V.Ger Travel Company - An Analysis

ITNBD4 Assignment

Student: 2710017

10 January 2025

Table of Contents

1 Summary

Summary of the report

2 Introduction

2.1 TO DO

appendices:

```
- _chapters/appendix_demand.qmd
- _chapters/appendix_customer_satisfaction.qmd
- _chapters/appendix_sarima.qmd
```

2.2 Background & Approach

The travel conglomerate V.Ger Travel (VGT) has a broad range of operations, including hotels, resorts, car rentals and also air travel through charter flights. Travel bookings originate primarily from VG's own travel web site which is supported by several operational information systems that cover all aspects of it's business from customer relations through to logistics and maintenance. There is a wealth of data available with over 10 years history of travel bookings, however, the use of modern Data Science methods to harness this data is in its infancy at VGT.

This report describes the recommendations of the new Chief Data Officer (CDO) as to how to implement modern Data Science techniques to utilise VGT's data to improve its efficiency and profit. Having completed an initial high level review, two opportunity areas where identified that appeared to offer the biggest opportunities for increasing efficiency and also that are achievable as a first step in implementing new techniques. The two use cases are:

- Hotel Demand Forecasting
- Customer Satisfaction

Each of these use cases are explored in the following sections, each of which describes:

- Business Scope & Benefits - What is proposed and why is this helpful
- Data Analysis Approach - The suggested analysis techniques and the data required
- Simulated Data Analysis - An example of the analysis using simulated data
- Conclusions & Next Steps - Findings and how to move forward

????? - Assumptions made about data prep/data wrangling .. eg normalisation of data etc - Any assumptions about company size, number of hotels etc ????

2.3 Results & Further Details

A summary of the approach taken and key results of the analysis are described for each use case in the individual sections of this report. Additionally, the all details for generating the simulated data and the analysis steps are described at the following:

- Apendices - Step by step approach, with data tables and plots
- Jupyter Notebooks - Full listings of Python code and results in a GitHub Repository
- Data Files - Generated CSV files in a GitHub Repository

2.4 Conclusions

???? ConclusionshHere as well as last past of each use case ????

3 Hotel Demand Forecasting

3.1 Business Scope & Benefits

The hotel operations of VGT are a significant part of its business and any efficiencies in this area have the potential to make large contributions to the overall profitability of the business. The profit contribution from individual hotels can be maximised by ensuring revenue is as high as possible and at the same time minimising the hotel's operating costs. One way to do this is to provide a hotel's management team with the tools to carry out reliable demand forecasting.

Hotel demand forecasting is the prediction of the demand for rooms and related hotel services to help a hotel's management team determine pricing, staffing and marketing strategies (Johansson, 2022). If the demand for hotel rooms can be reliably forecast then this enables:

- Dynamic Pricing - Adjust future prices in response to forecast demand. When high demand is expected then future room rates can be increased; when low demand is expected then discounts can be offered or packages can be advertised. And marketing strategies can be determined to respond to the demand forecasts. xx increase occupancy and revenue
- Staffing Levels - Hotel staffing can be adjusted to maintain customer service levels but not over staff when demand is expected to be lower. xx control costs
- Inventory Management - Similarly inventory can be adjusted, for example catering supplies maintained just sufficient to meet the forecast number of customers using catering facilities. xx control costs

There are many factors that will influence the demand for hotel rooms, some may remain stable whilst others are less so and will vary over time or in response to external factor, these include:

- Location & Market - Is the hotel a budget or a boutique hotel? Is is in a business district, near a beach or a ski resort etc. Are customers mainly business travelers or tourists?
- Economic - Macro level impact from the state of the economy.
- Local Competition - Competition with local hotels.
- Seasonality - When are the high and low seasons? Is it a summer or a winter resort? What are the local weather patterns and seasons. Are there weather related attractions?
- Local Events - When are any local festivities, sports events, school holidays, religious events, music conferences, business conferences?

3.1.1 Occupancy & Revenue Indicators

An important indicator of demand is the occupancy rate (Jeffrey and Hubbard, 1994) and (FHA, 2023) which is simply the percentage of the total rooms occupied for a given time period. The occupancy rate varies across the industry but a target of 60% to 80% is typical.

Occupancy is used alongside revenue related indicators to provide a measure of revenue health, the three main indicators (for a given time period, eg daily) are:

- Occupancy Rate (OCC%) - Percentage of available rooms that are occupied, or expected to be occupied.
- Average Daily Rate (ADR) - Average revenue per room occupied, across all room price bands.
- Revenue Per Available Room (RevPAR) - Revenue reflecting all available rooms. Calculated by: OCC% * ADR. A good overall indicator of revenue.

3.1.2 Forecasting Occupancy & Business Benefits

The time period used for forecasting will vary depending on the objectives desired (Lighthouse, 2024) and (Lighthouse, 2023) for example:

- Short Term - Forecast occupancy for next month so room pricing can be adjusted appropriately.
- Long-term - Forecast occupancy for next year so price bands and packages can be set, marketing strategies defined and required staffing levels determined.

To recap, if occupancy (and the associated revenue indicators) can be reliably forecast then plans can be put in place to maximise revenue by adjusting pricing and marketing strategies and controlling costs by flexing staffing and inventory levels.

3.2 Data Analysis Approach

At VGT, no rigorous forecasting is currently in place so to begin with a relatively simple approach will be implemented in a small number of hotels. If the benefits of this are confirmed, then it can then be extended in sophistication using more complex forecasting models and for longer time periods. It can then be implemented across all hotels in VGT.

The objective of this first step is to establish a model that can be used to forecast the daily occupancy (OCC%) at an individual hotel for the coming month, ie the forecast is for 30 to 40 days in the future. The forecast will then be used by hotel's management team to: i) adopt dynamic pricing; ii) execute supporting short-term marketing; iii) fine-tune staffing rotas and holiday leave for the coming weeks. This should improve the efficiency of the hotel's operations by increasing room bookings whilst ensuring staffing costs are controlled at an appropriate level.

3.2.1 Forecasting Model & Data Required

The scope of the envisaged forecasting model is to calculate a month of daily OCC% for each room category in an hotel. The output of the forecast model will be provided in spreadsheet form so that the hotel management team can manually make adjustments to try to improve revenue and staffing levels. The actual occupancy and revenue can then be tracked against the forecast throughout the month in order to assess how accurate the forecasting model is and to help refine it.

Forecasting Model Data

The data required for the forecasting model is the last 4 years of daily room occupancy. In this analysis a single hotel with two classes of room (standard and premium) will be used. A 4 year history of daily room bookings was selected with 3 years used to be used for training and 1 year for validation. This length of history was chosen as a starting point because a previous occupancy study (Phumchusri and Suwatanapongched, 2023) found that the choice could be quite dependent on the scenario; however, there needs to be a a balance of too short and missing seasonality vs too long and not being sufficiently responsive. Phumchusri and Suwatanapongched (2023) also found that using 4 years history of daily occupancy to forecast 2 to 8 weeks was a good approach.

