Robin Kjär
25.10.2020
github.com/RobinKj/Coursera_Capstone

PREDICTING TRAFFIC ACCIDENT SEVERITY IN SEATTLE, WA

IBM Data Science Professional Certificate
Capstone Project

Road accidents are a growing issue worldwide, with millions of deaths and injuries every year all over the world

- Road traffic injuries result in **1.35 million deaths** on roadways each year globally
- This effects people across all age groups with **more than half** of the victims being pedestrians, motorcyclists and cyclists

Seattle is the **25th biggest city in the U.S.** with almost 3,8 million people living in its metropolitan area

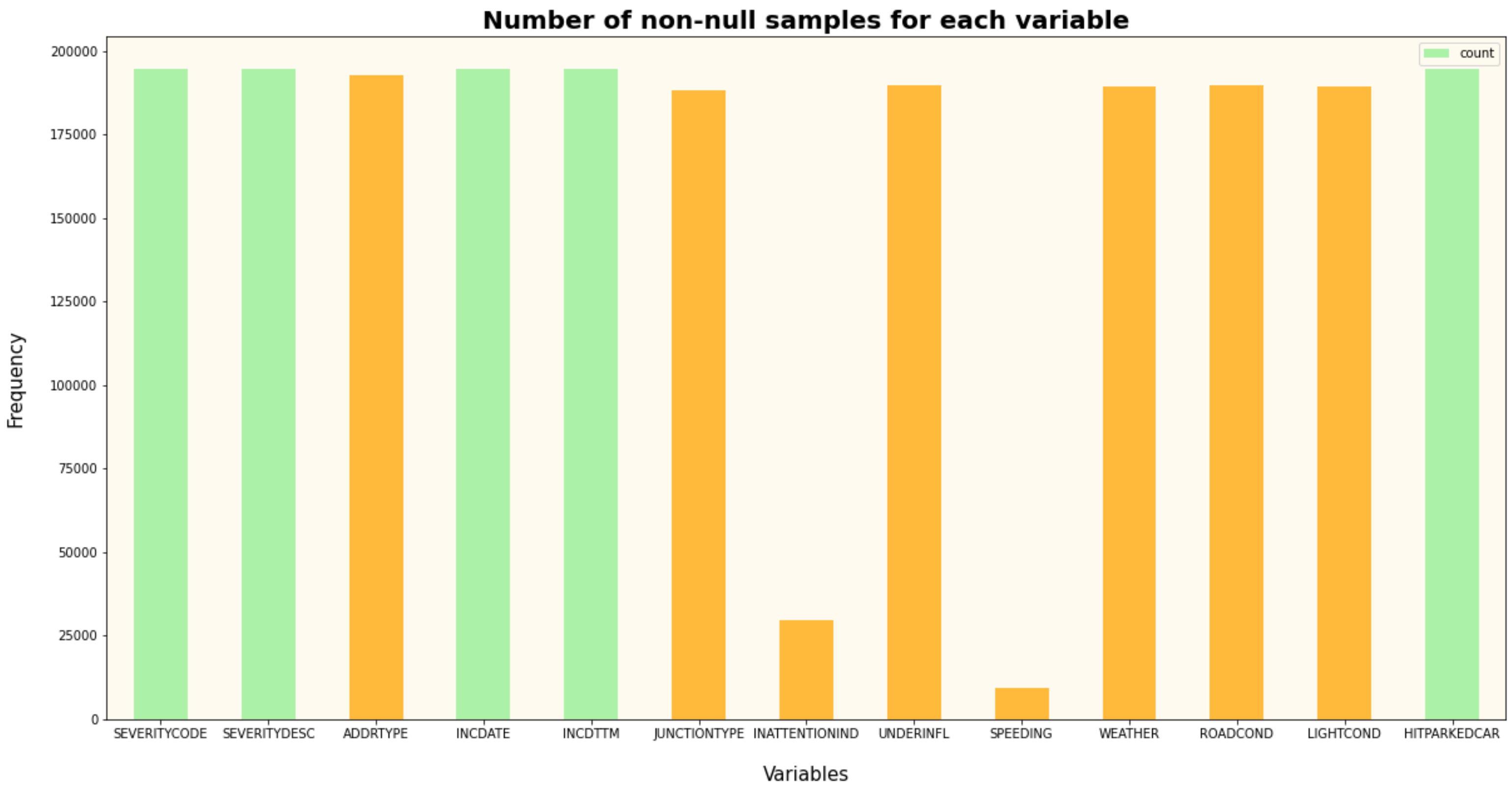
- A report from 2019 **ranks Seattle seventh** in the whole United States **in time stuck in traffic**
- 167,384,000 hours of delay resulted from traffic in 2017
- Annual cost of $3.1 billion or $1,408 for each single commuter
- The annual collision lists a crash every 4.5 minutes and a fatal crash every 16 hours

What if we could actively reduce the number of traffic accidents or their severity?

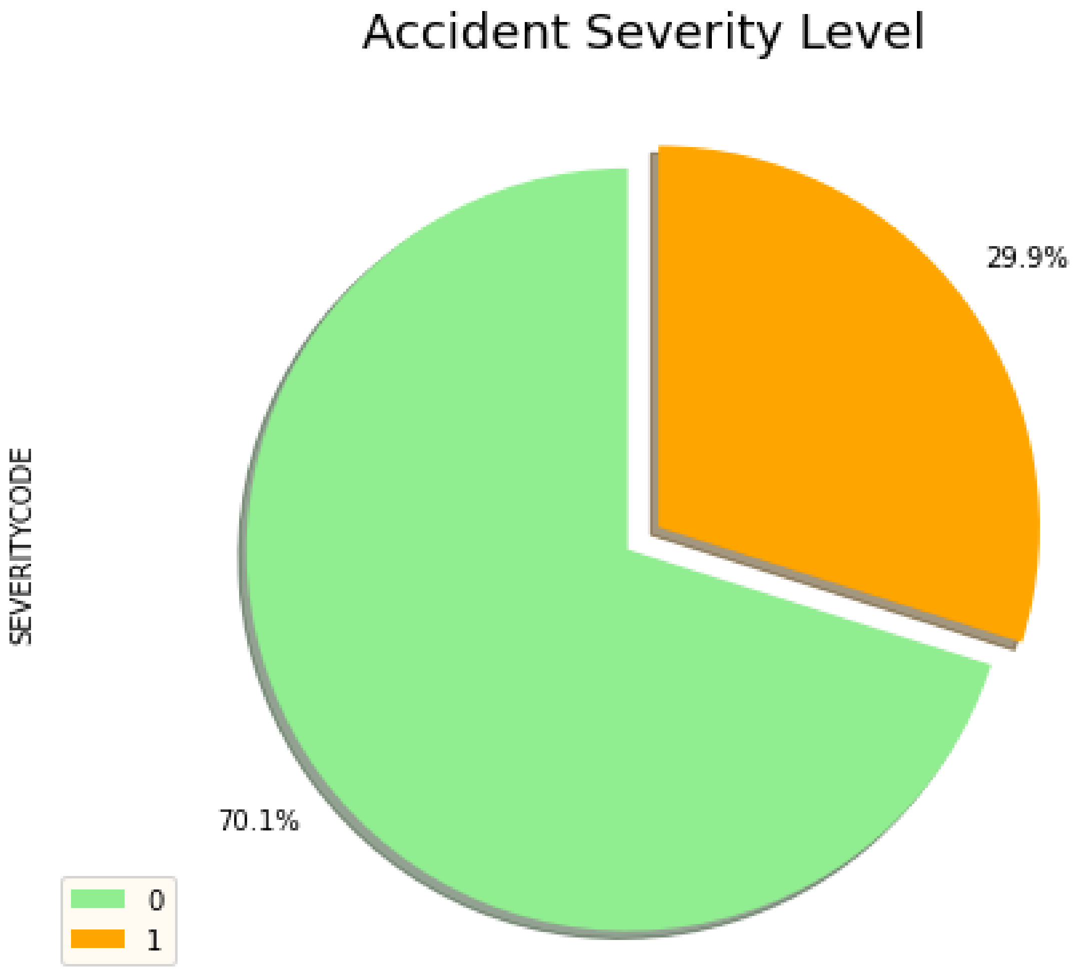
- We want to identify which variables have the biggest impact on the severity of accidents
- We want to train models to accurately predict accident severity
- Learning more about what causes severe crashes should interest everyone, as it leads to making our roads safer

