Identifying the causes and contributing factors of road collisions with injuries and fatalities in the City of Seattle, USA

Rob Highett-Smith

15 September 2020

1. Introduction

1.1 Background

According to ASIRT, road collisions are the leading cause of death in the U.S.A. for people aged under 55, with more 38,000 fatalities annually and an additional 4.4 million injuries serious enough to require medical attention. However, not only is there an obvious human cost of these incidents. There is also significant economic cost as a consequence of these incidents.

According to the CDC, the cost of medical care and productivity losses associated with occupant injuries and deaths from motor vehicle traffic crashes exceeded $75 billion in 2017. Despite the significant impact of road collisions, travel is a critical part of life. Therefore it is important to understand the causes of and contributors to road collisions in order to develop effective strategies and policies for reducing these incidents, as well as the human and economic consequences.

For this project, the focus is on the City of Seattle, in the state of Washington on the North West cost of the United States of America.

1.2 Problem

Given the sheer number of road collisions that happen each year in a city like Seattle, there are a wide range of possible causes and factors. However, quantifying and better understanding these will be critical to developing effective strategies and policies to reduce them.

The objective of this study is to build a robust, statistical model to identify the factors that predict which accidents result in injuries and death compares to those that do not (or result only in property damage).

By developing an understanding of the historical factors contributing to road collisions resulting in injury or death in the city of Seattle, this analysis can then be used by the Authorities to develop appropriate strategies and policies to reduce these outcomes.

This will be achieved through a statistical analysis of data collected on road collisions in order to identify the leading causes and contributing factors of road collisions that result in injuries or death.

1.3 Interest

Given the significance of the consequences and costs associated with road collisions, there are many stakeholders who would be interested in reducing them by an improved understanding of the causes and contributing factors. Specifically in this case, the local Government authorities of Seattle would be interested in the findings from this project, as

they have responsibility for implementing policies and strategies to reduce the impact and costs of road collisions in the City.

2. Data acquisition, variable selection & cleaning / processing

2.1 Data source

The data for this project was obtained from the City of Seattle Open Data portal.

Data was downloaded on the 6 September and includes 221,266 records relating the road collisions between 2004 and present updated on 5 September 2020).

2.2 Dependent (Target) variable

The target variable is 'SEVERITYDESC' which represents a code describing the severity of the accident. This variable has 5 levels:

- 0 - Unknown
- 1 - Property damage (only)
- 2 - Injury
- 2b - Serious Injury
- 3 - Fatality

For this analysis, the records with 0s (Unknown) will be discarded and levels 2, 2b and 3 will be combined to represent any collision that resulted in injury or death. One record was NaN for this variable, assumed to be 'unknown' and discarded. The resultant binary variable included 199,630 records and was recoded as:

- 0 - Property damage
- 1 - Injury / Fatality

2.3 Independent variable selection

The downloaded dataset consisted of 40 variables, including the Dependent Variable.

The metadata for this dataset was downloaded and examined. For all of the 40 variables, the number and meaning of the unique values present for each variable was examined in the context of the analysis objective, and a judgement made on whether it should be included in the analysis.

Consequently, the following variables were discarded for the following reasons:

Variable deleted	Rationale for deletion
'SEVRITYDESC'	This is the same as the DV and so not required
'X' & 'Y' co-ordinates	Variable not required for the analysis
'LOCATION'	Variable not required for the analysis
'INCKEY','COLDETKEY'	All keys / IDs not required for the analysis
'SEGLANEKEY', 'CROSSWALKKEY'	All keys / IDs not required for the analysis
'REPORTNO','INTKEY'	'SEGLANEKEY'
'STATUS','EXCEPTRSNCODE'	No metadata / not required for the analysis
'EXCEPTRSNDESC', 'SDOTCOLNUM'	No metadata / not required for the analysis
'INCDTTM'	Has same values as 'INCDATE' so not required
'SDOTCOLCODE', 'SDOTCOLDESC'	Contain a lot of information (40 levels) but do not add much to the analysis
'ST_COLCODE', 'ST_COLDESC'	Similarly contained a lot of detail (63 levels) with nothing useful to be extracted
'INJURIES', 'SERIOUSINJURIES', 'FATALITIES'	Include counts which are not required for this analysis
'HITPARKEDCAR', 'COLLISIONTYPE'	Considered more of an outcome of a collision thus not required
'JUNCTIONTYPE'	Contains similar information to 'ADDRTYPE' – not required
'PERSONCOUNT', 'PEDCOUNT'	Not considered useful for this analysis
'PEDCYLCOUNT', 'VEHCOUNT'	Not considered useful for this analysis

Table 1: Variables discarded with rationale for deletion

After discarding these variables, 9 Independent variables were retained for the analysis (in addition to the DV):

- 'ADDRTYPE', 'INCDATE', INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', PEDROWNOTGRNT', and 'SPEEDING'.

