Car Accident Severity Prediction

Alex Lee

7 Oct 2020

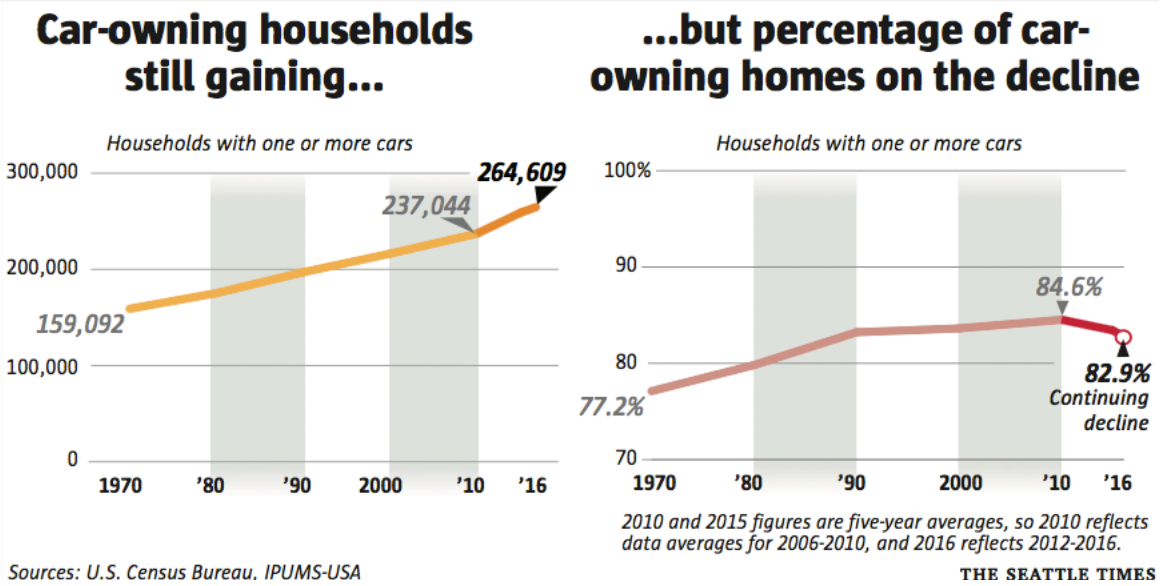
INTRODUCTION

1. The IBM Data Science Professional Certificate Capstone Project is built based on the collision dataset from the Seattle Police Department. This collision set will be used to predict the severity of traffic accidents.

BUSINESS UNDERSTANDING

2. Seattle, Washington. Seattle[1] is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th largest in the United States.

3. About Seattle's Car Population. The total number of personal vehicles in Seattle hit a new high of nearly 444,000 in 2016. And about 265,000 Seattle households have at least one car, also a record[2]. The increase in car density in Seattle streets meant a higher probability of accidents, which also necessarily led to higher occurrences of severe collisions. Such collisions are experienced as damages to property at the lower end of the severity scale, and loss of lives at the other end of the spectrum.



