Capstone Project

Traffic Volume Prediction using Machine Learning

By
Navin Sanjay
Date: 19/05/2022

Table of Contents

Problem statement

The transportation industry has a crucial role in the economic and social development of New Zealand. It provides connections between people, goods, and services, facilitating trade, tourism, and overall movement. That's why it's important to continue to develop solutions that will aid the transportation industry.

Huge demands are placed on the current transportation infrastructure as economic, and population continue to grow. Travel tends to increase as the population increases[1].
The use of new technologies and information can continuously help address issues faced in the transportation industry.

Efficient traffic flow management plays a key role in the impact of economic productivity, quality of life, and environmental sustainability. Predicting traffic flow volume may play a significant role in optimising transportation systems, aiding in decision-making processes, and increasing overall efficiency. Through the use of data science techniques and analytics, it is possible to gain insights into traffic patterns, identify issues, and devise strategies for effective management in the transportation industry.

The problem discussed in this project is to develop a predictive model that can accurately forecast the traffic volume at a specified location in Auckland, New Zealand for a specific date and time. Historical traffic data was leveraged along with relevant features such as date, time, day of the week, month, and weather conditions to build a model. Machine learning techniques were used to capture patterns, trends and seasonality in the traffic data, enabling it to make accurate predictions for future traffic volumes. The predicted traffic volume will aid in optimizing transportation planning, traffic management and resource allocation. The ultimate goal of enhancing efficiency, reducing congestion and improving overall traffic flow in a target area.

Industry/ domain

Predicting traffic flow can provide significant benefits to the transportation industry and all associated industries. By accurately forecasting traffic patterns, transportation agencies in New Zealand can make informed decisions to improve efficiency, safety, and the overall performance of the transportation systems. The overall value is to provide reliable traffic flow predictions, which agencies can use to optimise traffic signal timings, infrastructure planning, road maintenance, etc. Ultimately, using predictive modelling can enable the transportation agency in New Zealand to create a more sustainable, efficient, and resilient transportation network that benefits both the economy and the quality of life for its citizens.

[1] https://www.transport.govt.nz/assets/Uploads/Report/TransportOutlookFutureOverview.pdf

Stakeholders

Various stakeholders are involved in this project. Transportation agencies, urban developers, communities, residents, and businesses can use this information with regards to their own context. The key stakeholders that this project targeted included transportation agencies and urban developers. Transportation agencies can plan road maintenance, traffic signal timing, and new infrastructure based on expected traffic demand. It will also aid in managing traffic flow and mitigating congestion. Urban developers can use this predictive modelling tool to gain insights on how to design and develop transportation infrastructure (e.g., road networks, highways, bridges, and public transportation systems) to handle expected traffic volumes.

Business question

The value added by being able to predict traffic volume offers various benefits to the transportation industry and related sectors. Some potential areas where value can be added include:

- Cost Reduction: Predicting traffic volume allows businesses to optimize their operations and reduce costs. Companies can minimise fuel consumption, and labour costs by minimising delays.
- Improved Decision Making: Predictive traffic volume makes way for more informed decisions and effective choices. Transportation authorities can use this information to allocate resources, plan infrastructure development, and implement targeted traffic management strategies.
- Economic Impact: The broader economy may also see the benefits of the results of traffic volume prediction. By reducing congestion and improving traffic flow, there may be an increase in productivity, stimulating more efficient operations and economic growth.

Data question

Predicting the traffic volume involves various external factors that may influence the number of cars seen on a given date and time. The most obvious being the traffic volume seen in a lagged interval (hourly, daily, weekly, etc.). The traffic count seen previously will give insight into what the traffic may be in the next interval. Other external factors that will affect traffic counts are holiday periods and weather. Holiday periods are expected to see fewer people on the roads as most people aren’t going to work. Weather may also affect the number of

people on the roads due to road conditions, safety concerns, events or festivals, and seasonal factors. The additional data would assist in the prediction of traffic.

Data

The data was sourced from NZ Transport Agency. Daily traffic volumes from various state highway count sites in 15-minute intervals, from 2013 to 2020. The data represents real world data. **https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::tms-traffic-quarter-hourly-jan-2013-to-sept-2020/about**
This project looks at the traffic count at Lincoln Rd., SH16 East Bound Interchange (Auckland, New Zealand); see Figure 1. The particular location is looking at traffic leaving the suburb and heading towards the city. This project also looks at the traffic count across 2017. It is important to note that since this is real-world data, there are a lot of impurities that need to be considered. Sensors were used to collect the number of cars crossing the particular point; therefore, the data is subject to technical and environmental factors. Also, due to the nature of recording, there may not be consistent recordings for every hour of every day across the year; however, in general, the recordings taken were consistent throughout the year of 2017.

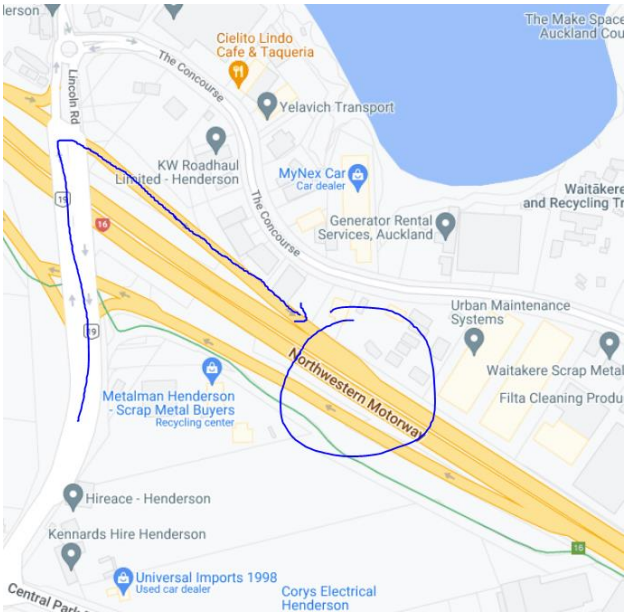


Figure 1: Location of recordings

Data science process

Data analysis

The dataset used included the traffic count for a given date and time. The raw dataset includes the vehicle class and direction; however, these were not important for this project, so they were removed. The raw dataset also recorded traffic in 15-minute intervals; however, this was compiled to show the traffic count in hourly intervals. Public holidays were also

included for the equivalent dates. The average rainfall and temperature for given months in Auckland were also added to the dataset. The data was loaded into Jupityer as dataframes, and initial checks were performed. Traffic count was the target variable. The data frames did not have null values. There were 7007 rows of data, representing the traffic volume for a given date and hour of the day. Looking at the raw data, there is no clear trend that is seen, and the data is regular. The date range is from March to December of 2017.

