Car Accident Severity: Analysis and Prediction using Machine Learning

Samir Gouhary

October 21, 2020

1 Introduction

1.1 Background and Problem

Hundreds of car accidents happen daily and they have a huge impact on both the people involved in the crash and the responding parties. On one hand, depending on the severity, the accidents result in physical and emotional trauma for the drivers and their passengers. On the other, there are substantial financial costs associated with car accidents. In this study we have attempted to find a way to accurately predict the severity of car accidents based on a number of factors, such as weather,road and light conditions, etc. using a number of machine learning models.

1.2 Applications/Interest

There are two main groups who directly benefit from this study - emergency response services and urban planners. Firstly, emergency services from all sectors (police, fire-fighters, paramedics, etc.) can all strategise their resource allocation on a daily basis. In turn this results in cost reductions. For medical staff, the reduced car accident cases also allows them to put their resources into other patients.

On the other hand, the findings in this study will aid urban planners in improving road safety. Road signs, speed bumps, street lights as well as other infrastructure can be installed in the optimal locations to reduce the frequency and severity of accidents.

2 Data

2.1 Data source and properties

The data that was used to train the models in this project is the Seattle city Collisions with records from 2004 to present date. It is provided by SPD and recorded by Traffic Records.

The data consists of 194673 recorded collisions with 38 feature columns. The data consists of almost entirely categorical data with the exception of the ID columns as well as the columns containing number of people involved in a collision and the recorded time. There are a few columns which include a word description of the collision, its type and a description of the severity. The target variable is given by the column *SEVERITYTYPE* which is categorical variable ranging from 0 to 3, with 3 being a fatality.

2.2 Data cleaning and formatting

As stated in 1.2 the two groups that are expected to benefit from this study are the emergency response services (ERS) and urban planners. However, the available information for both groups when they plan their resource allocation is completely different - the ERS only posses information about the external conditions. That includes weather conditions, lighting conditions, road conditions and the time of day. These factors are described in the data by the *WEATHER, LIGHTCOND, ROADCOND and INCDTTM* columns respectively. All columns except *INCDTTM* contain categorical values as well as missing ones, thus they were encoded and the rows with missing values were dropped (values were dropped due to the abundance of data).

On the other hand, urban planners are more interested in urban configurations when devising plans for safety improvements. These include the address type where collisions occur, type of junction at which these collisions occur, etc. In the data, these categories are given by the *ADDRTYPE and JUNCTIONTYPE* columns. Again, both of these columns only contain categorical data, therefore it was encoded and missing value rows were dropped.

The goal of this study is to predict accident severities before they happen based on different factors. However, most of the columns contain data which describes collisions which have already happened, meaning that they cannot be used as features in the model.
Therefore, all columns other than *WEATHER, LIGHTCOND, ROADCOND,*

INCDTTM, ADDRTYPE, JUNCTIONTYPE and SEVERITYTYPE were dropped.

Analysis of the our target variable *SEVERITYTYPE* shows that there are only two unique values, namely 1 and 2 which represent property damage and injury respectively. Thus, this problem becomes a binary classification problem. Moreover, the counts of the target variable shows us that the data is imbalanced with severity type 1 making up nearly 70% of the whole dataset. The data was therefore downsampled, resulting in a final dataset with 56638 entries for each severity type for a total of 113276 entries and 6 features.

3 Methodology

3.1 Exploratory data analysis

3.1.1 Relationship between time and accident severity

One of the dataset features we are interested is the date and time of the collision (Section2.2). However we are not interested in the exact date but rather the components of the time and date - hours, days of the month, the months and the years. Having separated those features, we can then see the distribution of the class 1 and 2 severity compared to these time components (Figure1).

Figure 1 very clearly outlines the fact that there is no relationship between any component of time and accident severity. Looking at the distribution of the hour by hour histogram (top left) for example, we see that both severity classes 1 and 2 have the exact same distribution. Since our goal is to be able to classify the severity of an accident and not the overall probability of one happening, we cannot use the hour of day as a classifier feature. A similar analysis can be applied to the rest of the 3 graphs, resulting in the same conclusion. The only exception can be considered to be the 4th (bottom right) plot. It is the only one showing a difference in distributions for the severity classes, with class 1 being more prevalent in the years before 2010-2012 and class 2 dominating after 2012. However even those differences, when compared as percentage difference are still too small, thus making years a bad classification feature as well.

