



IBM DATA SCIENCE CAPSTONE

SEATTLE CAR ACCIDENT SEVERITY PREDICTION

Barche Blaise Forgwa

9th September 2020

Table of Contents

Introduction	1
Background	1
Business problem	2
Data understanding	2
Methodology	5
Results	6
Logistic regression model	6
K-nearest neighbors	8
Decision tree (DT) classifier	9
Discussion	11
Conclusion	12
Recommendations	12
References	12

Introduction

Background

Seattle represents the largest city in Washington state(<https://en.wikipedia.org/wiki/Seattle>). It has a population of 3.4 million as of 2020([Macrotrends](#)). Each day, 3,700 people die from a road traffic crash ([CDC injury prevention](#)). In 2017, the total car accidents recorded was 120,993 (<https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state.cfm>). And these numbers are said to increase with the increasing number of cars in this state.

Business problem

Road traffic accidents often involve damages ranging from minor injuries to serious injuries including mortality. The outcomes of road traffic accidents are often influenced by the reaction time of traffic surveillance systems to contact emergency health care workers in case of human injuries or construction teams in case of property damage. This project sought to build a prediction model for severity of road traffic accidents (human injuries or property damage) such that traffic surveillance systems may prioritize the teams (health care teams or construction teams) to deploy to an accident site in future accidents.

Data understanding

The dataset for this will be the Seattle accident data which holds data on accident severity ([Seattle car collision dataset](#)). Overall there are 194673 rows and 38 attributes

The selected features from the dataset following data cleaning are presented in the table below with their respective descriptions

Table 1. Selected features for classification models

Feature	Description
INATTENTIONIND	Whether or not collision was due to inattention (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol Feature Description
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision (Dry, wet)
LIGHTCOND	The light conditions during the collision
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ADDRTYPE	Collision address type: (Alley, Block, Intersection)

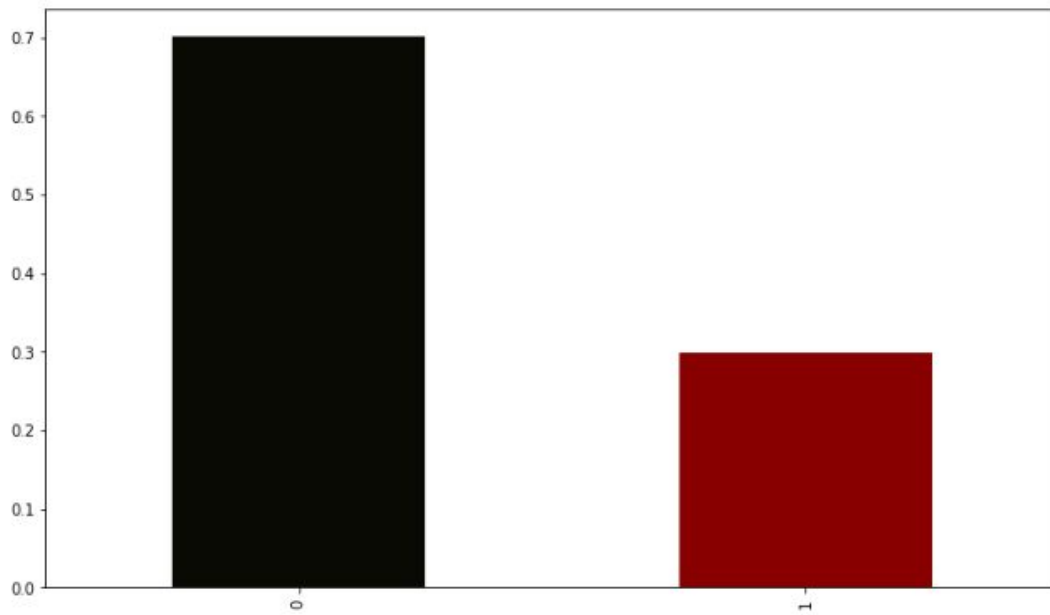
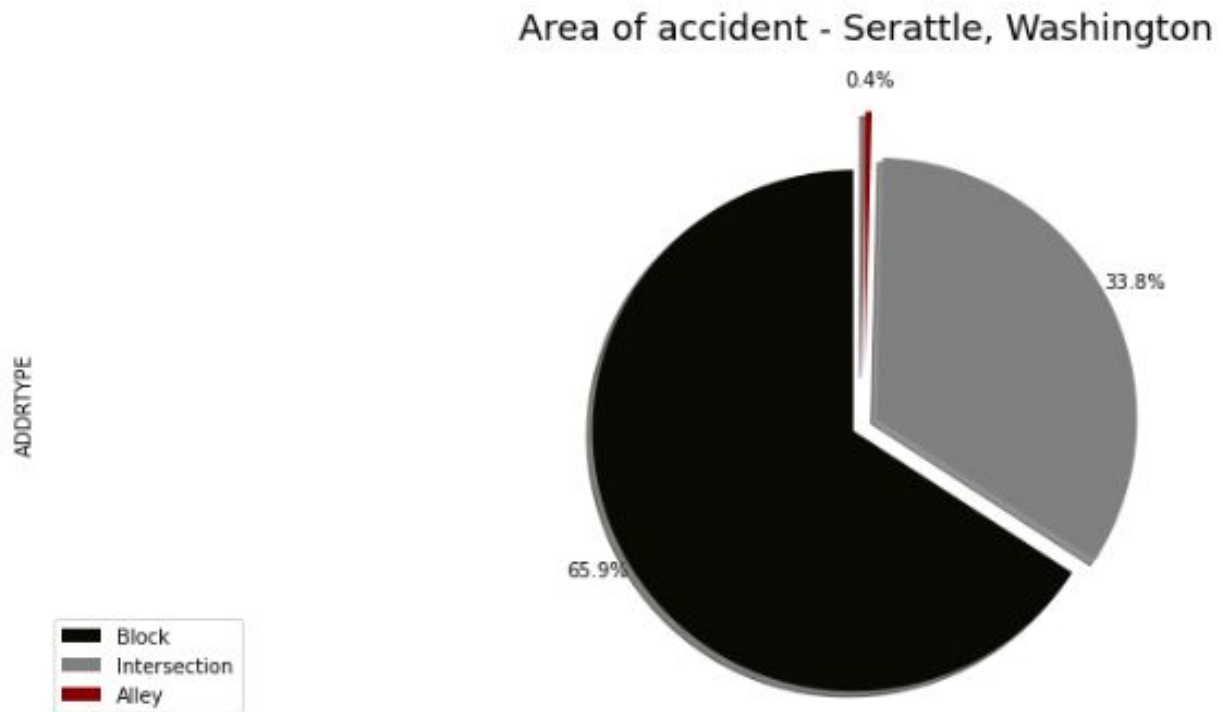


