Predicting the Severity of Accidents

Data Science Capstone Project
IBM/Coursera

Stefanie Welcker

Introduction

- Year-round accidents happen especially in and around big cities like Seattle causing traffic jams, damage and sometimes severe injuries or deaths
- Can historical data be used to predict the severity of accidents?
- Knowledge of severity could be beneficial to rescue stations to plan rescue forces, insurances to calculate future costs, radio stations to better inform their listeners and have therefore potentially more income through ads

Data

- **Data Source**:
  - Seattle provides data on collisions and a detailed description (1)
  - Data contains records starting 01.01.2004 and is updates weekly (2)

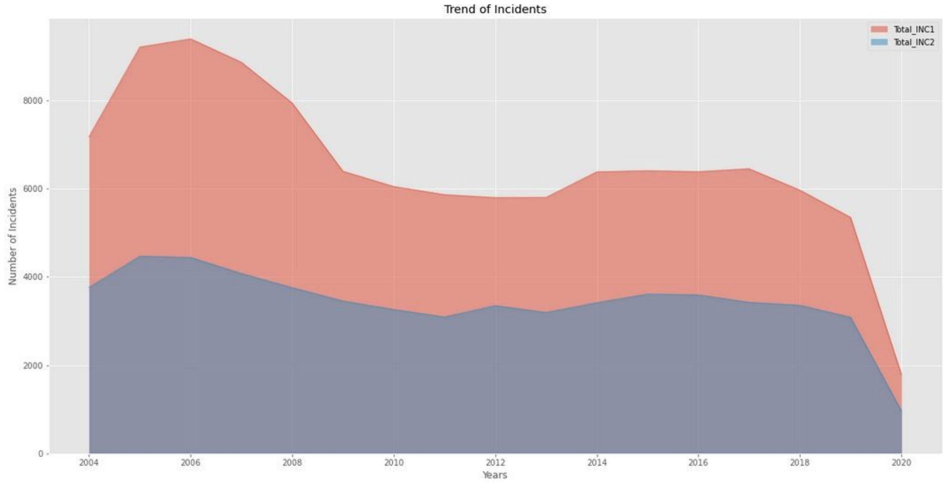
(1) https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data

(2) https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

- **Data Cleaning and Preparation**:
  - Data is reclassified in two categories: 1= minor accidents (propriety damage); 2= severe accidents with injured persons; most are minor accidents
  - Rows with unknown features are deleted because no information can be gained (most of them minor or unknown severity)
  - The cleaned data is still imbalanced: 65.6% minor and 34.4% severe accidents

