Auckland ICT Graduate School

Internship Final Report

22-02-2024

Forecasting Air Quality based on Traffic and Weather Data

Author: Yi Yang
Student ID: 802877420
Academic Supervisor: Philipp Skavantzos
Company: Auckland ICT Graduate School
Industry Mentor: Sean Zeng

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Yi Yang

***Abstract*—This report provides an overview of a ten-week internship project aimed at forecasting air quality using traffic and weather data, conducted at the Auckland ICT Graduate School. By employing machine learning and data analysis, this project analyzes patterns in air quality, traffic congestion, and weather data to predict air quality indices. With a focus on neural networks and other predictive models, the project seeks to enhance our understanding of the environmental factors impacting air quality and propose ways to lessen negative impacts.**

***Keywords*—Air Quality Forecasting, Traffic Data, Weather Data, Machine Learning, Data Analysis, Neural Networks, LSTM**

I. Introduction

This report documents a ten-week internship project, the Green Future Sustainable Development Challenge, an ambitious initiative aimed at creating a sustainable future through technology. It emphasizes the need to address numerous challenges, such as just transition, effective resource allocation, sustainable food production, waste management, biodiversity loss, and climate change, to ensure a sustainable development path. The pressing global challenge of reducing emissions is highlighted, with the impact of climate change being devastating. The future's sustainability will depend on corporate responses, where growth must not come at the expense of the planet and humanity. This report details the challenge of "Forecasting Air Quality Based on Traffic and Weather Data," utilizing machine learning and data analysis of air quality, traffic congestion, and weather-related data to identify patterns and predict air quality indices. Predicting climate and air quality is complex, with many existing models being singular in their approach. However, emissions from vehicles and weather events directly affect air quality, offering a basis for predictions. By collecting extensive data on traffic congestion and weather events, studying them holistically, and identifying correlations, AI and machine learning are used to analyze the relationships between these data sets, enabling air quality predictions. These findings aim to improve air quality forecasting.

The remainder of the report is structured as follows: The ICT background of the internship is discussed first, followed by a review of existing and new technologies in the second part. The third part discusses the project background, followed by research and analysis of the problem in the fourth part. The fifth part details the project's target audience, the sixth part discusses the design and facilities, and the seventh part presents the project outcomes. Finally, the report concludes with reflections in the eighth section and conclusions in the ninth section.

A. About the Company

The Auckland ICT Graduate School is a collaborative effort between the University of Auckland and the University of Waikato, designed to meet the demand for skilled ICT professionals. It emphasizes both academic excellence and strong industry connections, promoting a culture of innovation and hands-on learning. The school's internship program is designed to let students apply their theoretical knowledge in practical settings, enhancing their skills and professional development through industry-led projects.

The school's mission is threefold, aiming to:

1) Facilitate students' seamless transition into ICT careers.

2) Produce high-quality, industry-ready graduates with technical and business acumen.

3) Drive the development of New Zealand's technology sector by nurturing the architects of future innovation.

B. Work Environment

The internship began in the Unleash Space, the school's workplace for creativity and entrepreneurship, where we collaborated closely with other teams, promoting a strong sense of teamwork and knowledge exchange. Later, we moved to dedicated workspaces on the fourth floor of the science building, enhancing our focus and communication.

Our team, one of eleven, tackled the challenge of forecasting air quality using traffic and weather data, blending environmental science with data analysis. To ensure progress, we held various meetings, including:

1) Stand-up meetings: Twice weekly updates with our mentor Sean Zeng to discuss progress and receive guidance.

2) Presentations: Weekly short presentations for cross-team updates and a mid-term report showcasing our project and progress.

3) Academic meetings: Biweekly meetings with our supervisor, Philipp Skavantzos, for progress updates and advice.

C. Operational Structure

Our team consisted of four members: Alvin focused on data collection, processing, and analysis; Morgan and I on modeling; and Bella, our developer, integrated the model with the website's front-end and back-end.

II. Literature Review

Understanding air pollution's causes and gaining insight into the various methods and strategies to tackle air pollution is essential for effectively planning and carrying out this internship project.

A. Machine Learning and Related Technologies

At the core of air quality prediction are systems that construct and assess models, heavily relying on machine learning and data mining. Machine learning focuses on developing models that learn and make predictions from data, a subset of AI that enables computers to learn from data and make decisions without explicit programming. Key to machine learning is developing algorithms that process vast amounts of data to learn patterns and features for predictions or classifications. It is extensively applied in areas like recommendation systems, speech recognition, image recognition, and medical diagnostics[1]. Scikit-learn, a Python module integrating advanced machine learning algorithms, is utilized in this project for modeling, emphasizing ease of use, performance, and documentation[2].

Machine learning tasks are categorized into supervised and unsupervised learning[1]. Supervised learning models learn from labeled training data, predicting outputs from given inputs, including regression and classification problems like

predicting house prices or identifying spam emails. Unsupervised learning, however, learns from unlabeled data to discover data structures, including clustering similar data points or reducing data dimensions with Principal Component Analysis (PCA).

Data mining extracts valuable information or knowledge from large datasets, involving statistical analysis, machine learning, pattern recognition, and database technologies to discover data patterns, trends, and associations, applied in market analysis, fraud detection, and customer relationship management[3].

B. Existing Technologies

The existing mainstream models in the field of air quality prediction include Neural Network (NN) models, Radial Basis Functions, Support Vector Machines, and Feedforward Neural Network models[4].

Research employs various methods, such as using machines to predict air quality time series, mostly utilizing multilayer neural networks instead of models specifically built for time series[4]. Deep learning, a broader branch of machine learning involving multiple processing layers, has recently designed advanced Artificial Neural Networks (ANN), such as Recurrent Neural Networks (RNN) for air quality prediction. ANNs have special features that enhance their learning, training, and prediction capabilities[4]. For example, the study by Pasero and Mesin shows that despite ANNs' reliance on meteorological factors and issues with data overfitting, it remains a valuable method in air quality prediction[5].

Lee et al. demonstrated how to use gradient boosting machine learning methods to accurately predict PM2.5 concentration in Taiwan[6]. This method, by analyzing historical PM2.5 data and related meteorological information, can predict air quality for the next 24 hours. Zhang et al. [7] proposed a prediction model combining Long Short-Term Memory networks (LSTM) and Graph Attention (GAT) mechanisms, which showed clear advantages in accuracy over baseline models in tests on public datasets. Hasnain et al. [8] collected data from air quality monitoring stations and predicted six air pollutants based on the Prophet Forecasting Model (PFM), comparing their results with the predicted and actual values using correlation coefficient (R), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

III. Project Background

With the arrival of spring, flowers bloom, and pollen spreads everywhere, causing thousands to suffer from pollen allergies and allergic rhinitis. Known for its stunning landscapes, New Zealand attracts many tourists. Before heading out, it is common to check the weather, traffic congestion, and air quality to determine if it is suitable for travel. Considering this, our team thought of an idea: could we provide timely reminders for people to travel by predicting future air quality, like weather forecasts?