Figure 1. Distribution of classes in the target variable(SEVERITYCODE)

Property damage was coded as 0 and injury as 1



Percentage of accident by Address type

Methodology

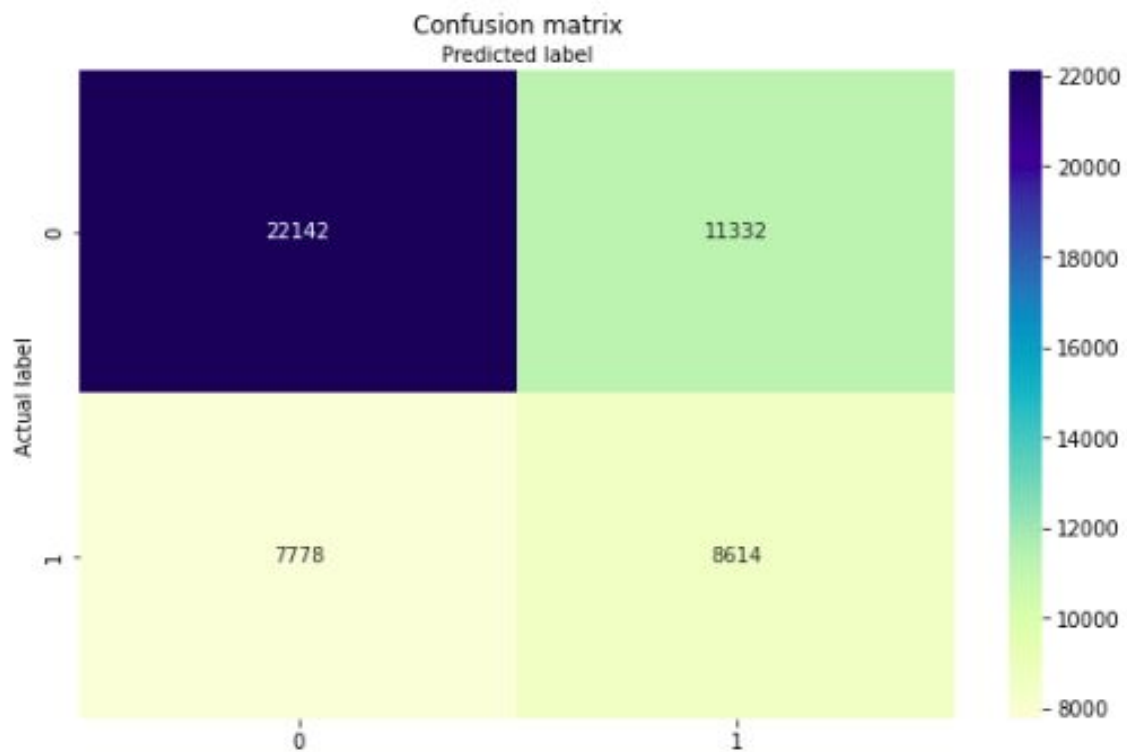
Pandas and Numpy libraries will be used for data importation and processing. The Matplotlib library will be used for data visualization. Exploratory data analysis using visualization and basic descriptive statistics will be used to determine the independent variables(e.g. weather condition, road condition, time of the day and day of the week) for building prediction models. Supervised machine learning classification models (K nearest neighbor , Logistic regression models, decision tree classifier) will be used to predict accident severity (injury or property damage). Scikit learn library will be used for model development and evaluation.