[1]The midnight hour (00:00) is excluded due to an anomalous count. It is most likely caused by negligent record keeping.

Currently the remaining features are the weather conditions, road conditions, lighting conditions, junction and address types which are all categorical. Because both the target variable and feature set are completely categorical, we perform a χ^2 contingency test to tell if the features and the target are dependent or independent.

	FEATURES					
Statistic	WEATHER	LIGHTCOND	ROADCOND	JUNCTIONTYPE	ADDRTYPE	STATUS(Benchmark)
Critical Value (CV)	18.307	15.507	15.507	12.592	5.991	3.841
χ^2 Statistic ($CHI2$)	3666.709	3636.771	3729.823	8056.589	6804.426	0.170
Independence criteria: ($CHI2 \leq CV$)	Dependent	Dependent	Dependent	Dependent	Dependent	Independent
P-value	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.681$
Independence criteria: ($p \geq \alpha = 0.05$)	Dependent	Dependent	Dependent	Dependent	Dependent	Independent

Table 1: Table of the χ^2 contingency test with the respective outcomes of the test. The benchmark STATUS is a column with categorical variables which is known to be independent of the target variable and the test correctly predicts that with high certainty of $\approx 70\%$

According to Table 1, the target variable is dependent on the whole feature set with p-values < 0.01 for all of them, meaning the dependence is statistically significant.

3.2 Feature selection

After having confirmed that our selected features and the target variable are statistically dependent, we have to check how important they are. For this, I have used sklearn's feature selection module. Using an extra-random trees classifier from sklearn, we train the tree classifier with the whole feature set and the target variables. I have then plotted the rankings as a bar graph to along with the standard deviation of the feature importance represented as black error bars (Figure 2).

Figure 2 shows that ranking is as follows: *1- JUNCTIONTYPE, 2- ADDRTYPE, 3-LIGHTCOND, 4- WEATHER, 5-ROADCOND.* The *ROADCOND* and *WEATHER* columns are inherently connected. When the weather is recorded as raining or overcast, the road conditions is recorded almost entirely as wet. Therefore, *WEATHER* and *ROADCOND* provide very similar information to the model. Due to its low feature importance,as well as it conveying similar information as WEATHER, I have decided to drop *ROAD-*

COND out of the feature set. This leaves the *WEATHER, LIGHTCOND, JUNCTIONTYPE and ADDRTYPE* as the final feature set which was used to train the classification model.

3.3 Classification through machine learning

I have applied 3 classification models to this problem - Logistic Regression (LR), Random Forest (RF) and Multilayer Perceptron (MLP). The hyperparameter tuning was done using sklearn's GridSearchCV module in combination with the Pipeline object to iterate over the list of models. The performance metric was chosen to be the logarithmic loss due to the probabilistic nature of the problem. Other metrics such as Jaccard, F1-score and Area Under ROC were also calculated.

Before training however, I examined the effect the test split size has on the accuracy of the model using a Random Forest Classifier (Figure 4) as well as a Logistic Regression Classifier (Figure 3). The metrics used were the Area Under ROC as well as accuracy score.

Comparing both the LR and RF performances, it is clear that the accuracy scores for both classifiers are around the same with a value of ≈ 0.60. The AUROC scores for RF are slightly higher by around 0.02.

In this project, we are interested in predicting the severity of new cases, therefore out-of-sample or test accuracy is more important than the train set accuracy. The LR's (Figure 3) test scores for both the accuracy and AUROC exhibit oscillatory behaviour. However, the amplitude of these oscillations is negligible. Therefore, it is constant around the oscillation's mean and as such the test split size has little to no effect on its accuracy.

On the other hand, the RF (Figure 4) has more clearly defined maximums at around 0.15 ± 0.05. Hence, for our model training we will be using a test split size of 0.15.

