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

Table of Contents

1. Introduction/ Business Problem

A description of the problem and a discussion of the background

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.

2. Data information

A description of the data and how it will be used to solve the problem

From the SDOT Traffic Management Division, Traffic Records Group, a 72MB dataset is collected. It has 194673 weekly-updated collision records from 2004 to present, in 37 columns.

The key variable to be use as predicted criteria (y) is Severity code, which has two unique values 1 and 2, representing for prop damage and injury collision severity, respectively. From the dataset, four variables: UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND, will be selected as the input for the models to analyze whether the severity is under the influence of drug/alcohol usage, weather condition, road condition, and/or light condition. Their n/a values, which are categorical, will be replace based on the highest frequency. Values under these selected variables will be transform into numeric values, and normalized.

Four supervised classification machine learning techniques are used in this study are K-Nearest Neighbor, Decision Tree, Logistic Regression, and Support Vector Machine. The four model results will be evaluated with three evaluation methods: F1-score, Jaccard index, and log loss. From that, the highest-accuracy model will be selected to predict and identify driven factors.

3. Methodology

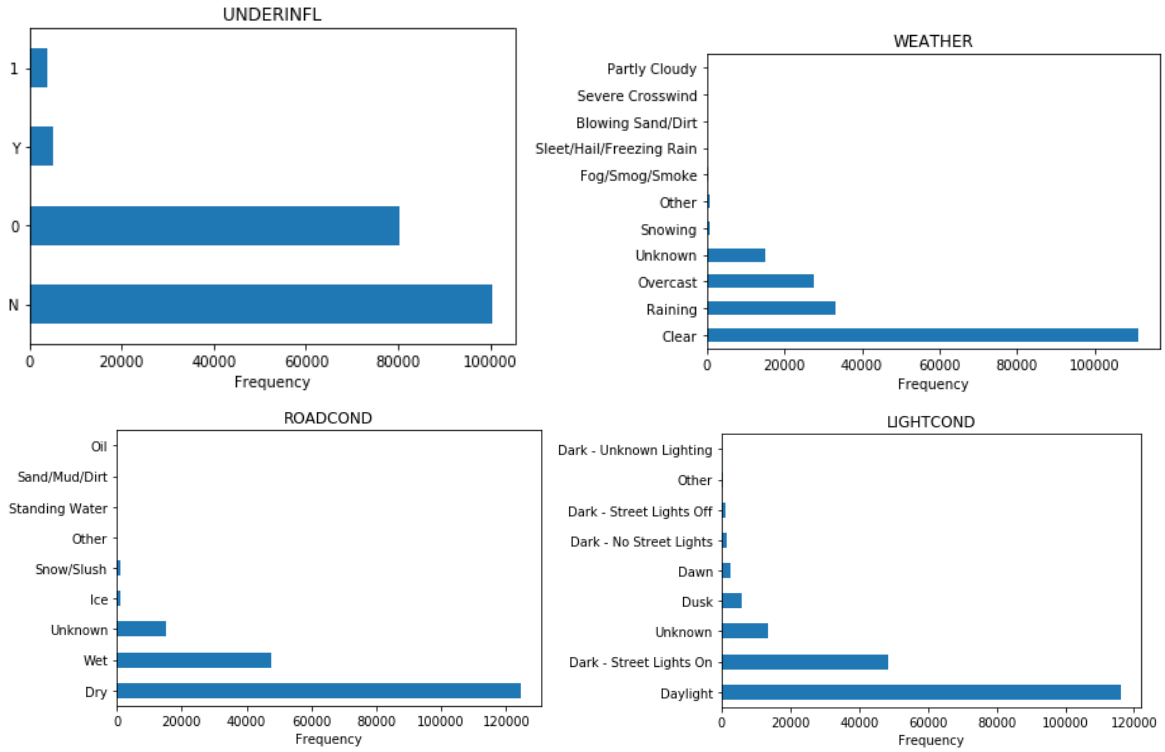
a. Data wrangling

i. Feature selection

Four key variables: 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND' are selected to analyze whether the severity is under the influence of drug/alcohol usage, weather condition, road condition, and/or light condition.

ii. Exploratory analysis

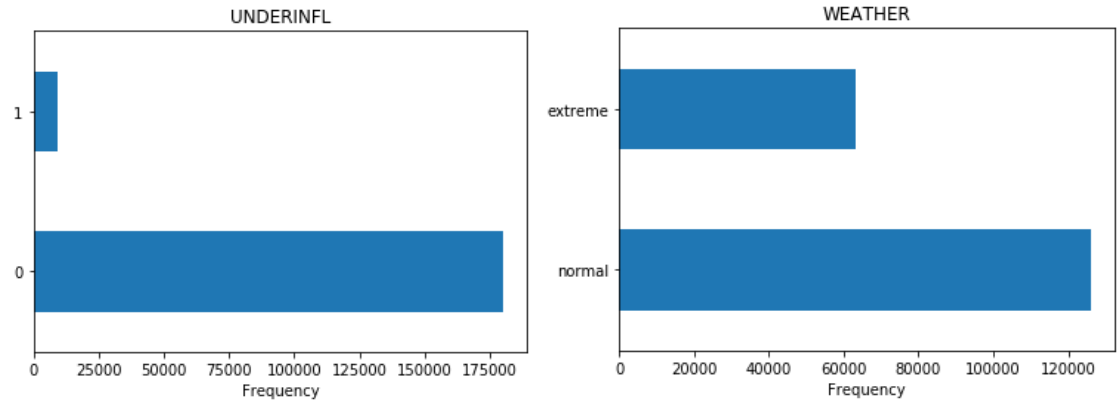
From the four charts below, values in selected attributes are distributed in many categories which can be grouped. In UNDERINFL, the boolean values are represented in two types, both Y(es)/N(o) and 1/0.

