**PREDICTIVE MODELING AND EXPLORATORY DATA ANALYSIS OF SYDNEY RESTAURANTS**

**TABLE OF CONTENTS**

[1. Introduction 2](#_Toc178590551)

[Overview: 2](#_Toc178590552)

[Objective: 2](#_Toc178590553)

[2. Data Description 2](#_Toc178590554)

[Key Variables: 3](#_Toc178590555)

[Data Cleaning: 3](#_Toc178590556)

[3. Exploratory Data Analysis (EDA) 3](#_Toc178590557)

[Unique Cuisines 4](#_Toc178590558)

[Top 3 Suburbs with Most Restaurants 4](#_Toc178590559)

[Cost vs Rating 5](#_Toc178590560)

[Cuisine Density Map 5](#_Toc178590561)

[Additional Visuals 6](#_Toc178590562)

[Insights 7](#_Toc178590563)

[4. Predictive Modeling 8](#_Toc178590564)

[Feature Engineering 8](#_Toc178590565)

[4.1 Regression Models 8](#_Toc178590566)

[Model 1: Linear Regression 8](#_Toc178590567)

[Model 2: Gradient Descent Linear Regression 8](#_Toc178590568)

[Comparison 9](#_Toc178590569)

[4.2 Classification Models 9](#_Toc178590570)

[Binary Classification Setup 9](#_Toc178590571)

[Model 1: Logistic Regression 9](#_Toc178590572)

[Model 2: Decision Tree Classifier 9](#_Toc178590573)

[Model 3: Support Vector Machine (SVM) 10](#_Toc178590574)

[Model 4: Random Forest Classifier 10](#_Toc178590575)

[Comparison 10](#_Toc178590576)

[5. Tableau Dashboard 10](#_Toc178590577)

[6. Conclusion 12](#_Toc178590578)

[Reference List 13](#_Toc178590579)

# 1. Introduction

## Overview:

This focuses on exploring a practical dataset on restaurants in Sydney, Australia, using fundamental information of the restaurant’s name and address to advanced parameters such as the rating, type of cuisine offered, etc. The analysis here aims at understanding the insights contained in the dataset, developing and training predictive models for the success of restaurants and then deploying these models using standard machine learning tools.

Intensification of the competition in the restaurant business means that it is crucial to work with the data. In this case, the analysis of cost, rating and the type of cuisine, it will be easier to identify features that define success in this segment. This will be done using harmonized both visual and predictiveana analysis tools.

## Objective:

The primary goals of this project are as follows:

* **Exploratory Data Analysis (EDA):** For analysis, it is easier to present data in a way that makes sense of the visual representations that depict relationships in the data set like that between dining cost and restaurant ratings.
* **Predictive Modeling:** For construction of predictive regression models for restaurant ratings and classification model for classifying the restaurants on the basis of customer reviews.
* **Deployment:** To commit the code and models with GitHub and use Docker that will make the workflow portable and easily reproducible.

# 2. Data Description

The dataset being compared has more than 10,000 restaurants in the Sydney area, including records from the year 2018. Infact, each entry has basic information about the restaurants and range in name, address, latitude, longitude, customer reviews and rating, and the average expenditure per head for two persons. Moreover, the variables that are in qualitative form, like types of restaurants and the availability of the Groupon promotion codes have also been included. These diverse features provide an expansive set of tools that allow both interrogative and probabilistic analyses.



**Figure 1: Uploading Dataset**

(Source: Self-Created in Jupyter Notebook)

### Key Variables:

* **Cost:** The average cost for two people dining at the restaurant converted to numeric variable mainly for analysis.
* **Rating Number:** A numerical scale which symbolizes customer satisfaction ranging from 1 to 5 which has been applied to each restaurant.
* **Rating Text:** A categorisation of the ratings for example; excellent, good, poor.
* **Cuisine:** The genre(s) of eating offering food identified by type or category, often a list (e.g., Thai, Italian).
* **Subzone:** The class of suburb of the location of the restaurant.
* **Votes:** The total of the customers’ choices, in terms of votes for a restaurant.