The data provided by the Seattle Department of Transportation (SDOT) contains all vehicular accidents between cars, pedestrians, cyclists and so forth

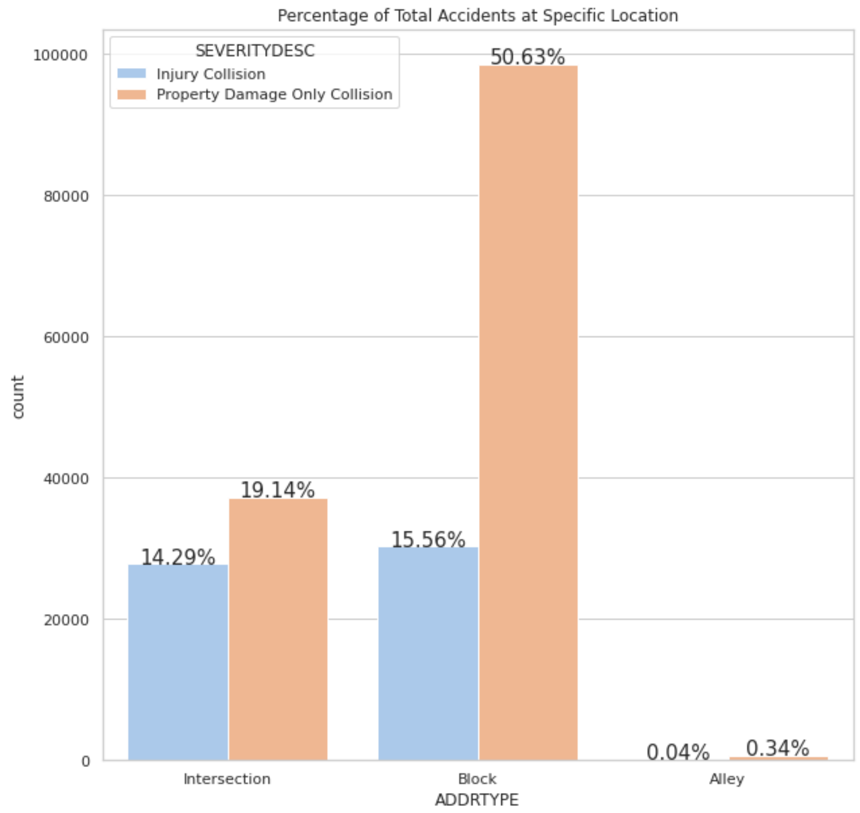
- Lists over 200,000 samples
- Includes 37 attributes
- Still needs some cleaning because of missing, similar or redundant information

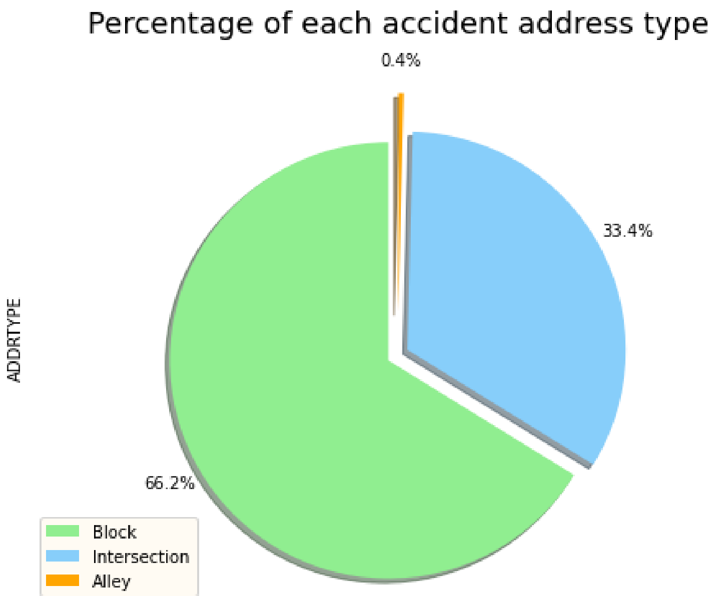


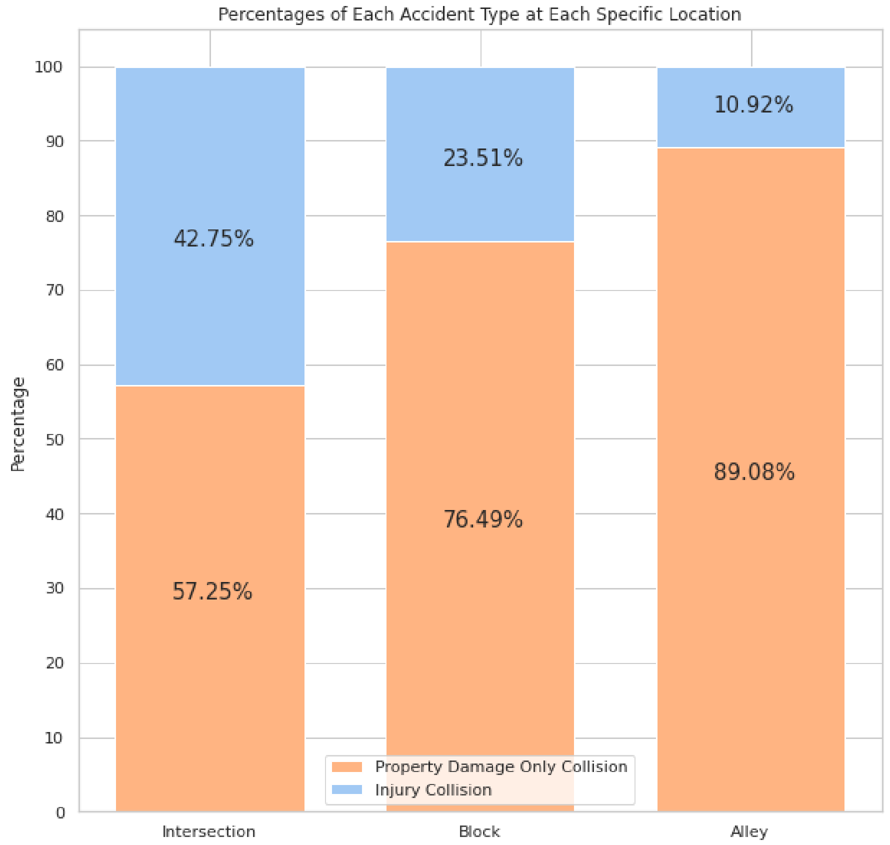
Imbalanced spread of target variable values may lead to bias in the machine learning models (Severe cases being 1, non-severe cases being 0)



