Machine Learning in Business

MIS710 – A1

Report on Factors Contributing to Black Spots

TAHSIN AFROZ
223137186

DEAKIN
BUSINESS
SCHOOL

CONTENTS

Introduction

Road safety is a critical concern for any society, impacting public health, infrastructure, and overall quality of life. Accidents and blackspots on roadways pose significant risks to both commuters and pedestrians, underscoring the need for a comprehensive understanding of the factors contributing to their occurrence. VicCrashAnalytics collaborates with Victoria's Department of Transport to analyse and predict blackspots – high-risk accident areas.

Machine Learning techniques simplify the strenuous and time-consuming task of classification. One can use Machine Learning techniques to perform data analysis and extract necessary information from the dataset.

This project is aimed to Black spot classification using Machine Learning model to identify whether an accident spot is a black spot or not.

Executive Summary

Reported to: MR. MICHAEL HOWARDS
(*Transport Analytics Manager*, VicCrashAnalytics)
Client: Depart of Transport

This report aims to provide a comprehensive analysis of blackspots- accident spots, aligning with DOT's road safety focus. By addressing the accurate identification and mitigation of accident-prone areas, this study enhances public safety and infrastructure integrity.

The analysis delves into demographics, road characteristics, and contributing factors, revealing vital insights through exploratory data analysis. Leveraging logistic regression, a predictive mode effectively discerns potential blackspots. Recommendations include converting intersections to roundabouts and implementing alcohol testing near commercial venues. The model offers interpretability, precision, and valuable insights. While effective for resource-friendly applications, its limitations include high-dimensional data challenges and sensitivity to outliers.

In essence, this analysis empowers informed decision-making and action for a safer transportation network, enabling DOT to prioritize interventions and road safety initiatives strategically.

Business Understanding and Problem to Address

This presented report details a thorough blackspot analysis, aligned with DOT's road safety focus. It addresses identifying and mitigating accident-prone blackspots, vital for public safety, infrastructure, and secure transportation commitment.

The business problem centres on the accurate prediction of blackspots by understanding their contributing factors. Through a comprehensive approach, this analysis explores the intersection of road segment demographics and specific characteristics that elevate the risk of accidents. The report delves into the complexities of blackspot occurrence, seeking to uncover actionable insights to guide strategic decisions.

The analysis drives business solutions with specific strategies for blackspot identification and road safety. Recommendations include converting high-traffic intersections to roundabouts and strict alcohol testing near licensed venues. Aimed at the DOT senior management, these actions align with targeted interventions like education campaigns, legislative reforms, and infrastructure enhancements.

In essence, this report serves as a comprehensive and data-driven tool for addressing the critical business problem of blackspot identification and mitigation. The recommended actions are poised to empower senior management in making informed decisions to prioritize road safety across Victoria's transportation network.

Business Understanding and Problem to Address

This section elaborates on the meticulous process employed for data comprehension, refinement, exploration, and the ensuing insights derived from the blackspot dataset, all within the framework of BACCM.

Data Understanding:

The initial steps of the analysis began by thoroughly understanding the dataset, including its composition, variables, and structure. It contains vital data on blackspots, demographics of nearby road segments, and relevant attributes. This groundwork informed the subsequent analytical steps effectively.

Data Cleansing and Preparation:

A crucial aspect of the approach was the rigorous cleansing and preparation of the dataset. By addressing instances of missing data (% age over 65 years and Liqour Licenses) and irrelevant information (ID, road names, family configuration, job position, people speaking only English at home, etc.) the integrity and reliability of the dataset was insured. This meticulous data refinement was pivotal in facilitating an accurate and unbiased analysis.

Exploratory Data Analysis (EDA) and Visualization:

Utilizing a refined dataset, an extensive exploratory data analysis (EDA) unveiled patterns and trends through statistical techniques and visualizations. Data distributions, correlations, and key features were assessed using tools like histograms and heatmaps, exposing hidden insights and relationships within the data.

