PREDICTION OF CAR COLLISION SEVERITY FOR CITY OF MONTREAL

1. INTRODUCTION

This report presents a machine learning approach to predict the severity of car collisions in the city of Montreal. The objective of this project was to develop a machine learning algorithm to predict the severity of a collision if a collision occurs. Such prediction can be used to plan for road developments or improvements in traffic regulations to lower possibility of collisions with severe outcomes. Another application is in map applications to give a warning of possible collisions with severe outcomes to the users and suggest alternative safer routes or other means of commute.

The car collision dataset used in this project was downloaded from:
https://open.canada.ca/data/en/dataset/cd722e22-376b-4b89-9bc2-7c7ab317ef6b

The primary language of the dataset is French, and it includes 190552 car collision data entries from year of 2012 to 2019. The dataset has 68 columns providing information about data quality, severity of collisions, date, hour, weather and road condition when the collision occurred, geographical location, and other relevant parameters that are discussed in section 2 of this report.

Note that this model cannot provide an estimate for the possibility of a car collision, as the dataset includes information about car collisions and does not include commutes that were safe with no accidents. Rather, the model can be used to predict the severity of the accident in case an accident happens.

2. DATA PROCESSING

In this section the three steps of data processing performed in this project are discussed: a. defining collision severity variable (dependent variable to be predicted), b. data cleaning and filtering, and c. defining features (variables used to predict severity).

2.1. Collision Severity

Several columns are present in the dataset that are relevant to the severity of the collision, such as number of vehicles and type of vehicles involved in the collision, and number of seriously or slightly injured people. For this project, the information in column '*GRAVITE*' is used to describe the severity of the collision. Data entries are categorized into 5 types based on the labels in this column:

1. **Minor Property Damage**:
   No casualties, and the damage assessment is lower or equal to the reporting threshold of $2,000
2. **Major Property Damage**:
   No casualties, and the damage assessment is above the reporting threshold of $2,000
3. **Non-Hospitalized Injury:**
   Only one or more victims slightly injured (injuries not requiring hospitalization, even if they require treatment from a doctor or in a hospital center)
4. **Hospitalized Injury**:
   No fatalities and at least one victim seriously injured (injuries requiring hospitalization, including those for which the person remains under observation in hospital)
5. **Fatal:**
   At least one victim died within 30 days of the accident

In this project, categories with any type of injury (3, 4 and 5) were assumed to be severe and were lumped together. Three flags 1, 2 and 3 were used to denote the three collision severity categories as summarized in Table 1.

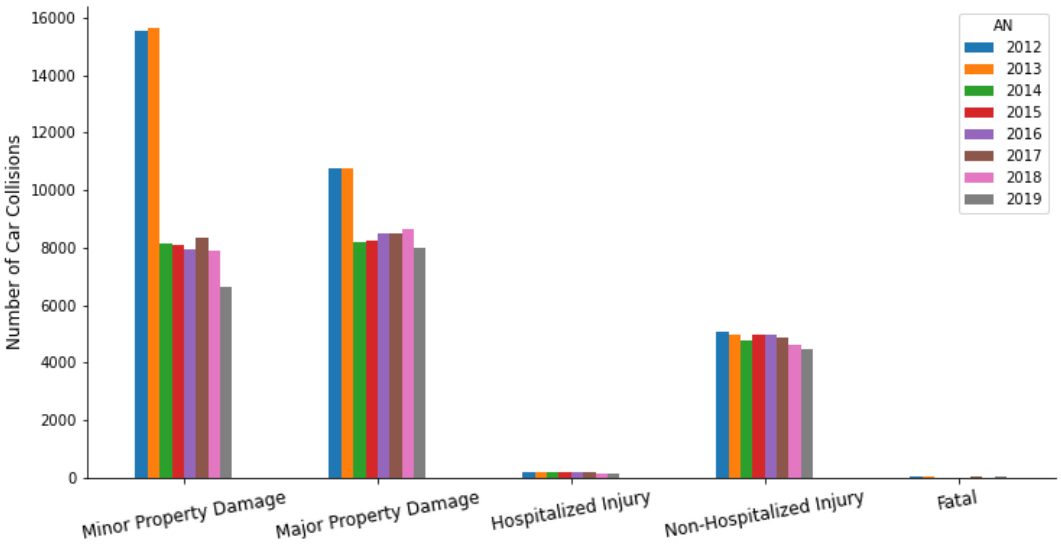
Table 1: Severity categories in the dataset and dependent variable defined in this project

