Case Studies in Data Science

Work Integrated Learning Project

Final Report

Bigfoot

City of Melbourne Foot Traffic number predictions

Group 3

Student ID	Full Name	Percentage Contribution
s3892540	Ashley Loong	16.66%
s3908162	Darren Yeo	16.66%
s3884280	Eleanor Cummins	16.66%
s3302044	Qi Lu	16.66%
s2105802	Nelson Cheng	16.66%
s3874566	James Twigg	16.66%

Executive Summary

Every customer-serving, bricks-and-mortar business faces the challenge of stock management and resource utilisation. Bigfoot looks to help with these problems by giving indications of active population in the City of Melbourne via foot traffic number predictions.

Daily Foot Traffic Data was sourced from the City of Melbourne website [1]. Then, basic weather, calendar, social and economic feature data was sourced for modelling.

Experiments were carried out to create a prototype model for predicting the foot traffic data based on the feature data. Modelling algorithms experimented with include Lasso Regression, Ridge Regression, XGBoost Decision Tree based Regression [2], ARIMA Time Series Forecasting, Azure Automated ML Ensembled Regression and Azure Automated ML Ensembled Time Series Forecasting [3].

The best performing model for the aggregated City of Melbourne foot traffic data was the Azure Automated ML Ensembled Time Series Forecasting model [1, 3], with the following results. Results reported in Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) in Table 1 :

Table 1 - Model Results Summary

MAE:	25322	Average Prediction Error (%):	5.26%
RMSE:	29177	Prediction Error Deviation (%):	+/- 6.06%

While these results are initially promising, evaluation of modelling and predictions on unseen data found that there was a significant gap between training predictions versus unseen data predictions, giving an indication that the model suffers from an overfitting issue. Methods are outlined for improving on this prototype model, the most important of which would be sourcing more data across domains such as weather, calendar, social and economic.

A prototype for the Bigfoot product application is presented, along with several potential use cases where Bigfoot could add value. Proposals for roadmap items and future scope expansions demonstrate the Bigfoot has significant growth potential beyond the prototype.

Introducing Bigfoot

Emerging AI and machine learning technologies can assist businesses to create value in a variety of ways. Our project, Bigfoot, utilizes machine learning models to help business owners to maximise profits by providing future foot traffic level predictions.

After three years of spread, the end of COVID-19 global pandemic is in sight. The state government of Victoria has announced the end of COVID-19 declaration in October 2022. With mask rules and other public health measures easing, we have entered the post pandemic era. People are returning to the Melbourne CBD in increasing numbers and the city council has setup strategies to lead Victoria out of the COVID-19 recession. According to a 2019 City of Melbourne councils report, the daily population in the City of Melbourne is around 1 million, who collectively generate $100 billion Gross Local Product in 2019 [4]. And this number is set to reach $150 billion by 2031 [4]. This thriving economic future provides businesses great opportunities to grow and develop. This project leverages data science and machine learning technologies to assist local business to seize opportunity, take advantages in the post pandemic boom and create value in an efficient way.

This new product, Bigfoot, is designed to help businesses potentially gain an advantage in a changeable environment. Bigfoot looks to help owners and employees to deal with day-to-day challenges such as stock management and roster planning by providing 7-day foot traffic level predictions in Melbourne city. The product uses data from various sources and utilises machine learning technologies to generate predictions. Bigfoot's input data includes weather attributes such as temperature, solar exposure and rainfall, significant calendar events such as weekends and holidays, as well as local economic performance, all to predict daily foot traffic, with foot traffic data provided from City of Melbourne council [1]. With feature engineering and optimised learning models, the prototype has already achieved high accuracy and will be able to provide users with useful insights. Businesses in Melbourne can use Bigfoot to anticipate visitor flow changes and be proactive to eliminate wastage and generate profit

effectively. Apart from local businesses, individual users such as commuters and visitors could also benefit from bigfoot in case of planning and get the most of their time in Melbourne.

Problem Definition

Melbourne was locked down for more than 260 days since pandemic began, which has made it the world's most locked down city. The COVID-19 led recession is sharper and deeper than any other in living memory and is having a disproportionate effect on Melbourne's economy.

Hundreds of discretionary services and retail businesses bankrupted or temporary closed in the past 2 and half years. Projection shows weak economic output and employment maybe continues beyond 2024 [6]. Based on the city council financial review, the food, beverage services and retail business account for more than 30% of local economy [4]. There are more than 75,000 employees in this small business sector in Melbourne [4]. Finding efficient ways to manage costs and enhance revenues is crucial for street-level businesses to survive and generate profit during the long recovery period.

Stock and labour cost management are the key concerns for private companies as they operate and grow. Australia has one of the highest salary rates in the world [5]. Many businesses were shut down because of the high labour costs and lack of cash flow. In addition, the Australian Bureau of Statistics (ABS) reported that the annual Australian inflation rate had reached at 6.1% in the June quarter, which was the highest recorded since 1990 [7]. In such a challenging period, lean running and efficiency may be a business's key to survival on the road back to 'Covid normal'.

Bigfoot will provide businesses in Melbourne the chance to project sales and increase revenue. High levels of foot traffic on the street means that many people will pass the shopfront. With reliable foot traffic predictions, owners and operators can use Bigfoot to forecast incoming customers, allowing them to sensibly manage their stock and staff to optimise revenue and profit.

While business owners are a prime focus, Bigfoot is expected to be useful to a wide variety of demographics. As a freely and publicly available website and mobile application, anyone can download and make use of Bigfoot's predictions, whether it be a parent trying to find the quietest time or spot to do their Christmas shopping, or an average grocery shopper wanting to avoid large crowds on their way to the store.

Prototype Methodology

As mentioned, the purpose of this prototype is to predict Foot Traffic numbers in the City of Melbourne based on modelling on publicly available weather, calendar and economic Melbourne and Australian data. In order to predict this, models were built for two different problems, which were to predict the aggregated City of Melbourne foot traffic predictions and to predict foot traffic numbers at specific street locations within the CBD.

This section discussed in detail the methodology of building the research and experiments carried out to create the prototype models. This includes information on the data collected, the models implemented, the method of evaluation on models and model results, the experiment results, and future recommendations on modelling.

The Data

The first step in the process was the collection of usable data, then preparation and analysis of the data so the data would be in a usable format for modelling and prediction.

Data Collection

For this model, the target data was daily foot traffic numbers from the City of Melbourne website. Feature data to train the model was pulled from a number of different sources as detailed below.

In all cases 9 years of data, going back to 2013, formed the basis of the model.

Target Data

- Data: Daily Foot Traffic Data for the City of Melbourne, by street locations [1]
    - Source: The City of Melbourne Council – Pedestrian Counting System site
    - Link: http://www.pedestrian.melbourne.vic.gov.au/#date=24-08-2022&time=17

Feature Data

- Data: Daily Weather Data, including Min and Max Temperatures, Rainfall Levels and Solar Exposure [8].
    - Source: The Australian Bureau of Meteorology (BOM) website
    - Link: http://www.bom.gov.au/climate/change/datasets/datasets.shtml
- Data: Annual Melbourne Population numbers [9].
    - Source: MacroTrends, Melbourne, Australia Metro Area Population 1950-2022
    - Link: https://www.macrotrends.net/cities/206168/melbourne/population
- Data: Daily Covid Lockdown Dates in Melbourne [10].
    - Source: Lockdown Stats Melbourne
    - Link: https://lockdownstats.melbourne/timeline/
- Data: Public Holidays in Melbourne [11-12].
    - Source: Australian Government Website
    - Link: https://data.gov.au/dataset/ds-dga-b1bc6077-dadd-4f61-9f8c-002ab2cdff10/details?q= , https://www.timeanddate.com/holidays/australia/2013
- Data: Monthly Melbourne Retail numbers [13].
    - Source: The Australian Bureau of Statistics
    - Link: https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/latest-release#data-download
- Data: Monthly Australian Historical Market Statistics [14].
    - Source: ASX website
    - Link: https://www2.asx.com.au/about/market-statistics/historical-market-statistics

Data Preparation and Manipulation

Once all this data was sourced, the next step in the process was to clean, transform and combine all the datasets into a unified set of data that could easily be processed by our machine learning models. The goal in this project was to have two different model predictions, one for an aggregated City of Melbourne foot traffic prediction, and another for predictions according to specific street locations as made available in the source Foot Traffic data.