SEIFA is a numeric measure of the socioeconomic status (SES) of people taking into account things such as education level, income level, house price, rent, mortgage, unemployment, job type etc. SES is classified as high, moderate or low based on the Australia 2001 Socio-Economic Index for Areas. Socioeconomic Index for Areas (1000= Average | >1000 = Above average | <1000 = Less than average).
According to a study published in *Journal of Epidemiology and Community Health (1978)*, October 2009, there is a high risk of crash-related hospitalisation for young drivers from low SES areas and the risk is independent of driving exposure and rural–urban differences. Similar relations can be drawn from the Blackspot data and its

visualisation below (Figure 1). It shows that there is availability of blackspot within 150 m of a road segment whose SEIFA is below average.

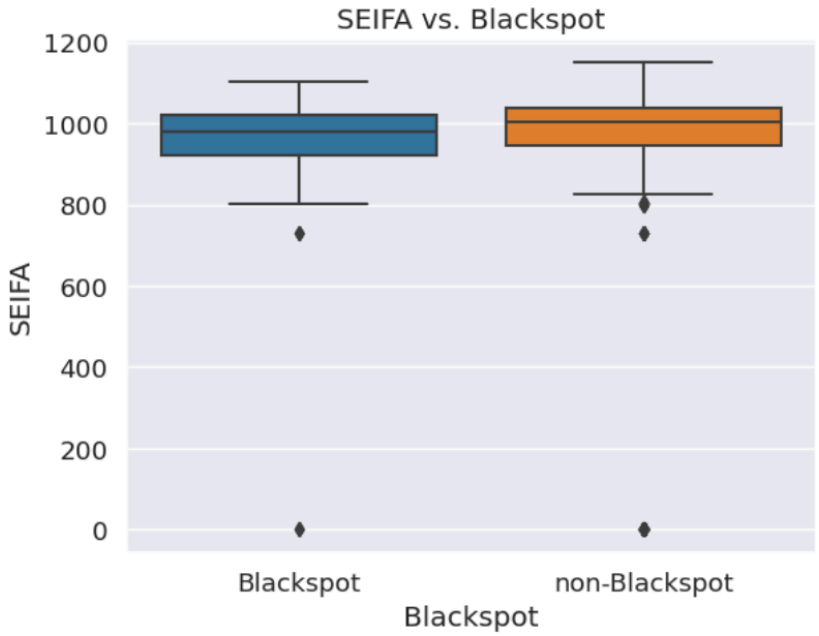


Figure: 1

Number of liquor licensed venues within a road segment broadly contributes to road accidents, therefore making road segments a blackspot. Logistic regression was used to analyze the relationship between number of liquor licensed venues and blackspot. It was found that there is a positive relation between number of liquor licensed venues and blackspot. Moreover, Hobsons Bay City Council in its *Hobsons Bay Liquor Licensing Policy Statement* mention that Research has also established a strong link between the density of liquor licenses and myriad of alcohol related harms (like drink driving accidents).

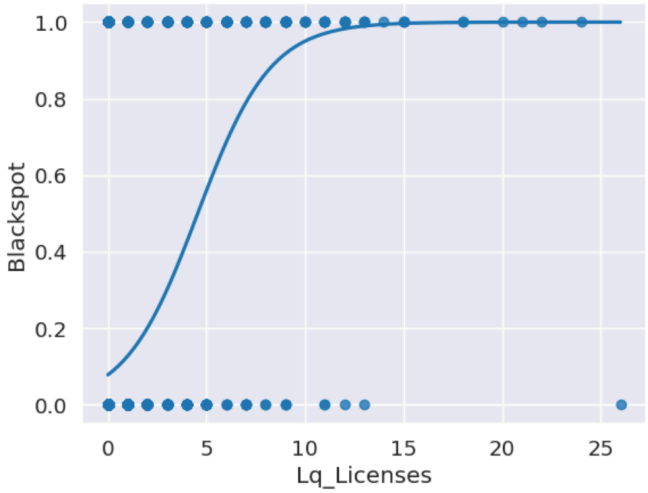


Figure: 2

On analysing the relationship between road types and whether there can be a possible occurrence of blackspot, the result showed that roads and streets are more prone to potential blackspot than the other road types, namely highway, drive, way freeway.