Column Name in Raw Dataset	Description	Extracted Dependent Variable
'GRAVITE'	Severity categories: 1. Minor Property Damage (counts: 4325) 2. Major Property Damage (counts: 5800) 3. Non-Hospitalized Injuries (counts: 3603) 4. Hospitalized Injuries (counts: 109) 5. Fatal (counts: 21)	SeverityFlag: 1: if collision is labeled as category 1 2: if collision is labeled as category 2 3: if collision is labeled as category 3 or 4 or 5

2.2. Data Filtering:

Variations of the number of collisions in different severity categories over the years in the dataset are shown in Figure 1. Number of collisions in 2019 has dropped compared to the previous years. A noticeable drop is also noticed from year 2013 to 2014. It is not known to the author if these changes were related to a change in the reporting regulations or due to driving regulations changes, road development advancements or other reasons. For the purpose of this project, the data from the most recent year (year of 2019) was selected.

Figure 1: Annual variations of car collision counts in different severity categories in Montreal



The accuracy of the geographical references is provided in four columns (‘*LOC_COTE_QD’, ‘LOC_COTE_PD’, ‘LOC_IMPRECISION’ and ‘LOC_DETACHEE’)* as described below. For this project the data entries with a quality rating of A and quality accuracy of 1 or 2 were used, and the rest of the entries were dropped.

1. *‘LOC_COTE_QD’:*
   Data quality rating based on the location of the accident:
   A: Location of the accident is recorded according to information in fields ‘*NO_CIVIQ_ACCN’*, ‘*RUE_ACCN’*, ‘*ACCDN_PRES_DE’*.
   B: Location of the accident is recorded but there are inconsistencies in the description of the site in relation to the road network in the fields ‘*NO_CIVIQ_ACCN’*, ‘*RUE_ACCN’*, ‘*ACCDN_PRES_DE’*.
   C: No information to locate in the fields ‘*NO_CIVIQ_ACCN’*, ‘*RUE_ACCN’*, ‘*ACCDN_PRES_DE’*.
2. *‘LOC_COTE_PD’:*
   Data accuracy rating based on Montreal's road network and the consistency of information in the fields ‘*NO_CIVIQ_ACCN’*, ‘*RUE_ACCN’*, ‘*ACCDN_PRES_DE’* to georeferenced:
   1. No ambiguity about the location on the network.
   2. Location based on a semi-automatic suggestion of the geolocation tool.
   3. Several inconsistencies in the information provided for geolocation. Manual validation must be made on the location found by the geolocator.
   4. Not accurate; accident located at the centroid on the edge of the neighbourhood where the event was reported.
3. *‘LOC_IMPRECISION’:*
   This field indicates that whether there is an inaccuracy in relation to the intersection where the accident is georeferenced.
4. *‘LOC_DETACHEE’:*
   This field indicates whether the accident location was connected to the road network based on the location treatment.

2.3. Feature Selection:

The first and second column of Table 2 present the column names and descriptions from the raw dataset that were used for feature selection. The last column summarizes the processing that was performed to reformat the categorical data into either binary variables or profiles that could then be input to the model.

A. Handling of date and hour:

The number of collisions in different severity categories changes from month to month, and during different days of the week. To keep the correlations related to this variation in the feature dataset, the following steps were performed:

1. A new parameter (‘MM_D’) was defined as a compound of Month (a number between 1 to 12) and Day of Week (a number between 0 to 6)
2. Total counts of collisions in each category were calculate for each MM_D value, as shown in Figure 2.
3. For each severity category a profile was defined as the number of collisions for each MMD value divided by the maximum number of counts (to get numbers between 0 an 1),
4. Each collision data entry was given a value from the frequency profiles for each severity categories based on the month and day of week that the collision occurred

A similar approach was taken for the hour of the day parameter as well. The variations of the collision counts as a function of hour of the day are shown in Figure 3.

