Accident Severity in Seattle from 2004 to 2020

1. Introduction

1.1 Problem background and description

Growth rate of population in Seattle ranks No.2 in US. As a fast growing city, the traffic congestion problems have become a terrible problem. Accourding to the 2019 Urban Mobility Report, Seattle rank 7th on the problem of time delay for auto commuters traveling during peak periods (6 a.m. to 10 a.m. and 3 p.m. to 7 p.m.). Increasing number of traffic accidents is one of the most contribution to traffic jams. If prediction of severity of traffic accidents can be made, traffic congestion problem can be relieved.

With the development of machine learning, we want to use the data of traffic attributes including road condition, weather and other factors to predict accident severity

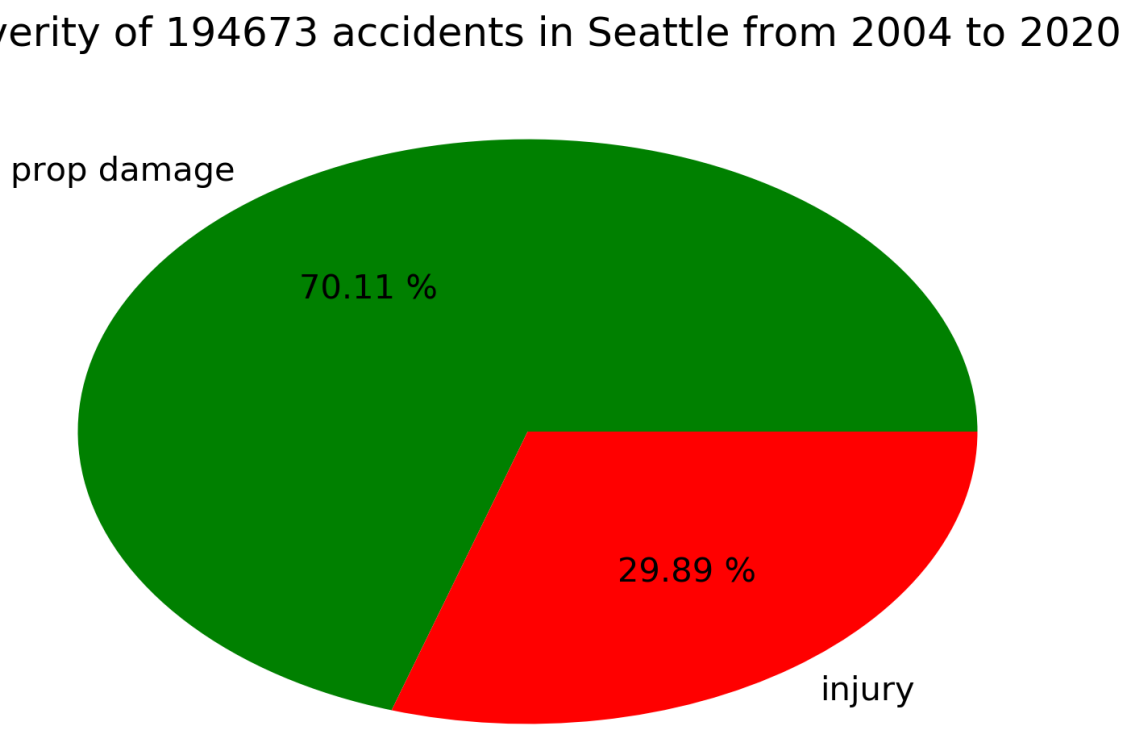
1.2 Target Audience

Our prediction of accident severity will be provided to the driver sufferred from traffic jams caused by traffic accidents and help them to decide whether they should change the routine or wait for the traffic jams. If the accident is serious and caused a large damage, we will recommend them to change their routine. Otherwise, they can still wait for traffic jams. Our prediction will help drivers to save their time and relieve traffic congestion.

2. Data

2.1 Data description

The data of severity of accident I used is from ArcGIS Metadata form. It concludes 194673 collisions happened in Seattle from 2004 to 2020. 37 attributes have been recorded on each collision. There are two kinds of severity have been recorded in this data set, injury and prop damage. 70% are prop damaged and 30% are injury.



Here, we analyzed features that have the protential relations with the severity of accidents which are shown below:

1. COLLISIONTYPE:Collision type
2. PERSONCOUNT: The total number of people involved in the collisions.
3. JUNCTIONTYPE: Category of junction at which collision took place
4. INATTENTIONIND: Whether or not the collision was caused by the inattention
5. WEATHER: Weather condition
6. ROADCOND: Road condition during the collision.
7. LIGHTCOND: Light condition during the collision.
8. SPEEDING: Whether or not speeding was a factor in the collision.

2.2 Data cleaning

Becaused the data are unbalanced labelled, I randomly select 5000 injured accidents and 5000 prop damaged accidents as the whole data set to balance the label. After cleaning the data, I began to choose the features.

3. Method

3.1 Feature selection

We have picked up 8 features that can be used. Here we will discuss which features can be related to the severity of the collisions.