4 Results and Discussion

After the models were trained and had their parameters optimised using sklearn's GridSearchCV (Section 3.3), their best respective performances with respect to logarithmic loss were recorded and graphed.

	LOG_LOSS	F1	JACCARD	AUROC
LogisticRegression	0.664819	0.626692	0.456349	0.627344
RandomForests	0.654687	0.604969	0.433673	0.642596
MultilayerPerceptron	0.651927	0.610904	0.439831	0.643755

Table 2: The table contains the performance metrics for each of the best performing machine learning models with respect to logarithmic loss. The highlighted values represent the best accuracy score for each metric.

Table 2 shows that the two best performing models are the Multilayer Perceptron as well as the Logistic Regression. MLP has the best logarithmic loss and area under ROC scores, where as LR has the best F1 and Jaccard scores.

Selecting the best final model requires that we look at which metrics are more important. We have already stated that the problem is probabilistic. That is, we are interested in the probabilities of the severity of an accident in order for ERS and urban planners to develop optimal resource allocation strategies. We therefore have to choose the Multilayer Perceptron. The area under ROC tells how likely our model is to distinguish between the positive and negative classes. Additionally, MLP also has the lowest logarithmic loss which is already our ranking evaluation metric.

However, this does not mean that any of the models perform very well. All three models have similar evaluation metric values, which are all close to, but above 0.5 for F1, logarithmic loss and area under ROC as well as low Jaccard similarity score. This can be interpreted as no model being very good at distinguishing between the two classes.

This can be for a number of reasons. One possible explanation is that the models don't have their hyper-parameters optimised very well. Large grid searches were avoided due to the computation time required. Therefore, to solve this problem, more rigorous training is required.

Another possible explanation could be the low impact of the variables chosen as a feature set. After all, vehicles are driven by humans and therefore the most likely cause for accidents and the scale of the severity is human error, which is impossible to measure.

The factors that we have used as our feature set only affect the drivers, increasing or decreasing the chance for a human error to occur.

5 Conclusion

In this study we attempted to create a machine learning model which predicts the severity of a car accident based on a few environmental factors and locational information. We achieved a minimum logarithmic loss of $0.65(2dp)$ and an area under ROC of $0.64(2dp)$. Considering our goal, this study has been an overall success. We created a model which can predict, though not extremely well, the severity of a car accident using just 4 features.

There are difficulties with this problem, due to the fact that arguably the main factor in a car accident's severity is human error which is not quantifiable. The features in our model simply affect the probability of human error occurring.

This is not to say that this model is completely unusable by the ERS and urban planners. Having even a little clue as to which environmental factors or which locations increase the probability of more severe accidents allows them to allocate resources and devise preventative strategies.

In future, this problem can be researched more deeply with more classifications models and deeper parameter optimisation. Moreover, more complex methods for feature extraction can be used, in order to improve the accuracy of the model.