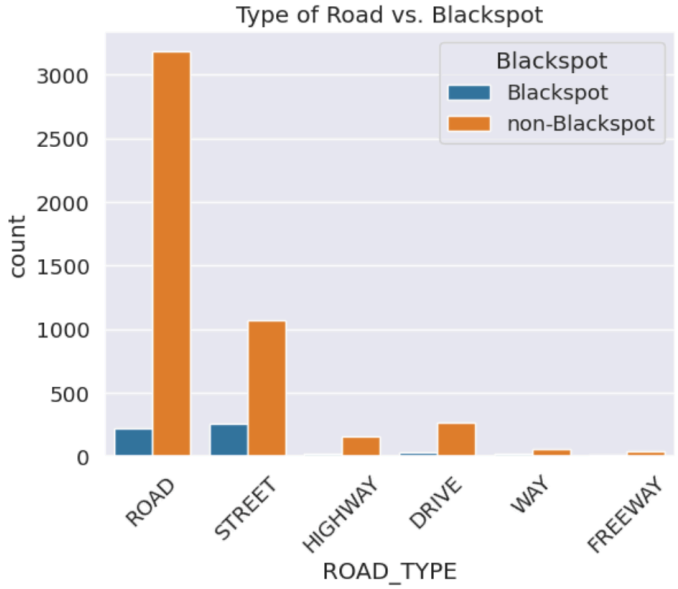


Figure: 3

Intersections are an integral part of the road system but are also one of the most dangerous because they represent a point where road users converge and potentially conflict with each other. In Australia, the majority of urban crashes and a substantial proportion of rural crashes occur at intersections (McLean et al., 2010). It can clearly be observed that the roads with no intersection mostly classify as non-blackspot or

substantially safer road than the roads that have intersections. Roads with intersection are potentially at higher likelihood of being classified into a Blackspot.

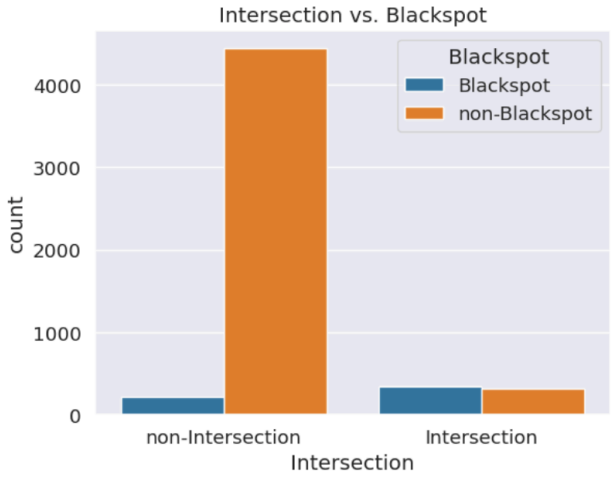


Figure: 4

Commercial land use contains a number of businesses such as office building with different tenancies, café and retail strips etc, and where possible, will have a zero population count. Road segments near commercial areas are accident-prone due to increased traffic, distractions, delivery vehicles, pedestrians, and complex patterns. A strong positive relation can be observed from the boxplot from figure…. below.

