CAPSTONE PROJECT

TRAFFIC FLOW PREDICTION

Esther Dantra

April 2023

Institute of Data

CONTENTS

INTRODUCTION

Traffic congestion is a major issue for cities around the world, and Auckland, New Zealand is no exception. The city is facing significant economic costs due to congestion, with estimates suggesting that it is costing $1.3billion annually. The cost is due to a variety of factors, including lost productivity.

In addition to lost productivity, congestion causes significant delays for freight movements which can impact the competitiveness of local businesses. Vehicles stuck in traffic burn more fuel than they would if they were moving at a normal speed. This increased fuel consumption is also a huge cost to freight and logistics companies. In addition, the wear and tear on vehicles from constant stop-and-go driving can increase maintenance cost, which is another cost for businesses and commuters alike.

Beyond these economic costs, traffic congestion is also impacting quality of life for Auckland commuters. Long commute times can lead to stress and reduced quality time with family and friends. It can also impact access to emergency services. To address these issues, Waka Kotahi needs to accurately predict traffic counts at various sites, to optimize traffic flow. By analysing traffic patterns over time, the system can identify areas that experience high traffic volume during certain periods of the day. Based on this information the system can adjust signal timings accordingly to ensure efficient traffic flow.

1.1 Business Problem

Waka Kotahi would like to investigate the traffic patterns at the Lincoln Road Interchange travelling East bound and build a traffic prediction model which predicts the traffic flow at various times of the day. The predictions are then classified as Heavy or peak hour times and off-peak times. The output from this traffic prediction model will be used as an input for the traffic signal cycling system which will optimize merging at on ramps and alleviate congestion. The stakeholders in this problem are Waka Kotahi, freight and logistics companies as well as everyday commuters. City planners can also benefit from this project as it helps them look at the current infrastructure and make decisions based on whether the infrastructure meets demand.

1.2 Data Question

- Predict traffic flow at the Lincoln Road East bound interchange
- Classify traffic flow at various time periods at Lincoln Road interchange into three levels of traffic volume, heavy medium and light.
- Create a metric to evaluate the efficiency of our model when implemented.

2. DATA OVERVIEW

2.1 Data Source

The dataset is from Waka Kotahi New Zealand Transport Agency Traffic Monitoring Sites(TMS) data. Road sensors record number of cars for every 15-minute interval and this is recorded in what is called the TMS software database. The data is then uploaded to the TMS internet site from where the dataset used for this project has been sourced.

The dataset had traffic counts for quarter-hourly times ranging from January 2013 to September 2020 for approximately 2000 sites across highways in New Zealand. This data was merged with the highway monitoring sites dataset which had the Annual Average Daily Traffic count for each site for the past five years, location information, percentage of Heavy vehicles travelling through this site and other identification information which was deemed unnecessary.

The data was then enriched with holiday information by webscraping https://publicholidays.co.nz/ to include whether the traffic count was a public holiday, school holiday or school day including also whether it was a weekend. To this data, the weather data was also included. The maximum temperature information for the day as well as whether it was a clear or a rainy or an overcast day was included.

2.2 Target Variable

Target variable is traffic count. The initial dataset had traffic count per 15 minute intervals and further classified into vehicle types. For the purpose of this project, the quarter-hourly traffic count has been totalled to an hourly count.

2.3 Predictor Variables

'siteRef': Reference id of the site.

'AADT5yearsAgo': Average Annual Daily Traffic count 5 years ago at this site.

'AADT4yearsAgo': Average Annual Daily Traffic count 4 years ago at this site.

'AADT3yearsAgo': Average Annual Daily Traffic count 3 years ago at this site.

'AADT2yearsAgo': Average Annual Daily Traffic count 2 years ago at this site.

'AADT1yearAgo': Average Annual Daily Traffic count 1 years ago at this site.

'Hol_type': The type of day it is, SchoolDay, SchoolHoliday, PublicHoliday or Weekend.

'day': The number of the date that day.

'month': The number of the month.

'hour': The hour of day.

'dayOfweek': The day of week in numbers.

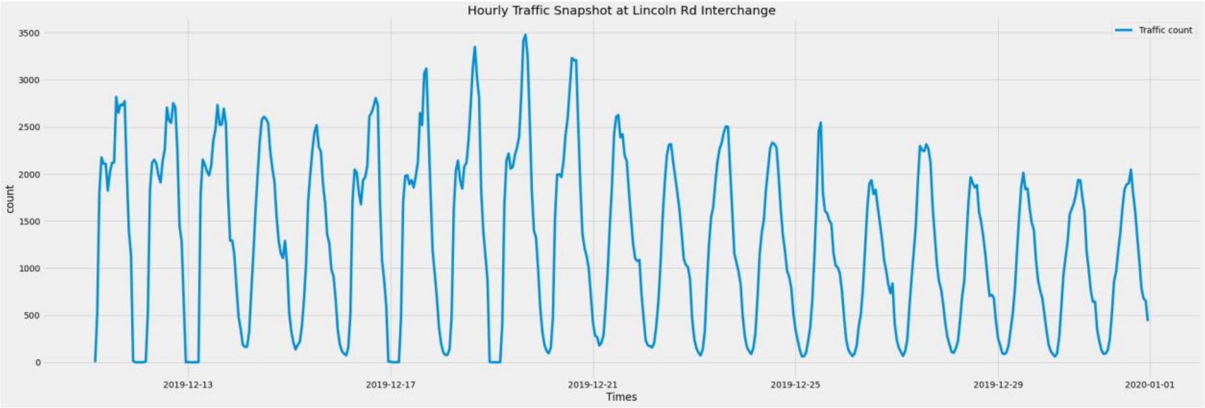
'Conditions': This was a categorical variable with six categories ranging from clear to rainy and overcast.

'lane': Categorical variable which described whether the recording was at an onramp or going eastbound or going westbound.

'Maximum Temperature': The maximum temperature in Auckland that day.

