Analysing Road Safety: Unveiling Factors Contributing to Blackspots in Victoria

CONTENTS

EXECUTIVE SUMMARY

The project addresses the pressing issue of road accident blackspots through data-driven insights to develop effective interventions, campaigns, and reforms. Utilizing a comprehensive dataset, the initiative aims to enhance road safety by identifying risk factors and patterns associated with blackspots. The Business Analysis Core Concept Model guides the process, emphasizing the need for targeted measures, stakeholder involvement, and practical value.

The stakeholders, the Victorian Department of Transport (DOT), the general public, and local authorities, are central to this endeavor. By leveraging a machine learning model, the analysis provides evidence-based decision-making for interventions, legislative changes, and tailored education campaigns. The outcomes have significant value, leading to safer roads and reduced accidents.

The dataset, derived from crash and demographic data, underwent rigorous cleansing and preprocessing. Visualizations highlighted key insights, such as the influence of factors like primary schools, traffic signals, and household car ownership on blackspot occurrences. A logistic regression model was employed due to its suitability for binary classification.

The model's evaluation showcased its strengths, including high precision and recall for non-blackspots, as well as a reasonable accuracy and interpretability. However, challenges such as low recall for blackspots and class imbalance were identified.

To address these challenges, recommendations include leveraging real-time monitoring, refining the model using more data, and fostering collaboration among stakeholders. The solutions encompass targeted interventions, tailored campaigns, and evidence-based legislative reforms. Additional data sources like weather and traffic flow can further enhance the model's accuracy. In conclusion, the project provides actionable insights to tackle road safety concerns and drive positive change.

BUSINESS UNDERSTANDING

The urgent problem of blackspots, or accident hotspots, on roads is addressed in the business case. The objective is to give actionable insights that will enable the development of efficient interventions, education campaigns, and legislative reforms by utilizing a dataset that contains information about blackspots and characteristics. We determine the following by using the Business Analysis Core Concept Model (BACCM) (International Institute of Business Analysis, 2015).

- **Need**
  The objective is to offer practical insights to direct the creation of successful measures, including public awareness campaigns, statutory changes, and interventions, aimed at lowering accidents and enhancing road safety.

- **Value**
  The initiative provides significant benefit to all parties involved. Data-driven insights help decision-makers make well-informed choices for successful interventions, evidence-based legislation reforms, and targeted education initiatives. Safer roads, fewer accidents, and more overall road safety benefit the general public.

- **Stakeholders**

1. Victorian Department of Transport (DOT): The body seeking actionable insights to enhance road safety and mitigate the risks associated with blackspots.
2. General Public: The ultimate beneficiaries, who will experience safer roads and reduced accidents.
3. Concerned Local Authorities: Collaborators in implementing education campaigns, legislative reforms and interventions.

- **Solution**
  The suggested solution entails performing a thorough study of a given dataset. To evaluate the danger of blackspots, a machine learning model will be created, allowing for evidence-based decision-making and targeted actions.

- **Change**
  Automating the process of detecting high-risk locations will facilitate effective resource allocation. To address concerns about road safety, put up legislative reforms based on the identified risk factors. Create ads that are tailored to the individual risk factors that exist in various locations.

- **Context**
  This factor is influenced by the Department of Transportation's budget, available resources, and willingness to distribute resources in accordance with data-driven recommendations. The viability and method of passing legislative reforms are influenced by the legal and political environment. The effectiveness of educational campaigns is influenced by the degree of implementation.

DATA UNDERSTANDING

The dataset has been developed by integrating crash data from Department of Transport and demographics data from the Australian Bureau of Statistics (ABS). It comprises of 5326 records, across 36 class attributes. The target variable is Blackspot. The categorical variables are EZI_ROAD_NAME, ROAD_NAME, ROAD_TYPE, Intersection and Blackspot.

Data Cleansing Process

In order to ensure the quality and accuracy of the dataset for our analysis, a thorough data cleansing process was conducted.

- **Dropping irrelevant columns**
  The 'ID','EZI_ROAD_NAME' and 'ROAD_NAME' columns have been dropped. Dropping irrelevant columns from a dataset is a common to improve model performance and reduce noise.

