UNIVERSITY OF CANBERRA

Assignment Cover Sheet

Unit No	**8696**
Unit Name	**Data Analytics and Business Intelligence**
Group No.	**Group 3**
Project Title	**Kai Mook Thai Restaurant Business - Weather Analysis**

We declare that this assignment is solely our work, except where due acknowledgements are made. We acknowledge that the assessor of this assignment may provide a copy of this assignment to another member of the University, and/or to a plagiarism checking service whilst assessing this assignment. We have read and understood the University Policies in respect of Student Academic Honesty.

Student ID							
U	3	2	1	0	5	9	0
U	3	2	0	7	7	6	7
U	3	2	0	1	8	6	3
U	3	2	1	3	4	9	6
U	3	2	0	6	2	0	1

Executive Summary

Kai Mook Thai Restaurant is located in Lower Plenty, Melbourne. Melbourne is known for its unpredictable weather which has prompted the owners of Kai Mook to investigate whether their sales are affected by rainfall. Understanding the likelihood of rainfall will allow them to efficiently roster staff in the appropriate areas at the appropriate times as understaffing and overstaffing can negatively affect daily sales. During typical day-to-day operations, one chef, one kitchen-hand, and one front-of-house staff are rostered. At a starting rate of ~$30 per hour, inefficient rostering can cost Kai Mook up to ~$150 per day, per employee. This is a considerable loss as the restaurant only operates from 5:00pm till late. Kai Mook also offers food delivery services such as UberEats, Menulog, and Deliveroo. An understanding of rainfall will also allow them to capitalise on the increased demand of these services during rainy periods.

Sales data was provided by Kai Mook detailing their daily sales, differentiating from in-store sales and delivery service sales. The weather data was sourced from the Bureau of Meteorology and was collected by Viewbank weather station which is roughly 1.1km from Lower Plenty. It details several attributes regarding climate including minimum and maximum temperatures, evaporation, the amount of clouds and sunshine present, and the pressures and humidities at 9am and 3pm. The sales data saw minimal preprocessing. Many NAs were present due to Kai Mook being closed, which meant that these values were 0. For the weather data, the few NAs that were present were imputed using the monthly averages. A correlation matrix was used to determine the variables most closely related to rainfall and total sales. These variables were then appended into a master data frame.

Modelling was conducted using logistic regression, and decision tree classifiers. Each model achieved an accuracy of 76% and 79% respectively. These models however, failed to meet the 80% accuracy threshold determined by Kai Mook. Despite this, the models were able to reveal a negative relationship between rainfall and total sales. However, when we separate in-store sales and delivery service sales, there is a positive relationship with rainfall and delivery service sales, particularly UberEats. Under the assumption that Kai Mook would utilise this model, they would be able to efficiently roster staff according to the likelihood of rainfall. For example, Kai Mook may consider operating during lunch hours during rainy periods whilst only offering food delivery services. This would allow them to maximise their daily profits whilst minimising the cost of daily wages as only one front-of-house employee would be needed to prepare orders for the food delivery drivers. Also, given the nature of these food delivery services, Kai Mook can actively open and close their online restaurant, giving finer control over their daily profits and expenditures.

Background

The project team has consulted a small restaurant business “Kai Mook Thai Restaurant", located in Lower Plenty, Melbourne. The data that the team has acquired and is used in this report is all legally obtained through communications with the business owners with no restrictions imposed. As the name implies, Kai Mook is a small Thai restaurant which seats up to 50 patrons with operational hours between 17:00 till late. Kai Mook offers dine-in, pick-up (takeaway) and outsourced delivery services utilising UberEats, Menulog, and Deliveroo. Kai Mook’s accountant was able to provide us with almost 23 months of sales data along with detailed sales between individual outsourced delivery services and in-store based sales.

Kai Mook typically has a chef, a kitchen hand and an employee working front-of-house taking orders, but usually requires more staff depending on the day. They need to ensure that the rostering is as efficient as possible as the cost of over-rostering for just one employee for the evening will be $150+ ($30 per hour incl. super) which can be costly overtime for a small business. Opportunity costs are more costly as under-hiring staff means that the restaurant can not keep up with orders, and the restaurant in the past have had to turn off their delivery services (UberEats, etc.) just to keep up with in-store demands. This can cost the restaurant up to $3000 per week in lost revenue if under-hiring occurred daily. This means that the restaurant will benefit financially and socially from knowing which specific staff it needs on the day by the model which aims at predicting rain occurrence.

Introduction

Business Geographical Problem

Kai Mook restaurant is located in Melbourne, which is well known to have variability in its weather, which some even describe as having “Four Seasons in One Day” (Reid 2017). Freak storms, heat waves and sudden drops in temperature of more than 20 degrees celsius are not uncommon as the city’s geographical position is at the intersection of a vast hot continent and the southern ocean. This results in an interaction between different air masses, as there is a need for a lot of energy to push a large body of warm air. Because of the collision of different air masses, there is a scenario where the city can observe a 15 or more degree celsius shift down in the weather within the hour (Burt 2019). According to Professor Iam Simmonds, from the School of Earth Sciences in the University of Melbourne, just because Melbourne’s weather is very variable, doesn't mean it's not predictable as there is an important distinction between predictability and variability. He also hints that there is a probable relationship between the Antarctic storms and the cold fronts Melbourne experiences (Simmonds 2019).

Melbourne has quite stable rainfall over extended periods of time, ranging from 45mm to 65mm on average during the wet months, with 139 days of rain per year (BOM, 2021). This

means that Kai Mook Thai Restaurant will be experiencing a significant amount of rainfall which the model predicts can have an effect on sales.

