IBM Capstone Project

Introduction

Car accidents costs life and huge economic loss every year in the US. In this project, I attempt to predict road accidents before it happens, so the drivers can take precautions to avoid traffic jam due to accidents or even prevent accidents from happening.

Data

The data set used in this project came from SDOT GIS Seattle(Data Set, Meta Data). It contains the speed, light, road condition, severity, etc. for the past road accidents. The idea is to use several supervised machine learning techniques to predict the severity given the various road conditions.

Figure1. Example data

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	S
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0	

For the scope of this project, I'll look at how "SPEEDING", "ROADCOND", "LIGHTCOND", "WEATHER" parameters affect "SEVERITYCODE". Keep only weather in (Clear, Raining, Overcast), road condition in (Dry, Wet) and light condition in (Daylight, Dark - Street Lights On) so that the training set will not be too skewed.

Figure 2. Data with selected attributes.

	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER	SPEEDING
0	2	Wet	Daylight	Overcast	N
1	1	Wet	Dark - Street Lights On	Raining	N
2	1	Dry	Daylight	Overcast	N
3	1	Dry	Daylight	Clear	N
4	2	Wet	Daylight	Raining	N

Methodology

Since this is a binary classification problem (severity code= 1 or 2), I'll use K-nearest neighbors and Logistic regression techniques. KNN is chosen because its performance on dealing with a large set of data. I also chose Logistic Regression because it provides the probability for detecting accidents.

To begin with, convert the attribute labels to numerical values, and then 5000 samples from each severity label were randomly picked.

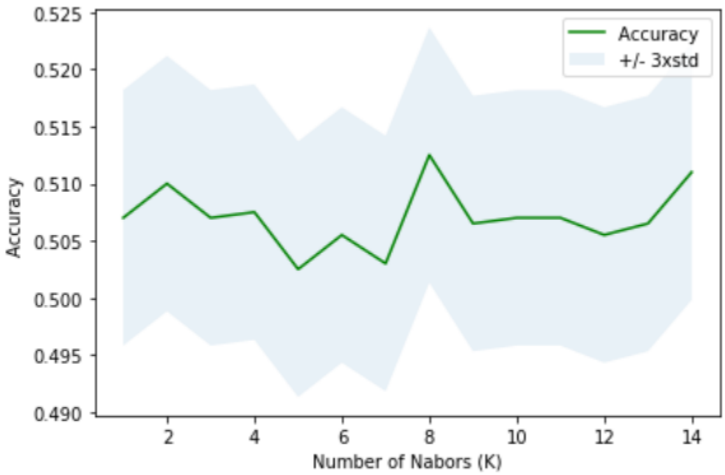
Figure 3. Numerically converted sample data

	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER	SPEEDING
0	2	1	0	2	0
1	1	1	1	1	0
2	1	0	0	2	0
3	1	0	0	0	0
4	2	1	0	1	0

KNN

For KNN, k=8 was found to give the best accuracy = 0.5125.

Figure 4. Accuracy vs k for KNN model



Logistic Regression

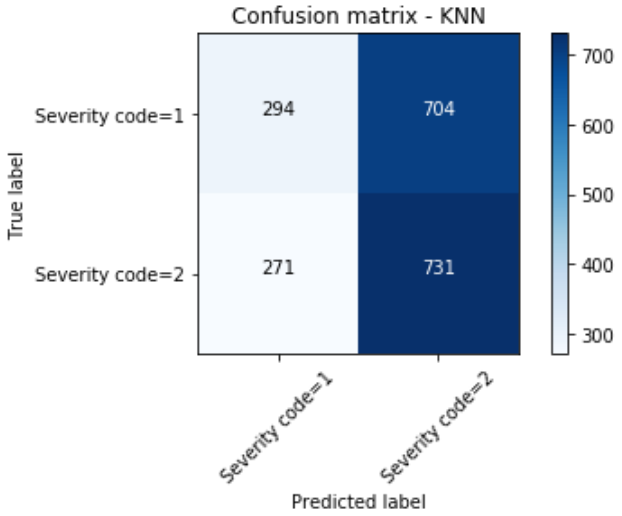
For Logistic regression, the regulation coefficient was chosen to be c=0.001, which yields a Jaccard similarity score of 0.5055.

Results

For results, confusion matrices were plotted for both KNN and LR models.

KNN

Figure 5. Confusion matrix for KNN



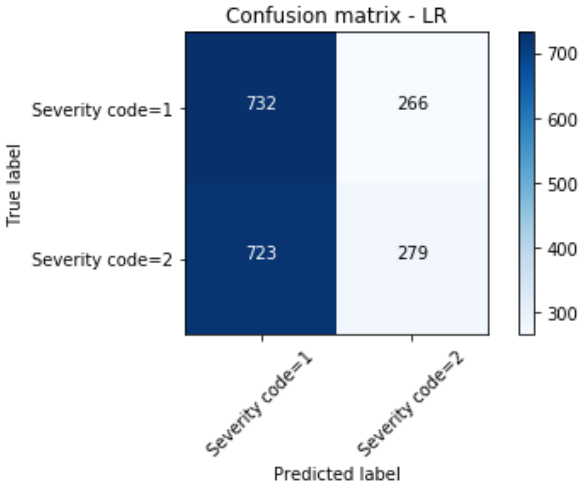
The classification report for KNN follows.

Figure 6. Classification report for KNN

	precision	recall	f1-score	support
1	0.52	0.29	0.38	998
2	0.51	0.73	0.60	1002
micro avg	0.51	0.51	0.51	2000
macro avg	0.51	0.51	0.49	2000
weighted avg	0.51	0.51	0.49	2000

Logistic Regression

Figure 7. Confusion matrix for Logistic Regression



The classification report for KNN follows.

Figure 8. Classification report for KNN

	precision	recall	f1-score	support
1	0.50	0.73	0.60	998
2	0.51	0.28	0.36	1002
micro avg	0.51	0.51	0.51	2000
macro avg	0.51	0.51	0.48	2000
weighted avg	0.51	0.51	0.48	2000

Discussion

As shown by the KNN and LR confusion matrices, the KNN model biases toward severity code = 2 while the LR model biases toward severity code = 1. It appears that with only selected road condition, light condition, weather, and speeding information is not enough for predicting the severity of accidents. A future direction will be to investigate the effects from location and day of the week on the severity.

Conclusion

This project is an exercise to apply ML techniques on predicting car accident severity using the SDOT GIS Seattle data set. K-Nearest Neighbor and Logistic Regression classification techniques were used to predict the severity of the accident with recorded attributes, including road condition, light condition, weather, and speeding. Each technique is biased against one severity. To improve the models, attributes like location and day of the week can be further investigated.