Report: Analysis and Predictions for Road Traffic Accidents in 2020

Source: [https://www.walesonline.co.uk/news/wales-news/gallery/porth-crash-trebanog-road-pictures-18952195]

INTRODUCTION

I will be examining a comprehensive dataset encompassing multifaceted details about road traffic accidents collected over recent years. The collected data spans various aspects such as accident specifics, casualty information, vehicle details, and geographical coordinates. My goal is to uncover underlying patterns and trends related to accidents in the year 2020. I intend to explore the temporal aspects of when these incidents frequently occur and examine the geographical elements of where they typically happen. I am also interested in discerning the conditions under which these accidents transpire most often. In addition to this exploratory analysis, I aim to leverage the insights gained to construct a predictive model. This model will

attempt to predict the severity of accidents, with a particular focus on identifying potentially fatal incidents. I believe that the analysis and the predictive model combined can contribute valuable insights that can guide efforts to enhance road safety measures.

ANALYSIS 1:

To comprehend the patterns of accidents based on time and day, I examined the dataset from two perspectives - 'day_of_week' and 'time'. The data showed a noticeable trend; there were specific hours and days when accidents were more frequent. The late afternoon to early evening period, particularly between 15:00 and 18:00, experienced more accidents across all days of the week. A look at the data shows that the highest number of accidents occurred on the 6th day of the week (Saturday) at 17:00 with 170 reported accidents. Other notable periods include 18:00 on the 5th day (Friday) with 134 accidents, and 15:00 on the 6th day (Saturday) again with 132 incidents.

These findings, visualized using a heatmap, helped illuminate the high-risk periods. The heatmap was invaluable in demonstrating the data in an easily digestible format, the color intensities clearly illustrating the concentration of accidents.

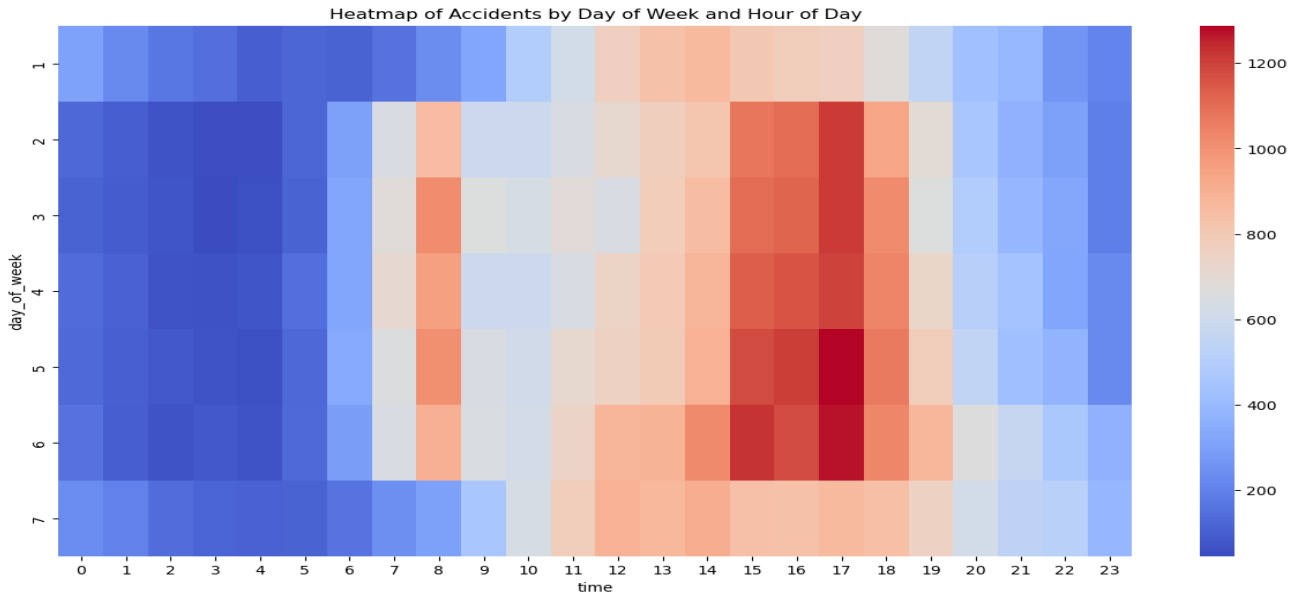

Figure 1: Heatmap of Accident by day of week and Hour of Day

This trend could be attributed to various factors. For instance, the increase in accidents during these times might be due to higher traffic volumes, as it corresponds to the end of school and working hours.