Due to imbalance in classes in the target feature, SMOTE function from the imblearn library was used to balance up the distribution of the classes in the training set of the selected data.

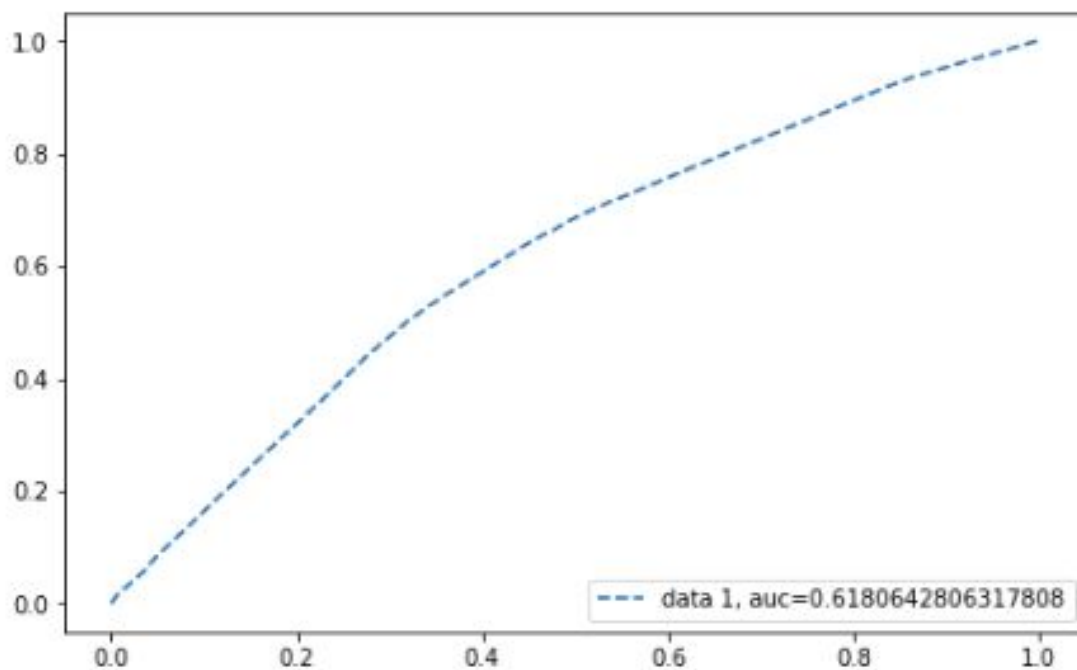
Results

Logistic regression model

The model was constructed with solver as 'liblinear', the figures below represents the confusion matrix and ROC curve respectively



Confusion matrix logistic regression model



ROC curve Logistic regression model.

The AUC from the logistic regression model was 0.62

The table below provides a classification report of the logistic regression model

Table 2. Classification report logistic regression model

	Precision	Recall	F1 score
0 (property damage)	0.66	0.74	0.70
1 (injury)	0.53	0.43	0.47
accuracy			0.62
Weighted average			0.61

The logistic regression model predicted property damage with more precision than an injury

K-nearest neighbors

The KNeighborsClassifier imported from the scikit-learn library was used to build the KNN model, a K of 4 was selected following iteration of different k values in the range of 1-8 and plotting their respective mean accuracy score (figure 4). The confusion matrix is presented in figure 5.