Note: 'OBJECTID' was also retained as an ID in case it was required, but will not be used in the analysis

2.4 Data cleaning / processing

A new dataframe ('modeldf') was created containing the Dependent Variable and the 9 Independent Variables. The datatypes, missing data and values were examined.

Several variables ('INATTENTIONIND', 'SPEEDING', 'PEDROWNOTGRNT' and 'UNDERINFL') were processed to replace 'N' and 'Y' values with 0 and 1 respectively. These variables were examined and it was ascertained that missing data in these cases were the same as 0 and thus missing data was transformed into zeros for these variables.

Several variables ('ROADCOND', 'LIGHTCOND' and 'WEATHER') were processed to combine values into a smaller set of values with similar / aggregated meanings. For example, the variable 'ROADCOND' was transformed from a variable with 9 values ('Dry','Wet','Unknown','Standing Water','Ice','Snow/Slush','Other','Sand/Mud/Dirt' and 'Oil') to a variable with 3 values ('Dry', 'Wet' and 'Other').

A new feature / variable ('Weekend') was extracted from INCDATE, representing Weekday (0) and Weekend (1). INCDATE was then dropped from the dataframe.

Missing data was reviewed across the dataset and in all variables represented about 10% of all values or less. Given the large number of records for analysis it was decided that records with missing data would be discarded from the dataset.

The datatype for all remaining variables was then reviewed and consisted of either Integers (Int) or categorical variables (Object). It was decided to retain these datatypes for exploration in the next phase, where some further processing or the creation of dummy variables might be undertaken.

Given there were no continuous variables in the dataset at this point, no normalisation or binning (beyond the aggregation described above) of variables was required.

The dataset will also be balanced in the next stage (EDA) prior to modelling.

At the completion of the Data cleaning / processing stage, the dataset consisted of 10 variables (a Dependent Variable and 9 Independent Variables) with a total of 174,452 observations.

3. Methodology

3.1 Exploratory Data Analysis

As a first step, the frequency distribution of all variables was examined. The Dependent Variable (SEVERITYCODE) had been aggregated in the Data Processing phase into two categories: ‘0’ representing collisions resulting in Property Damage only and ‘1’ representing collisions resulting in injury or death. 34% of the 174,452 collisions in the data file had resulted in injury or death, meaning that the data file was imbalanced, although not severely so.

Figure 1: Frequency distribution of Dependent Variable

The frequency distributions of the Independent Variables were also examined.