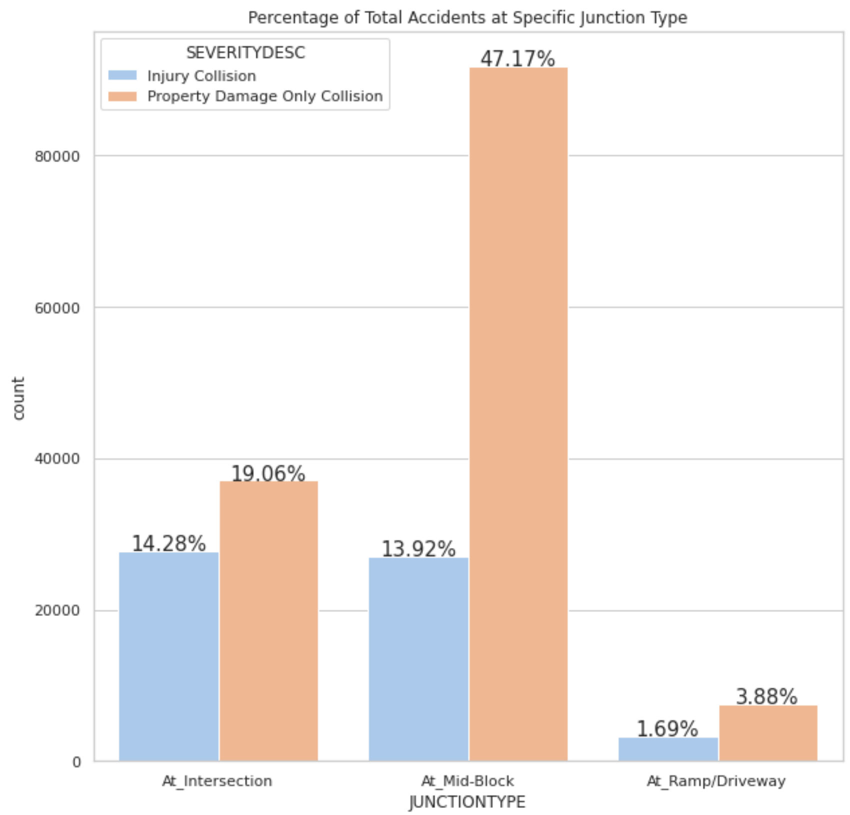
5. The Effect of the Address Type

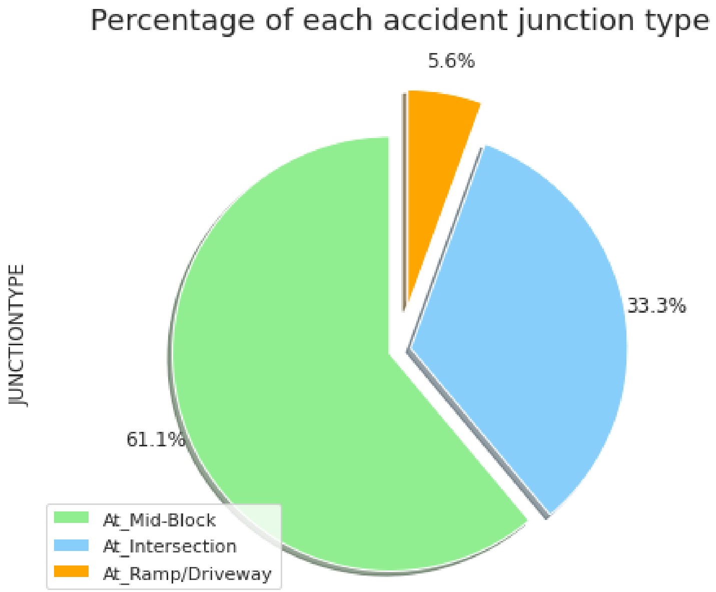


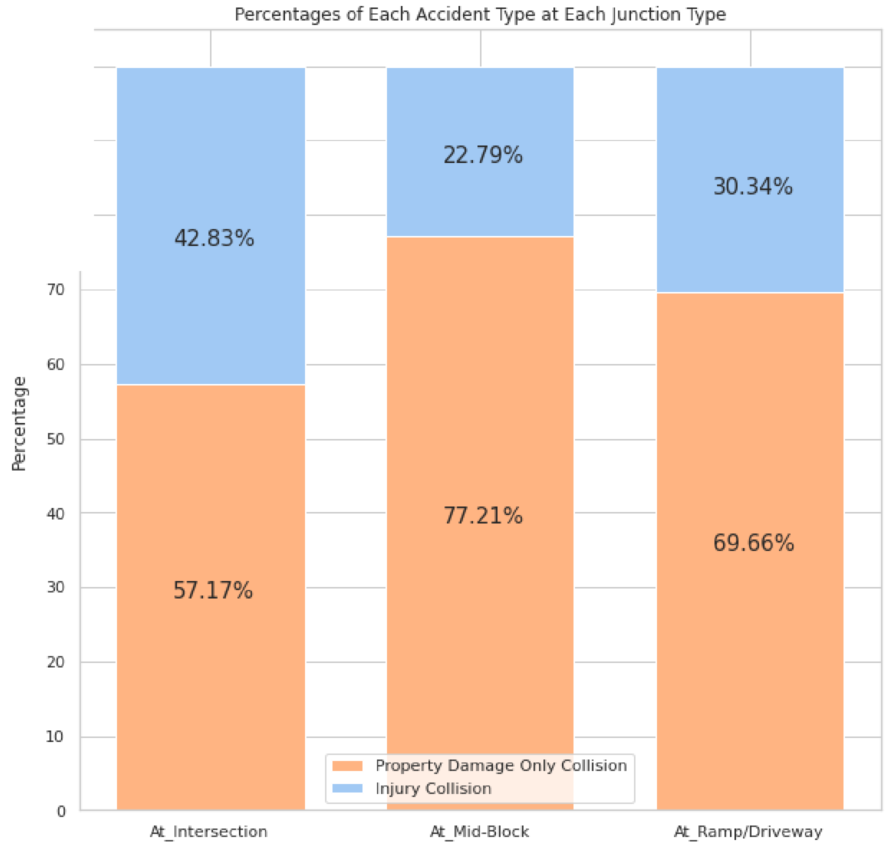




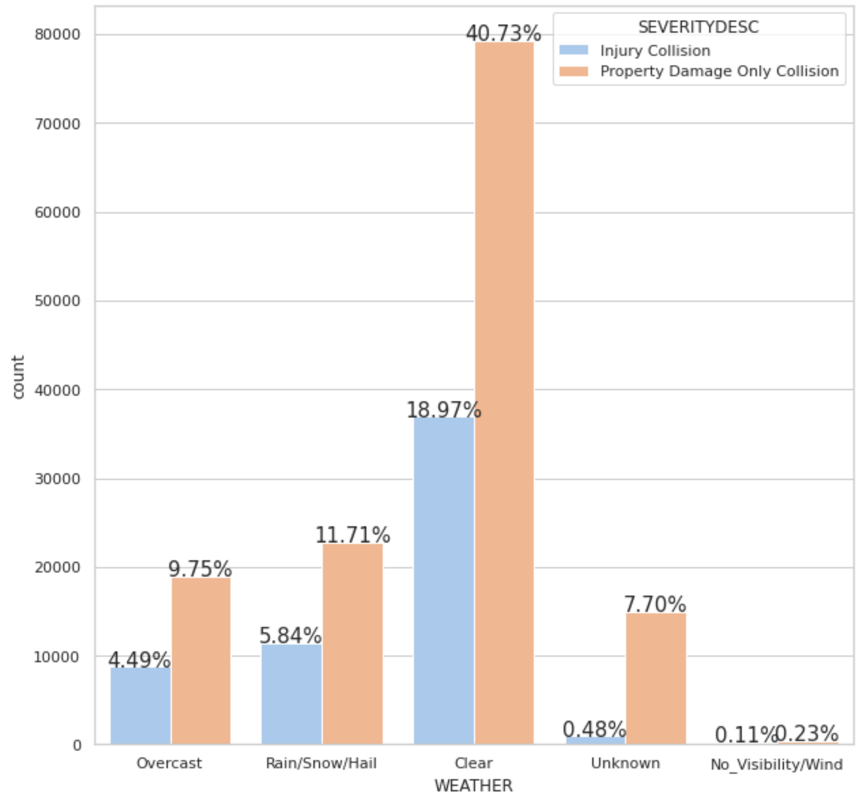
6. The Effect of the Junction Type

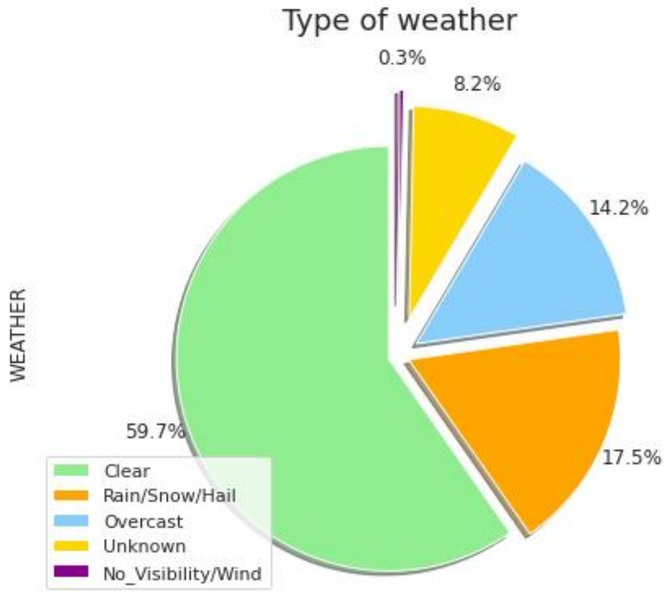


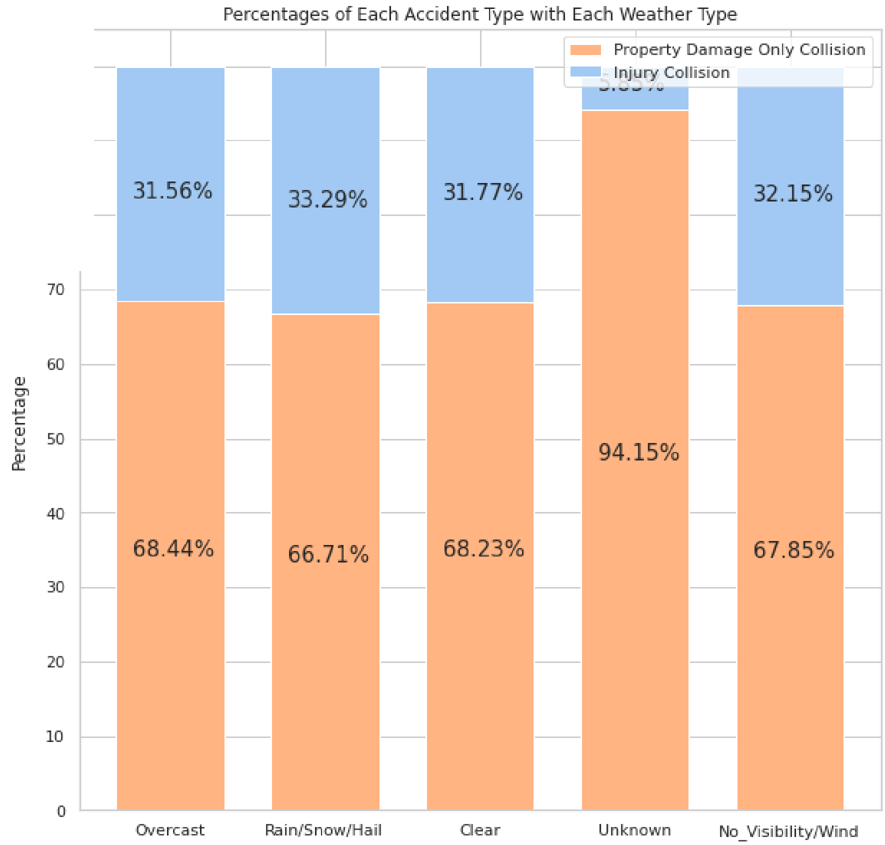




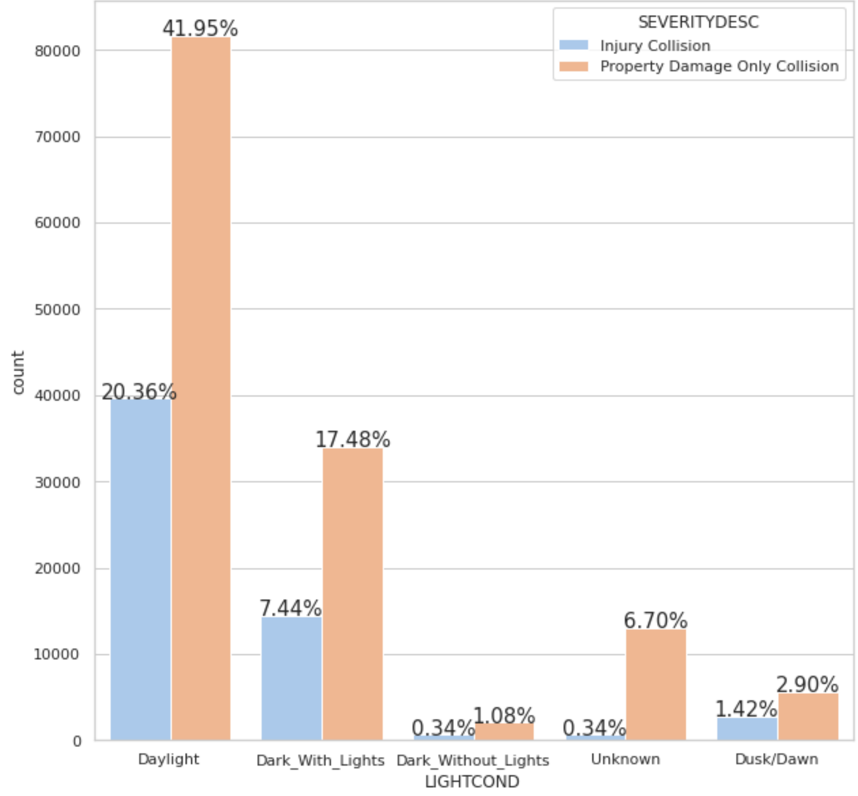
7. The Effect of the Weather Condition