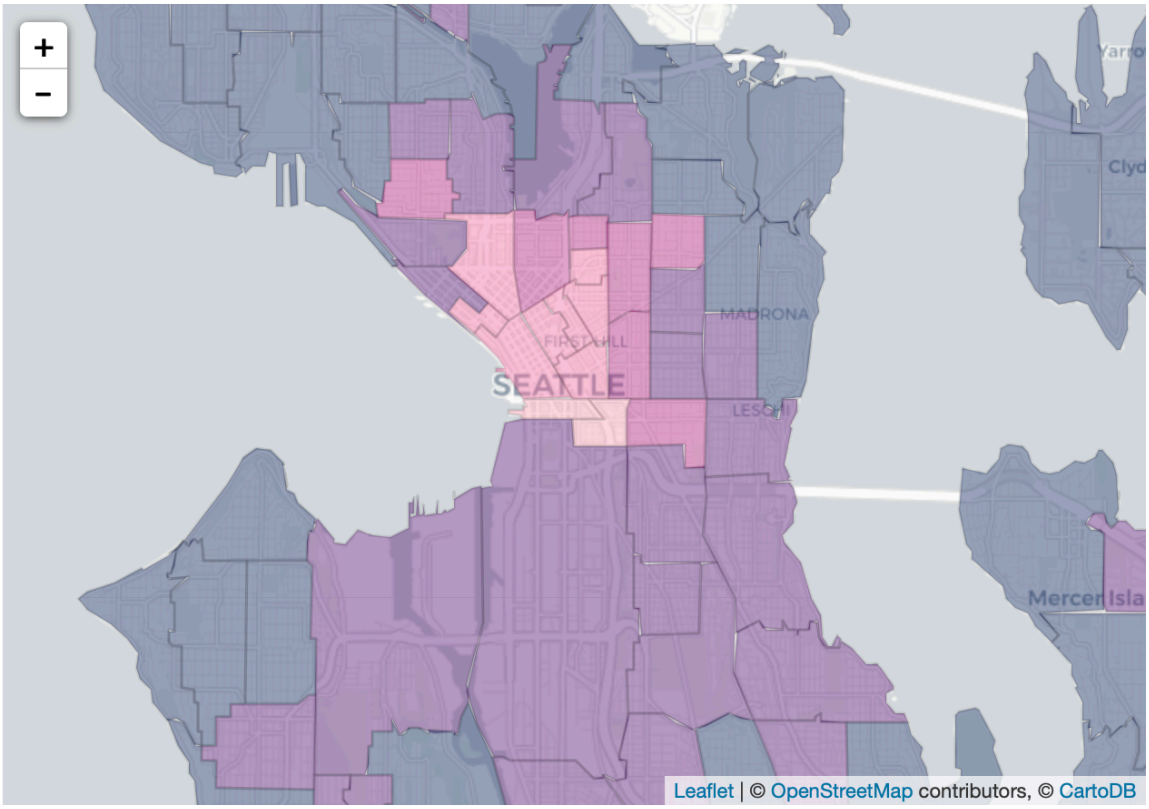
[1] https://en.wikipedia.org/wiki/Seattle, extracted 4 Oct 2020.

[2] Housing cars or housing people? Debate rages as number of cars in Seattle hits new high (https://www.seattletimes.com/seattle-news/data/housing-cars-or-housing-people-debate-rages-as-number-of-cars-in-seattle-hits-new-high/), Updated April 9, 2018 at 9:59 am, extracted 4 Oct 2020.

4. Business Objectives. The economic and societal harm from motor vehicle crashes amounted to a $871 billion in a single year, according to the National Highway Traffic Safety Administration. In a study which examined the economic toll of car and truck crashes in 2010, when 32,999 people were killed, 3.9 million injured and 24 million vehicles damaged, $277 billion was attributed to economic costs. Harm from loss of life, pain and decreased quality of life due to injuries was pegged at $594 billion[3].

Where are our cars?

Downtown Seattle and the University District contain the only census tracts in the city where the majority of households don't own a car.



HOUSEHOLDS WITH ONE OR MORE CARS

91% and up | 76- 90% | 61-75% | 46-60% | 0-45%

Source: U.S. Census

EMILY M. ENG / THE SEATTLE TIMES

5. Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. Accurate predictions of severity can provide crucial information for emergency responders to

[3] Traffic accidents in the U.S. cost $871 billion a year, federal study finds (https://www.pbs.org/newshour/nation/motor-vehicle-crashes-u-s-cost-871-billion-year-federal-study-finds), extracted 6 Oct 2020, last updated 29 May 2014.

evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures[4].

6. Business Value. Knowing the contributing factors to an accident, as well as the corresponding severity can help direct public resources to mitigate the risks and probability to where it matters most, and reduce the occurrence of accidents. Sharing processed data of contributing factors with the public through education and awareness programs, including the use of appropriate street signs in accident-prone areas will also contribute to reducing accident rates.