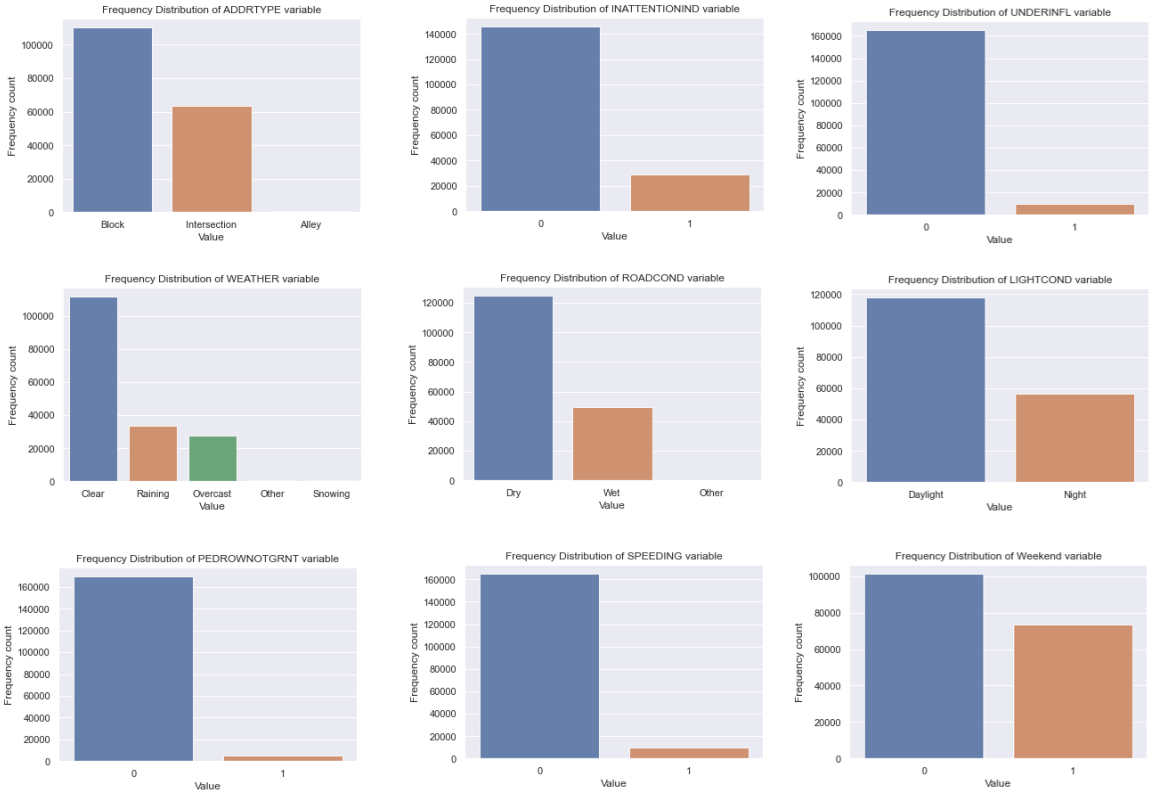

Figure 2: Frequency distribution of Independent Variables

These distributions showed some interesting patterns. Most of these variables were dominated by the ‘0’ value, indicating an absence of the factor in question.

The frequency distribution of Independent Variables compared to the Dependent Variable were then plotted using Bar Charts to visually examine their relationships:

