THE
GEORGE WASHINGTON
UNIVERSITY

CAPSTONE PROJECT

BEITBRIDGE BORDER:
TRAFFIC VOLUME ESTIMATING MODEL

DECEMBER 18, 2023

Prepared by **SHUMEL SIRAJ**

1. Introduction

The Beitbridge border post, a pivotal gateway between South Africa and Zimbabwe, stands as a cornerstone in facilitating trade and commerce in the Southern African region. This report delves into the development and application of a "**Beitbridge Border: Traffic Volume Estimation Model**" a project executed in collaboration with the World Bank's ambitious initiative, "**Regional Integration Using Remote Sensing and Artificial Intelligence to Measure Trade Node Activity and Beyond (P178398).**" The primary objective of this project is to foster international trade and enhance human prosperity by augmenting trade transparency and elucidating patterns and specific features of trade node activities.

Our focus at the Beitbridge border is underpinned by its strategic significance in regional trade. As one of the busiest border crossings on the continent, it serves as a vital conduit for a substantial volume of commercial traffic, epitomizing the trade dynamics between South Africa and Zimbabwe. The study utilizes toll station data as its main source, providing a comprehensive count of traffic volume, predominantly consisting of transportation vehicles pivotal to trade activities. This data is instrumental in analyzing, understanding, and eventually predicting traffic flow patterns, which are essential for efficient border management and planning.

This report outlines the methodology employed in developing the Traffic Volume Estimation Model, discusses the insights derived from the toll station data, and highlights the potential implications of these findings on regional trade. Furthermore, it demonstrates how this model aligns with the broader goals of the World Bank project, contributing to a more transparent, efficient, and prosperous trade environment across the Southern African region.

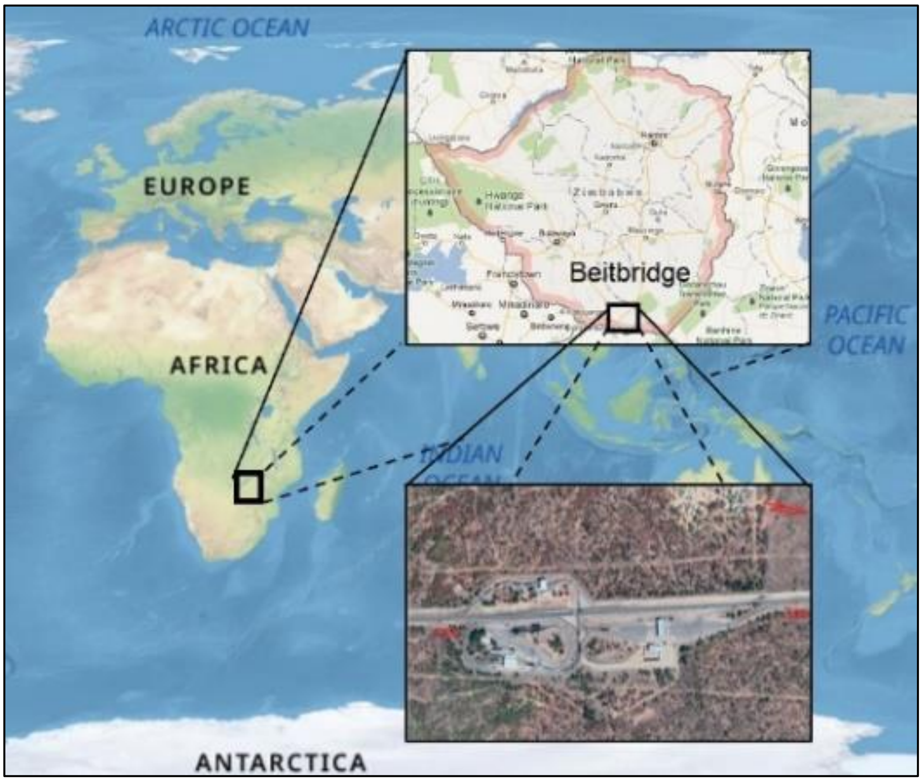


Fig 1.1 Beitbridge station 3201 And Beitbridge station 3202

In essence, the Beitbridge Border Traffic Volume Estimation Model stands as a testament to the innovative use of technology in fostering economic growth and regional integration, setting a precedent for similar initiatives in other trade-critical areas across the globe.

2. Use Cases of the Traffic Prediction Model

The development and implementation of the traffic prediction model have a broad range of practical applications that can significantly impact various aspects of transportation management and urban development. Here are the key use cases for the model:

1. **Congestion Management:** The model is instrumental in identifying potential bottlenecks and high-traffic zones, enabling traffic controllers to implement preemptive measures to mitigate congestion. By predicting traffic flow patterns, the model allows for the

optimization of signal timings and the strategic planning of lane closures, thereby reducing delays and improving travel times.