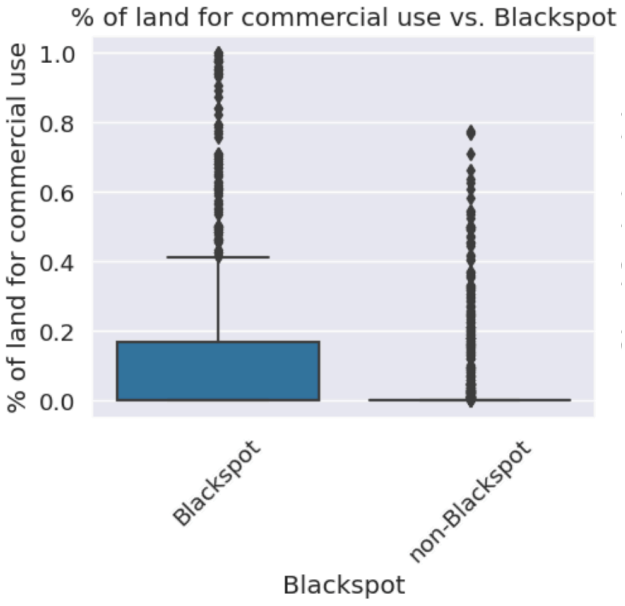


Figure: 5

Heatmap was used for identifying patterns, trends, and correlations within Blackspot dataset, to analyse the complexity of this data to make it understandable and actionable.