The specific data required is:

- For an individual hotel and each room category (standard and premium)
- Daily
- Room capacity
- Room rate
- Rooms occupied

Tracking Spreadsheets

The data comprising the revenue forecasting and tracking spreadsheet is:

- Forecast OCC% (derived from the forecasting model)
- Daily room rates by room category that can be manually adjusted
- Daily Revenue, ADR, RevPAR (derived from the OCC% and room rates)
- Special events, a facility to mark local events that may impact demand. For example, sports events, concerts, conference, unusual weather forecasts
- Actual OCC%, Revenue, ADR, RevPAR

The data provided for the staffing forecasting spreadsheet is:

- Forecast OCC% (from the forecasting model)
- Averaged staff requirements per room
- Staffing levels (derived from the OCC%)
- Actual staffing levels

3.2.2 Techniques Considered

There are several potential tools that can be used with historical time series data to forecast future occupancy rates, ranging from established 'simpler' techniques such as linear regression through to more sophisticated machine learning and neural network models (Huang and Zheng, 2022). However, given the relative infancy of Data Science at VGT, the use of more sophisticated tools will be prioritised for future work. In this investigation, the following techniques will be examined:

1. Ordinary Least Squares (OLS) linear regression
2. ARIMA, SARIMA - Use to fully incorporate seasonality
3. SAIMAX - If exogenous factors need to be accounted for
4. ***?? LightGBM …. if time allows!***

OLS Linear Regression

Investigating techniques for reliably identifying the factors influencing hotel occupancy has been ongoing for several years. See Andrew, Cranage and Lee (1990) and Jeffrey and Hubbard (1994) for earlier work focusing on the use of regression analysis of time series data. Although it is likely that the nature of hotel business means that occupancy will be seasonal, linear regression will still be investigated first.

ARIMA, SARIMA

Occupancy forecasting using ARIMA using factors such as room capacity and marketing expenditure has successfully been used (Chow, Shyu and Wang, 1998). A comparison of forecasting methods (Weatherford and Kimes, 2003) included historical time series analysis of occupancy using ARIMA. The best time period used was not clear and it is a balance of too short and missing seasonality vs too long and not being sufficiently responsive. Also the best analysis method appears to depend on the characteristics of individual hotels and hotel chains. Using a SARIMA approach using 4 years history of daily occupancy to forecast 2 to 8 weeks was found to be a strong approach (Phumchusri and Suwatanapongched, 2023).

SARIMAX

?? LightGBM .... if time allows!

3.3 Simulated Data Analysis

3.3.1 Data Summary

A dataset was created to simulate the data as defined in the earlier section. The data elements consist of:

	Count	Missing	Empty	Unique	Type	String	Int	Float	List
Standard_OCC	1461	0	0	177	int64	0	1461	0	0
Standard_Capacity	1461	0	0	1	int64	0	1461	0	0
Standard_Rate	1461	0	0	1	int64	0	1461	0	0
Premium_OCC	1461	0	0	101	int64	0	1461	0	0
Premium_Capacity	1461	0	0	1	int64	0	1461	0	0
Premium_Rate	1461	0	0	1	int64	0	1461	0	0

The first few elements of the loaded data:

```
            Standard_OCC  Standard_Capacity  Standard_Rate  Premium_OCC  \
Date
2020-01-01           129                254            325           65
2020-01-02           126                254            325           53
2020-01-03           137                254            325           63
```

Date	Premium_Capacity	Premium_Rate
2020-01-01	100	575
2020-01-02	100	575
2020-01-03	100	575

And key descriptive statistics:

	Standard_OCC	Standard_Capacity	Standard_Rate	Premium_OCC	Premium_Capacity	Premium_Rate
count	1461.000000	1461.0	1461.0	1461.000000	1461.0	1461.0
mean	135.646133	254.0	325.0	56.141684	100.0	575.0
std	50.216654	0.0	0.0	29.349613	0.0	0.0
min	38.000000	254.0	325.0	0.000000	100.0	575.0
25%	88.000000	254.0	325.0	31.000000	100.0	575.0
50%	136.000000	254.0	325.0	57.000000	100.0	575.0
75%	183.000000	254.0	325.0	83.000000	100.0	575.0
max	254.000000	254.0	325.0	100.000000	100.0	575.0

3.3.2 Time Series Characteristics

The occupancy time series for the two categories of room are shown in the figure below. This indicates that the occupancy has a strong seasonality with an annual peak and trough; there are also regular spikes to full capacity bookings. The premium room occupancy hits maximum and zero occupancy on several occasions.

The two occupancy time series were examined further (see details at the appendix) using lag plots, ACF plots and ADF tests which indicated autocorrelation, annual seasonality and a positive trend. Differencing was also completed to confirm non-stationarity. Finally a decomposition was completed and this confirmed the seasonality and trend, see the figure below.