2. **Urban Planning:** Urban planners can utilize the data provided by the model to design more efficient road networks. Insights into traffic volume trends can inform the development of roads, the placement of intersections, and the allocation of zones for public transportation facilities. This data-driven approach to urban infrastructure development supports sustainable growth and better living conditions in urban settings.
3. **Policy Decisions:** The model's predictive capabilities support policymakers in making informed decisions regarding transportation-related policies. By understanding traffic patterns, authorities can set appropriate fare rates, toll collections, and regulations that balance revenue generation with traffic flow efficiency.
4. **Resource Allocation:** The predictive model aids in the effective distribution of resources such as emergency services and road maintenance crews. With advanced knowledge of traffic conditions, these services can be dispatched proactively, ensuring timely response to incidents and efficient maintenance operations to keep the traffic system functioning smoothly.
5. **Real-time Updates:** In the realm of navigation and GPS systems, the model's real-time traffic predictions are crucial. They provide drivers with up-to-date route recommendations, helping to avoid traffic jams and reach destinations via the most efficient paths. This not only saves time for commuters but also contributes to reducing fuel consumption and lowering emissions by preventing vehicles from getting stuck in traffic.

By harnessing the power of predictive analytics, this model offers a comprehensive tool for enhancing the mobility and efficiency of transportation networks, shaping smarter cities, and contributing to the overall well-being of the commuting public.

3. Data Description

In our quest to accurately estimate traffic volumes at the Beitbridge border, our analytical model integrates comprehensive traffic count data sourced from two strategically located toll road stations near Beitbridge. This data is crucial in providing a near-ground truth perspective of the actual traffic flow dynamics in the area and forms the foundation of our analysis.

Data Coverage: The dataset spans from the beginning of 2018 to December 31, 2022. This extensive time coverage allows for an in-depth analysis of traffic patterns, capturing seasonal variations and identifying long-term trends.

Interval-Based Counts: Traffic counts are meticulously recorded at three-hour intervals. This granularity offers an insightful view into the diurnal and nocturnal traffic flow, presenting a more nuanced understanding than standard daily or weekly summaries.

Station Specifics:

- **Beitbridge TCC Station 3201:** This station records traffic in two directions. Direction 1 captures toward Musina traffic, while Direction 2 monitors toward Beitbridge traffic. This station's data is pivotal in understanding the volume and trends of vehicles entering and leaving Zimbabwe.
- **Beitbridge TCC Station 3202:** Similarly, this station tracks traffic flow in two directions. Direction 1 is responsible for toward Beitbridge traffic, and Direction 2 for toward Musina. The insights gleaned from this station are critical for assessing traffic entering South Africa from Zimbabwe.

Vehicle Classification: The dataset segments traffic counts into various vehicle types, including Light vehicles, and Short Heavy, Medium Heavy, and Long Heavy vehicles. This classification aids in detailed traffic composition analysis and specific vehicle-type predictions.

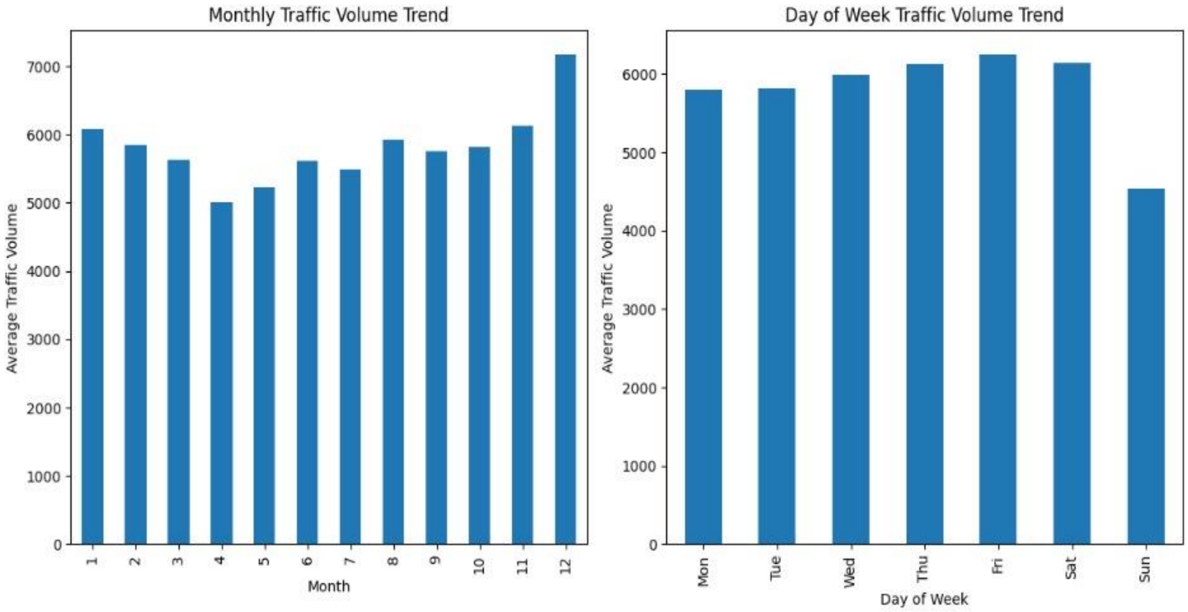


Fig 2.2: Monthly Average Traffic Volume Trend (Left) and Daily Average Traffic Volume Trend (Right)

While traffic flow is influenced by numerous factors, not all of which can be directly incorporated into our data (such as unpredictable events like wars or protests), we focus on consistent and quantifiable economic indices. Our analysis primarily operates on a daily scale, aligning with our traffic data frequency. However, most publicly available economic indices are on a monthly or quarterly scale. To bridge this gap, we have identified key companies in South Africa and Zimbabwe that can serve as economic indicators, along with daily scale data for fuel prices.

