IBM Capstone Project: Prediction Model of Accident Severity

Chiawen Kao

September 5, 2020

1. Introduction

1.1 Background

In recent news, several car accidents were causing severe traffic jams. Whether or not people involved in those accidents were unprepared and affected by those accidents, such as cancellation of their important meetings, holiday plan, etc. Therefore, the prediction model of accident severity will expect to be a beneficial topic to study and implement if possible.

1.2 Problem

As accidents generally happen in a rush, people have no choice but to deal with those unexpected incidents without any preparation. If there is a machine learning model can predict and warn people, then people can drive more carefully or change their travel plan beforehand instead of rushing into or being affected by some accidents. Therefore, the objective of this report is to avoid the impact of traffic when any accident happens in future.

1.3 Interest

Drivers and commuters would be very interested in an accurate prediction model of accident severity. With the help of the model, they are supposed to prevent themselves from getting involved in or affected by any possible accidents. They can pay the most careful attention if the possibility of an accident is particularly high someday.

2. Data Acquisition and Cleaning

2.1 Data sources

I choose to use the shared dataset in this IBM Capstone Project[1] because the data with many observations will be really beneficial for predicting accident severity. I plan to find the relationship between severity and other factors. The data status is as below:

- Label: “Severity Code” column for accident severity.

[1] Data source link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv provided by IBM Data Science Course at Coursera

- Features: other 37 features relate to previous accidents for predicting accident severity.

2.2 Methodology

Following the process of CRISP-DM, I will clean and explore the data and then determine to use which features for training my machine learning model. After I fully access those chosen features, I will start the process of modeling and evaluation by using the train-test split method and other accuracy measurement metrics. Finally, I will propose my solution to avoid the impact of traffic when any accident happens in future.

2.3 Data Cleaning

During the process of data cleaning, 20 features are removed from my dataset for four reasons summarized in Table 1:

Table 1. Data Cleaning Outcome

No.	Reason for Dropping	Dropped Features	Details
1	Unique information	OBJECTID, INCKEY, COLDETKEY, INTKEY, REPORTNO	Identifier for only a case
		INCDATE, INCDTTM	Past date/time for a case
2	Repeated information	SEVERITYCODE.1	Same as SEVERITYCODE
3	Skewed information	EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, SPEEDING	These would not be valued features, for over 50% of NA values in each feature.
		UNDERINFL, X, Y,	Meaningless information
4	Weak correlation	SDOT_COLDESC, ST_COLDESC, LOCATION	They have very weak correlations with severity. They can also be replaced by other features, including SDOT_COLCODE, ST_COLCODE.

2.4 Feature Selection

As stated in Table 2, only 17 out of 37 features are kept in the dataset after data is cleaned, and meaningful categorical data are all transformed into numerical data for further analysis.

Table 2. Summary of Simple Feature Selection and Transformation

No.	Selected Features	Data Type	Sample Qty	Data Transformation
1	STATUS	Categorical	194,673	Matched = 1, Unmatched = 0
2	ADDRTYPE	Categorical	192,747	Block = 1, Intersection = 2, Alley=3
3	SEVERITYDESC	Categorical	194,673	Property Damage Only Collision = 1, Injury Collision = 2
4	COLLISIONTYPE	Categorical	189,769	Parked Car =1, Angles=2, Rear Ended=3, Other= 4, Sideswipe=5, Left Turn=6, Right Turn=7, Head On=8, Pedestrian=9, Cycles=10
5	PERSONCOUNT	Numerical	194,673	Not applicable
6	PEDCOUNT	Numerical	194,673	Not applicable
7	PEDCYLCOUNT	Numerical	194,673	Not applicable
8	VEHCOUNT	Numerical	194,673	Not applicable
9	JUNCTIONTYPE	Categorical	194,673	At Intersection (intersection related) =1, Mid-Block (not related to intersection) =2, Mid-Block (but intersection related) =3, Driveway Junction=4, At Intersection (but not related to intersection) =5, Ramp Junction=6, Unknown=7
10	SDOT_COLCODE	Numerical	188,344	Not applicable
11	WEATHER	Categorical	189,769	Good weather (clear = 1), Nothing to do with weather (unknown and other = 0), Bad weather (the rest of samples = 2)
12	ROADCOND	Categorical	189,592	Dry=1, Wet=2, Unknown=3, Ice=4, Snow/Slush=5, Other=6, Standing Water=7, Sand/Mud/Dirt=8, Oil=9
13	LIGHTCOND	Categorical	189,503	Daylight=1, Dark - Street Lights On=2, Unknown=3, Dusk=4, Dawn=5, Dark - No Street Lights=6, Dark - Street Lights Off=7, Other=8, Dark - Unknown Lighting=9
14	ST_COLCODE	Categorical	189,769	Transform original sting number to float
15	SEGLANEKEY	Numerical	194,673	Not applicable

16	CROSSWALKKEY	Numerical	194,673	Not applicable
17	HITPARKEDCAR	Categorical	194,673	N=0, Y=1

In addition, you may have found that some of features have different sample quantities. Indeed, the complete sample quantity for each feature should be 194,673. Let's take the feature "SDOT_COLCODE" that has the least sample quantity as an example. Its existing sample quantity still has around 97% completeness compared to 194,673. Therefore, I just use the mean of each incomplete feature to replace original missing values so that further measurement can be done.