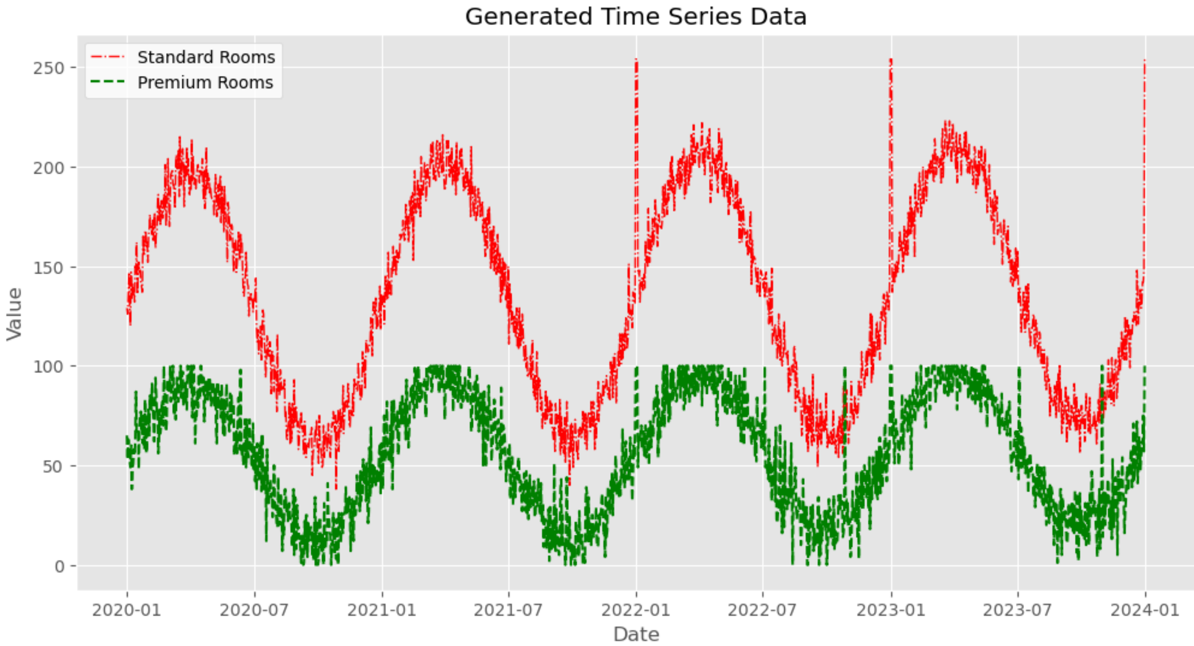


Figure 3.1: Occupancy for Each Room Category

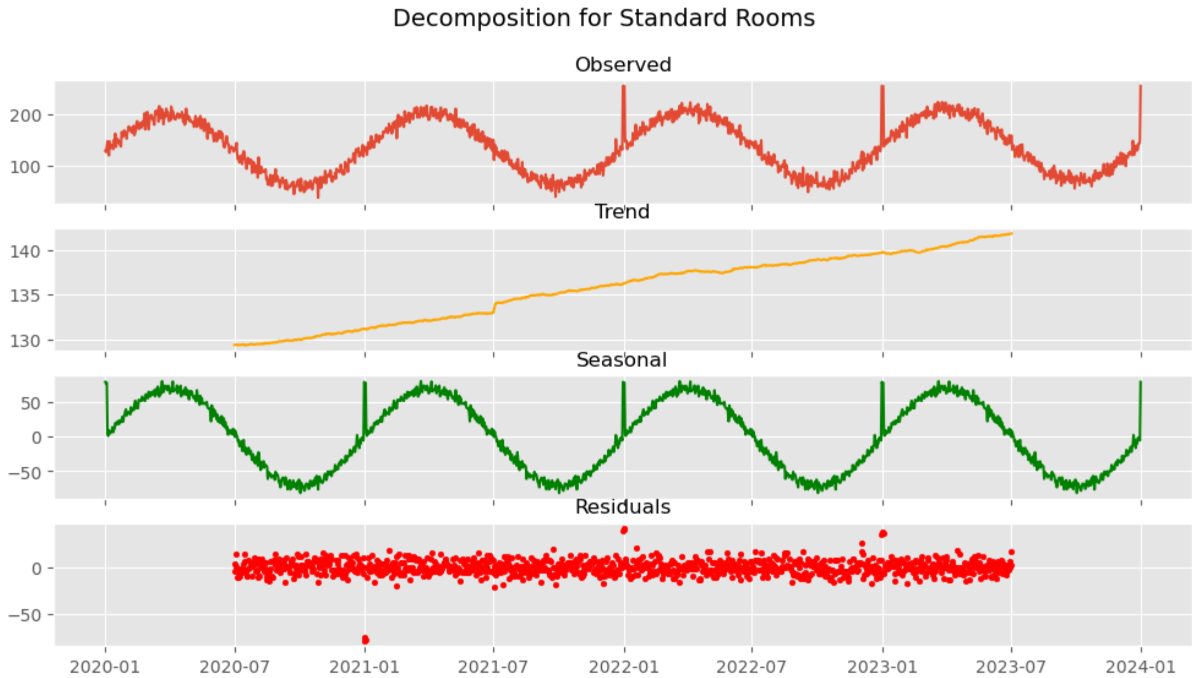
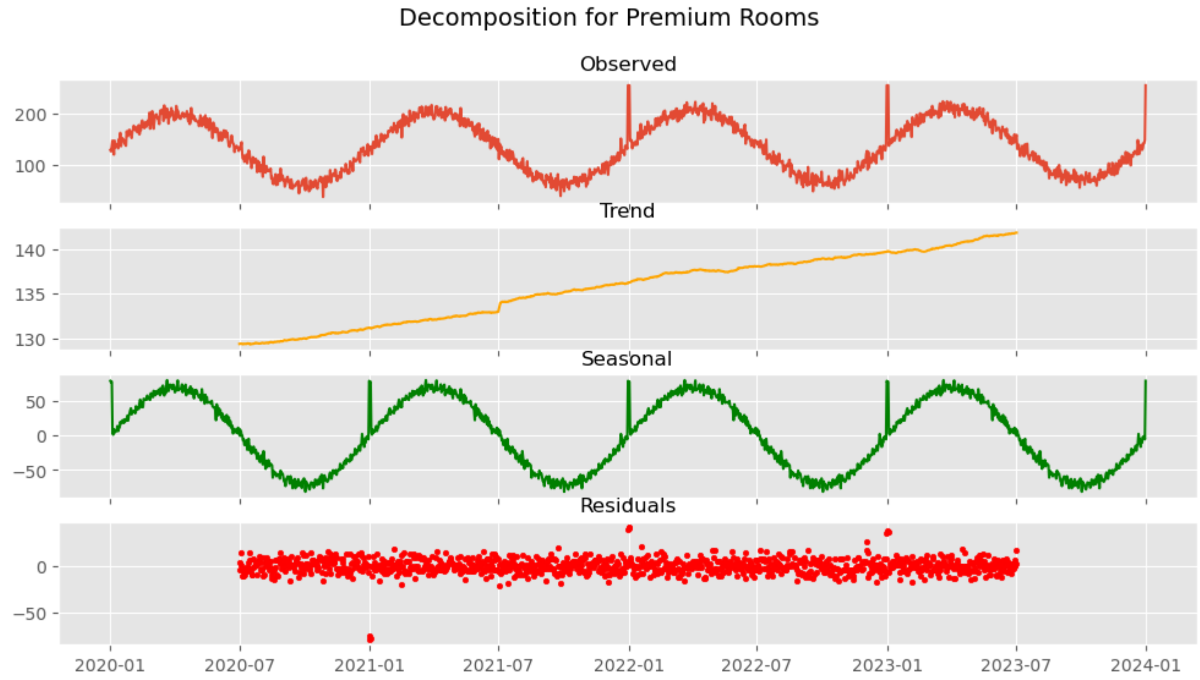


Figure 3.2: Occupancy Time Series Decomposition



3.3.3 OLS Linear Regression

Given the strong seasonality, linear regression is unlikely to be a good model for forecasting, however a regression fit was calculated to double-check. See the plot below which shows that the fit lines could not reliably provide a forecast and even show a downward trend. And the Durban Watson statistics are less than 2 which confirms evidence of autocorrelation.

3.3.4 ARIMA, SARIMA

3.3.5 SARIMAX

Test see earlier Section 3.3.3

and earlier **?@fig-test**

3.3.6 Execution - Data Analysis

- Use simulated data to carry out analysis
- Show the results, forecasts ….
- ***?? carry out PACF to determine the autoregressive order for ARIMA etc***