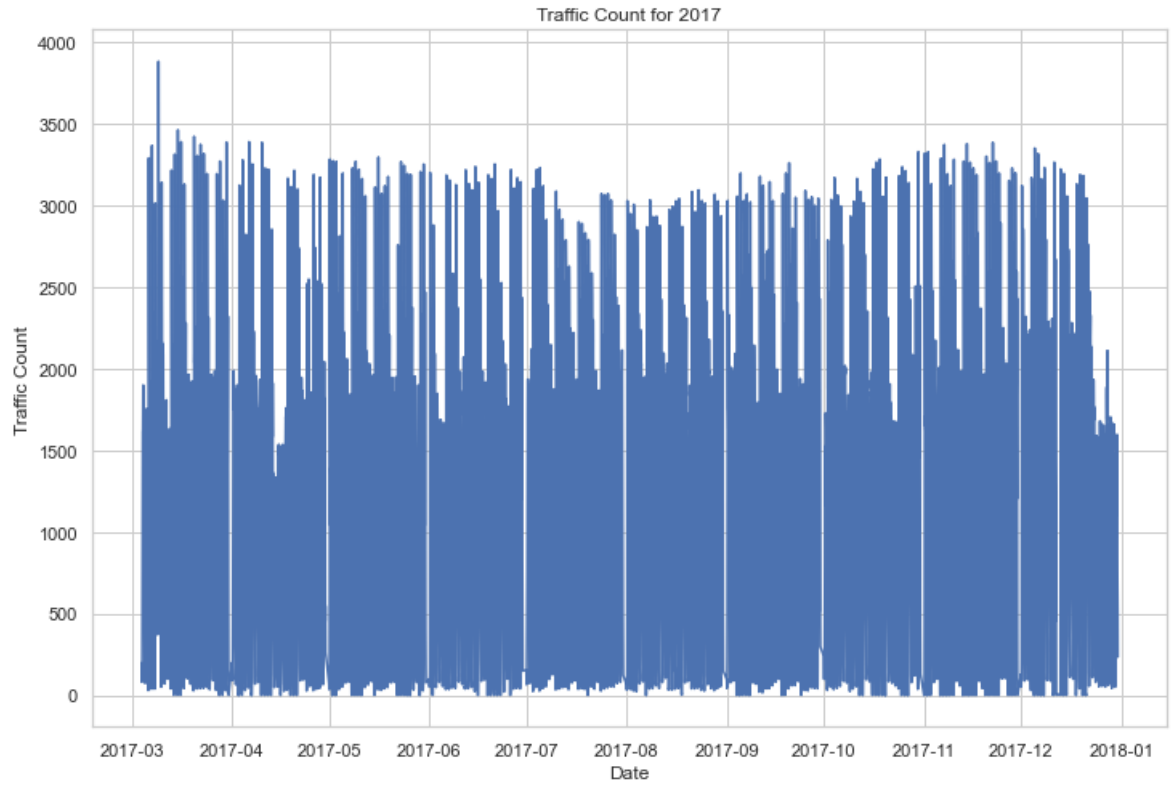


Figure 2: Traffic Count for 2017

The average traffic count per hour of the day across 2017 is shown in Figure 3. This gives us an insight into the seasonality and trend of the data. There is an overall increase in the average traffic count per hour across the year. The constant dips seen are on weekends, and larger dips represent public holidays. (shown in Figure 4). The highest average hourly traffic for 2017 was 1600 cars. In total, 8,840,597 vehicles were recorded in 2017. The total number of cars seen in a given month ranges from 750,000 to 920,000.

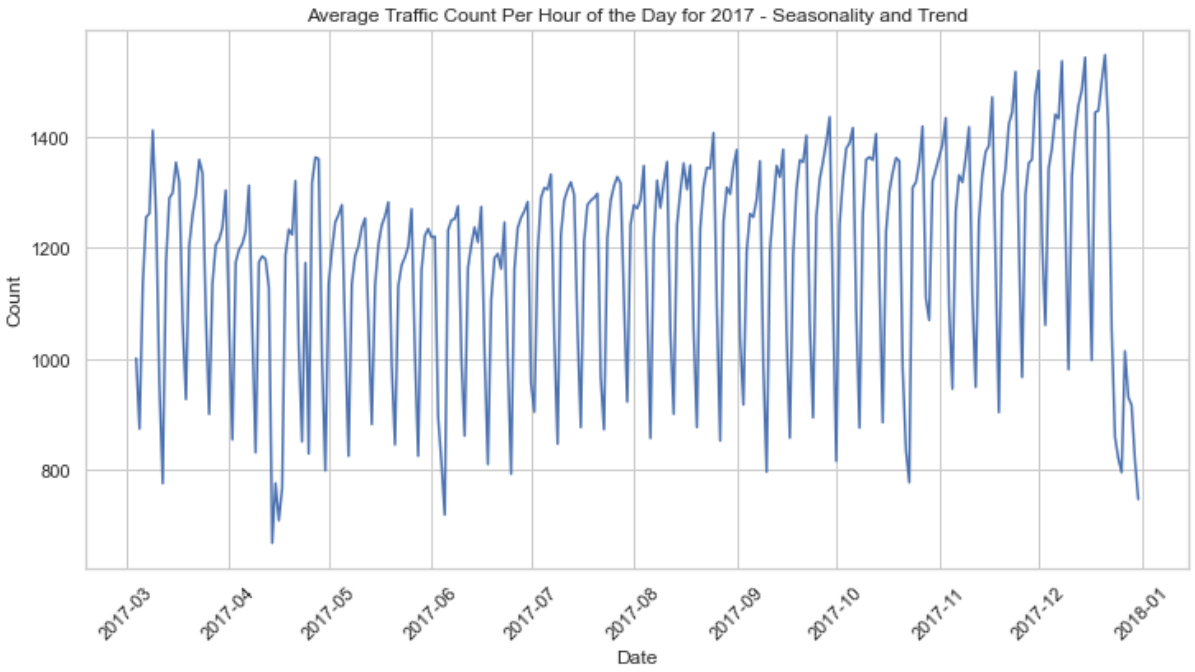


Figure 3:Average Traffic Count per Hour across 2017

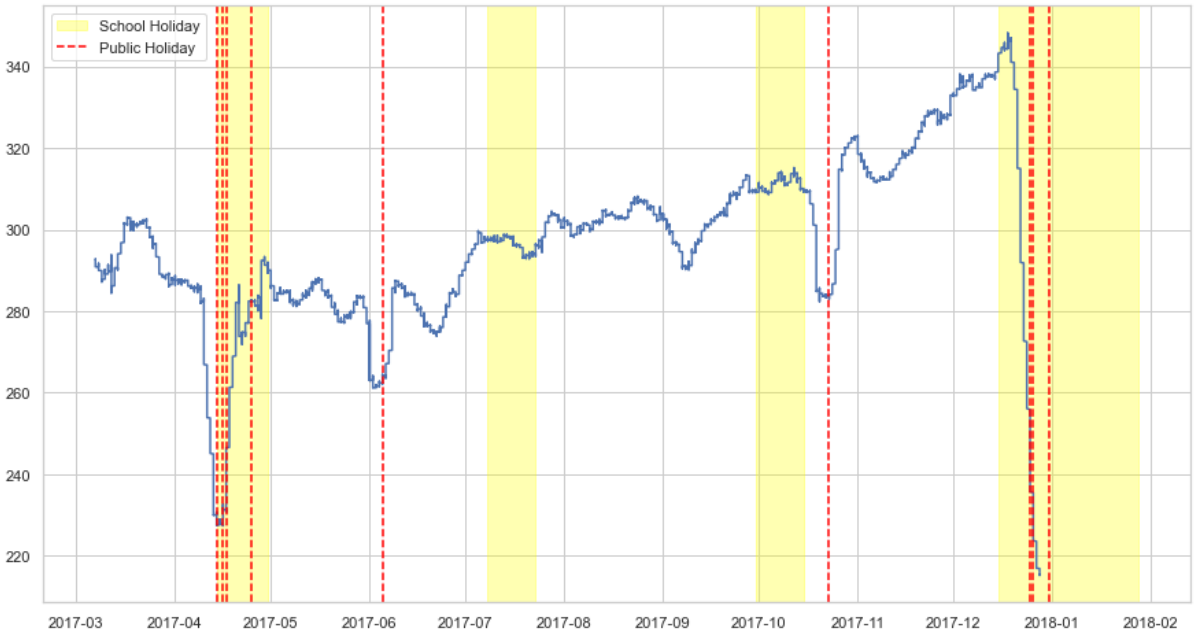


Figure 4: Traffic Count Trend 2017

Fridays see the highest traffic count per weekday. Mondays are the lowest. Various months experience different traffic counts for a given day of the week, however all months follow the same trend shown.