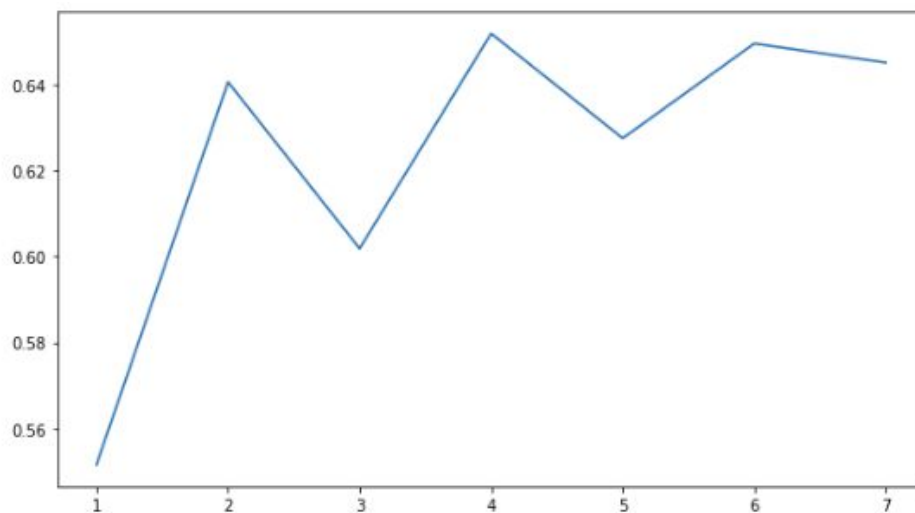


figure 4. Best K for KNeighborsClassifier model

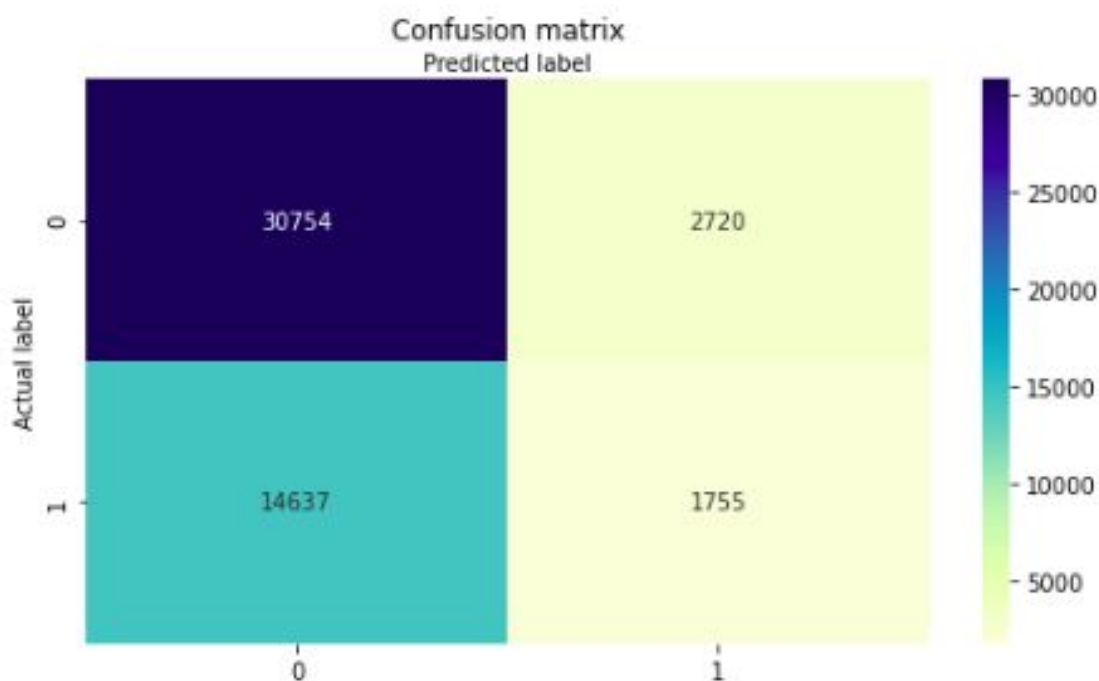


figure 5 : Confusion matrix KKN model

Table 3. Classification report KNN classifier

	Precision	Recall	F1 score
0 (property damage)	0.92	0.68	0.78
1 (injury)	0.11	0.39	0.17
accuracy			0.65
Weighted average			0.73

Overall, the precision of the KKN model was 0.39 with a recall rate of 0.11

Decision tree (DT) classifier

The criterion for node selection was entropy and the maximum depth was 4. The figure below presents the confusion matrix of the decision tree classifier.