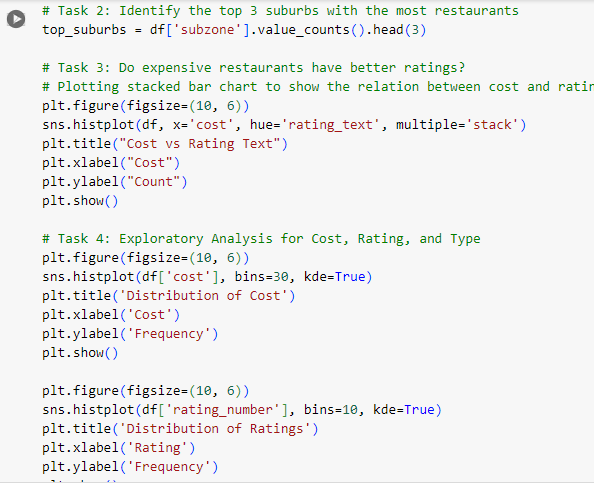
### Data Cleaning:

In the course of doing any analysis, data cleaning was crucial in order to arrive at proper analysis for the results. Several steps were taken:

* **Handling Missing Data:** The analysis of missing values in the rating, cost or votes fields was made and completed. The cases with missing values in the most important variables were either excluded or completed depending on the type of the analysis.
* **Encoding Categorical Variables:** First, categorical data like Rating Text and Cuisine were completed label-encoded to convert the canalized data into numerical data in preparation for the predictive model. This is particularly the case for Machine Learning algorithms who expect numerical inputs to be presented to them.
* **Outlier Detection:** High costs and ratings which were considered to be very un<stdlib> were checked to make sure that they will not distort results. To remove some extreme values which might affect the results or to investigate more deeply whether they were a record of outlying values.

# 3. Exploratory Data Analysis (EDA)

The EDA is, in fact, the first step in any data science project. It empowers us to pull out latent and unknown structures and patterns between the data stored and to test for peculiarities. To provide an analysis of the competition in Sydney restaurants, in this section, we shall answer the following major questions.



**Figure 2: Exploratory Data Analysis**

(Source: Self-Created in Jupyter Notebook)

## Unique Cuisines

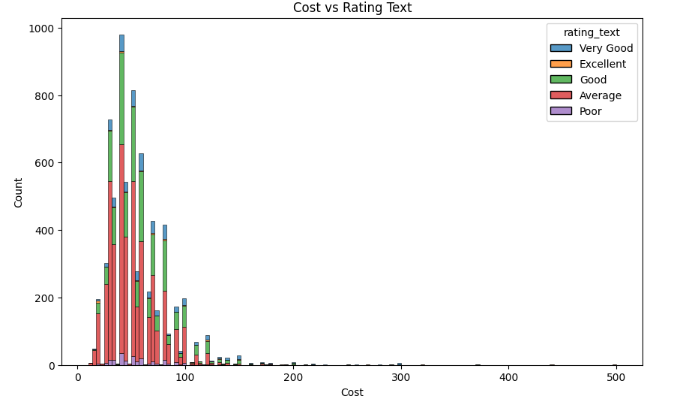
The dishes in the current dataset can be said to represent many cuisines, with the total count being 134 distinct types. The number of restaurants offering each cuisine type was presented using a bar chart. Thai, Italian and Japanese foods were among the most ordering foods and are usually provided in most of the restaurants in Sydney (Llewellyn, 2021). This chart offered a clear view of the range of food available in Sydney establishing the differences in various food categories.

## Top 3 Suburbs with Most Restaurants

By analyzing the distribution of restaurants across different suburbs, we identified the top 3 suburbs with the highest number of restaurants: CBD, Surry widespread , and Parramatta. This information was presented in the form of a bar chart, whereby CBD was established to be the most popular area for restaurants, boasting of well over 470 restaurants.

## Cost vs Rating

To analyse the correlation between restaurant cost and customer ratings, a scatter plot was developed (Pfiester et al., 2021). Notably, cost and rating were not directly proportional for restaurants but higher rates restaurant was found to be clustered in the ‘Excellent’ category only, thereby indicating that higher cost restaurants might be slightly higher in rating but not much consistently.