Data Modelling Goal

The model will combine the sales data with the matching 23 months of rainfall data from a nearby weather station in Melbourne, provided by the Bureau of Meteorology. This allows a model to be created which can help the restaurant management predict whether it will rain on that given day, if at all. This may influence customer demand on any given day during operations which will aid the daily rostering for the business. The project also hopes to achieve a model yielding 80% accuracy. This threshold was agreed upon by the business owners after consideration of their risk tolerance.

It is hypothesised what sales will initially form a weak negative correlation against rainfall due to noise and environmental factors that cannot be accounted for. Through cleaning and refining, a prediction model yielding 80% accuracy will be attainable and delivered to the business.

Data

Although full permission is granted for access to the restaurant sales data, a nearby weather tower was located and the team was able to source 14 months of raw weather data to complement the project brief (BOM 2022). The weather data set records daily weather conditions including the date, rainfall (in millimetres), evaporation, minimum temperature, maximum temperature, sunshine hours, direction of maximum wind gust, speed of maximum wind gust, time of maximum wind gust, 9am temperature, 9am humidity, 9am cloud amount, 9am wind direction, 9am wind speed, 9am pressure, 3pm temperature, 3pm humidity, 3pm cloud amount, 3pm wind direction, 3pm wind speed, 3pm pressure. It consists of 422 observations and 22 variables. The data set contains very few NA values. A histogram of rainfall indicates that rainfall is heavily skewed to the right. The box plot also reveals numerous outliers. We can also see that the median rainfall is 0 and that the mean rainfall is 1.95mm.

The sales data set consists of 414 observations and 9 variables. This data set records daily sales and details the source of sales including UberEats, Menulog, Deliveroo sales, and cash and eftpos sales. A histogram of sales reveals that the data is also skewed to the right, but we can see a slightly normal distribution. The box plot also reveals numerous outliers. The mean sales is $1814.42 and the median sales is $1518.18.

Pre-processing for data modelling techniques

To support the productivity of the pre-processing and modelling procedure, it came to the consensus that both R and Python will be applied to this project. To create a master dataset which aims at diving into the relationship between the rainfall and the business revenues, R would be an appropriate language to start off.

Firstly, the variable names for both data sets were renamed to ease pre-processing. Next, the date columns of both data sets were formatted to the “Date” format. This allowed us to extract the month and the year and create each variable respectively. Using these two variables, a new date variable in the format “Jan 2000”, was created. This variable will also allow the creation of time-series graphs, and eliminates the need to use the gather() function in later stages of the project. The eight most recent observations in the weather data set were removed as the sales data set only records data eight days prior. The NAs in the weather data set were imputed by grouping the data by the year and month, and calculating their respective means.

Two separate csv files were used to create the sales data set. One recorded sales data for 2021 and the other for 2022. The 2022 file contained an extra variable recording eftpos sales including the surcharge. This variable was the “Card” variable multiplied by the surcharge (1.011%). This variable was removed as it was deemed unnecessary. The sales data contained numerous NA values. The majority of these NAs occurred in the “Uber”, “Menulog”, and “Deliveroo” variables. These NAs exist due to the restaurant being closed so it makes sense intuitively to replace them with 0.

Next we combined the weather and sales data sets. The “Uber”, “Menulog”, “Deliveroo”, and “total” variables were appended to the weather data to create a master data frame. The “rain_today” and “rain_tmrw” variables were encoded into binary values. Then, we removed the observations where total sales equals zero as this does not reflect the effect of rainfall on sales. Finally, we created a subset of the master data frame containing only numeric variables to allow for the creation of a correlation matrix. After constructing all the necessary plots, a .csv file of cleaned data will be launched for Python data execution by the write.csv() function.

Using Python, the data pre-transformation can now turn to be more specific for the modelling purpose. To kick off the pre-transformation process, relevant libraries such as numpy, Pandas, Matplotlib.pyplot, seaborn, SimpleImputer, StandardScaler, LabelEncoder, train_test_split, LogisticRegression , DecisionTreeClassifier, confusion_matrix, classification_report, roc_auc_score, and accuracy_score, confusion_matrix, roc_curve, auc should be implemented. Then, the cleaned data from R would be imported into the notebook. To narrow down the attributes which are unnecessary for the model, some irrelevant columns such as 'Unnamed: 0', 'date','month’ have been removed by the “drop” function. Because only numeric variables would be acceptable for modelling, LabelEncoder was built-in to transform all the categorical variables including 'tm_max_wind', '9am_wind_dir', '9am_wind_spd', '3pm_wind_dir', and '3pm_wind_spd' into numeric value type. Accordingly, those values were made ready to be evaluated in the modelling section.

In the data visualisation section, Seaborn has been used to plot count plots between ‘yr_month’ and ‘rain_tmrw’, ‘rain_today’ and ‘rain_tmrw’ for testing possible data trends and rain_today-rain_tmrw relationship which can be used in the predictive model. Also, a pair plot was established to provide a quick look on the interactions amongst variables in which the scatter points were presented as hue (colour encoding) - ‘rain_tmrw’ values (‘0’ and ‘1’). After considering the pairplot, scatter plots were set up between 9am temperature and 3pm temperature, 9am humidity and 3pm humidity, 9am wind speed and 3pm wind speed, Max temperature and Min temperature to have a closer insight at the relationship considered for the models based on the rain_tmrw attribute (colour encoding - hue).

Turning now to the features engineering part, the mean would be used for representing the similar variables. In detail, 'Wind_spd' calculated as the mean of 9am wind speed and 3pm wind speed became a new variable replacing 9am wind speed and 3pm wind speed. The same case happened for ‘Hum’ - a new variable for 9am humidity and 3pm humidity, ‘Temp’ - a new variable for 9am temperature and 3pm temperature, ‘Pres’ - a new variable for 9am pressure and 3pm pressure. After that, the variables which were already replaced would be removed from the current dataset. After variable encoding and features engineering, a new dataset which comprises only numeric data was created as weather_data_num.