Examples of these specific street locations might be “Melbourne Central”, “Town Hall”, or “Flinders Street and Swanston Street (North)”. From a user perspective, an indication of foot traffic by specific location would provide greater utility.

As the data provided is for the specific street locations, scripts were implemented to aggregate the numbers together to provide aggregated totals for the foot traffic for the City of Melbourne.

Data Cleaning

When exploring and reviewing the data sets, the data was found the be very well structured and clean. Very little data cleaning was required, with only minimal processing required (like replacing empty values with zeroes). No significant operations like imputation were required.

All numerical fields were also investigated for outliers. However, no significant outliers were removed or cleaned. Even though some outlier values were found, for example in weather, such as large total rain amounts, these are important data points, as large rainfall events should be considered in modelling.

Crucially, there were no significant outliers for the Foot Traffic data to be filtered, as can be seen from the boxplot in Figure 1.

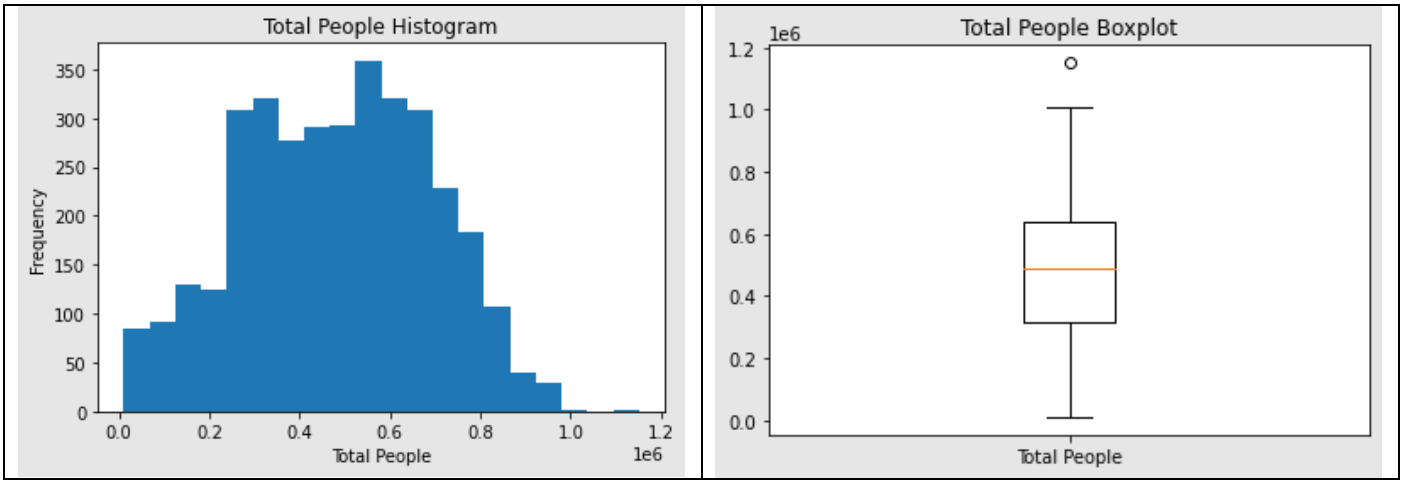

Figure 1: Example of outlier analysis for Total People

Data Joining

The purpose of the model is to predict Foot Traffic levels by day, ultimately to forecast 7 days ahead. While the target foot traffic data, as well as other data sets had daily data points, other data could only be found to have monthly or annual data numbers. Therefore, data sets like the Monthly Offline Retail data [13], ASX Historical Markets Statistics [14] and Melbourne Population [9] numbers were joined to the daily data accordingly to their respective month/year, where appropriate.

The results were two aggregated data sets, containing the respective data by day, with data spanning January 2013 to July 2022, which could be used for modelling. The first is for the aggregated City of Melbourne data set, an example of which can be seen below. The second is for the street specific location data, which is the same in format to the first, except there is an additional feature column which contains the name of the street location.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	date	total_people	total_rain	rain_qu	max_temp	max_te	min_temp	min_temp	solar_exp	WeekDay	populatior	populatior	is_holiday	is_lockdov	OfflineRetail_Orig	OfflineRetail_Se	Offline	all_ords	sp_asx200	dom_equity_m
2	31/07/2022	327383	0	N	14.7	Y	4.3	Y	4.8	6	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
3	30/07/2022	462115	0	N	13	Y	2.1	Y	11.3	5	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
4	29/07/2022	405511	1	N	12.7	Y	6.5	Y	11.2	4	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
5	28/07/2022	334858	1	N	13.2	Y	9.3	Y	9.3	3	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
6	27/07/2022	340569	3	N	15.3	Y	9.3	Y	7.7	2	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
7	26/07/2022	316316	4.4	N	13.2	Y	8.8	Y	6.4	1	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
8	25/07/2022	274106	0	N	16.8	Y	8.1	Y	5.1	0	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	
9	24/07/2022	406977	7.8	N	19.3	Y	10.4	Y	10.1	6	5151000	1.78	0	0	8562.7	8947.3	7173.8	6945.21	2453645	

Figure 2: Example of aggregated City of Melbourne Data Set

Exploratory Data Analysis

Several of the features of the data were explored and analysed with their respect to each other as well as to the target “total_people” target feature. In the interest of report length, not all results from data investigations are covered in this report, however anything of note is discussed within this report. An example plot showing the relationship between foot traffic and rainfall is displayed in Figure 3.

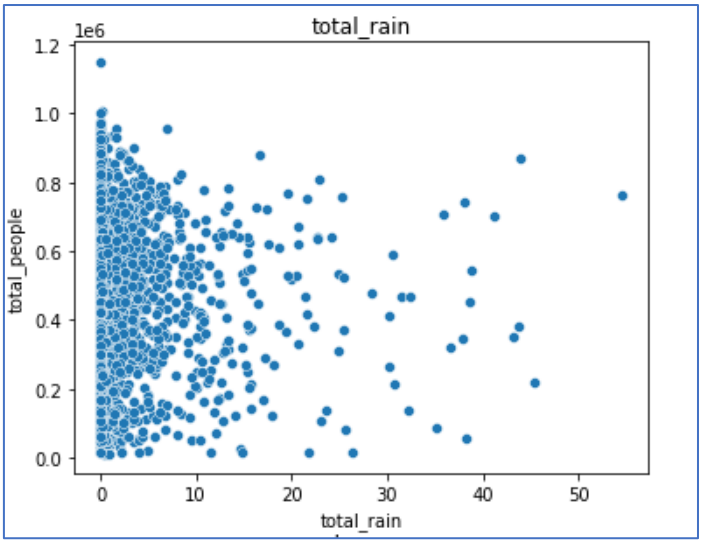

Figure 3: Scatter plot of "total_people" vs "total_rain"

Figure 3 illustrates a correlation between "total_people" and "total_rain" which seems to support a hypothesis that, unsurprisingly, people are less likely to be wandering the city streets when there is heavy rainfall.