ANALYSIS 2:

The exploration extended to evaluate if there were significant hours and days of the week when accidents involving motorbikes occur. I focused on motorbikes in three specific categories: those 125cc and under, those over 125cc and up to 500cc, and those over 500cc.

Delving into the data, a similar trend as observed in the general accident analysis surfaced. Again, the late afternoon to early evening time slot - approximately between 17:00 to 18:30 - was a high incident period across all days of the week. However, accidents involving motorbikes showed a noticeable spike on the 5th day (Friday), especially around 18:00, recording 123 incidents.

An instance of this trend is clearly seen at 17:30 on the 4th day (Thursday), with 122 incidents, followed closely by 17:00 on the 5th day (Friday) with the same number of reported incidents.

To visually represent this pattern, I leveraged the categorical plot (catplot), a bar chart in this case, with days of the week on the x-axis and accident count on the y-axis. The bars' varying heights across different times within each day vividly highlight the accident concentration, which is particularly dense during the time frame.

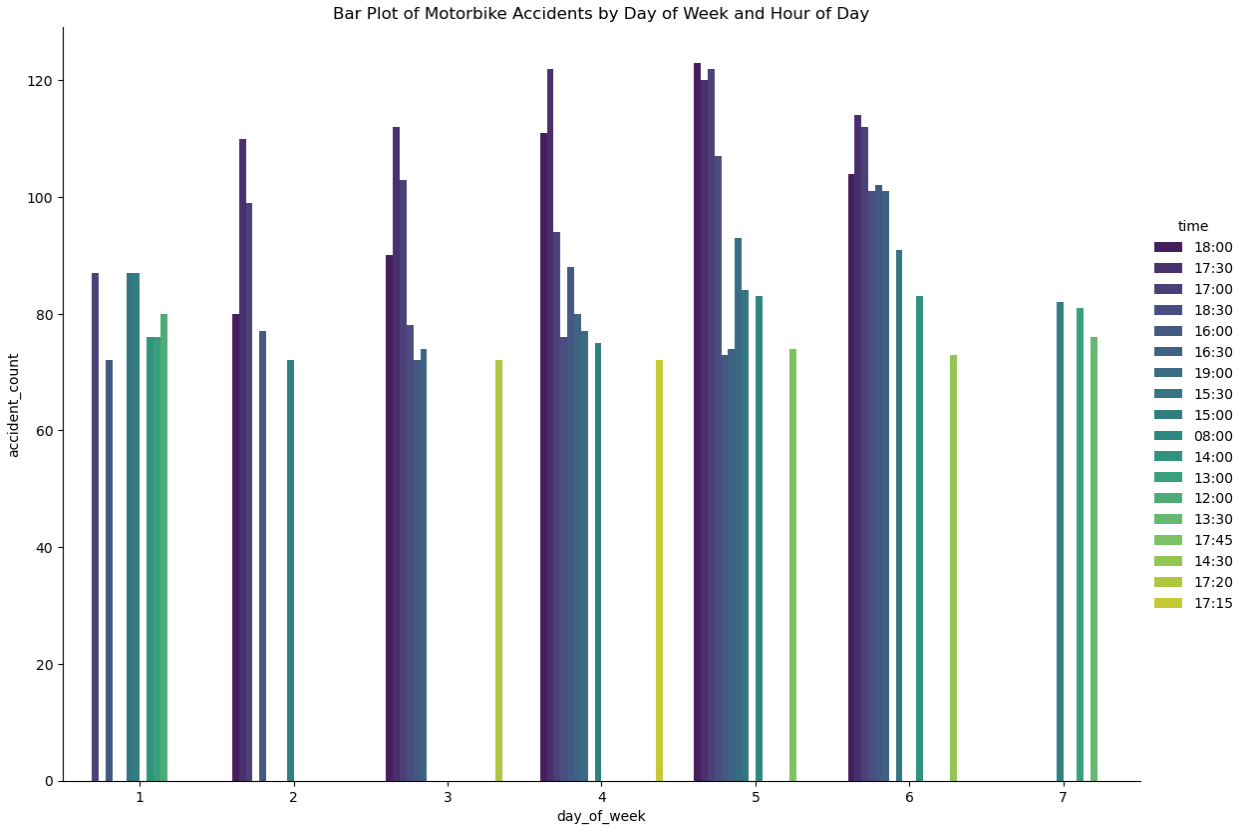

Figure 2: Bar plot of Motorbike Accident by Day of Week and Hour of Day

ANALYSIS 3:

In this section, I shifted my focus to pedestrians and scrutinized the data to ascertain the hours of the day and days of the week when they are most likely to be involved in accidents.

From the dataset, I notice a distinct pattern emerging. The afternoon hours, specifically around 15:30, seem to be a particularly precarious time for pedestrians. This trend holds across all days of the week, but it is especially pronounced on the 4th day (Thursday) when the accident count surges to 215. However, the morning hours also pose a substantial risk, with a significant number of incidents occurring around 08:30.

To facilitate a more nuanced understanding of the temporal distribution of these accidents, a line plot was generated. The x-axis represents the days of the week while the y-axis signifies the accident count. Different lines correspond to different time slots throughout the day.