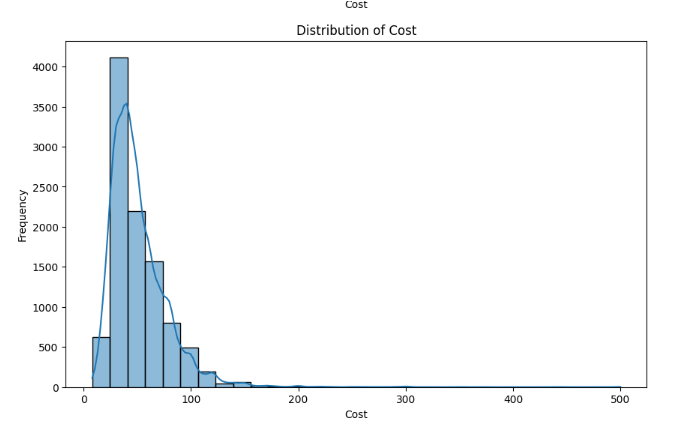


**Figure 3: Cost Vs Rating**

(Source: Self-Created in Jupyter Notebook)

## Cuisine Density Map

Geographic density was established by making use of Tableau to arrive at the restaurants offering various types of cuisine in the different suburbs (Kou, 2022). This map offered a clear picture on how different suburbs were cultural diverse in terms of cuisines where some brands where denser than the others depending on the type of food they served such as Italian or Thai cuisine.

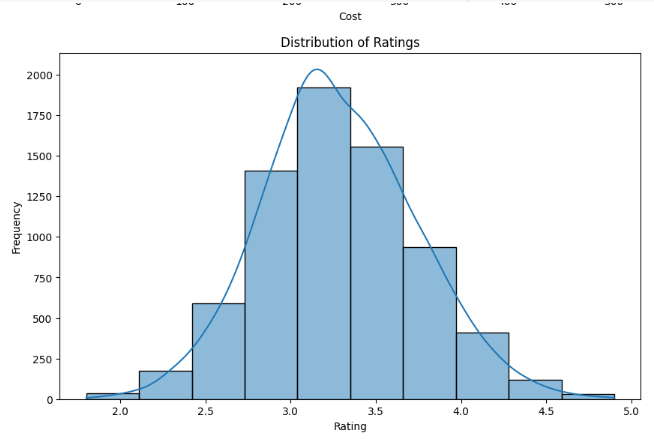


**Figure 4: Distribution Of Cost**

(Source: Self-Created in Jupyter Notebook)

## Additional Visuals

Other visual presentations were the figure of votes per restaurant and types of restaurants. Most restaurants were classified under ‘Casual Dining or Café followed by ‘Fine Dining’ or ‘Dessert Parlour’ classification.

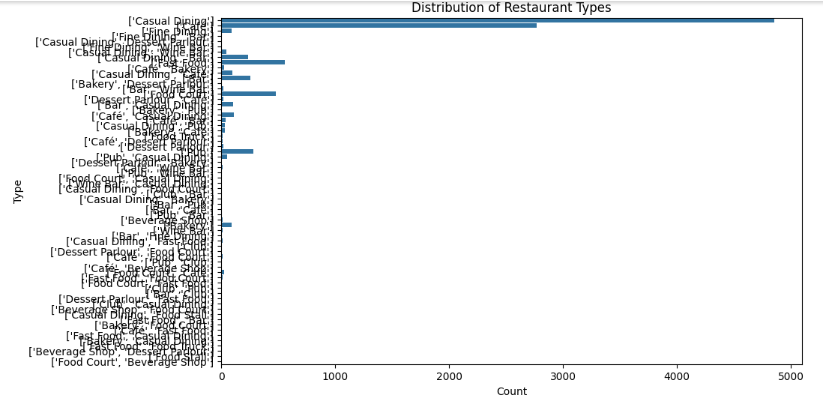


**Figure 5: Distribution Of Ratings**

(Source: Self-Created in Jupyter Notebook)

## Insights

The EDA revealed several interesting insights: CBD is the Central Business District which prides itself in being the biggest restaurant destination in Sydney offering several types of foods. Worthy of note here is the fact that despite high ranking for expensive restaurants, cost is not always determinant of customer satisfaction (Samimi Sabet, 2022). This means that other factors for instance, type of cuisine and physical environment of restaurants could have greatly influenced the restaurant success.