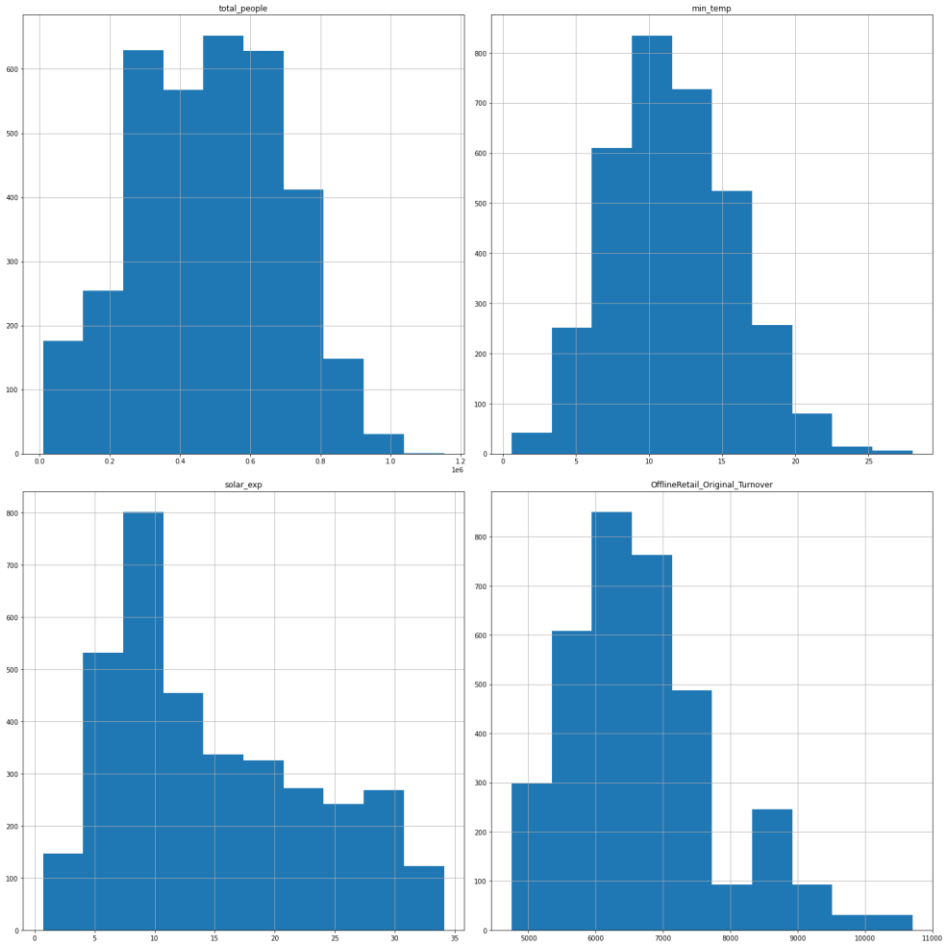

Figure 4: Histograms for various features

While exploring the data, the distributions of the different features and target of the data set were checked. This step is important prior to the building machine learning models as understanding the distribution of the data is key knowing whether to use parametric or non-parametric methods.

Data Modelling

This section provides a brief description of each modelling packages and models used in experiments.

Scikit-Learn Lasso & Ridge

The scikit-learn package is a standard Open-Source python library for Machine Learning models and functionality, and one from which several models were experimented with for our prototype [15].

The Lasso model is a linear model trained with L1 as the priority regularizer. If there are only a few relevant parameters and the rest are near to zero, this model tends to perform well. L1 is used for feature selection and has the ability to drop variables linked to coefficients that reaches 0.

The Ridge model was also experimented with, where Ridge is a linear model trained with L2 as the priority regularizer. If there are numerous large parameters with about the same value, this model performs well. When you have collinear/co-dependent characteristics, however, L2 is helpful.

Lasso and Ridge both have limitations in that they utilise different penalty terms, L1 and L2 respectively. Both regularization methods have differing advantages, L1 will be useful for feature selection, and for L2, if there is multicollinearity present. However, given this time series dataset, there might be insufficient features and data to be represented entirely, as well as encoding date variables to ordinal variables, but not be the best option for a time series forecast.

If there are opportunities in the future, other algorithms such as ElasticNet could be explored more which finds the harmony between these two penalties, potentially yielding a better result.

See also https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html and https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html for more information [16-17].

XGBoost Gradient Boosted Decision Tree Regression

This model uses the XGBoost library, which uses Gradient Boosted Decision Trees for machine learning [2]. While Decision Trees are often more associated with Classification problems, XGBoost uses real-value scores on the leaf nodes in the Decision Tree to be able to make use of the trees for prediction in Regression problems. In addition, XGBoost uses in-built Ensembling with Gradient Boosting to achieve accurate predictions.

See also https://xgboost.readthedocs.io/en/stable/get_started.html for more information [18].

Pmdarima ARIMA

This model uses the open source Pmdarima Python package to build Time Series Forecasting models [19]. This model uses Auto Regressive Integrated Moving Average as a way of modelling time-series data for forecasting.

ARIMA has limitations on predicting seasonal trends, further models which can be used to test on seasonal data would be an extension of the ARIMA model, the SARIMA model, which supports the direct modelling of the seasonal component in time series datasets. It also assumes the process underlying to be stationary, which means the mean and standard deviation or variance of the data are assumed to be small, this can be tested using several methods like the Dickey Fuller test and stationarity can be achieved by differencing the models.

Azure Automated Machine Learning Regression and Time Series Forecasting

Azure Automated Machine Learning is a cloud-based platform by Microsoft that can be used for building machine learning pipelines [3]. In this project, the system was used to model for Regression and Time Series Forecasting problems.

The Azure Auto ML job generated a VotingEnsemble solution that incorporated several models, including models from XGBoost [2] and LightGBM [20], applies different strategies of feature engineering including Standard Scaling and Min Max Scaling, and applies different strategies for hyperparameter tuning. Deep Learning was not configured for these experiments.

While the highest performing model for the aggregated City of Melbourne Foot Traffic predictions was the Azure Automated Machine Learning Time Series Forecasting Model, there are some other factors to consider before recommending using this model for a functioning production product [3]. Firstly, the model would need to be trained/deployed within the Azure Machine Learning platform, and therefore there would be the associated costs for the business to run the system in that environment [3].

Additionally, since the model source is not available, certain usage is not easily developed within python, making some activities such as evaluation with Learning Curves more difficult than with other models.

Findings

Using the mentioned models, a series of experiments on modelling for foot traffic numbers were carried out, and the results were evaluated and analysed to find the strongest models for both aggregated City of Melbourne Foot Traffic predictions and Street Specific Foot Traffic predictions.

Model Evaluation

To evaluate the models, K-Folds Cross Validation using 5 Folds was applied. In brief, this involved 5 iterations of modelling, where 80% of the full data was used for training the model, then the remaining 20% would be used as a validation set to predict via the model, and those predictions can be compared against the true values from the validation set. After the 5 iterations were completed, the resulting evaluation metrics are averaged based on the 5 results from the iterations.