3. EXPLORATORY DATA ANALYSIS

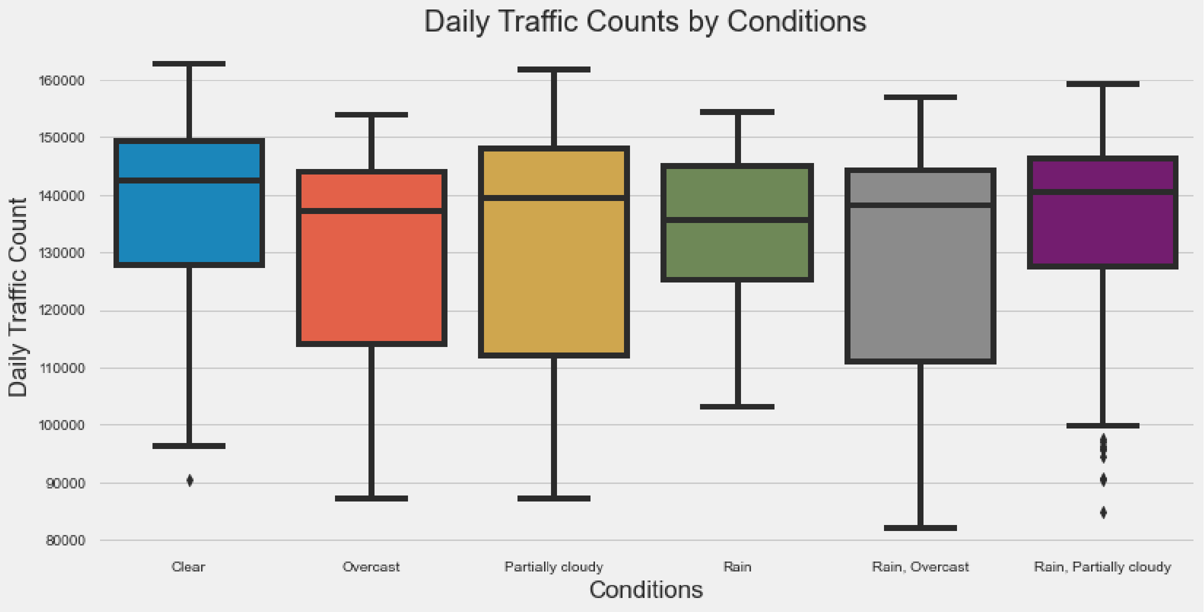
3.1 Initial Observations



To get a clear idea of the data, we began by zooming into a month’s data of traffic flow arbitrarily choosing 11th dec 2019 to 1st Jan 2020. We noticed that traffic flow has a cyclic pattern.

3.2 Traffic counts by Weather

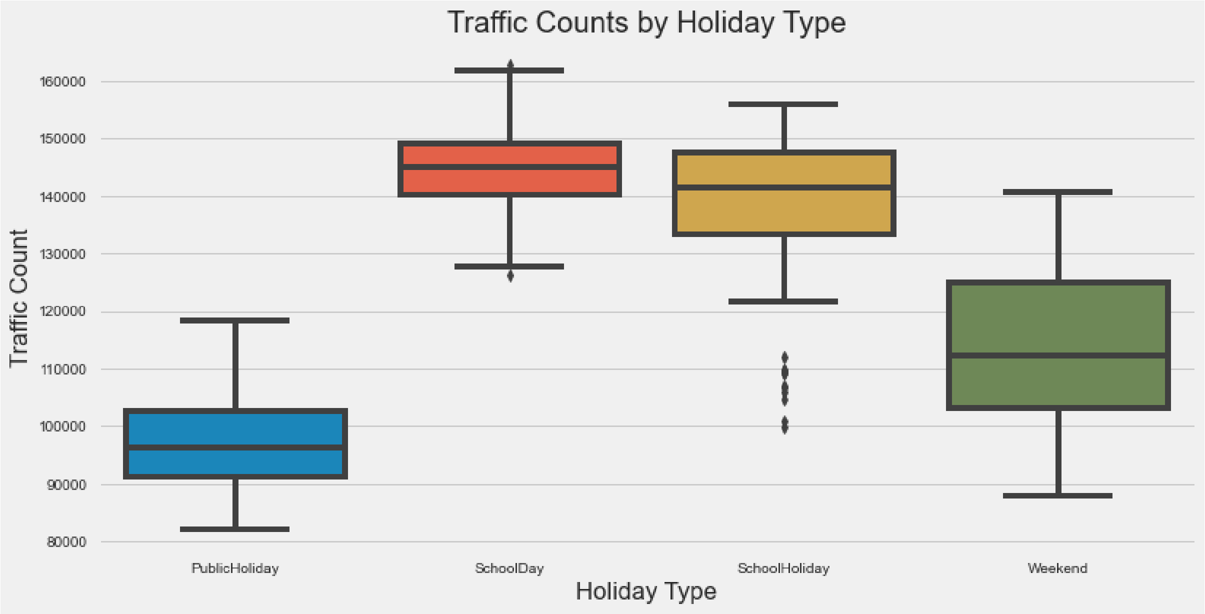
Weather is an important factor in traffic flow prediction because it can significantly impact driving conditions causing traffic congestion and delays. Rain can lead to decrease in visibility leading to difficult driving conditions which can decrease traffic speed and cause congestion. However, the commonly held premise that traffic flow is heavier during rainy and overcast conditions turned out to be incorrect at this particular site. The boxplot below shows the traffic count on clear days was on average higher than compared to rainy or overcast conditions. This is perhaps because people are more likely to put off travel during rainy conditions.