This issue is critical to me. In China, smog is a common phenomenon that severely affects people's health. I would like to know if I will face the same air pollution problems if I plan to go out tomorrow in New Zealand, especially in Auckland, where traffic congestion and air quality issues have gradually caught my attention. As a major metropolis in New Zealand, Auckland has seen a continuous increase in traffic flow in recent years, accompanied by a decline in air quality. I observed a significant improvement in Auckland's air quality during the implementation of winter traffic restrictions. However, once the restrictions were lifted, more people chose to drive, resulting in traffic congestion and air pollution returning to pre-restriction levels. Vehicle exhaust emissions not only cause traffic congestion but also negatively impact air quality.

Particulate matter (PM) includes solid and liquid particles in the air. Children, the elderly, and those with heart or lung problems are at higher risk of health issues due to exposure to particulate matter. These issues include decreased lung function, heart attacks, and mortality[9]. PM2.5, composed of extremely small particles with a diameter of less than 2.5 micrometers, is particularly harmful as it can linger in the deep lung and affect health. The main human sources of PM2.5 in New Zealand include domestic fires (e.g., wood and coal used for home heating) and motor vehicles. Through our study of air quality in Auckland, we found that PM values (particulate index) increased by 2-3% annually, exceeding the safety standards set by the World Health Organization[10]. Even in areas with relatively good air quality, such as the Queen's area, PM values exceeded the average by 28.7%. This phenomenon has begun to affect people's respiratory health.

For example, according to a study published in 2022[11] (HAPINZ 3.0 study), approximately 1,292 New Zealanders died prematurely in 2016 due to anthropogenic PM2.5 air pollution, and 4,626 were hospitalized for cardiovascular and respiratory diseases. This study highlights the significant impact of fine particles on health.

Therefore, our team focused on air quality prediction, collecting, and analyzing traffic, weather, and other data to build models for early warning of air quality changes. By providing timely and accurate air quality information, we are committed to creating healthier, sustainable urban living.

IV. Research & Discussion

To accomplish this task, we conducted research and divided the discussion into different levels. Alvin was responsible for data collection and processing analysis. We focused on three iconic cities in New Zealand: Auckland, Wellington, and Christchurch. By merging different datasets from each city, Alvin created a comprehensive dataset, which was then passed on to Morgan and me to build air quality prediction models. Our modeling approach produced two different versions, a demo version, and a professional version, creating separate models for the three cities to address their unique environments. This means a total of six models were developed to meet the specific needs and data of Auckland, Wellington, and Christchurch.

I was responsible for developing the three demo version models, one for each city. These models are designed with the public in mind, using an easy-to-understand set of variables such as temperature and rainfall. This approach ensures that anyone, regardless of their technical background, can easily

use our website to input these values and receive a tailored air quality forecast for any day they want in any of the three cities.

On the other hand, Morgan was responsible for creating the three pro version models. Suppose the introduction of the demo version is aimed at enhancing user experience, making our model understandable and usable by everyone. In that case, the Pro version selects all features for model training. This version involves more comprehensive data, increasing the model's complexity. We put these and all historical data on our website for researchers and government officials to use.

V. Project's Target Audience

My team provide valuable information to governments, the academic community, and the public.

For government and academic researchers, we use curated datasets to build diverse machine learning predictive models. This provides essential information and recommendations for government departments, like the Department of Environment and Health, Transportation Authority, and local councils in New Zealand, but they can also use our models to make decisions that help protect ecosystems and reduce environmental issues. For example, by integrating air quality, weather, and traffic data, this information can give the government a comprehensive understanding of the factors affecting air quality and support policymakers in developing effective policies related to air quality and environmental protection, promoting long-term environmental health. It also enables urban planners to make wise decisions regarding urban infrastructure, green spaces, and zoning, promoting environmentally sustainable and resilient urban development.

Additionally, we can provide case studies for research groups. By providing datasets and related literature reports, we hope to make a small contribution to the academic community for others to use.

Moreover, not just for governments and research groups, we also focus on serving the public. We have designed a website to host our models for public prediction, where people only need to enter simple information, such as the desired date, temperature, and traffic flow, to get the air quality index for the day. This way, the public can also use our models. At the same time, we will continue to upload well-organized data resources to the website and perform data visualization for everyone's reference.

VI. Project Implementation

A. Early Stage of the Project

In the early stages of the project, we focused on data collection. We first collected three types of data from recent years in New Zealand: traffic data (such as daily traffic flow number of large and small vehicles), weather data (such as rainfall, temperature, humidity, wind speed, wind direction, sunrise, and sunset times), and air quality data (such as PM2.5, PM10, AQI (Air Quality Index)).

In the first week, we emailed the government's data departments and environmental agencies to inquire about relevant data. At the same time, we searched for the required weather, traffic, and air quality data on websites. I also learned to use Git and VS Code for code management and version control, essential tools for our project's collaboration and data management. We found all the data within two days and attempted to merge them.

During the merging process, we encountered some problems: the data was messy, with mismatched dates and locations. Some data were from Auckland, others from Wellington or other places, and some even detailed down to every single road. It was clearly impractical for us to check tens of thousands of road data to determine their district. Therefore, we held a team discussion and decided first to lock down the data from Auckland, selecting and merging data from 2020 to 2022 into a daily format. I discovered that many datasets we found were missing and unsuitable, so I searched again for traffic data and found a complete set of traffic data on the opendata-nzta website.

After reorganizing the dataset, we successfully obtained a CSV file containing 1096 rows of data. The features included 'Timestamp', 'Heavy,' 'Light,' 'WDir.Deg.,' 'WSpd.m.s.,' 'GustDir.Deg.,' 'GustSpd.m.s.,' 'WindRun.Km.,' 'Rain.mm.,' 'Tdry.C.,' 'TWet.C.,' 'RH…,' 'Tmax.C.,' 'Tmin.C.,' 'Tgmin.C.,' 'ET10.C.,' 'ET20.C.,' 'ET100.C.,' 'Pmsl.hPa.,' 'Pstn.hPa.,' 'Sun.Hrs.,' 'Rad.MJ.m2.,' 'trafficCount,' 'Value_O_3 (μg/m³)_x,' 'Value_ SO_2 (μg/m³)_y,' 'Value_NO_x (μg/m³),' and we used 'PM10,' 'PM2.5,' 'Value (AQI)' as labels for prediction.

Fig. 1. First Organized Data Set

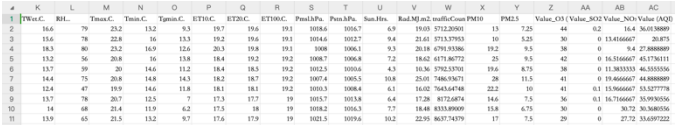


Fig. 2. First Organized Data Set

In the second and third weeks, based on the dataset organized in the first week, we preliminarily established a neural network model. We found that neural network models are most suited for time series data. For example, we learned that using artificial intelligence-based algorithms like CNNs

(Convolutional Neural Networks) to capture features is better than specifying them directly, making our model less singular. Therefore, we first embarked on this complete dataset, checking the correlation of features and tuning parameters before modeling with CNNs, eventually achieving a model with good performance.