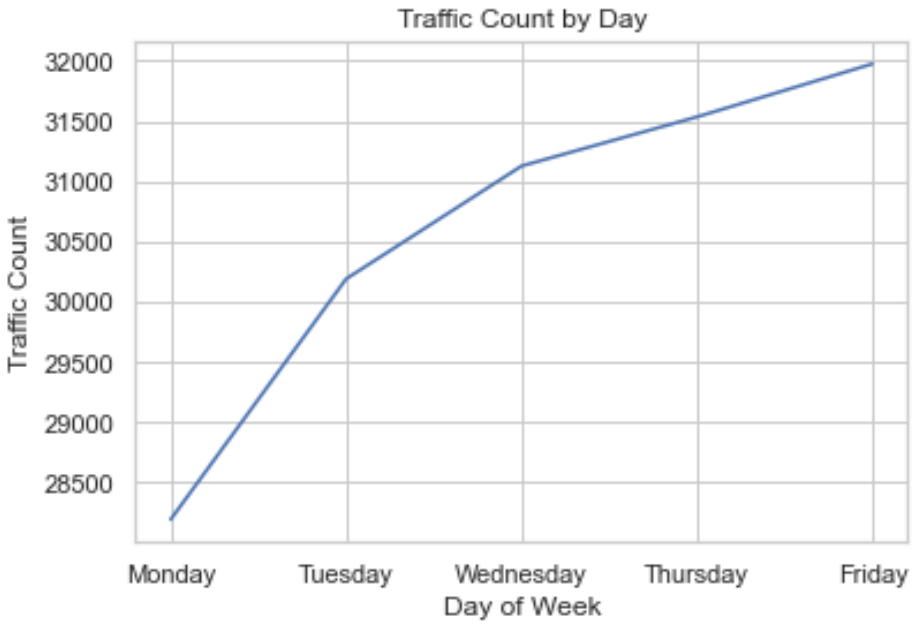


Figure 5: Traffic Count by Day

The average traffic count for a given time and day show that the peak times for weekdays are 5am - 9am which represent people leaving the suburb to go to work, see Figure 6. The next peak is the working crowd again, from times 4pm – 6.30pm. Fridays experience slightly less traffic during peak hours, however they have the highest post work traffic (in the night-time), which backs up Friday having the highest traffic count per day of the week.

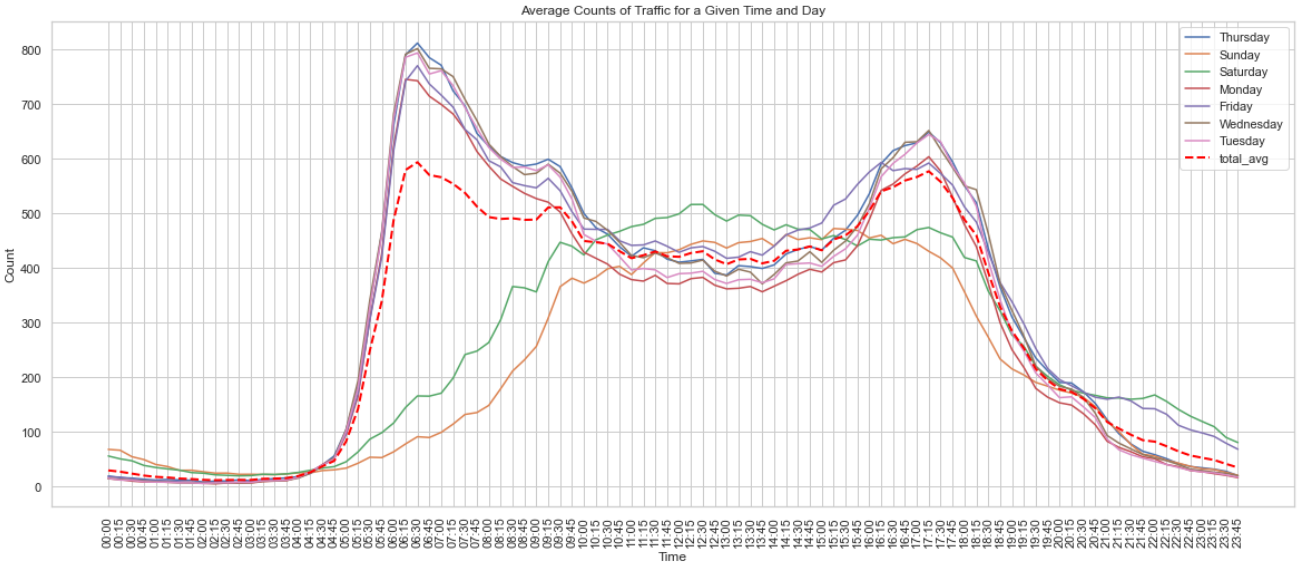
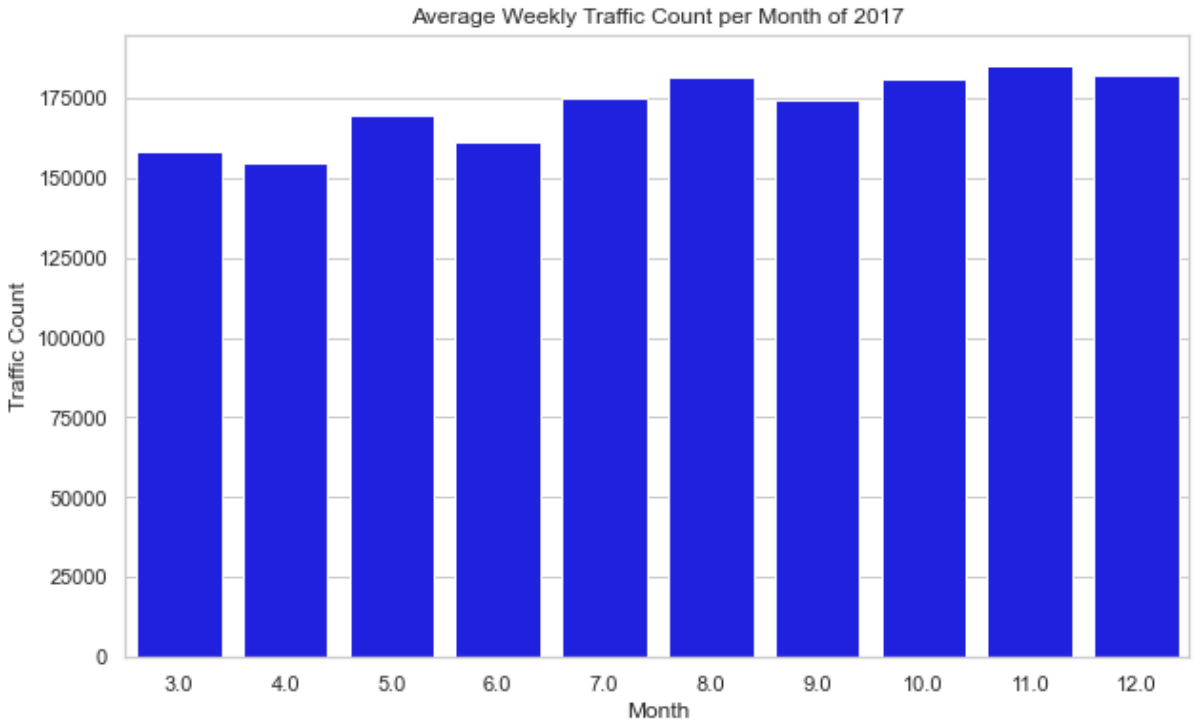
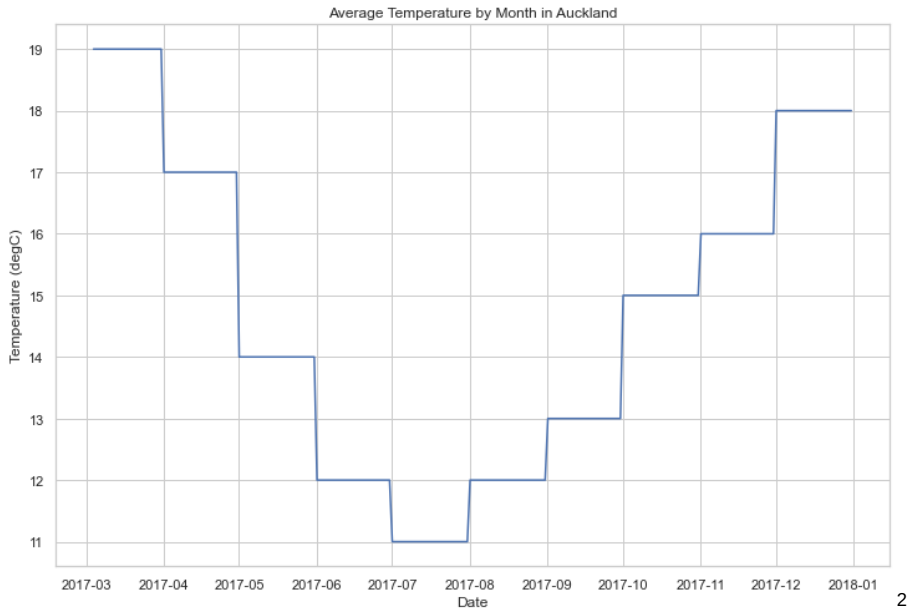


Figure 6: Average Traffic Count for a given time and day

The average weekly traffic per month shows that November has the highest traffic count per week. Holistically, the earlier half the year experience less traffic than the latter. This makes sense as the latter half of the year, especially November/December, is heading towards the holiday season. There will be more people out shopping, going to events etc. Also, the weather is generally better than the earlier months, see Figure 7 and Figure 8.