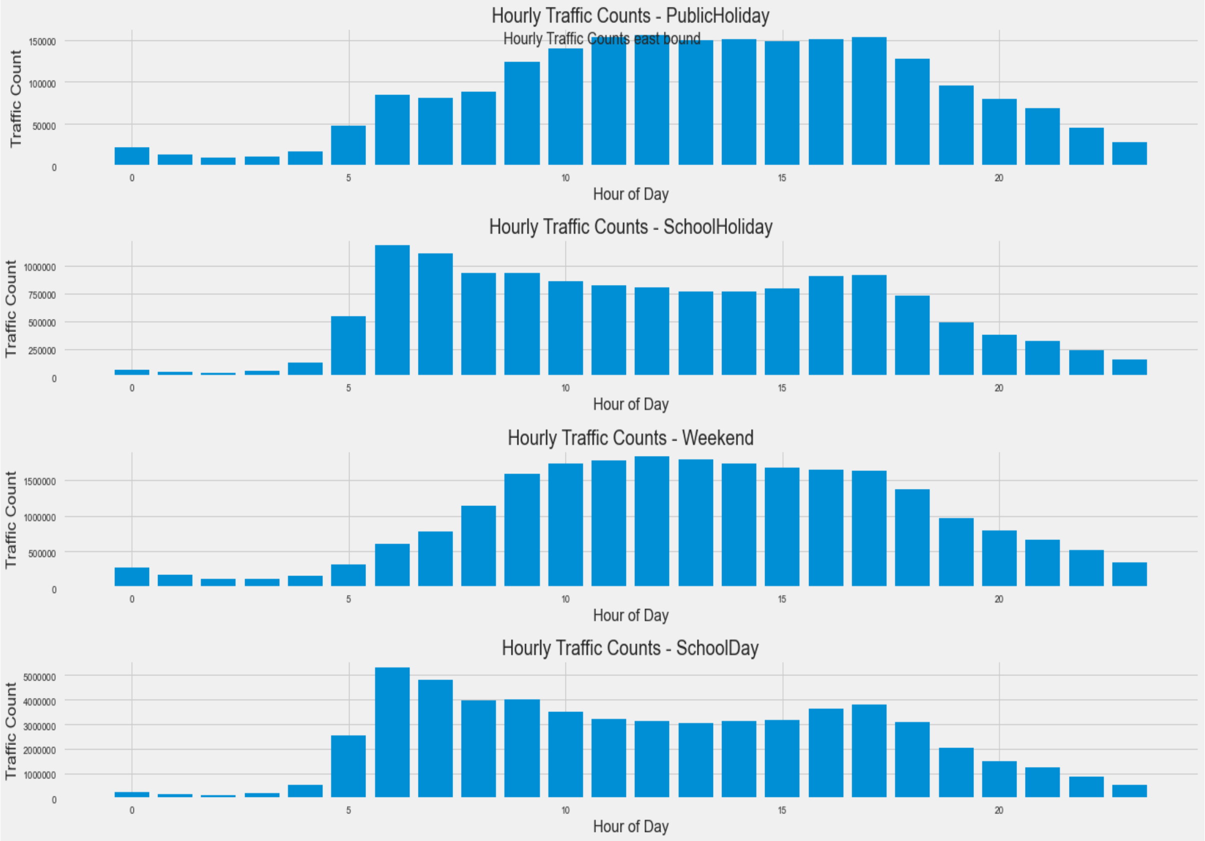
3.3 Traffic counts by type of day

Holidays and school days are important factors that can affect traffic flow. During holidays, traffic patterns may differ from regular weekdays due to changes in commuting habits, travel patterns and business operations. In addition, school days can also have a significant impact on traffic, as parents tend to drop off and pick up their children and school during specific hours, causing increased traffic congestion. Overall, understanding the impact of holidays and school days on traffic flow is crucial to predicting and managing traffic congestion.

In the boxplot below we noticed that traffic count is significantly lower on average on public days while there isn't a big difference between traffic counts on school days compared to school holidays. This could be attributed to the fact that the monitoring site for this analysis is on a motorway and motorway traffic is more dependent on business holidays than school holidays. Weekend traffic is also significantly lower than school day or school holiday traffic. However, it is greater than traffic on public holidays.



3.4 Traffic counts by Hour of day



To dive deeper into the traffic flow patterns, we looked at the traffic counts broken down by the hour of day for different types of day. The key observation was that peak hour traffic at this site was 6am-8am on school days and school holidays. For the rest of the day traffic is at a consistent flow 7pm when it begins to drop away significantly. It is important to note that this traffic is east bound travelling towards the CBD and therefore the peak is in the early hours of the day and no heavy traffic is observed in the evening peak hour times. The traffic flows for public holidays and weekends is normally distributed with peak traffic flow occurring at the midday. There is consistent traffic flow from 10am to 6pm on these days.

4. MODELLING

4.1 Imputing missing values

Missing values were present in the dataset possibly due to the section of the motorway being closed for maintenance. It is essential to fill in missing values in order to avoid biases and inaccuracies in the analysis of the data. To preserve the integrity of the data and to make sure that the data is representative of the actual traffic flow patterns, the average traffic flow for each day of the week, holiday type and hour of the day was calculated and imputed into missing rows.

4.2 Encoding of Categorical variables

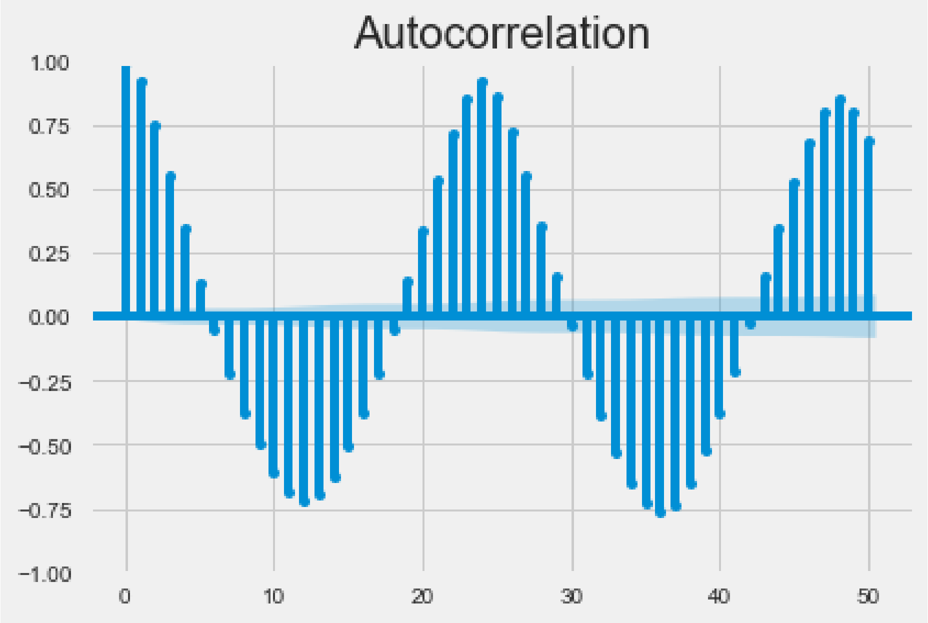
The categorical variables in this dataset were conditions of the day, type of day and lane type. The categorical variables were given one hot encoding.