- **Identifying and handling missing values**
  The columns AGE_65YRS_OVER_PCNT and Lq_Licenses have missing values. Both Lq_Licenses and AGE_65YRS_OVER_PCNT are right skewed. So, we replace the missing data with the median due to its resistance to outliers and its ability to provide a representative value.

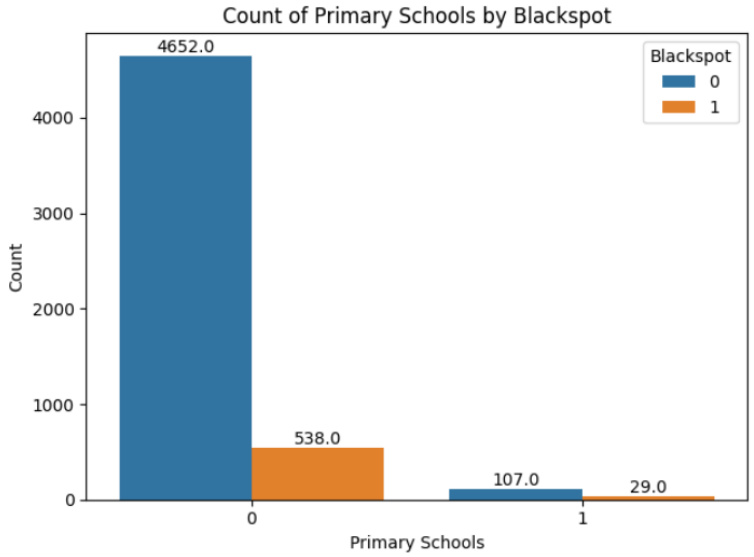
- **Converting categorical data to numerical data**
  Intersection, Blackspot and ROAD_TYPE variables data has been converted from categorical to numerical data for better visualisation and the regression model.

- **Set formatting for floating numbers**
  The number of decimal places in the float data type has been set to 3 so that the data remains constant.