When evaluating the results of the Regression Models, there are several metrics that can be used. In this case, a common method of evaluation was chosen, which is to consider the following two metrics:

- **Mean Absolute Error (MAE):** This represents the mean amount of error between predicted values vs true value on the validation set.
    - This metric was chosen as a simple and easily understood method of measuring the accuracy of predictions
- **Root Mean Squared Error (RMSE):** This is the root of the mean of the square of the errors between the predicted values vs true values on the validation set.
    - This metric was chosen as this gives a measure of the variance in the errors of the predicted values.

Therefore, the target is to reduce both the MAE and RMSE, to ensure accuracy of predictions without large fluctuations in the error of these predictions.

Model Results

As mentioned, the models were trained on a dataset with daily records spanning every day from 1st January 2013 to 31st July 2022. The predictions, MAE and RMSE are therefore units of total pedestrians in foot traffic.

Aggregated City of Melbourne Model Results

The following results are from the modelling of the aggregated foot traffic over Council Area of the City of Melbourne. In this dataset:

- Min Daily Foot Traffic (total_people): 10,390
- Max Daily Foot Traffic (total_people): 1,151,467
- Mean Daily Foot Traffic (total_people): 481,666.69
- Standard Deviation on Foot Traffic (total_people): 209,496.18

In the following Model Results, a basic measure of Prediction Accuracy and Prediction Error is calculated, according to the following formulas:

- Average Prediction Error = (Model MAE / Mean Daily Foot Traffic) * 100
    - Meaning: This represents the error from the Mean Daily Foot Traffic that predictions are on average, according to the Mean Absolute error (as a percentage)

    - This is a simplified attempt to communicate the accuracy of the models, where the goal is to minimise this value.
    - A more standardised way of displaying this is with the Normalized MAE.
- Prediction Error Deviation = (Model RMSE/ Mean Daily Foot Traffic) * 100
    - Meaning: The amount of deviation in predictions

Table 2 - Azure Automated ML Time Series Results

Models	MAE	RMSE	Average Prediction Error (%)	Prediction Error Deviation (+/- %)
Lasso Regression	135633.22	174650.88	28.16%	36.26%
Ridge Regression	135870.87	174671.77	28.21%	36.26%
XGBoost Decision Tree Regression	32913.73	49544.43	6.71%	10.1%
Azure Automated Machine Learning Regression	29617	44775	6.15%	9.3%
ARIMA	194924.19	207185.50	40.47%	43.01%
Azure Automated Machine Learning Time Series Forecasting	25322	29177	5.26%	6.06%

As can be seen, the best performing model for Aggregated City of Melbourne modelling was the Azure Automated Machine Learning Time Series Forecasting model [3], with the following further results metrics in Table 3.

Table 3 - Additional Azure Automated ML Results

Azure Automated Machine Learning Time Series Forecasting Model			
Mean Absolute Error	25322	**Normalized MAE**	0.025568
Root Mean Squared Error	29177	**R2 Score**	0.39597
Explained Variance	0.55090		

Street Location Specific Model Results

After modelling predictions on the total Foot Traffic on the City of Melbourne, further modelling was done for street specific locations. As mentioned, a selection of streets was selected based on consistency of data across the whole time period.

While Azure Time Series Forecasting models performed well for the Aggregated City of Melbourne dataset, the standard types of Time Series Forecast models could not be applied for models on the Street Specific data [3]. This is because this data has multiple records for each date (that is, one per street location per date), and these Forecast models required unique data records per date.

Therefore, the two strongest regression models were selected for experiments to be carried out.

In this dataset:

- Min Daily Foot Traffic by street (total_people): 0
- Max Daily Foot Traffic by street (total_people): 58,960
- Mean Daily Foot Traffic by street (total_people): 14,693.91
- Standard Deviation on Foot Traffic by street (total_people): 12,187.38

Table 4 - XGBoost vs Azure Auto ML

Models	MAE	RMSE	Average Prediction Error (%)	Prediction Error Deviation (+/- %)
XGBoost Decision Tree Regression	1147.07	2163.58	7.81%	14.72%
Azure Automated Machine Learning Regression	1225.5	2201.5	8.34%	14.98%

As can be seen, the best performing model for Street Location Specific modelling was the XGBoost Regressor model [2], with the following further results metrics:

Table 5 - Street Level XGBoost Decision Tree Regressor Model

XGBoost Decision Tree Regressor Model			
Mean Absolute Error	1147.07	Normalized MAE	0.019551
Root Mean Squared Error	2163.58	R2 Score	0.96
Explained Variance	0.96		

Results Analysis

As observed, initial results from the prototype models appeared to be quite promising, with error in predictions being low. The results of the modelling experiments were analysed, firstly to attempt to find the most important features in modelling, then to determine whether the model would generalize well when predicting on new, unseen data.

Feature Importance

The best results were produced by the Azure Automated Machine Learning Time Series Forecasting Models [3]. Based on the modelling and results of the aggregated City of Melbourne data, a Feature Importance Graph was produced in Figure 4.

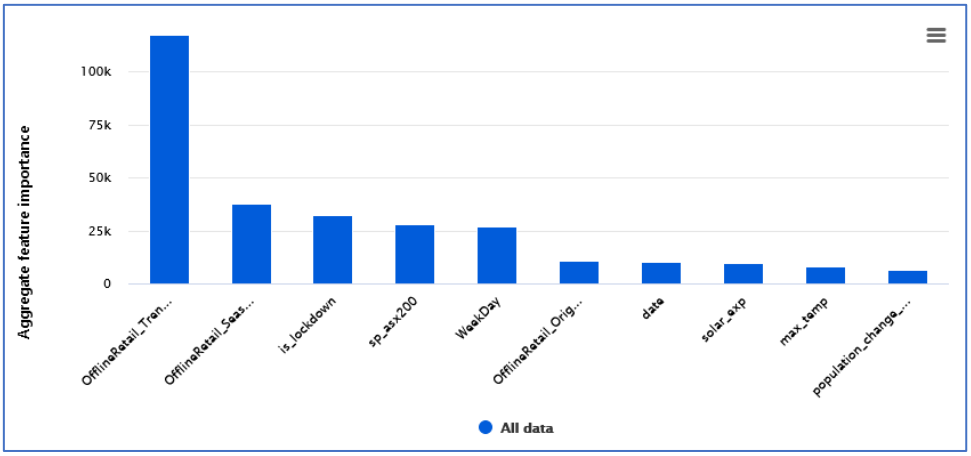

Figure 4: Feature Importance Graph for Azure ML Time Series Forecast model on Aggregated City of Melbourne data

This model found that the economic and Covid lockdown features of the dataset had the greatest correlation with the target foot traffic feature. Data such as the Monthly Offline Retail Turnover and even the S&P ASX 200 market performance were also of high importance to the model [14]. Unsurprisingly, the Covid Lockdown dates also had a large impact when predicting foot traffic numbers.

Another important feature was the day of the week. This aligns with the expectation that weekends in the city are likely to be busier with shoppers. Of the weather-based features, solar exposure and maximum temperature were the most important. Interestingly these features seemed to be considered more important than the holiday feature.

Model Bias and Variance

While the modelling results appear to be quite promising, the results were also analysed from the context of model bias and variance. In general:

- Model bias should be kept minimal to avoid building models that are too simple
- Model variance should be kept at a minimum to avoid overly complex models and overfitting.

The method used to analyse this and to determine if the models are suffering from overfitting or underfitting is through a Learning Curve graph.

As the Azure based models are built on the platform without access to source code, the XGBoost Regression model results were analysed to give overall indications of this [2-3].