4.3 Training and Test sets

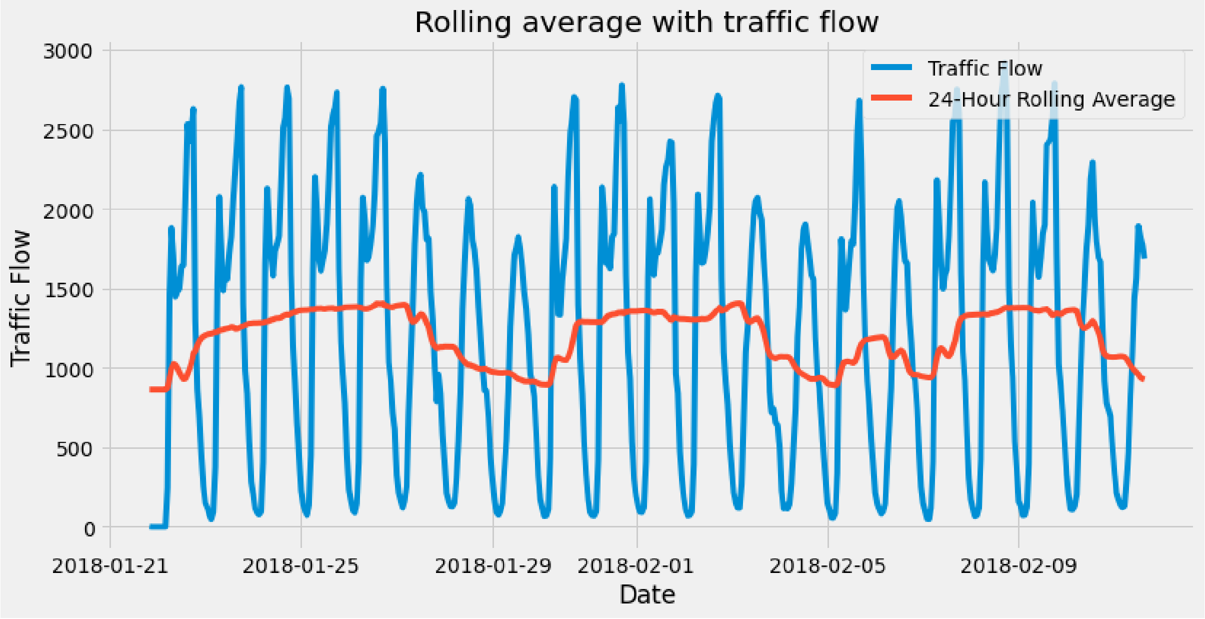
As indicated earlier the dataset had traffic counts from January 2013 to September 2020. Owing to computational constraints, only data between January 2018 and January 2020 was chosen. This was done to choose the data which was closest to current dates while avoiding COVID lockdown times.

4.4 Time series analysis

The Auto-correlation curve draw for the first 50 time intervals which is a little over two days shows that our data has significant periodicity. The Augmented Dickey fuller test returned a p-value of less than 0.05 suggesting that this dataset is stationary. This was expected because the traffic counts were from a two-year time period. We do not expect traffic counts to have an increasing trend over the course of two years. However, if we were to include more historical data perhaps going back at least five years, we will probably notice an increasing trend. So the resulting model can only be used to predict traffic counts in the short term.



To identify underlying trends, the daily rolling average was plotted and this showed a smoother curve removing the noise and short-term fluctuations. It confirmed the ADF test that the data is stationary and it is cyclical with a 24 hour period.

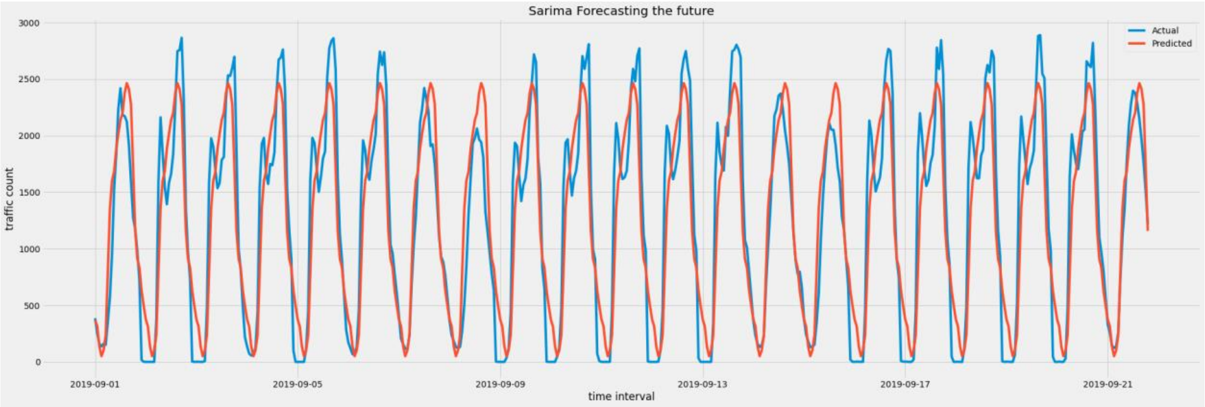


4.5 Time Series Model – Sarimax

Seasonal AutoRegressive Integrated Moving Average with eXogenous factors commonly know as SARIMAX seemed like the logical choice to predict traffic flow. This model analyses and forecasts time series data, especially when there is seasonality or other complex patterns in the data. When SARIMAX was applied with the features, the model did not converge. So, SARIMAX was applied without any features. In the absence of exogenous variables, the model is essentially SARIMA. This model captures the patterns and seasonality in the data such as daily, weekly or hourly fluctuations in the traffic. However, in the absence of exogenous variables such as holidays or weather, the model is not able to incorporate those effects and thereby not as accurate as we need it to be.