3. Exploratory Data Analysis

After I have accessed the condition of those chosen features by examining patterns, correlations and skewed information, I determine to keep only 8 features (see Table 3) that have over 0.15 positive/negative correlation for further phase.

Table 3. Selected Features for Modeling

No.	Selected Features	Feature Description[2]	Data Type	Sample Qty	Correlation
1	ADDRTYPE	Collision address type: Alley, Block or Intersection	Numerical	194,673	0.1854
2	SEVERITYDESC	A detailed description of the severity of the collision	Numerical	194,673	1.000
3	COLLISIONTYPE	Collision type	Numerical	194,673	0.3069
4	PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.	Numerical	194,673	0.2463
5	PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.	Numerical	194,673	0.2142
6	SDOT_COLCODE	A code given to the collision by SDOT.	Numerical	194,673	-0.1630
7	ST_COLCODE	A code provided by the state that describes the collision.	Numerical	194,673	-0.1630
8	CROSSWALKKEY	A key for the crosswalk at which the collision occurred.	Numerical	194,673	0.1751

[2] ArcGIS Metadata Form provided by IBM Data Science Course at Coursera

4. Predictive Modeling

Classification model will be my first choice to build a prediction model of accident severity as classification models can predict discrete class labels, find patterns and groupings. In the shared dataset, all accidents are classified into two levels of severity: 1 and 2, which should be the best fit for classification model. In addition, I will also try regression model, which is used to predict a continuous trend and variables, may receive some additional insights.

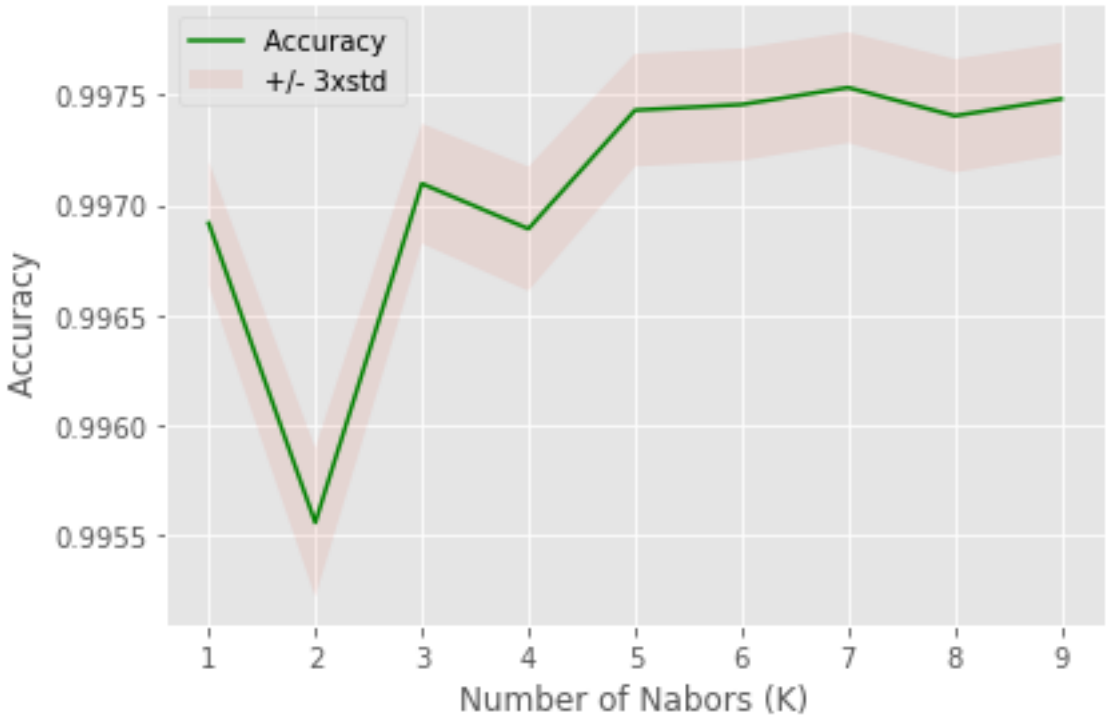
4.1 Classification models

With out-of-sample train and test data, I will then apply algorithms of K Nearest Neighbor (KNN), Decision Tree, Support Vector Machine (SVM) and Logistic Regression to the processed dataset, and then evaluate their accuracy by the measurement metrics, including Jaccard, F1-score and Log loss.