Meanwhile, we had a meeting with our team supervisor, Philipp, who provided us with some essential guidance, such as needing to clarify what the team project's issue was, how to solve it, where its value lies, what makes it different from others, and how to tell this story to others. Thus, in the process of building our model, we continuously improved our direction, with team members promptly pulling back anyone who strayed, ensuring the project stayed on the right track. For instance, in the week, we had doubts about the value of our project and needed clarification about where our focus should be. Our initial assumption was on machine learning modeling and data analysis, but we realized this did not reflect the uniqueness of our project. Therefore, some team members gradually shifted the focus to the website, leading to some confusion. However, after discussion, we clarified our team's value. The data science part is provided to governments and the academic community, while the website part is made available for public use, with the focus still on the former. The website is a means to make our project models more understandable and to present the data more clearly.

Based on such values and goals, we decided I would build another demo version of the prediction model for public use, focusing only on features like time, whether it rains, the day's temperature, and traffic volume. Inputting these four variables alone would yield the PM2.5 and values for the day. Morgan continued to train the Pro version (i.e., the initial version integrating all features), considering the Pro version has many features that ordinary people cannot understand and input one by one. Therefore, we selected a few simple and easy-to-understand variables for refinement.

Additionally, regarding the dataset, after modeling, we believed the data integrated in the first week needed to be more. Although the data quality was good, the timeline only extended to December 2022 without including 2023, which could affect the model's prediction accuracy. Moreover, the old dataset's air quality-related features only included PM2.5 and PM10. Thus, Alvin searched for data again, finding data updated to the end of 2023, including SO_2, NO_2, etc., making the data more comprehensive. At this point, everyone's roles became gradually clear.

In the fourth week, we preliminarily built our website using VS Code and React and created a backend service with Flask to load and run our prediction models. Then, we provided an API (Application Programming Interface) to receive requests from the front end and return the model's prediction results. My team members integrated the front-end and back-end, ensuring smooth data exchange and communication. The demo model I was responsible for building also showed initial success. In the fourth week, I made several models and compared their performance:

First, I examined the correlation between features (variables) and found low correlation among them:

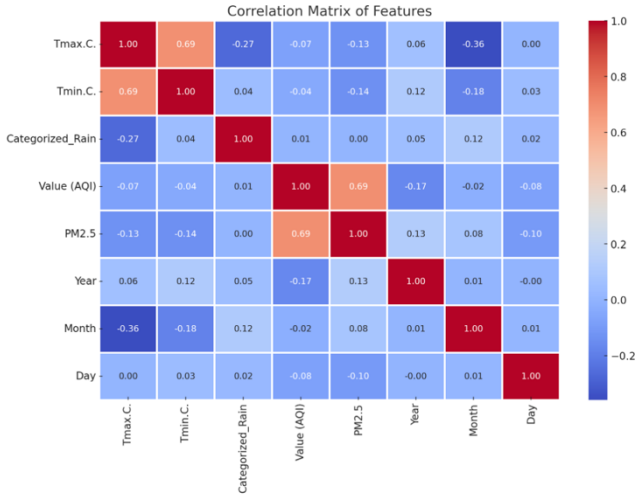


Fig. 3. Correlation Matrix of Features

Low correlation between data features usually means that no single feature can predict the target variable well. In this case, choosing a model capable of capturing complex patterns and relationships from the data is crucial:

1) Random Forest

Advantages: Random forest is a powerful ensemble learning method, combining multiple decision trees to improve prediction accuracy and stability. It handles low feature intercorrelation well and is robust for various data types and distributions.

Limitations: Random forest may not capture nonlinear relationships between features, especially in large or complex feature spaces.

MSE for AQI prediction: About 39.15

MSE for PM2.5 prediction: About 2.38

R^2 score for AQI prediction: About 0.32

R^2 score for PM2.5 prediction: About 0.30

2) Neural Networks

Advantages: Neural networks, particularly deep neural networks, excel at learning complex nonlinear patterns from large datasets. They often perform best in highly complex and nonlinear problems.

Limitations: Neural networks require a lot of data for effective training, and their training process can be time-consuming and compute-intensive. Additionally, their results might not be easily interpretable.

LSTM: Attempted to use LSTM but found the model performance very poor, likely because LSTM is suited for continuous data rather than categorical data. Although our time is continuous data, we categorized traffic volume and rainfall.

```
AQI MSE: 141.42171508712036
PM2.5 MSE: 7.2863992883067565
AQI R^2: -2.0343533162731795
PM2.5 R^2: -0.565858424204376
AQI MAE: 10.351245028824714
PM2.5 MAE: 2.189606372866087
```

Fig. 4. LSTM Model's Performance

3) Gradient Boosting Machine (GBM)

Gradient Boosting Machines (like XGBoost and LightGBM): These are powerful ensemble learning methods that sequentially build trees to improve model performance, ideal for complex datasets.

When attempting to use GBM, we encountered a problem: this algorithm does not natively support multi-output regression. Therefore, I used MultiOutputRegressor to wrap GBM, allowing us to maintain a single model structure while predicting two target values. Here are the performance metrics:

MSE for AQI prediction: About 45.57

MSE for PM2.5 prediction: About 2.68

R^2 score for AQI prediction: About 0.21

R^2 score for PM2.5 prediction: About 0.21

4) Support Vector Machine (SVM)

SVM performs well on medium and small-scale datasets, especially when the relationships between data features are complex.

Since SVM also does not natively support multi-output regression, we used MultiOutputRegressor to wrap it, allowing simultaneous prediction of AQI and PM2.5.

MSE for AQI prediction: About 58.02

MSE for PM2.5 prediction: About 3.44

R^2 score for AQI prediction: About -0.0036

R^2 score for PM2.5 prediction: About -0.0095

These results indicate that the SVM model performs poorly on this specific dataset, actually worse than simple average predictions (R^2 scores close to 0 or negative).

Considering neural networks require a lot of data for training, training time might be long, and their results are not easily interpretable. We need an easy-to-implement, interpretable model with lower computational costs. Combining our trials, the random forest model performs best among these algorithms.

Eventually, I chose the Random Forest model for prediction. I converted continuous data into categorical data for processing, for example:

- Traffic Count: Use quartiles for categorization:
  a) Low traffic (0): Less than the 25th percentile
  b) Medium traffic (1): Between the 25th and 75th percentiles
  c) High traffic (2): Greater than the 75th percentile
- Rainfall:
  a) Most values are 0, indicating no rainfall.
  b) Values greater than 0 are categorized as 1, indicating rainfall.
- Data processing: Extract year, month, and day from the Timestamp column.

So, the features columns are: 'Year,' 'Month,' 'Day,' 'Categorized_Traffic,' 'Tmax.C.,' 'Tmin.C.,' 'Categorized_Rain,' and the labels' columns are: 'PM2.5,' and 'PM10.'