Figure 5: Learning Curve Plot for XGBoost Regressor model on Aggregated City of Melbourne Data

The minimum error gap between predictions on the Training Data vs predictions on the unseen Validation Data is **MSE of 171,261.26**. Note that the Learning curve uses a different metric, Mean Squared Error (MSE), rather than Mean Absolute Error (MAE). Mean Squared Error is the standard metric for learning curves to penalize higher error values.

As can be seen by the Learning Curve and the resulting minimum error gap, the model has an issue with predictions against unseen data and is therefore the model is not generalizing well.

This is an example of an issue of overfitting**.** This result is not entirely unexpected. This prototype has only sourced some of the most basic data across the fields of weather, calendar, social and economic data. The challenge of predicting city foot traffic numbers would likely require a lot more sophisticated data to better predict.

The best recommendation for improving these models would be to source more feature data for the model. Please see the section **Further Development and Improvements > Model Accuracy and Reduce Overfitting** for more detailed recommendations.

App and Website Development

For the IOS and Android apps, the application will be built as native mobile applications. The apps will be developed using Microsoft's Xamarin Framework [21], which will allow the application to be built for both iOS and Android. Visualizations within the app will be powered by Machine Learning and Deep Learning models. The prototype is modelled on an iPhone 13 Pro Max and has dynamic stylings to adapt to all types and phone sizes. This design layout follows the good UX/UI practices, such as consistent visual and functionality, as well as an easy navigation interface and buttons to allow user to gain control of the application activity. The application features easy-to-understand terminologies and are kept short and precise.

For the website, a Python Flask [22] application with HTML, JavaScript, NodeJS [23], CSS can be used.

Customer Journey Map

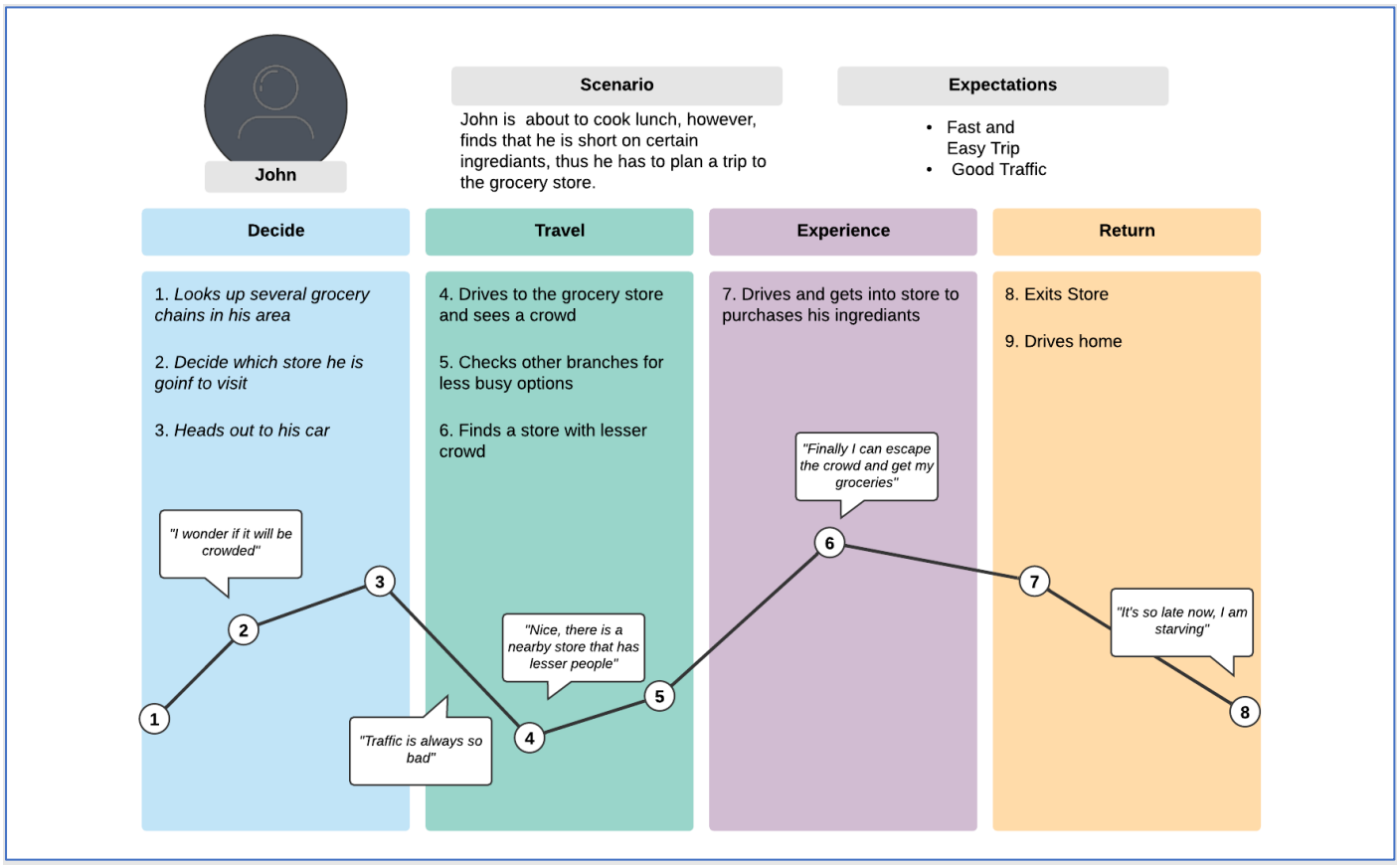

Figure 6: Customer Journey Map

Figure 6 is the created customer journey map that visualizes the end-to-end consumer experience. This persona depicts a man in his 20s, who is famished and about to cook lunch when he finds out that he is short on certain ingredients. Thus, he is wanting to visit the grocery store to get ingredients for a meal, however, he is unsure if the place will be crowded when he gets there or if the entire experience will be smooth sailing.

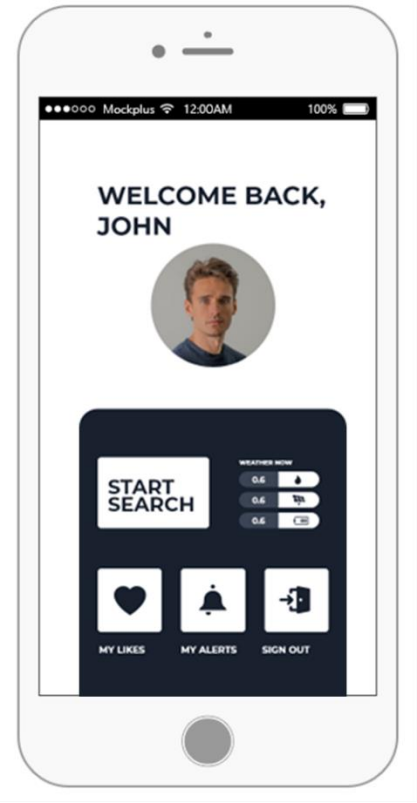

Figure 7: App Dashboard

As per Figure 7, when the user logs into their account, it leads to a profile page that has several icons. They can click on the “Start Search” icon to begin searching for the location to access the crowd watch feature. The user can also access their “Liked” locations and locations where they have set “Alerts” on for, in the case where a particular area has lesser crowds or when the crowd dissolves. Lastly, the weather data is for that day is shown beside the search button, and the chance of rain is shown in case they must prepare for wet conditions and take an umbrella out.

Figure 8: Street location search

When the user starts a search, they will be let on this “Home” page in Figure 8 where they have a choice to search for today, tommorrow or for the week. The user can type in a location in the search bar where they will be brought to the next page.