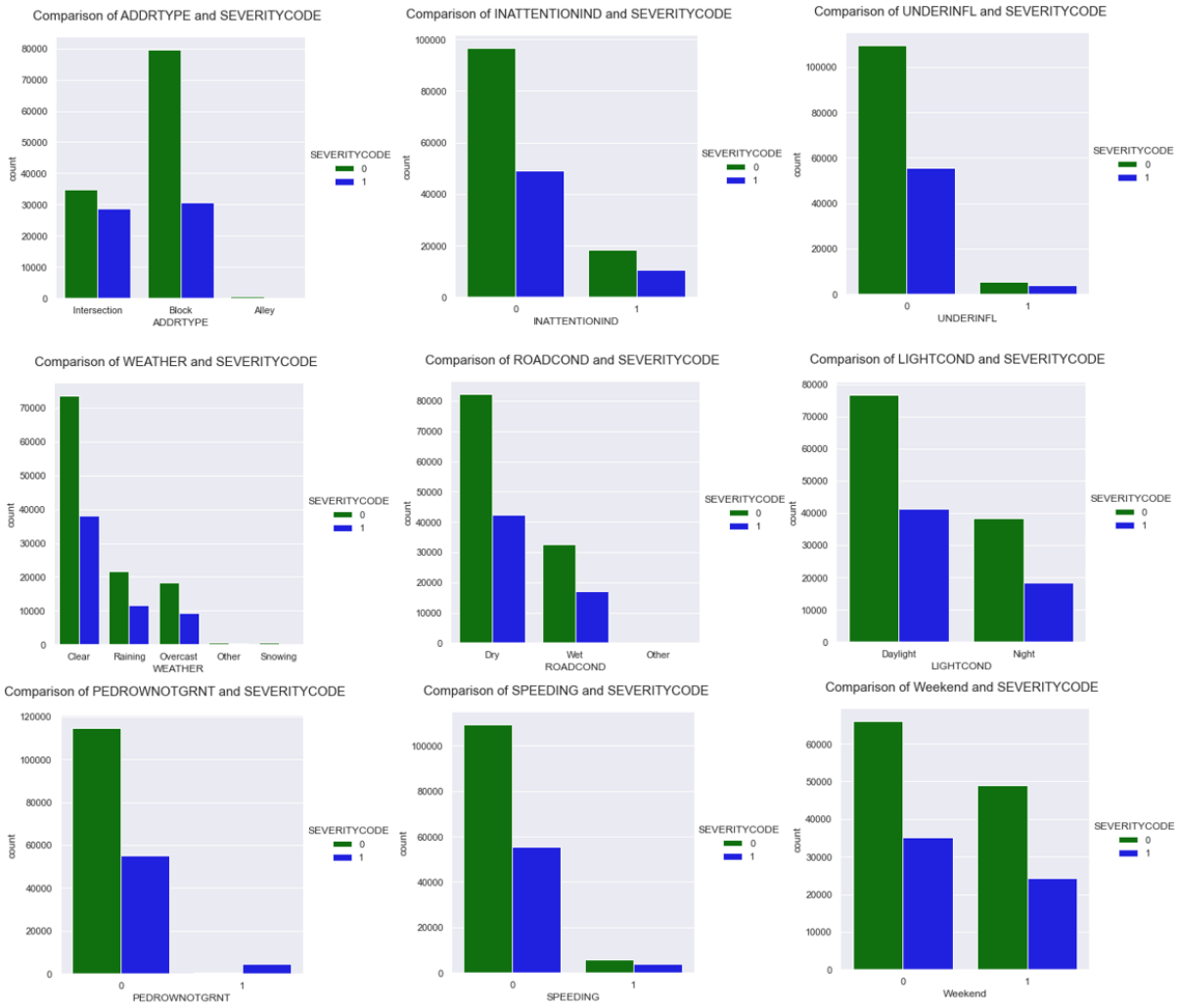

Figure 3: Frequency distribution of Independent Variables compared with the Dependent Variable

As shown in the chart above, these bar charts visually indicated that there may be a relationship between some of these variables. For example, the proportion of collisions resulting in Injury / Death (i.e. SEVERITYCODE = 1) were a visually greater proportion of Intersections (in the ADDRTYPE variable), Driving under the Influence and Speeding.

After visually inspecting these distributions, dummy variables were created for the remaining multi-class categorical variables (ADDRTYPE, WEATHER and ROADCOND).

Pairwise Variable Comparison

Given the binary nature of the variables, a pairwise comparison matrix was run using Hamming distance from the Scikit-Learn library. The results are shown on the following page.

As the results in the table show, many of the variables show high levels of similarity on the Hamming metric. However, this is likely due to the sparsity of the dataset (i.e. the large proportion of ‘0’ values in each variable).