4.1.1 Built Models

First, I used train-test split function to split the processed dataset, which split the whole data with 194,673 samples into random train and test subsets. I had applied KNN algorithms to the train dataset and found that the best K with the best accuracy will be 7 (see Figure 1). Therefore, I built the KNN model with K=7.

Figure 1. Accuracy Table of Different K in KNN



Then, I applied the standard algorithms of Decision Tree, SVM and Logistic Regression to build the processed dataset for further evaluation.

4.1.2 Model Evaluation Results and Performance

With all candidate models built, I would then figure out how well those models performed by using Jaccard, F1-score and Log loss scores. Let's look at Jaccard score first. If the entire predicted labels for a sample perfectly match the true labels, the Jaccard accuracy score will be 1.0. In comparison with all other candidates, as you can see in Figure 2, SVM is the best one for prediction.

Figure 2. Accuracy Score of Candidate Models

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.92	0.91	NA
Decision Tree	0.92	0.91	NA
SVM	1.00	1.00	NA
LogisticRegression	0.92	0.91	0.35

As for F1-score, if a model makes perfect prediction and recalls the target labels, then F1-score will be 1.0 after calculation. If the F1 score is near 0, the more near it reaches 0 the worst precision the model is made. SVM is still in the lead and make the best precision among all the candidates. Therefore, SVM should be the best classifier for the prediction model of accident severity.

The last metric is Log loss only applying to Logistic Regression, which output can also indicate probability. In other words, Log loss can calculate the precision between a true label and a probability value. With the candidate Logistic Regression model, its Log loss outcome does not seem to be positive. This could be a topic to take a deep look and investigation.

4.2 Regression Models and Results

Next, I applied multiple linear regression algorithm with a common metric: explained variance regression score. The explained variance, the square of the standard deviation, regression score can reveal the accuracy of a regression model. If a model reaches a score of 1.0, then it will be the best possible solution ever had. The more a model is near a score of 0, the worse it will be. With several rounds of validation, I found out very different results as shown in Figure 3. When

I used all of eight chosen features, then the explained variance score was very close to the result that I had done in classification models (see section 4.1). Then, I experimented on different feature combination and data indicated that I could use only the 'SEVERITYDESC' feature to reach the highest explained variance score in simple linear regression model. For further investigations, I tried to use different feature groups with higher correlation scores without the 'SEVERITYDESC' feature, the explained variance scores decreased drastically and unexpectedly. For instance, the feature group with correlation score over 0.2 turned out to be worse prediction model than other feature groups, including groups with correlation score over 0.15 and 0.13. However, no matter feature group had higher correlation score or a bit lower or not, they all received relatively very low explained variance score. This may mean that this processed shared dataset is suitable for simple linear regression model with the crucial feature 'SEVERITYDESC' and other features are not that meaningful as 'SEVERITYDESC' does.

Figure 3. Accuracy Score of Candidate Models

Features	Explained Variance
All	1.0000
Corr over 0.2 w/o SEVERITYDESC	0.1247
Corr over 0.15 w/o SEVERITYDESC	0.1536
Corr over 0.13 w/o SEVERITYDESC	0.1650

In addition to simple and multiple linear regression models, Polynomial regression may be another way to solve the situation. Therefore, I built a Polynomial regression model for all of eight chosen dataset, and its R-squared score is -0.14, which indicates very low correlation even with 'SEVERITYDESC'. That is, in regression models, simple linear regression will be a better solution instead of multiple linear and polynomial regression.

5. Conclusions

In summary, classification model will be the best possible solution to build a prediction model of accident severity because severity levels have already been classified in original dataset. As for regression models, simple linear regression performs way better than multiple linear and polynomial regression and it also suggests the crucial feature is 'SEVERITYDESC'. Therefore, as long as people hear about 'SEVERITYDESC' information (Property damage only collision or Injury Collision) about any abrupt accident in future, they can immediately guess the severity level of the accident and prevent themselves from being affected by possible traffic jams.

6. Future Directions

In future, there will be two ways to improve the performance of an accident severity prediction model. First, the features studied in this report have relatively lower correlation with severity level, except 'SEVERITYDESC'. Therefore, if you would like to do supervise learning on the said prediction model, you should gather much more new features that have stronger correlation with accident severity. Last, various features that are commonly supposed to have strong connection with accident severity prediction, such as weather, light condition or road condition, actually have very weak relationships with accident severity in the machine learning results. This means that the prediction model of accident severity may require some unknown and unexpected features that people can never come up with. Therefore, unsupervised learning and reinforcement learning would be possible directions for further study.