In modelling procedure, outlier handling can be regarded as an indispensable stage to enhance the model accuracy. Again. Seaborn would be recommended to plot distributions of the whole dataset variables. That would not be sufficient to focus only on the distributions, hence a combined figure of box whisker plots was created to test the outliers. Then, it would be of high importance to construct a correlation heatmap for deciding on which variables to ignore in case most of them contain outliers. This action was taken to prevent a huge loss of data since the weather data is not large enough. After identifying the variables that need to be considered for outliers, Quantile 1 (Q1) and Quantile 3 (Q3) would be calculated by the numpy library, specifically the percentile function at 25% and 75%. Having known Q1 and Q3, calculating the value of IQR and deduced the values of Min and Max. The outliers could be smaller than Min or bigger than Max, therefore the`box-whisker plots would be helpful to determine the area of outliers should be subtracted. When the outliers were successfully excluded, a new correlation heatmap would be established for the modelling variables chosen process.

Analysis

The modelling procedure will be based on Logistic regression and Decision tree classifiers to predict the rain tomorrow. To begin with, the chosen variables names are framed into a list called num_vars (numeric variables). Then, the list would be used to extract the corresponding variables from the main table weather_data_num to be a new table assigned to x. Accordingly, x already comprised all the explanatory variables, so there should be an y to include the target variable rain_tmrw (rain tomorrow). After that, the standard scaler was used for scaling x variables which ended up a new scaled table x_scaled. Next, the train_test_split was utilised for partitioning the dataset into different data pieces y train, y test, x train and x test with train_test_split configured as x_scaled, y, test size of 0.25 and random state at 0.

Turning now to the first model which is the Logistic regression, the maximum number of iterations would be set as 200. Then, x train and y train would fit into the logistic model. After fitting, the x train will be predicted for identifying the y train prediction. After that, x test would also be predicted to be predicted for y pred. By using accuracy_score for y test and y pred, the model accuracy would be printed. Also, a classification report with multiple metrics such as precision, recall, F1-score and support would be launched with the application of the library classification_report. To build up the heatmap for the confusion matrix, the seaborn library would also be used with y test and y pred. The ROC curve was depicted by different measures such as y probability train (y_prob_train), y probability test (y_prob_test), x train, x test, false positive rate train (fpr_train), false positive rate test (fpr_test), true positive rate train (tpr_train), true positive rate test (tpr_test), thresholds_train, and thresholds_test. Firstly, x_train would be implemented in the function predict_proba to get y_prob_train, and x_test would be applied into predict_proba to obtain y_prob_test. Afterward, the y_train and y_prob_train were implemented into the roc_curve function to figure out the fpr_train, tpr_train, and thresholds_train. However, the statistics which are needed to draw the ROC curve for test data are fpr_test, tpr_test, and thresholds_test. Therefore, y_test and y_prob_test were applied into the roc_curve function to gain these values. Finally, the ROC curve would be printed out for test data by setting up the fpr_test and tpr_test as the plot metrics. Since fpr_test and tpr_test were available during the ROC curve set-up, they could also be used for AUC by the use of auc function with fpr_test and tpr_test being included.

Moving onto the Decision tree model, all the metrics for the model would be set as default. X train and y train would fit into the decision tree model, which will be the only difference as compared to the modelling procedure of the Logistic regression model. Afterward, the modelling and metrics evaluation procedure would be similarly applied on the dataset as how they are implemented in the case of Logistic regression model. Therefore, replication of the above process would provide all the items including the accuracy rate, precision, recall, F1-score, AUC, confusion matrix, and roc curve for the model.

After considering the above models, they have the same procedure to build up as well as find all the relevant metrics for the evaluation. Therefore, this procedure can be iterated for the optimised decision tree. However, to optimise the decision tree, the metrics should be configured inside the decision tree classifiers are criterion="entropy" and max_depth=30. Except for this configured part, the rest of the procedure would stay the same as depicted in the Logistic regression model.

To have better insights into the decision tree model (optimised version), it was required to take a look at the decision tree visualisation. The first step is to import and install if needed all the relevant libraries which are StringIO, Image, export_graphviz, and pydotplus. Afterward, set up the function export_graphviz with the metrics such as decision tree model, out_file (dot_data), filled (True), rounded (True), special_characters (True), feature_names (num_vars), class_names (['0','1']) being configured. The graph will then be established by the application of pydotplus and written down as a png file. To illustrate the graph, image() function would be utilised in this case.

Results

In the results section, a classification report has been established to evaluate the model in detail. Meanwhile, a confusion matrix and ROC curve with AUC score are also of high importance to illustrate the model performance. By the use of these evaluation methods, the chosen models can be compared to put forward the most efficient model for rainfall prediction.

Modelling techniques	**Precision**	**Recall**	**F1-score**	**Accuracy**	**AUC**
Logistic regression	0.80	0.92	0.86	0.76	0.75
Decision tree	0.76	0.72	0.74	0.60	0.46
Decision tree (Optimization)	0.87	0.87	0.87	0.79	0.70

Figure 1: Metrics table

As is depicted by the table, precision, recall, F1-score, accuracy and AUC reach 0.80, 0.92, 0.86, 0.76 and 0.75 in the logistic regression model respectively. The Recall value is significantly higher than other metrics which should be considered. Overall, the metrics are over the threshold of 0.70.