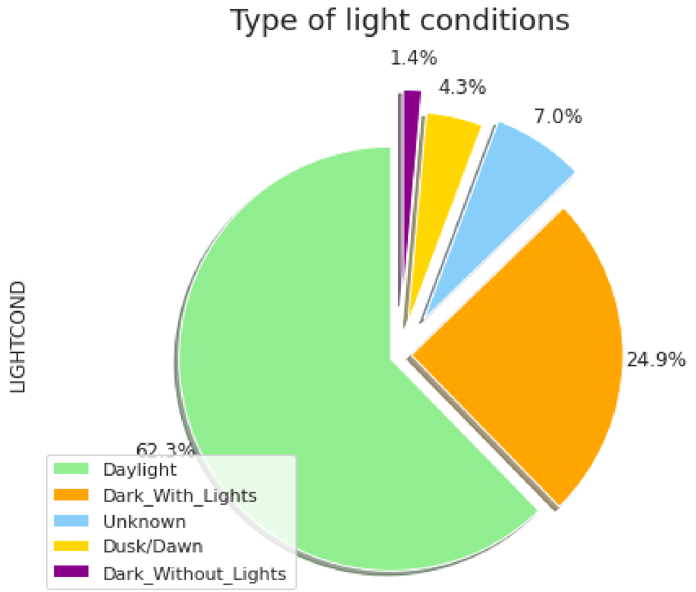


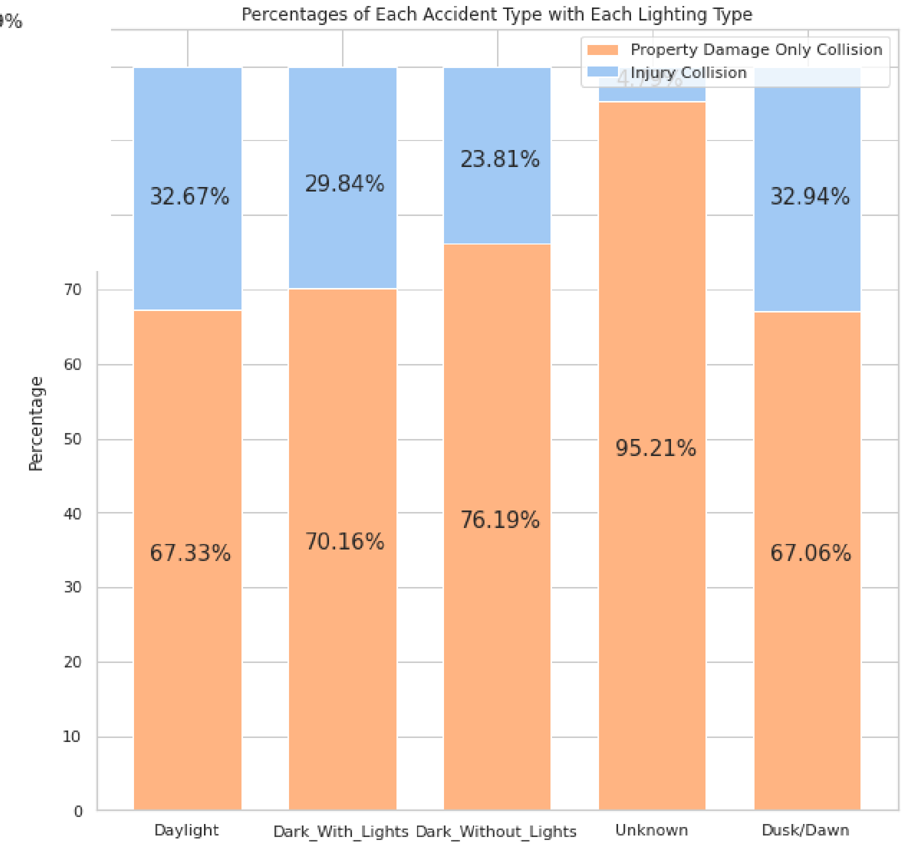




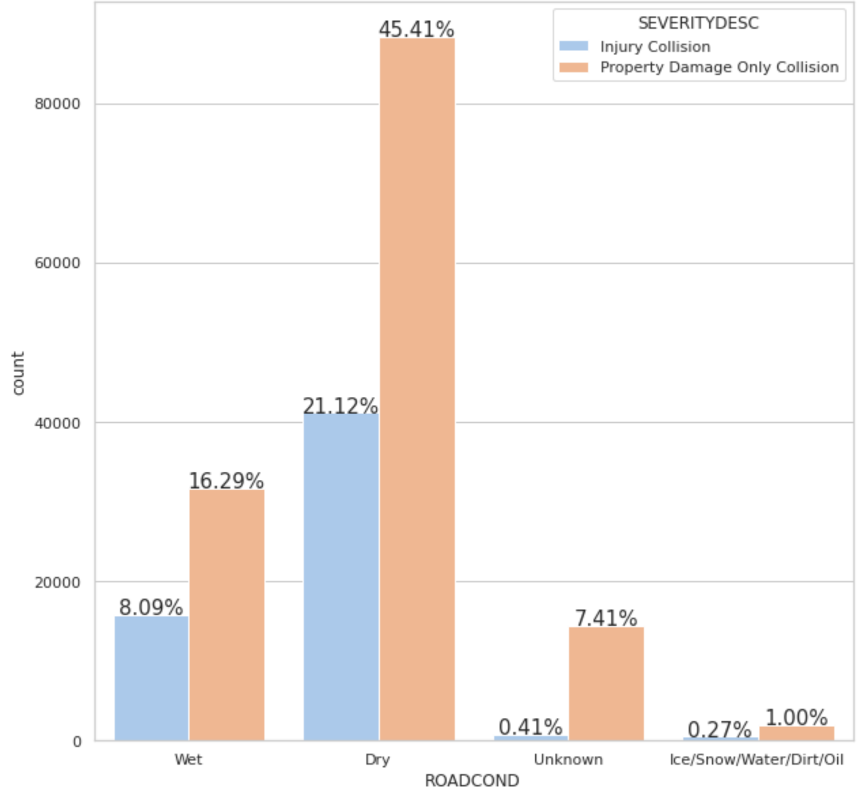
8. The Effect of the Light Condition

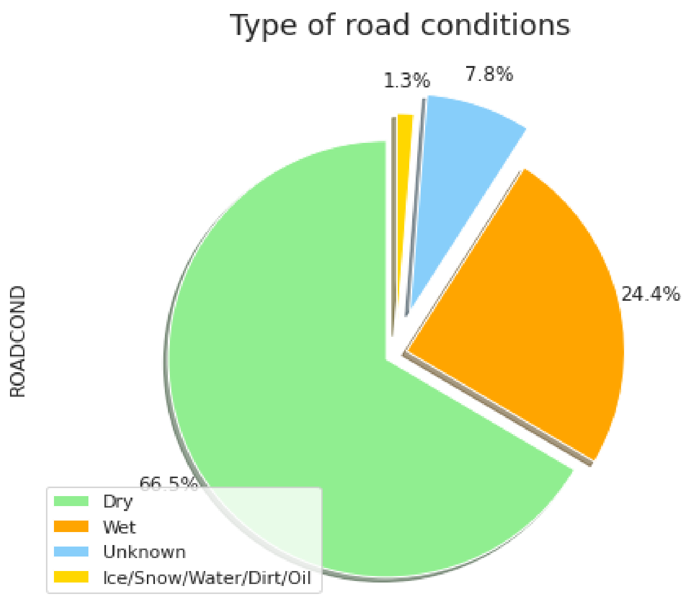


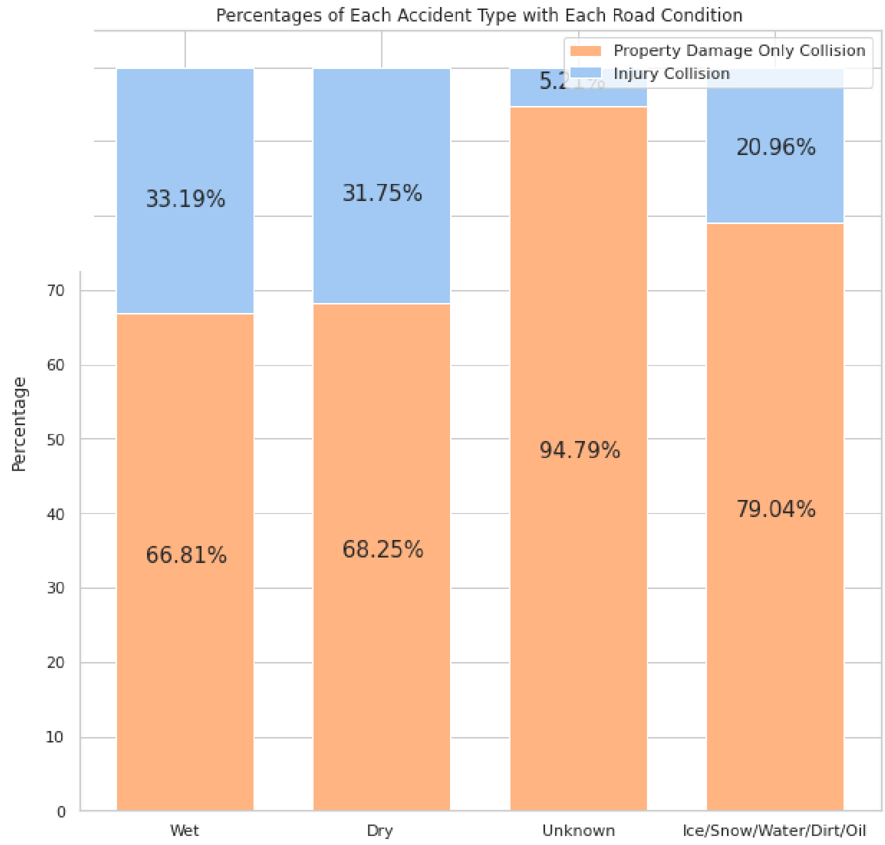




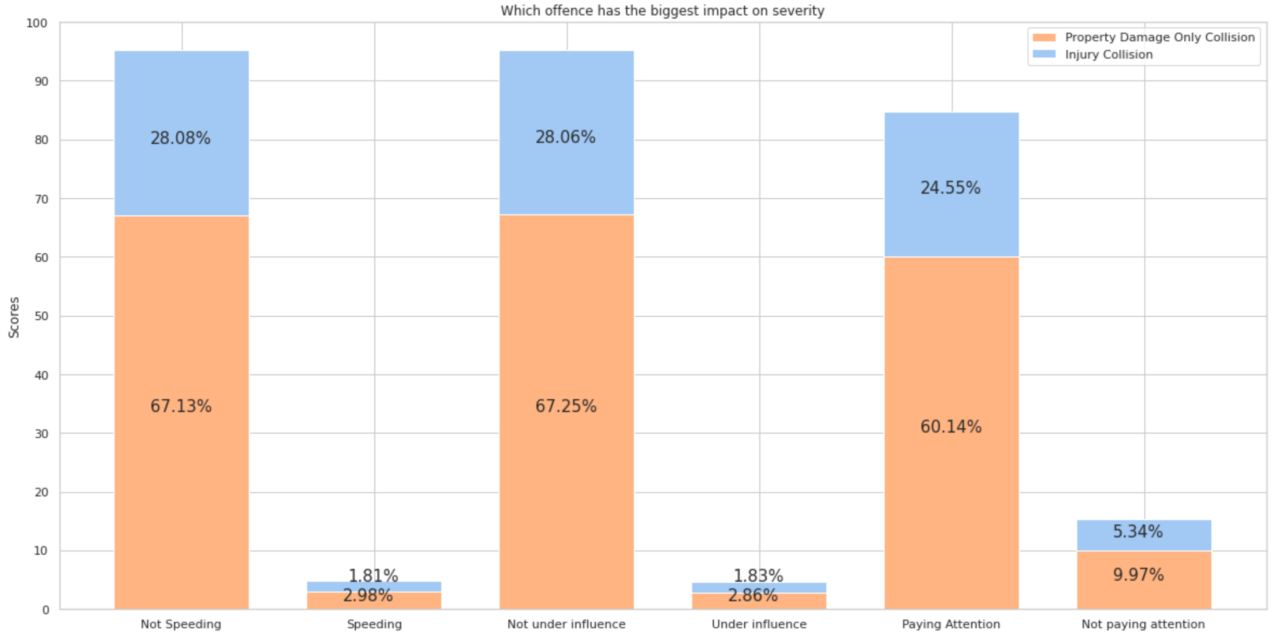
9. The Effect of the Road Condition

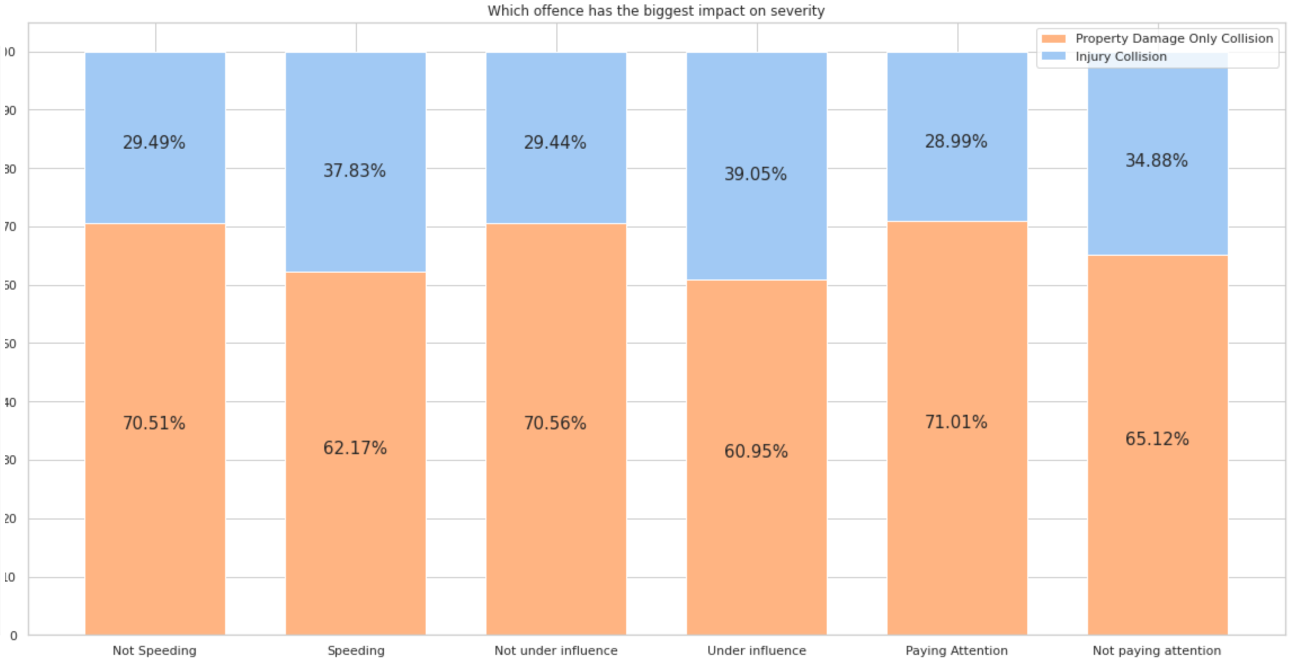






10. The Effect of the True/False Variables



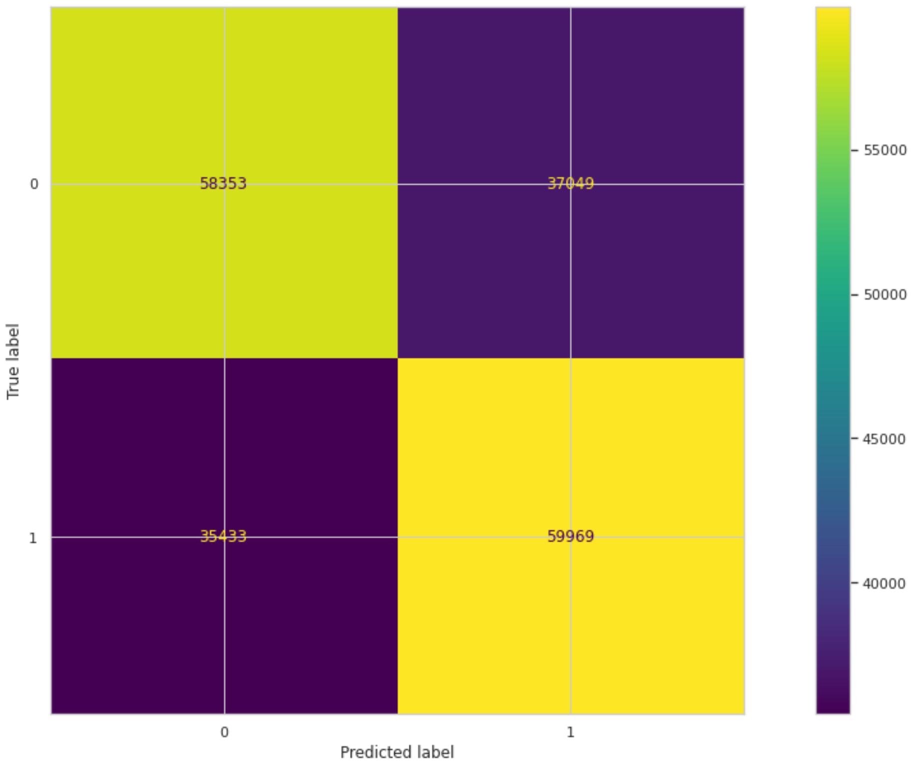


