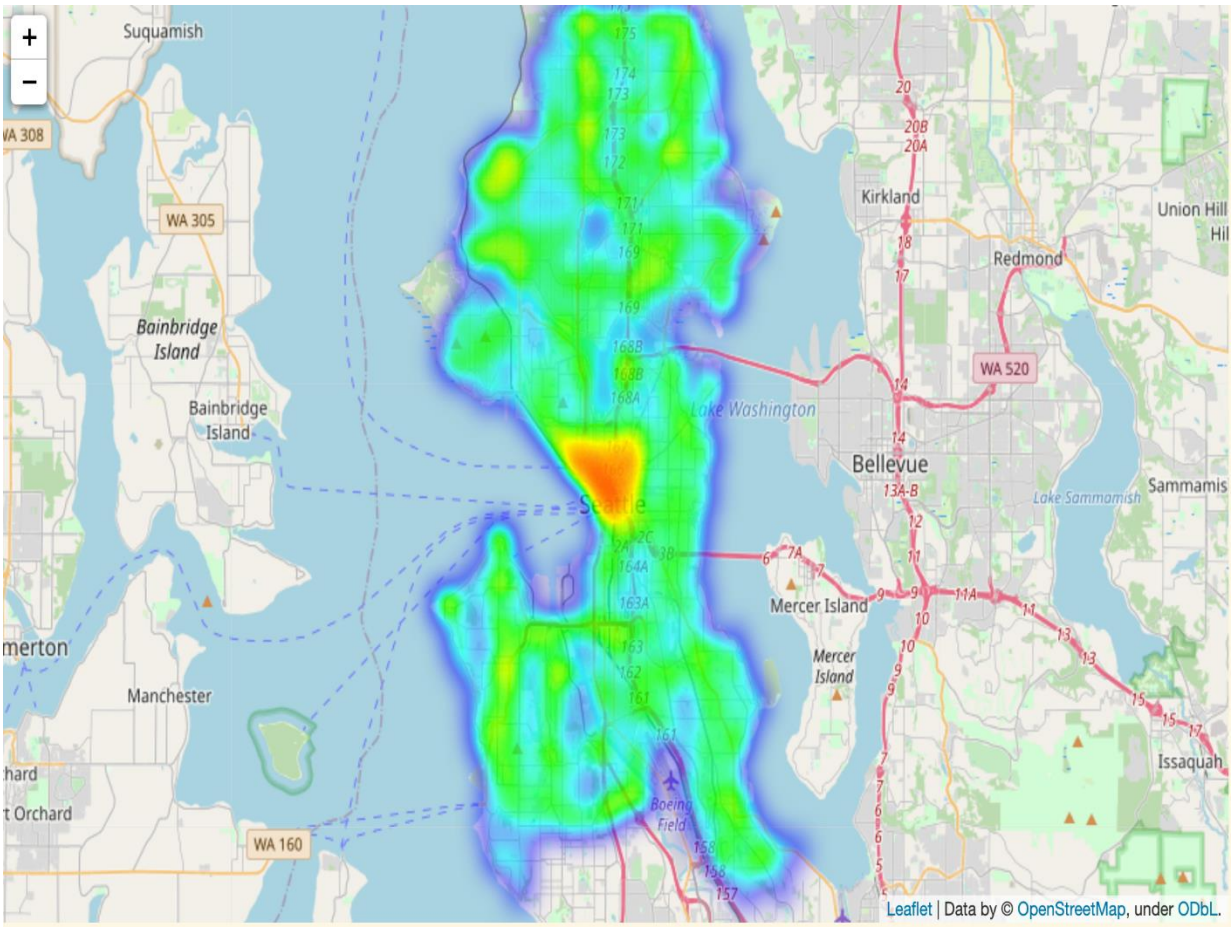
IBM Coursera Capstone: Seattle Car Collision Severity Presentation

Business Understanding:

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

Describe project

It is always rainy and windy in Seattle, and on the way, you always come across a terrible traffic jam on the other side of the highway, with long lines of cars barely moving. It would be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.



Objective

Luckily enough, The Seattle Police Department (SPD) has recorded all car collision accident from 2004 to present. Basing on those historical data ,we can create a map and information chart to help us understand the high-risk areas, understand car injury factors to avoid accident, and plan our next trip to Seattle better.

Algoritms

Overall Result

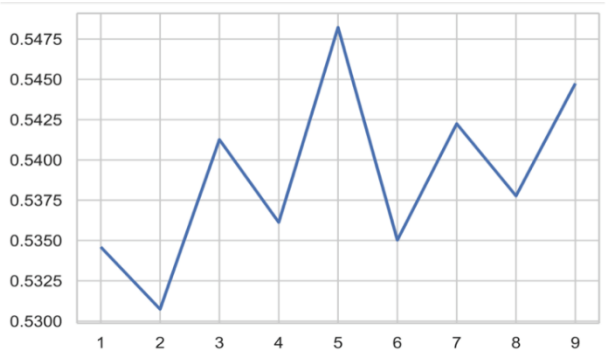
Alogorithm	Average F-1 Score	Type	Precision	Recall
Decision Tree	0.61	Property collision	0.80	0.56
		Injury Collision	0.39	**0.67**
k-Nearest Neighbor	0.60	Property collision	0.74	0.65
		Injury Collision	0.36	**0.46**
Logistic Regression	0.61	Property collision	0.71	0.97
		Injury Collision	0.45	**0.06**

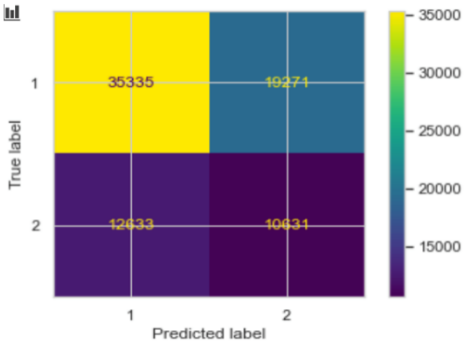
K nearest neighbors

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

5.1 K-NEAREST NEIGHBOR

The best K, as shown below, for the model where the highest elbow bend exists is at 5.