[2]

Figure 7: Average Temperature by Month in Auckland

[2] https://www.holiday-weather.com/auckland/averages/

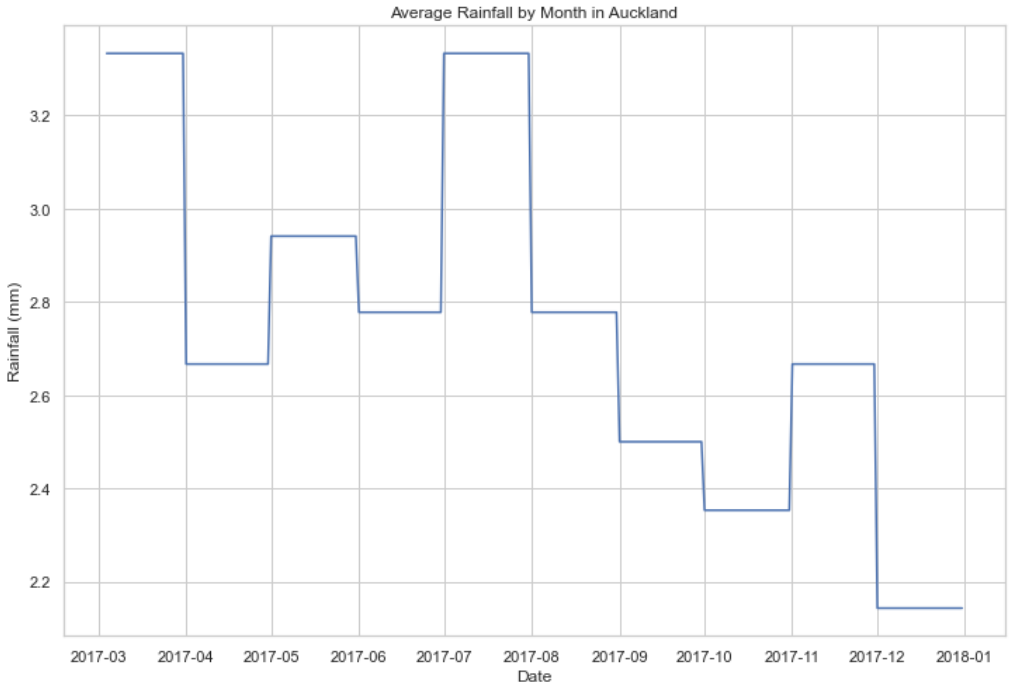


Figure 8: Average Monthly Rainfall in Auckland

Representing the traffic data into different components, Trend, Seasonality and Residuals give us additional insights.

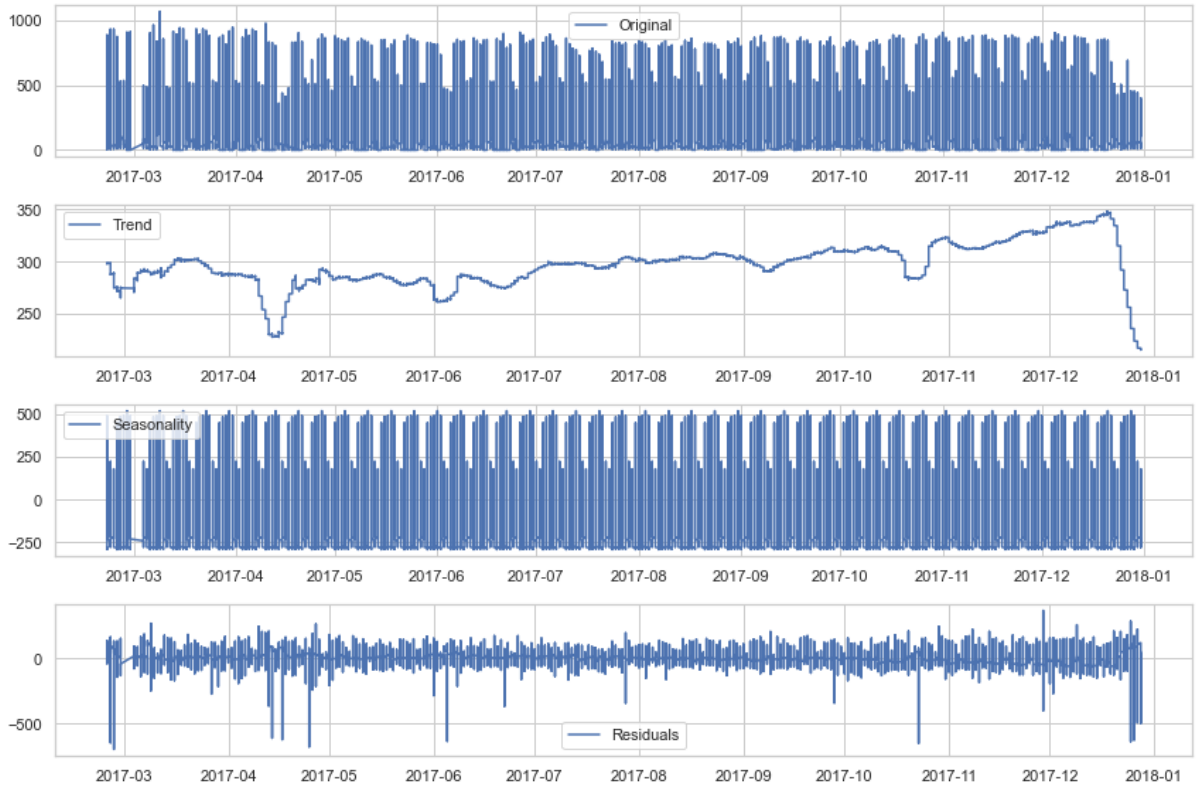


Figure 9: Seasonal Decomposition

The trend graph shows us the long term, persistent behaviour of the traffic count. Generally, we see an increase in traffic as we go throughout the year but there are some points in

months where there is a steep drop. These indicate various holidays as displayed previously. Public holidays have more of an impact on traffic count than school holidays.

The seasonal graph tells us the repeating patterns/cycles that occur over a given period (Weekly). The Seasonality that we see are the general increase from Monday-Friday in traffic count, but the steep decline when we get into the weekends for each month.

The residuals represent the error component, which shows the random fluctuations or noise that the trend and seasonal can't explain. It shows the unpredictable variations. For the traffic count, we mainly see the noise in the negative direction. I think this can be explained by the inconsistency of the recordings of the traffic. Some days or times the sensor may be broken, or they are re-calibrating it, etc.

A Pearson correlation heatmap was used to find the correlations between the target variable traffic count and features (Figure 10). This heatmap includes feature-engineered variables. Here we see that the lagged traffic counts are the highest correlated, which is expected. The average temperature and rainfall also have a moderately positive correlation.