	SEVERITYCODE	INATTENTIONIND	UNDERINFL	PEDROWNOTGRNT	SPEEDING	Weekend	Road-Dry	Road-Other	Road-Wet	Weather-Clear	Weather-Other	Weather-Overcast	Weather-Raining	Weather-Snowing	LightCond-Daylight	LightCond-Night	ADDR-Alley	ADDR-Block	ADDR-Intersection
SEVERITYCODE	1.00	0.61	0.65	0.68	0.65	0.52	0.43	0.66	0.57	0.46	0.66	0.61	0.60	0.66	0.46	0.54	0.66	0.38	0.62
INATTENTIONIND	0.61	1.00	0.79	0.81	0.79	0.55	0.37	0.83	0.63	0.42	0.83	0.73	0.70	0.83	0.40	0.60	0.83	0.45	0.56
UNDERINFL	0.65	0.79	1.00	0.92	0.91	0.59	0.31	0.94	0.69	0.38	0.94	0.80	0.78	0.94	0.29	0.71	0.94	0.39	0.61
PEDROWNOTGRNT	0.68	0.81	0.92	1.00	0.92	0.57	0.30	0.97	0.70	0.37	0.97	0.82	0.79	0.97	0.34	0.66	0.97	0.35	0.65
SPEEDING	0.65	0.79	0.91	0.92	1.00	0.58	0.28	0.94	0.72	0.36	0.94	0.80	0.79	0.94	0.33	0.67	0.94	0.40	0.60
Weekend	0.52	0.55	0.59	0.57	0.58	1.00	0.46	0.58	0.54	0.47	0.58	0.55	0.56	0.58	0.42	0.58	0.58	0.48	0.52
Road-Dry	0.43	0.37	0.31	0.30	0.28	0.46	1.00	0.28	0.00	0.87	0.29	0.31	0.10	0.28	0.65	0.35	0.29	0.56	0.44
Road-Other	0.66	0.83	0.94	0.97	0.94	0.58	0.28	1.00	0.72	0.36	0.99	0.84	0.81	0.99	0.32	0.68	1.00	0.37	0.63
Road-Wet	0.57	0.63	0.69	0.70	0.72	0.54	0.00	0.72	1.00	0.13	0.71	0.69	0.90	0.72	0.35	0.65	0.71	0.44	0.56
Weather-Clear	0.46	0.42	0.38	0.37	0.36	0.47	0.87	0.36	0.13	1.00	0.36	0.20	0.17	0.36	0.61	0.39	0.36	0.54	0.46
Weather-Other	0.66	0.83	0.94	0.97	0.94	0.58	0.29	0.99	0.71	0.36	1.00	0.84	0.80	0.99	0.32	0.68	0.99	0.37	0.63
Weather-Overcast	0.61	0.73	0.80	0.82	0.80	0.55	0.31	0.84	0.69	0.20	0.84	1.00	0.65	0.84	0.38	0.62	0.84	0.41	0.59
Weather-Raining	0.60	0.70	0.78	0.79	0.79	0.56	0.10	0.81	0.90	0.17	0.80	0.65	1.00	0.80	0.34	0.66	0.81	0.41	0.59
Weather-Snowing	0.66	0.83	0.94	0.97	0.94	0.58	0.28	0.99	0.72	0.36	0.99	0.84	0.80	1.00	0.32	0.68	0.99	0.37	0.63
LightCond-Daylight	0.46	0.40	0.29	0.34	0.33	0.42	0.65	0.32	0.35	0.61	0.32	0.38	0.34	0.32	1.00	0.00	0.33	0.53	0.47
LightCond-Night	0.54	0.60	0.71	0.66	0.67	0.58	0.35	0.68	0.65	0.39	0.68	0.62	0.66	0.68	0.00	1.00	0.67	0.47	0.53
ADDR-Alley	0.66	0.83	0.94	0.97	0.94	0.58	0.29	1.00	0.71	0.36	0.99	0.84	0.81	0.99	0.33	0.67	1.00	0.36	0.63
ADDR-Block	0.38	0.45	0.39	0.35	0.40	0.48	0.56	0.37	0.44	0.54	0.37	0.41	0.41	0.37	0.53	0.47	0.36	1.00	0.00
ADDR-Intersection	0.62	0.56	0.61	0.65	0.60	0.52	0.44	0.63	0.56	0.46	0.63	0.59	0.59	0.63	0.47	0.53	0.63	0.00	1.00

Table 2: Pairwise comparison between all variables using Hamming Distance metric

3.2 Modelling

After the creation of dummy variables, the modelling dataset consisted of 1 Dependent Variables (SEVERITYCODE) and eighteen Independent Variables. All variables were binary variables.

Train and test samples

Prior to modelling and sample balancing, the data was split into training and test samples, using a ratio of 5:1 (i.e. 20% test sample). This resulted in the following sets:

A training set of 139,561 records where 91,822 cases of the Dependent Variable were '0' and 47,739 were '1'.

A test set of 34,891 records where 23,127 cases of the Dependent Variable were '0' and 11,764 were '1'.

Sample balancing

As noted earlier, the dataset is imbalanced with 34% of cases representing a road collision resulting in injury / death (SEVERITYCODE=1) and 66% representing collisions resulting in property damage only.

It is recommended that imbalanced datasets are balanced in respect of the values in the Dependent Variable. Based on a review of the literature, three strategies were identified for balancing a dataset.

- Randomly **Over-sampling** the minority class (i.e. SEVERITYCODE=1)
- Randomly **Under-sampling** the majority class (i.e. SEVERITYCODE=1)
- Generating synthetic data to up-sample the minority class (using SMOTE)

These methods were all called from the imbalanced-learn library. In order to assess the impact on the accuracy of the prediction algorithms, each algorithm was run using sample generated by each of these sample balancing strategies, in addition to the original, unbalanced sample.

Given the relatively large proportion of the minority class and the relatively large size of the data set, it was considered interesting to see the extent to which sample balancing strategies had a positive impact on the accuracy scores for our prediction algorithms.

Prediction algorithms

In order to develop an algorithm to predict collisions causing injury / death rather than just property damage, 4 prediction algorithms were explored:

- K Nearest Neighbour (KNN) classification
- Support Vector Machines (SVM)
- Decision Tree
- Logistic Regression

The objective of this part of the project was to compare the accuracy result of each of these algorithms to each other and select the best performing algorithm, which would then be used to understand the relative contribution of different factors in causing collisions with these outcomes. The 'F1 score' is compatible with all four algorithms and was selected as the main measure of accuracy performance for this project.