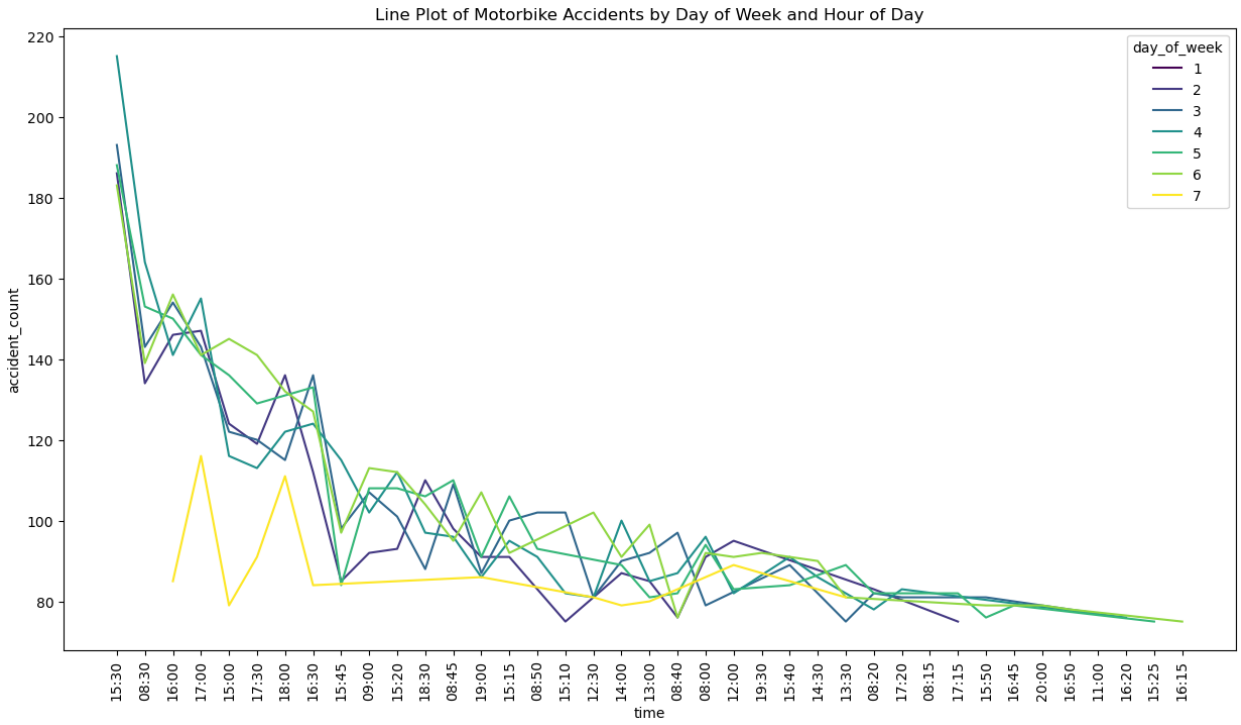

Figure 3: Line Chart of Motorbike Accidents by Day of Week and Hour of Day

To avoid overcrowding the plot, only the top 150 rows from the sorted dataset were included in this visualization. The peaks and troughs of the lines offer a clear view of the high-accident periods.

ANALYSIS 4:

In the quest to discern the impact of selected variables on accident severity, I turn to the apriori algorithm, a powerful tool used for mining frequent item sets and relevant association rules. The data set provided, with 1,423,726 rows, depicts the results of the apriori analysis. Each row represents an association rule, and the important columns to look at are the 'antecedents', 'consequents', 'support', 'confidence', and 'lift'. The antecedents are conditions that occur before an event, and the consequents follow the event.