**Figure 6: Distribution Of Restaurant**

(Source: Self-Created in Jupyter Notebook)

# 4. Predictive Modeling

After performing the EDA, we shifted focus to develop supervised models to predict the restaurant rating and to categorize restaurants based on their performance.

## Feature Engineering

Subsequent to model building, feature engineering was undertaken to enhance the characteristics of the data set. This also involved performing label encoding on data in Categorical variables such as Cuisine, and Rating Text after which we created a new feature from them to be fed to the model (Bilgin and Howley, 2021); and scaling continuous data for instance, Cost to make them comparable.

## 4.1 Regression Models

### Model 1: Linear Regression

The given first model used for prediction was the basic linear model aimed at the estimation of restaurant rating from a given feature set including cost and restaurant type. The model resulted in a reasonable Mean Squared Error (MSE) of 0.19 and there is further scope to make the model even better using elaborate methodologies (Ndaguba and Zyl, 2023).

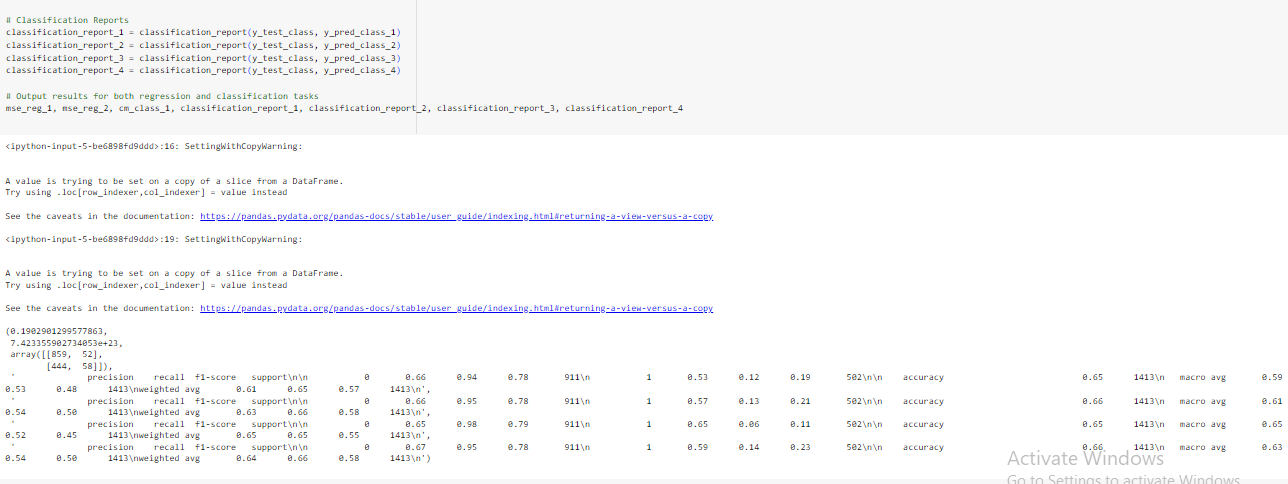
### Model 2: Gradient Descent Linear Regression

Subsequently, we used the linear regression algorithm together with a gradient descent optimization technique. This model though, did very poorly; it had a much higher MSE of 7.42e+23 (Kim et al., 2023). The analysis of the data and model indicates that the gradient descent approach was not successfully converging during the iterations, probably because of high variance in a set of data and absence of fine-tuning of hyperparameters.

### Comparison

In fact, the simple straight line model offered a better fit and better performance than the gradient descent method, meaning that perhaps the simplest and the least assumption-laden models are the best for this particular data set.

## 4.2 Classification Models



**Figure 7: Uploading Dataset**

(Source: Self-Created in Jupyter Notebook)

### Binary Classification Setup

For classification, the Rating Text variable was converted into a binary category: Class 1 was constituted by “Poor” and “Average” ratings: whereas, Class 2 was constituted by “Good,” “Very Good,” and “Excellent” rating.