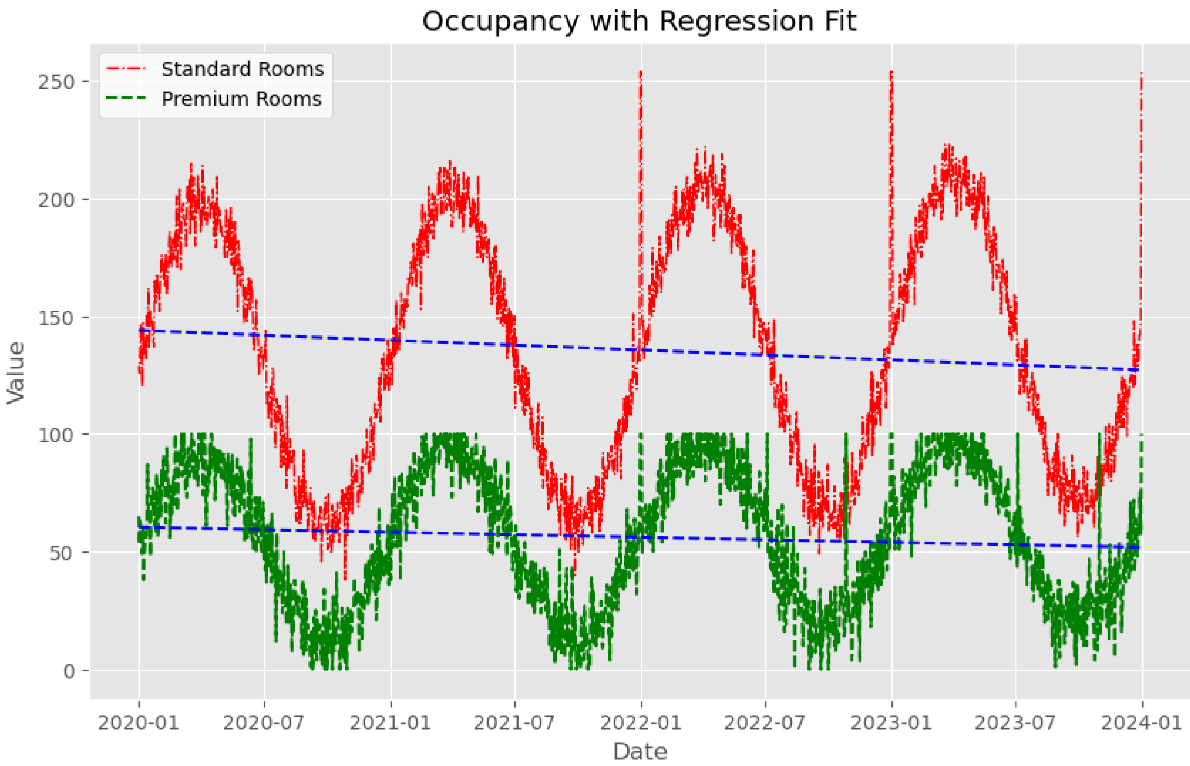


Figure 3.3: OLS Linear Regression

3.4 Conclusions & Next Steps

3.4.1 Findings

xxx

3.4.2 Next Steps

- ?? how are special events impactful on revenue predictions, can these be better used in the forecast going forward or in an improved forecasting model
- ?? recent flight searches, enquiries … for the area

xxxx

- ?? track actual OCC% etc and compare to forecast to allow model refinement as well as intra-month fine tuning
- ?? Get competitors OCC% to compare and improve forecasting models
- ?? automate the manual pricing changes etc, ie no need for manual adjustment of the forecast spreadsheet
- xx LightGBM ..
- xx CNN, LSTM, RNN …
- xx transformers, LLM ..

xxx improvements

- ?? isolate pricing strategies and refine the model
- ?? identify the main parameters / impacts on the forecasting … what are the demand indicators?
- ?? categorisation of customers, families, demographics etc, business, tourist
- ?? identify correlations with holiday patterns, events ……
- ?? identify seasonal bands, high, low, shoulder etc
- ?? identify links between revenue and room price bands …

4 Customer Satisfaction

4.1 Business Scope & Benefits

For the hotel operations within VGT, customer satisfaction is an important contributor to the overall efficiency and profitability of their business. This has been recognised by the hotel industry for several years: *"Service quality and customer satisfaction have gradually been recognized as key factors used to gain competitive advantage and customer retention"* (Yang, Jou and Cheng, 2011). If the levels of satisfaction across different aspects of a customers experience can be quantified, then action can be focused on improving those areas that are lacking. Clearly, identifying the areas that contribute most is part of the challenge. Not surprisingly, the use of data science to assist in this area is increasingly important (Zarezadeh, Rastegar and Xiang, 2022).

4.1.1 Measuring Customer Satisfaction

There are several aspects to quantifying customer satisfaction, for example:

- The specific areas of interest. Such as cleanliness, service experience, quality of facilities, local amenities and even things such as wifi availability
- Customer satisfaction ranking. Obtaining a customer's views on how poor to excellent a specific area was
- Importance to the business. The appropriate weighting of the importance of each area.

It has become standard practice and easy to gather customer feedback on how they rate their experience after a stay at a hotel through check-out questionnaires and feedback requests (Li, Ye and Law, 2013), and also increasingly utilising reviews in services such as Booking.Com and as Tripadvisor. VGT already has some feedback mechanisms and some historical customer satisfaction data that can be used initially.

Previous studies have looked at a way of pulling the above aspects together and an example is the use of a matrix, an "I-S Model" (Yang, Jou and Cheng, 2011), see an example at Figure 4.1. This categorises the areas of interest and helps highlight those areas performing well and those not, whilst also reflecting their importance to the business. In particular, areas in the "To-be-improved" quadrant are important but have not met customer expectations and so these are areas that a hotel's management team can focus on to improve the business.

4.2 Data Analysis Approach

In a similar way to Demand Forecasting, VGT does not currently formally analyse the experience of its customers across all hotels and then use this to inform management actions to improve its services. However, VGT does have a good history of customer ratings data that

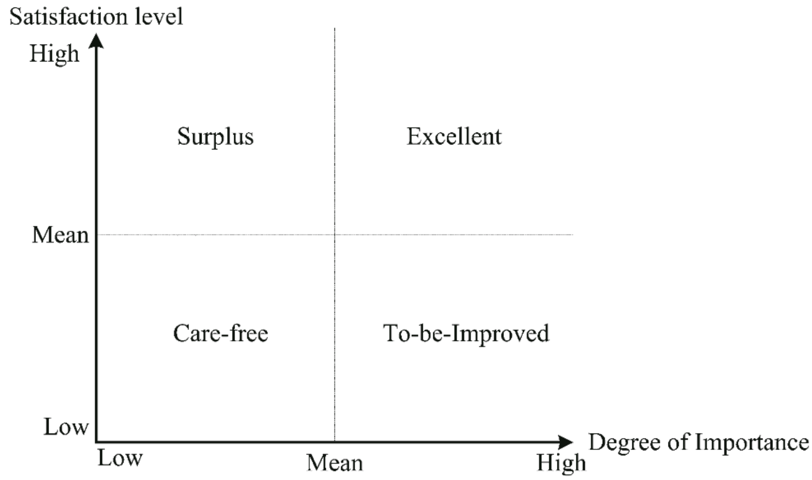


Figure 4.1: Example of an Importance-Satisfaction Model

can be utilised initially to experiment with ways of more formally measuring customer satisfaction. If positive, then this can then be extended in sophistication and scope of data collection (see Section 4.4).

The objective of this initial experimentation step is to establish a process that can utilise existing historical customer review data for a single hotel and create an IS-Model to determine possible actions. In summary, the approach is to:

- Collate customer details and satisfaction ratings
- Utilise these to identify what are the areas that customers regard as important
- Establish a CSAT prediction model
- Combine the above to create an IS-Model for historic data and also for prediction
- Use the IS-Model to identify actions that can improve customer service.

4.2.1 Data Required

The scope of the data is all the historical customer satisfaction ratings that VGT customers have provided after staying at a hotel. For consistency a Customer Satisfaction Score (CSAT) will be used, this is a 5 point Likert scale from "very dissatisfied" to "very satisfied" (Bishop, 2022). Data points required are:

- Overall customer rating
- Customer demographics (Age, Gender, Residence)
- Trip Background (Purpose, Travel Type, Booking Type)
- Individual service ratings (Amenities, Staff, Cleanliness, Wifi etc).

4.2.2 Classification Techniques to Determine Importance

The aim is to use a classification model to determine how the different factors contribute to the overall customer rating, ie to determine the relative importance of each factor. The target variable 'y' is the overall CSAT and the input feature set 'X' consists of the demographics, trip background and individual service ratings. The created model will provide an initial indication of the relative importance across the feature set. The model can then also be used to provide predictions of the CSAT in response to value changes for individual features and this can help hotel management teams focus in planning what actions to take to improve customer service

Potential classification approaches for multi-class classification include:

- Support Vector Machine (SVM)
- K-Nearest Neighbour (KNN)
- Naive Bayes (NB)
- Decision Tree (DT)
- Random Forest (RF)
- Extreme Gradient Boosting (XGBoost).

A comparison of the classification of customer reviews, (Noori, 2021), examined several models, including: SVM, KNN, NB, DT; a DT approach was found to be the most accurate. Another examination of how to identify the important predictors of customer reviews found DT to be an accurate approach(Baouchi, 2018). Extending the DT approach, XGBoost has found to be a successful model in many situations (Chen and Guestrin, 2016). In this initial examination, XGBoost was used.

