Predicting the severity of car accidents

Business Problem

- Car accidents are a big modern society problem and cost millions of dollars to government and affect people's lives;
- In 2017, Seattle police reported 10,959 motor vehicle collisions on city streets;
- In 2017 there were a total of 187 fatal and serious injury collisions on Seattle streets;

Data Source

- The dataset was provided by The Seattle Department of Transportation's here and the metadata was founded here. The dataset has 38 columns (37 possible features) and 194,673 rows.

Methodology

- The dataset has lots of missing values throughout the whole dataset
- According to the metadata some features are related to coordinates, address, and other unique keys thar government agencies uses to identify each accident. Though other features like road condition, light condition and weather look promising;

Feature	Correlation
Road Condition (ROADCOND)	0.027938
Light Condition (LIGHTCOND)	0.028834
Weather (WEATHER)	0.027361

Methodology

- Only two of out 5 different labels are found in the dataset (0 and 1);

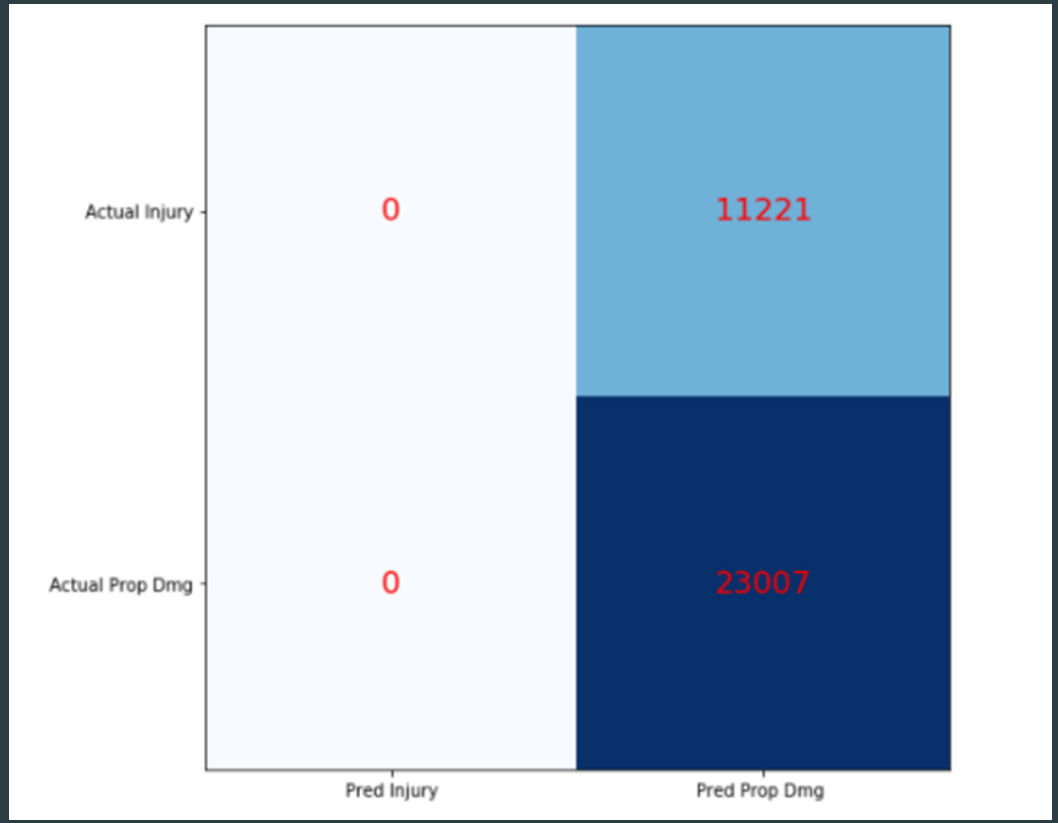
Severity Code	Description
0	Unknown
1	Prop Damage
2	Injury
2b	Serious Injury
3	Fatality

Methodology – Selecting Models

- KNN, Logistic Regression, SVM and Decision Tree were used to find the best one;
- LR show the best results when compared to the others;

Algorithm	Jaccard	F1-Score	Logloss
KNN	0.67	0.54	NA
Decision Tree	0.67	0.54	NA
SVM	0.55	0.55	NA
Logistic Regression	0.67	0.54	0.63

Results


Actual Injury
Actual Prop Dmg
0
11221
0
23007
Pred Injury
Pred Prop Dmg

Results

- 100% of Injury were classified as Prop Damage besides the train and test data had Injury cases;

	precision	recall	f1-score	support
injury	0.00	0.00	0.00	11221
prop damage	0.67	1.00	0.80	23007
micro avg	0.67	0.67	0.67	34228
macro avg	0.34	0.50	0.40	34228
weighted avg	0.45	0.67	0.54	34228

Conclusion

- Model is not optimal enough to delivery a good result that can provide information to the people and government agencies;
- To improve the model and implement an optimal solution we need more data with all labels instead 2 out of 5;
- More precise data to the ones that we already have (less missing values);