DATA UNDERSTANDING AND PREPARATION

7. Data Source. The dataset used in this analysis is provided for, and is available as an open source download from the Seattle Police Department. The collisions data spans from 2004 to present.

8. General Overview. The dataset consists 38 columns, and 194,674 rows of data. This translates to an average of 10,800 accidents per year. In the interest of this project, not all features are useful in the prediction of accident severity, and will be managed accordingly during the data preparation section.

9. Data Wrangling. Based on the 38 columns of data, a significant amount of it are present as strings, and will require conversion to numerical formats for correlation and model fitting. Data will significant amounts of missing fields will also be dropped as they will skew the outcome of the results.

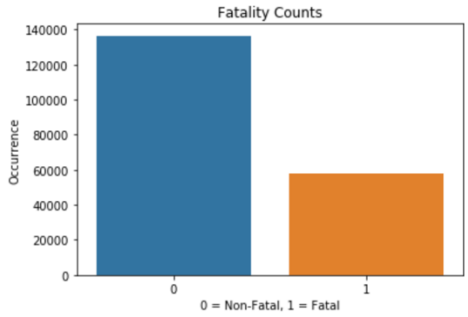
10. Initial Data Processing. The following tables showed the columns that were retained for further analysis, as well as those that were removed from the raw data set as they do not possess significantly useful information.

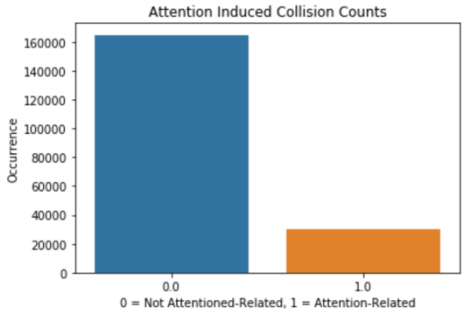
FEATURE SETS RETAINED		
X	Y	ADDRTYPE
LOCATION	SEVERITYCODE.1	SEVERITYDESC
COLLISIONTYPE	PERSONCOUNT	PEDCOUNT
PEDCYLCOUNT	VEHCOUNT	JUNCTIONTYPE
INATTENTIONIND	UNDERINFL	WEATHER
ROADCOND	LIGHTCOND	SPEEDING

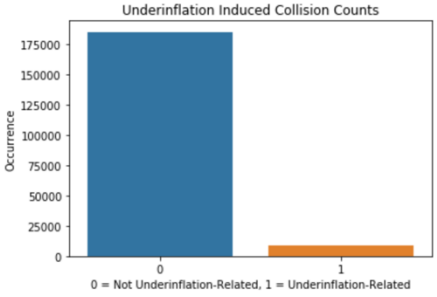
COLUMNS REMOVED		
OBJECTID	INCKEY	COLDETKEY
REPORTNO	STATUS	INTKEY
EXCEPTRSNCODE	EXCEPTRSNDESC	INCDATE
INCDTTM	SDOT_COLCODE	SDOT_COLDESC
PEDROWNOTGRNT	SDOTCOLNUM	ST_COLCODE
ST_COLDESC	SEGLANEKEY	CROSSWALKKEY
HITPARKEDCAR		

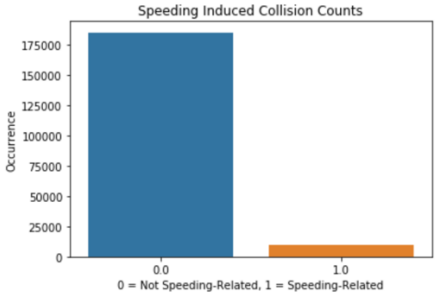
[4] Fang Zong, Huiyong Zhang, Hongguo Xu, Xiumei Zhu, Lu Wang, "Predicting Severity and Duration of Road Traffic Accident", Mathematical Problems in Engineering, vol. 2013, Article ID 547904, 9 pages, 2013. https://doi.org/10.1155/2013/547904

11. Fatality Counts against Contributing Factors. According to the four histograms plotted above, fatality count is significantly higher from attention-induced collisions than speeding and under-influence factors.