The table indicates that the values for precision, recall, and F1-score are presented as 0.76, 0,72, and 0.74 in the decision tree respectively. Accuracy exhibits a relatively low score of 0.60. The AUC score (0.46) is seen as being under the threshold value of 0.50. .

Considering the classification statistics of the optimised decision tree, precision, recall, and the F1-score have the same score (0.87), and accuracy is 0.79. However, the AUC is far higher than the threshold of 0.50 with the value of 0.70 in this report.

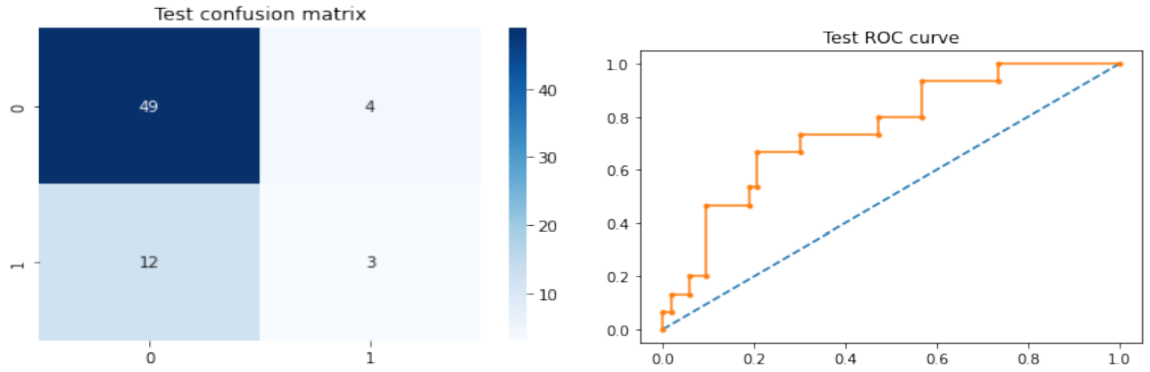


Figure 2: Confusion matrix and Roc curve for Logistic regression

The confusion matrix comprises the values of 49, 4, 12, and 3 for true positive, false positive, false negative and true negative in that order. The highest belongs to true positive, while the lowest value is at true negative position. Also, false negative is also the second highest with the value of 12.

Turning now to the ROC curve visualisation, the curve line has been seen being inclined toward the top left corner.

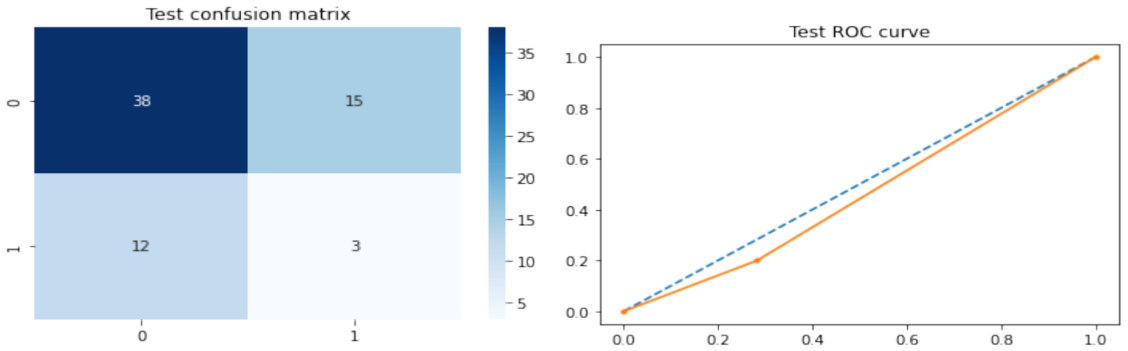


Figure 3: Confusion matrix and Roc curve for Decision tree

The values of 38, 15, 12, and 3 are recognized as True positive, False positive, False negative and True negative in the order given. In the matrix, False positive and False negative are also recorded as 15 and 12 for each metric. The highest and least values for this confusion matrix are 38 and 3 respectively.

Considering the ROC curve, it can be seen that the line leans slightly to the right side of the dot line.

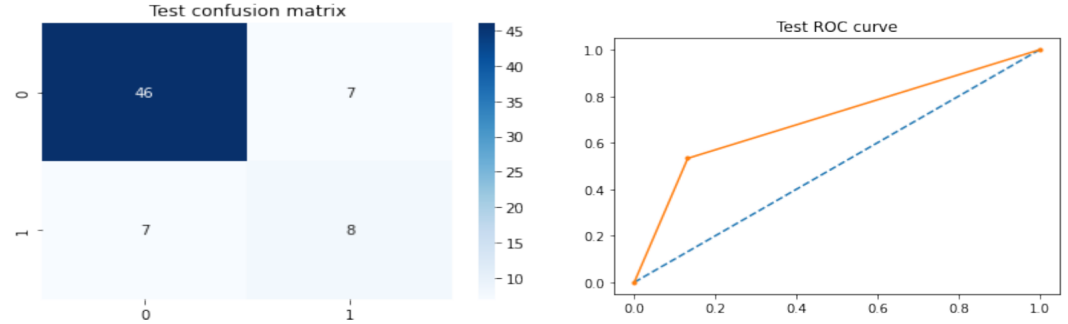


Figure 4: Confusion matrix and Roc curve for the optimised decision tree

As for the confusion matrix, it illustrates that 7 is seen at both of the metrics False positive and False negative. Moreover, True positive and True negative are displayed as 46 and 8 correspondingly.

Regarding the ROC visualisation, its curve line has been seen as moving more towards the left side corner of the graph.