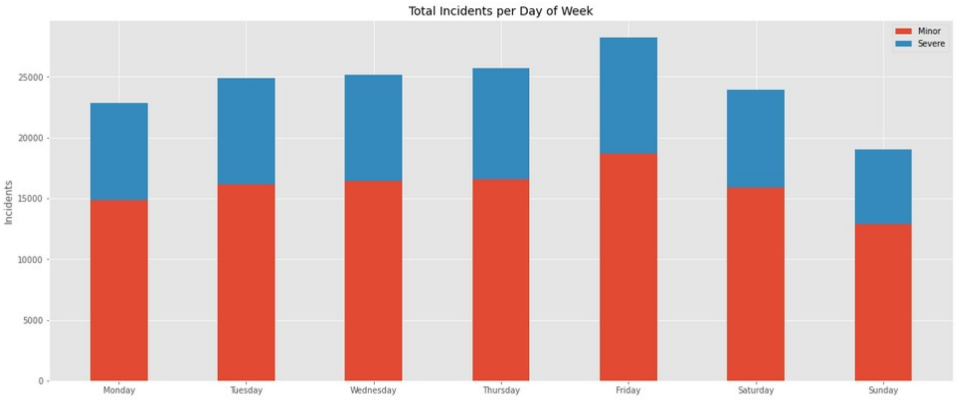
Data Exploration

- Trend of Years

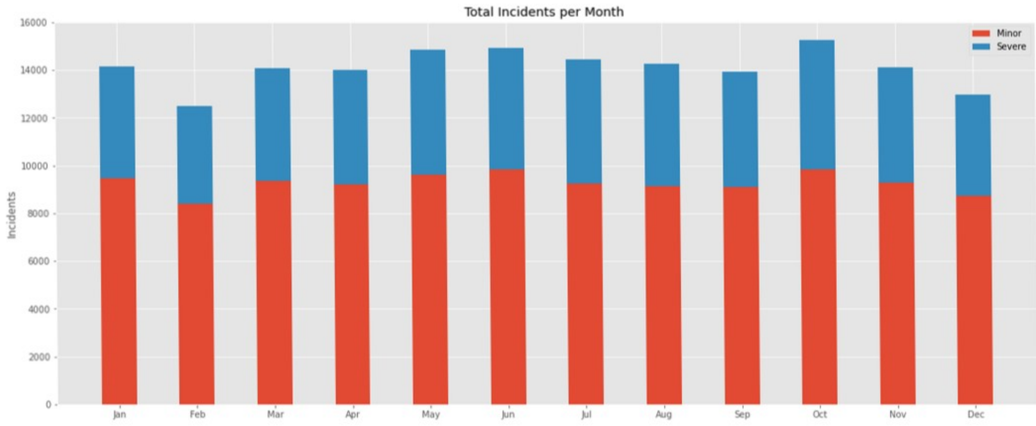


- No dependency on Day of Week or Month could be detected
- The proportion of minor accidents shows the same ratio like the complete data set
- Trend of Years shows an increased amount of minor accidents in years 2005 to 2008