4.3 Simulated Data Analysis

4.3.1 Data Summary

A dataset was created to simulate the data as defined in the earlier section. The data elements consist of:

	Count	Missing	Empty	Unique	Type	String	Int	Float	List
id	103904	0	0	103904	int64	0	103904	0	0
gender	103904	0	0	2	category	103904	0	0	0
age	103904	0	0	75	int64	0	103904	0	0
purpose_of_travel	103904	0	0	5	category	103904	0	0	0
type_of_travel	103904	0	0	2	category	103904	0	0	0
type_of_booking	103904	0	0	3	category	103904	0	0	0
score_wifi	103904	0	0	5	int64	0	103904	0	0
score_transport	103904	0	0	5	int64	0	103904	0	0
score_booking	103904	0	0	5	int64	0	103904	0	0
score_location	103904	0	0	5	int64	0	103904	0	0
score_restaurant	103904	0	0	5	int64	0	103904	0	0
score_staff	103904	0	0	5	int64	0	103904	0	0
score_parking	103904	0	0	5	int64	0	103904	0	0
score_checkin	103904	0	0	5	int64	0	103904	0	0
score_local_sites	103904	0	0	5	int64	0	103904	0	0
score_housekeeping	103904	0	0	5	int64	0	103904	0	0
score_overall	103904	0	0	5	int64	0	103904	0	0

The first few elements of the loaded data:

```
       id  gender  age purpose_of_travel   type_of_travel type_of_booking  \
0   70172    Male   13          aviation  Personal Travel     Not defined
1    5047    Male   25           tourism     Group Travel  Group bookings
2  110028  Female   26           tourism     Group Travel  Group bookings

   score_wifi  score_transport  score_booking  score_location  \
0           2                3              2               1
1           2                2              2               2
2           2                2              2               2

   score_restaurant  score_staff  score_parking  score_checkin  \
0                 4            4              4              3
1                 1            1              1              1
2                 4            4              4              3

   score_local_sites  score_housekeeping  score_overall
0                  4                   4              1
1                  3                   1              0
2                  3                   4              3
```

And key descriptive statistics:

```
                  id            age     score_wifi  score_transport  \
count  103904.000000  103904.000000  103904.000000    103904.000000
mean    64924.210502      39.379706       2.179839         2.425912
std     37463.812252      15.114964       0.943846         1.131421
min         1.000000       7.000000       0.000000         0.000000
25%     32533.750000      27.000000       2.000000         2.000000
50%     64856.500000      40.000000       2.000000         2.000000
75%     97368.250000      51.000000       3.000000         3.000000
max    129880.000000      85.000000       4.000000         4.000000

       score_booking  score_location  score_restaurant    score_staff  \
count  103904.000000   103904.000000     103904.000000  103904.000000
mean        2.199935        2.333192          2.538324       2.698991
std         1.011443        0.909634          0.965042       0.975912
min         0.000000        0.000000          0.000000       0.000000
25%         2.000000        2.000000          2.000000       2.000000
50%         2.000000        2.000000          2.000000       3.000000
75%         3.000000        3.000000          3.000000       4.000000
max         4.000000        4.000000          4.000000       4.000000

       score_parking  score_checkin  score_local_sites  score_housekeeping  \
count  103904.000000  103904.000000      103904.000000       103904.000000
mean        2.648127       2.552443           2.818900            2.569901
std         0.976920       0.944624           0.898211            0.968967
min         0.000000       0.000000           0.000000            0.000000
25%         2.000000       2.000000           2.000000            2.000000
50%         3.000000       2.000000           3.000000            2.000000
```

```
75%          3.000000      3.000000          4.000000          3.000000
max          4.000000      4.000000          4.000000          4.000000

       score_overall
count  103904.000000
mean        2.211590
std         1.296201
min         0.000000
25%         1.000000
50%         2.000000
75%         3.000000
max         4.000000
```

4.3.2 XGBoost Model Creation & Evaluation

An XGBoost model was created using the above dataset randomly split 80:20 into training and testing data. The initial model was evaluated and a Confusion Matrix, Shap Values and F-Score were generated. The F-Score of 52% was not high whcih indicates the prediction are not very accurate and so more hyper-parameter tuning will be required; however, as an initial examination the model was used.

The relative feature importance ranking was calculated, shown in the figure below. Age clearly has a much larger importance than any of the other features.

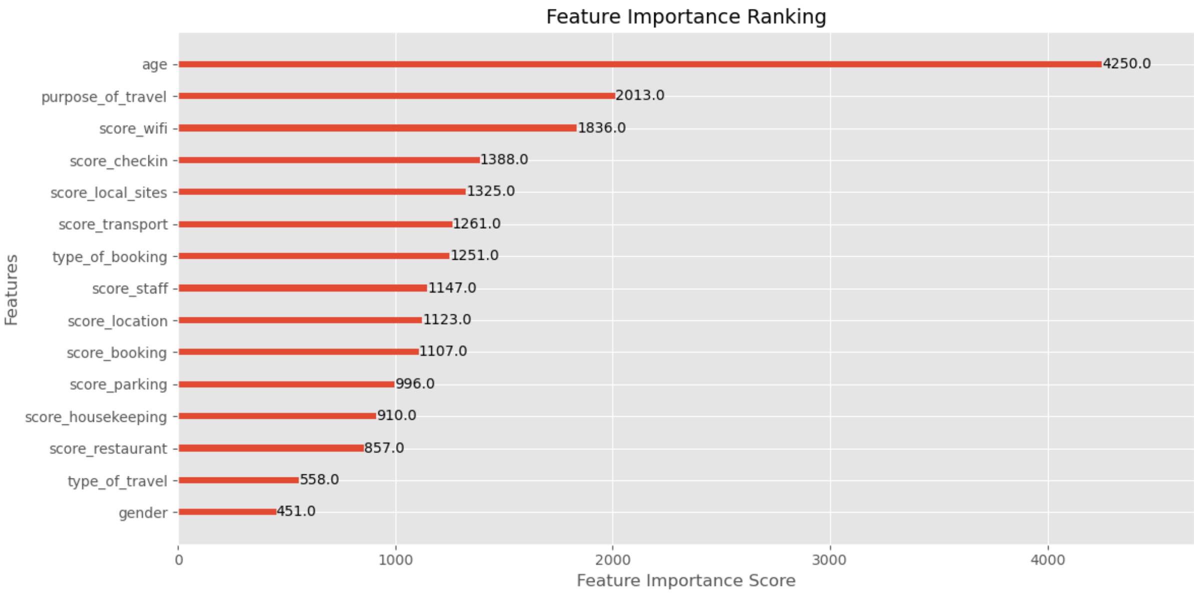


Figure 4.2: XGBoost Feature Importance Results - All Features

It is useful to see what customer demographic factors and travel type impact the overall CSAT, however these can not be directly changed by the hotel's management team, so a second model was generated using just the individual service scores and this feature importance ranking is shown in the figure below. There is a closer spread of importance, with wifi and transport being the most important.