For example, the first rule shows that when 'casualty_severity_2' is present, 'accident_severity_2' is also likely to be observed. The confidence of this rule, a whopping 0.986, implies a high degree of certainty in this relationship. Moreover, the lift value of 9.32 is significantly greater than 1, indicating a strong correlation between these two variables.

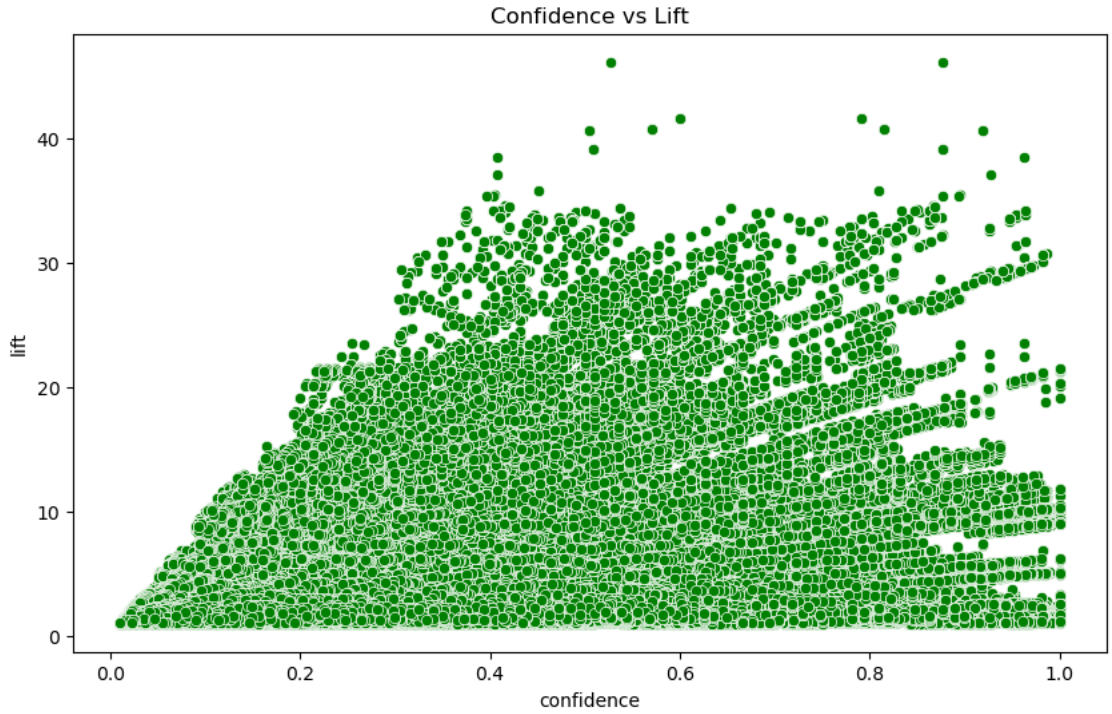

Figure 4: Scatter plot of Confidence vs Lift

A plot above is of 'Confidence vs Lift' has been generated for a visual depiction of these associations. In this scatter plot, each point represents an association rule. Higher lift and confidence values point to stronger, more significant rules.