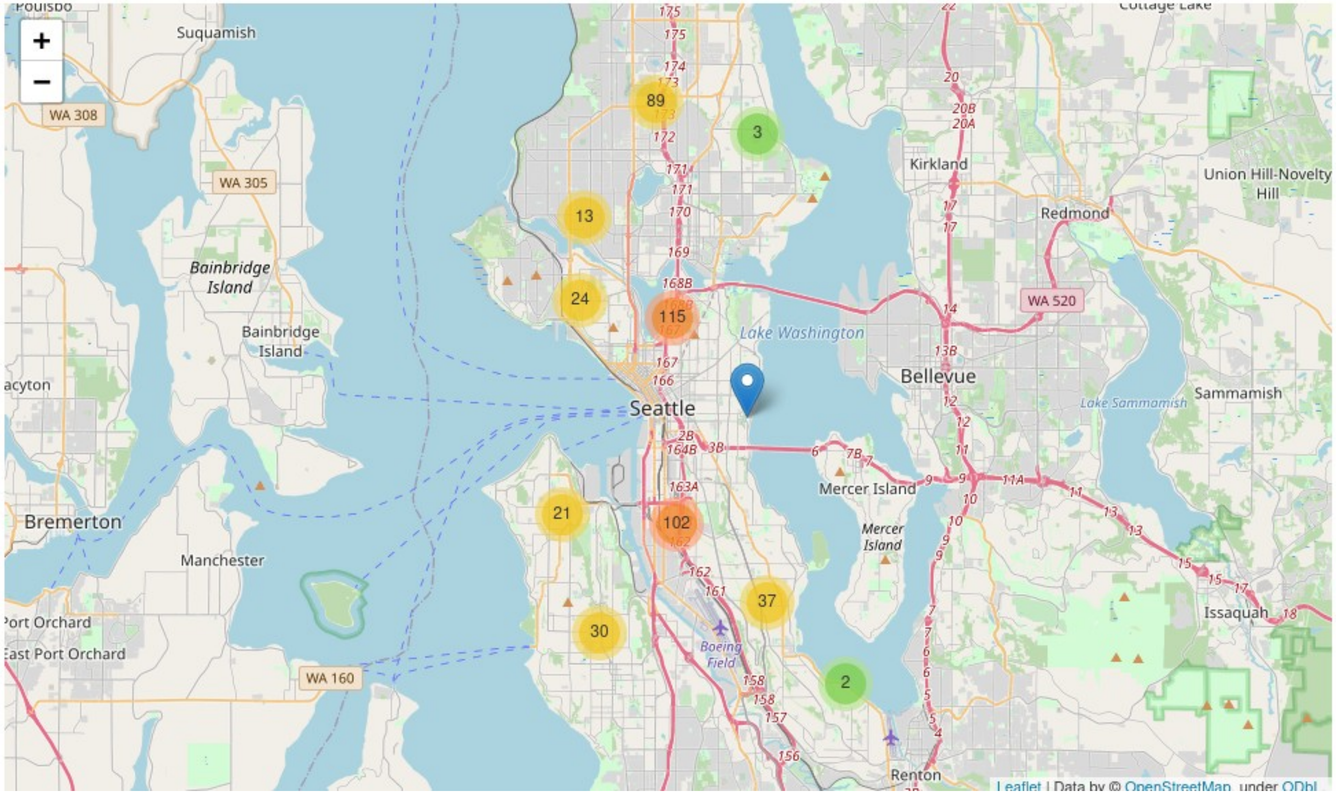
Day of Week



- Month



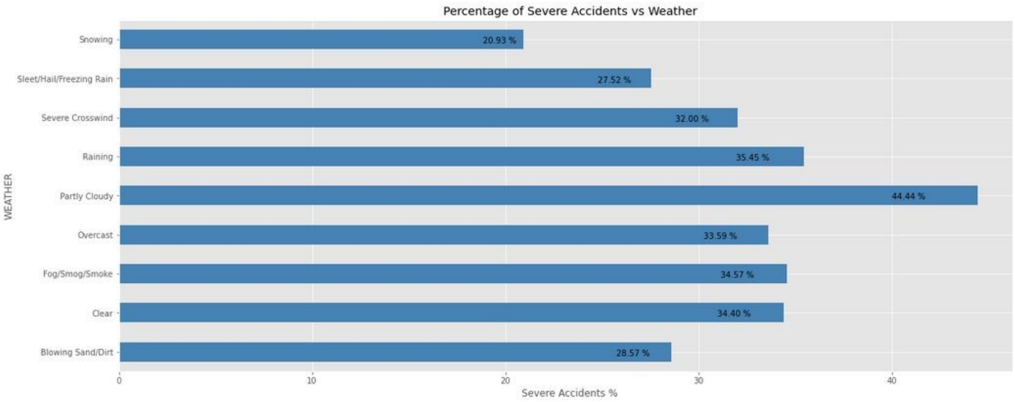
Data Exploration



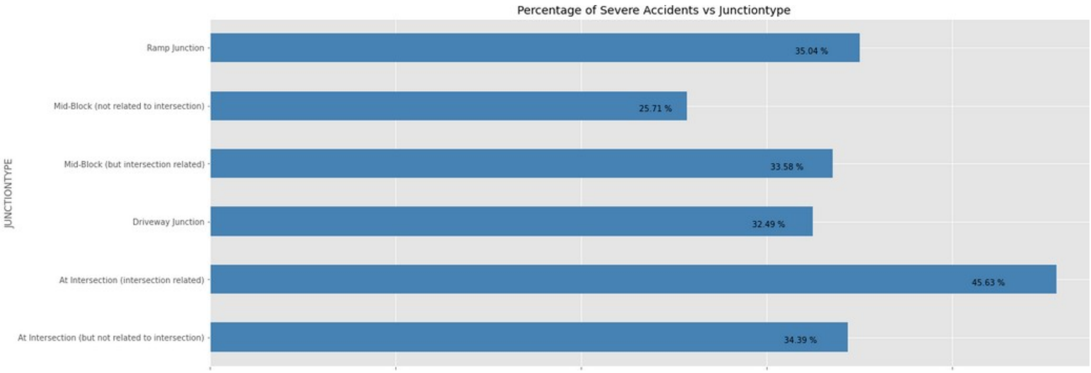
- Accidents from 2020/06/01 until 2020/07/29
- Location may influence the outcome of accidents since there are clusters of higher density