However, one of the objectives for this study was to identify the relative contribution of different factors and this requires the coefficients to be interpretable. Based upon my reading of the literature, this is not true of the KNN algorithm, so this was only included to provide accuracy scores that could be used to compare with the other algorithms. Even if it proved to be the best performing algorithm, it would not be selected as the 'winner'.

As mentioned earlier, all four classification / prediction algorithms were run four times, on each of the four training data sets:

- The original unbalanced dataset,
- The over-sampled dataset,
- The under-sampled dataset, and
- The synthetically up-sampled dataset (using SMOTE)

In order to establish the best performing **KNN** algorithm, the classifier was run with default settings for every k 1-12. The accuracy score, Jaccard Score and F1 Score were calculated for each and the best performing k was selected.

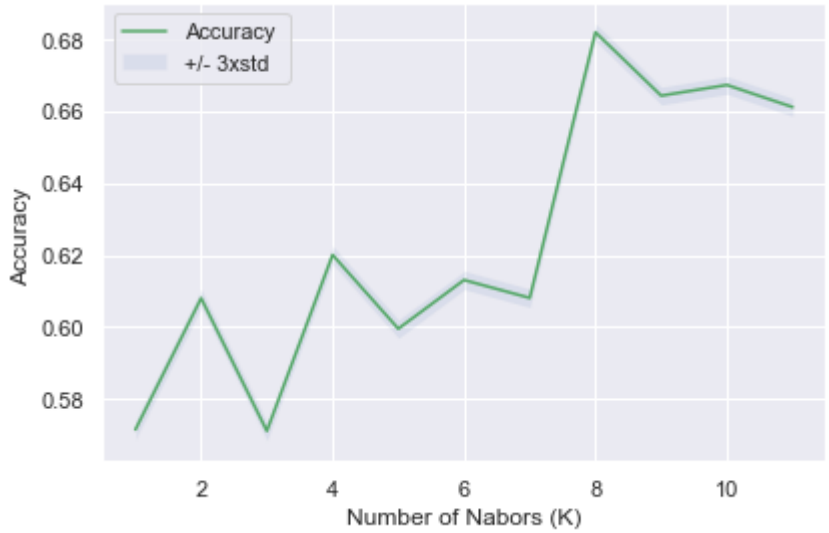

Figure 4: Accuracy of the KNN algorithm run on the Original dataset by Number of Nabors

This was done for all four datasets. The best performing classifier was the one run on the Original (unbalanced) dataset where k=8 which delivered a F1 score of 0.6228

Scores by Sample	Jaccard Score	F1 Score
Original / Imbalanced	0.194961	0.622836
Oversampled	0.161489	0.609429
Undersampled	0.083977	0.575888
SMOTE	0.176027	0.614389

Table 3: Jaccard and F1 scores for KNN algorithm

To establish the best performing **SVM** algorithm, the 'rbf' kernel was used with the default settings, with the Jaccard Score and F1 Scores calculated. Again this was done for each of the four training datasets. The best performing SVM algorithm was the one run on the Under-sampled training set with a F1 score of 0.6240.

Scores by Sample	Jaccard Score	F1 Score
Original / Imbalanced	0.098366	0.594983
Oversampled	0.334384	0.610955
Undersampled	0.327735	0.624016
SMOTE	0.332221	0.614423

Table 4: Jaccard and F1 scores for SVM algorithm

To establish the best performing **Decision Tree** algorithm, the 'entropy' criterion and max depth of 8 nodes was used with the default settings, with the Accuracy score, Jaccard Score and F1 Scores calculated. Again this was done for each of the four training datasets. The best performing Decision Tree algorithm was the one run on the SMOTE synthetically up-sampled training set with a F1 score of 0.6244.

Scores by Sample	Jaccard Score	F1 Score	Accuracy Score
Original / Imbalanced	0.096147	0.593774	0.686653
Oversampled	0.344958	0.599331	0.588175
Undersampled	0.334785	0.599331	0.601530
SMOTE	0.330017	0.624390	0.615861

Table 5: Jaccard, F1 and Accuracy scores for Decision Tree algorithm

To establish the best performing **Logistic Regression** algorithm, the 'lgfbs' solver and C=0.01 was used with the default settings, with the LogLoss, Jaccard Score and F1 Scores calculated. Again this was done for each of the four training datasets. The best performing Logistic Regression algorithm was the one run on the SMOTE synthetically up-sampled training set with a F1 score of 0.6278.