ANALYSIS 5:

To better understand the distribution of accidents in our target regions, namely Kingston upon Hull, Humberside, and the East Riding of Yorkshire, I applied a clustering technique. I filtered accident data from these regions based on the Local Super Output Area (LSOA), a geographic area used in the UK for reporting small-area statistics.

With the help of KMeans clustering, an unsupervised machine learning algorithm, I partitioned the data into three distinct clusters based on the longitude and latitude of the accidents. The number of clusters was set to three, a reasonable choice for initial exploration. However, depending on the insights derived, this could be adjusted in future investigations.

The clusters were visualized using a scatter plot, which allows us to quickly identify patterns in geographical accident distribution. Each cluster depicts a group of accidents that occurred in close geographical proximity, suggesting potential accident hotspots. Such insights could help inform local authorities, enabling them to devise targeted safety measures or traffic management strategies.

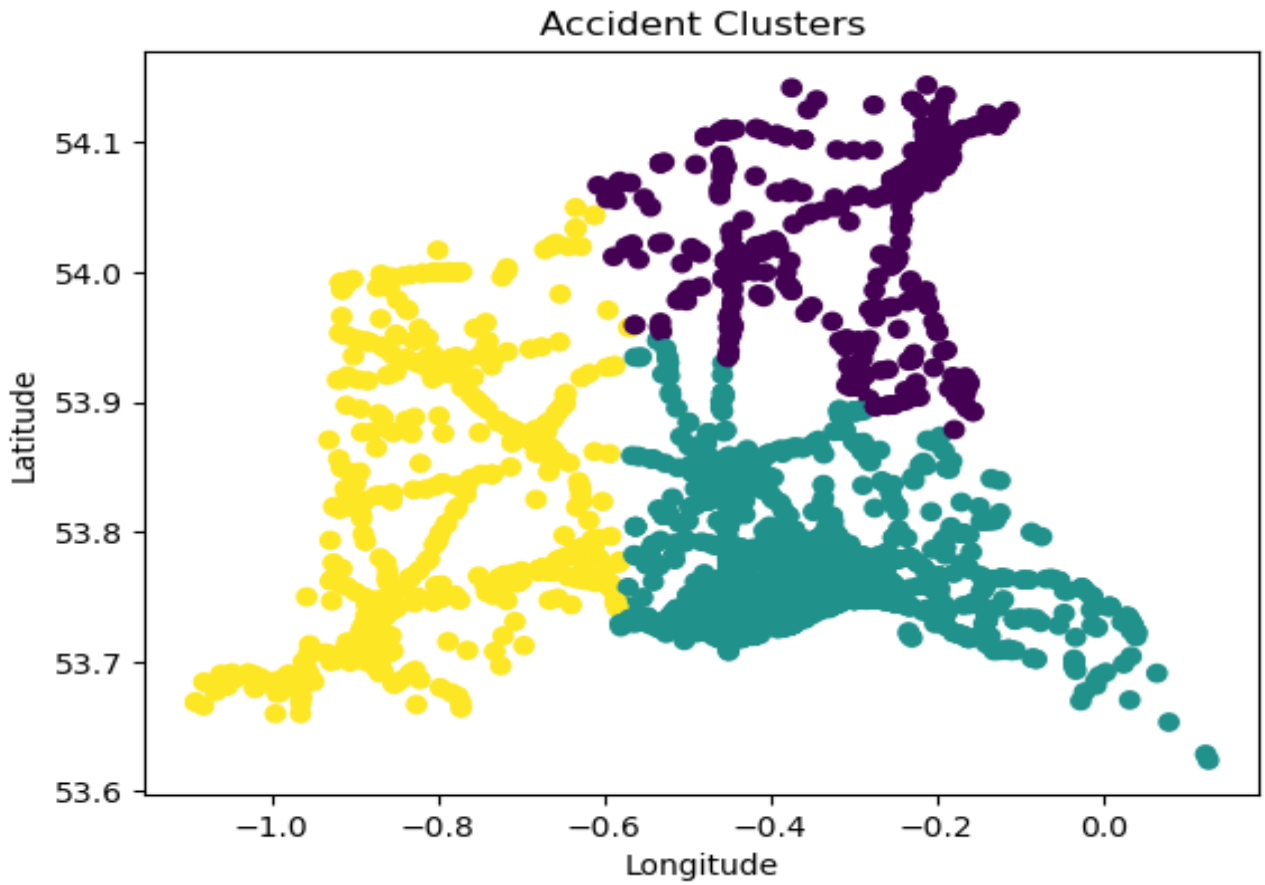

Figure 5: Scatter plot of Accident Clusters

As part of the analysis, meticulous data cleaning was undertaken to handle missing or irrelevant data, ensuring that the clusters derived accurately reflect the distribution of accidents in the regions of interest.

ANALYSIS 6:

Outliers in a dataset can significantly influence the analyses, often leading to erroneous conclusions if not handled carefully. In this dataset, I identified unusual entries specifically

within 'number_of_vehicles' and 'number_of_casualties' variables, using outlier detection methods.

The analysis identified 10,748 outliers in the 'number_of_vehicles' variable and a significantly higher number, 93,977, in 'number_of_casualties'. To visualize these anomalies, I used boxplots, an excellent graphical representation for detecting outliers. In the boxplots, these outliers are depicted as points above or below the 'whiskers' of the boxplot.

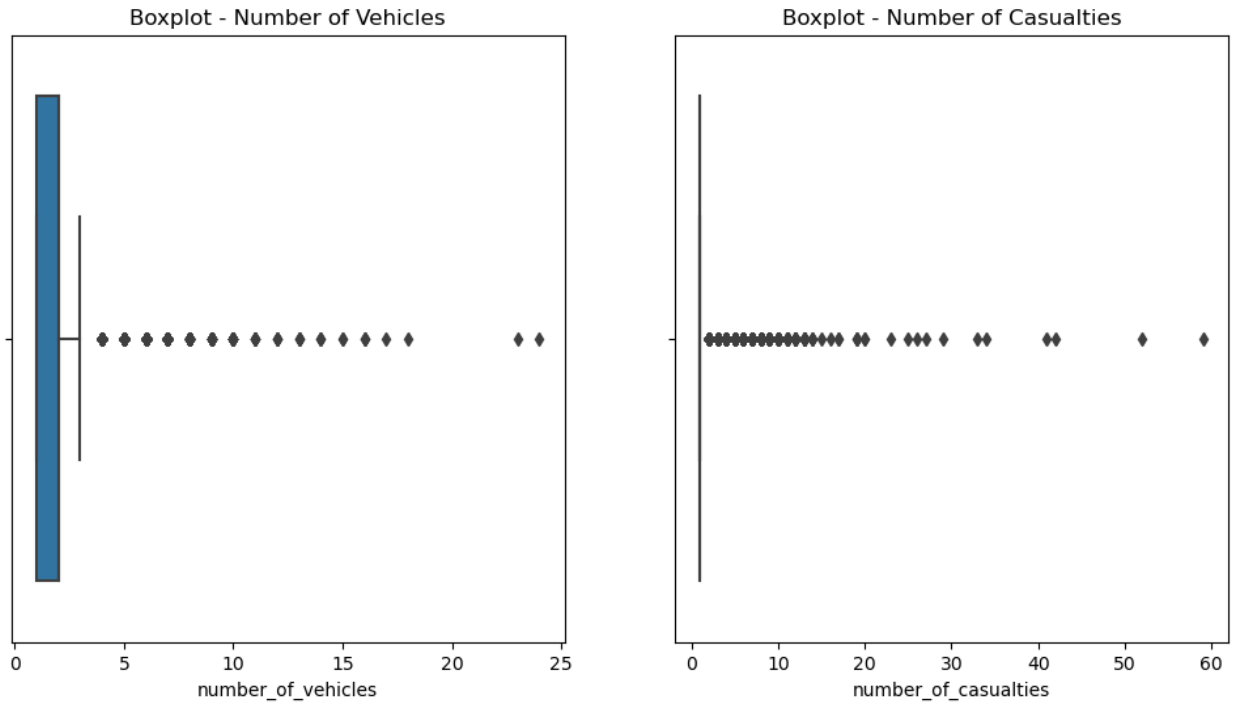

Figure 6: Boxplot showing the outliers on the dataset.