Data Exploration

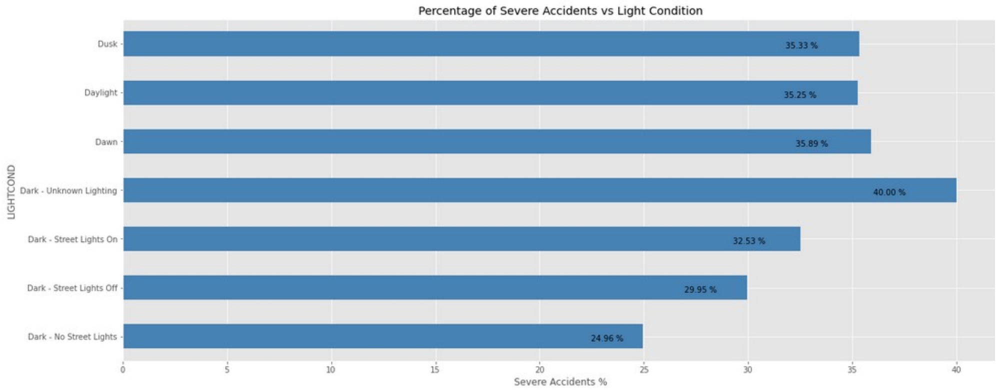
- Weather Condition



- Junction Type



- Light Condition

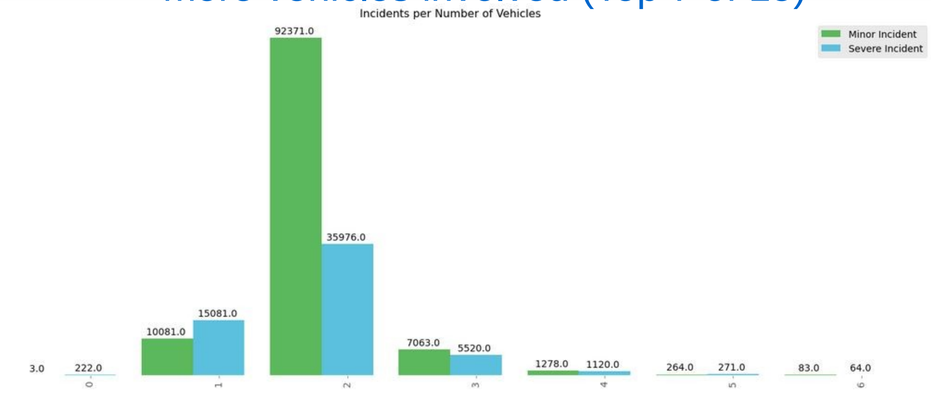


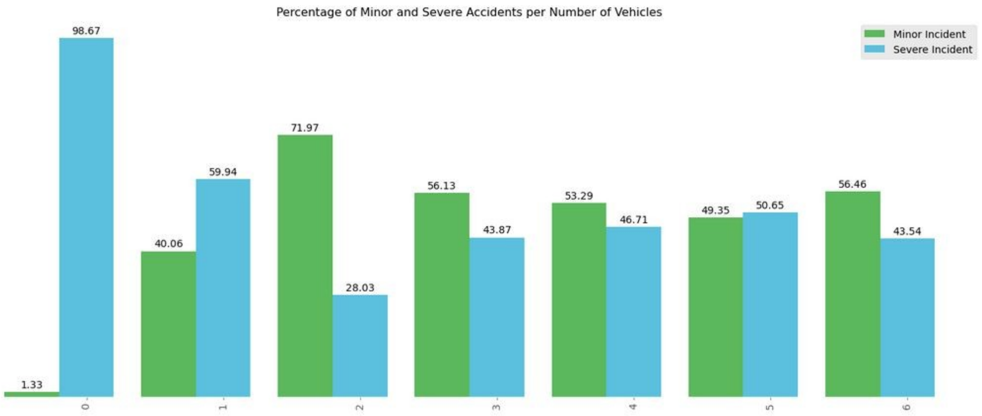
- The severity of accidents depend on external factors like light condition or type of junction
- Main influences are
  - partly cloudy sky
  - darkness
  - intersections or related areas