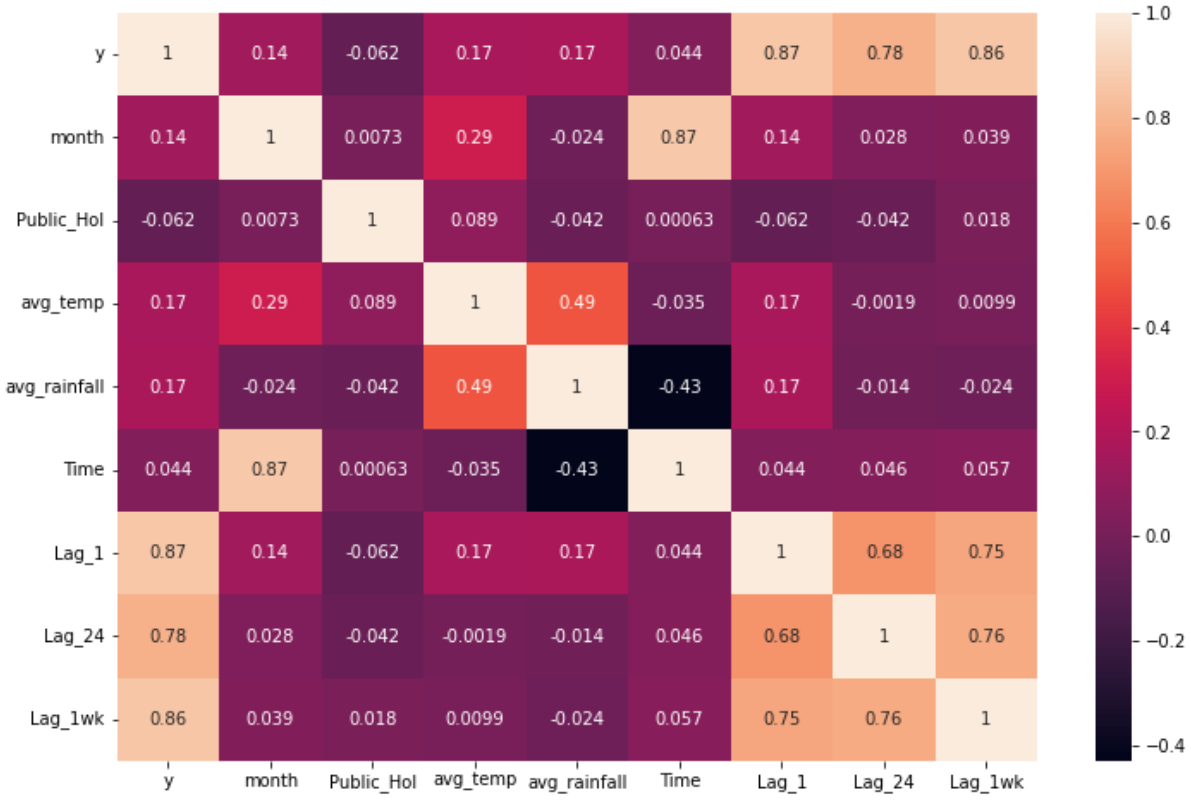


Figure 10: Pearson Correlation Heatmap

To succeed in solving the various business problems, predictions would ideally take place a week in advance. Therefore, even though having 1- and 24-hour lags would significantly improve predictions, it is not useful from a business case standpoint. A lag of 1 week was used to aid in predictions only. This gives ample time for a stakeholder to plan and allocate resources accordingly.

Modelling

The features used to predict the target variable traffic count include traffic count 1-week prior, average temperature, average rainfall, and public holidays. Some of the features have a stronger correlation than others with the target variable. Features were selected with a correlation map, and conclusions were drawn about these features through feature importance and forward feature selection. Four models were used to predict the traffic count. The problem at hand is a regression problem, and the models chosen were linear regression, decision tree regression, random forest regression, and prophet. The metrics used to evaluate these models included root mean squared error (RMSE), mean absolute error (MAE), and R2. These models were evaluated against each other based on the performance metrics, and a final model was chosen as the best-performing.

The data was split into a training set and a testing set when modelling. The split was done in an 80/20 split, where the training set represented the first 10 months and the model was evaluated on the last 2 months of the 2017 data.

Linear Regression:

Linear regression was the base model used in this prediction. It captures the relationship between the traffic count (dependant variable) and the set of predictor variables (independent variables) that can influence traffic volume. The data was split into training and test sets. A linear regression model was trained using the training set, and the performance was evaluated based on the metrics mentioned earlier. Linear regression can effectively capture the relationship between factors such as time of day, day of the week, weather conditions, and traffic volume. It provides insights into the direction and magnitude of the impact of each predictor variable on traffic flow count, allowing transportation planners to make informed decisions based on the model's coefficients.
A limitation of the linear regression model is that it assumes a linear relationship between the predictors and traffic count, therefore it can't capture the non-linearity that may be present in the data. Refer to Table 1 to see the models scores.

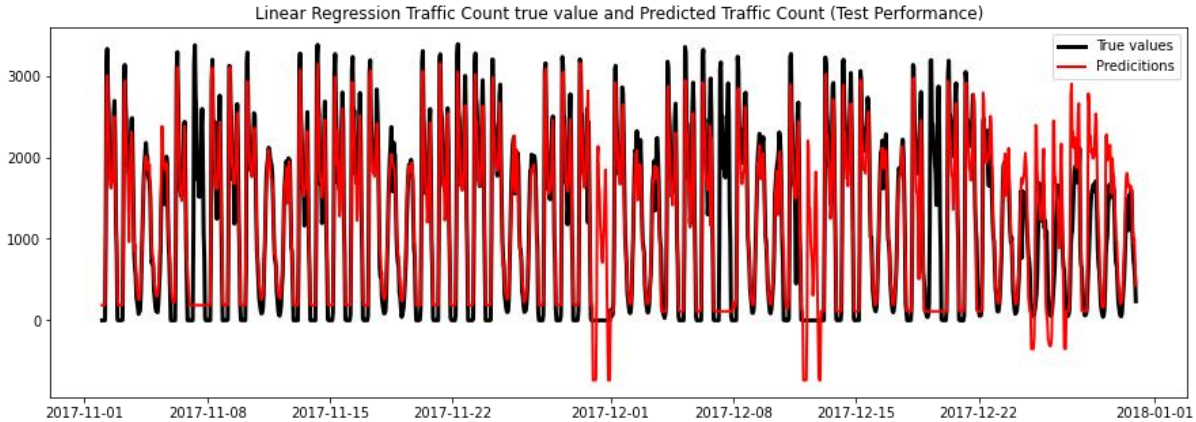


Figure 11: Linear Regression Test Performance

Decision-Tree Regression:

Decision-tree regression was a model used to predict traffic count as an improvement to the limitations of linear regression. It constructs a tree-like structure based on the dataset. A decision tree algorithm creates nodes and branches that represent conditions and possible outcomes. Each node is examined, and the algorithm identifies the predictor variable that best splits the data based on reducing variance or maximising information gain. The decision tree model can capture complex interactions and non-linear relationships between the predictors and traffic count. Traffic flow is influenced by various factors, including time, weather, road conditions, and events. Decision trees can effectively split the data based on these factors and create a hierarchy of conditions that lead to different traffic flow outcomes. By recursively partitioning the data, decision trees can capture complex relationships and interactions, making them a suitable choice when the relationship between predictors and traffic flow counts is non-linear. Decision trees are prone to capturing noise and exhibiting high variance.