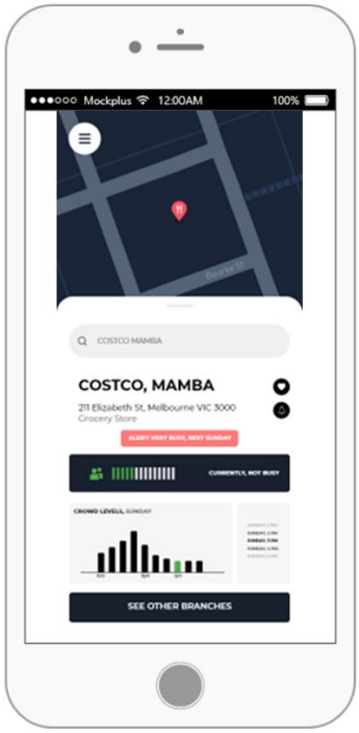

Figure 9: Foot Traffic Information

When the user enters a search, they will be brought to the “Information” page in Figure 9 which consists of several features. Firstly, the description of the location will be under the search bar and there are two icons beside that, the “Like” icon and the “Bell” icon where the user can save the place or get alerts of the crowds of that particular area. Under the description is a label that gives an alert for when this place might be busy, particularly on a public holiday. Since the place is currently not busy, it will be indicated by then green lines. And beside the graph on the crowd levels of time throughout the day, the user can adjust the time to see if it will be busier later in the day. Lastly, at the bottom of the screen, the user can click to view other branches.

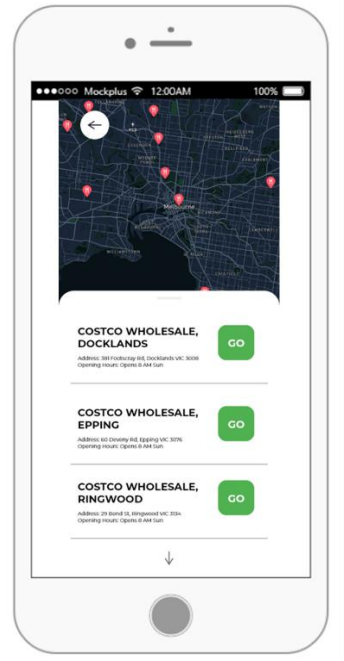

Figure 10: Location Select screen

After clicking on “See other branches”, user will be brought to this second “Branches” page as per Figure 10 where it displays the other branches in Melbourne, basic information like the store address and operating hours are stated on the listing.

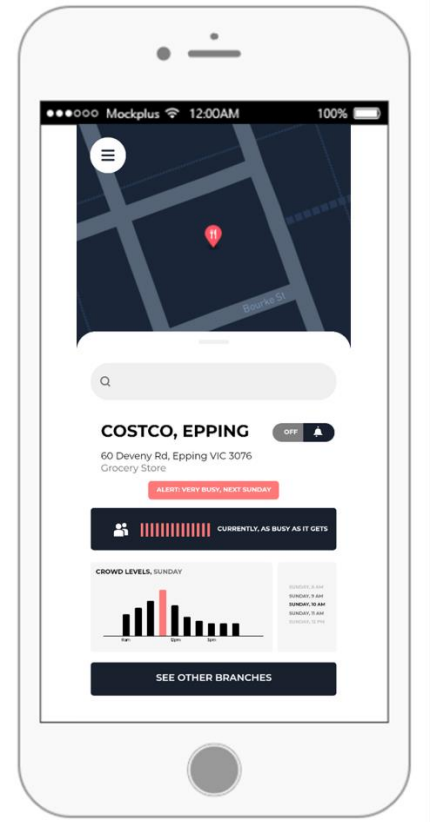

Figure 11: Other branches location

Lastly, when the user clicks on another branch, they will be brought back to the “Information” page in Figure 11. The description of the location will be under the search bar. Under the description is a label that gives an alert for when this place might be busy, particularly on a public holiday. Since the place is currently busy, it will be indicated by then red lines. And beside the graph on the crowd levels of time throughout the day, the user can adjust the time to see if it will be busier later in the day. Lastly, at the bottom of the screen, the user can click to view other branches, if they are unsatisfied with the crowd levels.

Impact and Significance of Results

It seems intuitive to most that if there are more people walking around near your business, it should translate to more revenue. However, the research seems to indicate that it is more complex than that; increased foot traffic provides the *opportunity* for higher sales, but that an increase to conversion rate (ratio of sales to customers) or basket value (average transaction value per sale) is more dependent on having appropriate staffing levels to capitalise on the extra customers [24]. To give a simplified example of a complex phenomenon; there is little benefit to having, for example, 40% more people than usual through your doors, if there is only one staff member on the floor. Customers will feel neglected, become impatient and likely leave without purchasing. It is also reasonable to assume that repeat business will also be impacted as people are disinclined risk the chance of another unsatisfactory shopping experience. The ability to predict which days are likely to be busier allows businesses to roster on sufficient staff and ensure optimal customer experience. It can avoid costly overheads such as having unnecessary staff rostered on days when the customer demand is predicted to be lower.

Bigfoot cannot tell owners how to run their businesses but anticipating your customer flow patterns allows owners to be proactive rather than reactive, like scheduling special events such as in-store only sales on predicted peak days. Events such as these often have benefits for both businesses and customers: business see a boost to sales and stock turnover, and customers feel like they have stumbled onto some good luck, that they are in on a secret or part of a community.

It is not only retail businesses that must find that delicate profit margin balance. As well as ensuring staffing is appropriate, hospitality venues face the added complication of food waste, which has both economic and environmental impacts. With the cost of produce being at the mercy of inflation and other external factors, Bigfoot's insights into potential customer numbers could help restaurants, bars and cafes walk that fine line between over- and under-stocking.

It can take years of trial and error for a business owner to "get a feel" for the ebbs and flows of customer numbers. Bigfoot aims to circumvent the potential revenue loss that can accompany this kind of experimentation, by coming pre-loaded with years of historical data. For newer businesses just starting out, or for established brands expanding to their first city location, Bigfoot can take away some of that guesswork.

While there is no doubt that the ability to shop online was immeasurably valuable during the COVID lockdowns, which has changed the way that we shop, possibly permanently, Mortimer *et. al.* [25] suggest that the age of the bricks-and-mortar store is not yet behind us. Google reports that searches for terms such as 'open now near me' increase ~400% year on year, indicating that customers still value the physical in-store experience [26].

In a world where data literacy and the use of technology are on the rise, Bigfoot could work alongside a range of other applications designed to help bricks and mortar stores stay relevant and competitive. Where Bigfoot can predict the peaks and troughs, novel 'people counting' tools such as Dôr and RetailNext enable business to monitor the effectiveness of their strategies as they respond to Bigfoot's forecast [27, 28].

It is not just commercial entities however that could benefit from Bigfoot's insights. City councils could utilise predictions to better manage timings for projects and works, as well as offer Bigfoot as a tool to help international and local tourists and visitors to get the most out of their time in the city by avoiding (or perhaps joining) the crowds. With city-specific tailoring, Bigfoot could also be a valuable tool at a streetwise level, able to offer a level of granularity such that a visitor could be warned to avoid Punt Road on days when there is a big game "at the 'G" or know where the crowds will be for the Boxing Day sales.

As Melbourne emerges out the other side of a pandemic that has had incomprehensible global economic impact, many businesses are operating on a knife's edge. Bigfoot can offer potential gains in efficiency which could mean the difference between survival and collapse. And although Melbourne is Bigfoot's birthplace and hometown there is clear scope for improvements and expansions beyond the CBD.

