Forecasting in the Shipping & Distribution Industry

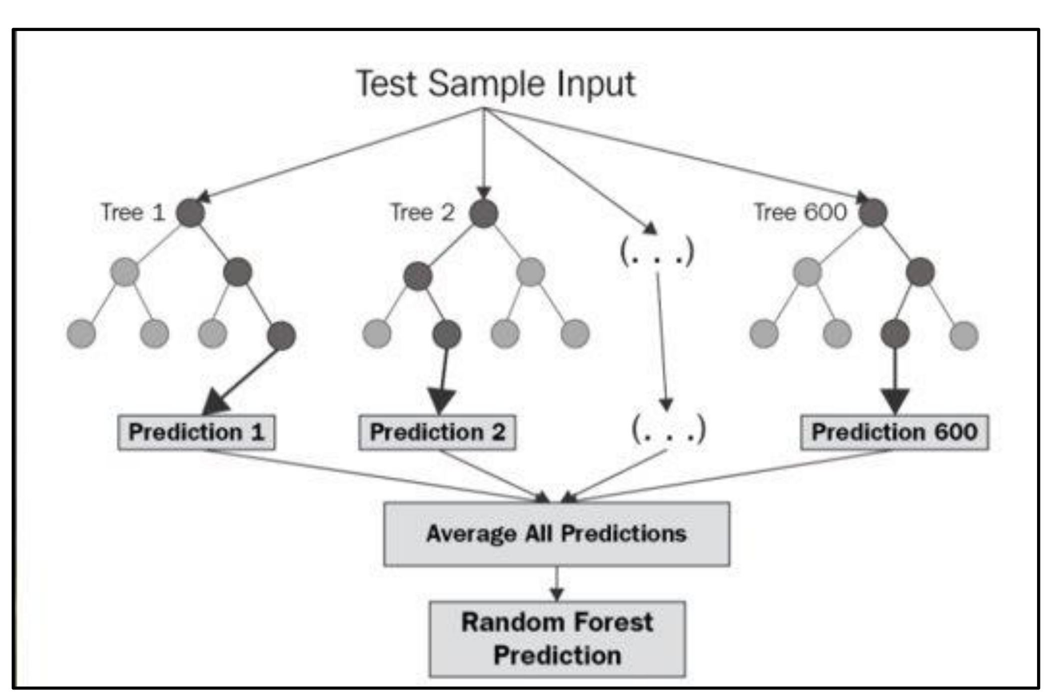
Paul Florio 2026 and George Thorp 2025

Forecasting Background

Analyzing shipping data from Notions Marketing, we created a model to predict shipping rate timings using forecasting models. Shipping rate forecasting involves predicting future shipping rates based on historical data, market trends, and various economic indicators. These forecasts are essential for planning and decision-making, helping companies anticipate market changes and make informed decisions about fleet management, chartering, and pricing strategies.