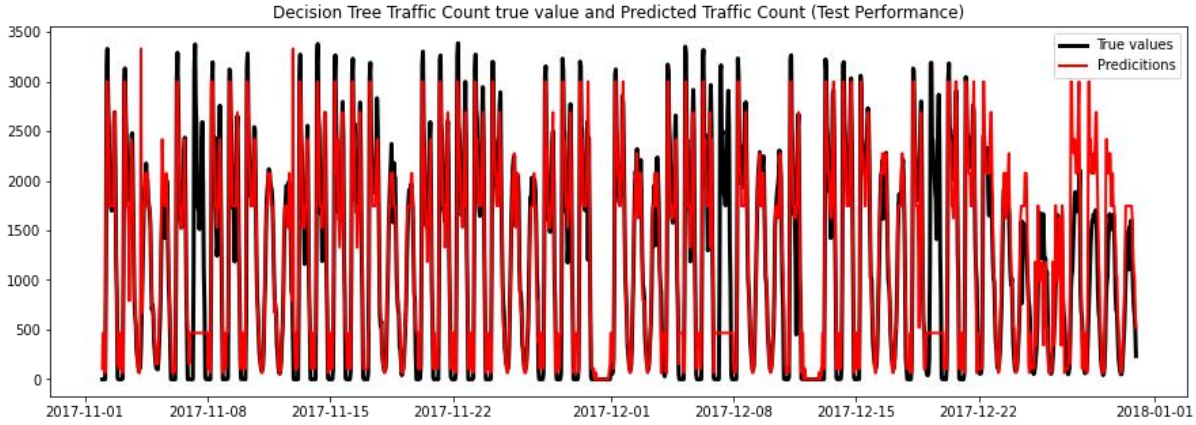


Figure 12: Decision Tree Test Performance

Random Forest Regression:

Random Forest Regression improves upon the decision tree regression algorithm by leveraging the power of ensemble learning and decision trees. It combines multiple weak learners (e.g. Decision tree stumps) to form a forest, where each tree was trained on a random subset of data and predictor variables. The randomness of this introduces diversity and reduces overfitting. Random forest regression effectively handled non-linear relationships, interactions, and outliers in the data. The averages are taken of multiple trees, and achieves greater generalisation. With its ability to handle complex data patterns and

mitigate overfitting, random forest regression proved to be a valuable model for predicting traffic count.

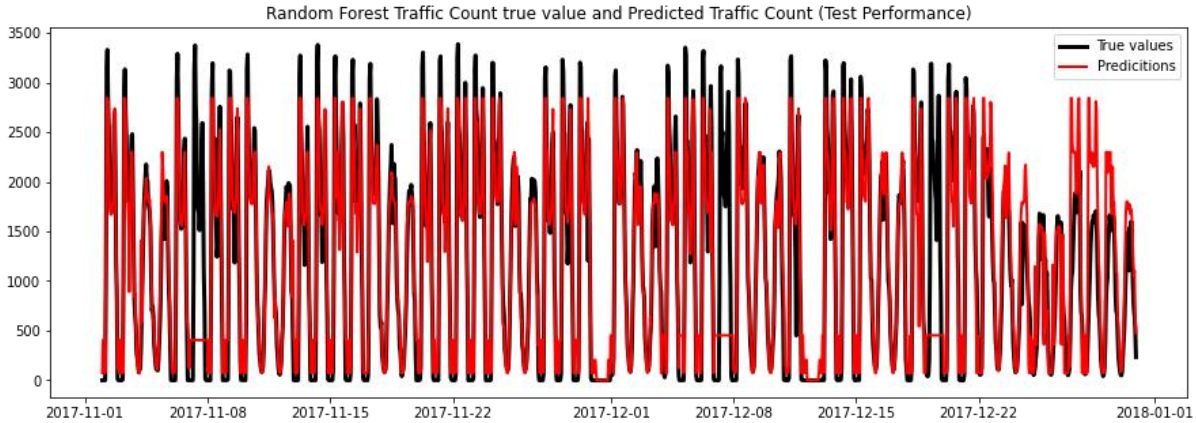
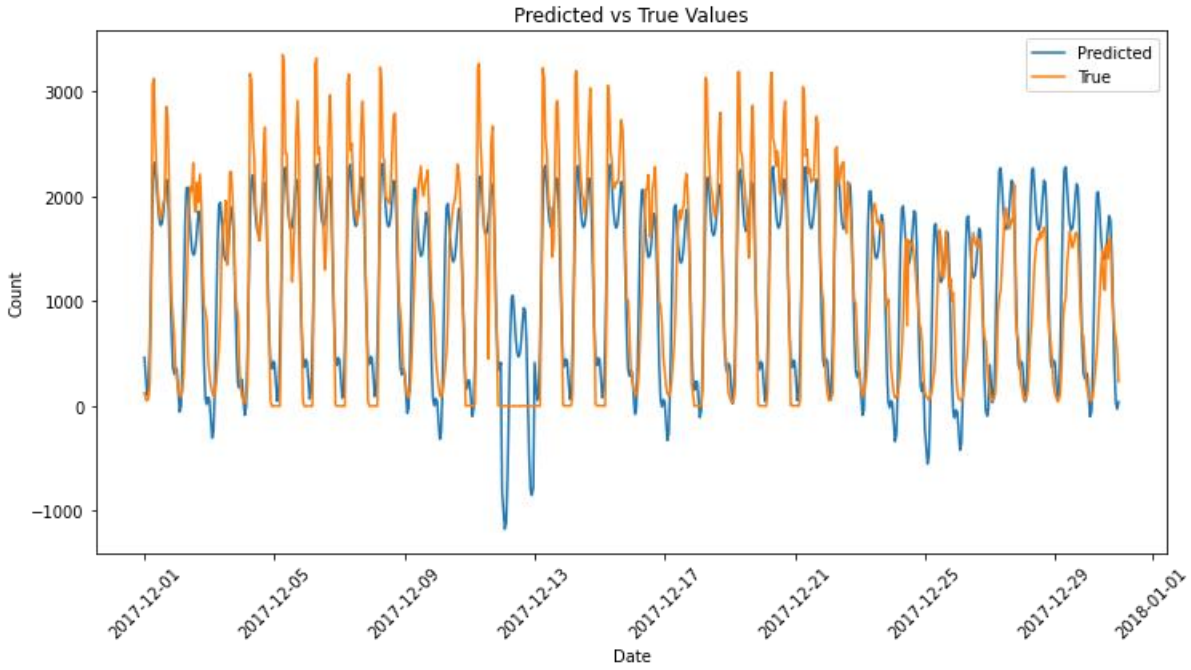


Figure 13: Random Forest Test Performance

Prophet Model:

The prophet model is a popular time-forecasting model developed by Facebook. It incorporates trends, seasonality, and holidays to capture the underlying patterns and dynamics in the data. Prophet was trained using historical traffic count data, including the corresponding dates and counts. It is able to consider both long-term and short-term fluctuations. It also accounted for seasonal patterns that repeat over a specific period; in this case, it's the daily and weekly variations in traffic. The included regressors were average temperature, average rainfall, and holiday periods. However, the prophet model has limited support for complex relationships and assumes external factors have an additive effect on the forecast, which may not always hold true, especially in the traffic count context. The prophet model works best with time series that exhibit strong seasonal effects and several seasons of historical data. Traffic count data does not always exhibit strong seasonality, and the data used was only for one year.



TRAINING				
Metric	Linear Regression	Decision Trees	Random Forest	Prophet
MAE	245.34	176.63	178.75	332.72
RMSE	408.63	321.66	324.05	498.45
R2	0.80	0.88	0.88	0.71

TEST				
Metric	Linear Regression	Decision Trees	Random Forest	Prophet
MAE	287.01	256.75	252.66	426.58
RMSE	524.54	461.95	447.64	584.18
R2	0.72	0.78	0.80	0.63

Table 1: Model Evaluation

Performance Metrics

Mean Absolute Error:

MAE is a commonly used metric when doing predictive modelling. It provides a straightforward measure of the average absolute difference between the predicted and actual values. MAE is easily interpretable as it directly represents the average absolute difference between the predicted and true values. Especially in the context of traffic count, we care about both overestimation and underestimation.

Root Mean Square Error:

RMSE is another commonly used metric for predictive modelling. It provides an overall measure of the average magnitude of the prediction errors. RMSE penalises higher errors more than MAE due to the squaring option, therefore it is more sensitive to outliers.