Stock Prices (Daily Scale)

- **Anglo American Platinum Ltd** and **Harmony Gold Mining Company Ltd:** Representing the mining sector and export earnings.
- **Barloworld Limited:** A reflection of the logistics, fleet management, and automotive retail sectors.

- **Shoprite Holdings Ltd:** Indicative of the retail and consumer goods industry.
- **Woolworths Holdings Ltd:** Impacting various segments including clothing, food, and home decor.
- **Standard Bank Group Ltd** and **FirstRand Ltd:** Representing the financial sector and banking services.

Crude Oil Price (Daily Scale)

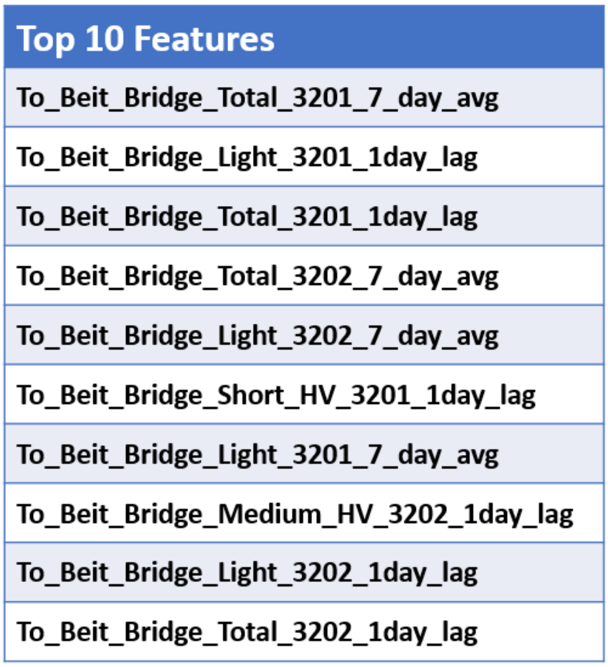
- **The OPEC Basket Price:** This is a weighted average of oil prices from various OPEC member countries. Given that Zimbabwe and South Africa import fuel from OPEC countries, this price is a significant indicator influencing traffic flow.

This comprehensive dataset, combining detailed traffic counts with economic indices and fuel prices, provides a robust framework for analyzing and predicting traffic volume trends at the Beitbridge border. This approach not only enhances our understanding of current traffic conditions but also aids in forecasting future patterns, crucial for effective border management and planning.

4. Methodology

Our study aims to predict the volume of traffic moving toward Musina. During the model development, we observed a counterintuitive yet statistically significant finding: the most predictive features for traffic volumes in the direction of Musina were measurements of traffic flow in the opposite direction. This revelation led us to a feature selection strategy that departs from conventional expectations.

Given the observed importance of opposite direction traffic volumes, we placed a greater emphasis on lagged variables and historical averages in our analysis. The rationale is that the traffic dynamics in one direction may have a delayed or historical impact on the traffic in the opposite direction, possibly due to return trips, correlated travel behaviors, or shared causative factors such as local events or seasonal migration patterns.



Top 10 Features
To_Beit_Bridge_Total_3201_7_day_avg
To_Beit_Bridge_Light_3201_1day_lag
To_Beit_Bridge_Total_3201_1day_lag
To_Beit_Bridge_Total_3202_7_day_avg
To_Beit_Bridge_Light_3202_7_day_avg
To_Beit_Bridge_Short_HV_3201_1day_lag
To_Beit_Bridge_Light_3201_7_day_avg
To_Beit_Bridge_Medium_HV_3202_1day_lag
To_Beit_Bridge_Light_3202_1day_lag
To_Beit_Bridge_Total_3202_1day_lag

Fig 4.2 Top 10 Features

The flowchart given below outlines the methodology for estimating traffic volume at the Beitbridge border using various data sources and analytical models.

1. Data Collection and Preprocessing:

Toll Station Data: Toll Station 3201 and 3202: Data from two toll stations are collected. These stations are responsible for recording the traffic volume in both directions, entering and exiting Zimbabwe and South Africa.

Economic Data:

- **Stock Price:** Daily stock prices of selected companies that represent economic activity in the two countries are collected.

- **Crude Oil Price:** Daily OPEC crude oil basket prices are also gathered, as they are a proxy for fuel costs which can influence traffic volume.

