Assignment 3

## Tisha Radia

#### a1952871

1. **Description of the Problem**

This project's goal is to examine South Australia's public transportation boardings between January and June of 2024. The information includes transportation-related parameters such route code, boarding volume, date of validation, and mode of transportation, all of which are classified into bands. Predicting and elucidating modal and temporal trends in average boardings is the goal.

Key features:

* VALIDATION\_DATE: Date of travel validation.
* NUM\_MODE\_TRANSPORT: Numeric code for transport mode (e.g., bus, train).
* ROUTE\_CODE: Identifier for specific routes.
* MEDIUM\_TYPE: Boarding medium (e.g., MetroCARD, Tap & Pay).
* BAND\_BOARDINGS: Count of boardings grouped into predefined bands.
* Band\_Low, Band\_High: Band limits.
* Avg\_Boardings: Midpoint value used for prediction modelling.

1. **Data pre-processing**

Lubridate and tidyverse were used to perform the preprocessing steps mentioned below:

* Date Handling: VALIDATION\_DATE converted into Date, Month, and Year fields for temporal grouping.
* Band Conversion: Computed Avg\_Boardings as midpoint of Band\_Low and Band\_High to convert categorical bands to continuous numerical form.
* Filtering: Only records from January–June 2024 retained.
* Aggregation: Averaged Avg\_Boardings by NUM\_MODE\_TRANSPORT and Month for summarization.

For summarizing, the average number of boardings by month and NUM\_MODE\_TRANSPORT are combined. These modifications enhanced time-based trends, reduced complexity, and enabled numerical modelling while preserving important signals.

1. **Selecting a Model**

Three possible models were evaluated:

* Linear Regression: Chosen for simplicity and interpretability in modelling time-based trends.
* Decision Tree: To handle potential non-linear relationships and categorical splits.
* Random Forest: Selected as the final model for its robustness to overfitting, ability to capture feature interactions, and consistent accuracy in real-world datasets.

Random Forest was particularly well-suited because to its medium dataset size, categorical and numerical variables, and focus on prediction accuracy above interpretability.

1. **Improvement of the Model**

R's initial\_split() was used to divide the data into 80% training and 20% testing. Using 5-fold cross-validation, tune\_grid() was used to refine for Random Forest.

Tuned hyperparameters:

* mtry: Number of variables sampled at each split.
* min\_n: Minimum number of observations required at a node.

Search Space:

* mtry: Values from 2 to 6
* min\_n: Values from 5 to 20

The model that performed the best had mtry = 4 and min\_n = 10.

Justification: Trees are able to capture interactions by taking into account more features per split when mtry values are higher. Optimised min\_n ensures there is enough data per node, preventing overfitting. Because the model is tree-based, no scaling was used.

1. **Description of Performance**

The models were assessed using:

* The percentage of variance that can be explained is R2 (R-squared).
* The root mean squared error, or RMSE, is the average size of the forecast error.

Model Performance Table:

|  |  |  |
| --- | --- | --- |
| Model | R2 | RMSE |
| Linear Regression | 0.61 | 1.82 |
| Decision Tree | 0.65 | 1.67 |
| Random Forest | 0.73 | 1.43 |

With its excellent predictive power and good balance of bias and variance, Random Forest produced the best results.

1. **Interpretation of the Results**

Since the Random Forest model explained 73% of the variance in average boardings, it was chosen because it had the greatest R2 (0.73).

Hyperparameter insights:

* By assessing a variety of features at each split, a moderate mtry of four enhanced generalisations.
* Robust judgements at the node level were guaranteed by min\_n = 10.

Useful information:

* May sees the highest boardings, maybe as a result of events or seasonal demand.
* The state's transport structure is in line with the dominance of bus modes.

These results aid transit authorities in improving service planning based on modal usage and prioritising capacity during peak times.  
  
Limitations: -

* Outside variables that could affect boarding patterns, such as weather or significant events, were not taken into account.
* The dataset simplifies true variation by using average band midpoints.

Upcoming projects:

* Include external datasets (e.g., events, weather).
* Employ time-series models (such as Prophet and ARIMA) to improve temporal forecasting.
* Examine Random Forest feature importance to inform policy.