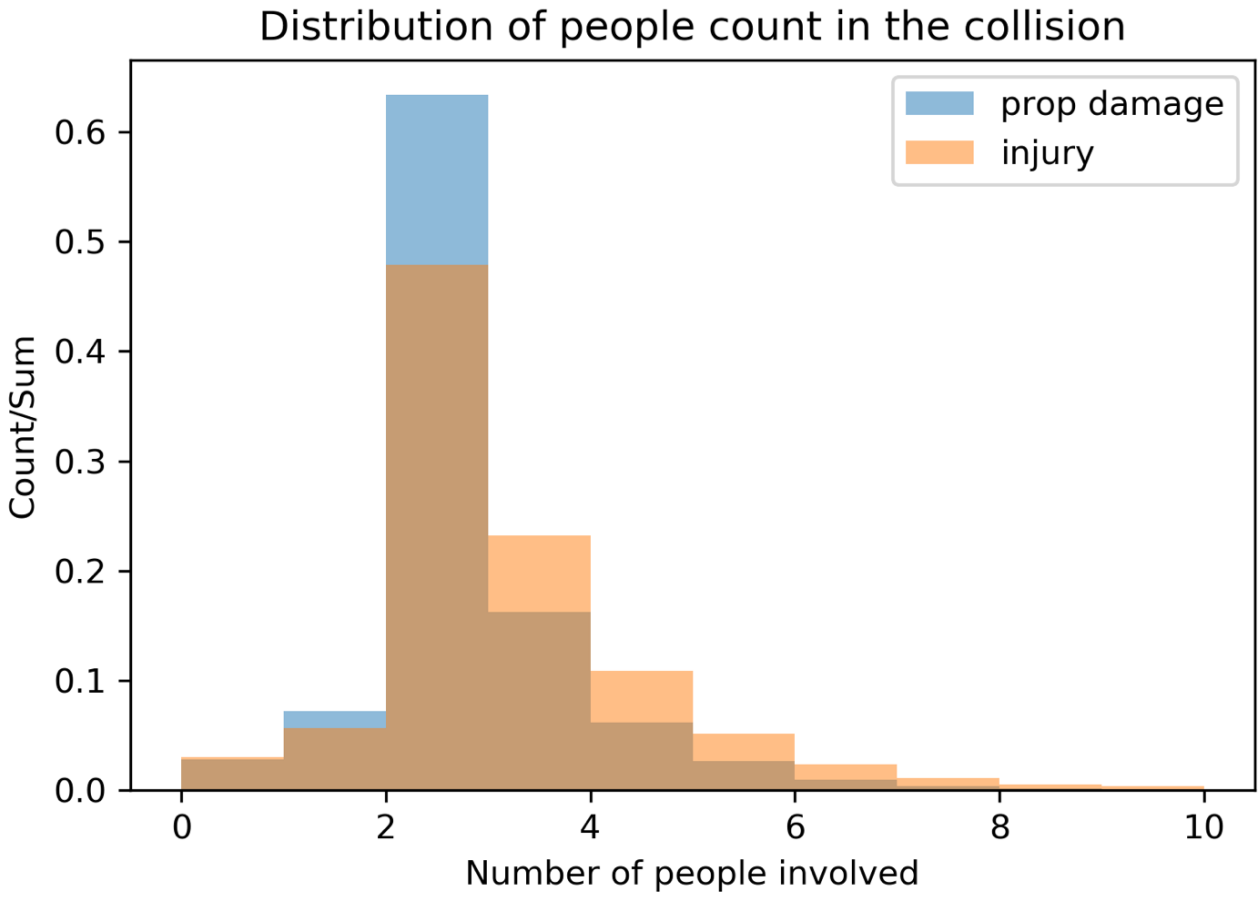
3.1.1 COLLISIONTYPE

There are 10 types of collision types recorded in the data. From the data we shown below, we can see that Rear ended, Angles and other can be three main factor caused the injured collision. Parked car collision have a high potential to caused a prop damaged collision.

SEVERITYCODE	COLLISIONTYPE	
1	Parked Car	0.341761
	Angles	0.158722
	Rear Ended	0.146424
	Other	0.132640
	Sideswipe	0.121420
	Left Turn	0.062524
	Right Turn	0.017697
	Head On	0.008686
	Pedestrian	0.005067
	Cycles	0.005059
2	Rear Ended	0.256724
	Angles	0.238403
	Other	0.106952
	Pedestrian	0.103872
	Left Turn	0.094686
	Cycles	0.083014
	Parked Car	0.046582
	Sideswipe	0.043852
	Head On	0.015259
	Right Turn	0.010657

3.1.2 PERSONCOUNT

We can see from the figure shown below, when accident with involved person number of 3 contributes to major accidents. For more serious accidents, high number of people involved in the accident showed a higher probability caused a more serious accident.



3.1.3 JUNCTIONTYPE

Here we find that the top 3 junction types where prop damaged accidents and injured accidents accurred are the same to each other and here we can assumed that junction type have a small relations to the severity of collision.

3.1.4 INATTENTIONIND

For most cases caused by inattention, it has a higher probability to cause a more serious accidents.