Data Exploration

Vehicles:

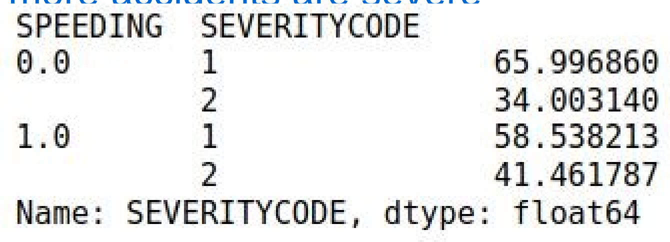
- Relatively more severe accidents with more vehicles involved (Top 7 of 15)





Speeding:

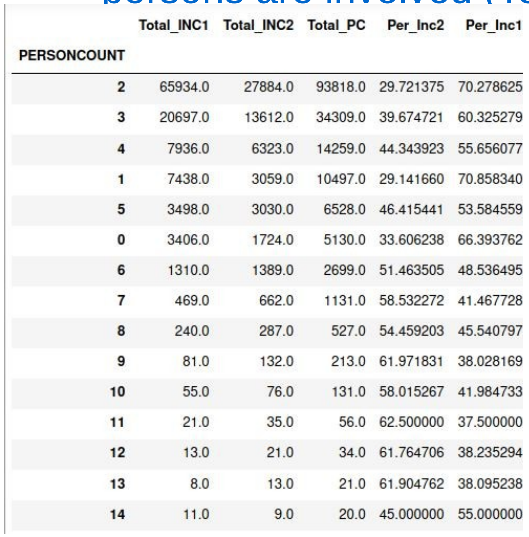
- With speeding involved, about 7.5 % more accidents are severe



```
SPEEDING  SEVERITYCODE
0.0       1               65.996860
          2               34.003140
1.0       1               58.538213
          2               41.461787
Name: SEVERITYCODE, dtype: float64
```

Persons:

- % of severe accidents is high when less persons are involved (Top 15 of 93)



	Total_INC1	Total_INC2	Total_PC	Per_Inc2	Per_Inc1
PERSONCOUNT					
2	65934.0	27884.0	93818.0	29.721375	70.278625
3	20697.0	13612.0	34309.0	39.674721	60.325279
4	7936.0	6323.0	14259.0	44.343923	55.656077
1	7438.0	3059.0	10497.0	29.141660	70.858340
5	3498.0	3030.0	6528.0	46.415441	53.584559
0	3406.0	1724.0	5130.0	33.606238	66.393762
6	1310.0	1389.0	2699.0	51.463505	48.536495
7	469.0	662.0	1131.0	58.532272	41.467728
8	240.0	287.0	527.0	54.459203	45.540797
9	81.0	132.0	213.0	61.971831	38.028169
10	55.0	76.0	131.0	58.015267	41.984733
11	21.0	35.0	56.0	62.500000	37.500000
12	13.0	21.0	34.0	61.764706	38.235294
13	8.0	13.0	21.0	61.904762	38.095238
14	11.0	9.0	20.0	45.000000	55.000000

Modeling and Results

Tested Models

- Decision Tree
- Logistic Regression
- Random Forest Classifier

Dealing with imbalanced data:

