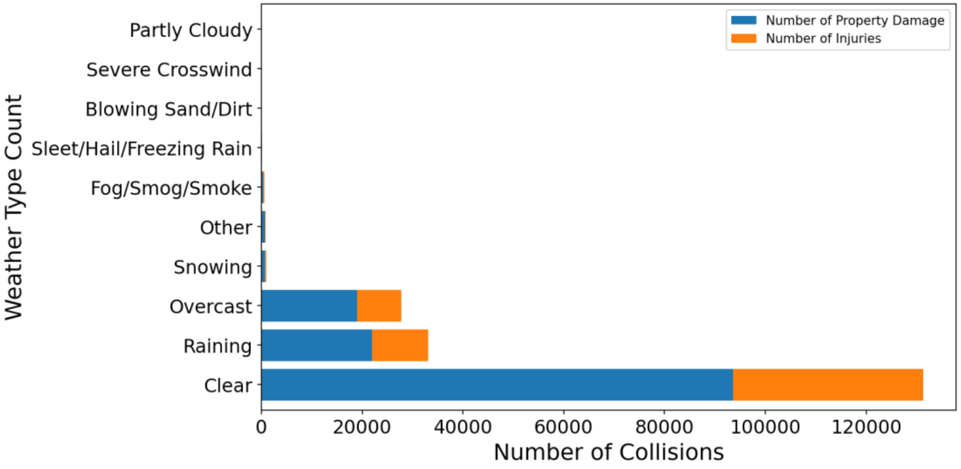


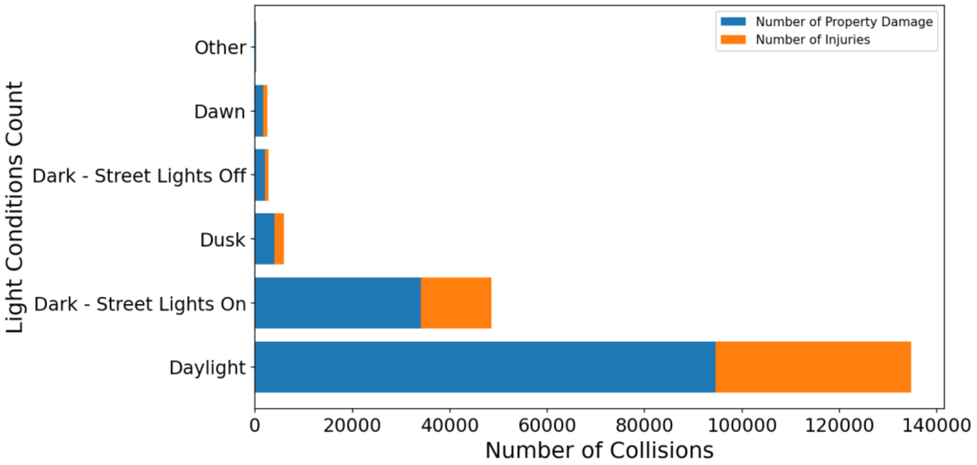




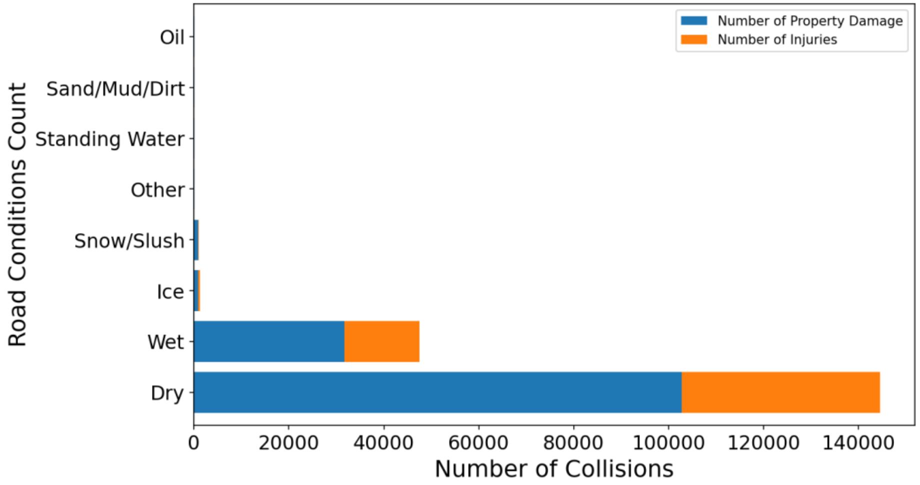
Exploratory Data Analysis

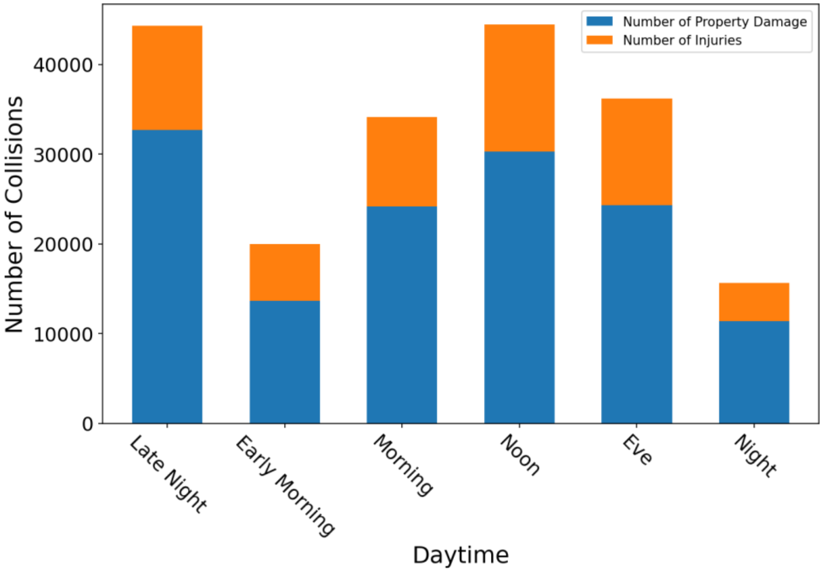


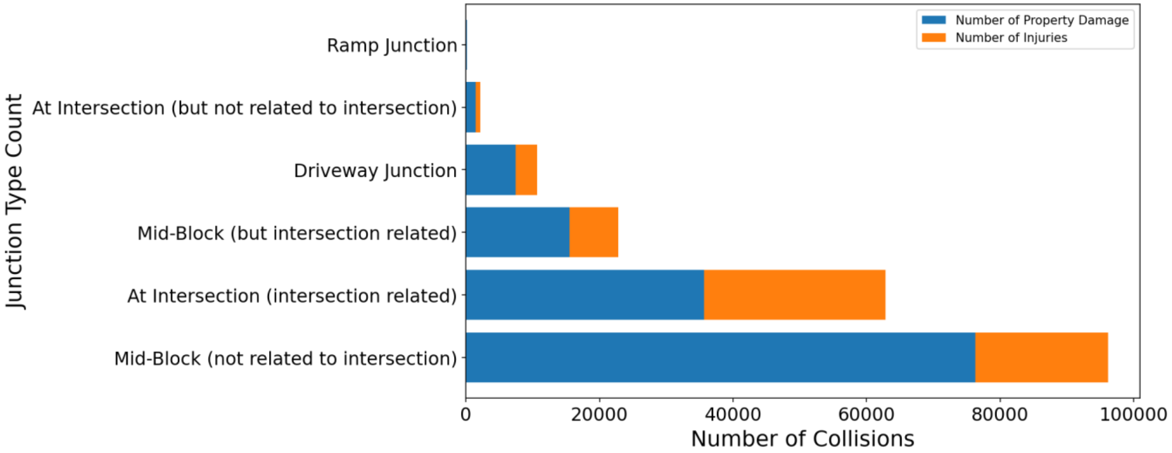




Exploratory Data Analysis