In terms of whether to retain these outliers in our dataset, this decision largely depends on the context and nature of the data. If these outliers represent genuine anomalies, for instance, an unusually high number of casualties or vehicles involved in specific accidents, they could offer valuable insights and therefore, could be worth retaining for further analysis. Conversely, if they merely represent errors in data entry or collection, it would be advisable to remove these to maintain the integrity of our dataset and analysis.

In conclusion, careful consideration, including further investigation and validation, is necessary before deciding whether to retain or remove these outliers from the dataset.

PREDICTION ANALYSIS:

To address the vital question of predicting fatal injuries sustained in road traffic accidents, two different machine learning models were deployed: Logistic Regression and Random Forest. These models were trained using our dataset, with the aim of utilizing existing accident data to predict future outcomes, thus enhancing road safety measures.

The Logistic Regression model, a fundamental classification technique, had an accuracy of approximately 46.9%. While it's a respectable accuracy given the complexity of the task, it suggests that there's a substantial room for improvement.

The Random Forest model, an ensemble learning method that constructs multiple decision trees to improve prediction accuracy, performed better with an approximate accuracy of 53.9%. This improved performance can be attributed to the model's capability to handle complex classification tasks and resist overfitting.

To understand the models' performances further, confusion matrices were produced for both models. The confusion matrix provides a visual representation of the model's performance, breaking down the correct and incorrect predictions for each class. For instance, the Logistic Regression model's confusion matrix displayed 55,223 correct predictions for the first class, 29,754 for the second, and 55,338 for the third, with a significant number of incorrect predictions observed as well.