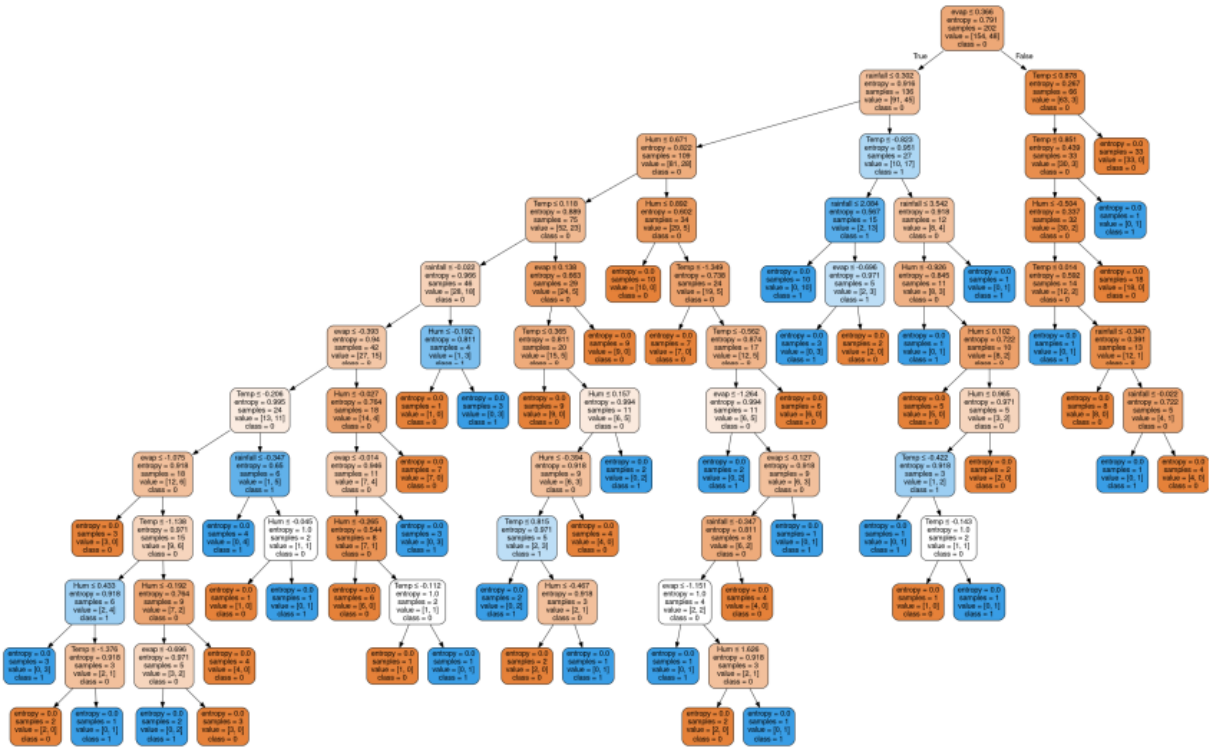


Figure 5: Decision tree visualisation

Based on figure 8, when evap is less than 0.34 it's more likely to rain. On the other hand, if the temperature is less than 0.88 it's more likely to rain as well as Humidity if it's less than 0.8. Moreover, evap was later found if it's less than -1.03 it's more likely to rain.

This decision tree was generated using 4 variables for modelling purposes, as a visualisation this causes a lot of noise and makes it difficult to draw conclusions. However, at a higher level it is still an interesting and noteworthy visualisation to include considering the choice to model decision tree structures.

- **Apparent trends in data are Identified:**

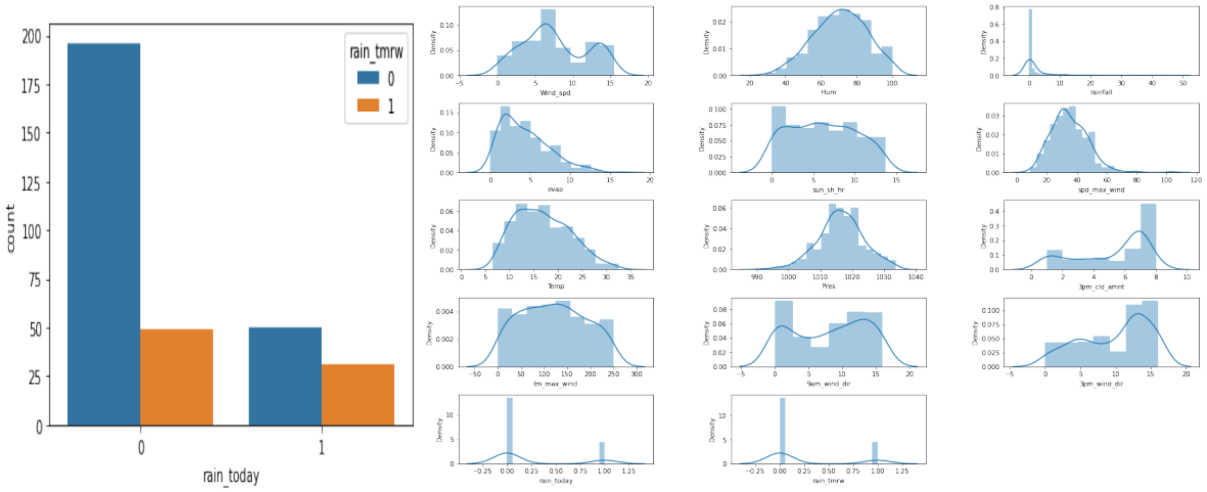


Figure 6: Decision tree visualisation Figure 7: Numeric variables distribution of the dataset

Figure 6 has demonstrated that when rain_tmrw is true, the rain today is extremely high. When rain today is low, there is low occurrences of rain tomorrow.