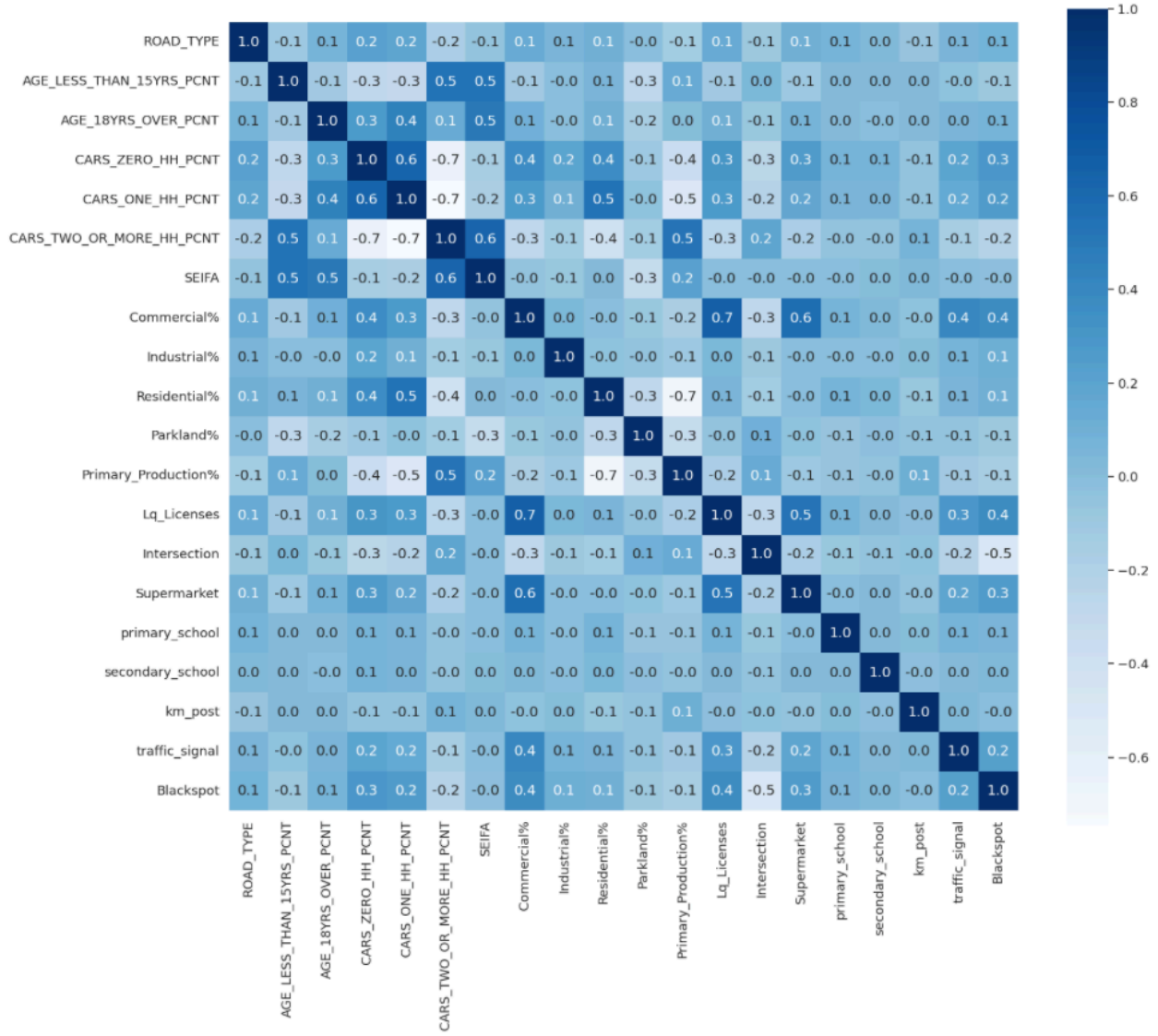


Figure: 6

Insights Gained:

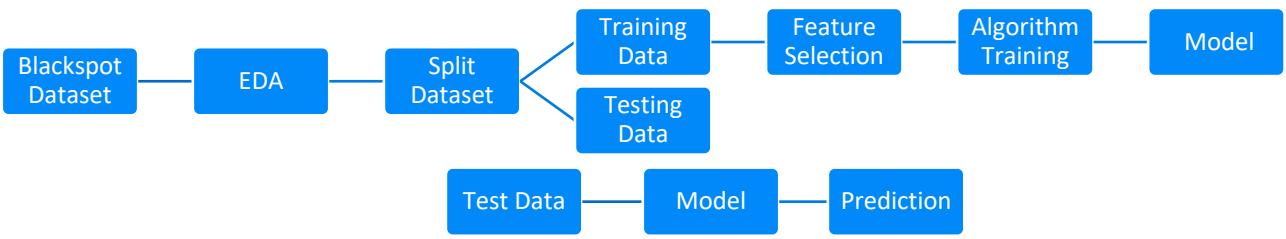
The EDA phase revealed significant insights pertaining to factors contributing to blackspot occurrences. Notably, locales characterized by higher **liquor licensed venues** densities and **intersected road segments** emerged as susceptible to blackspot designation. Correlations between specific road attributes **such as percentage of commercial land in a road segment, road types, SEIFA** and accident frequency

offered crucial insights into potential focal points for intervention. These insights form a cornerstone for informed decision-making and will guide subsequent stages, including predictive **modelling** and strategic recommendations.

The insights derived will underpin the development of predictive models and serve as the bedrock for the formulation of actionable recommendations aimed at addressing the challenges posed by blackspots.

Machine Learning Approach

The endeavour to predict and comprehend blackspots involved a rigorous machine learning approach that seamlessly integrated advanced techniques with comprehensive data analysis. The Machine Learning used in the corresponding analysis involved Supervised Learning wherein the model was trained to identify the attributes of the output, in this case blackspots or non-blackspots from a proportion of the data. The model prepared is based on the algorithms of Logistic Regression. The machine learning model was imported using *sklearn*, implemented them for the dataset, and compared their results. The classification algorithm logistic regression is used to assign observations to a discrete set of classes. The logistic sigmoid function translates logistic regression output into a probability value. Logistic Regression is a Machine Learning algorithm that is used to solve classification problems. It is a predictive analytic approach that is based on the probability concept. The classifier gives a set of outputs or classes based on probability when the inputs are passed through a prediction function, and it returns a probability score between 0 and 1.