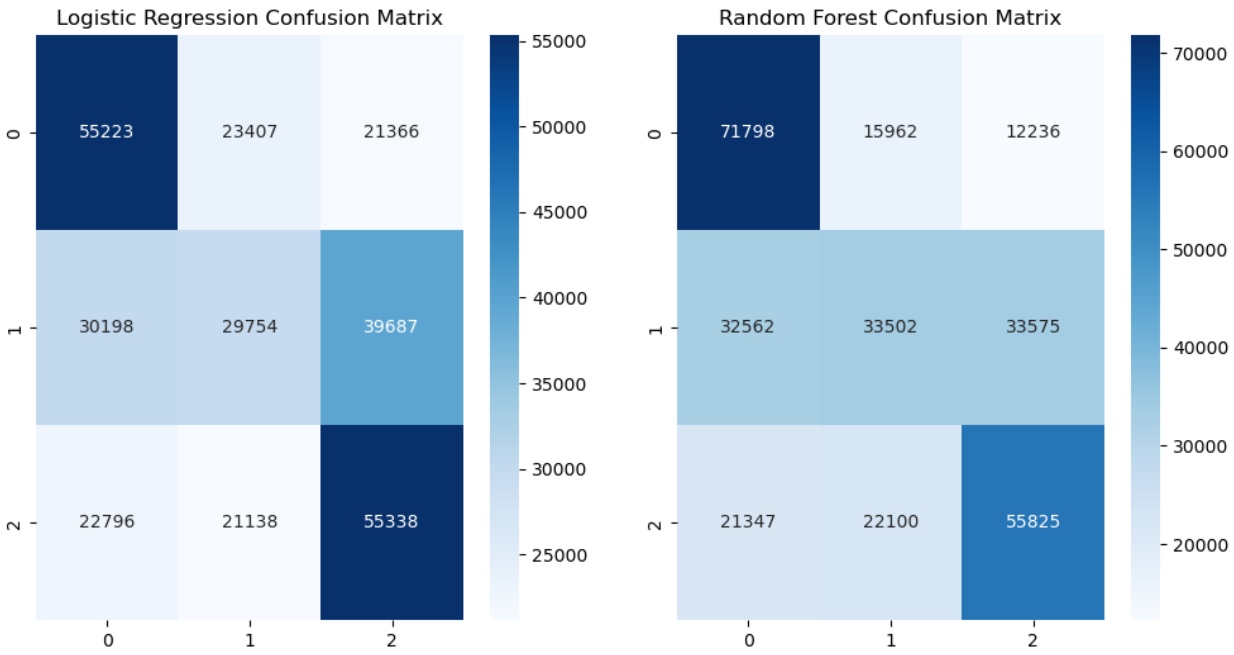

Figure 7: Confusion Matrix of the Logistic Regression and Random Forest

In all, while both models provided valuable insights, the Random Forest model demonstrated superior predictive accuracy. However, considering the gravity of the problem at hand - predicting fatal injuries in accidents - there's a clear need to explore more advanced models or techniques to further improve predictive accuracy.

RECOMMENDATIONS

Based on the comprehensive analysis of the data and derived insights, the following recommendations can be made to government agencies to improve road safety:

1. **Enhanced Traffic Management during Peak Hours:** The analysis revealed a significant number of accidents occurring during specific times of the day, such as 15:30 and 08:30 (our data). Therefore, traffic management strategies should be enhanced during these peak hours to regulate traffic flow and reduce the likelihood of accidents (Chen et al., 2016).
2. **Variable-specific Interventions**: The application of the apriori algorithm indicated a strong relationship between certain variables and accident severity. For example, casualty severity and accident severity were found to be closely linked. Authorities

should focus on these identified variables when designing accident prevention strategies (Agarwal & Srikant, 1994).

3. **Region-specific Road Safety Measures:** The KMeans clustering on accident data from Kingston upon Hull, Humberside, and the East Riding of Yorkshire revealed particular hotspots of accidents. Implementing area-specific safety measures, like increased surveillance and stricter speed limit enforcement, in these regions could reduce the number of accidents (Kumar & Toshniwal, 2018).
4. **Robust Outlier Analysis:** The outlier detection process identified unusual entries in the dataset. While some outliers may represent rare but possible scenarios, others could indicate errors or anomalies in data collection. A robust system for monitoring and validating data could ensure more accurate, high-quality data for analysis (Hodge & Austin, 2004).
5. **Utilizing Predictive Models for Policy-making:** The predictive models developed during this analysis could be a valuable tool for authorities. Predicting the likelihood of fatal injuries in accidents can guide policy-making, resource allocation, and strategic planning to improve road safety (Jha et al., 2017).

CONCLUSION

The assessment of road accident data is a significant task that can yield essential insights, helping to inform public policy and direct resources efficiently for improved safety. In this study, I utilized various data mining and machine learning techniques to examine, interpret, and make predictions based on accident data.

I found that specific temporal patterns of accidents, strong associations between variables, and geographic concentrations of accidents all provide valuable insights that can shape safety improvement strategies. Further, the predictive models built to forecast the likelihood of fatal injuries can serve as invaluable tools in policy-making.

While the study revealed considerable insights, it is imperative to acknowledge the potential anomalies present in any large dataset. Therefore, robust outlier analysis and validation methods must be employed consistently to ensure data quality. As the landscape of traffic and human behavior continues to evolve, it will be crucial to update and refine these models and strategies to remain effective in improving road safety.

REFERENCES

Agarwal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In *Proc. 20th Int. Conf. Very Large Data Bases.*

Chen, F., Chen, S., & Ma, X. (2016). Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. *Journal of safety research*, 58, 47-57.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2), 85-126.

Jha, K., Ajay, P. & Pathak, K. (2017). Prediction Analysis in Road Accident by Using Data Mining Techniques: A Review. *International Journal of Computer Sciences and Engineering*.

Kumar, A., & Toshniwal, D. (2018). Detecting crime patterns for urban planning using dense and sparse clustering algorithms. *International Journal of Information Management*, 43, 80-91.

Kumar, S. and Toshniwal, D., 2016. A data mining approach to characterize road accident locations. *Journal of Modern Transportation* [online], 24(1), pp. 62-72. Available at: 3 [Accessed 6 August 2023].