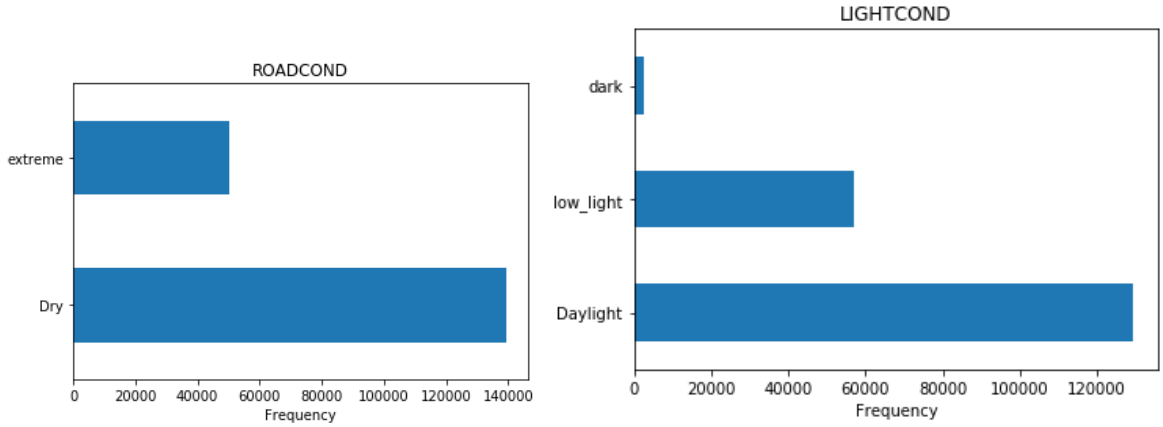


iii. Data re-labeling and solving n/a

Re-labeling

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





Unknown values were replaced by the highest frequency value in each columns, identify by print(df.describe(include='all')) command

Solving n/a
Initial approach was to replace n/a values by the highest frequency values in each attributes, however, due to the limitation in CPU cores and speed which make handling large data difficult, all rows with n/a values were removed by df.dropna(0,'any',inplace=True) command.

b. Pre-modeling

i. Data balancing using down-sample method
Using resample function from sklearn.utils, two sampling methods were consider: Upsample minority class and Downsample majority class. Due to the limitation in CPU cores and speed, this study conducted with downsampled data - at 57052 samples.

ii. Convert categorial values into to numerical ones and Normalizing data
From the preprocessing function of sklearn library, categorial values in the Feature matrix (X) were converted into numerical ones for suitable in further machine learning models.
The data is also normalized using preprocessing.StandardScaler function.

iii. Train/test split
The dataset is split into 7/3 ratio for train and test set using train_test_split function from sklearn.model_selection

c. Modeling

Four supervised classification machine learning techniques K-Nearest Neighbor, Decision Tree, Logistic Regression, and Support Vector Machine are used in this study. For each technique, the most accurate parameter is tuned and selected to ensure the highest model performance for each method.

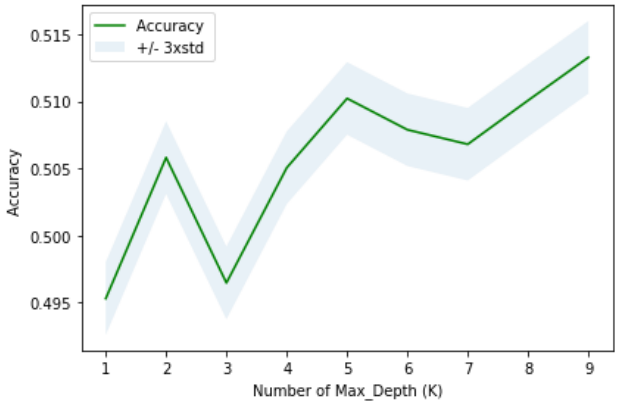
Machine Learning methods	Tuning parameter
K-Nearest Neighbor	K value
Decision Tree	Max depth value
Support Vector Machine	Kernel types ('linear', 'poly', 'rbf')
Logistic Regression	Solver types ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga')

d. Evaluation

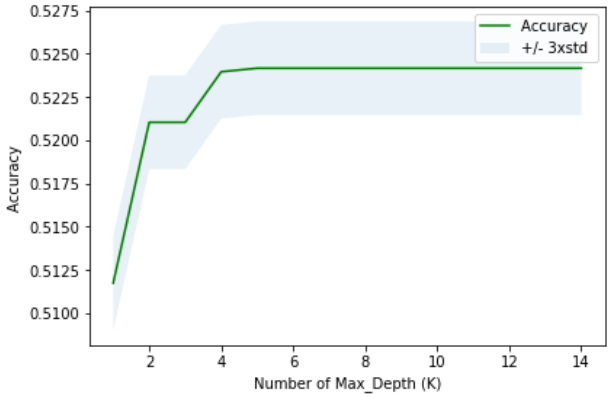
Three evaluation method were used in this study are F1 Score, Jaccard score, and Log Loss.

4. Result and discussion

Each machine learning method has to time-consumingly run through a loop of different parameters in order to find the most accuracy one.
For K-Nearest Neighbor (KNN) methods, after having 10 irrigations, k = 5 is observed having highest accuracy.



The same process applied to Decision Tree ML model to find max depth value = 4 having the highest accuracy, before overfitting.



For Support Vector Machine, RBF-type kernel is observed having the highest accuracy among the three tested kernels, considering both F1 and Jaccard score

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

Regarding Logistic Regression ML model, all solvers perform similarly, except for sag-type solver is slightly more accurate in log loss evaluation, which will be used for the main modelling parameter.

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

Decision Tree and Logistic Regression method have the highest score

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.

5. 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.