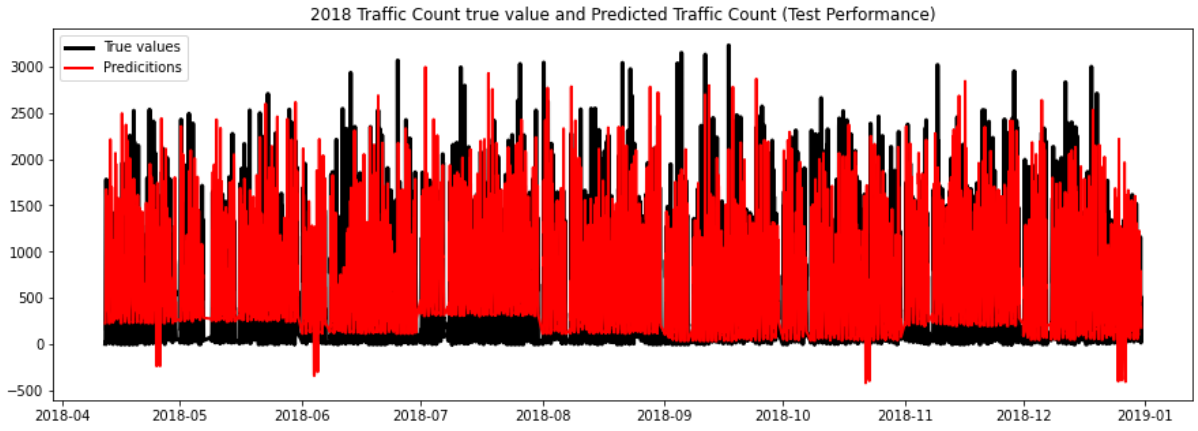
R2 (Coefficient of Determination)

The R2 coefficient is a value performance metric as it measures the proportion of the variance in the target variable that is described by the predictors. It displays how well the model fits the observed data. A higher R2 is a better fit, ranging from 0 to 1. It does not consider the magnitude of errors and is therefore used in conjunction with MAE or RMSE to provide a more comprehensive evaluation of models' performances.

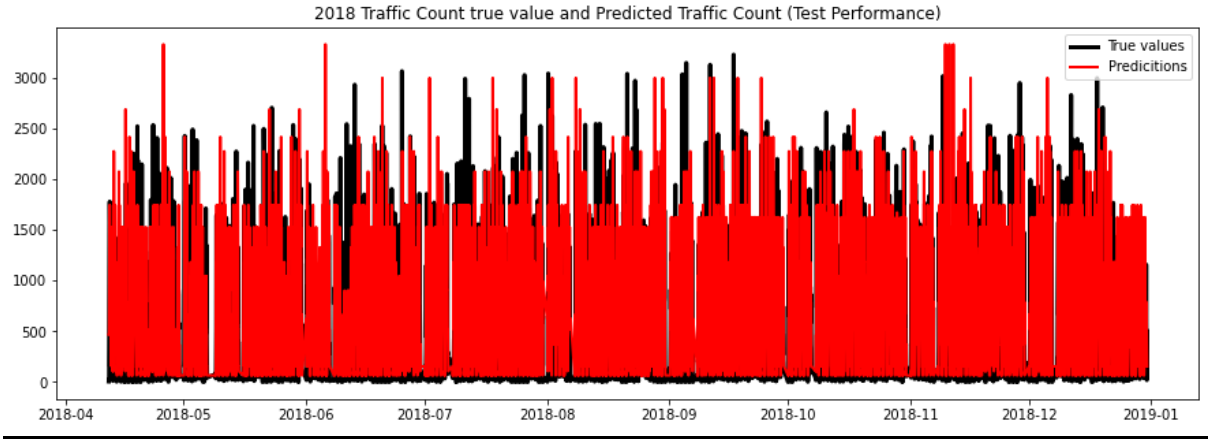
Hyperparameter tuning using GridSearchCV was used, but it didn't improve the models significantly.

The models were then tested on unseen data. The models were used to predict the traffic count across 2018 and were compared with the actual traffic count.

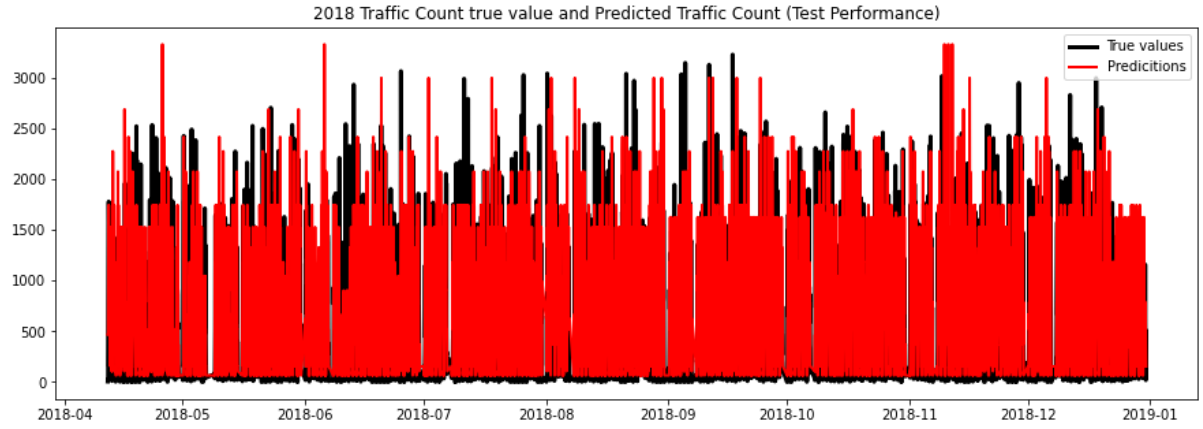
Linear Regression 2018:



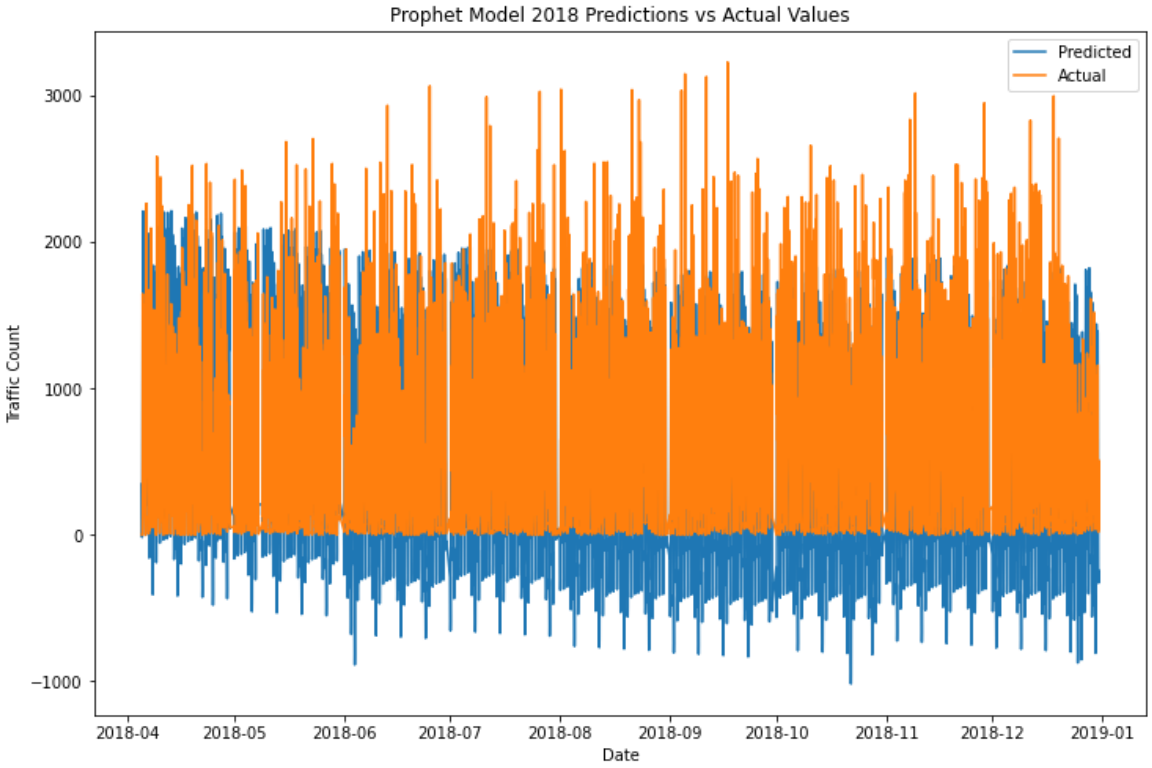
Decision Tree Regression 2018:



Random Forest Regression 2018:



Prophet Model 2018:



2018 Data				
Metric	Linear Regression	Decision Trees	Random Forest	Prophet
MAE	356.28	371.96	364.95	475.63
RMSE	506.37	548.80	535.96	606.54
R2	0.30	0.17	0.21	0.01

Table 2: Model Performance 2018 data

The top two performers were the Linear Regression and Random Forest models. Interestingly, linear regression performed better than the other models in the 2018 data compared to the initial performance on 2017 data. This could be due to the more complex models overfitting the nuances of the 2017 data and failed to generalise well to the 2018 data.