Subsequently, I conducted hyperparameter tuning to optimize the model (from 2020-01-01 to 2022-12-31):

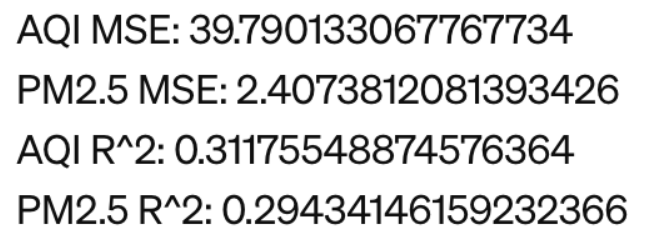


Fig. 5. Random Forest Model's Performance (from 2020-01-01 to 2022-12-31)

Then we changed the dataset, we have added some new dates into the dataset, and the current performance of the random forest (from 2020-01-01 to 2023-08-23):

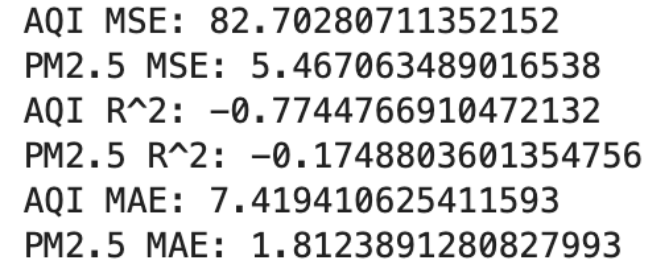


Fig. 6. Random Forest Model's Performance (from 2020-01-01 to 2023-08-23)

The modeling process and uploading the model to the website went smoothly.

In the fourth week, we further designed our website to make it more aesthetically pleasing and concise. Then, we continued to expand our scope, searching for regional data from Christchurch and Wellington to model and upload these prediction models to our website.

B. Late Stage of the Project

From the fifth to the tenth week, our cooperation efficiency accelerated due to a clear division of labor and unified goals. Meanwhile, when the project requirements shifted.

While analyzing time series data, I realized the need for more complex models to capture the temporal correlations in the data. Thus, in the later stages, I overturned the Random Forest model and tried the LSTM and CNN models, like Morgan.

After several attempts, I successfully obtained three demo prediction models for three cities. The features of the models included:

-Date (converted into total nanoseconds since 1970-01-01)
-WDir (Deg) (wind direction in degrees)
-WSpd (m/s) (wind speed in meters per second)
-Rain (mm) (rainfall in millimeters)
-RH (%) (relative humidity in percentage)
-Tmax (C) (maximum temperature in Celsius)
-Tmin (C) (minimum temperature in Celsius)
-LightCount (light vehicle count)
-HeavyCount (heavy vehicle count).

Here are the models' performance:

1) Christchurch

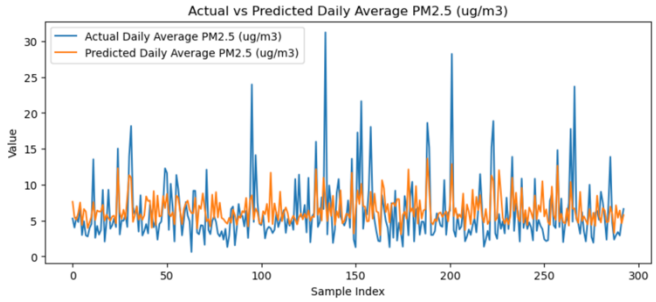


Fig. 7. Actual vs Predicted Daily Average PM2.5

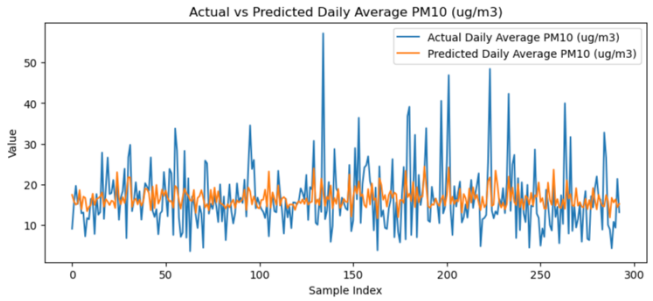


Fig. 8. Actual vs Predicted Daily Average PM10

Fig. 9. Training and Validation Loss

Test Loss: 0.008163809776306152
Mean Squared Error (MSE): 33.23885165953058
Mean Absolute Error (MAE): 3.8868524803740616
R^2 Score: 0.2085922168933888

2) Auckland

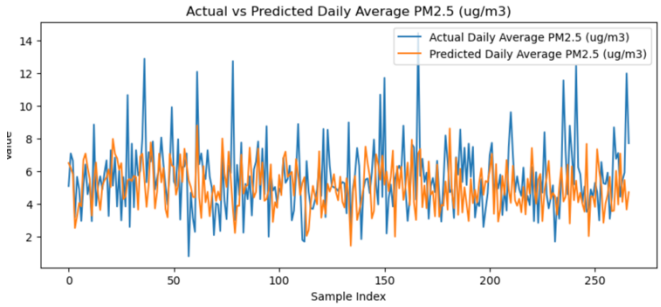


Fig. 10. Actual vs Predicted Daily Average PM2.5

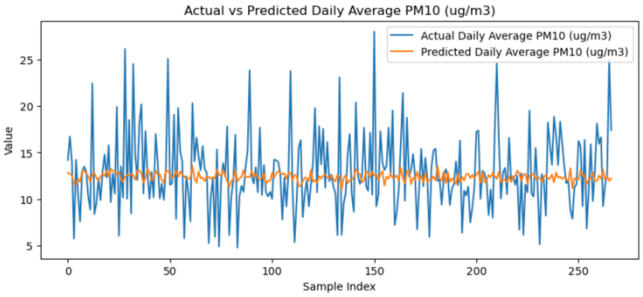


Fig. 11 Actual vs Predicted Daily Average PM10

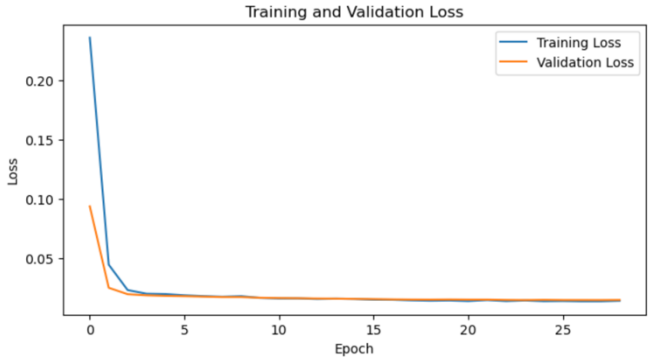


Fig. 12. Training and Validation Loss

Test Loss: 0.01175309345126152
Mean Squared Error (MSE): 10.960694776449197
Mean Absolute Error (MAE): 2.3347147475276238
R^2 Score: -0.07054848358981725