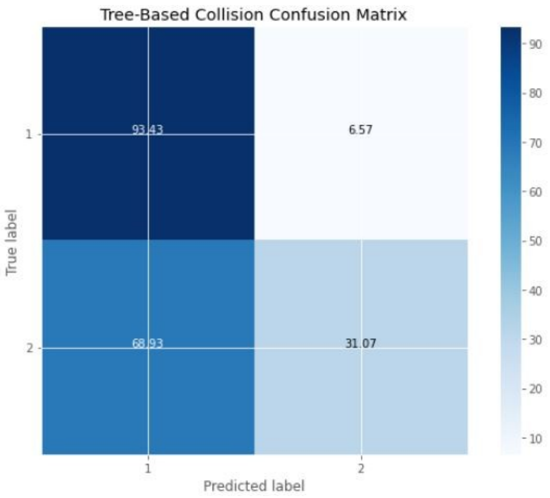
- SMOTE + **R**andom **F**orest **C**lassifier
- SMOTE + **L**ogistic **R**egression
- **B**alanced **R**andom **F**orest **C**lassifier

Results

Model	Jaccard	F1-Score	LogLoss
• Decision Tree	0.5474478619	0.68459020752	NA
• Logistic Regression	0.5089967867	0.6424616253	0.6085578323
• Random Forest Classifier	0.5174818500	0.66963022803	NA
• SMOTE + RFC	0.4897052853	0.6501701188	NA
• SMOTE + LR	0.4986759004	0.63312561455	0.611919708461
• BRFC	0.4550778172	0.6215009477	NA

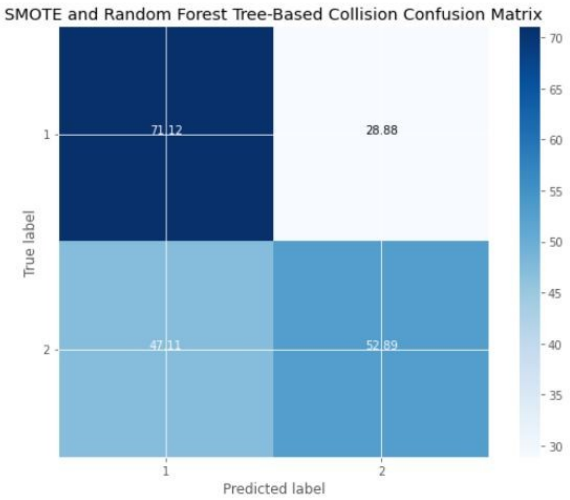
SMOTE (Synthetic Minority Over-sampling Technique)

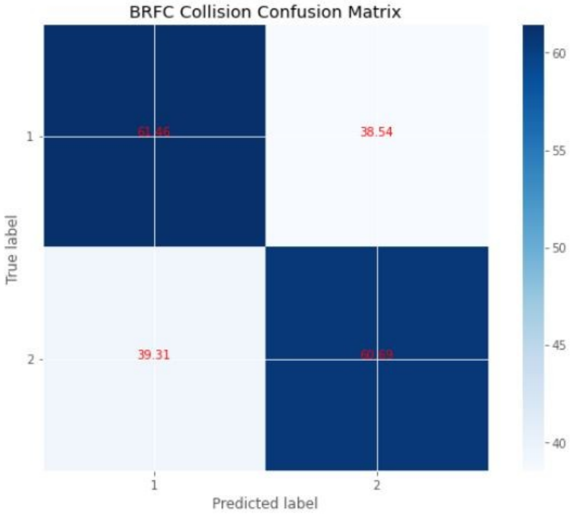
Discussion



Regarding Jaccard and F1-score the Decision Tree performed best followed by Random Forest Classifier

- The “severe” prediction for accidents is still not good with Decision Tree (31.07%)
- Using SMOTE (Synthetic Minority Over-sampling Technique) enhances the “severe” prediction but only for Random Forest Classifier (52.89%)
- BRFC predicts severe outcome best with 60.69% correct.





Conclusion

- Severity of accidents can be predicted by using machine learning algorithms
- You need to know many features (number of persons, number of vehicles, type of junction) that lead to an accident beforehand to make the prediction, so using the prediction as forecast like in live traffic news is not likely
- Forecast calculations can be done when the features are already known, but not the outcome, insurances may quickly calculate costs
- The time of day when accidents occur may be an interesting aspect for future study