3.1.5 WEATHER

It is very clearly to see that when the weaher is clear, smoke, rain, it has a higher potential to cause a injured collision.

3.1.6 ROADCOND

When there're ice and snow, the accident will prefer to be prop damaged.

3.1.7 LIGHTCOND

During the daylight, there will be a higher potential to have a serious accident.

3.1.8 SPEEDING

A speeding will cause a more serious accidents.

Therefore, 7 features are used to predict the severity of the collisions.

3.2 Model: K nearest neighbors algorithms

In our problems, it is a surpervised machine learning models with labelled data. Motivation of the project is to predict the severity of collision. We can determine this problem to be a classification problem. Based on the features, we can predict the collision severity. Here we choose K nearesr model.

The k-nearest neighbors algorithm (k-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

3.3 Data preprocessing

We choose 7 features which are:

1. COLLISIONTYPE:Collision type
2. PERSONCOUNT: The total number of people involved in the collisions.
3. INATTENTIONIND: Whether or not the collision was caused by the inattention
4. WEATHER: Weather condition
5. ROADCOND: Road condition during the collision.
6. LIGHTCOND: Light condition during the collision.
7. SPEEDING: Whether or not speeding was a factor in the collision.

We then separate the features into columns so that can be used to calculate:

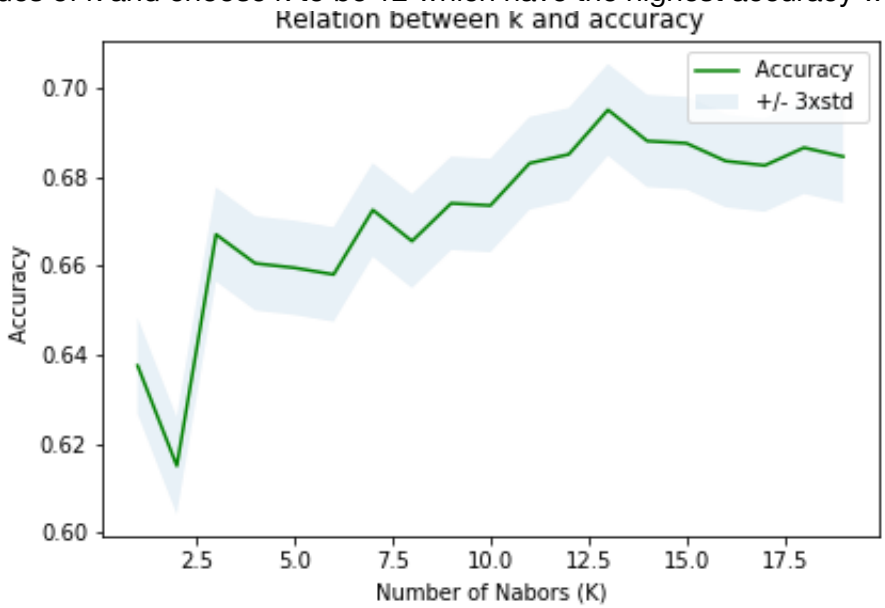
```
t_Feature = tr[['PERSONCOUNT','INATTENTIONIND','SPEEDING']]
t_Feature = pd.concat([t_Feature,pd.get_dummies(tr['COLLISIONTYPE'])], axis=1)
t_Feature = pd.concat([t_Feature,pd.get_dummies(tr['WEATHER'])], axis=1)
t_Feature = pd.concat([t_Feature,pd.get_dummies(tr['ROADCOND'])], axis=1)
t_Feature = pd.concat([t_Feature,pd.get_dummies(tr['LIGHTCOND'])], axis=1)
```

We also normalize each values in the features as shown below:

```
: from sklearn import preprocessing
  t_X = t_Feature
  t_y = tr['SEVERITYCODE'].values
  t_X= preprocessing.StandardScaler().fit(t_X).transform(t_X)
```

3.4 Number of k selection

We tried different values of k and choose k to be 12 which have the highest accuracy which is 0.69.



4 Result and Discussion

Once the model is trained we can obtain a model using 7 features to predict the severity of collision. To evaluate our results, we use the rest data to test our model.

```
from sklearn.metrics import f1_score
from sklearn.metrics import jaccard_similarity_score
f1n = f1_score(y_test,yhat,average='weighted')
jan = jaccard_similarity_score(y_test,yhat)
```

And the evaluation is shown below:

Algorithm	Jaccard	F1-score
KNN(k=12)	0.685	0.6849747983870969

As we known, when f1-core and Jaccard values close to 1 is the better. I think the value around 0.7 is good. It shows that using 7 features as I have mentioned above can predict the severity of the collision with an accuracy of 0.69.

5 Conclusion

We have developed an K nearest neighbor model to predict the severity of the collisions. We used 7 features which are highly related to the severity of the collisions, including: collision type, person involved, reason caused the collision, weather, road condition and driving condition. It appears that when the driving condition is good, it will cause a more serious accidents. The reason might be when the weather is good and road condition is good, people will be more careless. Here we recommend that the drivers should always be careful to save people from accidents.