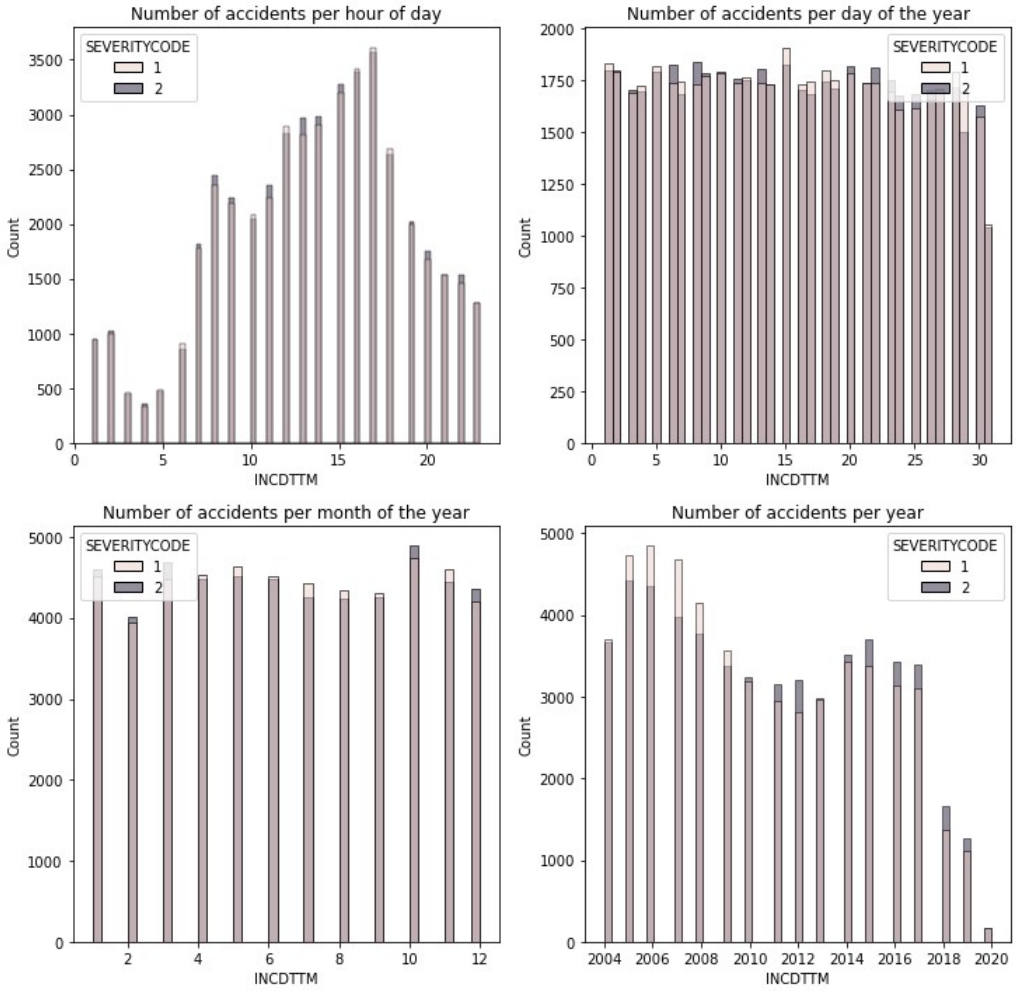

Figure 1: Relationship between different scales of time and the severity of the accident. Each figure represents the distribution of each severity with respect to, starting from the top: hour of day [1], day of the month, month and finally year.