Further Modelling Development and Improvements

While this prototype has produced some promising initial results, there is room for improvement around model scope, more sophisticated models, model accuracy and ability for predictions to generalise with new data.

Model Scope of Areas

This model is currently only predicting according to the Foot Traffic data of the Council Area of the City of Melbourne, where the foot traffic numbers have been segmented according to certain streets within this area [1]. As such, there is great potential to expand the model to include Foot Traffic data from other council or metropolitan areas within Greater Melbourne, then other towns within Victoria and then Australia.

Model Accuracy and Reduce Overfitting

While the initial results are promising, the model accuracy can be improved. Additionally, as seen with the Learning Curve, there appears to be an amount of overfitting that has occurred in the prototype model, meaning the model's performance is suffering when predicting on new, unseen data.

The following are some recommended actions that would both improve the accuracy and reduce the amount of overfitting:

- **More Feature Data:** While sourced some base datasets for the model were sourced, clearly, this data is very general and only very basic when considered against predicting foot traffic. More data would have a definite affect for improving the model, such as:
    - **Weather Data:** for example, wind speed, wind direction, cloud cover and other data would be beneficial.

    - **Calendar and Social Data:** This is an area where there is great opportunity to add data. Some recommendations include School Holidays, Sporting/Cultural/Social events such as the Australian Open, F1 or Black Friday sales. Perhaps a way to differentially codify Public Holidays, e.g., ANZAC Day vs Boxing Day could provide starkly different retail opportunities.
    - **Economic Data:** More sophistication in retail data, segmentations of retail data by sector would be useful. Adding more macro-economic data besides market statistics would be interesting. For example, employment statistics, interest rates, or conversion rates with other currencies.
    - **Social and Infrastructure data:** Other data would be useful if it is possible to source, such as the concentration of construction jobs or public works, the proximity of locations to parks, markets, retail or hospitality businesses. Tourism related data such as incoming and outgoing flight numbers, hotel occupancy and other statistics would also be of interest.
- **Further Experimentation on Model Improvement:** Further experimentation could be conducted on our modelling process, such as further feature engineering, data transformations, and hyper-parameter tuning.
- **Better Models:** Further experimentation can be carried out by other modelling techniques and different packages. Enabling Deep Learning on the Azure Automated Machine Learning platform might also yield improved results [3].
- **Ensembling:** A powerful technique to increase accuracy and to reduce overfitting is to apply ensembling. While the Azure Machine Learning models are all Ensembles, other Model Ensembles could be created by stacking the other models that were created in order to create new Models that would likely have better accuracy and be able to generalise better on new, unseen data [3].

Geospatial Machine Learning

Another possibility of improvement would be to move on to more sophisticated modelling techniques. An area of great interest is in geospatial machine learning. Currently, we have target data based on specific street locations. Research and experimentation are recommended for geographical based modelling that will be able to predict on generalized areas rather than just these specific locations.

In addition, moving to geospatial machine learning may allow for the interpolation of predictions. The simplest case of interpolation is the ability to predict according to a location spatially between two different points, where target data is known for these two points.

Some methods and libraries that may be good options for this include:

- Spatial Analysis and Interpolation using GIS tools and spatial data. The premise to explore would be whether data curated could be sourced or represented in a GIS spatial data format. For example, some weather data from the BOM can be sourced in GIS format. Then, prediction and interpolation models could be trained using standard Regression techniques, some of which have already been used in this report, such as Lasso, Ridge and XGBoost [2].
- **PyTorch TorchVision** is a prime possible solution to be explored [29]. PyTorch open-source machine learning library and TorchVision is the package with model architecture for computer vision modelling [29].
- **TorchGeo** would be another library to be explored [30]. TorchGeo is a python library built using PyTorch with transforms and pre-trained models specific to geospatial data [30].

Project Management

The project was managed using several Agile tools and strategies. The team met on a weekly basis, every Wednesday at 6:30 PM, during which ideas, challenges, progress, and steps moving forward were discussed. A Jira board was used to lay out tasks and ideas and to assign task responsibilities to team members. In addition, the weekly catchups and communication over the Microsoft Teams chat was also used which was found to be more direct, flexible, and convenient.

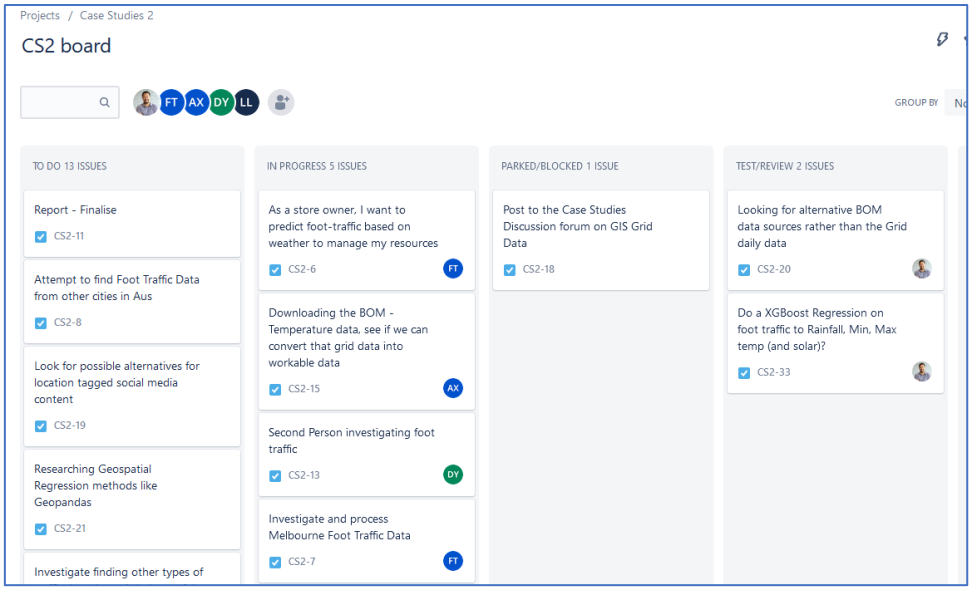

Figure 12: The team's Jira board

A GitHub repo was used to store data and share code. The desktop app Sourcetree was also utilised for a good GUI interface to Git. It was here that the code for data analysis, transformation and predictive modelling was iteratively developed.

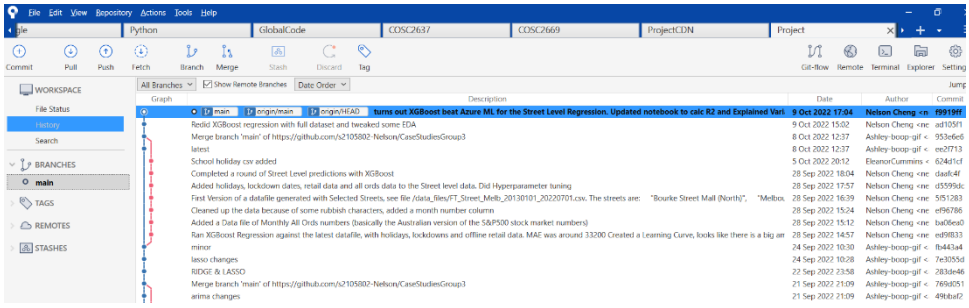

Figure 13: The team's GitHub

Using the Agile scrum methodology, the project was managed through a concept of 'sprint' to set overarching targets, providing a useful way to divide development periods. The team laid out a twelve-week plan to develop the product which matured with progress. This can be seen in Figure 14.