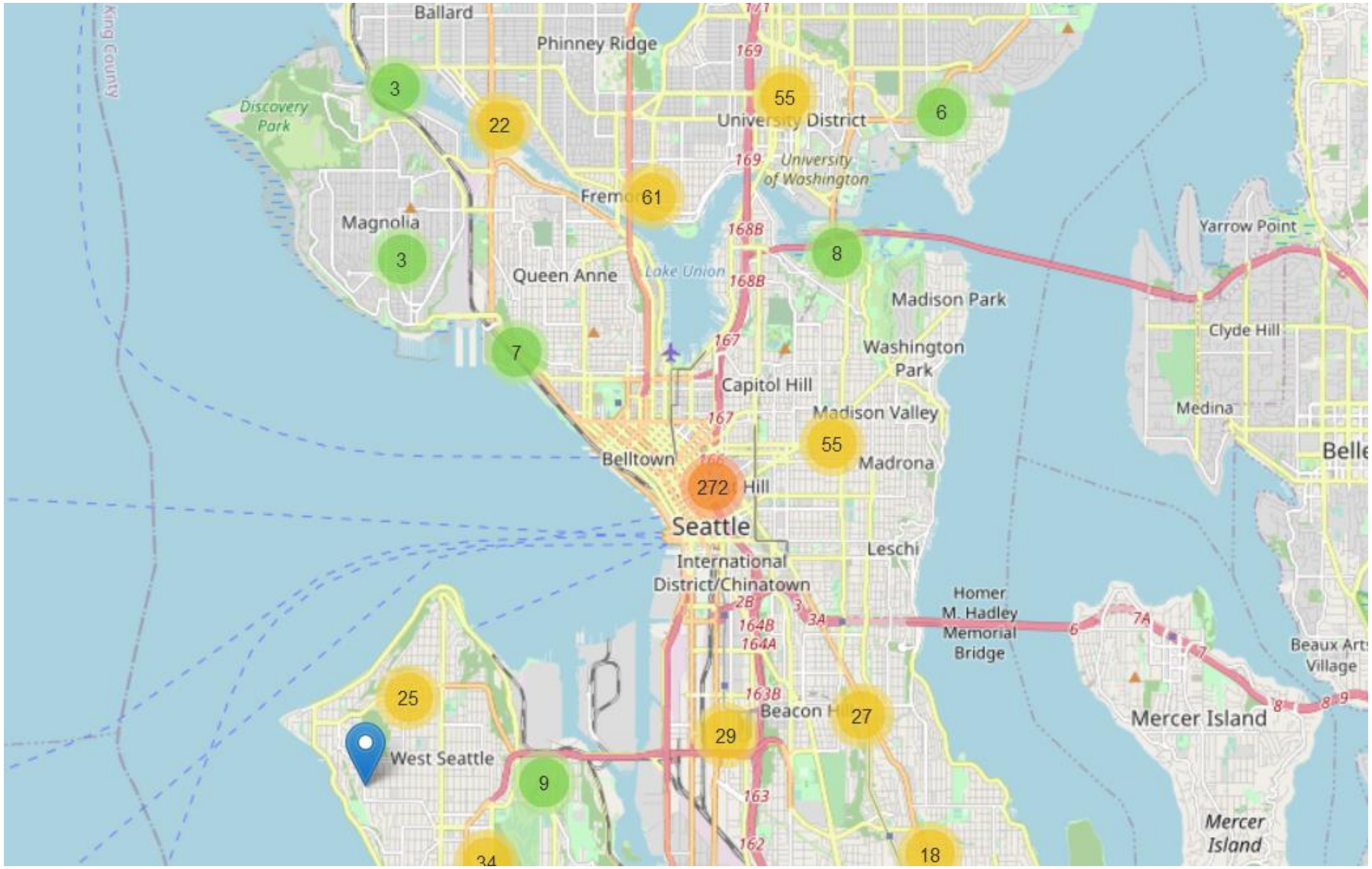






Exploratory Data Analysis

Majority of collisions
in the city centre



Exploratory Data Analysis

Summary

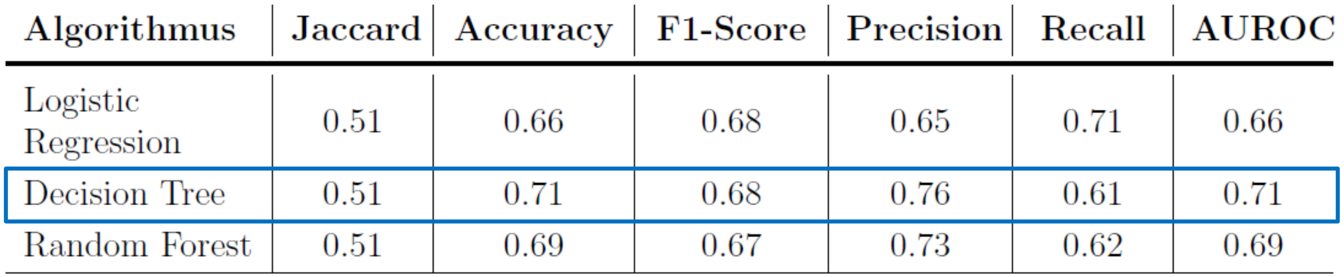
- Most collisions occured:
  - in a block or intersection
  - under no influence of drugs/alcohol
  - without pedestrians or bicycles
  - two persons and two vehicles
  - colliding at angles or rear ended
  - at daylight
  - at clear weather
  - on dry road
  - at late night or noon
  - in the city centre
  - in a mid block