3) Wellington

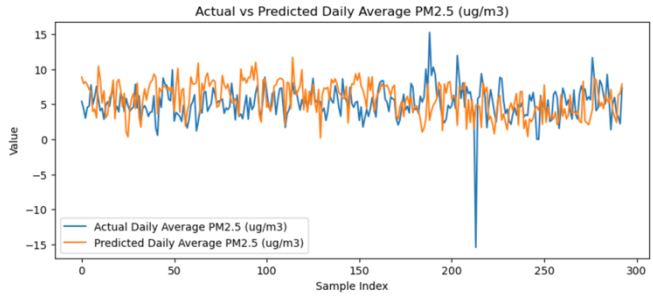


Fig. 13. Actual vs Predicted Daily Average PM2.5

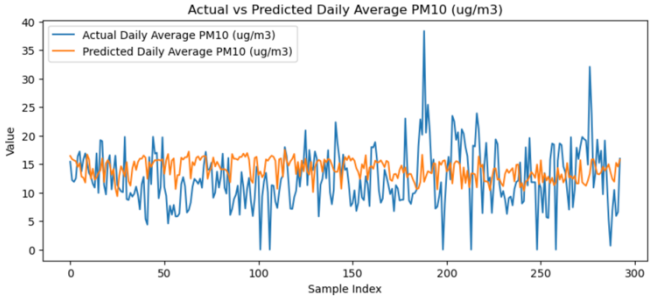


Fig. 14. Actual vs Predicted Daily Average PM10

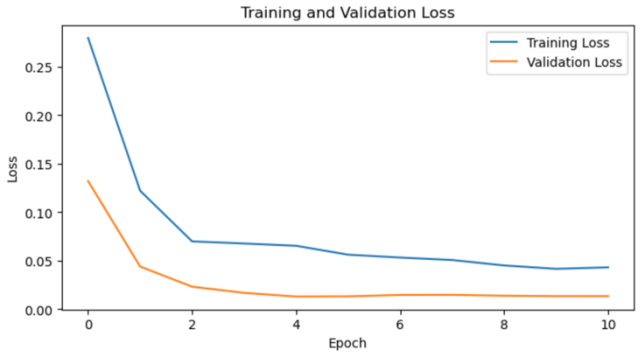


Fig. 15. Training and Validation Loss

Test Loss: 0.016157513484358788
Mean Squared Error (MSE): 22.156964887411377
Mean Absolute Error (MAE): 3.588935712221948
R^2 Score: -0.6454006946256792

I also gained a deep understanding of the practical operation of LSTM and CNN.

Meanwhile, Alvin used Python to fill in the data visualization part on our website, allowing the public and researchers to access and understand it easily.

Here are some examples of data visualization:

Start Date	Site Alias	Region Name	Site Reference	Class Weight	Site Description	Lane Number	Flow Direction	Traffic Count
2018/1/1 01:00	39	05 - Gisborne	00200444	Light	200 m Nth of Bell Rd	2	2	2,468
2018/1/1 01:00	39	05 - Gisborne	00200444	Heavy	200 m Nth of Bell Rd	2	2	73
2018/1/1 01:00	39	05 - Gisborne	00200444	Heavy	200 m Nth of Bell Rd	1	1	74
2018/1/1 01:00	39	05 - Gisborne	00200444	Light	200 m Nth of Bell Rd	1	1	3,806
2018/1/1 01:00	40	05 - Gisborne	00200448	Heavy	Sth end of Whatatuna Bridge	1	1	58.5
2018/1/1 01:00	40	05 - Gisborne	00200448	Light	Sth end of Whatatuna Bridge	1	1	2,915.5
2018/1/1 01:00	40	05 - Gisborne	00200448	Light	Sth end of Whatatuna Bridge	2	2	1,569
2018/1/1 01:00	40	05 - Gisborne	00200448	Heavy	Sth end of Whatatuna Bridge	2	2	51
2018/1/1 01:00	50	06 - Hawkes Bay	00200646	Light	Napier Airport Combined	2	2	6,499
2018/1/1 01:00	50	06 - Hawkes Bay	00200646	Light	Napier Airport Combined	1	1	7,827
2018/1/1 01:00	54	06 - Hawkes Bay	05100002	Light	Hyderabad : Taradale - Georges ...	4	2	4,667
2018/1/1 01:00	54	06 - Hawkes Bay	05100002	Light	Hyderabad : Taradale - Georges ...	1	1	5,153
2018/1/1 01:00	54	06 - Hawkes Bay	05100002	Light	Hyderabad : Taradale - Georges ...	2	1	3,085

Fig. 16. Data Source: TMS daily traffic counts API from NZTA

Item	Datatype Option	Definition
WDir(Deg)	Extended	Wind direction at 9am local (spot value)
WSpd(units users choice)	Extended	Wind speed at 9am local (spot value)
GustDir(Deg)	Extended	Direction of Max Gust over 24 hours from midnight to midnight of current local day
GustSpd(units users choice)	Extended	Speed of Max Gust over 24 hours from midnight to midnight of current local day
WindRun(Km)	Extended	The number of km of wind run. Placed back one day from time of measurement (see above). Note Km, wind_run / 24 will give the mean wind speed for 24 hours).
Rain(mm)	Basic	Rain over 24 hours 9am to 9am local. Placed back one day from time of measurement (see above)
T9hr(C)	Basic	Spot value of dry bulb temperature (air temperature) at 9am local
TWet(C)	Basic	Spot value of wet bulb temperature at 9am local. Often calculated from relative humidity
RH(%)	Basic	Spot value of relative humidity at 9am local.
Tmax(C)	Basic	Maximum temperature over last 24 hours to 9am local. Placed back one day from time of measurement (see above).
Tmin(C)	Basic	Minimum temperature over 24 hours to 9am local.
Tgmin(C)	Basic	Minimum grass temperature over 24 hours to 9am local.
ET05(C)	Extended	Spot value of earth temperature at 5cm depth at 9am local.
ET10(C)	Extended	Spot value of earth temperature at 10cm depth at 9am local.
ET20(C)	Extended	Spot value of earth temperature at 20cm depth at 9am local.
ET30(C)	Extended	Spot value of earth temperature at 30cm depth at 9am local. Note that this F301 output is for 30cm depth which is different to the "Daily 0000" output which is for 50cm depth. Most manual stations use 30cm, most EWS and AWS stations use 50cm.
ET100(C)	Extended	Spot value of earth temperature at 100cm depth at 9am local.
Pmsl(hPa)	Extended	Spot value of mean sea level pressure at 9am local.
Pstn(hPa)	Extended	Spot value of station level pressure at 9am local.
Sun(Hrs)	Extended	Hours of sunshine for 24 hours from midnight to midnight of current local day
Rad(MJ)	Extended	Amount of solar radiation in MJ/m2 over 24 hours from midnight to midnight of current local day.