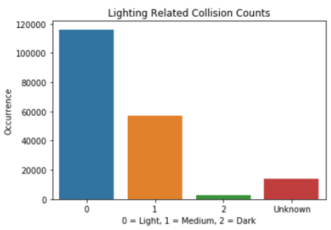


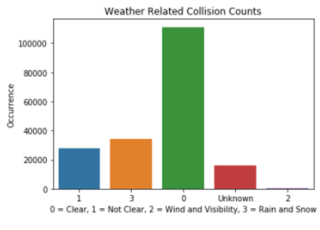
12. Collision Occurrences against Lighting, Weather and Road Conditions. The following observations could be made regarding the various contributing conditions.

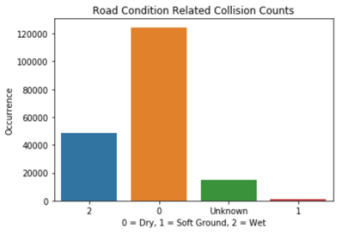
    a. Lighting Condition. Interestingly, most collisions happen during daylight hours, when lighting should be most optimum. Correspondingly, fatality rate was also higher during the day, contributing almost 50% of fatality for all daylight collisions. Collisions at night was significantly lower.

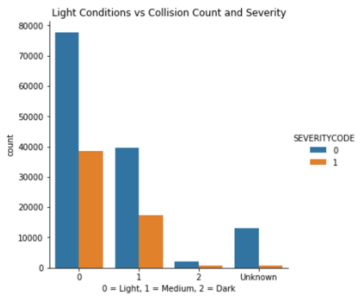
    b. Weather Condition. Most collisions happen when weather was clear, contrasting greatly with the number of collisions during foul weather.

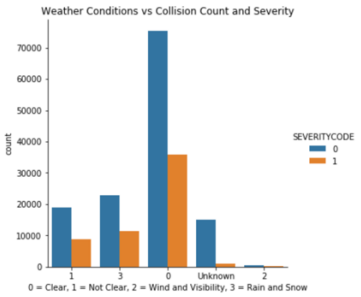
    c. Road Condition. When road conditions were less than ideal, accident rates were much lower compared to when the road was dry.

