Capstone Project - Car Accident Severity

1. Introduction: Problem and Background

Traffic accidents are threats to public safety as they cause injuries to people and damages to properties. Reducing traffic accidents is both meaningful and challenging. Prior research has proposed new frameworks or models on collision prediction using Montreal and US data [1][2]. However, neither papers predicted collision severity for Canada as a whole.

Data science is an up-trending field of exploring, manipulating, and analyzing data to provide insights for real world problems. It uses data to answer questions or make recommendations. Data analytics using machine learning would be the key to provide insights to collision severity. The dataset that this project uses is the National Collision Database (NCDB). The NCDB recorded nearly 290,000 cases of police-reported motor vehicle collisions on public roads in Canada during the year of 2017.

The goal of this project is to predict the collision severity using machine learning algorithms using the Canada-wide dataset. Analysing accident data and predicting accident severity may help improve public safety by identifying accident occurrence patterns, predicting risk of accidents and optimizing public transportation. Accident prediction suggests safer routes, which benefits society as a whole.

2. Data: Cleaning and Visualization

National Collision Database (NCDB) is a public database containing all police-reported motor vehicle collisions on public roads in Canada. It contains collision related dataset from 1999 to 2017. I downloaded the csv dataset and fed it into a panda dataframe. There are 23 columns and 289,841 rows of data.

Table 1 shows the data elements and related information. The variable C_SEV (collision severity) is the target variable that this project will predict. The collision severity variable takes on two values. It equals 1 if collision produced at least one fatality and it equals 2 if collision produced non-fatal injury.

	Data element	Definition	Levels after Feature Engineering	Included in Modelling?
Collision level data elements				
	C_YEAR	Year	19xx-20xx	No
	C_MNTH	Month	0-11	No
	C_WDAY	Day of week	0-6	No
	C_HOUR	Collision hour	1,2,3,4	Yes
	C_SEV	Collision severity	1,2	Yes
	C_VEHS	Number of vehicles involved in collision	1,0	Yes
	C_CONF	Collision configuration	1,2,3	Yes
	C_RCFG	Roadway configuration	1,0	Yes
	C_WTHR	Weather condition	1,0	No
	C_RSUR	Road surface	1,0	Yes
	C_RALN	Road alignment	1,0	Yes
	C_TRAF	Traffic control	1,0	Yes
Vehicle level data elements				
	V_ID	Vehicle sequence number	01-99	No
	V_TYPE	Vehicle type	1,0	Yes
	V_YEAR	Vehicle model year	19xx-20xx	No
Person level data elements				
	P_ID	Person sequence number	01-99	No
	P_SEX	Person sex	1,0	Yes
	P_AGE	Person age	1,2,3	Yes
	P_PSN	Person position	1,0	Yes
	P_ISEV	Medical treatment required	1,2,3	Yes
	P_SAFE	Safety device used	1,0	No
	P_USER	Road user class	1,0	No

Table 1. Data elements and related information.

This project uses the most recent 2017 dataset to predict the severity of motor vehicle collisions by building machine learning models. Since this is a labeled dataset, supervised learning algorithms will be implemented. The model that produced the highest accuracy rate will be used for future prediction.

2.1 Data Imbalance Issue

The NCDB dataset suffers from the data imbalance issue, since it is fortunately relatively rare for collision to result in fatalities, compared to non-fatal injuries. Out of the total 289,841 cases of collisions reported, only 4,468 of them were fatal cases. This might cause bias in machine learning model predictions. Therefore, I balanced the dataset by having 5,000 non-fatal rows randomly selected from the entire dataset. Together with the 4,468 fatal rows, the balanced dataset has 9,946 rows of data.

2.2 Data Cleansing and Pre-processing

In the NCDB Data Dictionary document provided together with the dataset, there are several invalid data cells (Table 2). I removed all the rows containing invalid data values from the balanced dataframe. The resulting cleaned dataframe contains 5207 rows, where fatal and non-fatal rows are 2304 and 2903, respectively. Note that the number of non-fatal rows could vary, due to the fact that they were randomly selected and may or may not contain invalid values.