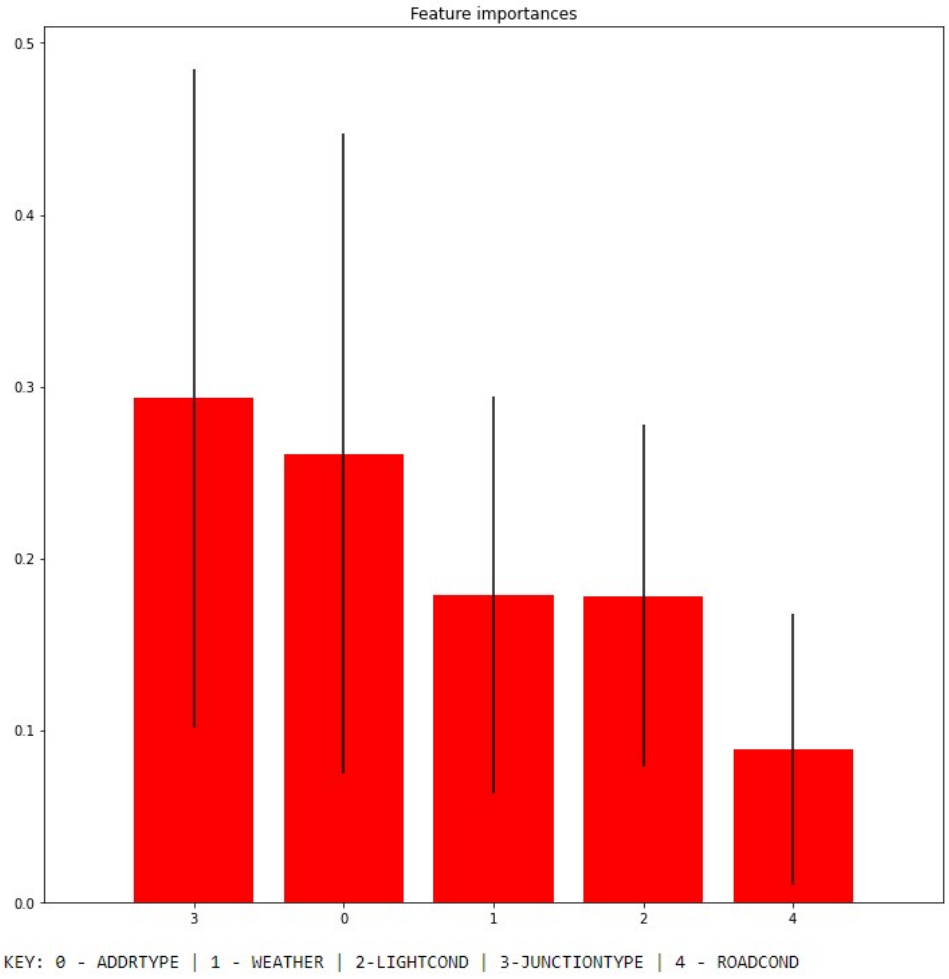

Figure 2: Feature importance ranking using sklearn "Feature Selection". The key is provided under the graph with the labels for each feature. The black error bars represent the standard deviation in the importance.

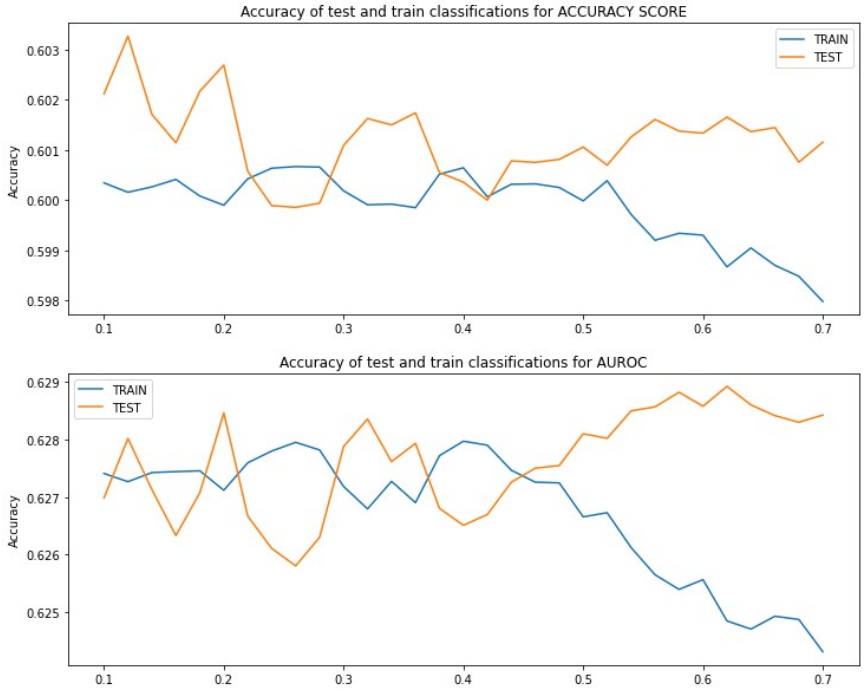

Figure 3: Variation of the Accuracy score and Area under ROC (AUROC) accuracy score with respect to data's test size split, using a LR model. Best performance is calculated as the average of the train and test set performance