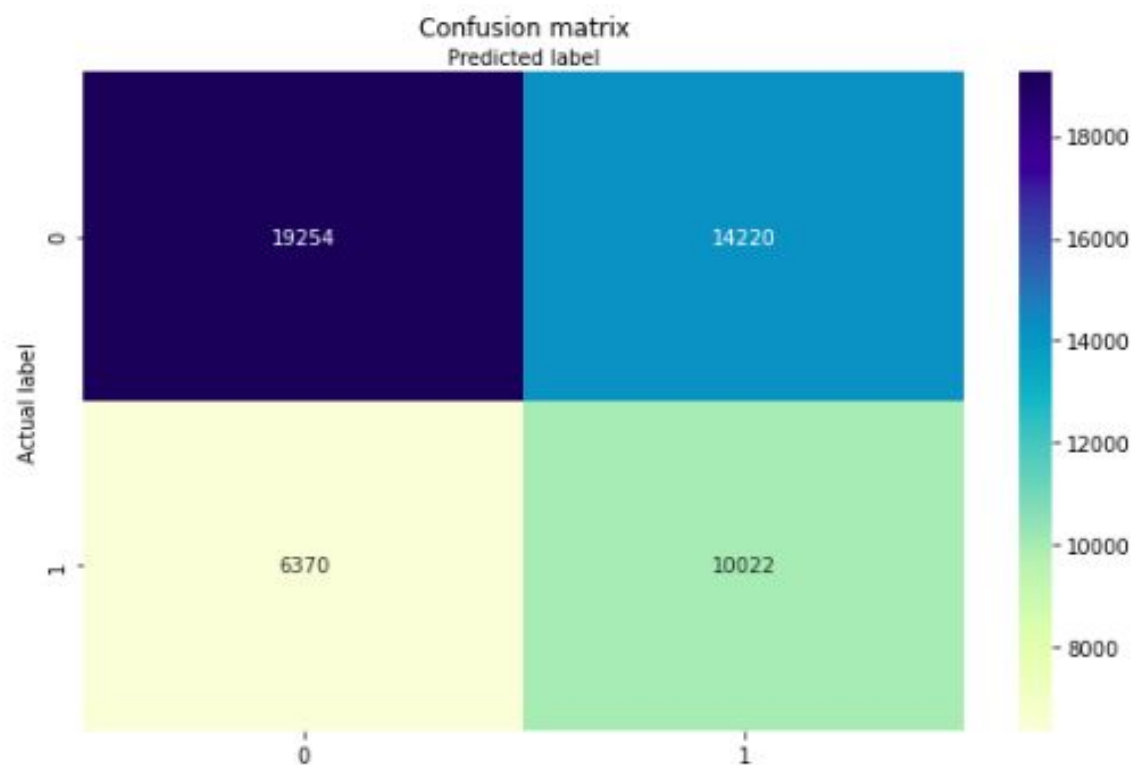


figure 6: Decision tree classifier confusion matrix

Table 4. Classification report for Decision tree classifier

	Precision	Recall	F1 score
0 (property damage)	0.58	0.75	0.65
1 (injury)	0.61	0.41	0.49
accuracy			0.59
Weighted average			0.57

Table 5. Comparing classification reports the three classification models

Model	F1 score	Accuracy	recall	precision
KNN	0.58	0.65	0.10	0.39
Logistic regression	0.62	0.62	0.43	0.52
Decision tree	0.60	0.58	0.61	0.41

Discussion

The precision simply means the percentage of relevant predictions by the model, whereas the recall implies the percentage of relevant predictions correctly predicted by the model. Though there is often a tradeoff between these two evaluation matrices, the F1 score obtains the harmonic mean between these two evaluators.

Overall, the three models had an accuracy between 58% to 65 % . The logistic regression model had a better precision than the other two models. The AUC (area under the receiver operator characteristic curve) measures the capability of distinguishing classes and ranges from 0 (no ability to distinguish classes) to 1 (can perfectly distinguish between classes). Thus an AUC of 0.61 for the logistic regression model stipulates the model is moderately efficient in separating the different classes of accident severity.

Conclusion

The three models were modest in predicting car accident severity.

This could be improved by:

- Obtaining more data regarding speeding condition of drivers
- Better classification of road, light and weather conditions

Nevertheless, the logistic regression model proved to be a better classifier model than the others

Recommendations

To Drivers


- Be cautious in major roads in the city
- Be cautious when driving along a block or when approaching an intersection

To traffic surveillance system:

- Obtain more information from drivers to enable creation of robust models

References

1. [Precision and recall](#)
2. [understanding ROC and AUC](#)
3. [Seattle geodata](#)
4. <https://en.wikipedia.org/wiki/Seattle>
5. [Macrotrends](#)
6. [Imbalanced dataset classifier](#)
7. [CDC injury prevention](#)

- 
8. https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state_cfm