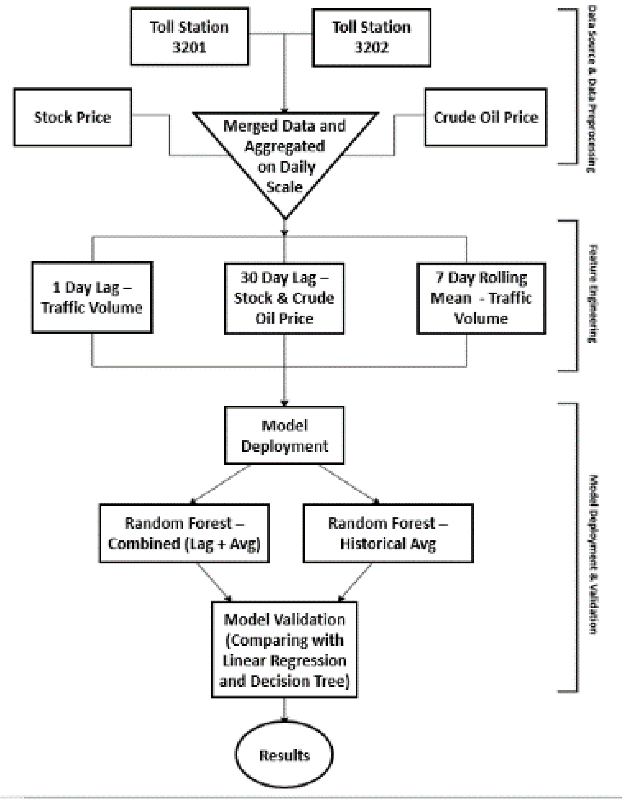

Toll Station 3201
Toll Station 3202
Stock Price
Merged Data and Aggregated on Daily Scale
Crude Oil Price
Data Source & Data Preprocessing
1 Day Lag – Traffic Volume
30 Day Lag – Stock & Crude Oil Price
7 Day Rolling Mean - Traffic Volume
Feature Engineering
Model Deployment
Random Forest – Combined (Lag + Avg)
Random Forest – Historical Avg
Model Validation (Comparing with Linear Regression and Decision Tree)
Model Deployment & Validation
Results

Fig 6.1 A Methodological Flowchart

Data Aggregation: All the collected data is merged and aggregated on a daily scale to align with the traffic volume data. This step ensures that all variables are on the same temporal resolution for accurate analysis.

2. Feature Engineering:

- **Lagged Traffic Volume:** In this study, three different lag durations – 1 day, 2 days, and 3 days – were tested to determine their predictive power for traffic volume forecasting. The 1-day lag, where the traffic volume of the previous day is used as a feature, showed the best and most consistent results. This supports the hypothesis that the previous day's traffic volumes are predictive of the following day's volumes. However, when extending the lag to 2 and 3 days, the predictive accuracy did not improve further. In fact, it was observed that beyond a 3-day lag, the model's performance began to decrease, indicating that older traffic data becomes less relevant for predicting immediate future volumes.

- **Lagged Economic Indicators:** A 30-day lag is applied to stock prices and crude oil prices. This longer lag period can capture more extended trends and the delayed effect of economic changes on traffic volume.
- **Rolling Mean Traffic Volume:** To mitigate the impact of short-term variances and enhance the detection of long-term trends in traffic data, a 7-day rolling mean algorithm was implemented. This approach involves calculating the mean traffic volume over a moving window of seven consecutive days. Each day, the window advances by one day, incorporating the current day's traffic volume while excluding the oldest day's volume from the calculation. This method effectively smooths out daily fluctuations due to factors like weekend effects, holidays, or other transient anomalies. By averaging over a week-long period, the rolling mean provides a more stable and reliable representation of traffic trends, facilitating a clearer understanding of longer-term traffic patterns and anomalies in the data set.

3. **Model Deployment:**

Two separate Random Forest models are deployed using the engineered features:

- **Random Forest - Combined (Lag + Average):** This model uses both the lagged traffic volumes and the rolling mean, along with the lagged economic indicators, to predict traffic volume.
- **Random Forest - Historical Average:** This model likely uses historical averages of traffic volume for prediction, serving as a baseline or comparison model.

4. **Model Validation:**

The models are then validated and compared with other machine learning algorithms, namely Linear Regression and Decision Tree models. This step is crucial to ensure the robustness and accuracy of the Random Forest models.

5. Results:

The outcome of the modeling process is analyzed, and the results are compiled, which likely include the model's predictive performance, feature importance, and possibly, the implications of the findings on traffic management and policy decisions.

This methodology combines time-series analysis with machine learning to provide a comprehensive approach to traffic volume estimation. It leverages both real-time traffic data and economic indicators to create a model with predictive capabilities that can be used for effective border management and planning.