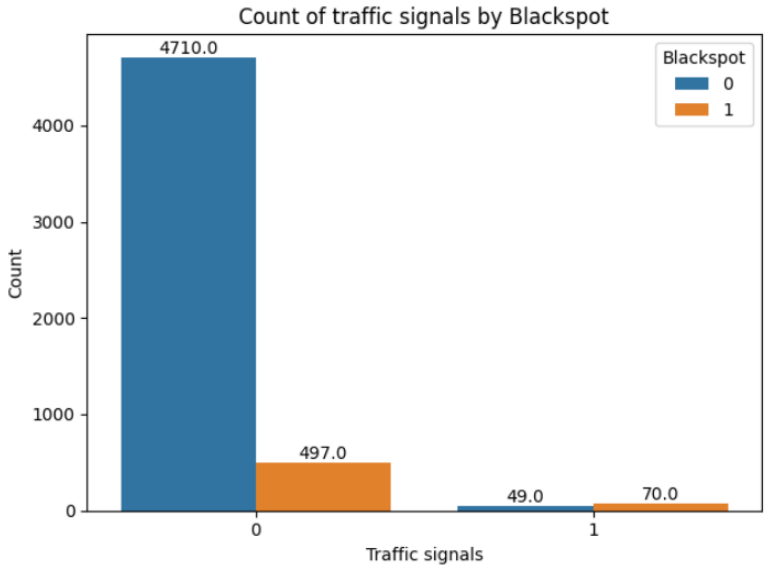
Data Visualisation

Bivariate Analysis

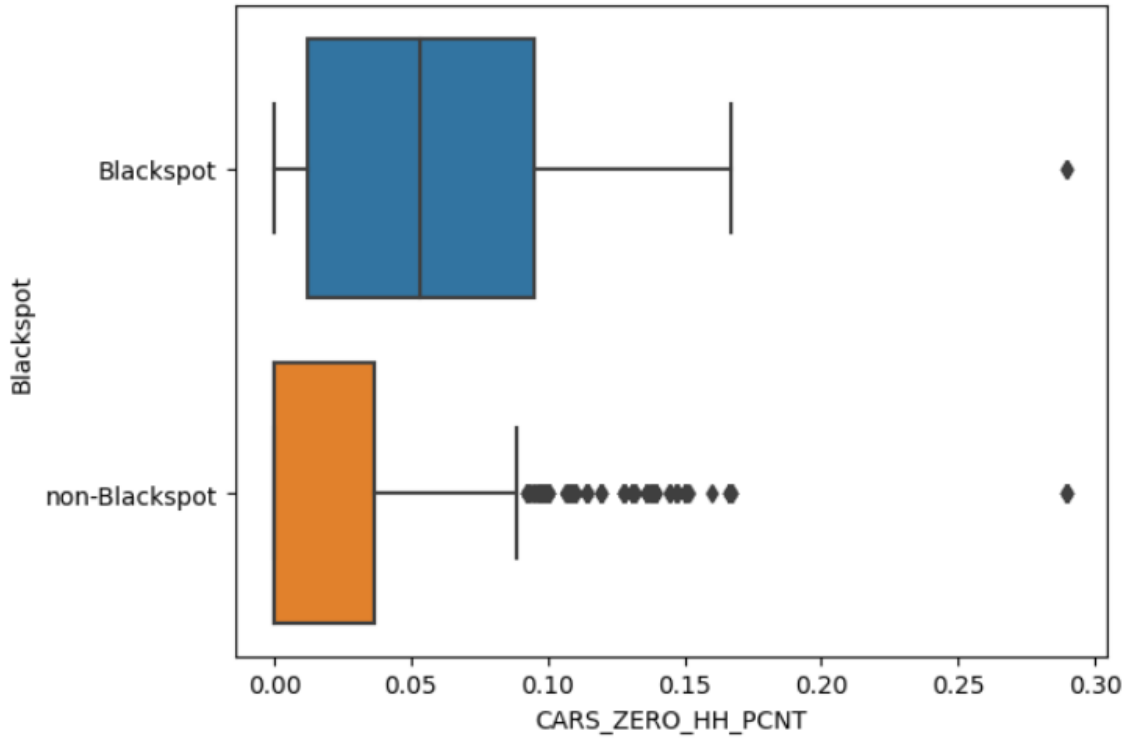
The graph below shows that the percent of blackspots significantly increases 10.36% to 21.32% when the road segment has a primary school nearby. So, it is a key feature for the model.



The graph below shows that the percentage of blackspots significantly increases from 9.54% to 58.8% when there is a traffic signal nearby. This is an unexpected observation.



The box plot below shows that the percentage of blackspots significantly increases when the % households with no cars increases. This is also unexpected observation.



Multivariate Analysis

The values in the heat map represent the correlation coefficients. Variables with strong positive and strong negative correlations have been used as feature for the model.

ML MODEL IMPLEMENTATION

For the prediction of blackspots and understanding the factors that contribute to them, using logistic regression is a suitable approach. Since we are dealing with identifying blackspots (accident hotspots), this is a categorical binary classification problem. Logistic regression is specifically designed for binary outcomes, making it a natural choice for predicting whether a location is a blackspot or not.

Feature Selection

The basis on which features that have been selected are:

- **Correlation Analysis**
  High Positive correlation
  The features ‘intersection’, 'Commercial% and Lq_Licenses' have the highest positive correlation with blackspots.

  High Negative Correlation
  The feature 'DWELLING_SEMID_RO_OR_TCE_H_TH_P' has high negative correlation with blackspots.

- **Exploratory Data Analysis**
  The features ‘traffic_signal', 'primary_school' and 'CARS_ZERO_HH_PCNT' showed pattern with respect to the presence of a blackspot.

- **Domain Knowledge**
  The feature ‘AGE_45_64YRS_PCNT' has been selected on the basis of the fact that according to CarExpert Australia the age group of 40-59 is most likely to be involved in road accidents.

Splitting Data

The data is split into a training set and a test set

- **Training Set (80%)**
  By allocating 80% of the data to the training set, the model has been provided with a substantial amount of data to learn from, which can lead to better parameter estimates and a more accurate model.

- **Test Set (20%)**
  Allocating 20% of the data to the test set ensures that you have a substantial amount of data to evaluate the model's performance.

- **Random State (2023)**
  The random state parameter is set to 2023, which means that the data splitting will be reproducible. This is important for consistency in your analysis.