Figure 2: Variations of the collision counts at different months and days of the week throughout year 2019

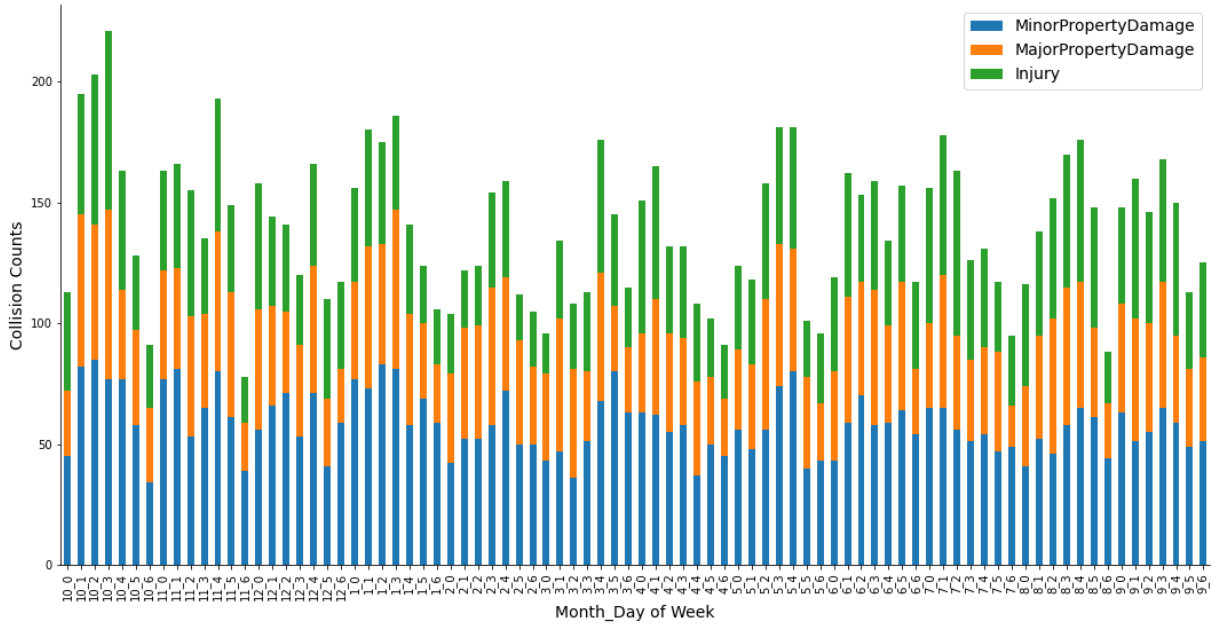
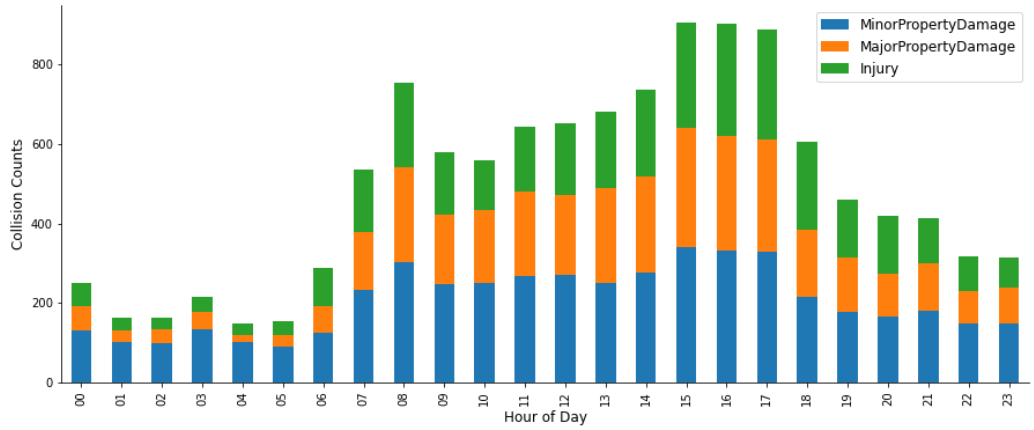


Figure 3: Variations of collision counts at different times during the day



B. Handling of categorical parameters:

All the parameters presented in Table 2 (except the date that was discussed above) are of categorical type. For each parameter, the One Hot Encoding approach was used to define a feature with binary variables for each of the top frequent categories. A different approach was taken for Road Condition, Weather and Daylight. For each of these parameters a binary variable was used to differentiate between normal conditions (e.g.., clear weather) and conditions with difficulties that could influence the severity of collisions (e.g., not clear weather).