Supervised Machine Learning approach applied in the analysis and Model preparation of Blackspot Data

Data Splitting and Scaling:

To effectively train and assess the model's performance, we divided the dataset into training and testing subsets. The training subset facilitated the model to learn patterns and relationships inherent in the data, while the testing subset gauged the model's generalization ability. The data was divided into 80% training data and 20% testing data. Feature selection was then implemented on the taring data set to exercise the deployment of a better functioning model.

Feature Selection and Engineering:

Relevant variables significantly affecting blackspots were meticulously pinpointed. Utilizing forward feature selection, the model exhaustively calculated average scores for potential combinations. Optimal combinations, chosen based on highest accuracy, were selected. Forward feature selection yielded an average score of 92.30%, encapsulating intricate data interactions and enhancing model predictiveness.

Model Selection and Training:

In selecting an appropriate machine learning algorithm, the intricate balance between interpretability and predictive power was deliberated. The Logistic Regression algorithm emerged as an ideal choice, offering a clear understanding of feature contributions while yielding reliable predictive outcomes, therefore, classifying whether a road segment is potential Blackspot or not based on the independent variable. The selected model underwent iterative training, during which it adapted to the inherent complexities of the dataset, ultimately refining its predictive capabilities.

Model Evaluation and Interpretation:

Model efficacy was rigorously assessed using tailored binary classification metrics—accuracy, precision, recall, F1-score, and ROC analysis. Feature importance interpretation revealed key blackspot drivers, deriving actionable insights for road safety improvements.

Because the dataset is unbalanced, blackspots are about 11%, non-blackspot are about 89%. Therefore, we cannot mainly use accuracy to evaluate our model. We

want to cover both recall and precision, since we think that the prediction of correct target is important. Because of an unbalanced dataset, we choose f1 score instead of any other metrics. We evaluate our machine learning model by testing data. The f1 score is 95% for non-blackspots and 46% for blackspots. In ROC curve, the true positive rate is higher than false positive rate. Therefore, the trained model can predict whether a road segment has potential blackspot or not.

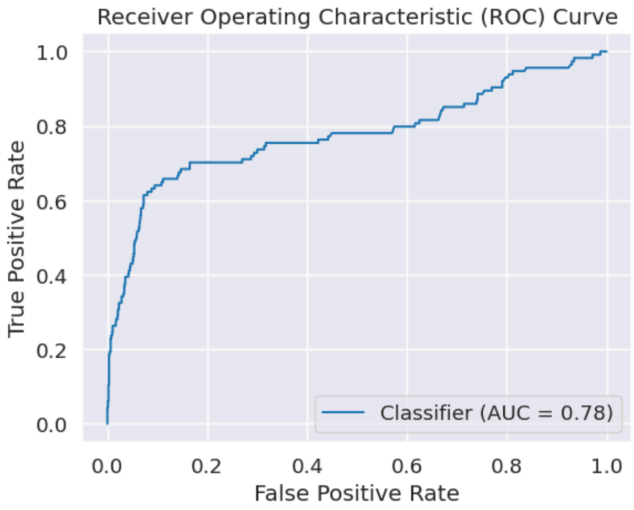


Figure: 7

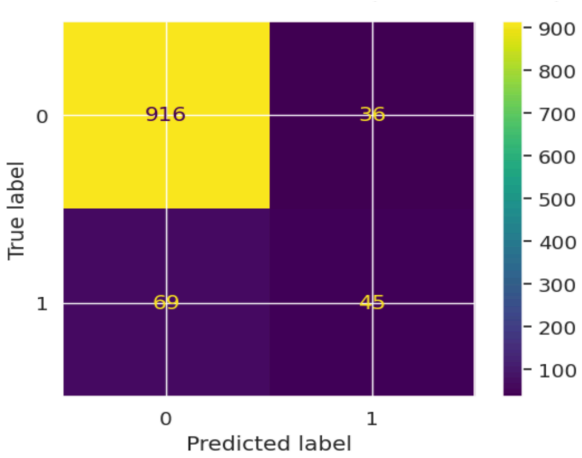


Figure: 8

	precision	recall	f1-score	support
0	0.93	0.96	0.95	952
1	0.56	0.39	0.46	114
accuracy			0.90	1066
macro avg	0.74	0.68	0.70	1066
weighted avg	0.89	0.90	0.89	1066