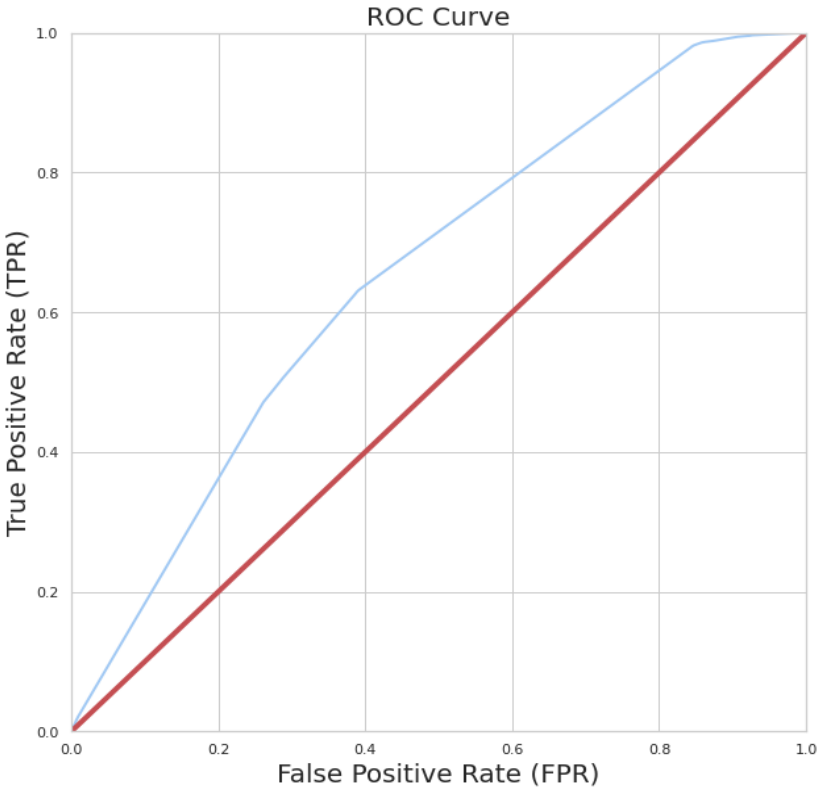
We decided on using scikit-learn's Decision Tree, Logistic Regression and K-Nearest Neighbor for our models

Our feature set looks like the following:

Variables	Description
ADDRTYPE	At what kind of location the accident took place (as a category)
JUNCTIONTYPE	At what kind of junction the accident took place (as a category)
INATTENTIONIND	Whether or not the driver was paying attention (1/0)