Table 2: Columns from the original dataset and the extracted features

Column Name in Raw Dataset	Description and Top Frequent Categories	Extracted Features Names and Description
'DT_ACCDN'	Date of the accident (YYYY-MM-DD)	Frequency profiles for each collision severity category defined based on collision counts on a specific month and day of week scaled by the maximum counts.
'HEURE_ACCDN'	Time of accident in 60-minute intervals, example: 20:00:00 - 20:59:59	Frequency profile for each collision severity category defined based on collision counts on hours of the day.
'CD_ETAT_SURFC'	Road Condition: 11: Dry (counts: 8718) 12: Wet (counts: 2416) 13: Water accumulation (counts: 24) 14: Sand, gravel on the road (counts: 16) 15: Slush (counts: 229) 16: Snowy (counts: 1276) 17: Hardened snow (counts: 219) 18: Icy (counts: 505) 19: Muddy (counts: 4) 20: Oily (counts: 3)	'DryRoad': 1 if road condition is dry 0 if road condition is not dry
'CD_ECLRM'	Degree of light at the time of the accident: 1: Day and light (counts: 9001) 2: Day and half darkness (counts: 758) 3: Night and lighted path (counts: 3644) 4: Night and unlit path (counts: 60)	'Daylight': collision occurred: 1 during daylight, 0 not during daylight
'CD_ENVRN_ACCDN'	Dominant activity of the zone where the accident occurred: 1: School (counts: 294) 2: Residential (counts: 6556) 3: Commercial (counts: 5865) 4: Industrial (counts: 657)	'Residential': 1 if Residential, 0 otherwise 'Commercial': 1 if Commercial, 0 otherwise 'Industrial': 1 if Industrial, 0 otherwise
'CD_CATEG_ROUTE'	Category of road on which the first physical event (impact) has occurred in the following hierarchical order: 11: Public road: slip road / motorway collector / service road (counts: 365) 12: Public path: numbered road (counts: 305) 13: Public road: main artery (counts: 7966) 14: Public road: residential street (counts: 4771)	'Motorway': 1 if collision is labeled as category 11, 0 otherwise 'MainArtery': 1 if collision is labeled as category 13, 0 otherwise 'ResidentialStreet': 1 if collision is labeled as category 14, 0 otherwise
'CD_LOCLN_ACCDN'	location of the first fact physical impact along the road: 32: Intersection (less than 5 meters) (counts: 6733) 33: Near an intersection (counts: 3514) 34: Between intersections (100 meters and more) (counts: 2787)	'Intersection': 1 if collision is labeled as location 32, 0 otherwise 'NearIntersection': 1 if collision is labeled as location 33, 0 otherwise
'CD_CONFG_ROUTE'	Road configuration: 1: One way (counts: 3456) 2: Two directions, one lane per direction (counts: 4222) 3: Two directions, more than one lane per direction (counts: 4618) 4: Separated by passable development (counts: 457) 5: Separated by impassable development (counts: 671)	'OneWay': 1 if collision is labeled as configuration 1, 0 otherwise 'TwoWay_OneLane': 1 if collision is labeled as configuration 2, 0 otherwise 'TwoWay_MultiLane': 1 if collision is labeled as configuration 3, 0 otherwise 'Passable': 1 if collision is labeled as configuration 4, 0 otherwise
'CD_COND_METEO'	Weather condition: 11: Clear (counts: 9549) 12: Overcast (cloudy / dark) (counts: 1709) 13: Fog / haze (counts: 8) 14: Rain / drizzle (counts: 1052) 15: heavy rain (counts: 118) 16: Strong wind (no snowstorm, no rain) (counts: 40) 17: Snow / hail (counts: 755) 18: Snowstorm (counts: 93) 19: Ice (counts: 69)	'ClearWeather': collision occurred: 1 during clear weather, 0 not during clear weather
'CD_GENRE_ACCDN'	Collision type: 31: Collision with road vehicle 32: Collision with pedestrian 33: Collision with cyclist 59: Fixed object	RoadVehicle: 1 if collision is labeled as type 31, 0 otherwise Pedestrian: 1 if collision is labeled as type 32, 0 otherwise Cyclist: 1 if collision is labeled as type 33, 0 otherwise