Figure 7, evap, spd max wind and temp are right skewed while wind speed is non symmetric bimodal distribution. On the other hand, Pres is a normal distribution while tm_max_wind,3pm_wind_dr and sun_sh_hr are uniform distributions. 3pm_cld_amnt and Hum are left skewed.

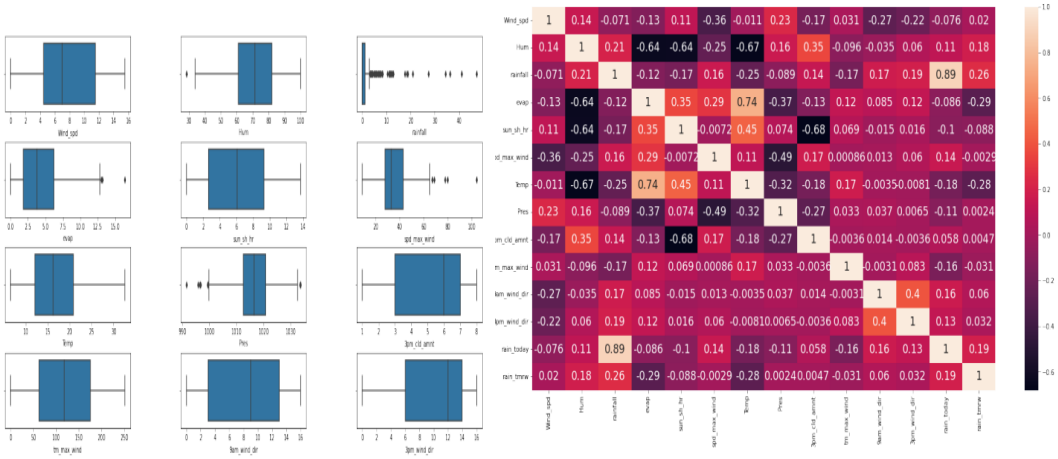


Figure 8: The box whisker plots of variables Figure 9: The correlation heatmap of the dataset variables

Based on figure 11, rainfall has many outliers while evap has few and Hum has one. Outliers were detected by using interquartile range and then removed to get the final dataset.

Based on figure 12, evap appears to have the strongest correlation to our target rain tomorrow while temp came second followed by rain_fall, rain_today and Hum while wind_spd and tm_max_wind appeared to have the lowest correlation. These strongest correlation features were chosen to train the models.

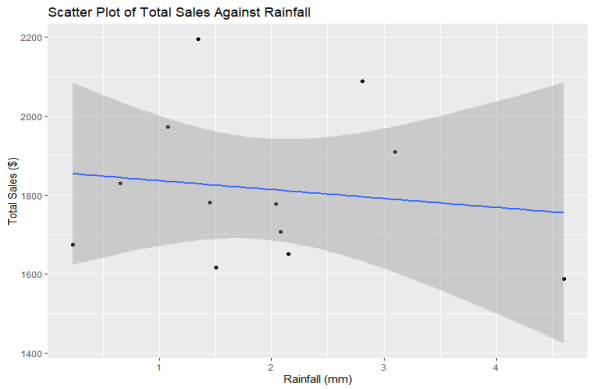


Figure 10: The scatter plot for total sales-rainfall

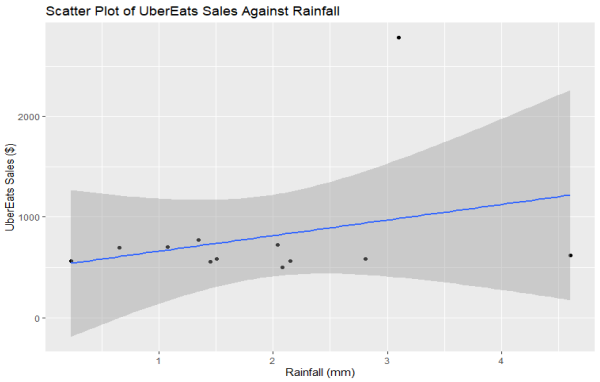


Figure 11: The scatter plot for UberEats-rainfall

Figure 10, total sales has a low negative correlation with the rainfall with few possible outliers.
Figure 11, UberEats has a low positive correlation with rainfall and there is one possible outlier which will be removed.