5. Algorithm and Statistical Methods Used in Traffic Volume Estimation

The core of our traffic volume estimation model is built on the Random Forest algorithm, an ensemble learning method renowned for its high accuracy and robustness. By harnessing the collective power of multiple decision trees, the Random Forest algorithm delivers a predictive model that surpasses the performance of individual trees.

Core Concept

The Random Forest: The Random Forest algorithm amalgamates numerous decision trees to create a single, composite model. This integration leads to more precise and stable predictions than what could be achieved by any single decision tree.

Mechanism of Operation

Bootstrap Aggregating (Bagging): Each tree within the Random Forest is constructed from a unique bootstrap sample drawn from the original dataset. This technique allows for some observations to be represented multiple times in a single tree, while others may not be represented at all, thereby ensuring diversity within the model's structure.

Feature Randomness: In the tree-building process, Random Forest introduces variability by selecting random subsets of features for splitting nodes. This randomness is pivotal in enhancing the robustness of the model by preventing overreliance on any single feature and promoting generalization.

Consensus Approach: For regression problems, the Random Forest algorithm predicts an outcome based on the average of all individual tree predictions. This collective decision-making process helps in reducing variance and mitigating the risk of model overfitting.

Advantages

Resistance to Overfitting: The algorithm's ability to average out biases among diverse trees significantly lowers the risk of overfitting, making it highly reliable for predictive analytics.

Feature Handling: Random Forest excels in managing datasets with a large number of input features, adeptly identifying and ranking the importance of each feature in the prediction process.

Application in the Project

Model Configuration for Traffic Prediction:

Lagged and Averaged Features: By incorporating lagged traffic volume data along with a 7-day rolling mean, the model takes into account both short-term fluctuations and longer-term trends.

Economic Influences: The inclusion of a 30-day lagged view of stock and crude oil prices allows the model to reflect the impact of economic conditions on traffic volumes.

Performance Evaluation

R-squared (R^2) Metric: We gauge the success of our Random Forest model using the R^2 statistic, a measure of the variance in traffic volumes that our model can predict based on the independent variables. A high R^2 value would indicate that the model has a strong predictive capability, accounting for a significant proportion of the variance in traffic volume.

In addition to the primary Random Forest model, we employ Linear Regression and Decision Tree models for validation to ensure the robustness and generalizability of our predictions.

Linear Regression

Fundamentals: Linear Regression is a statistical method used to model the relationship between a scalar response and one or more explanatory variables. It is one of the simplest and most traditional forms of predictive analysis.

Application: In this context, Linear Regression serves as a baseline model to validate the performance of the Random Forest. Since Linear Regression assumes a linear relationship between variables, it provides a contrast to the more complex relationships captured by Random Forest.

Evaluation: The performance of Linear Regression is assessed by its R^2 metric, which indicates how much of the variance in traffic volume can be explained by the model. Comparing this metric across models helps to understand the benefits of using more sophisticated methods over simpler ones.

Decision Tree

Structure: A Decision Tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

Function: The model makes decisions by splitting the data into subsets based on the value of input features. This process is repeated recursively, resulting in a tree with decision nodes and leaf nodes that represent the final predictions.

Comparison: While a single Decision Tree is often prone to overfitting, it can provide valuable insights into the decision-making process and the data's underlying structure. By comparing it to the Random Forest model, which is essentially a collection of Decision Trees, we can assess the improvement in prediction accuracy gained by the ensemble approach.

Model Comparison and Validation Strategy

Cross-Model Evaluation: By comparing the Random Forest with Linear Regression and Decision Tree models, we can validate the traffic volume estimation model's predictive accuracy and stability.

Metrics for Comparison: While R^2 is a primary metric for comparison, other metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) may also be used to evaluate and compare the models' performance in terms of prediction accuracy and error rate.

Interpretation of Results: A superior performance by the Random Forest model, indicated by a higher R^2 value and lower error metrics compared to Linear Regression and Decision Tree models, would validate its use as the primary model for estimating traffic volumes.

By employing a robust validation framework that leverages multiple modeling approaches, we can confidently ascertain the predictive power of the Random Forest model and its efficacy in estimating traffic volumes for effective border management and strategic planning.