3. MODELLING

The dataset was split into three sets:

1. **Training:** This dataset was used to fit the model parameters. 80% of the samples were used for the training set.
2. **Test:** This dataset was used to optimize model parameters to maximize the accuracy. 20% of the samples were used for the test set.
3. **Evaluation:** This dataset was used to evaluate the accuracy of the model. Model fitting and optimization was done on the training and test sets and independent from this set for an unbiased evaluation. 20% of the samples were used for evaluation.

Two model approaches were used: logistic regression and decision tree. Model steps and performance measures are presented in the following sections.

3.1. Logistic Regression:

The accuracy of the prediction using f1-score was calculated using different regularization parameters *C* (inverse of regularization fraction) and are shown in Figure 4. A value of 0.6 was used for the logistic regression model. The model performance is measured using the f1-score metrics on the evaluation set and is presented in Table 3.

Figure 4: Accuracy of the logistic regression model vs regularization parameter *C*

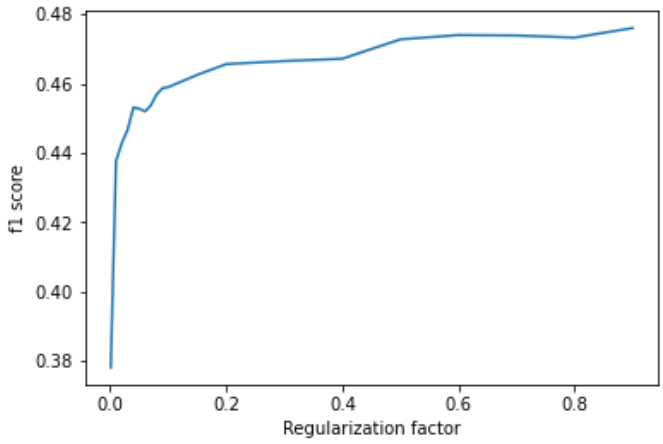


Table 3: Logistic Regression model performance using C = 0.6

	Precision	Recall	f1-score	Support
Minor Property Damage	0.49	0.81	0.61	435
Major Property Damage	0.43	0.22	0.29	309
Injury	0.79	0.46	0.58	315
accuracy			0.53	1059

3.2. Decision Tree:

Figure 5 shows the accuracy of the decision tree model on the test set for various maximum number of layers (max depth). The optimum number for max depth is found to be 6 and is used for evaluation. The model performance using the f1-score metrics is presented in Table 4.

Figure 5: Decision tree prediction accuracy vs max depth

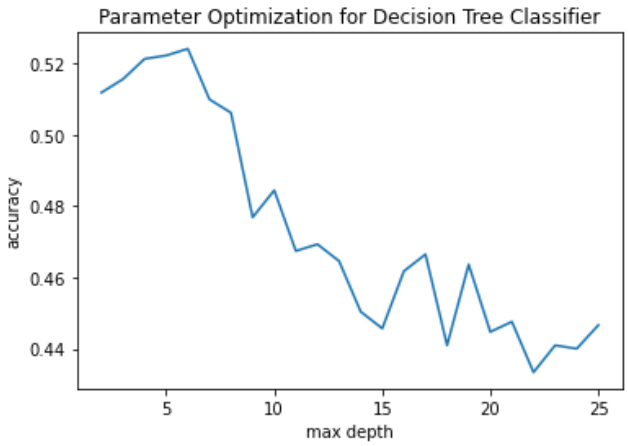


Table 4: Decision Tree model performance using max depth = 6

	Precision	Recall	f1-score	Support
Minor Property Damage	0.49	0.81	0.61	435
Major Property Damage	0.43	0.24	0.31	309
Injury	0.79	0.43	0.56	315
accuracy			0.53	1059

4. DISCUSSION OF RESULTS:

Both the decision tree and logistic regression approaches result in a similar accuracy. Both models predict collisions with minor property damage and collisions with injury reasonably well, but they are less confident in predicting collisions with major property damage. The overall accuracy of both models is 0.53.