To get a better understanding of how the model performed, plotting the predictions along with the actual data while also looking at metrics like mean squared error, mean absolute error and R^2 score is the best method for Time series data as this gives us a better picture of how the model is performing.

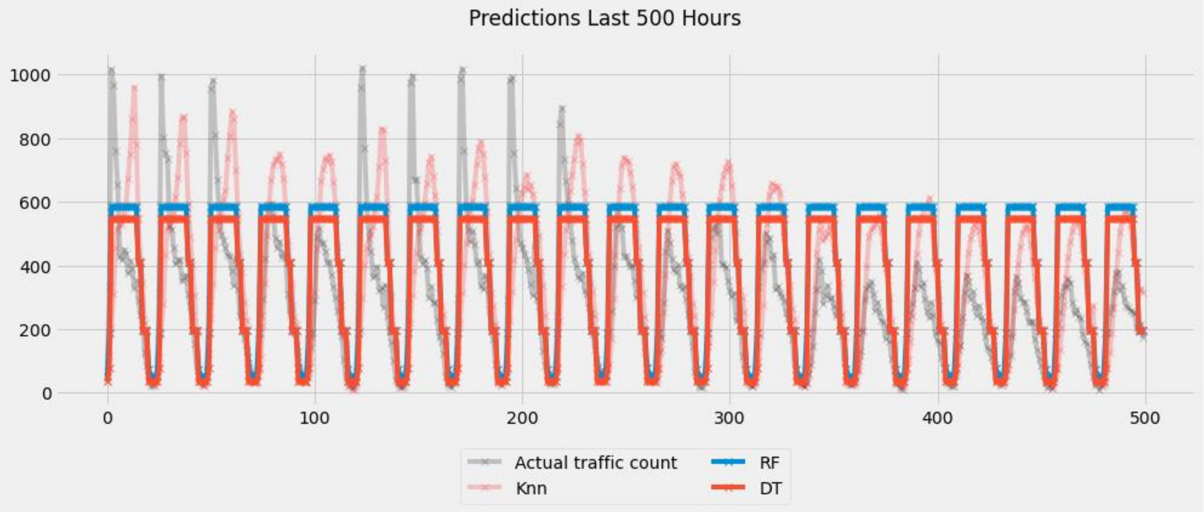
With this model, we noticed that Sarimax did capture the cyclic patterns quite well and for intuitive feel about the fit, an R^2 score 0.82 is in the right direction.



Model	Mean Squared Error	Mean Absolute Error	R2 score
Sarimax	163009	309	0.82

4.6 Regression Models

In predicting traffic flow, various models were tested to find most accurate and suitable for the problem. The first model applied was a linear regression model. Linear regression is a simple and fast model to apply for predictive modelling. Its interpretation is straightforward and it works well with continuous variables, making it a good fit for traffic flow prediction. The second model tested was the random forest. Random forest is a popular ensemble learning algorithm that works well with both categorical and continuous data. It is known to be less sensitive to outlier and can handle interactions between features making it a strong candidate for traffic flow prediction. The third model was a decision tree. Decision trees are simple to understand and can easily visualise and interpret the results They also work well with both categorical and continuous data. The model splits the data based on the most significant features, which can help identify the most influential factors affection traffic flow prediction. Finally, k-nearest neighbours(K-NN) was applied. The model works by identifying the k-closest points in the training data to the test data point and using their average to predict the target value.

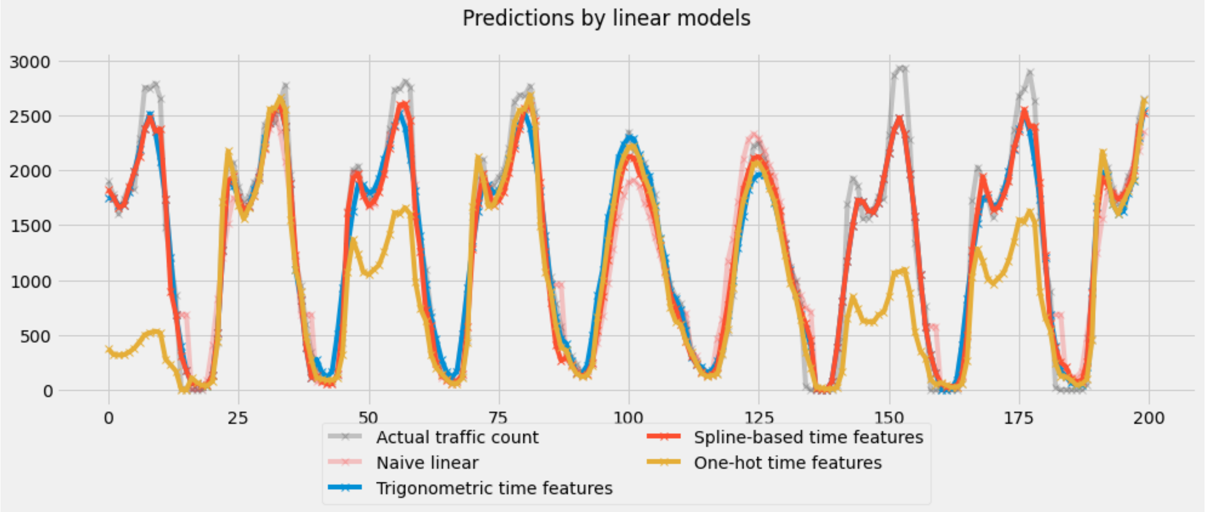


Model	Mean Squared Error	Mean Absolute Error	R2 Score
Linear Regression	115980	281	0.32
Random Forest Regression	68344	200	0.60
Decision Tree Regression	67700	195	0.61
K-NN Regression	65140	180	0.62

Based on the model performances, the K-NN regression model seems to have the lowest Mean Squared Error and Mean Absolute Error, indicating better predictive accuracy than the other models The R^2 score for all models is moderate although the K-NN model has the highest at 0.62. So at this stage it appears that K-NN has better predictive power than Decision Tree or Random Forest.