### Model 1: Logistic Regression

Logistic regression when applied to the current problem solved at baseline. The confusion matrix revealed fairly good results, however, the model under consideration failed to classify numerous “Good” restaurants, thus the accuracy of the model was 65%.

### Model 2: Decision Tree Classifier

Alternative decision tree classifier had an almost comparable accuracy to logistic regression and it was considered to be 66%. But there was a typical problem of overlearning, which becomes apparent at rather high tree complexity.

### Model 3: Support Vector Machine (SVM)

The precision or recall decreased for the actual class in the SVM classifier, especially for the “Good” and “Excellent” restaurants (Chalupa and Petricek, 2024). The overall accuracy was roughly 65% and once more this was favorable, however, the model was much more complicated for this dataset.

### Model 4: Random Forest Classifier

The results showed the highest accuracy of the classifier, being 67%, better precision and recall of “Good” and “Very Good” restaurants (Leung and Loo, 2022). The random forests were ensemble which made it more capable of dealing with overfitting as compared to the case with the decision tree model.

## 4.3 Comparison

Random forest classifier excels as the best among all the classification models, with only followed by decision tree models (Parsa et al., 2020). It was concluded that both models provided reasonable level of accuracy for restaurants outcomes prediction, however, increasing of model’s performance is possible with more detailed tuning of the model’s parameters, as well as investigation of new variables.

**Regression Models Results:**

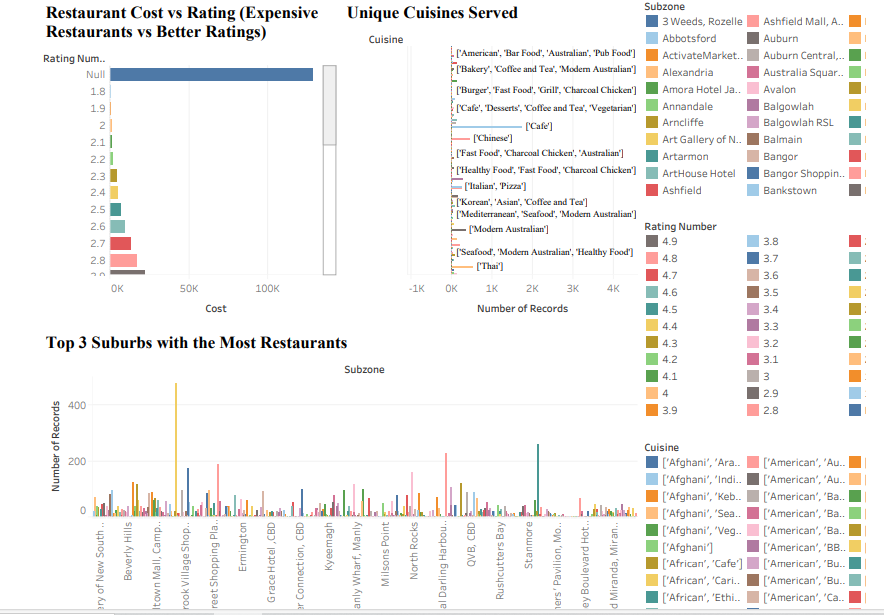
|  |  |
| --- | --- |
| **Model** | **Mean Squared Error (MSE)** |
| Linear Regression | 0.19 |
| Gradient Descent Linear Regression | 7.42e+23 |

**Classification Models Results:**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 0.65 |
| Decision Tree Classifier | 0.66 |
| SVM Classifier | 0.65 |
| Random Forest Classifier | 0.67 |