Project Activities	Wk 1	Wk 2	Wk 3	Wk 4	Wk 5	Wk 6		Wk 7	Wk 8	Wk 9	Wk 10	Wk 11	Wk 12
Group formation													
Idea Generation													
Sprint 1													
Collect data													
Prepare Data													
Explore Data													
Sprint 2													
Modelling													
Web and IOS demo													
Sprint 3													
Oral presentation													
Report													

Figure 14: Gantt chart showing development plan

Sprint 1

With a rudimentary conception of the foot-traffic predictor, the initial focus was to collect the foot-traffic and weather data and understand how it could be wrangled into a useable format. Here as well, ideas for what other sorts of data beyond weather were identified as useful features. The data was cleaned and explored, then stored on GitHub.

Sprint 2

Having collected the data and grounded the concept for Bigfoot, this sprint was focussed on testing and developing the models and graphical user interfaces (GUI) demo.

Sprint 3

The final sprint was dedicated to collating the team's progress into an oral presentation and report to share with stakeholders. It was useful to develop elements for the presentation as the model and GUI demos were being built. Conversely, it was also useful to continue improving the product as the presentation was being made. Hence, there was overlap with the previous sprint period.

Conclusion

As Australia is looking to move out of the COVID-19 recession, the State of Victoria is implementing strategies to aid the City of Melbourne to increase local economic activity and travel on a path of growth and development. Some common challenges for all retail and hospitality businesses are questions of stock management, resource utilisation and rostering.

Prototype models were created for Bigfoot to predict on both the aggregated foot traffic levels for the City of Melbourne as well as some street specific locations. The results from these models are promising for an initial prototype. However, the models have observable overfitting, resulting in the model predictions not generalizing well to unseen data.

This issue is not unexpected, as these prototype models are built on a relatively small data set, while one might theorise that foot traffic levels in Melbourne City and at specific points will likely be dependent on a much larger myriad of factors. There are many opportunities to improve this product with more data, different modelling techniques and a larger scope of functionality.

In conclusion, Bigfoot looks to address the common challenges for City of Melbourne customer-serving businesses by enabling people, businesses and organisations to see indications of the active CBD population through foot traffic number predictions. Businesses will then have a more informed position when answering these questions by having an indication of whether local foot traffic to their businesses is expected to be busier or quieter than usual in the coming days.

References

[1] City of Melbourne, *Pedestrian Counting System*, City of Melbourne, 2022. [Online]. Available: http://www.pedestrian.melbourne.vic.gov.au/#date=24-08-2022&time=17

[2] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, ACM, 2016. [Online]. Available: https://github.com/dmlc/xgboost

[3] Microsoft Azure. "Automated Machine Learning". Microsoft. https://azure.microsoft.com/en-au/products/machine-learning/automatedml/ (accessed September 5, 2022).

[4] City of Melbourne. "Economic Development Strategy 2031." Melbourne.vic.gov.au. https://www.melbourne.vic.gov.au/about-council/vision-goals/Pages/economic-development-strategy-2031.aspx (accessed September 26, 2022).

[5] OECD Data. "Average wages." OECD.org. https://data.oecd.org/earnwage/average-wages.htm (accessed September 26, 2022).

[6] City of Melbourne. "Melbourne is open!", Melbourne.vic.gov.au. https://www.melbourne.vic.gov.au/about-melbourne/melbourne-is-open/Pages/covid-19-recovery.aspx (accessed September 26, 2022).

[7] Australian Bureau of Statistics. "Consumer Price Index." ABS Website. https://www.abs.gov.au/statistics/economy/price-indexes-and-inflation/consumer-price-index-australia/latest-release (accessed September 21, 2022).

[8] Australian Government - Bureau of Meteorology. "Australia's climate change datasets." BOM Website. http://www.bom.gov.au/climate/change/hqsites/about-hq-site-data.shtml (accessed August 10, 2022).

[9] Macrotrends. "Melbourne, Australia Metro Area Population 1950-2022." Macrotrends.net. https://www.macrotrends.net/cities/206168/melbourne/population (accessed September 21, 2022).

[10] Lockdown Stats Melbourne. "The timeline." Lockdown Stats Melbourne. https://lockdownstats.melbourne/timeline/ (accessed September 9, 2022).

[11] Australian Government, *Australian Public Holidays Dates Machine Readable Dataset*, Data.gov.au, 2022. [Online]. Available: https://data.gov.au/dataset/ds-dga-b1bc6077-dadd-4f61-9f8c-002ab2cdff10/details?q=

[12] Time and Date. "Holidays and Observances in Australia in 2013." Time and Date. https://www.timeanddate.com/holidays/australia/2013 (accessed September 21, 2022).

[13] Australian Bureau of Statistics. "Retail Trade, Australia". ABS Website. https://www.abs.gov.au/statistics/industry/retail-and-wholesale-trade/retail-trade-australia/latest-release#data-download (accessed October 9, 2022).

[14] ASX. "Historical market statistics." ASX. https://www2.asx.com.au/about/market-statistics/historical-market-statistics (accessed September 21, 2022).

[15] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[16] F. Pedregosa *et al.* "sklearn.linear_model.Lasso Documentation." Scikit-learn Website. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html (accessed September 10, 2022).

[17] F. Pedregosa *et al.* "sklearn.linear_model.Ridge Documentation." Scikit-learn Website. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html (accessed September 10, 2022).

[18] XGBoost Developers. "Get Started with XGBoost." dmlc XGBoost. https://xgboost.readthedocs.io/en/stable/get_started.html (accessed September 10, 2022).

[19] Python Community. "Pmdarima 2.0.1." Python Software Foundation. https://pypi.org/project/pmdarima/ (accessed September 17, 2022).

[20] LightGBM. “Welcome to LightGBM’s documentation!” Microsoft Corporation. https://lightgbm.readthedocs.io/en/v3.3.2/# (accessed September 5, 2022).

[21] Microsoft. “Xamarin.” Microsoft. https://dotnet.microsoft.com/en-us/apps/xamarin (accessed October 17, 2022).

[22] Pallets. “Flask.” Pallets. https://github.com/pallets/flask (accessed October 17, 2022).

[23] NodeJS Contributors. “NodeJS.” Open JS Foundation. https://nodejs.org/en/ (accessed October 20, 2022).

[24] O. Perdikaki, S Kesavan, J. M. Swaminathan, “Effect of Traffic on Sales and Conversion Rates of Retail Stores,” *Manufacturing & Service Operations Management,* vol. 14, no. 1, Aug. 2011, doi: 10.1287/msom.1110.0356

[25] G. Mortimmer, L. Grimmer and P. Maginn, 2020. “The suburbs are the future of post-COVID retail”. The Conversation. https://theconversation.com/the-suburbs-are-the-future-of-post-covid-retail-148802 (accessed October 9, 2022).

[26] A. Thygesen. “2022 Retail Marketing Guide: Drive foot traffic and in-store sales.” Think With Google. https://www.thinkwithgoogle.com/consumer-insights/consumer-journey/increase-foot-traffic-and-in-store-sales/ (accessed October 19, 2022).

[27] Dôr. “People Counter.” Dôr. https://www.getdor.com/solutions/people-counting (accessed October 19, 2022).

[28] RetailNext Inc. “RetailNext.” RetailNext Inc. https://retailnext.net/ (accessed October 9, 2022).

[29] The PyTorch Foundation, 2022. “PyTorch”. The Linux Foundation. https://pytorch.org/ (accessed October 17, 2022).

[30] Microsoft, 2022. “TorchGeo”. Microsoft. https://github.com/microsoft/torchgeo (accessed October 17, 2022).