Capstone Project Report - Predicting severity of injuries sustained in car crashes
A Binary classification supervised machine learning project

By Conor Owens-Walton

Table of Contents

1. Introduction *1*
1.1 Who would be interested in project 1
1.2 Data 2
3. Methodology *2*
3.2 Normalizing the data 4
3.3 Train and Test Sets 4
3.4 Classification Modelling 5
4. Results *5*
4.1 Model performance 5
4.1.1 F1 Scores: Precision and Recall 5
4.1.2 Weighted F1-Scores and Jaccard Index 5
5. Discussion *6*
6. Conclusion *6*
6.1 Limitations 7
6.2 Future Directions 7

1. Introduction

36,560 people died in motor vehicle crashes in the United States in 2018. Understanding whether someone who has been in a car accident is likely to be gravely injured, may enable our medical and emergency services teams to better prepare when there is an accident.

For this project, we have built a number supervised machine learning models to predict the severity of road accidents. These models will enable us to predict, based on information from the crash scene, whether the people involved are likely to be severely injured, and thus in need of more immediate and thorough care. This is a classic binary classification supervised machine learning task where we use a number of independent variables to predict a binary outcome, in this case whether someone is severely injured or not in a car accident.

1.1 Who would be interested in project

Information like this will enable hospitals and emergency officials to better allocate resources when there are car collisions, with the hope that such a move can reduce the chances that people involved in those car collisions will die from their injuries.

1.2 Data

Based on our business problem, this project used the 'Seattle Accident Collisions' dataset, bringing together data from 2004-2015. This variables that we will use from within this dataset relates to certain physical characteristics of the accident scene and the environment, factors that could all be easily discerned by first responders, and related to emergency services who would input this data in their systems and predict the likelihood that the injuries sustained by those involved in the crash would be mild or severe.

This project thus uses data on the type of collision that took place; what the lighting was like around the collision scene; what the weather was like at the time of the collision; and finally, whether a driver was speeding.

Taken together, our models can identify the severity of the injuries sustained in road collisions based on those factors.

3. Methodology

3.1 Feature Engineering

To begin we needed to perform significant feature engineering on our dataset to get it to a point where we could train meaningful and accurate machine learning models on it.

This involved removing a number of columns that were of no interest, or that would only tell us data that a first responded would not be able to discern at the scene of a car collision, such as whether inattention was involved or whether the driver was under the influence of alcohol.

This process left us with 5 variables to use, 4 independent variables and one dependent variable. Our independent (or predictor variables) which we use to predict our dependent variable (outcome variable) are as follows:

- Independent variables
    - Collision type
        - Whether the collision happened at an angle
        - Whether a bicycle was involved
        - Whether the collision happened head-on
        - Whether the vehicle was turning left
        - Whether the vehicle was turning right
        - Whether the collision involved a parked car
        - Whether the collision involved a pedestrian
        - Whether the collision involved someone being rear-ended
        - Whether the collision involved someone getting side-swiped
    - Weather
        - Was the weather clear at the time
        - Was there fog/smoke/smog
        - Was the weather overcast

        - Was it raining
        - Was it snowing
        - Was it dry
    - Road condition
        - Was the road dry
        - Was the road icy
        - Was there oil on the road
        - Was the road sandy/dirt
        - Was there snow on the road
        - Was there pooled water on the road
    - Lighting
        - Was it dark (without streetlights)
        - Was it dark (with streetlights)
        - Was it dawn at the time of the accident
        - Was it daylight at the time of the accident
        - Was it daylight
        - Was it dusk

- Dependent variable
    - Severity of the injury sustained in the accident

We wanted out models to predict if an accident causes severe injuries, or just mild ones. Because of this we removed rows with a '0' value (unknown) or 1 (property damage). We also combined the ‘Serious injury’ and ‘fatality’ data into a single category of ‘severe’ injury. All other injuries were considered ‘Mild’.

After this we dropped all of the values that were not meaningful, such as missing data, or data that would not be interpretable by a first responder.

This left us with a dataset with 50157 instances of ‘Mild’ injuries and 2713 instances of ‘Severe’ category injuries, or a total N = 52,870.

Because of this severe class imbalance, any machine learning algorithms would ignore our minority class which we cannot have. The simplest approach to deal with this is to synthesize new examples from the existing examples using a data augmentation for up-sampling the minority class ('Severe' injury instances) (Figure 1).