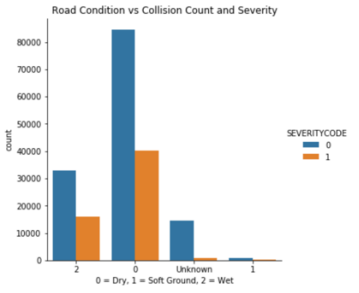




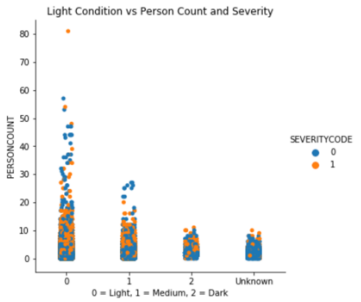


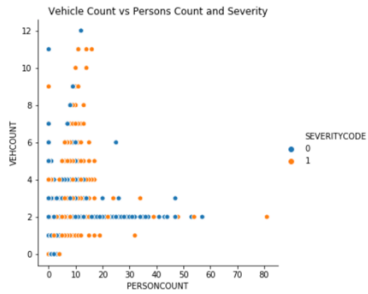


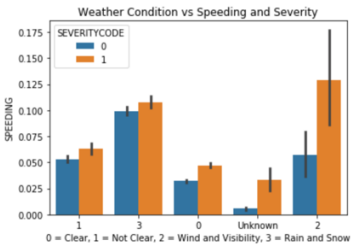




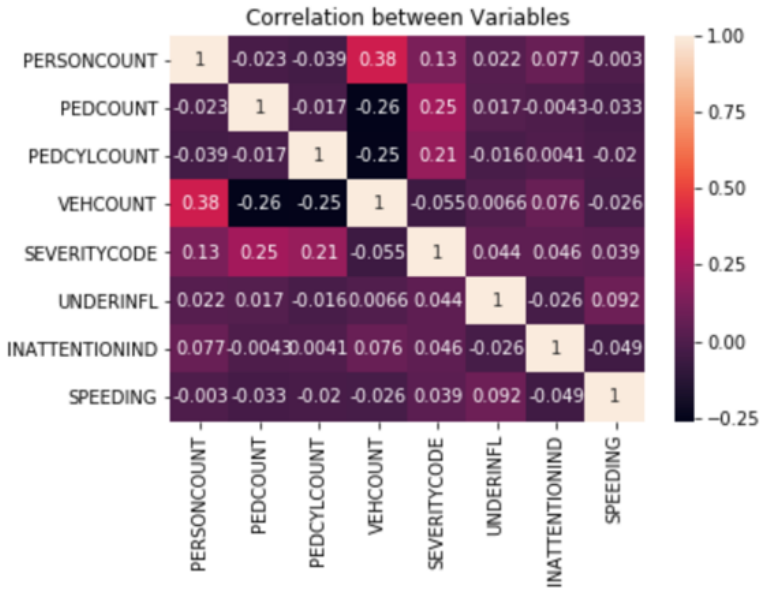
13. Severity of Collisions versus Multiple Variables. The accidents with the highest head count involved happened mostly in daylight hours, with vehicle counts hovering between 1 to 6. Speeding is a more significant contributor when weather conditions are due to wind and low visibility, with a higher count of fatality.



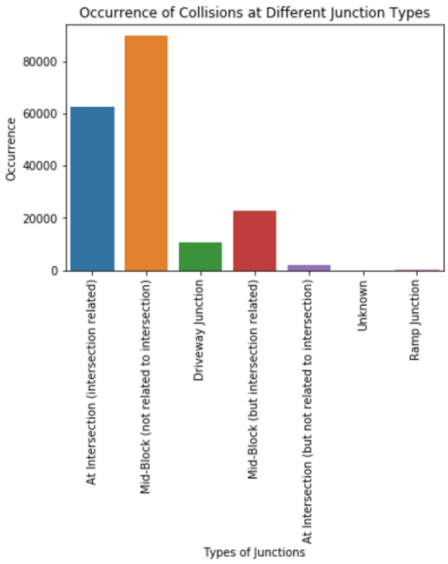


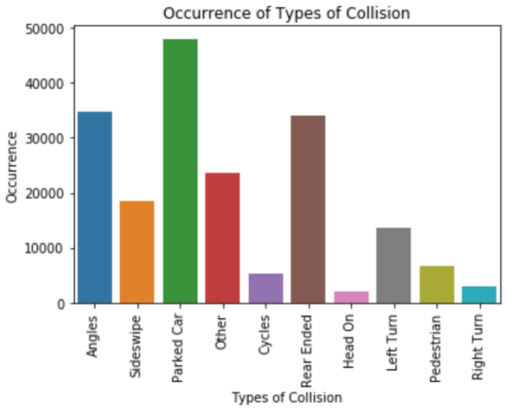


14. Correlation between Variables. Despite what seems to be a logical deduction, the data does not indicate strong correlation between the severity of accidents to the presence of inattention, under-influence, nor speeding. The strongest correlation between the variables shown exist only with vehicle counts and person counts at 0.38, which is sensible, since more vehicles means more people involved.



15. Categorical Visualisation of Data. From the three bar plots, most collisions happen at mid-block sections, with the highest occurrences involving parked cars.