ML MODEL EVALUATION

Performance Metrics

Performance metrics are crucial for evaluating the effectiveness and reliability of a logistic regression model. The performance metrics that have been used are

- **Confusion Matrix**
  The model correctly predicted 931 instances as not being accident blackspots (TN) and 43 instances as accident blackspots (TP). However, it incorrectly predicted 16 instances as accident blackspots that were not (FP), and it missed 76 instances that are accident blackspots (FN).

- **Classification report**
  The classification report provides a summary of various classification metrics for both classes (0 and 1). The overall accuracy of the model is 0.91, which is the proportion of correct predictions over all predictions.
  Class 0 (negative class) has high precision (0.92) and recall (0.98), indicating that it's well-predicted by the model.
  Class 1 (positive class) has lower precision (0.73) and recall (0.36), indicating that it's less well-predicted by the model.

- **Receiver Operating Characteristic**
  The more that the ROC curve hugs the top left corner of the plot, the better the model does at classifying the data into categories. The ROC curve for this model is away from the 45-degree diagonal. It is further validated by AUC

- **Area Under the Curve (AUC)**
  The AUC here is 0.87 which indicates that it has the high area under the curve and is quite accurate in classifying observations into categories.

Pros and Cons

Pros

- **High Precision for Class 0 (Not a Blackspot)**
  The model has a high precision of 0.92 for class 0, meaning that when it predicts a place is not a blackspot, it is correct about 92% of the time. This is valuable when false alarms are costly.

- **High Recall for Class 0 (Not a Blackspot)**
  The model's recall for class 0 is 0.98, indicating that it is effective at identifying most of the actual non-blackspot places. This is important for ensuring that potential non-blackspots are not missed.

- **Reasonable Accuracy**
  The overall accuracy of the model is 0.92, which indicates that it is correctly predicting the class for a large portion of the instances.

- **Interpretability**
  The coefficients provide insights into the features' impact on the outcome. This can help stakeholders understand the factors influencing blackspot determination.

- **High AUC**
  The model has a strong capacity to distinguish between positive and negative cases, according to an AUC of 0.87. It suggests that the predictions made by the model are typically distinct and highly ranked.

Cons

- **Low Precision and Recall for Class 1 (Blackspot)**
  The model's precision for class 1 is 0.73, and its recall is 0.36. This suggests that the model's ability to correctly predict blackspots is limited. The low recall means that the model is missing a substantial number of actual blackspots.

- **Class Imbalance Impact**

The class imbalance (119 instances of class 1 compared to 947 instances of class 0) could be influencing the model's performance. The model might be biased toward predicting the majority class (not a blackspot) more accurately.

- **F1-Score Imbalance**
  The F1-score for class 1 is lower than for class 0 (0.48 vs. 0.95). This imbalance reflects the trade-off between precision and recall for class 1.

SOLUTION AND RECOMMENDATION

SOLUTIONS

Data-Driven Interventions

Utilize the developed logistic regression model to identify high-risk blackspots accurately. Prioritize interventions and resource allocation based on the risk predictions provided by the model. Implement targeted road improvements, signage updates, and safety measures in identified high-risk areas.

Education Campaigns

Tailor campaign content to address specific risk factors identified by the model, such as primary schools or traffic signals.

Legislative Reforms

Present evidence-based findings from the model to the Victorian Department of Transport to support proposed legislative reforms. Advocate for changes in road safety regulations that align with the identified risk factors and patterns.

Real-Time Monitoring System

Implement a real-time monitoring system that integrates with the developed model. Continuously assess and update risk predictions based on new data to ensure accurate and up-to-date insights.

RECOMMENDATIONS

Additional Data Collection

Weather Data: Including weather conditions at the time of accidents can provide insights into accident patterns.
Traffic Flow Data: This can help understand traffic congestion and how it affects accident occurrences.
Road Design Data: Information about road geometry, signage, condition of roads, street lighting and lane markings can impact accident risk.

REFERENCES

https://www.carexpert.com.au/car-news/what-age-group-causes-the-most-car-accidents

https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/

https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

https://www.statology.org/interpret-roc-curve/