6. Results

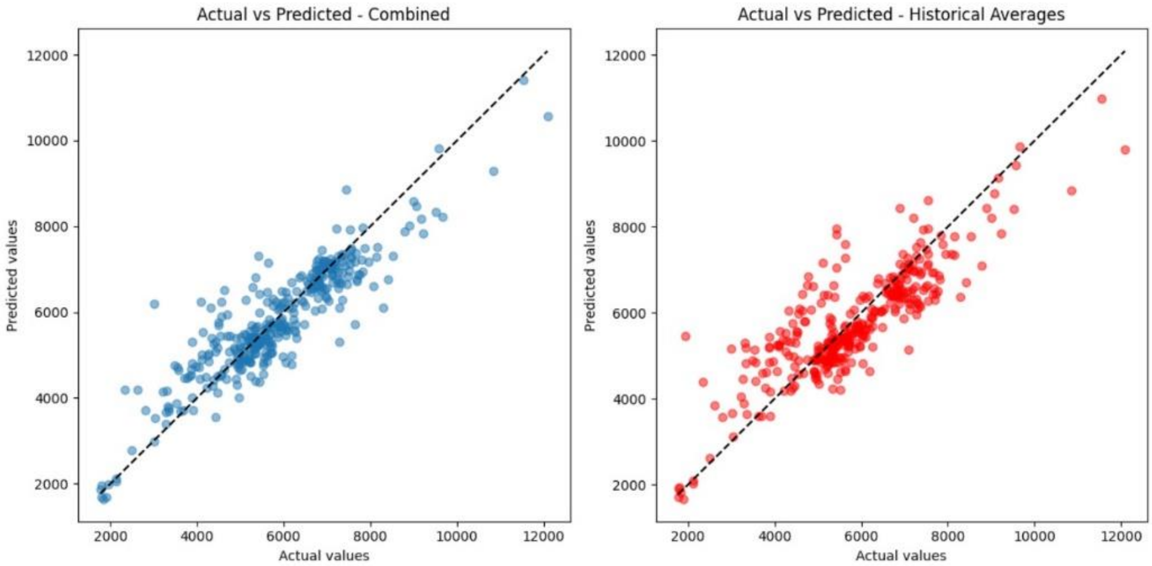


Fig 6.1 Actual VS Predicted For Both Combined (Left) and Historical Averages (Right) – Random Forest Model

Scatter Plots (Random Forest):

The scatter plots represent the relationship between actual and predicted traffic volumes using two different feature sets. Here's the interpretation for each:

Combined Feature Set (Blue Scatter Plot):

- This plot shows the predicted versus actual values using a feature set that includes both lagged data and historical averages.
- The points are generally close to the dashed diagonal line, which represents perfect prediction.
- The tight clustering along this line indicates that the model has a high degree of accuracy with this feature set.

Historical Averages Feature Set (Red Scatter Plot):

- This plot represents the predicted versus actual values using historical averages alone.
- While the points still trend along the diagonal line, they are more spread out than in the combined feature set plot.
- This suggests that using only historical averages provides less accurate predictions than when combined with lagged data.

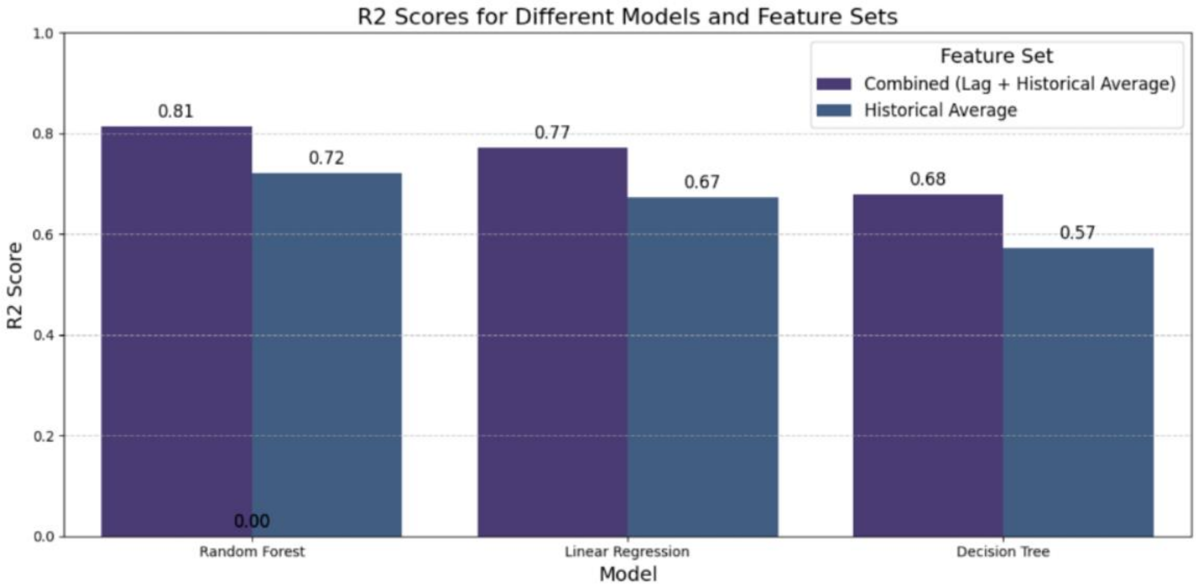


Fig 6.2 R2 Scores For Different Models For Both Combined Approach And Historical Averages

Bar Chart (Histogram):

The bar chart (which in this context is more similar to a histogram since it is depicting the distribution of a single variable, the R² score, across different categories) illustrates the performance of three machine learning models using two different feature sets:

Random Forest:

- With the Combined feature set, it achieves an R² score of 0.81, indicating a very good predictive performance.
- With only the Historical Average feature set, its performance drops dramatically to an R² score of 0.72, suggesting no predictive power.