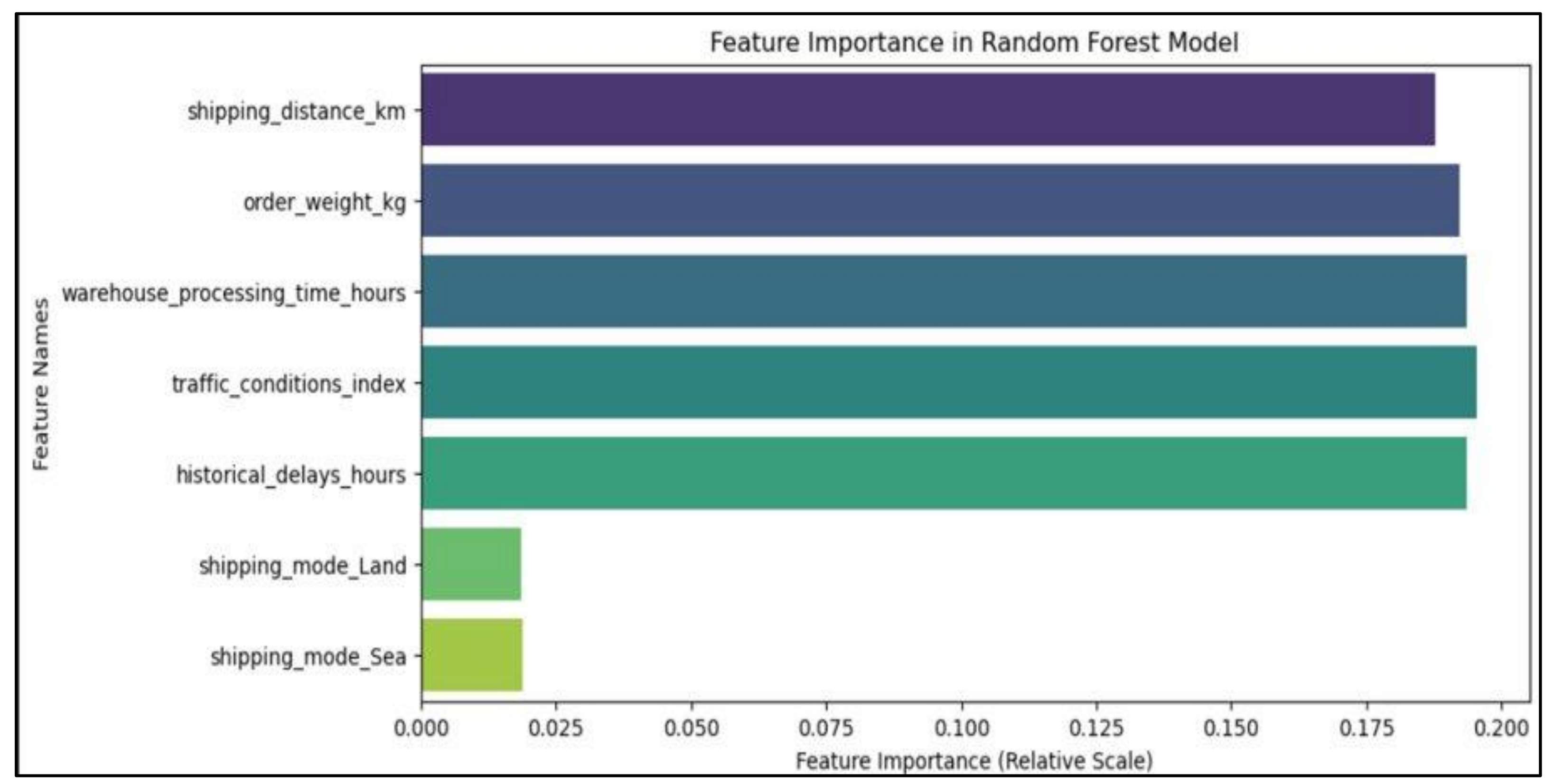
Data and Methods

10,000 records were pulled containing their Shipping Distance in kilometers, Weight in kilograms, Time Processed in hours, Traffic Conditions Index, Delay Times in hours, and Total Time of Delivery in hours. We utilized the Random Forest model which is an ensemble learning technique that builds multiple decision trees and combines their outputs to make a prediction. In regression tasks, it averages the predictions of all trees to provide a more stable and accurate estimate.