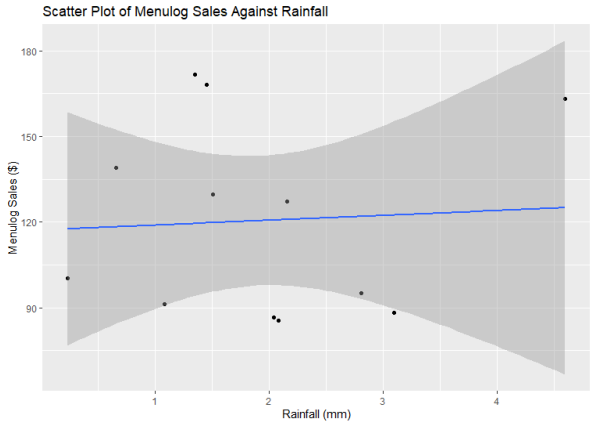


Figure 12: The scatter plot for menulog-rainfall

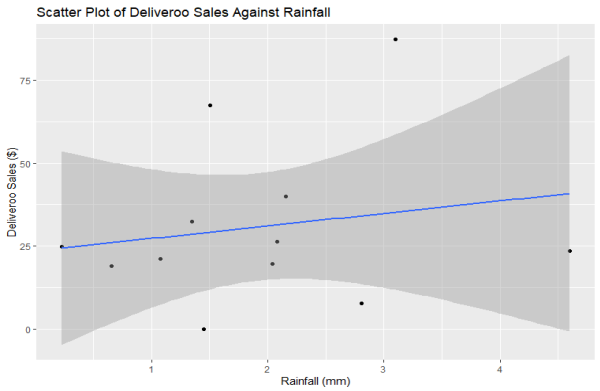
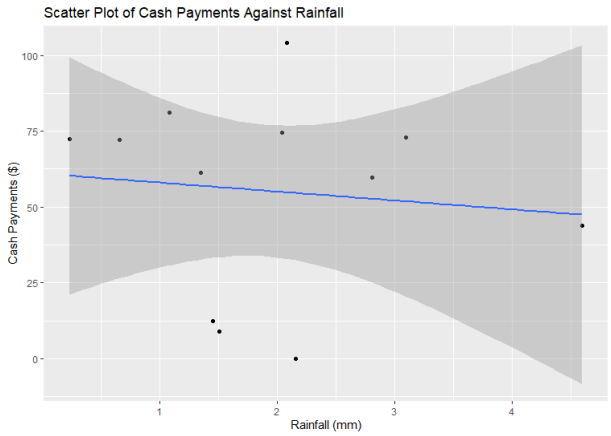


Figure 13: The scatter plot for deliveroo-rainfall

Figure 12, menulog has a low positive correlation with rainfall and there are few possible outliers which will be removed.
Figure 13, deliveroo has a positive correlation with rainfall and there are few possible outliers which will be removed.



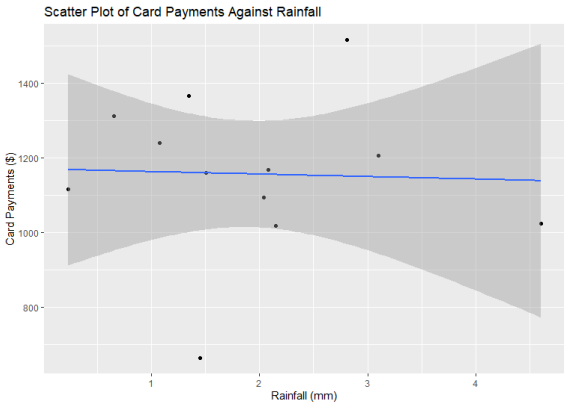


Figure 14: The scatter plot for cash payment-rainfall

Figure 15: The scatter plot for card payment-rainfall

Figure 14, cash payment has a negative correlation with rainfall and there are few possible outliers which will be removed.
Figure 15, card payment has a low negative correlation with rainfall and there are few possible outliers which will be removed.

Discussion

a. Results interpreted and explanations offered for trends or patterns as well as anomalies in the data.

As stated above, the chosen metrics for model evaluation include precision, recall, F1-score, accuracy, and AUC. Precision is used for estimating the number of true positives amongst the total of both true positives and false positives. Recall is the assessment for the percentage of true positives predicted by the models amongst the actual true positives. The F1-score is the harmonic mean which represents the efficiency of both precision and recall. Accuracy is also a good metric since it shows the prediction capability of the model. The last metric mentioned in the table is AUC (Area Under Curve), which will be analysed for the ability to classify observations into classes. Other additional statistical methods that could be used are R^2 and RMSE (Root Mean Squared Error). The model is considered as highly efficient when it has a high R square and a low RMSE.
As for the ROC visualisation, it illustrates the trade-off relationship between sensitivity and specificity. Therefore, the roc line which is closer to the top-left corner would depict a better performance.
By assessing these metrics on Logistic regression, it can be concluded the model performs quite well with a moderate to high rate for all the metrics. The precision rate (0.80), recall rate (0.92), F1-score (0.86), and Accuracy (0.76) all passed the threshold of 0.70 which proves that this model is effective at predicting the rainfall. Also, AUC with 75% is a good indicator for the model power to classify 0 (Yes) and 1 (No) for tomorrow 's rainfall.
On the contrary, the Decision tree is considered as ineffective at predicting the target variable which can be seen from the metrics score. Although Precision, Recall and F1-score surpass 70%, its accuracy and AUC seems trivial, especially AUC with only 0.46. This means the model cannot predict a good number of true positives and negatives, and its ability to differentiate the outputs is extremely low. Also, the ROC curve proves this since it falls below the dot line to the right side corner, indicating a low performance.
After pre-tuning the decision tree model, a higher model performance has been attained when all the metrics skyrocket with a fairly good accuracy (nearly 80%). All the other metrics (Precision, Recall, and F1-score) skyrocket to a higher score (87%) compared to just around 70% in the regular decision tree model. In the meantime, the classification capability of the model has been witnessed to reach 70%, which is a comparatively good rate.

b. Put forward the best model for rainfall prediction

By looking at the metrics table, confusion matrix and ROC curve, it is undoubted that the optimised Decision tree model outperforms the other models, especially the score of almost 80% for Accuracy. Although the pre-tuning model receives lower values at Recall and AUC compared to Logistic regression, it still shows a superior score in light of Precision, the harmonic mean F1-score and Accuracy. Significantly, the model is recognized as having the least value for false positives and false negatives (7 and 7) compared to the other models regarding the confusion matrix.

c. The meaning of the results to the original goal/hypothesis

The original goal is to find out the relationship between rainfall and the business performance, therefore establish a predictive model to predict whether it will rain tomorrow to help the business make decisions. Throughout the whole process, the relationship has been recognized as negative between the rain fall and the total revenue. One more finding should be addressed is that the delivery services (Menulog, UberEat, and Deliveroo) perform highly efficiently when the rainfall increases, which indicates a positive relationship for all of the delivery modes. Also, the model results indicate that it is of absolute capability of predicting whether it will rain or not with high accuracy. By interpreting these trends and forecasting the coming rains beforehand, the business owner will be able to make business decisions on staff allocation as well as delivery services control at the right time of the year (seasonal patterns). By doing so, the business performance will be less influenced by special events such as rainfall and pandemics, overstaffed and understaffed circumstances.

d. The most important variables in the models

The most important variables chosen in the models are rain_today, temp, hum, evap and rainfall.

