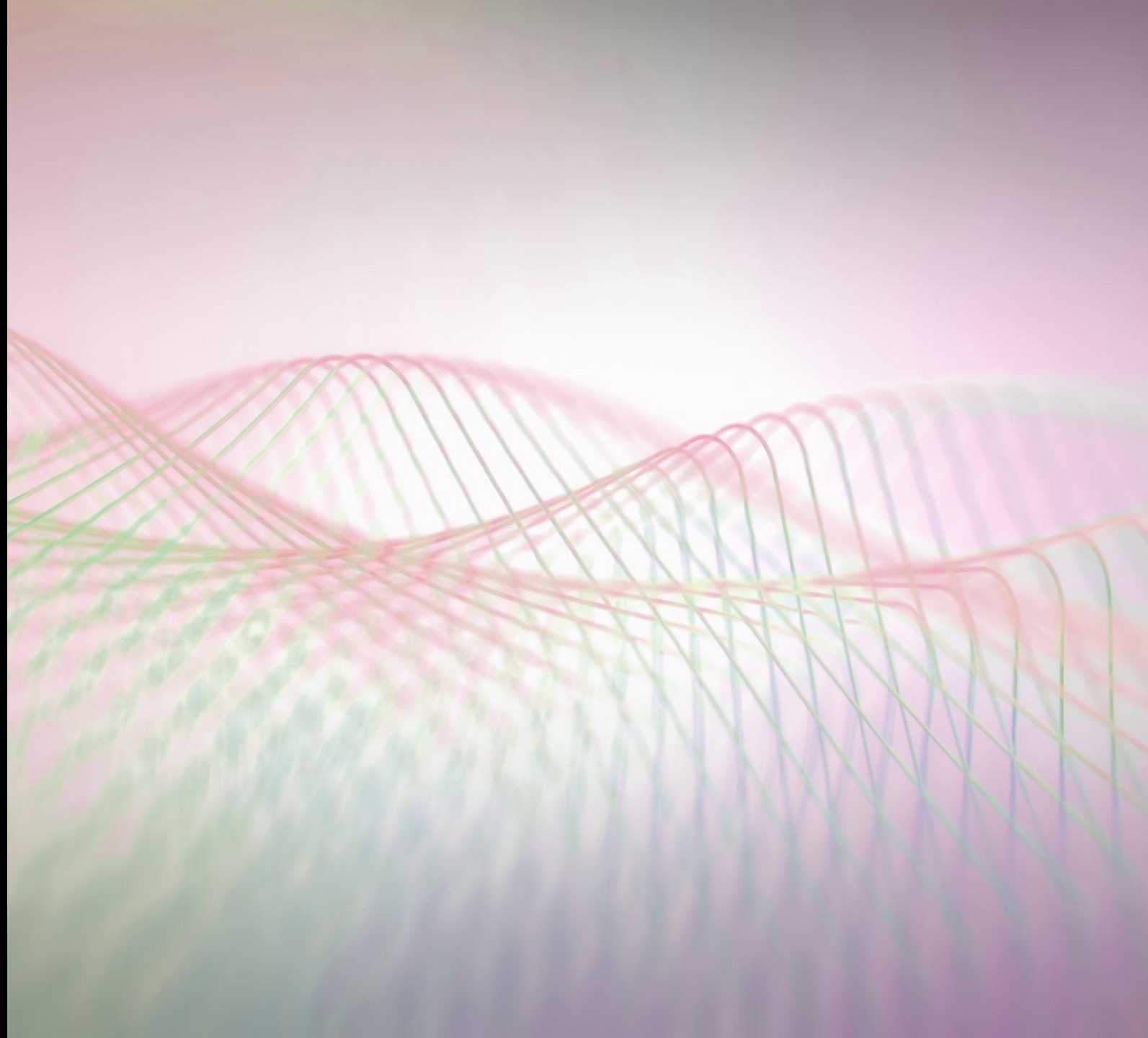
IBM Data Science Capstone Project

Modelling Seattle transportation collision using Classification Machine Learning

Long Phan – Climate Change Data Analyst, SNV Vietnam

September 2020

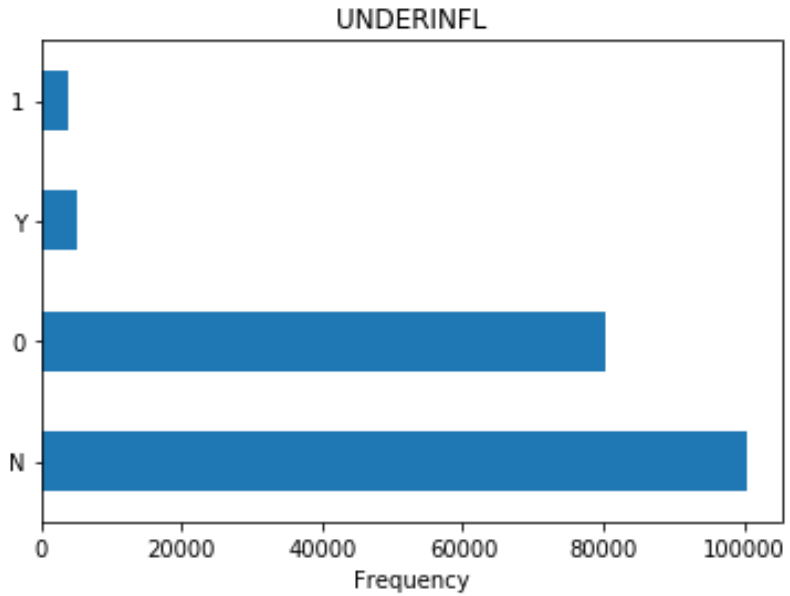
tuanlongphannguyen@gmail.com

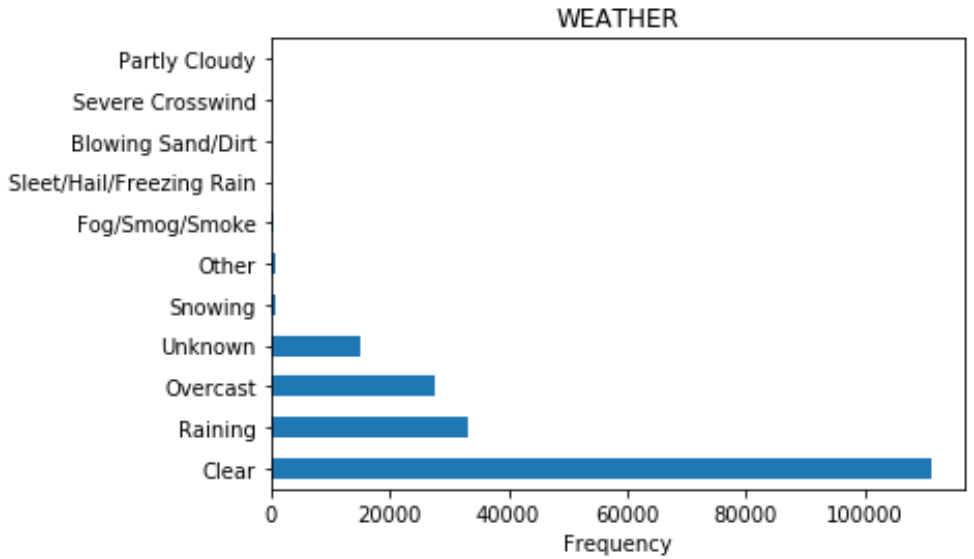


