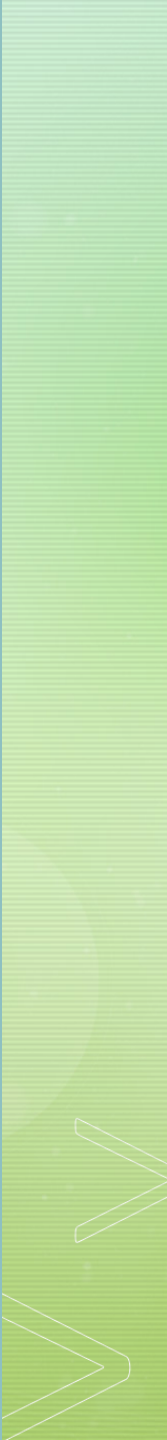


Prediction of Accident Severity by using ML

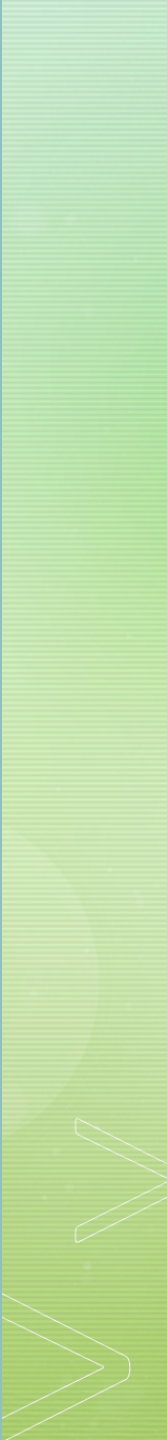


Introduction

- Cars are a big part of our daily life.
 - By 2019, there were a car per two persons in the most of the developed countries while in some of them it is even more.
 - There are more and more cars on the roads as the demand increases.
- 



Business Problem

- With the increase of cars on the roads, the accident probability increases.
 - Prediction of car accident severity is important to decrease deaths, injuries and damage to vehicles.
 - Drivers and passengers are the primary stakeholders that would be interested in knowing accident probability to make their transportation plans.
 - Additionally, insurance companies and public authorities would be interested in such a model to avoid loss of money due to damages and hospital costs.
- 

Data & Methodology

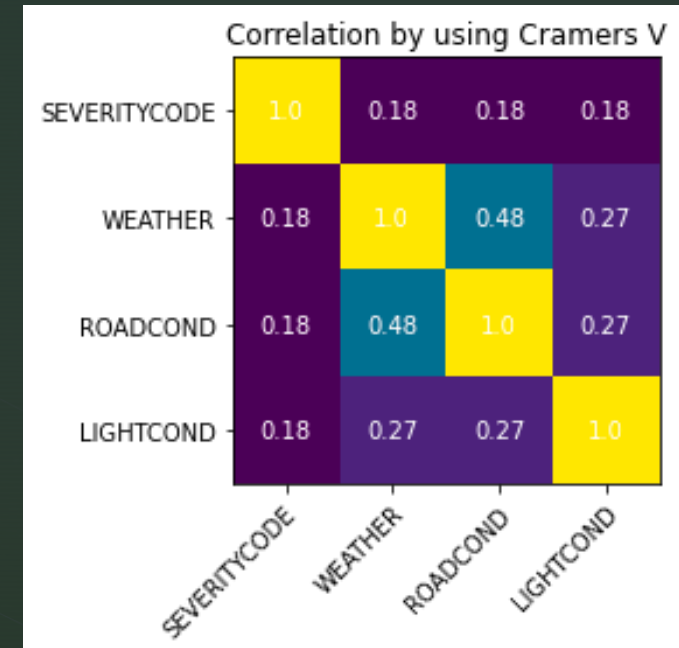
- To solve this problem, the data collected for Seattle city is used.
- The data includes all collisions provided by SPD and recorded by Traffic Records from 2004 to 2019.
- The dataset contains 7397574 entries for accident information with variable name of SEVERITY CODE and some related data such as location, road conditions etc in 38 columns (194673x38 table).
- The prediction will be done as

SEVERITY ~ WEATHER + ROADCOND + LIGHTCOND

- These attributes would mostly available for any time of the year; so, the model would work with their values.