Model Development

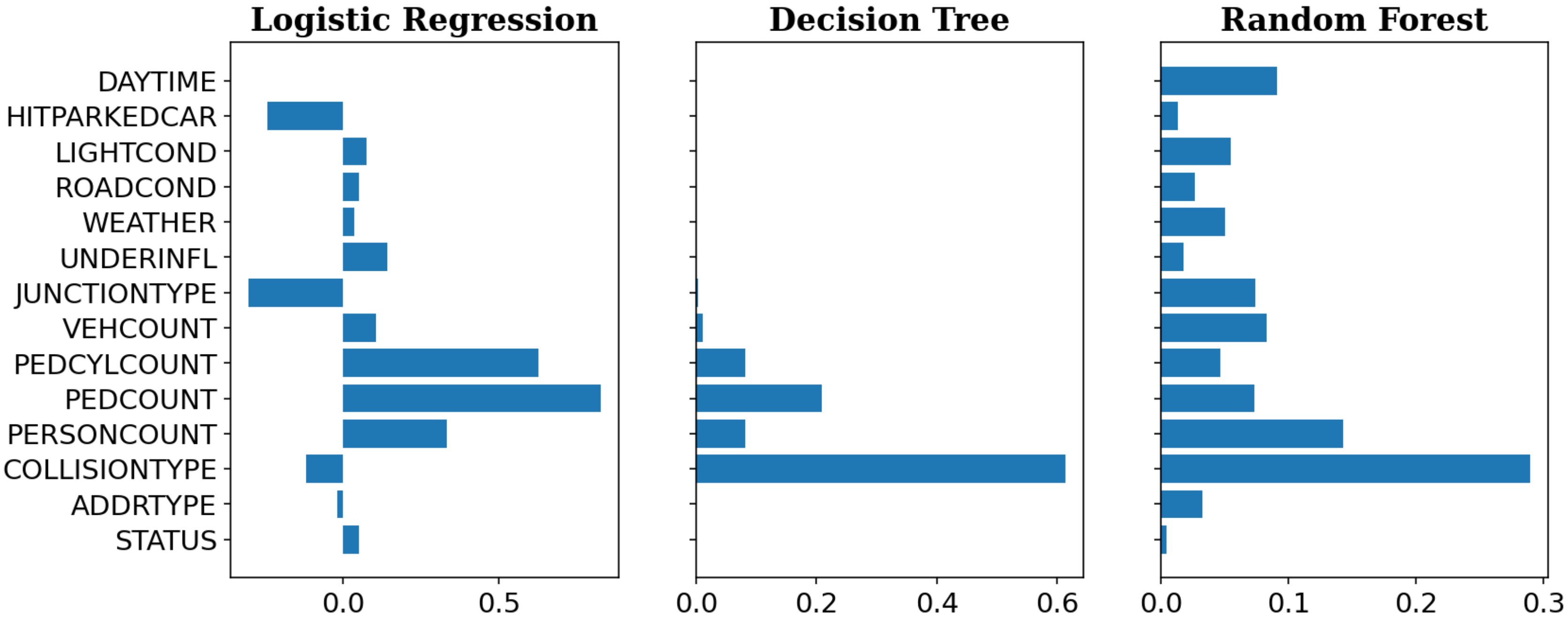
- Data preparation
  - Balanced data by down-sampling target label
  - Converted categorical to numerical data using LabelEncoder()
  - Normalization with StandardScaler()
  - Splitting data set into 70 % training and 30 % test data

- Machine-learning models
  - Logistic Regression
  - Decision Tree
  - Random Forest

Results



Algorithmus	Jaccard	Accuracy	F1-Score	Precision	Recall	AUROC
Logistic Regression	0.51	0.66	0.68	0.65	0.71	0.66
Decision Tree	0.51	0.71	0.68	0.76	0.61	0.71
Random Forest	0.51	0.69	0.67	0.73	0.62	0.69



Discussion & Conclusion

- Decision Tree is best model
- 71 % accuracy is not satisfying
- Dependencies and influences on accident severity not found

- Further training of prediction model is needed
- Data collection should be checked due high amount of missing values