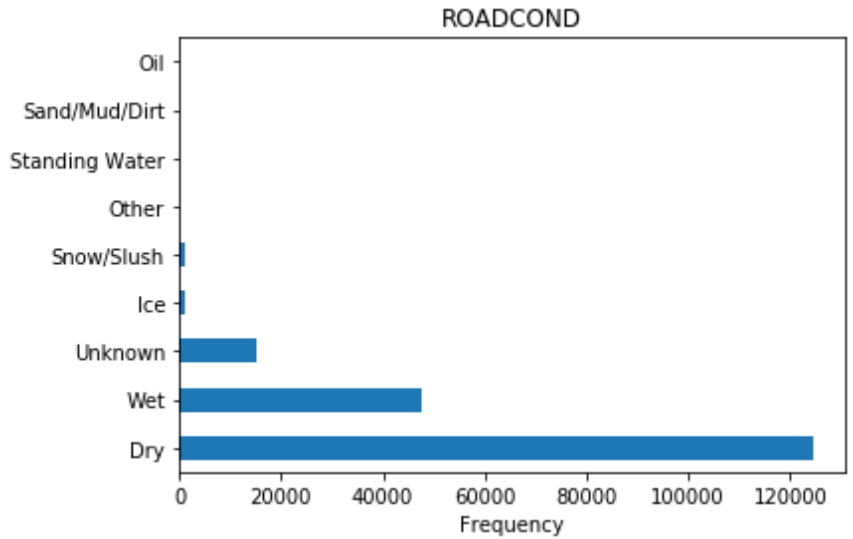
Introduction/ Business Problem

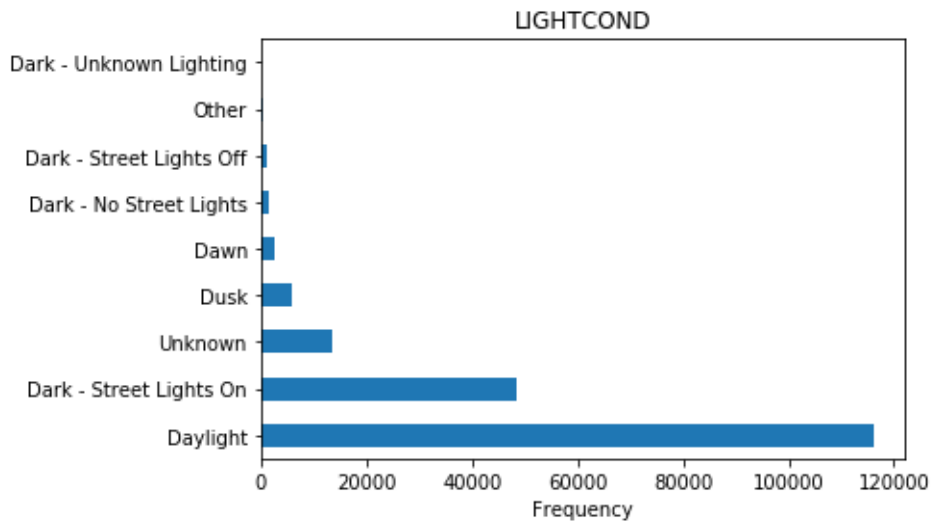
In this capstone project, I will choose to work on a transportation problem in Seattle by analyzing the accident characteristic. Transportation accident is one of the most headache issue in any big cities. This problem, fortunately, thanks to the power of data, can be addressed and mitigated. The goal of the project is to help predict future accidents base on key variables, thus finding the ways to reduce collision rate.
Beneficiaries from this study are policy makers, drivers, and people working in the transportation sector.