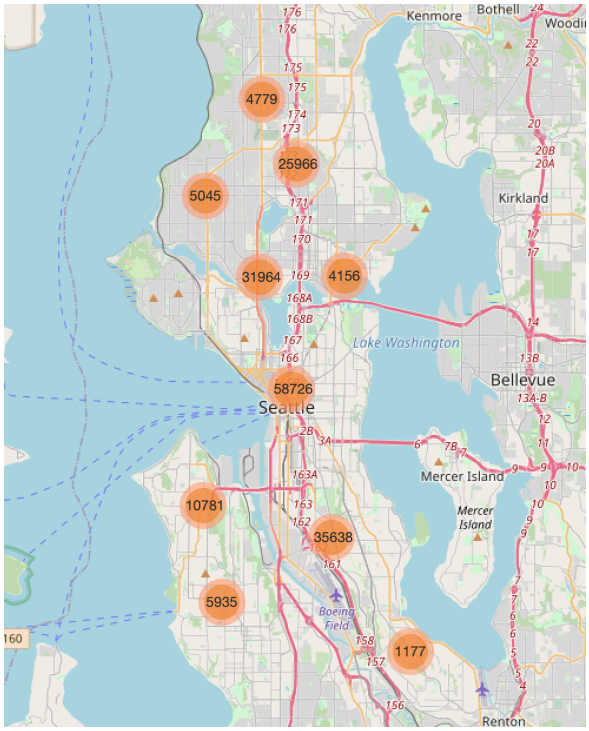




16. Folium Cluster Visualisation. Folium was used to visualise the occurrence of all 194,673 datasets on the map of Seattle to help build an appreciation of the locations of the collisions from 2004 till present.

17. It can be seen that a majority of the collisions occur along the freeway, which may be a consequence of the expected higher traffic volumes around those roads.

18. Downtown Seattle also boast the highest numbers at 58,726, evident of its higher road density and vehicular volume as well. The further we go from the city centre, the lower the number is, unless it is in vicinity of the freeway. The number was lowest at 1,177 collisions in the outskirts, between the areas of Boeing Field and Renton.



DATA MODELLING

19. Data Split. Train-test-split was used to separate the data into 90% training and 10% testing data sets.

20. Model Selection. The following models are considered for prediction of fatality outcomes Y from the multi-variable input of the values of X, where X refers to the status of the road, light and weather conditions.

a. K-Nearest Neighbour Classifier (KNN)[5]. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. The quality of the predictions depends on the distance measure, and is appropriate when sufficient domain knowledge is available. Although this method increases the costs of computation compared to other algorithms, KNN is still the better choice for applications where predictions are not requested frequently but where accuracy is important.

b. Decision Tree Classifier[6]. Decision trees belong the most accurate classifiers. Unlike other algorithms, decision trees can handle discrete attributes and numerical attributes by using them under split conditions that are represented by symbols. Another advantage is the human readability of decision tree models. Particularly applications that require both, high accuracy and interpretability of the classification model, can use decision trees. Such applications are, for example, customer classification, fraud detection, and diagnostics.

c. Logistic Regression Classifier[7]. Logistic regression is an algorithm that is used in solving classification problems. It is a predictive analysis that describes data and explains the relationship between variables. Logistic regression is applied to an input variable (X) where the output variable (y) is a discrete value which ranges between 1 (yes) and 0 (no). It uses logistic (sigmoid) function to find the relationship between variables. The sigmoid function is an S-shaped curve that can take any real-valued number and map it to a value between 0 and 1, but never exactly at those limits.

d. Support Vector Machines (SVM)[8]. The support vector machine is a model used for both classification and regression problems though it is mostly used to solve classification problems. The algorithm creates a hyperplane or line (decision boundary) which separates data into classes. It uses the kernel trick to find the best line separator (decision boundary that has same distance from the boundary point of both classes).

21. Model Evaluation Metrics. The following metrics are used to assess the accuracy of the models with the test data.

    a. Jaccard Index[9]. 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

[5] Usage of KNN (https://www.ibm.com/support/knowledgecenter/SSCJDQ/com.ibm.swg.im.dashdb.analytics.doc/doc/r_knn_usage.html), extracted 7 Oct 2020.
[6] Usage of Decision Trees (https://www.ibm.com/support/knowledgecenter/SSCJDQ/com.ibm.swg.im.dashdb.analytics.doc/doc/r_decision_trees_usage.html), extracted 7 Oct 2020.
[7] Logistic Regression vs Support Vector Machines (SVM) (https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16), extracted 7 Oct 2020.
[8] Logistic Regression vs Support Vector Machines (SVM) (https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16), extracted 7 Oct 2020.
[9] Jaccard Index (https://deepai.org/machine-learning-glossary-and-terms/jaccard-index), extracted 7 Oct 2020.