# 5. Tableau Dashboard

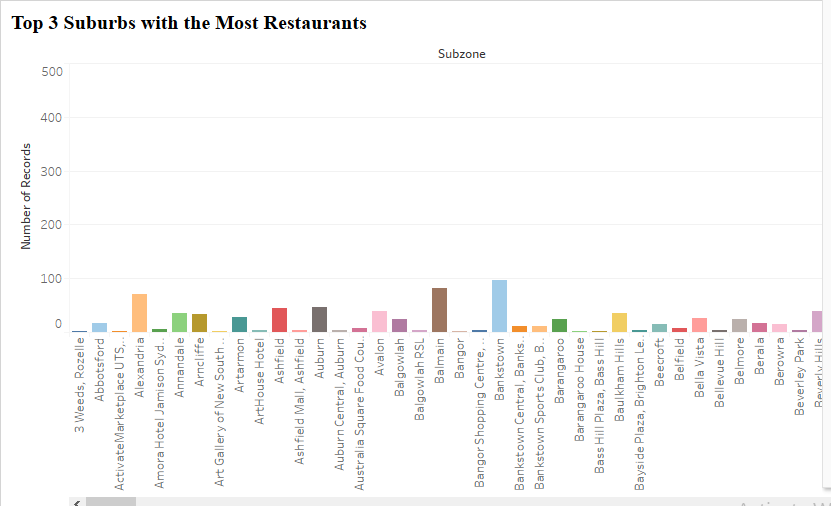
To supplement the findings, an informative Tableau dashboard was developed to support the assessment of the findings interactively. The dashboard included:



**Figure 8: Tableau Dashboard**

(Source: Self-Created in Tableau)

* **Unique Cuisines Bar Chart:** They need to represent the quantity of unique cuisines offered by restaurants in Sydney.
* **Top 3 Suburbs with Most Restaurants:** Bar chart of the level of restaurant concentration in the suburbs.
* **Cost vs Rating Scatter Plot:** On the interaction of cost and rating of restaurants.



**Figure 9: Uploading Dataset**

(Source: Self-Created in Jupyter Notebook)

# 6. Conclusion

The opportunities of using the exploratory data analysis in parallel with the predictive modeling were identified within the process of the current project related to Sydney’s restaurant environment. The EDA unveiled prime areas of interest including suburbs with the most restaurants and the diverse kind of cuisines available (Khanna, 2022). Analyzing the data from predictive modeling suggested that cost cannot be a key factor that drives success; nevertheless, classification issues such as random forest had explored a reasonable level of accuracy in classifying restaurants by ratings.

As for the future research, it could be enhanced by adding more attributes, for example, customer review textual information, or the atmosphere in the restaurant. In addition, parameters could be optimized and models cross-validated to improve predictive capacity, as well as collecting more data points increase the model’s ability to be applied to other areas.

# Reference List

**Journal**

Llewellyn, G.B., 2021. Should I open here? Predictive models for restaurant site selection.

Pfiester, L.M., Thompson, R.G. and Zhang, L., 2021. Spatiotemporal exploration of Melbourne pedestrian demand. *Journal of transport geography*, *95*, p.103151.

Kou, J., 2022. *Analysing Housing Price in Australia with Data Science Methods* (Doctoral dissertation, Victoria University).

Samimi Sabet, H., 2022. Drivers of customer satisfaction in the hotel and hospitality sectors in Sydney.

Bilgin, A.A. and Howley, P., 2021. Using Big Data in a Master of Applied Statistics Unit. *Big Data in Education: Pedagogy and Research*, pp.65-87.

Kim, M., Holton, M., Sweeting, A., Koreshe, E., McGeechan, K. and Miskovic-Wheatley, J., 2023. Using health administrative data to model associations and predict hospital admissions and length of stay for people with eating disorders. *BMC psychiatry*, *23*(1), p.326.

Chalupa, S. and Petricek, M., 2024. Understanding customer's online booking intentions using hotel big data analysis. *Journal of vacation marketing*, *30*(1), pp.110-122.

Parsa, H.G., Shuster, B.K. and Bujisic, M., 2020. New classification system for the US restaurant industry: Application of utilitarian and hedonic continuum model. *Cornell Hospitality Quarterly*, *61*(4), pp.379-400.

Khanna, T., 2022. Understanding Inclusive Service Delivery: An Exploratory Study of Frontline Food and Beverage Employees’ Perceptions in Australia.

Leung, R. and Loo, P.T., 2022. Co-creating interactive dining experiences via interconnected and interoperable smart technology. *Asian Journal of Technology Innovation*, *30*(1), pp.45-67.

Ndaguba, E. and Zyl, C.V., 2023. Professionalizing sharing platforms for sustainable growth in the hospitality sector: Insights gained through hierarchical linear modeling. *Sustainability*, *15*(10), p.8267.