Feature selection





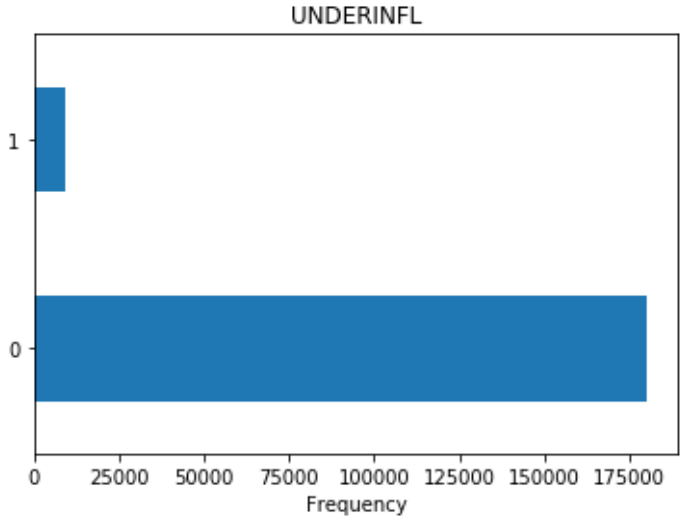


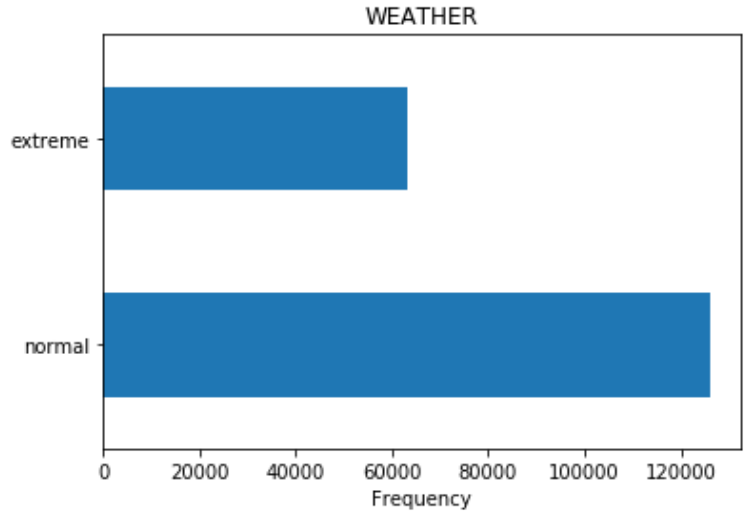


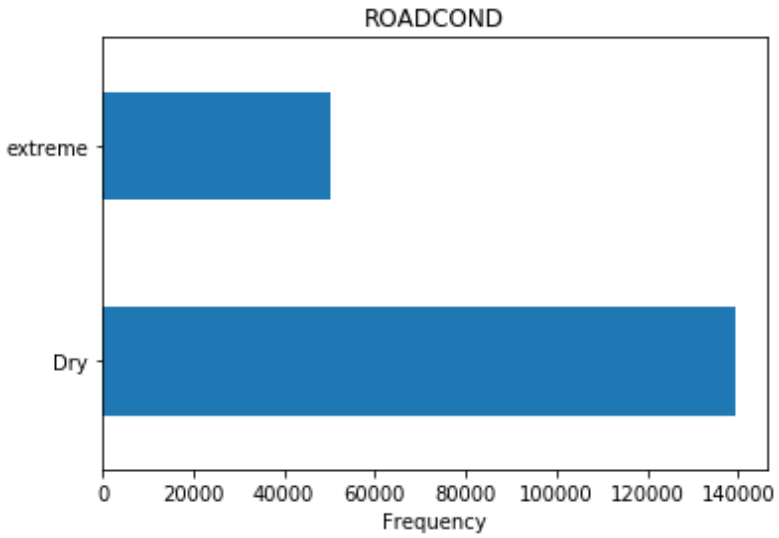
Data relabeling

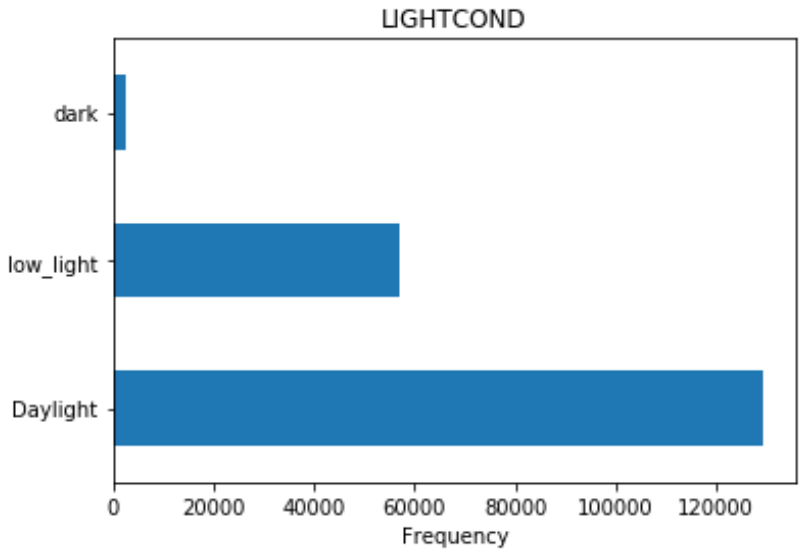
Attributes	Old value	New re-labeled value
UNDERINFL	Y/N	1/0
WEATHER	'Clear', 'Partly Cloudy'	normal
	'Raining', 'Overcast', 'Snowing', 'Other', 'Fog/Smog/Smoke', 'Sleet/Hail/Freezing Rain', 'Blowing Sand/Dirt', 'Severe Crosswind'	extreme
ROADCOND	'Wet', 'Ice', 'Snow/Slush', 'Other', 'Standing Water', 'Sand/Mud/Dirt', 'Oil'	extreme
LIGHTCOND	'Dark - Street Lights On', 'Dusk', 'Dawn', 'Other', 'Dark - Unknown Lighting'	low_light
	'Dark - No Street Lights', 'Dark - Street Lights Off'	dark