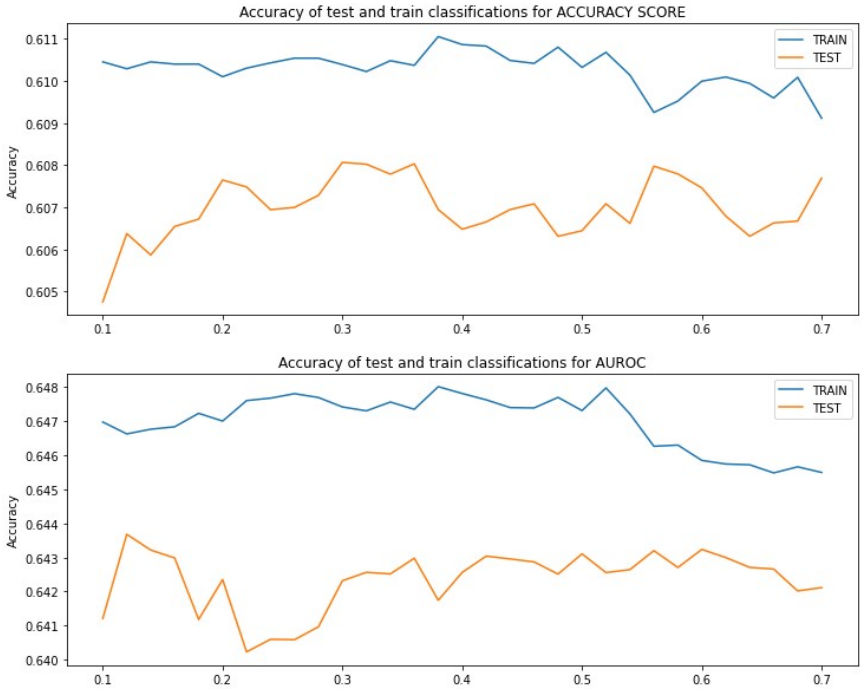

Figure 4: Variation of the Accuracy score and Area under ROC (AUROC) accuracy score with respect to data's test size split, using a RF model. Best performance is chosen as the highest test set performance

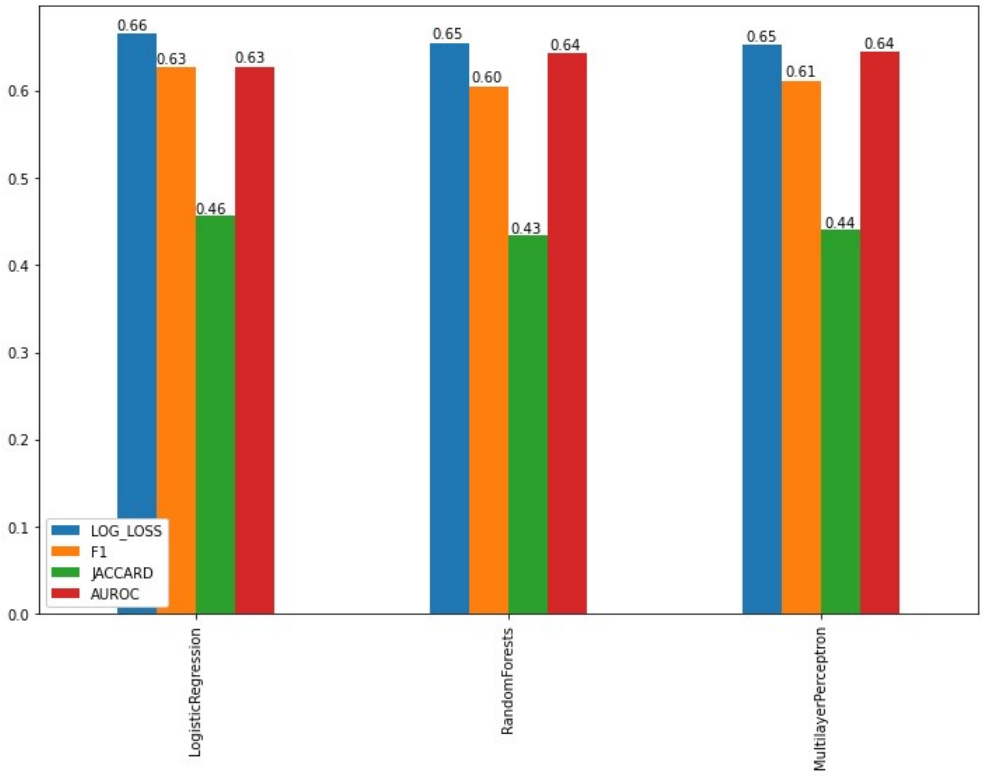

Figure 5: A graph of the performance metrics of the best performing machine learning models with respect to logarithmic loss.