Fig. 17. Data Source: Climate database from NIWA

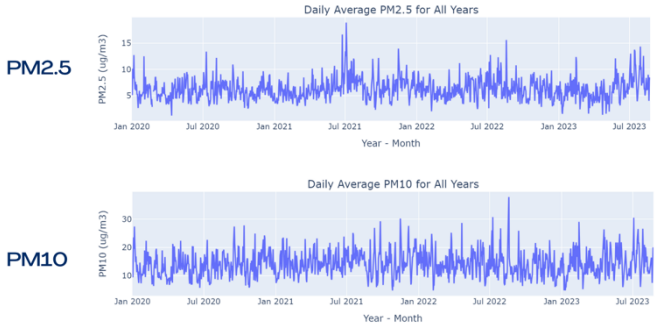


Fig. 18. EDA for Auckland

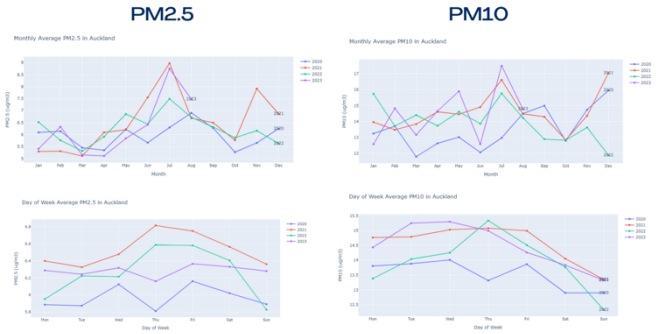


Fig. 19. EDA for Auckland

VII. Project Results

We put our prediction model on our website, and the final presentation of the results is dominated by the website. It is divided into Home, Model, and Dataset interfaces.

Fig. 20. Home Interface

This is the main functional interface, where users can choose the desired functions.

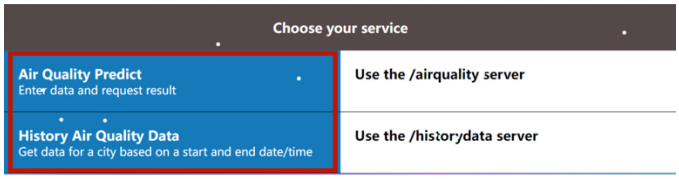


Fig. 21. Main Functional Interface

You can choose the form of output (JSON/XML), the version: demo (popular) and Pro (professional) and the city.

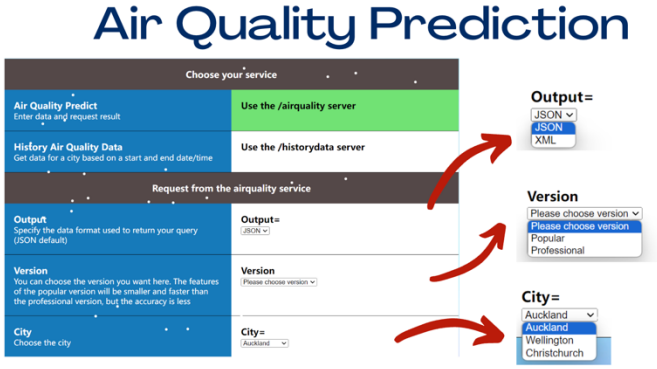


Fig. 22. Main Function Interface Expanded

A. Popular Version

For the Popular Version, users are given an example input, which they can follow to input data, or they can autofill other columns by entering data in the Date field.

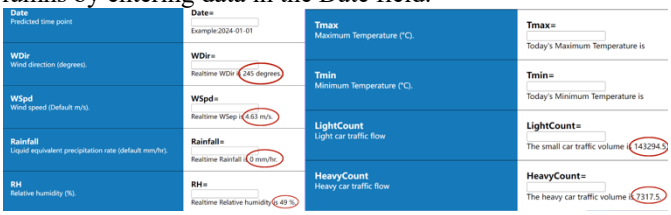


Fig. 23. Popular Version Interface

The output results include the city, PM10, and PM2.5 values.


```
{
  "prediction": {
    "city": "Auckland",
    "pm10": 14.294424057006836,
    "pm2.5": 17.936315536499023
  }
}
```


Fig. 24. Output Results for the Popular Version

B. Professional Version

This version requires users to input 19 additional features. Unfortunately, we were unable to find an application programming interface (API) that provides this data. It requires more time to run but is more accurate.

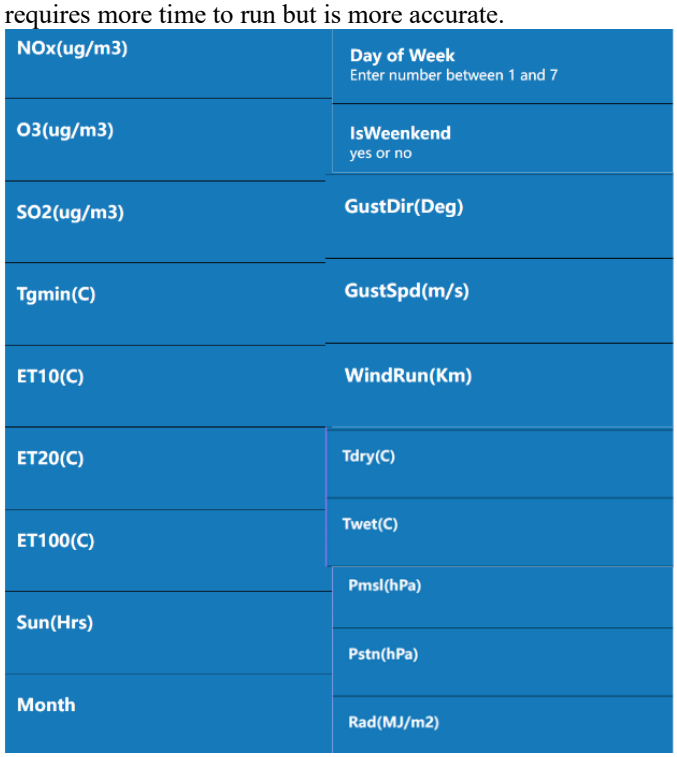


Fig. 25. Professional Version's Features

The output results include the AQI, city, PM10, and PM2.5 values.


```
{
  "prediction": {
    "AQI": 67.62008666992188,
    "city": "Auckland",
    "pm10": 16.65230941772461,
    "pm2.5": 2.6362922191619873
  }
}
```


Fig. 26. Output Results for the Professional Version

This interface allows users to view historical data related to air quality. Users can select a specific time range and view the corresponding data.

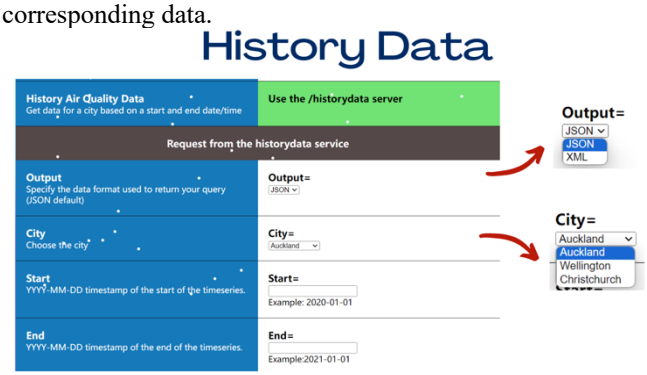
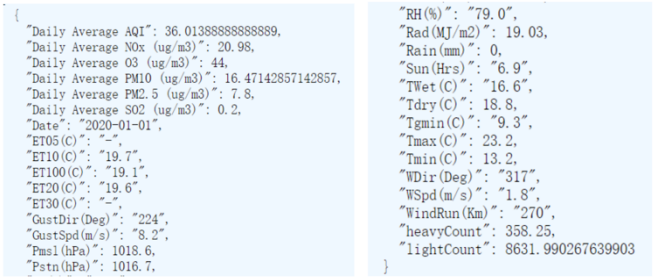