It is clear to see from Figure 14, that the 2018 data is irregular. This is a major limitation to the data used. Since it is real world data, the data between years may have completely different distributions, which in turn will make predicting very hard.

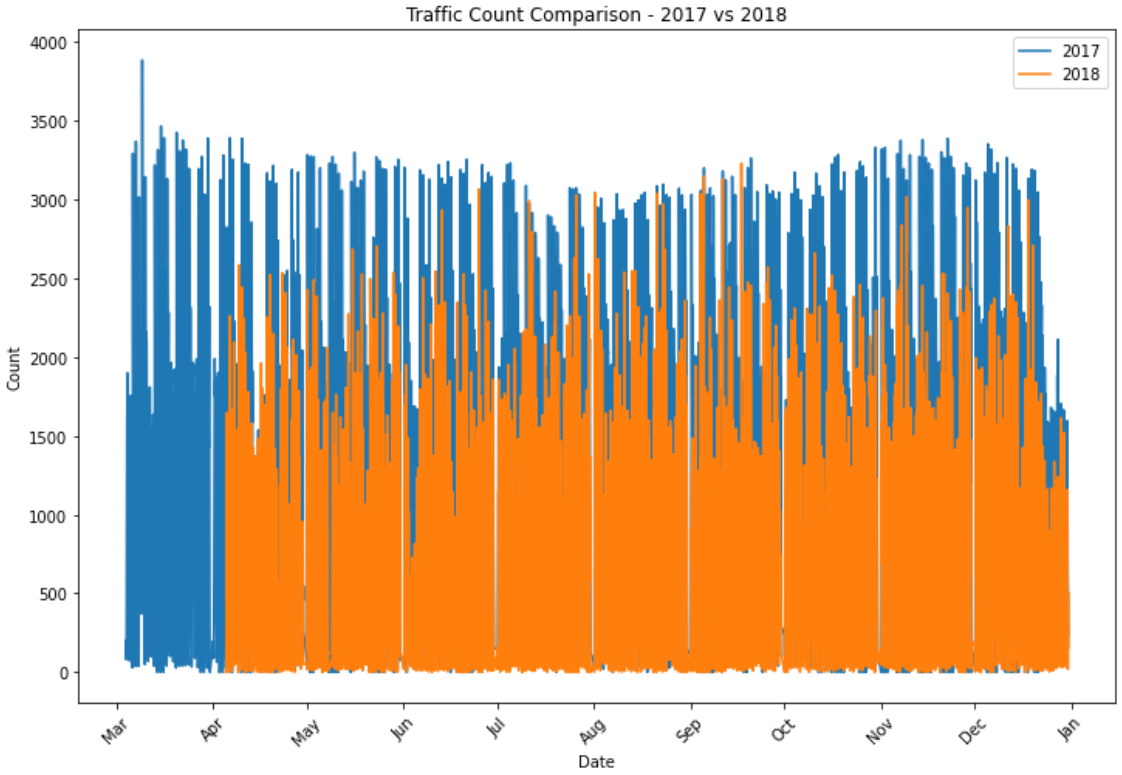


Figure 14: Traffic Count Comparison 2017 vs 2018

Outcomes

The EDA and feature engineering processes led to the identification of correlated features. Multiple models were evaluated and compared. The random forest regression proved to be the best model, outperforming all other models in each evaluation metric. The models were then tested on unseen 2018 data, and linear regression and random forest performed the best. However, 2018 has a huge limitation as it is irregular data, therefore making the predictions difficult. Further work needs to be done to ensure that the data used is comparable and consistent.

Further Work

There are various additional implementations that can be made to improve the accuracy and effectiveness of traffic count predictions. Firstly, improvements to the datasets used need to be made. The data obtained between years is inconsistent; therefore, the models are more sensitive to specific nuances. Having regular and relatively consistent data is key to making accurate predictions. Incorporating real-time traffic information could provide valuable insights into the dynamics of traffic patterns. These features can be obtained from the NZTA web API. Furthermore, additional modelling techniques can be used. Using RNN (LSTM) models or ARIMA models may yield performance increases. Finally, considering different spatial and temporal scales, such as analysing traffic patterns at the neighbourhood or city level, could lead to more granular and localised predictions, allowing for targeted traffic management strategies. Overall, by expanding the scope of features, refining modelling

techniques, and incorporating contextual information, future work can contribute to more accurate and robust traffic flow models.

Data answer

The dataset had enough features and volume of data to provide relatively accurate traffic count predictions. However, further improvements need to be made. Currently, there are limitations on the data that can be used. Ideally, several years of historical traffic count data would be used. This project only uses 2017 data and predicts 2018 data based on that. Attempts at using more historical data were made, but they resulted in problems that could not be resolved within the given time frame of the project scope. Regardless, models were still made that have a confidence level of 80%.

Business answer

The business question was answered satisfactorily. The proposed solution provides 80% accuracy to predict traffic count.

Response to stakeholders

The project is promising and expected to achieve its objective in predicting traffic count. It is recommended to proceed with attempting to add additional historical data before implementing the project.

End-to-end solution

To implement the end-to-end solution the organisation needs to have the necessary infrastructure for data acquisition and storage, systems to run the model and display the traffic count.

Deploying a predictive model in real-world scenarios requires careful planning and consideration of various factors. In this section, we outline an implementation plan for integrating the developed predictive model into existing transportation systems and provide recommendations to ensure successful deployment.

Implementation

The considerations for implementing the solution will be to develop systems for data acquisition, storage, and programme execution. Detailed below is a step-by-step plan on how to successfully implement this project:

1. Data Acquisition
    a. Relevant data sources is required for the prediction model. This should include historical traffic count data, weather data and any other potential information. The data must be comprehensive, regular, accurate and cover a

significant timeframe to capture the different traffic patterns. Data acquisition can be done using the NZTA traffic count API, which gets real time data of traffic counts at various state highways.

2. Data Pre-processing
    a. Clean and pre-process the collected data. Handle missing values, outliers, and inconsistencies. Feature engineering should be done to extract meaningful predictors and additional features if necessary.
3. Model Selection
    a. Develop and train the selected models.
4. Model Evaluation
    a. Evaluate all models and compare them with each other. Evaluation metrics can be used that were discussed previously.

Note that steps 1–4 can be done using the code I have already made for this project. The appropriate data must be collected, but all other steps will be relatively the same as what is already done.

5. Deployment:
    a. Integrate the selected model into an interface that users can easily interact with. This will provide users with the capability to see traffic flow predictions and enter in specific dates and times they want to predict traffic. The system should be scalable, efficient and capable of handling large amounts of data.
6. Monitoring
    a. Monitor the models performances regularly so it can be as accurate as possible. This will help prevent any degradation issues. The models should periodically be updated with new data too.

References

https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::tms-traffic-quarter-hourly-jan-2013-to-sept-2020/about

https://www.holidays-info.com/new-zealand/calendar/auckland/2017/

https://facebook.github.io/prophet/

https://www.transport.govt.nz/assets/Uploads/Report/TransportOutlookFutureOverview.pdf

https://matplotlib.org/

https://seaborn.pydata.org/