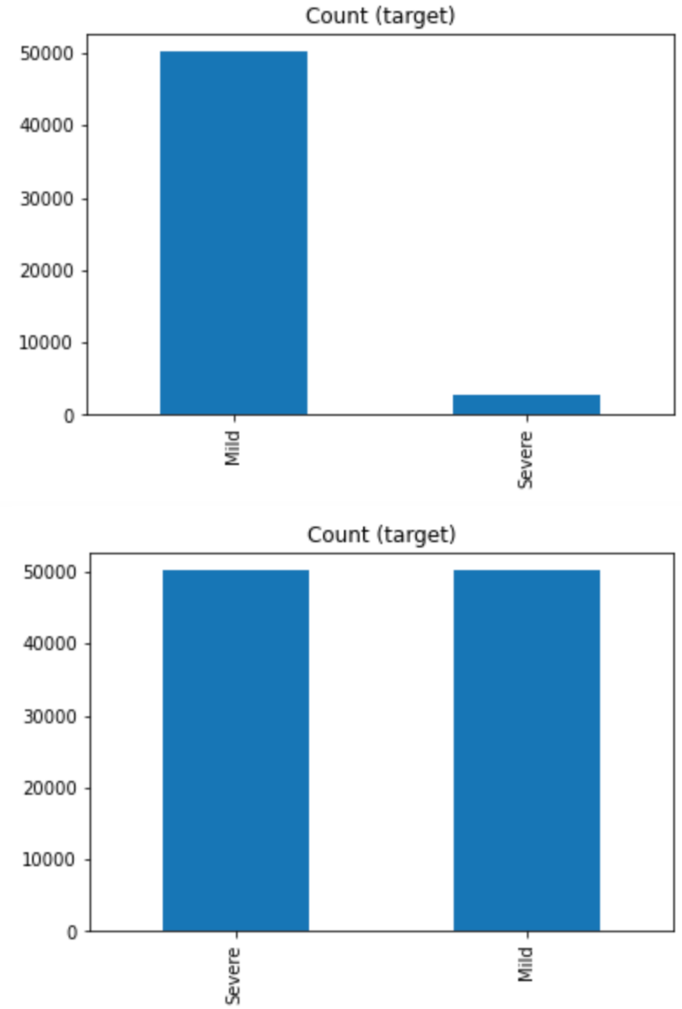

Figure 1: Here we can see the major class imbalance between instances of our outcome variable of interest in the top. To fix this issue we up-sampled our 'Severe' cases using a synthetic data-augmentation approach. In the bottom figure you can see the new data showing that both outcome categories are even.

After this data augmentation approach, we finished up with 50,157 instances of 'Severe' and 50,157 instances of 'Mild' injuries sustained in car collisions.

3.2 Normalizing the data

After this step we performed data standardization givin our data a zero mean and a one unit variance.

3.3 Train and Test Sets

We then broke up the data into train and test sets, training our models on %80 of our data and testing it on the remaining %20 of the data that was unseen.

3.4 Classification Modelling

This project employed four supervised machine learning models to predict our binary outcome measure.

1. K-Nearest Neighbor
2. Decision Tree
3. Support Vector Machine
4. Logistic Regression

To train our K-Nearest-Neighbor (KNN) we first ran an series of simulations with varying values of K to work out the optimal number.

From that we ascertained that using the closest 3 neighbors, L=3, is the best approach.

4. Results

4.1 Model performance

To measure the effectiveness of our models we used two key statistics as our performance metrics. The first is an F1-Score, and the second is a Jaccard Index.

4.1.1 F1 Scores: Precision and Recall

Our models obtained precision, recall and F1-scores for predicting Mild and Severe injuries from car collisions, presented in Table 1. The most effective score, that balances precision and recall, and is represented by an F1-score of 0.7 is our Decision Tree model when predicting Mild injuries.

Table 1. Precision and recall model performance of machine learning models predicting the severity of an injury

Model	Injury Severity	Precision	Recall	F1-Score
KNN	Mild	0.65	0.61	0.63
	Severe	0.63	0.67	0.65
Decision Tree	Mild	0.64	0.78	0.70
	Severe	0.72	0.56	0.63
SVM	Mild	0.66	0.73	0.69
	Severe	0.70	0.62	0.66
Logistic Regression	Mild	0.65	0.74	0.69
	Severe	0.70	0.60	0.65

4.1.2 Weighted F1-Scores and Jaccard Index

Taking both performance evaluation metrics into account, our most effective model in predicting whether a car collision results in ‘mild’ or ‘severe’ injuries is the SVM model, with a Jaccard Index of 0.68 and an F1-Score of 0.67.

Table 2. F1-Scores and Jaccard Index

Model	Jaccard Index	F1-Score
KNN	0.64	0.64
Decision Tree	0.67	0.66
SVM	0.68	0.67
Logistic Regression	0.67	0.67

5. Discussion

Here we can discuss any observations we made as well as recommendations on we can make on the results.

To measure the effectiveness of our models we used two key statistics as our performance metrics. The first is a Jaccard Index, which is a statistic that helps demonstrate the similarities between sample sets. It defines the size of the intersection between two sets divided by the size of the union between them. The Jaccard Index, also known as the Jaccard similarity coefficient, is a statistic used in understanding the similarities between sample sets. The measurement emphasizes similarity between finite sample sets, and is formally defined as the size of the intersection divided by the size of the union of the sample sets.

The second statistic we use to measure effectiveness is the F1 statistic. This statistic gives us a measure of the balance between the precision and the recall of our models. Precision is the positive predictions divided by the total number of positive class values predicted. It is commonly referred to as the positive predictive value. Precision can be thought of as a measure of a classifiers exactness. A low precision can also indicate a large number of false positives, which would be our model incorrectly labelling a car collision as 'Severe' or 'Mild' based on a set of factors. Recall is the number of positive predictions divided by the number of positive class values in the test data. It is commonly referred to as the 'sensitivity' of a model. Recall can be thought of as a measure of the completeness of a classifier, a low recall indicates many false negatives.

Based on information about the type of collision, the weather, the condition of the roads and the lighting that was available at the time of the incident, our best SVM model can predict with an accuracy of 0.7 whether an 'severe' or 'mild' injury would likely have been sustained based on these factors. This SVM model had a Jaccard Index of 0.68 and an F1-Score of 0.67.

6. Conclusion

The goal of this project was to develop machine learning models that emergency services could implement to better predict the likelihood of a mild or severe injury occurring after a car accident, based on a range of factors of the crash site. While our model was able to perform better than chance, with a positive predictive rate of 0.7, there is still significant improvement if such a model was to ever be able to be used in the real world.

6.1 Limitations

Possible limitations of the model may be that we used artificially inflated data to cater for the class imbalance problem.

6.2 Future Directions

Future development of similar models might benefit from more thorough feature engineering that can better uncover relationships between variables, and how they can be combined to best predict the outcomes of car collisions.