Scores by Sample	Jaccard Score	F1 Score	LogLoss Score
Original / Imbalanced	0.102892	0.596860	0.603583
Oversampled	0.326980	0.625437	0.653892
Undersampled	0.325912	0.627413	0.654485
SMOTE	0.325386	0.627835	0.654021

Table 6: Jaccard, F1 and LogLoss scores for Logistic Regression algorithm

5. Results

Having evaluated 4 classifier / prediction algorithms across 4 sample balancing strategies, the F1 score for each combination was compared and is shown in the table below:

F1 Scores by Sample & Algorithm	Original Imbalanced	Oversampled	Undersampled	SMOTE upsampled
KNN	0.622836	0.609429	0.575888	0.614389
SVM	0.594983	0.610955	0.624016	0.614423
Decision Tree	0.593774	0.599331	0.611951	0.624390
LogisticRegression	0.596860	0.625437	0.627413	0.627835

Table 7: F1 scores for all algorithms and Sample Balancing strategies

The best performing combination was the Logistic Regression algorithm applied to the SMOTE up-sampled training set, which delivered a F1 score of 0.6278.

The following Classification Report for the Logistic Regression based on the SMOTE up-sampled training set shows that the algorithm performed moderately well. The Precision score of 0.45 (in predicting '1' – collisions resulting in Injury / Fatality) indicates that the model was identify a relatively large proportion of False Positives – over half (55%) of the predicted '1's were incorrectly classified.

	precision	recall	f1-score	support
0	0.74	0.66	0.70	23127
1	0.45	0.54	0.49	11764
accuracy			0.62	34891
macro avg	0.59	0.60	0.59	34891
weighted avg	0.64	0.62	0.63	34891

Table 8: Classification report for Logistic Regression using SMOTE up-sampling

The Recall score of 0.54 (in predicting '1' – collisions resulting in Injury / Fatality) indicates the algorithm performed marginally better in terms of identifying the actual collisions resulting in injury or death; 54% of these were correctly classified as such while 46% were incorrectly classified as causing property damage only.

The following Confusion Matrix shows these same results in a volumetric format for the test set using the Logistic Regression developed using the SMOTE up-sampled training set:

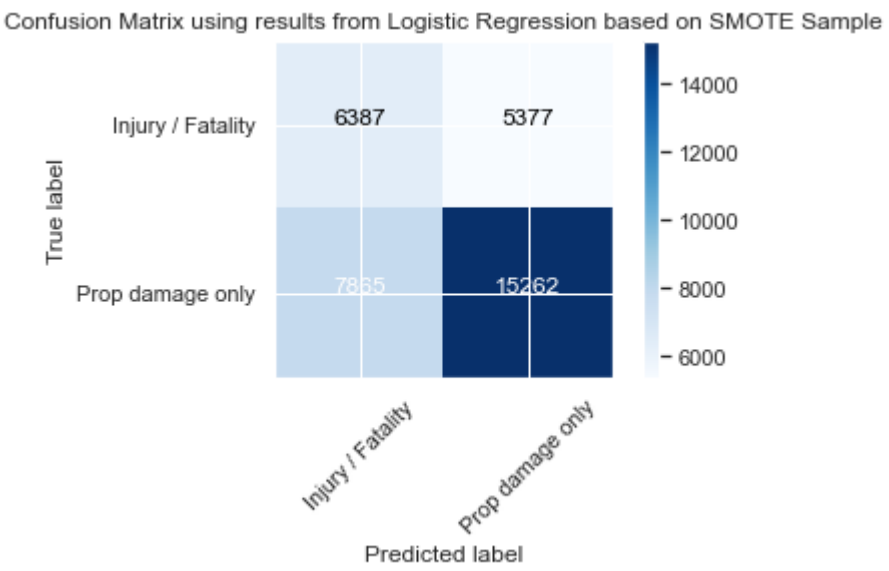

Figure 5: Confusion Matrix for Logistic Regression using SMOTE up-sampling

Given the purpose of this model is to help us understand the contributing factors towards serious car collisions, the ‘cost’ of false positives and false negatives is not great (unlike a medical trial, for example). However, these measures do speak to a model that has only moderate effectiveness in predicting these outcomes of interest. This will be discussed further in the conclusions.