Code	Description
U or UU or UUUU	Unknown
X or XX or XXXX	Jurisdiction does not provide this data element
Q or QQ	Choice is other than the preceding values
N or NN or NNNN	Data element is not applicable

Table 2. The code and description for invalid data values.

I changed the data types of most variables from object to integer, in order to fit the model. The variables which remain as type objects, are V_ID, P_ID and P_SEX. I binarized the P_SEX variable in the feature engineering section below.

2.3 Data Visualization

I visualized the data to identify the most important features using seaborn. Figure 1 shows the histograms of four categorical features, with the y-axis being the number of collisions, separated by gender, with label being the collision severity. The feature is deemed important to predict collision severity, when there is significant pattern shown on the histograms. For example, one or few levels have high collision severity rate, while other levels do not.

The top-left 2 histograms show the distribution of collision for the feature Day of week, there is roughly even distribution of collisions, so I deemed this feature to be not important and excluded it from the modeling. There were clear patterns of collision for the 3 other features. The top-right histograms show the distribution of collision for Road surface, with 1 being Dry, normal, 2 being Wet. We can see that most collisions happen on dry and normal roads. Same for the bottom-left graphs: most collision happened when Roadway configuration are 1 Non-intersection and 2 At an intersection of at least two public roadways. My suspicion was that drivers were most careless when driving on even roads in good weather days. Drivers drove more attentively when facing rough roads or tough weather. Lastly, the bottom-right graphs show the hourly distribution of collisions. Most collisions happened during daytime, and females were more likely to involve in non-fatal accidents.

I also ran histograms for all other features. The features deemed to be important and selected to be modeled are listed in Table 1 column 4.

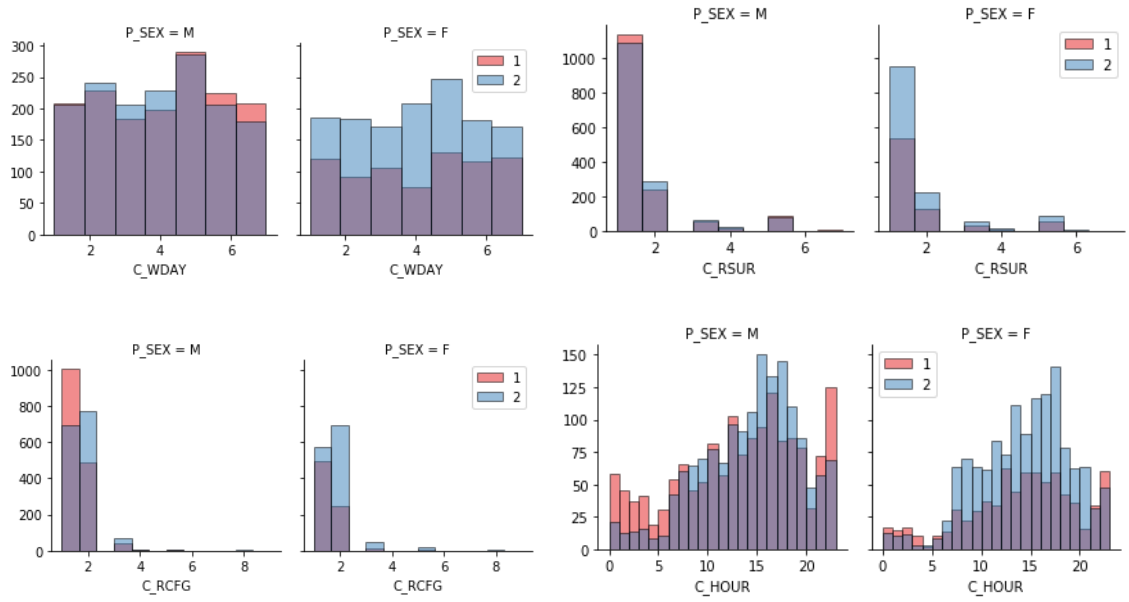

Figure 1. Histograms of four features. (top-left) Day of week. (top-right) Road surface. (bottom left) Roadway configuration. (bottom-right) Collision hour. Grey areas indicate the overlap of classes 1 (fatal) and 2 (non-fatal).