As for the rain today depicted in the bar plot (figure 6), there is a correlation between the rain today variable and rain tomorrow variable when the rain today is ‘Yes’, then the rain tomorrow value will be obviously much higher in the area of ‘Yes’ and vice versa. In the heatmap (figure 9), it shows that the correlation between rain_today and rain_tmrw is considerable with the value of 0.19 - a comparatively high score compared to the others. Comparing the heatmap results, rainfall with a high correlation score (0.26) is regarded as one of the most vital explanatory values to be used in the model regardless of the fact that it contains a significant amount of outliers. For those reasons. it could be declared that rain today and rainfall can be seen as important in the modelling process.

Depending on the box-plot whisker plots, it can be deduced that evap and hum do not have many outliers, which are considered as having least impacts on the modelling outcome. Also, they also have significant correlation percentages with the target variable with -0.29 for evap and 0.18 for hum. For those purposes, evap and hum are standard enough to be considered amongst pivotal variables.

Considering the box and whisker plot of Temp, it is observed with no outliers which is suitable for modelling purposes. Moreover, the correlation rate of Temp against the target variable is witnessed with up to -0.28, which is also a strong statistical value compared to the others. Therefore, it is undeniable that Temp should be one of the most essential variables to conduct the modelling procedure.

e. Further questions and current limitations

There are a lot of questions and limitations which should be addressed for possible improvements in the future works.

Except for Logistic regression and Decision tree models, there are questions about how effective the other prediction models will perform on the same dataset structure. There are a range of classification models such as random forests classifier, Kneighbors classifier, xgb.XGB classifier, etc. that could be implemented in this dataset, which might be highly promising at predicting the rain. This experiment with these models should be established to enhance the metrics performance in the future.

The second question is raised if there are other variables for enhancing model predictive capability. Different datasets are outputted everyday, which might make it unstable for specific variables to perform well on all the datasets, even if the variables considered are the same. Therefore, this problem should be enlightened in the future with different methods to conclude a group of long-lasted variables which can be inputted in every model.

These future scenarios should be executed for a broader range of applications in not only business aspects but also other industries.

Current limitations: The current dataset is not large enough for performing all the data transformations. For example, in the outliers removal process, only three variables were conducted but the data has incurred a loss of 17% over the original total of data observations. Therefore, this has resulted in the other variables being unable to be accessed properly, which can be a hindrance for a worse decision of variable choices.

Conclusion

In summary, the report documents the process of consulting a small business based in Melbourne 'Kai Mook Thai Restaurant', seeking to optimise their rostering and HR costs in two ways: minimising losses from overstaffing and by limiting opportunity costs from having insufficient resources. Upon delivery of their business needs with respect to their geographical context, it is decided to generate a model to predict rain conditions, based on sales trends.

To begin, the data was cleaned yielding 414 observations with 9 restaurant variables and 22 weather variables. NA's were handled, dates formatted and 2 extra binary variables 'rain_today' and 'rain_tmrw' were generated for modelling purposes. After cleaning, it was decided to use python in addition to r. This would help us take advantage of different packages and cater to individual skills within the group. In python for modelling purposes, outliers were detected by using interquartile range and then removed to get the final dataset.

Using Seaborn and ggplot, visualisations were generated to assess the correlations between variables. True to the hypothesis, total sales was indeed slightly negative against rainfall with a relatively weak scatterplot. However, it is noteworthy that deliveries in isolation flipped the trend and generated a slightly positive correlation.

The selected modelling procedure is based on Logistic regression and Decision tree classifiers to predict the rain tomorrow. After all the relevant python packages have been configured, the models can be trained. Our chosen evaluation metrics are precision, recall, F1-score, accuracy, and AUC with Decision tree consistently outperforming the logistics regression model at an accuracy of almost 80%, just falling short of expectations.

Throughout the process, it was discovered that delivery services across all modes generate more revenue in rainy conditions. While total sales remain negative against rainfall. This indicates that catering to outsourced delivery services should be prioritised over in-store sales in rainy conditions. This could see the restaurant shifting staff allocation away from wait staff giving the kitchen more resources to process demand.

Reference

Enviro Friendly 2021, Melbourne Victoria Average Rainfall [2021], Enviro Friendly, viewed 21 October 2022, <https://enviro-friendly.com/information/average-rainfall/melbourne-victoria-rainfall/#:~:text=Melbourne>.

Burt, S 2019, Explaining Melbourne's crazy but predictable weather, The University Of Melbourne, Pursuit, viewed 21 October 2022, <https://pursuit.unimelb.edu.au/articles/explaining-melbourne-s-crazy-but-predictable-weather>.

Reid, K 2017, Four Seasons in One Day – Scientific Scribbles, The University Of Melbourne, viewed 20 October 2022, <https://blogs.unimelb.edu.au/sciencecommunication/2017/08/06/four-seasons-in-one-day/>.

Zomato 2015, Zomato.com, viewed 16 October 2022, <https://www.zomato.com/melbourne/kai-mook-thai-restaurant-viewbank>.

Bureau of Meteorology 2022, Daily Rainfall - 086282 - Bureau of Meteorology, Bom.gov.au, viewed 21 October 2022, <http://www.bom.gov.au/jsp/ncc/cdio/weatherData/av?p_nccObsCode=136&p_display_type=dailyDataFile&p_stn_num=086282&p_startYear=>.