UNDERINFL	Whether or not the driver was under the influence (1/0)
SPEEDING	Whether or not the driver was speeding (1/0)
WEATHER	Weather conditions during the accident
ROADCOND	Road conditions during the accident
LIGHTCOND	Light conditions during the accident
HITPARKEDCAR	Whether or not the accident involved a parked car (1/0)

12. Decision Tree Analysis



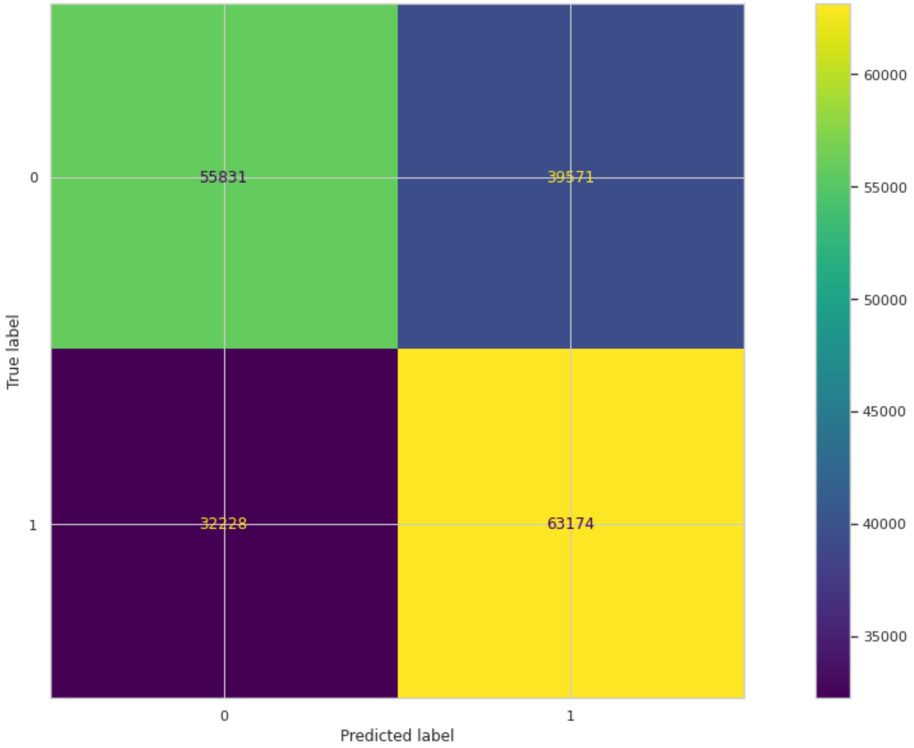