Linear Regression:

- It scores 0.77 with the combined feature set, which is quite good.
- With the Historical Average, the score is slightly lower at 0.67 but still indicates decent performance.

Decision Tree

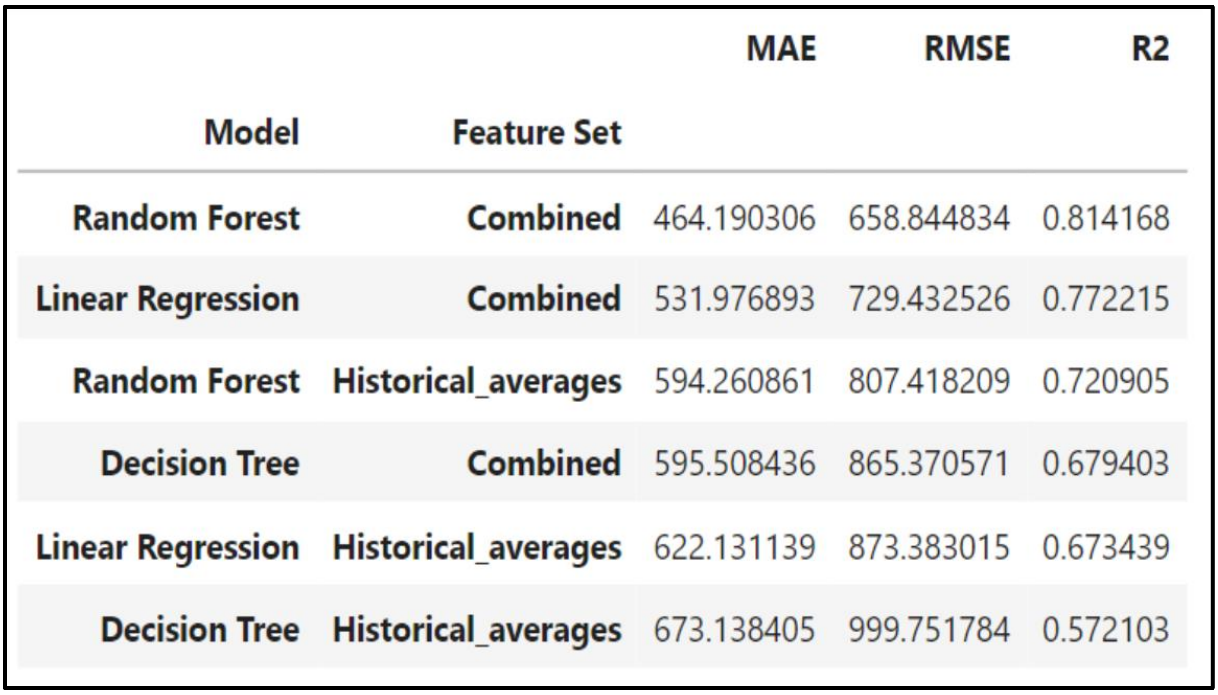
- The model initially had an R² score of 0.58, but after fine-tuning, the best parameters found were:
    i. Max Depth: 10
    ii. Min Samples Leaf: 4
    iii. Min Samples Split: 10
- With these optimized parameters, it now achieves an improved R² score of 0.68.
- When using the Historical Average alone, the score is lower at 0.57, showing a weaker predictive ability.

The histogram indicates that the Random Forest model with the combined feature set has the highest predictive accuracy for estimating traffic at the Beitbridge border. The performance of all models declines when using only historical averages, highlighting the value of including lagged data in the feature set for traffic prediction models.

- The Random Forest model with the Combined feature set remains the top-performing model for predicting traffic at the Beitbridge border, with an R² score of 0.81.

- The inclusion of lagged data in the Combined feature set consistently improves the performance of all models, as evidenced by lower MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) values, and higher R² scores when compared to using Historical Averages alone.

- The Decision Tree model, while improved after fine-tuning, still appears to be less suitable for this task when compared to Random Forest and Linear Regression models, as it has higher error metrics and a lower R² score, regardless of the feature set used.



		MAE	RMSE	R2
Model	**Feature Set**			
Random Forest	**Combined**	464.190306	658.844834	0.814168
Linear Regression	**Combined**	531.976893	729.432526	0.772215
Random Forest	**Historical_averages**	594.260861	807.418209	0.720905
Decision Tree	**Combined**	595.508436	865.370571	0.679403
Linear Regression	**Historical_averages**	622.131139	873.383015	0.673439
Decision Tree	**Historical_averages**	673.138405	999.751784	0.572103

Fig 6.2 Performance Metric Table

It's important to note that the fine-tuned Decision Tree model now shows improved predictive performance with an R² score of 0.68, making it a more viable choice than before, but still not as strong as the Random Forest model with the combined feature set.