Data & Methodology

- 9333 rows with SPEEDING answered yes, 5126 (entry Y) + 3995 (entry 1) = 9121 rows with UNDERINFL answered yes and 29805 rows with INATTENTIONIND answered yes are excluded to avoid person-related accidents.
- The final data frame has shape of 144089x4 excluding 3 columns (SPEEDING, INATTENTIONIND and UNDERINFL) used in the preparation step and null entries for predatory variables.
- Associations between the predictors between themselves and with target values are explored via using Cramer's V statistic since all variables are categorical.

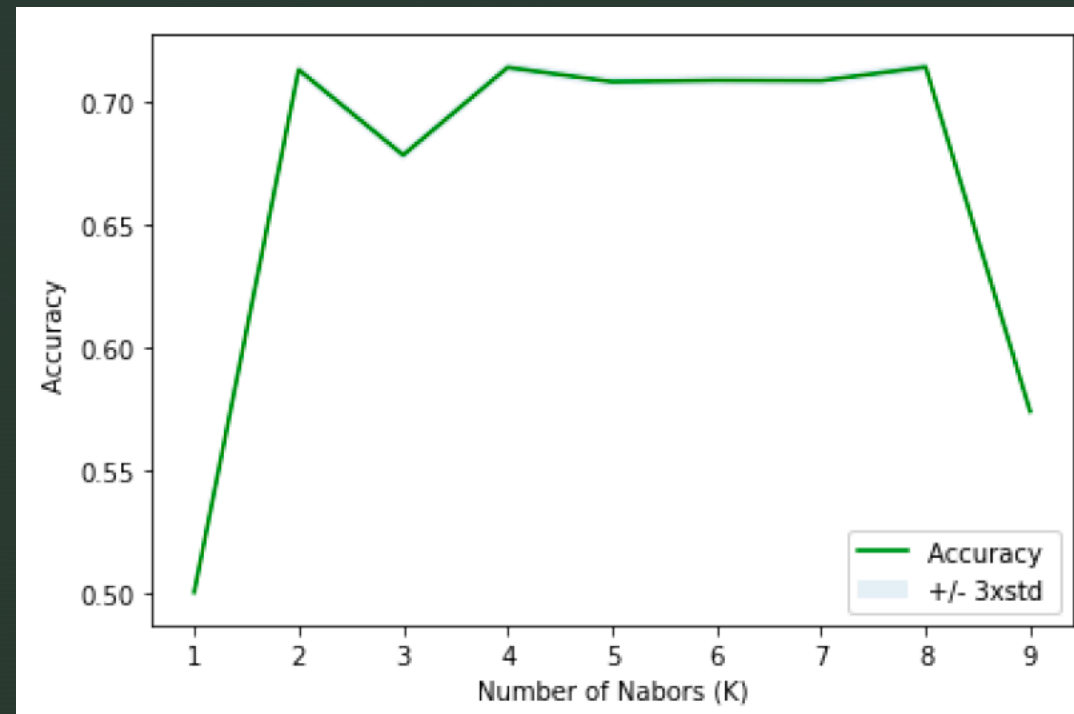


Data & Methodology

- The problem is a classification problem with two different output values for the target variable. Applied methods are:
 - KNN,
 - decision trees,
 - logistic regression and
 - support vector machines.
- Before the method applications the dataset is divided into training (70%) and test sets (30%).
- The predictor variables are converted to dummy variables.

Results

K nearest neighbors method gives best accuracy of 71.5% with $k = 8$.



Results

- All methods are compared by using Jaccard score. There is no method which result in very good accuracy.
- KNN with $k = 8$ is selected as the final model with mean accuracy of about 50%.

	precision	recall	f1-score	support
1	0.72	1.00	0.83	30917
2	0.27	0.00	0.00	12310
accuracy			0.71	43227
macro avg	0.49	0.50	0.42	43227
weighted avg	0.59	0.71	0.60	43227

Conclusion

- The predictive model for accident severity using weather, road and light conditions provides good accuracy.
- The resulted model can be implemented in any algorithm used to calculate insurance rates.
- However, there should be some improvement of the model for low number of more severe accidents and maybe additional predictive variables.