The Confusion Matrix shows that the KNN model have an accuracy of 59%, and the recall rate is 46%.

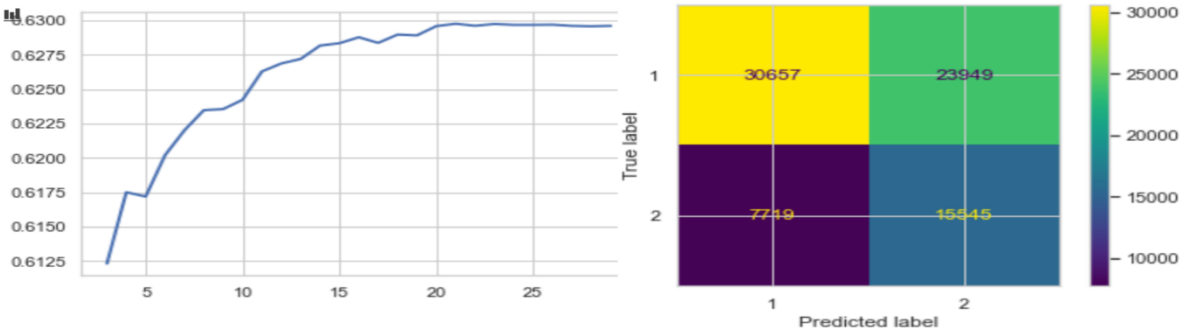
	precision	recall	f1-score	support
Property collision	0.74	0.65	0.69	54,606
Injury Collision	0.36	**0.46**	0.40	23,264
accuracy			0.59	77,870
macro avg	0.55	0.55	0.54	77,870
weighted avg	0.62	0.59	0.60	77,870

The Decision Tree Analysis

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

5.2 DECISION TREE ANALYSIS

The criterion chosen for the classifier was 'entropy' and the **max depth was 21** with **best accuracy of 63%.**



Confusion Matrix shows that the Decision Tree model have an accuracy of 59%, and the recall rate is 67%.

	precision	recall	f1-score	support
Property collision	0.80	0.56	0.66	54606
Injury Collision	0.39	**0.67**	0.50	23264
accuracy			0.59	77870
macro avg	0.60	0.61	0.58	77870
weighted avg	0.68	0.59	**0.61**	77870

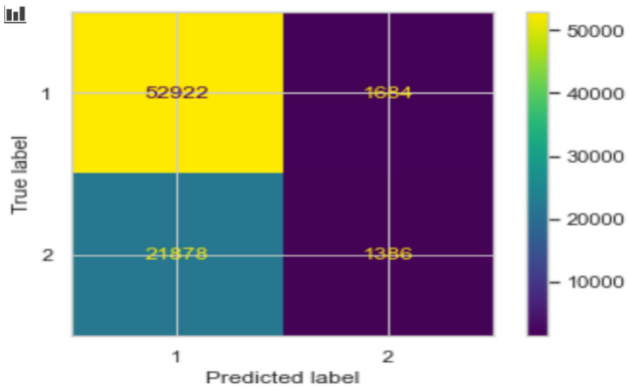
Logistic regression

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

5.3 LOGISTIC REGRESSION

We use GridSearchCV to search the best parameters.

The C used for regularization strength was '0.01' and penalty was "l2", whereas the solver used was 'liblinear'.



	precision	recall	f1-score	support
Property collision	0.71	0.97	0.82	54606
Injury Collision	0.45	0.06	0.11	23264
accuracy			0.70	77870
macro avg	0.58	0.51	0.46	77870
weighted avg	0.63	0.70	0.61	77870

Although the Logistic Regression model has an accuracy ratio of 70%, but the recall ratio is only 6%, which is not what we want to predict.

7. CONCLUSION

In this study, I analyzed the factors which may lead to injury (severity of a car collision). I identified address type/intersection type, weather condition, light condition, whether the driver had drug/alcohol, among the most important features that affect the severity of a car collision. I built three classification models to predict what condition would cause injury collision. These models can be very useful in helping the society in a number of ways. For example:

• drivers can use this model to adjust their behavior to avoid injury;

• insurance company can use this model to adjust auto insurance premium level;

•Traffic management department can use the model to help decrease future injury collision by providing better light and road infrastructure.

FUTURE DIRECTIONS

I was able to achieve a 67% recall ratio using the Decision Tree Model. However, the other two model didn't do better and can only predict "Property collision" well. This may be due to unbalanced dataset. If we can have a better dataset, future performance of these two other models might be different.

We can see from individual feature analysis part that there are some "unknown" types of weather, road condition, etc. have a significant difference with other conditions. This means that there maybe some unrecognized factors which need us to dig deeper.