To check for potential bias or variance issues, a learning curve is plotted in Figure 6 for the logistic regression model. The model was run several times using a subset of the training set with *m* samples and the accuracy of the training set (in-sample) and the test set (out-of-sample) were plotted against the number of samples used in the training set, *m*.

A high variance is expected when too many features are used to fit a dataset that does not have high variations. For a case with high variance, the in-sample frequency is expected to remain high with increasing *m* and a large gap is expected between the in-sample and out-of-sample frequencies. This case is not observed in Figure 6, as both accuracies converge at larger *m* values.

A high bias is expected when too few features are used to fit a complex dataset with high variations. By increasing *m*, the accuracy does not get better as the selected features cannot capture the high variations in the sample. Both the in-sample and out-of-sample accuracies converge quickly and remain unchanged by increasing *m*. As is seen in

Figure 6, the accuracies lie around 0.5 at around m = 500, and remain unchanged afterwards. This behaviour suggests the model would benefit by addition of more features.

The logistic regression model was used with the addition of polynomial features with a degree of 3 and the performance on the evaluation set is presented in Table 5. The largest improvement is seen for collisions with major property damage outcome, with recall changing from 0.22 to 0.26. The overall accuracy, however, remains almost unchanged. Another approach to account for the high variance in the dataset would be applying a neural networks model with multiple layers. For the purpose of this capstone project, the author leaves this as the future work.

Figure 6: In-sample and out-of-sample accuracies vs training sample size *m*

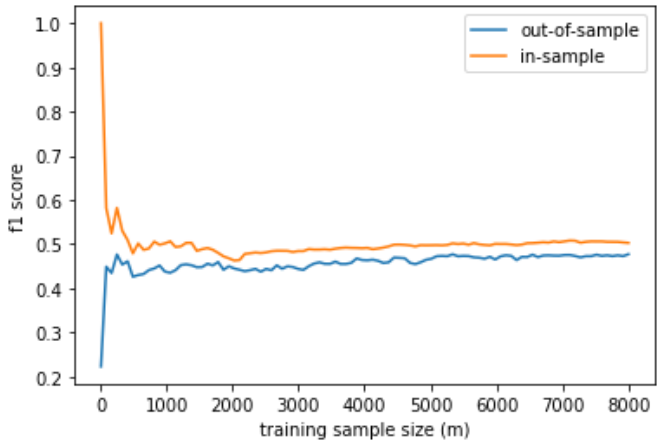


Table 5: Logistic Regression model performance using *C* = 0.6 and polynomial features of degree *n* = 3

	Precision	Recall	f1-score	Support
Minor Property Damage	0.50	0.70	0.58	435
Major Property Damage	0.41	0.26	0.32	309
Injury	0.65	0.51	0.57	315
accuracy			0.52	1059

4. CONCLUDONG REMARKS

Two machine learning algorithms, the logistic regression and the decision tree, were used to predict the severity of a collision given circumstances such as weather and road condition, hour, day and month of the accident, road type and configuration, and several other features as summarized in Table 2. Both models performed similarly, with an acceptable accuracy around 0.6 for predicting light collisions with property damages less than 2000$ and no injury, and more severe collisions involving injury. However, the confidence of predictions falls for collisions with an intermediate outcome, with no injury but damages higher than 2000$.

One reason for such poor performance could be due to the wide range of collisions that are flagged as 'Major Property Damage'. For example, a collision with slightly higher damage than the reporting threshold (e.g., 2500$) and a more serious collision with much higher damage but no injury, all are lumped into this category. If this is the case, a more refined classification is required to differentiate collisions with more severe outcomes from those with lighter consequences in the Major Property Damage category. One approach could be to combine the

Severity flags with other pieces of information, such as number of vehicles involved in the crash, type of vehicles, etc., to refine the collision severity classes further.

Further improvements might also be possible by applying a more sophisticated machine learning algorithm. As the learning curve suggests in Figure 6, the number of features might not be enough to capture variations in the dataset, and a polynomial feature extension did not result in a significant improvement. A neural network approach with multiple layers might result in a better performance.