Fig. 27. Historical Data Interface



```
{
  "Daily Average AQI": 36.01388888888889,
  "Daily Average NOx (ug/m3)": 20.98,
  "Daily Average O3 (ug/m3)": 44,
  "Daily Average PM10 (ug/m3)": 16.47142857142857,
  "Daily Average PM2.5 (ug/m3)": 7.8,
  "Daily Average SO2 (ug/m3)": 0.2,
  "Date": "2020-01-01",
  "ET05(C)": "-",
  "ET10(C)": "19.7",
  "ET100(C)": "19.1",
  "ET20(C)": "19.6",
  "ET30(C)": "-",
  "GustDir(Deg)": "224",
  "GustSpd(m/s)": "8.2",
  "Pmsl(hPa)": 1018.6,
  "Pstn(hPa)": 1016.7,
  "RH(%)": "79.0",
  "Rad(MJ/m2)": 19.03,
  "Rain(mm)": 0,
  "Sun(Hrs)": "6.9",
  "TWet(C)": "16.6",
  "Tdry(C)": 18.8,
  "Tgmin(C)": "9.3",
  "Tmax(C)": 23.2,
  "Tmin(C)": 13.2,
  "WDir(Deg)": "317",
  "WSpd(m/s)": "1.8",
  "WindRun(Km)": "270",
  "heavyCount": 358.25,
  "lightCount": 8631.990267639903
}
```


Fig. 28. Output Results for Historical Data Interface

This interface allows users to select features they want to include in their query or analysis. It could be a form with checkboxes or dropdown menus for selecting different environmental parameters, such as temperature, humidity, wind speed, etc.

Fig. 29. Data Interface Provided to Users

VIII. Reflection

A. Professional Attributes Developed & Applied

1) Hard Skills

a) Machine Learning

Firstly, I had only theoretically studied neural networks before, but now I'm proficient in using LSTM. Through this process, I've not only learned to implement complex machine learning models but also understood how to adjust them for specific datasets and business needs.

b) Data Collecting

Secondly, I was responsible for finding and organizing data for the first time. Previously, I worked with datasets provided by teachers. During this internship, I realized that this is a very time-consuming process, requiring patience and mental preparation, as it's likely that not all data will be found or be suitable. This process is a test of mindset, like optimizing a model.

c) Frontend and Backend Development

Thirdly, I also learned some front-end and back-end knowledge, such as how to initially build our own website using VS Code and react and create a back-end service using Flask to load and run our prediction models.

d) Ability to Learn Quickly

Lastly, throughout the internship, I discovered my ability to learn new knowledge and skills quickly. I believe a self-driven learning attitude and the capability for continuous development are essential in any rapidly evolving industry. For instance, when transitioning from Random Forest to LSTM models, I quickly learned how to build and optimize LSTM models and assess model performance by observing the differences between predicted and actual values in trend graphs. This adaptability is essential in a rapidly changing work environment.

2) Soft Skills

a) Presentation Skills

This internship has significantly enhanced my presentation skills. Before this, I was uncomfortable presenting in front of many people, finding it was torturous. Over these weeks, with frequent large and small presentations, I gradually gained confidence, from avoiding eye contact and stuttering to becoming more assured. Adequate preparation and constant practice can boost confidence. Furthermore, Deb taught us presentation skills, such as starting with an interesting story. I also stepped out of my comfort zone and interacted with the audience. This progress makes me more optimistic about future presentations, focusing on mindset and presentation skills, which are crucial when dealing with real clients in a professional setting.

Delegating Appropriately

The second skill I developed is communicating with people of different temperaments and learning to delegate. Previously, I tended to be anxious, worrying about others not performing their tasks well. However, through extended group collaboration, I have learned to treat others as equals and trust them to do their part, making cooperation easier and more efficient.

B. Knowledge Applied from Taught Courses

1) COMPSCI 762: Foundations of Machine Learning

This introductory course on machine learning fueled my passion for the field. I learned the basic principles of algorithms and the challenges of teaching computers to learn from data; I also developed practical skills to solve different learning problems and critically evaluate modeling results. This course allowed me to focus further on advanced areas in data science, machine learning, and artificial intelligence. It taught me the processes and principles of data analysis, including the workflow and mental adjustments required. During my internship, I frequently applied the professional knowledge from this course, such as using scikit-learn, data preprocessing, understanding neural networks, model evaluation and validation (e.g., cross-validation, performance measurements like accuracy, precision, recall, ROC/AUC), and theories of supervised (e.g., decision trees, support vector machines, neural networks) and unsupervised learning methods (e.g., clustering, association rule mining). This process was very interesting and was a key reason I decided to undertake a data-related project for my internship.

2) COMPSCI 747: Computer Education

Writing skills are crucial, especially for someone who has never written an English paper before and is unfamiliar with how to write a literature review. This course taught me how to write a literature review professionally, which was immensely helpful in understanding and integrating knowledge from related research fields. Not only did it improve my academic writing skills, but it also deepened my understanding of the theories and practical knowledge related to my project. By applying the theoretical knowledge learned in class to real-world projects, I enhanced my technical skills and ability to solve practical problems.

3) GLMI 709: Creating Global Ventures

This course on entrepreneurship involved many presentations and taught me how to introduce my product to clients, create effective slides, and many other soft skills. It also gave me an understanding of business models and strategies, how to identify and evaluate opportunities, and how to use frameworks and tools to assess risks and develop viable business models. I realized that transferable skills related to teamwork, problem-solving, and communication are developed by working in teams to turn business ideas into reality.

4) INFOSYS 700: Digital Innovation

This course developed my critical thinking and understanding of digital businesses. Exploring the theory and practice of digital innovation, its potential impact, and its disruptive effects on business and society, I gained insights into emerging technologies such as the sharing economy, blockchain, digital innovation, and the Internet of Things.

C. *Reflection & Learning*

The challenges faced during the project provided invaluable learning opportunities. From initially selecting the Random Forest model to later exploring LSTM and CNN models, each step was accompanied by discoveries and understandings. For instance, I initially used the Random Forest model, chosen for its efficiency in handling classification and regression problems, its capability to process large volumes of data, and its robustness to outliers and non-linear relationships. This model was selected as the best performer among multiple models. The entire modeling process with Random Forest went smoothly, and the prediction results were good. Passing the model to Bella for integration into the website was also seamless.

However, as the project shifted towards analyzing time series data, I realized the need for more complex models capable of capturing the temporal correlations in data. Consequently, I ventured into the unfamiliar territory of LSTM and CNN models. LSTM models are particularly suited for time series data as they can learn long-term dependencies in data. Although typically used for image processing, CNN models can also be applied to sequence data, improving prediction accuracy by identifying patterns within the time series. Despite these models offering better performance potential, they also introduced more complex challenges in model tuning and interpretation.