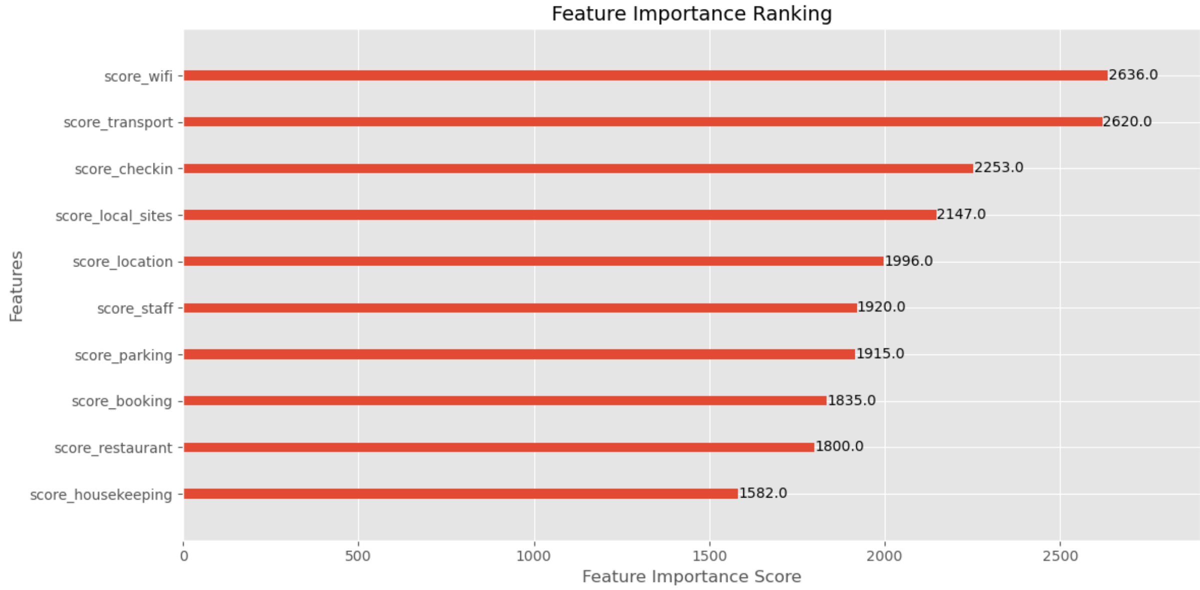


Figure 4.3: XGBoost Feature Importance Results - Service Features Only

4.3.3 I-S Model

Using the predicted feature rankings, an I-S Model can be compared, and an I-S Model created.

	feature	importance	satisfaction_mean
0	score_wifi	66	3.179839
1	score_transport	66	3.425912
2	score_booking	46	3.199935
3	score_location	50	3.333192
4	score_restaurant	45	3.538324
5	score_staff	48	3.698991
6	score_parking	48	3.648127
7	score_checkin	56	3.552443
8	score_local_sites	54	3.818900
9	score_housekeeping	40	3.569901

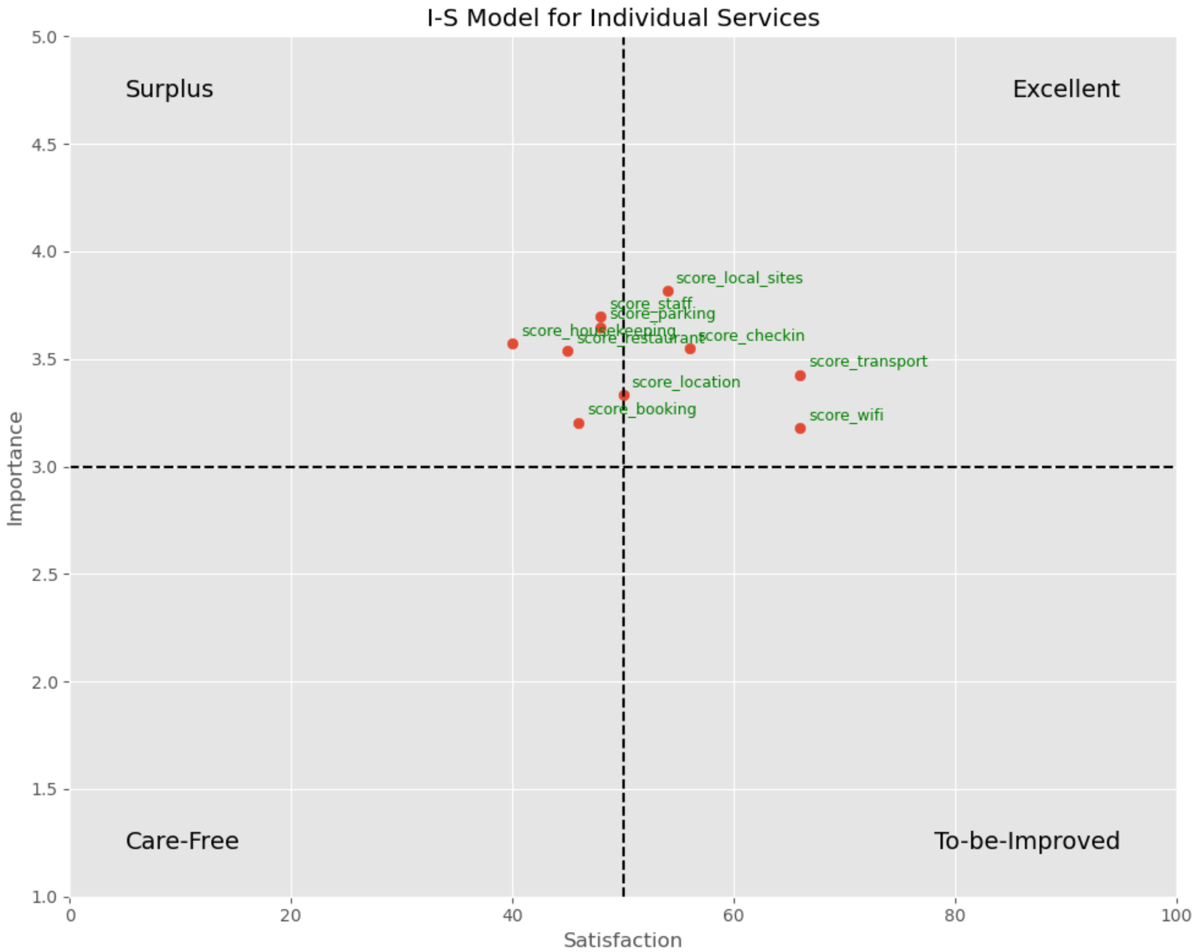


Figure 4.4: Importance-Significance Predictions

4.4 Conclusions & Next Steps

4.4.1 Findings

4.4.2 Next Steps

Use the Model / Predictions / Business Use

- ***?! More processing capacity***
- Look at action areas
- Change feature values and consider impact on overall CSAT
- Monitor changes and impact on scores etc
- Extend to hotels
- ***?? Find a way to plot over time eg Yang, Jou and Cheng (2011) improvement index, p 353***

Modelling Improvements / Refine

- additional customer features that might be useful
- Enhance customer collection approach, standardise across hotels
- third-party / other hotels to help rank importance and provide a comparative/competitor scoring/comparison … Having utilised its own internal data on customer reviews, VGT will be like much of the hotel industry in not significantly incorporating third-party review data from sources such as Tripadvisor and a next-step is to investigate how to extract, analyse and incorporate this data (Park, 2023).
- compare across hotels and also competitors for satisfaction factors etc

5 Conclusions

5.1 Conclusions

5.2 Next Steps

?????? - Overall conclusions - Or possibly conclusions within each section

References

Andrew, W.P., Cranage, D.A. and Lee, C.K. (1990) 'Forecasting Hotel Occupancy Rates with Time Series Models: An Empirical Analysis', *Hospitality Research Journal*, 14(2), pp. 173–182. Available at: https://doi.org/10.1177/109634809001400219.

Baouchi, S. (2018) *A text analytical approach: Predicting and understanding customer satisfaction by making use of customer reviews*. PhD thesis. Available at: https://thesis.eur.nl/pub/53903/thesis-hotel-reviews-430727sb-laatsteversie-.pdf.