defined as the size of the intersection divided by the size of the union of the sample sets.

b. F1-Score[10]. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

   i. Precision. Precision is the fraction of true positive examples among the examples that the model classified as positive. In other words, the number of true positives divided by the number of false positives plus true positives.

   ii. Recall. Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples. In other words, the number of true positives divided by the number of true positives plus false negatives.

c. Log Loss. A loss function measures the discrepancy between the true values (observations) and their estimated fits, for a given instance of data. The greater the log loss, the greater the discrepancy.

22. Model Evaluation Results. Tests were generated with the following accuracy results from the above mentioned tests. It is noted that the models performed the same for the input data set and test data. Only SVM with Sigmoid kernel performed less optimally than the rest in comparison.

Metrics	KNN	Dec Tree	Log Reg	SVM	
Solver	N=2	D=7	All solvers	Others	sigmoid
Jaccard	0.699	0.699	0.699	0.699	0.648
F1	0.823	0.823	0.823	0.823	0.602
Log Loss			0.611		

EVALUATION

23. Evaluation of Data Sets. It is important to note that the data set that we have from SPD was not a balanced dataset for the target variable, skewed in favour of non-fatal accidents. Larger datasets could also aid in the training of the models, which can yield better results. There were also a large amount of information missing from SPEEDING and UNDERINFL. Some other information, if present, can also provide greater insights, such as traffic volume and density, drivers' age and experience, whether drivers practise defensive driving/riding, etc.

24. Evaluation of Models. As all models provided similar results, the ideal choice of modeller to be used shall be the decision tree model, since the data sets consists discrete values, and the logic is simple to understand by the human user. Though logistic regression would have been a good model as well, the log loss experienced at 0.611 is high, and further refinement will be required. KNN is also a plausible choice since N=2 is computationally viable. Both Decision Tree and KNN classifiers can be used in tandem if desired.

[10] F-Score (https://deepai.org/machine-learning-glossary-and-terms/f-score), extracted 7 Oct 2020.

PLAN DEPLOYMENT

25. Recommendations. With the insights gained from the analysis of the data set, we can consider the following interventions to reduce the occurrences of accidents in the high probability areas and conditions.

    a. Education of Drivers. With a majority of collisions happening during daylight hours, and in good weather and road conditions, it is highly possible that they arise due to drivers' complacency and inattention on the roads. This can be approached with proper education of the drivers, and with focus on regular road safety campaigns. Proper education can also reduce the occurrence of drivers being subjected to speeding, inattention and driving under influence.

    b. Improvement to Road Conditions. Where appropriate, continued road and infrastructure maintenance must take place to ensure that there is sufficient drainage for surface runoffs during rain and snow. Post-accident treatment of road surfaces to reduce the presence of oil should also be done.

    c. Improvement to Lighting Conditions. Street lightings should adhere closely to sunrise and sunset timings to enable smooth transition from light to dusk conditions, so that drivers' eyesight can perform optimally under shifting light conditions.

    d. Collisions with Parked Cars. This is an interesting phenomenon that should be investigated. Enforcement of vehicles keeping within parking spaces, checking that the roads are appropriately wide enough for vehicles to pass in vicinity of parked cars, should also be carried out. Vehicles can also have a mandatory reflector at their corners to improve visibility, especially during conditions of low-lighting so that drivers can be visually alerted to their presence If they are parked in proximity.

    e. Placement of Warning Signs. Road signages can be installed in locations of higher traffic accidents to warn drivers of the dangers on the roads and remind them to be more alert in those areas.

CONCLUSION

26. Analysis of the traffic accident dataset from the SPD has revealed certain insights into the causes of accidents, as well as prompted recommendations on what can be done to improve the situation. The next steps in this project shall be to implement the recommendations, and monitor improvements over time.