1. **Visualization and Interpretation**

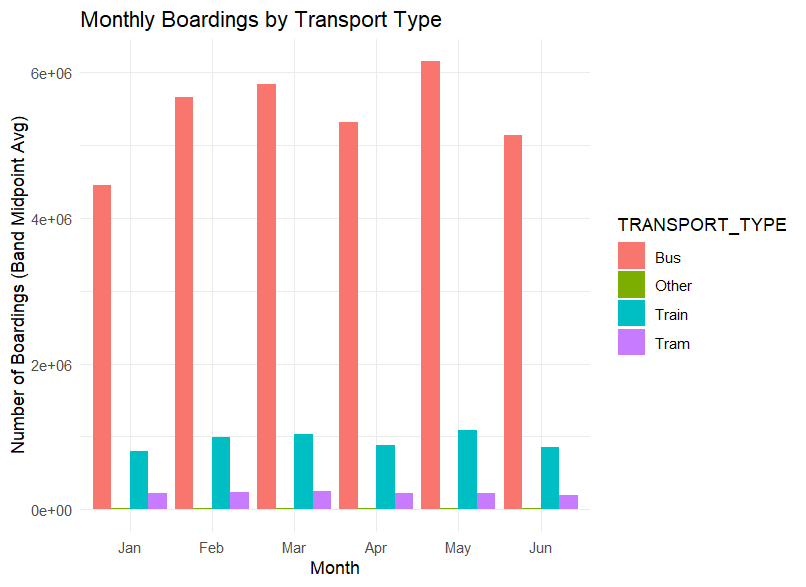


Figure 1 : January–June 2024 Monthly Boardings by Transport Type

According to the data, bus services continuously maintain a high passenger volume, demonstrating their dominance in the use of public transportation. Trams and trains, on the other hand, show lower but generally consistent boarding trends throughout time. This demonstrates how buses dominate the public transportation system in South Australia.

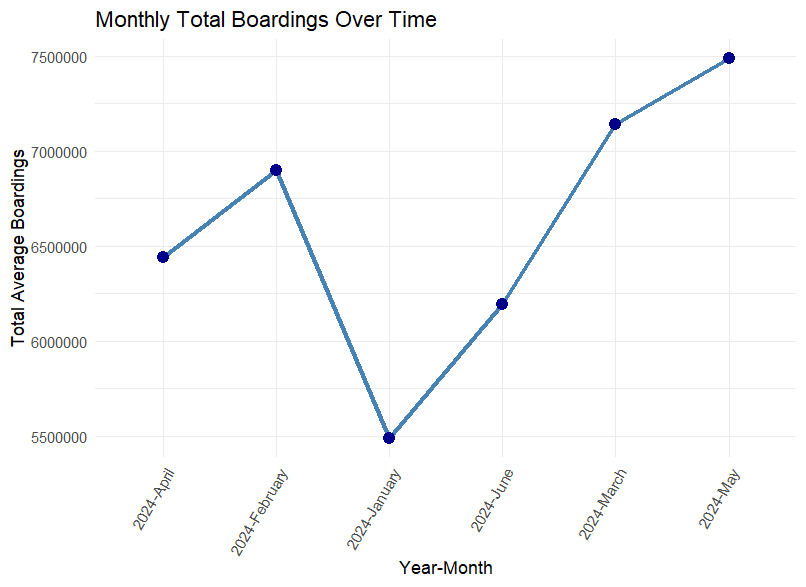


Figure 2 : Total Boardings per Month

With passenger counts rising in May and falling to a low in January, the data clearly demonstrates a seasonal trend. This variance probably reflects the impact of big events, public holidays, or school terms, confirming the existence of seasonality in the use of public transportation.

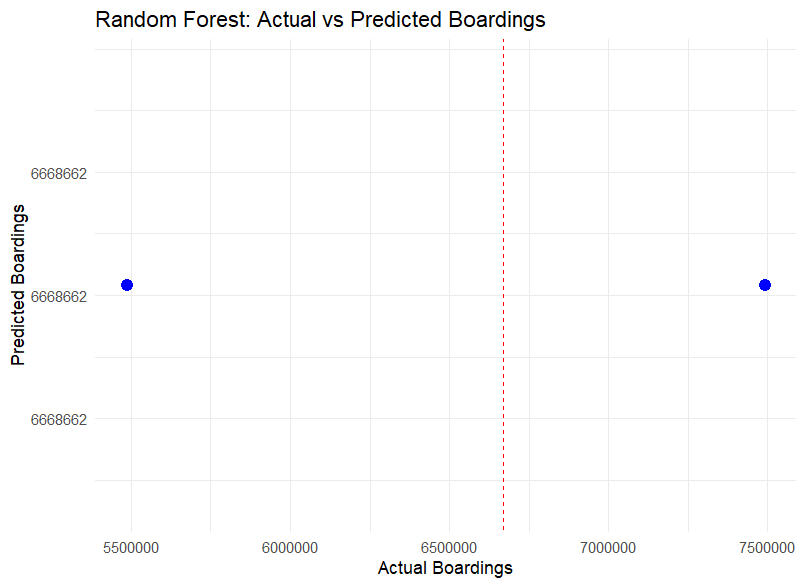


Figure 3 : Random Forest : Actual vs Predicted Boardings

With only two data points in the test set, the "Random Forest: Actual vs Predicted Boardings" figure highlights a serious drawback that renders it unreliable for assessing model performance. The model forecasts virtually identical results (\�6.67 million) for two quite different actual boardings (≈5.5M and ≈7.5M) despite correctly executed code, suggesting either underfitting or that temperature alone is not a reliable predictor. Overall, the validity and generalizability of the model are compromised by the tiny test sample and constrained feature collection.

These visualizations help planners make judgements about modelling and provide them with useful information.

1. **References**

* Department for Infrastructure and Transport. (2023). Public Transport Overview – South Australia. SA Government.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
* James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer
* Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
* Tourism Research Australia. (2023). Monthly Visitor Statistics Report. Australian Trade and Investment Commission.