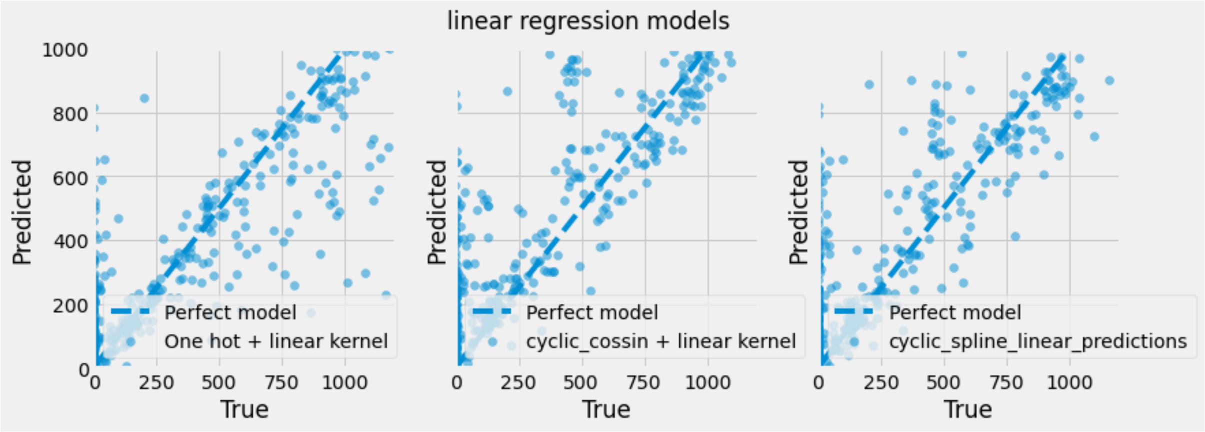
4.7. Generalised Additive Models

To further improve on the K-NN regression model which performed the best in the previous iteration of modelling, a series of Generalised Additive Models were developed. The first model assumed the categorical columns were ordinal and transformed them to numerical data based on this order. Similar encoding was applied to such as day of the week, month and hour. The transformed data was then scaled and applied to a K-NN regression Model. To capture the periodicity and phase of the time series, the time features were transformed using sin and cosine transformations, the transformed data is then scaled and applied to a K-NN regression model. The next approach was to apply a spline based transformation to the time features instead of a trigonometric transformation and then scaled before being applied to a K-NN regression model.



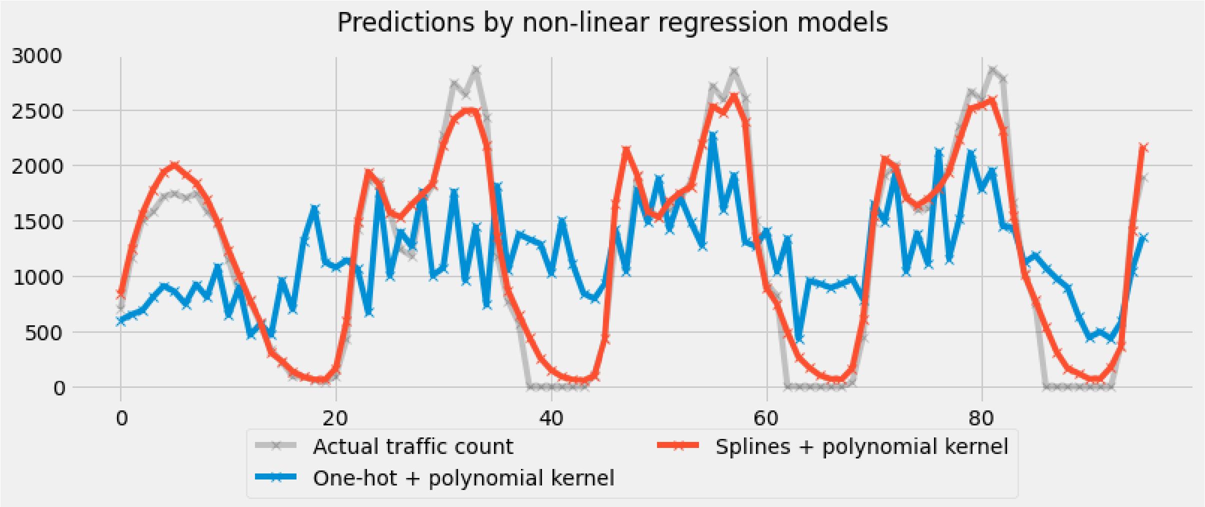
Model	Mean squared error	Mean Absolute Error	R2 Score
GAM 1	20768	258	0.76
GAM 2	44627	153	0.94
GAM 3	34233	129	0.96

In these three models the first model performed best with respect to the mean squared error and the second model has a higher mean squared error and a high R^2 suggesting that this model may have overfit the data. The third model seemed to perform best with low mean absolute error and a relatively low mean squared error but a high R^2.



On inspection of the actual vs predicted graphs, we can see the third model performs reasonably well. However, there is considerable variance across the board.