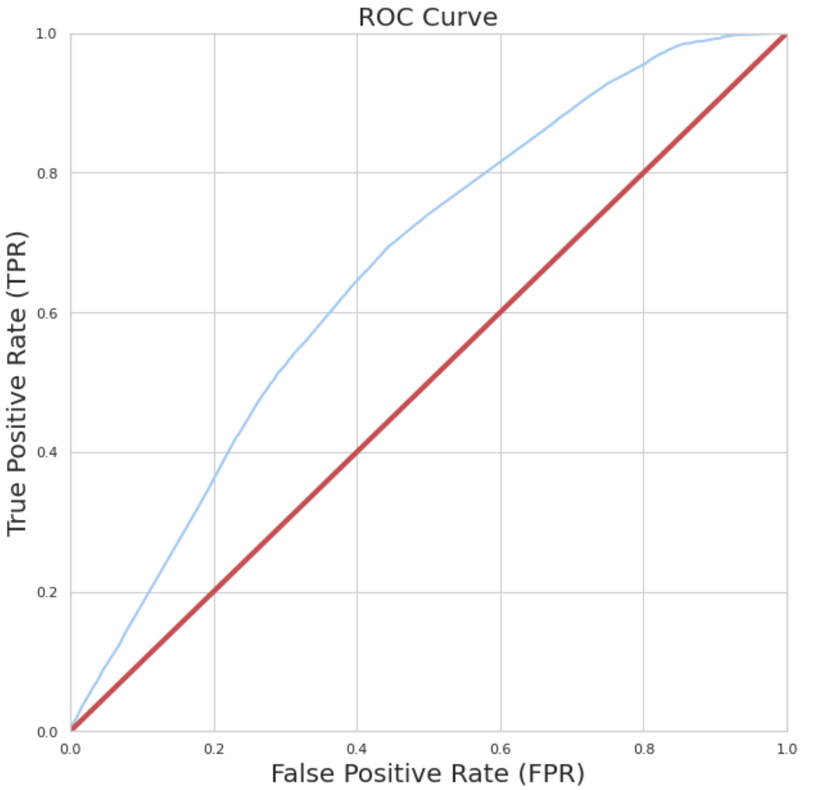

```
Jaccard score for the decision tree model: 0.3277469080480373
              precision    recall  f1-score   support

           0       0.80      0.61      0.69     41083
           1       0.40      0.63      0.49     17319

    accuracy                           0.61     58402
   macro avg       0.60      0.62      0.59     58402
weighted avg       0.68      0.61      0.63     58402
```