After trials, it was evident that their performance was not up to expectations, largely because LSTM is not suited for categorical data; it is more appropriate for continuous data. Bella then suggested that she could find different data APIs (which could provide real-time air quality application programming interfaces), such as providing the latest real-time traffic flow and rainfall data, eliminating the need for online searches. With access to continuous data, I rebuilt the LSTM model. I encountered numerous challenges during the technical implementation, including data preprocessing, model tuning, and technical difficulties related to frontend and backend integration. Data preprocessing was particularly enlightening, teaching me how to clean and organize complex time series data, handle missing and outlier values, and perform feature engineering to extract useful information. Model tuning was a trial-and-error process requiring continuous experimentation with different parameter combinations. I have become very familiar with this model, from package adjustments to learning the reasons behind code errors. Some critical observations during this process were:

The initially built models for the three cities relied on different variables, leading to disparities in model performance. For example, Christchurch's dataset included additional SO_2 and NO columns; Auckland's dataset had extra SO_2 and AQI columns, but Wellington had neither, only PM2.5 and PM10 (which the other two cities also had). After training three separate LSTM models, it was clear that Christchurch's model had the best predictive capability, followed by Auckland, with Wellington's model performing the worst. I tried to standardize the variables across the three models at Bella's suggestion to ensure data consistency. Removing the extra data from Auckland and Christchurch to match Wellington worsened the model performance. Despite efforts to adjust parameters for better results, the experiments showed that the model had learned as much as possible from the available data. However, the existing features had little correlation with PM10, so the model performance was unsatisfactory no matter how much the parameters were adjusted. This highlighted the issue of dataset quality. Although this adjustment reduced the model's predictive performance, it taught me how to balance model accuracy and generalizability.

Additionally, technical issues encountered during model deployment, such as software package compatibility and data processing accuracy, presented challenges that needed to be overcome. For example, the results produced after uploading the model (i.e., PM2.5 and PM10) were unreasonably high and clearly incorrect. This taught me the importance of correctly processing time series data (dates) and how to test model results to evaluate performance. The model can only handle data suitable for it, regardless of the original data or features used; it concerns itself only with the data produced during its processing. For instance, though the input data is raw, it requires standardization, and the data used for predictions must also be standardized. This process was cumbersome, as all data processing steps needed to be saved in a separate .pkl file to be sent to the backend team for processing.

```
[1]: import numpy as np
     import pandas as pd
     import joblib
     from sklearn.preprocessing import MinMaxScaler
     from tensorflow.keras.models import load_model

     # Load MinMaxScaler instance for input data
     input_scaler = joblib.load('/Users/yangyi/778/Auckland_input_scaler.pkl')

     # Load MinMaxScaler instance for output data
     output_scaler = joblib.load('/Users/yangyi/778/Auckland_output_scaler.pkl')

     # Load model
     model = load_model('/Users/yangyi/778/Auckland_model.h5')

     # Input data
     input_data = np.array([[1706140800000000000, 218, 5.66, 0, 62, 28, 22, 8000, 100]])

     # Standardize input data
     input_data_scaled = input_scaler.transform(input_data)
     input_data_scaled_reshaped = input_data_scaled.reshape((1, -1, 1))

     # Use model to make predictions
     predictions = model.predict(input_data_scaled_reshaped)

     # Inverse standardize prediction results
     predictions_inversed = output_scaler.inverse_transform(predictions)

     # Output prediction results
     print(predictions_inversed)

     WARNING:absl:At this time, the v2.11+ optimizer `tf.keras.optimizers.Adam` runs slowly on M1/M2
     Macs, please use the legacy Keras optimizer instead, located at `tf.keras.optimizers.legacy.Ada
     m`.
     1/1 [==============================] - 0s 68ms/step
     [[13.149524   6.5863066]]
```

Fig. 30. Testing if the model operates normally

IX. Conclusion

In conclusion, this internship was very successful, as we achieved all our initial project goals and learned many things independently. This project taught me a deeper understanding of combining theoretical knowledge with practical skills to solve complex problems. This project was not only a technical challenge but also a learning and growth process. From transitioning between Random Forests to LSTM and CNN, from data processing to model deployment, I have gained a more comprehensive understanding and knowledge of data science. After this internship, I will continue to learn about machine learning. I enjoy working in this field and hope to apply it further in the future, possibly moving into the field of artificial intelligence, where I hope to make a significant contribution.

Acknowledgment

I would like to express my gratitude to my academic supervisor, Philipp Skavantzos, and my mentors, Sean Zeng and Andrew Meads, for their support and guidance throughout this internship. Their insights have prompted me to think about my career planning and have illuminated the path for my future professional endeavors.

References

[1] T. G. Dietterich, "Machine learning," in *Encyclopedia of Computer Science*, GBR: John Wiley and Sons Ltd., 2003, pp. 1056–1059.

[2] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[3] M. Bharati and B. Ramageri, "Data mining techniques and applications," *Indian J. Comput. Sci. Eng.*, vol. 1, Dec. 2010.

[4] K. Mehmood *et al.*, "Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives," *J. Clean. Prod.*, vol. 379, p. 134656, Dec. 2022.

[5] E. Pasero, L. Mesin, E. Pasero, and L. Mesin, "Artificial Neural Networks for Pollution Forecast," in *Air Pollution*, IntechOpen, 2010.

[6] M. Lee *et al.*, "Forecasting Air Quality in Taiwan by Using Machine Learning," *Sci. Rep.*, vol. 10, no. 1, p. 4153, Mar. 2020.

[7] K. Zhang, X. Zhang, H. Song, H. Pan, and B. Wang, "Air Quality Prediction Model Based on Spatiotemporal Data Analysis and Metalearning," *Wirel. Commun. Mob. Comput.*, vol. 2021, p. e9627776, Aug..

[8] A. Hasnain *et al.*, "Time Series Analysis and Forecasting of Air Pollutants Based on Prophet Forecasting Model in Jiangsu Province, China," *Front. Environ. Sci.*, vol. 10, 2022, Accessed: Dec. 18, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fenvs.2022.945628

[9] "$PM_{2.5}$ concentrations | Stats NZ." Accessed: Feb. 21, 2024. [Online]. Available: https://www.stats.govt.nz/indicators/pm2-5-concentrations

[10] "WHO Global Air Quality Guidelines." Accessed: Feb. 21, 2024. [Online]. Available: https://www.who.int/news-room/questions-and-answers/item/who-global-air-quality-guidelines

[11] "EHINZ." Accessed: Dec. 17, 2023. [Online]. Available: https://www.ehinz.ac.nz/projects/hapinz3/key-findings-from-hapinz/

[12] S. Ilarri, R. Trillo-Lado, and L. Marrodán, "Traffic and Pollution Modelling for Air Quality Awareness: An Experience in the City of Zaragoza," *Sn Comput. Sci.*, vol. 3, no. 4, p. 281, 2022.

[13] P. Louridas and C. Ebert, "Machine Learning," *IEEE Softw.*, vol. 33, pp. 110–115, Sep. 2016.

[14] L. A. Díaz-Robles *et al.*, "A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile," *Atmos. Environ.*, vol. 42, no. 35, pp. 8331–8340, Nov. 2008.

[15] K. P. Singh, S. Gupta, A. Kumar, and S. P. Shukla, "Linear and nonlinear modeling approaches for urban air quality prediction," *Sci. Total Environ.*, vol. 426, pp. 244–255, Jun. 2012.