Figure: 9

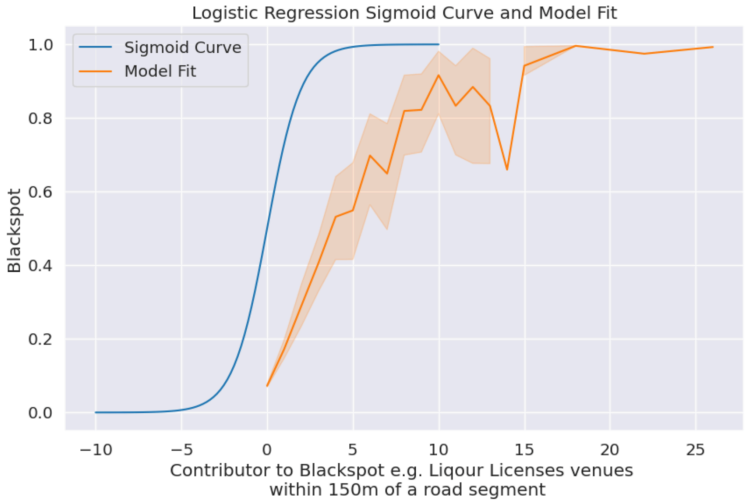


Figure: 10

Pros and Cons of the model

Pros:

- The Logistic Regression model will provide probability predictions and not only classification labels.
- It is very easy to interpret.
- It is not a resource hungry model, and this makes it suitable for simple applications.
- It unlikely to overfit on low dimensional data.
- The efficient algorithm ensures fast and resource-friendly scalability.

Cons:

- The Logistic Regression model is likely to overfit on a high dimensional data.
- It fails to capture complex relationships.
- Model's predictive value will degrade on including irrelevant features.
- Logistic Regression algorithm is sensitive to outliers.

Recommendation

The model's predictions can be used to prioritize safety interventions, policy development and allocate resources effectively in high-risk road segments that are identified as potential blackspots.

- As intersections are prone to road accidents, it can be recommended to convert high-traffic intersections to roundabouts. Trim vegetation and maintain road signage and markings for optimal visibility.
- It can be recommended to take strict traffic control measures such as regular alcohol testing around liquor licensed venues in busy road segments such as those lining in commercial lands.

References

Chen, H. Y., Ivers, R. Q., Martiniuk, A. L. C., Boufous, S., Senserrick, T., Woodward, M., Stevenson, M., & Norton, R. (2009). Socioeconomic status and risk of car crash injury, independent of place of residence and driving exposure: results from the DRIVE Study. *Journal of Epidemiology & Community Health*, *64*(11), 998–1003. https://doi.org/10.1136/jech.2009.091496

Curtin-Monash, & Hobday, M. (2017). *The effect of alcohol availability on road crashes at varying distances from the Central Business District in Perth, Australia from 2005 to 2015*.

Factors Contributing to Road Accidents. (2009). *The Handbook of Road Safety Measures*, 35–80. https://doi.org/10.1108/9781848552517-003

Graph Embeddings for Predicting Traffic Accident Black Spots. (n.d.). School of Science and Technology - Hong Kong Metropolitan University. Retrieved August 11, 2023

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior therapy*, *51*(5), 675–687. **https://doi.org/10.1016/j.beth.2020.05.002**

LaValley, M. P. (2008). Logistic Regression. *Circulation*, *117*(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Tay, R., & Rifaat, S. M. (2007). Factors contributing to the severity of intersection crashes. *Journal of Advanced Transportation*, *41*(3), 245–265. https://doi.org/10.1002/atr.5670410303