Bishop, C. (2022) 'What is customer satisfaction score? (+ how to measure CSAT)'. Available at: https://www.zendesk.co.uk/blog/customer-satisfaction-score/.

Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', in. New York, NY, USA: Association for Computing Machinery (Kdd '16), p. 785794. Available at: https://doi.org/10.1145/2939672.2939785.

Chow, W.S., Shyu, J.-C. and Wang, K.-C. (1998) 'Developing a Forecast System for Hotel Occupancy Rate Using Integrated ARIMA Models', *Journal of International Hospitality, Leisure & Tourism Management*, 1(3), pp. 55–80. Available at: https://doi.org/10.1300/J268v01n03_05.

FHA, F. (2023) 'Understanding occupancy rate in hotels: A comprehensive guide'. Available at: https://fhahoreca.com/glossary/occupancy-rate/.

Huang, L. and Zheng, W. (2022) 'Hotel demand forecasting: a comprehensive literature review', *Tourism Review*, 78(1), pp. 218–244. Available at: https://doi.org/10.1108/TR-07-2022-0367.

Jeffrey, D. and Hubbard, N.J. (1994) 'A model of hotel occupancy performance for monitoring and marketing in the hotel industry', *International Journal of Hospitality Management*, 13(1), pp. 57–71. Available at: https://doi.org/10.1016/0278-4319(94)90059-0.

Johansson, A. (2022) 'Benefits and challenges in forecasting hotel demand'. Available at: https://www.demandcalendar.com/blog/benefits-and-challenges-in-forecasting-hotel-demand.

Li, H., Ye, Q. and Law, R. (2013) 'Determinants of customer satisfaction in the hotel industry: An application of online review analysis', *Asia Pacific journal of tourism research*, 18(7), pp. 784–802. Available at: https://www.tandfonline.com/doi/full/10.1080/10941665.2012.708351.

Lighthouse, A. (2024) 'Hotel ADR: Everything you need to know about Average Daily Rate'. Available at: https://www.mylighthouse.com/resources/blog/hotel-adr-everything-you-need-to-know-about-average-daily-rate.

Lighthouse, B. (2023) 'The hotelier's ultimate guide to occupancy forecasting'. Available at:

https://www.mylighthouse.com/resources/blog/how-to-forecast-hotel-occupancy.

Noori, B. (2021) ‘Classification of customer reviews using machine learning algorithms’, *Applied Artificial Intelligence*, 35(8), pp. 567–588. Available at: https://doi.org/10.1080/08839514.2021.1922843.

Park, J. (2023) ‘Combined text-mining/DEA method for measuring level of customer satisfaction from online reviews’, *Expert Systems with Applications*, 232, p. 120767. Available at: https://doi.org/10.1016/j.eswa.2023.120767.

Phumchusri, N. and Suwatanapongched, P. (2023) ‘Forecasting hotel daily room demand with transformed data using time series methods’, *Journal of Revenue and Pricing Management*, 22(1), pp. 44–56. Available at: https://doi.org/10.1057/s41272-021-00363-6.

Weatherford, L.R. and Kimes, S.E. (2003) ‘A comparison of forecasting methods for hotel revenue management’, *International Journal of Forecasting*, 19(3), pp. 401–415. Available at: https://doi.org/10.1016/S0169-2070(02)00011-0.

Yang, C.-C., Jou, Y.-T. and Cheng, L.-Y. (2011) ‘Using integrated quality assessment for hotel service quality’, *Quality & Quantity*, 45(2), pp. 349–364. Available at: https://doi.org/10.1007/s11135-009-9301-4.

Zarezadeh, Z.Z., Rastegar, R. and Xiang, Z. (2022) ‘Big data analytics and hotel guest experience: A critical analysis of the literature’, *International Journal of Contemporary Hospitality Management*, 34(6), pp. 2320–2336. Available at: https://doi.org/10.1108/IJCHM-10-2021-1293.

A Hotel Demand Forecasting - Jupyter Notebook Output

A.1 Data Load & Characteristics

- Load time series data and look at its characteristics
- Determine autocorrelation, seasonality, stationarity
- Decomposition,
- OLS Regression

A.1.1 Characteristics

	Count	Missing	Empty	Unique	Type	String	Int	Float	List
Standard_OCC	1461	0	0	177	int64	0	1461	0	0
Standard_Capacity	1461	0	0	1	int64	0	1461	0	0
Standard_Rate	1461	0	0	1	int64	0	1461	0	0
Premium_OCC	1461	0	0	101	int64	0	1461	0	0
Premium_Capacity	1461	0	0	1	int64	0	1461	0	0
Premium_Rate	1461	0	0	1	int64	0	1461	0	0

```
            Standard_OCC  Standard_Capacity  Standard_Rate  Premium_OCC  \
Date
2020-01-01           129                254            325           65
2020-01-02           126                254            325           53
2020-01-03           137                254            325           63

            Premium_Capacity  Premium_Rate
Date
2020-01-01               100           575
2020-01-02               100           575
2020-01-03               100           575
```

```
       Standard_OCC  Standard_Capacity  Standard_Rate  Premium_OCC  \
count   1461.000000             1461.0         1461.0  1461.000000
mean     135.646133              254.0          325.0    56.141684
std       50.216654                0.0            0.0    29.349613
min       38.000000              254.0          325.0     0.000000
25%       88.000000              254.0          325.0    31.000000
```

50%	136.000000	254.0	325.0	57.000000
75%	183.000000	254.0	325.0	83.000000
max	254.000000	254.0	325.0	100.000000

	Premium_Capacity	Premium_Rate
count	1461.0	1461.0
mean	100.0	575.0
std	0.0	0.0
min	100.0	575.0
25%	100.0	575.0
50%	100.0	575.0
75%	100.0	575.0
max	100.0	575.0

A.1.2 Autocorrelation, Seasonality, Stationarity

- Determine autocorrelation, seasonality, stationarity .. lag plot, ACF plot, ADF test, Differencing
- Decomposition