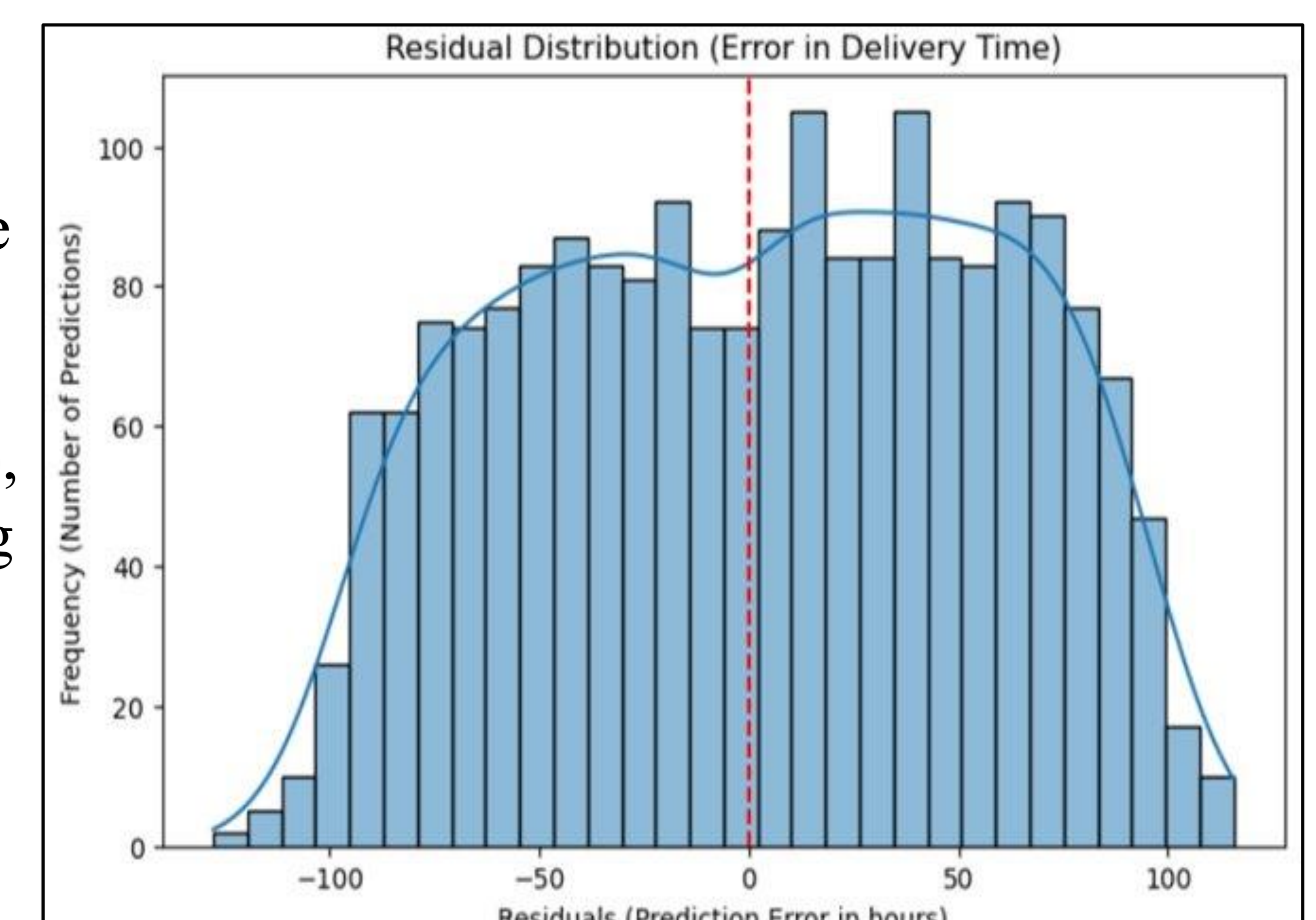
Results

For our results, our algorithm outputs the Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error for the expected delivery time. On average, the model's predictions are off by 48.15 hours (~2 days) from the Mean Absolute Error. A Mean Squared Error of 3131.28 suggests that a few extreme mispredictions are skewing the results. Lastly, for the Root Mean Squared Error the model's typical error is almost 56 hours (~2.3 days).



This Feature Importance chart shows which features impact delivery time (hours) the most, and the higher values indicate more influence on delivery time. The three most important features are Traffic Conditions Index, Warehouse Time Processed, and Historical Delay Times.

This histogram shows the distribution of residuals prediction errors ranging from around ±110 hours. The x-axis represents the error in hours, while the y-axis shows how often each error range occurred. The distribution is roughly symmetric around zero, as indicated by the red dashed line, suggesting the model does not consistently over or under predict. The blue curve reinforces this balanced shape. While most errors cluster around zero, some are quite large, indicating that while the model is generally unbiased.



Our algorithm was developed to mitigate risk and increase strategic planning but there are outside factors that would be difficult to incorporate into the model. Regulatory changes and geopolitical events could arise at any moment to which our model can't quantify.

Forecasting Importance

Businesses utilize forecasting models to improve delivery estimates, optimize logistics, and reducing costs. The utilization of a Random Forest model is significant for how it handles the complex interactions such as the ability to account for how features impact each other. Forecasting with a Random Forest model works well with missing data, since the model can handle missing values better than other models and avoids overfitting, which is critical with new shipping data.

Prospective Refinements

With our model's prediction error being calculated around 2 days, we could look to improve the model by adding other features such as weather conditions, construction delays, etc. Other additions to our model could be a cost benefit analysis for fuel prices, supply & demand pricing, and other cost driven expenses. This could be beneficial for comparing the delivery predictions with pricing to optimize the most effective way to forecast.

Acknowledgements

We would like to give a special thank you to Notions Marketing for assisting us with their data contribution and the Data Analytics Department for guiding us on our project.