Relabeled Attributes







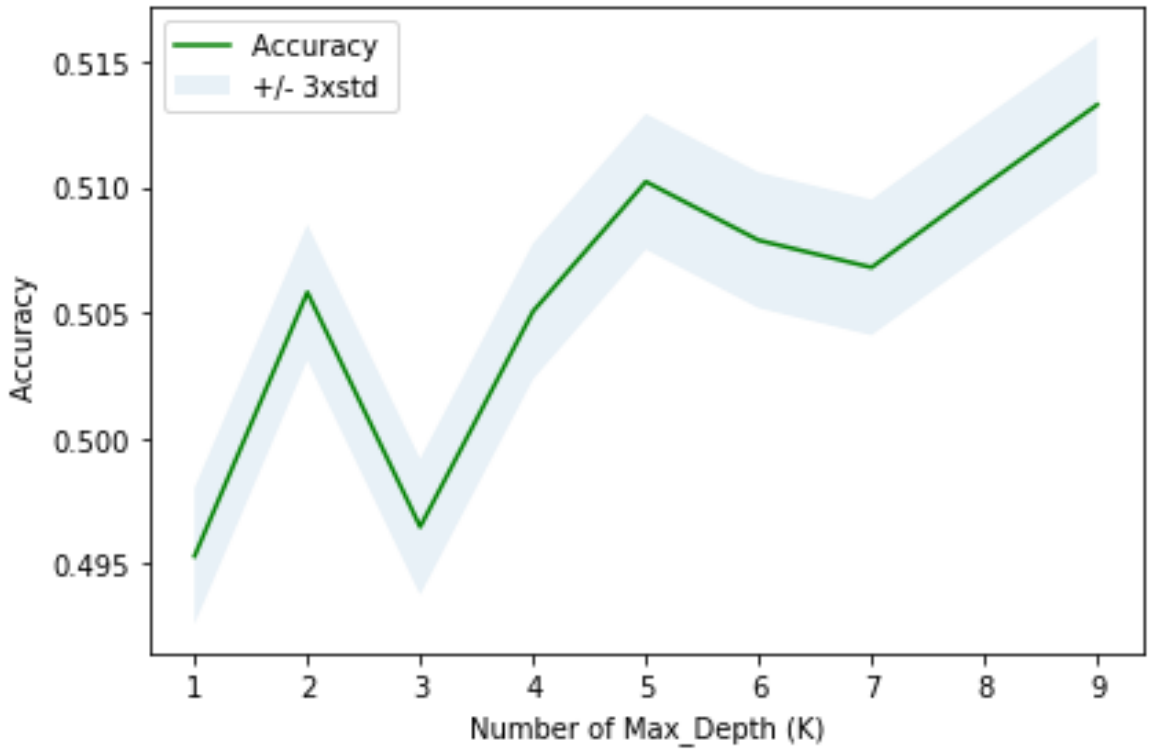


Classification Modeling methods

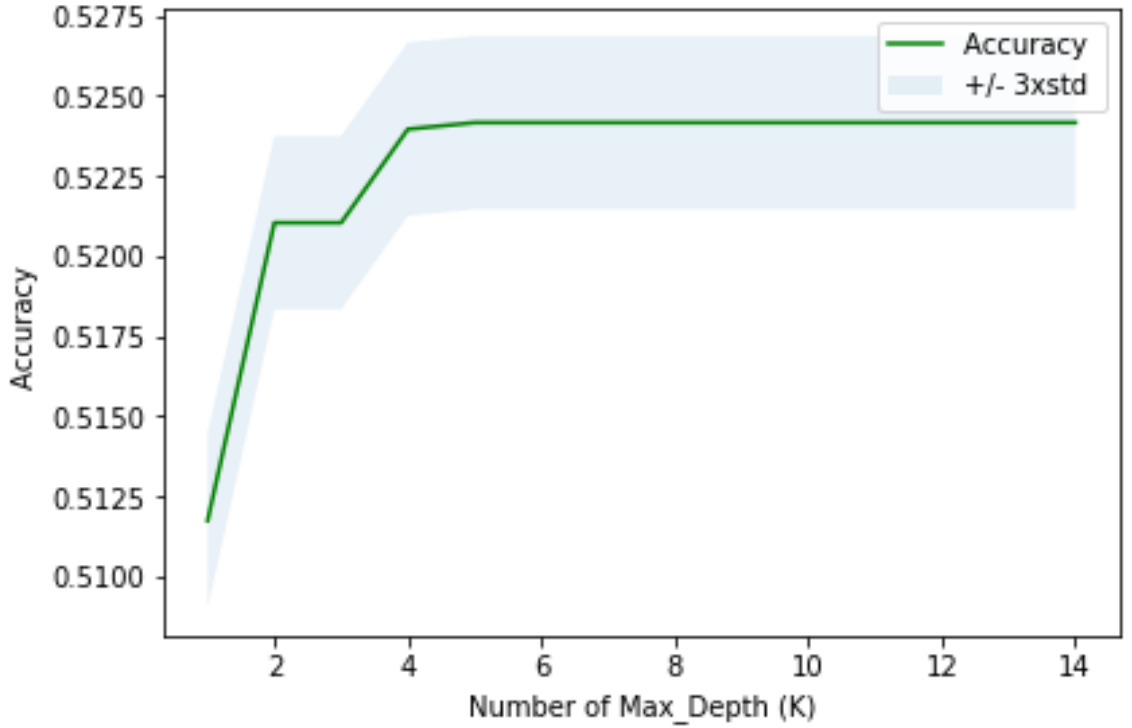
Machine Learning methods	Tuning parameter
K-Nearest Neighbor	K value
Decision Tree	Max depth value
Support Vector Machine	Kernel types ('linear', 'poly', 'rbf')
Logistic Regression	Solver types ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga')

Classification Modeling tuning result

K-Nearest Neighbor (KNN)



Decision Tree



Classification Modeling tuning result

Support Vector Machine

Kernel	F1 score	Jaccard score
Linear	0.5146	0.5235
Poly	0.5142	0.5239
RBF	**0.5145**	**0.5241**

Logistic Regression

Solver	F1 score	Jaccard score	Log loss
newton-cg	0.51467686	0.523516	0.69147211
lbfgs	0.51467686	0.523516	0.69147219
liblinear	0.51467686	0.523516	0.691472038
sag	0.51467686	0.523516	**0.6914720301**
saga	0.51467686	0.523516	0.69147204

Final Evaluation Result

ML model technique	F1 score	Jaccard score	Log loss
K-Nearest Neighbor	0.4827	0.5102	-
Decision Tree	**0.5149**	**0.524**	-
Support Vector Machine	0.5145	0.5242	-
Logistic Regression	**0.5147**	**0.5235**	**0.6914**

Conclusion

The study has successfully determined the best machine learning model technique in the Seattle transportation collision dataset.

It uses Severity code as the response vector (target) and four variables (UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND) as Feature Matrix. By running through the well-tuned models, Decision Tree and Logistic Regression are proved to be the most appropriate techniques and can be use in further analysis and study.

Recommendation

Due to the limit in CPU cores and speed, this study could not carry the largest extractable dataset for modelling. To achieve this highest possible dataset an up-sample method and replacing n/a with highest frequency are needed. This approach may result better accuracy.

Also due to the computer processing speed, other attribute is currently not taking into account. Further studies may need to approach this issue.

Thank you.