Finally, the coefficients generated by the Logistic Regression based on the SMOTE up-sampling strategy were reported:

Logistic Regression Coefficients	Coefficients
PEDROWNOTGRNT	2.313794
ADDR-Intersection	0.602323
UNDERINFL	0.528675
SPEEDING	0.475095
INATTENTIONIND	0.302761
Weather-Raining	0.130155
LightCond-Daylight	0.085672
Weather-Clear	0.071903
Weather-Overcast	0.063506
Weather-Other	0.06108
Road-Other	0.026697
Road-Dry	-0.0342
Weekend	-0.04642
Road-Wet	-0.07716
LightCond-Night	-0.08358
ADDR-Block	-0.12589
Weather-Snowing	-0.35971
ADDR-Alley	-0.5608

Table 9: Logistic Regression Coefficients based on SMOTE up-sampling strategy

These results indicate that the following variables, in particular, have a notable contribution to road collisions in Seattle that result in Injury / Fatality:

- Pedestrian Right of Way Not Granted (PEDROWNOTGRNT)
- Collisions at an Intersection (ADDR-Intersection)
- Driving under the influence (UNDERINFL)
- Speeding (SPEEDING)
- Inattention (INATTENTIONIND)

6. Discussion

In developing an algorithm to predict the road collisions in Seattle that result in Injury / Death rather than property damage only, four different algorithms were evaluated using four different sample balancing strategies.

Interestingly, the results were broadly similar, ranging from 0.5759 to 0.6278 with an average score across the 16 combinations of 0.6109.

Overall the SMOTE up-sampling approach performed best of the sample balancing strategies producing the highest F1 score on 2 occasions and 2nd highest on the other 2. In contrast, the unbalanced original sample performed worst (4th highest F1 score on 3 occasions and highest on one).

In terms of the algorithms, the Logistic Regression algorithm performed best delivering the highest F1 score on 3 occasions (2nd on the other) whilst the Decision Tree performed least well (4th highest F1 score on 2 occasions, 3rd on one and 2nd highest on one).

The best performing combination was the Logistic Regression algorithm applied to the SMOTE up-sampled training set, which delivered a F1 score of 0.6278. It is likely that this result could be further improved through the use of Grid-Search to optimise the Logistic Regression hyper-parameters, however this is out of scope for this project.

However, further analysis of this model indicated that it performed only moderately well, with relatively high proportions of false positives and false negatives (as shown in the Classification report and Confusion matrix). Although, as noted, given the purpose of the model primarily explanatory – to help understand the contributing factors towards serious car collisions – the ‘cost’ of false positives and false negatives is not as great as in some use-cases (such as a medical trial, for example).

One potential reason for the moderate performance of the model is that the dataset is not comprehensive and many contributing factors are not included. The causes of road collisions are often complex and it may not be possible to accurately record or even collect these factors.

However, the model is useful in highlighting a number of key contributing factors that can be demonstrated to have a statistical relationship with the outcome we are trying to predict, and these can be used to inform the policy and strategy development of relevant authorities in this domain.

7. Conclusions

In this project, I analysed data on road collisions in Seattle in order to identify and better understand the contributing factors associated with collisions that result in Injury or death. I explored a range of possible prediction algorithms and several sample balancing strategies in order to optimise the accuracy of the algorithm. Despite this, the model only achieved a moderate level of performance. This is likely due to the fact that the dataset itself does not comprehensively capture all of the possible contributing factors to the severity of a road collision.

However, the model has proven useful. It identified a number of important contributing factors that have a statistical relationship with road collisions resulting in serious (negative) outcomes. These included:

- Pedestrian Right of Way Not Granted
- Collisions at an Intersection
- Driving under the influence
- Speeding
- Inattention

These insights can now be used by the relevant authorities in Seattle to help guide the process of developing policies and strategies to reduce the number and severity of road collisions in the city.

Further future improvements to this model could be achieved in a number of ways. In particular through:

- Reviewing the data collected by authorities on road collisions. If it is possible to collect more data on the potential causes and factors present in road collisions, it may be possible to improve both the accuracy of the model and the direction it provides to authorities.
- Using Grid Search or other similar techniques to tweak the hyper-parameters used in the modelling process. This will also help further optimise the accuracy of the model predictions.