13. Logistic Regression



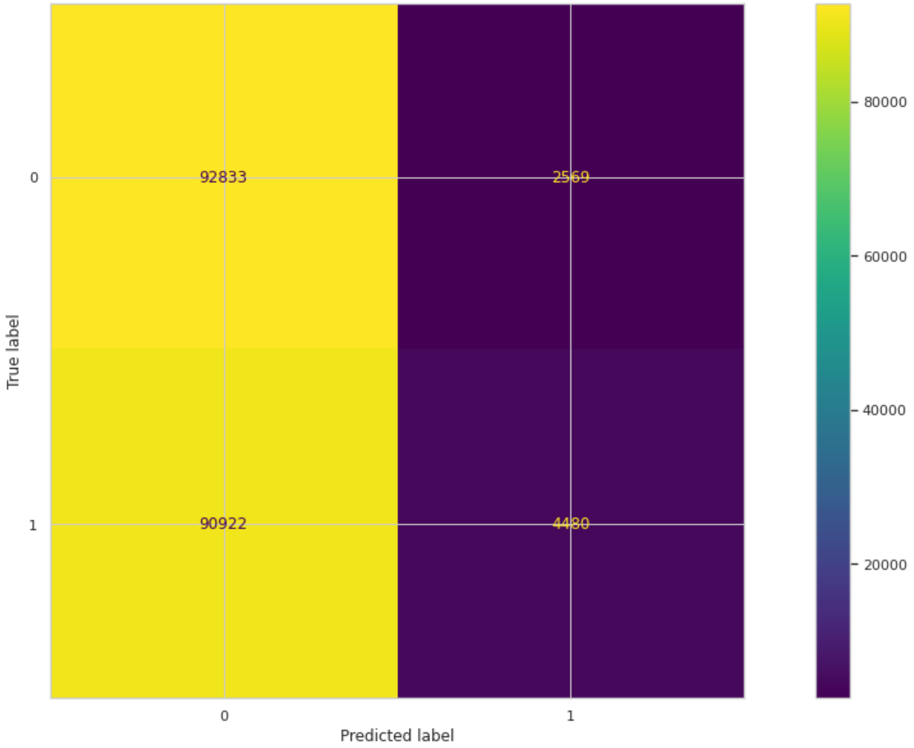


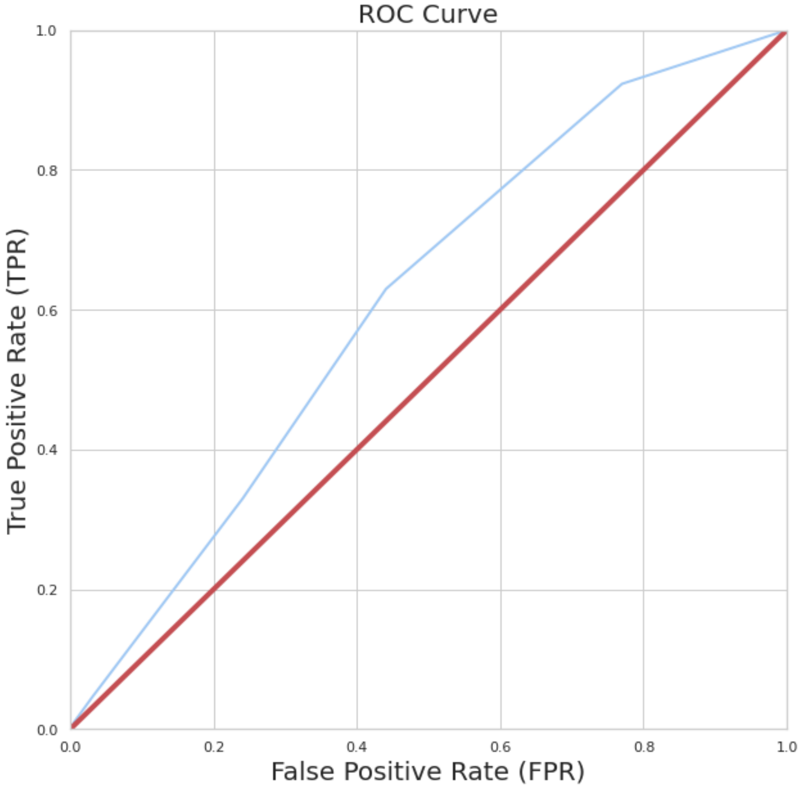

```
Jaccard score for the logistic regression model: 0.3338735818476499
              precision    recall  f1-score   support

           0       0.80      0.58      0.67     41083
           1       0.40      0.67      0.50     17319

    accuracy                           0.61     58402
   macro avg       0.60      0.62      0.59     58402
weighted avg       0.69      0.61      0.62     58402
```

14. K-Nearest Neighbor

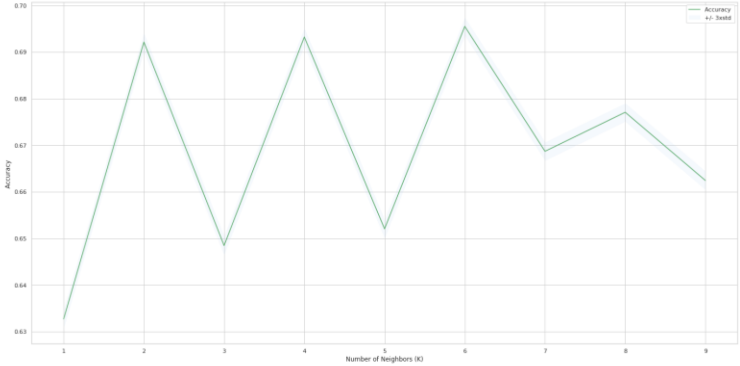





```
Jaccard score for the logistic regression model: 0.0425325724130505
              precision    recall  f1-score   support

           0       0.71      0.97      0.82     41083
           1       0.39      0.05      0.08     17319

    accuracy                           0.70     58402
   macro avg       0.55      0.51      0.45     58402
weighted avg       0.61      0.70      0.60     58402
```



The best model is the Logistic Regression model

- Similar good values for most metrics
- Leads in Recall which is our most important metric for our problem case
- Recall tells us the percentage of how many severe cases the model predicted correctly out of all severe cases it should have gotten
- This is the most important metric, because we want to be on the safe side and prefer to have too many false positives, as our models missing on predicting some severe cases

Algorithm	Accuracy	Jaccard	F1-score	Precision	Recall	Time (s)	LogLoss
Decision Tree	0.614688	0.327747	0.631087	0.404446	0.633466	0.324444	NA
Logistic Regression	0.605904	0.333874	0.622953	0.400987	0.666089	0.552270	0.644885
KNN	0.695490	0.042533	0.599260	0.386308	0.045615	94.956030	NA

Following steps could improve the result

- A bigger sample size would lead to a more refined model
- Reduce the number of missing values Recall tells us the percentage of how many severe cases the model predicted correctly out of all severe cases it should have gotten
- This is the most important metric, because we want to be on the safe side and prefer to have too many false positives, as our models missing on predicting some severe cases

Following steps could improve the result

- A bigger sample size would lead to a more refined model
- Reduce the number of missing values
- Add more diverse features to the dataset

The result is already helpful for following use cases

- Drivers can be taught about which factors lead to severe accidents and can then adjust their driving appropriately
- Hospitals, police and ambulances can prepare and be on alert for potential severe accidents that are predicted to happen by our model
- Governmental agencies can improve upon or eliminate factors leading to severe accidents
- Navigation app developers can implement the variables and send live information about high probability of severe accidents