3. Methodology: Data Preprocessing and machine learnings Algorithms

3.1 Pre-processing: Feature Engineering

After selecting the important features, in this section, I performed feature engineering and feature selection to make the features either be binary or categorical, and have integer data types. This improved the predictability of the models.

Table 1 column 3 shows the feature levels after conducting feature engineering. For example, for the C_RSUR (Road surface) feature, from Figure 1 (top-right) we know that most collision happen on dry and normal road surfaces. I changed the C_RSUR from a categorical feature with 9 levels to a binary feature, with 1 being dry and normal, 0 being all other road surfaces.

I then converted categorical feature with object dtype to numerical values. The only feature that required this conversion was C_SEX (Person sex). The levels used to be F (Female) and M (Male). After replacement, 1 denoted F, and 0 denoted M.

Last but not least, I performed one hot encoding to convert categorical features to binary features.

3.2 Pre-processing: Feature Selection and Data Normalization

I created a new dataframe named Feature which contains all the features that will be used in modeling. Table 1 column 4 shows which features were selected. I defined feature set X, which equals dataframe Feature.

I then normalized the dataframe X. Data standardization gave data zero mean and unit variance. The target feature is y, which is C_SEV (collision severity).

3.3 Modeling with Machine Learning and Classification Algorithms

The supervised machine learning models were trained and tested using the X dataframe. Training set and testing set contained 80% and 20% of dataframe X, respectively. Four supervised machine learning algorithms were used and compared in order to find the one with highest prediction accuracy. The four algorithms are KNN, Decision Tree, Support Vector Machine (SVM) and Logistic Regression. Accuracy, Jaccard similarity score and F1 score were calculated for each of the algorithms. Log loss results were also calculated for Logistic Regression.

4. Results

Table 3 reports accuracy and other evaluation metrics for the four algorithms. We can see that the SVM algorithm produced the highest accuracy score, which was 80%. The accuracy score resulting from the other three algorithms were slightly lower.

Algorithm	Accuracy	Jaccard	F1-score	LogLoss
KNN	0.7564	0.7564	0.7558	NA
Decision Tree	0.7838	0.7838	0.7833	NA
SVM	0.8033	0.8033	0.8028	NA
LogisticRegression	0.7661	0.7661	0.7651	0.5160

Table 3. Results for the classification algorithms.

5. Discussion

SVM produced the highest accuracy score. We can tell that our model performed reasonably well on the test set. The F1 score is very close to the accuracy score, indicating that the two classes are balanced. The other 3 models also performed well, achieving more than 75% of accuracy. The high LogLoss of Logistic Regression is a concern.

With the features selected for modeling (Table 1), we can see that those are good indicators of a potential car accident severity.

6. Conclusion

Vehicle collisions are common and it is a public safety issue which worth conducting research on. In this project, I aim to predict the severity of vehicle collisions in Canada in year 2017. I used the National Collision Database (NCDB) public database which recorded nearly 290,000 cases of police-reported motor vehicle collisions on public roads in Canada during the year of 2017. I tackled the data imbalance issue and cleaned the dataset. Through data visualization and pre-processing, I have identified and selected the important features to include in to machine learning models. After splitting the data into training and testing sets, four classification algorithms were implemented and compared. All the models predicted with a high accuracy of above 75%, with SVM has highest accuracy of 80%. This model could be useful in predicting future potential collision severities, promote safe measures and save lives and properties.

7. References

[1] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42).

[2] Hébert, A., Guédon, T., Glatard, T., & Jaumard, B. (2019, December). High-Resolution Road Vehicle Collision Prediction for the City of Montreal. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 1804-1813). IEEE.