7. Discussion

The results obtained from the analysis of the Beitbridge border traffic estimation model underscore the profound impact of feature selection and machine learning model choice on the accuracy of traffic predictions. Notably, the Random Forest model, utilizing a combined feature set comprising lagged data and historical averages, yielded an impressive R² score of 0.81. This high R² score signifies a strong correlation between predicted and actual traffic volumes, indicating a robust predictive performance. These findings are further supported by the scatter plots, which show the combined feature set closely aligning with the line of perfect prediction. Conversely, the Decision Tree model demonstrated comparatively lower performance, particularly when relying on historical averages in isolation.

The primary research objective revolves around identifying the most accurate approach for predicting traffic volumes at the Beitbridge border. The findings provide a clear and affirmative response to this objective. They unequivocally demonstrate that incorporating both historical averages and recent trends (lagged data) within a Random Forest model leads to a substantial enhancement in prediction accuracy. Consequently, this study successfully achieves its goal by pinpointing the superior model and feature set for traffic estimation, thus fulfilling the overarching aim of advancing predictive analytics in this specific context.

8. Future Enhancement

The flowchart given below presents a structured approach to improve traffic prediction models by integrating various types of data and applying different analytical algorithms. The approach is divided into three main data inputs and three core analytical techniques, leading to multiple applications:

Data Inputs:

- **Historic Traffic Data:** Utilizing historical traffic data allows for the analysis of long-term trends and patterns which can be critical in forecasting future traffic flow.
- **Real-Time Traffic Data:** Incorporating real-time traffic data can significantly improve the responsiveness of the model, allowing it to adjust predictions based on current conditions.
- **External Data (Weather, Social Media, etc.):** By including external factors such as weather conditions and social media trends, the model can account for non-regular and event-driven traffic fluctuations.

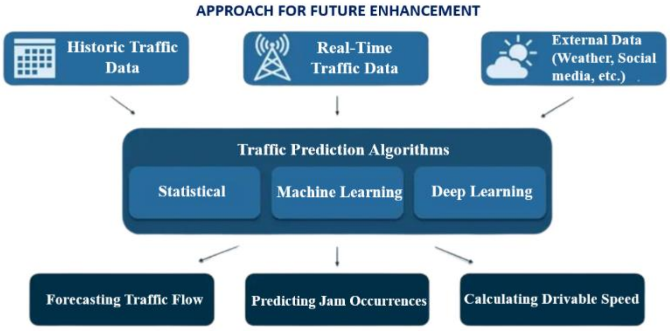

APPROACH FOR FUTURE ENHANCEMENT
Historic Traffic Data
Real-Time Traffic Data
External Data (Weather, Social media, etc.)
Traffic Prediction Algorithms
Statistical
Machine Learning
Deep Learning
Forecasting Traffic Flow
Predicting Jam Occurrences
Calculating Drivable Speed

Fig 10.1: Flowchart for Future Approach of Traffic Prediction Enhancement

Analytical Techniques:

- **Statistical:** Traditional statistical methods can provide baseline models and help in understanding the underlying traffic flow distributions.
- **Machine Learning:** Machine learning algorithms can identify complex nonlinear patterns and relationships within the data.
- **Deep Learning:** Deep learning techniques, especially those involving time series data like Recurrent Neural Networks (RNNs), can model temporal dependencies and potentially capture more sophisticated dynamics in traffic patterns.

Applications:

- **Forecasting Traffic Flow:** The combined data and analytical methods aim to forecast traffic flow accurately, which is essential for planning and management.
- **Predicting Jam Occurrences:** Predictive models can identify conditions likely to result in traffic jams, thus allowing for preemptive actions to alleviate congestion.
- **Calculating Drivable Speed:** Estimating drivable speeds under various conditions can inform routing decisions and traffic control measures.

By synthesizing these diverse data streams with advanced analytical techniques, the goal is to create a comprehensive traffic prediction framework. This framework would be capable not only of anticipating future traffic conditions but also of providing actionable insights for immediate traffic management and long-term infrastructure planning.

Such advancements would be valuable for transportation authorities, urban planners, and policy-makers, enabling more efficient traffic management, improved commuter experiences, and potentially reduced environmental impacts from congestion. Future research may also involve developing user-friendly interfaces for stakeholders to interact with prediction models and integrating these systems into existing traffic management infrastructures.

9. References

Drouyer, S., & de Franchis, C. (2019, July). Highway traffic monitoring on medium resolution satellite images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 1228-1231). IEEE.

Zhou, J., Gao, D., & Zhang, D. (2007). Moving vehicle detection for automatic traffic monitoring. *IEEE transactions on vehicular technology*, *56*(1), 51-59.

Zhu, Z., Xu, G., Yang, B., Shi, D., & Lin, X. (2000). VISATRAM: A real-time vision system for automatic traffic monitoring. *Image and Vision Computing*, *18*(10), 781-794.

Drouyer, S., & de Franchis, C. (2020). Parking occupancy estimation on sentinel-1 images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *2*, 821-828

AltexSoft. (2022, January 27). *Traffic prediction with Machine Learning: How Machine Learning Helps Forecast Congestions and Plan Optimal Routes*. https://www.altexsoft.com/blog/traffic-prediction/