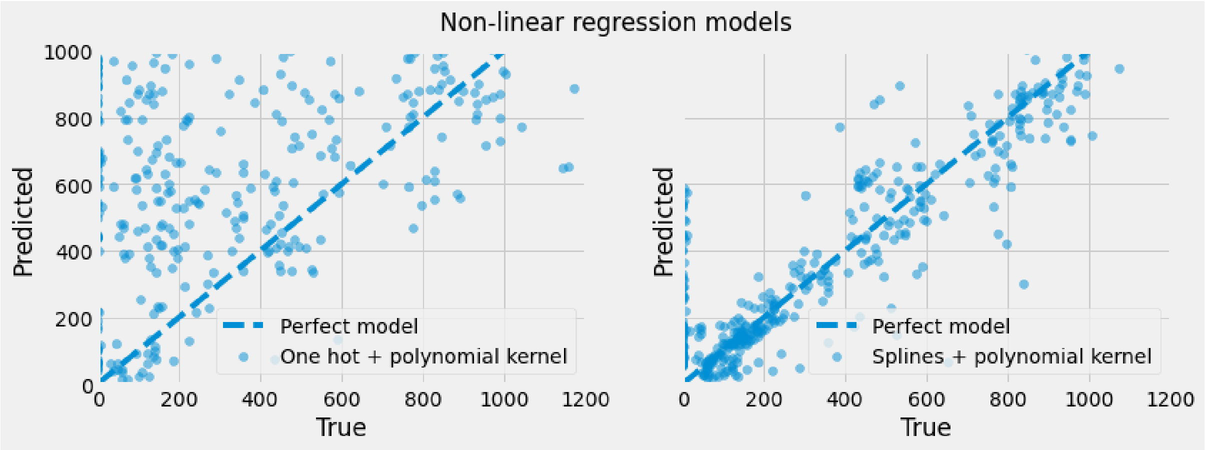
In the fourth model, the Nystroem Kernel Approximation allows for efficient approximation of high dimensional feature spaces. So this approximation was applied after encoding time features and then the transformed data is applied to a K-NN model. Finally, the Nystroem Kernel Approximation was also applied after time features were transformed using spline transformations. Then the data was applied to a K-NN regression model.



Model	Mean squared error	Mean Absolute Error	R2 Score
GAM 4	404723	483	0.52
GAM 5	27376	115	0.97

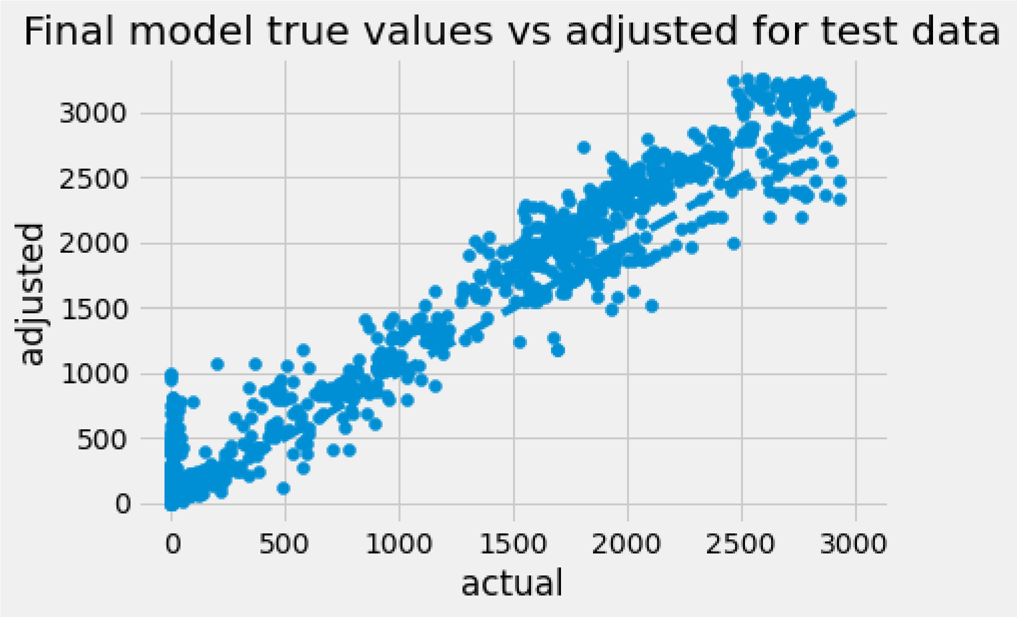
The fourth model had a very high mean squared error and a low R^2 indicating that it did not perform well in terms of predicting the traffic flow values. It also had a relatively high mean absolute error, suggesting that its predictions were quite far from the actual values. In contrast, the fifth model had a low mean squared error and a high R^2 score, indicating that it performed well in terms of predicting the traffic flow values and capturing the underlying patterns in the data. It also had a low mean absolute error, suggesting that its predictions were quire close to the actual values. So at this stage, this model performed the best and showed high prediction power.

When we looked at the true values vs the predicted values, it shows greater variance in some areas and high predictability at low traffic count regions.



4.8 Final Model

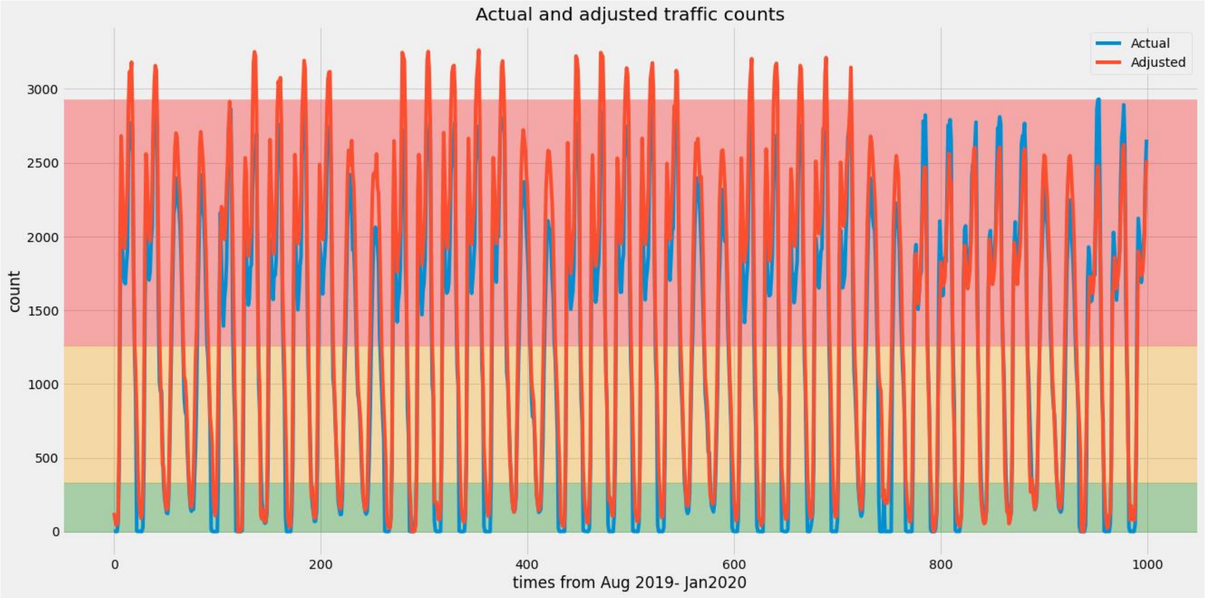
To improve on the model further, the error were investigated further and found the times when the model was predicting lower than the actual traffic. This was found to be primarily during peak hours. So, a correcting factor was applied so during those times predicted count is scaled up by 15%. This gave rise to the below true vs adjusted values plot.



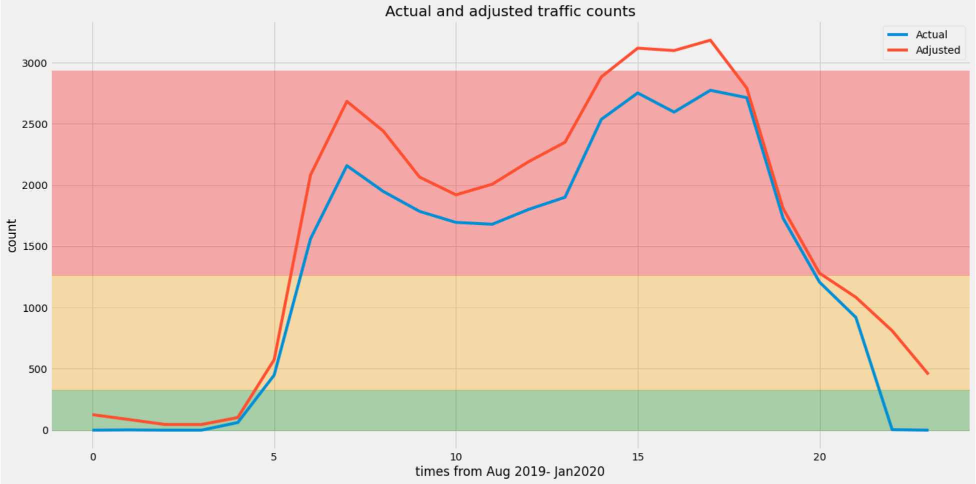
This plot shows good predictability with expected variance across the board.

5. IMPLEMENTATION

To prevent gridlock, the prediction from the traffic model indicates at what times high volumes of traffic is expected, medium volumes and low volumes. This was determined by looking the mean of the predicted values. Any point the traffic volume is predicted to be greater than the overall mean of the predicted values, it is considered to be heavy.



Looking at our models closely, we can see that traffic is predicted to be a high volume from 6am to 8pm while average traffic is expected between 4am to 6am and between 8pm to 12am. So the traffic signals are set to high cycle during the heavy times, medium cycle during the medium times and low cycle to during the low times. During high cycle, the on ramp lights cycle a little slower to allow traffic on the motorway to flow without congestion while during moderate traffic times the lights cycle faster and during low times, the lights are turned off.



6. MODEL EFFICIENCY RATE

The model is considered to be efficient if it predicts the traffic to be heavy or greater than the actual. For example, if the actual traffic is moderate but the prediction suggests it is heavy, the model is still considered efficient. Based on this definition, the number of observations in the test set that meet the efficiency mark were divided by the total observations in the test set and then converted to a percentage. The model is considered to be efficient close to 90% of the time. The 10% inefficiency is possibly from the times when an incident has occurred causing increased traffic flow.

7. BUSINESS SOLUTION

With a high efficiency rate, the final model is a good input for on ramp lights at the Lincoln Road Interchange to determine at what length of the time the lights need to cycle for. This model can be trained for other junctions also and used as input for different traffic junctions. Use of this model as input for various traffic light systems can alleviate traffic congestion and bring down the estimated loss of $1.3 billion dollars to the Auckland economy.

8. CONCLUSION

The project's objective was to address traffic congestion issues at Lincoln Road East Bound interchange through the development of an accurate traffic flow prediction model. To achieve this goal, various models were tested beginning with the SARIMAX model. After rigorous evaluation, the final model selected was a Generalised Additive model that involved data transformation, polynomial kernel approximation and K-NN application. The developed model achieved an efficiency level of 90%.

The next step is to compare the performance of this model against other sites and explore the inclusion of incident data to further enhance its accuracy. The addition of incident data has the potential to significantly improve the model's predictive ability. Overall, the project has resulted in the development of an efficient model that can help mitigate traffic congestion at the Lincoln Road East Bound Interchange, and its potential applications can be extended to other sites.

9. RESOURCES AND LIBRARIES USED

Python

numpy

pandas

sklearn

matplotlib

datetime

statsmodels

itertools

10. REFERENCES

https://www.stuff.co.nz/business/95383973/traffic-congestion-costs-aucklands-economy-13b-a-year-report

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

https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::state-highway-traffic-monitoring-sites/about