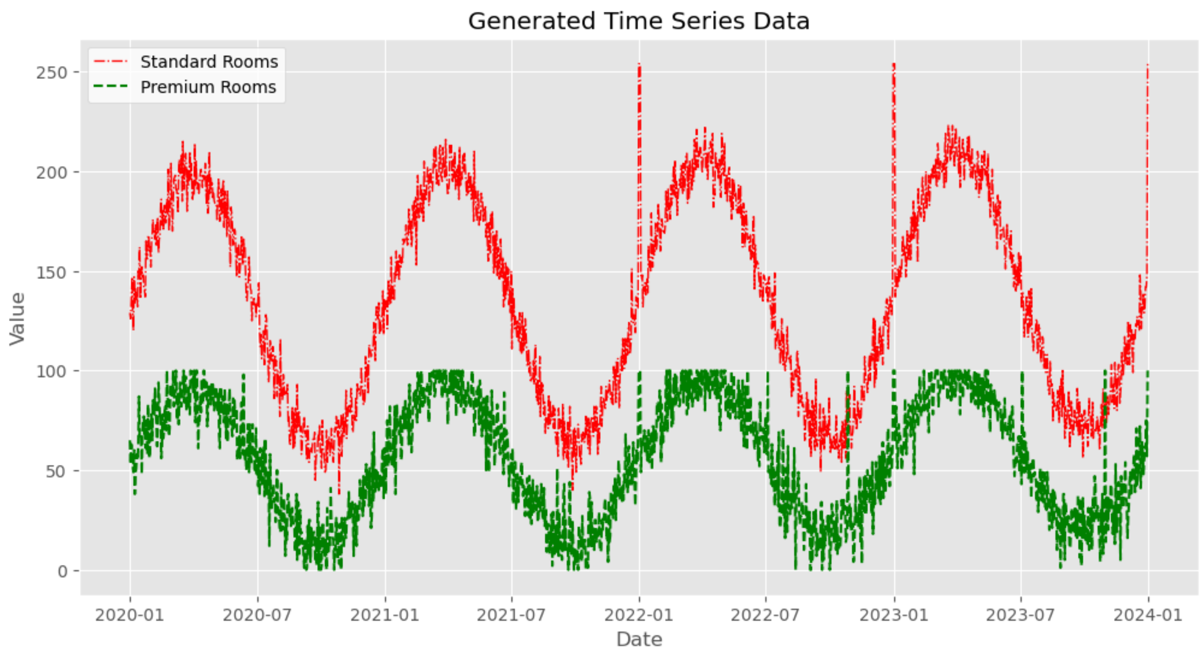


Figure A.1: Occupancy for Each Room Category

- Shows definite annual seasonality with peak high and low seasons
- Also some infrequent spikes in bookings
- Possibly a small upward tend over time
- Premium rooms hit max and zero bookings several times ...

- Both categories of room show definite autocorrelation
- Premium rooms bunched up at max value and autocorrelation may be slight less strong
- Some outliers when rooms are fully booked

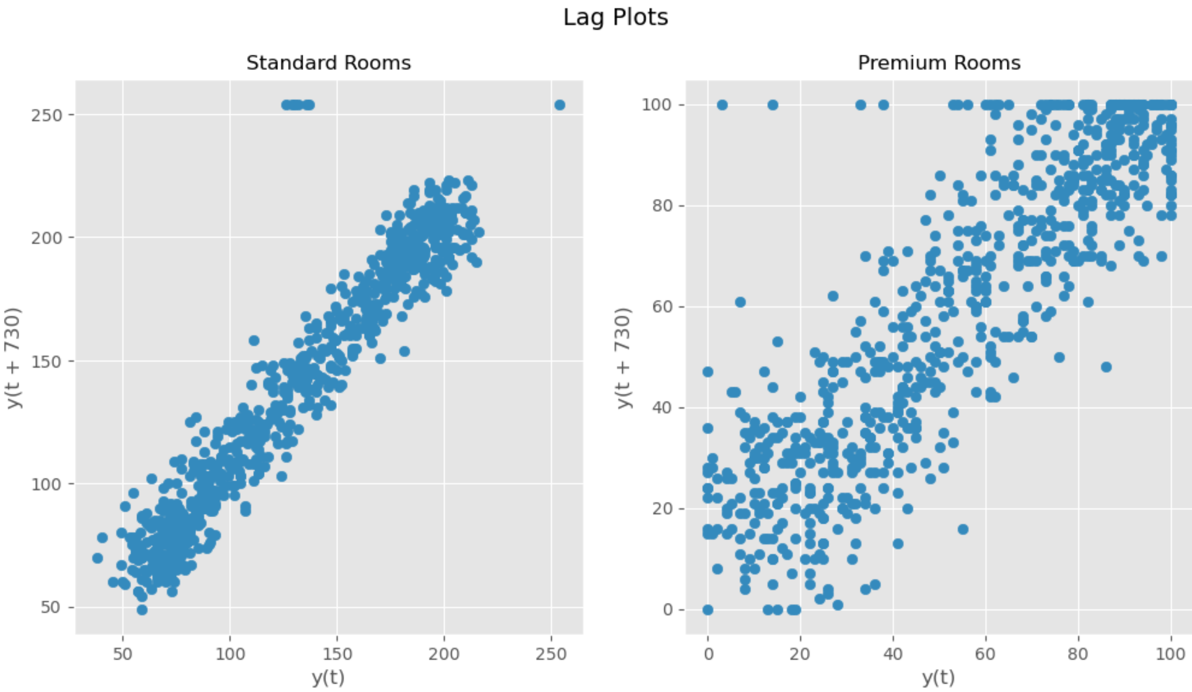
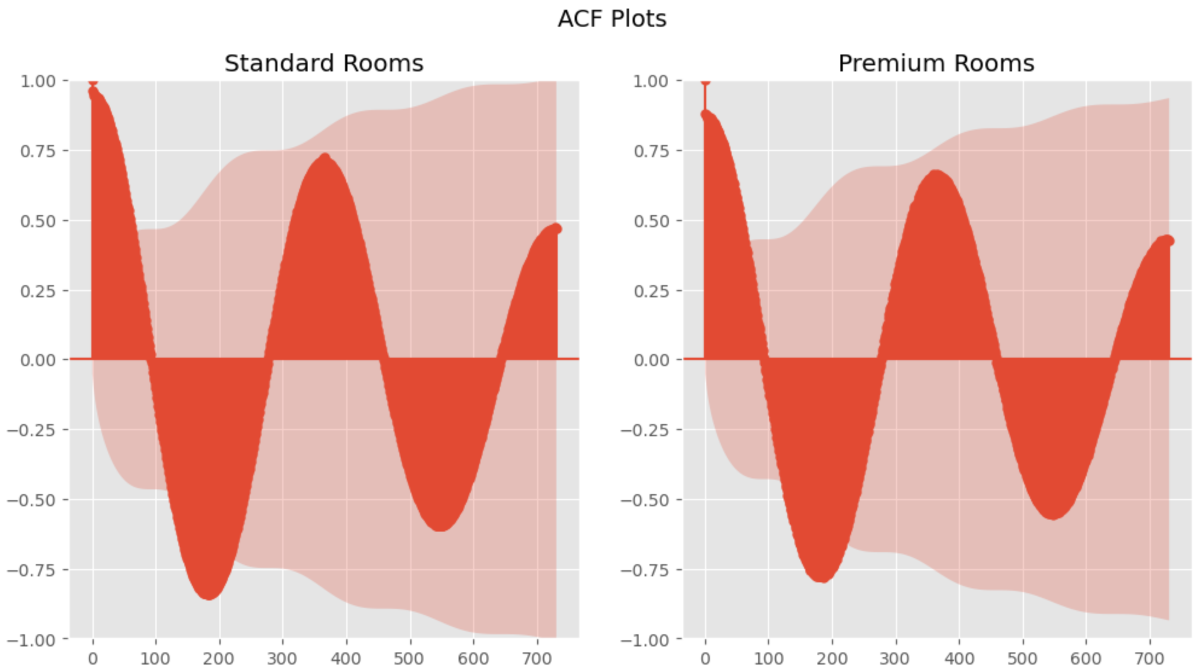


Figure A.2: Lag Plots - Test



- Both exhibit strong autocorrelation that diminishes slowly after approximately 250 days
- A positive trend is suggested by the slowly diminishing autocorrelation
- Multiple peaks at 350 days indicates annual seasonality
- ?? carry out PACF to determine the autoregressive order .... does indicate that it is autoregressive

`ADF Test for Standard Rooms`

```
ADF Statistic: -2.362457191578427
p-value: 0.15261408046089647
Critical Value 1%: -3.434908816804013
Critical Value 5%: -2.863553406963303
Critical Value 10%: -2.5678419239852994
Conclusion: Non-Stationary

ADF Test for Premium Rooms
ADF Statistic: -2.1359470749871265
p-value: 0.2303078418058474
Critical Value 1%: -3.434911997169608
Critical Value 5%: -2.863554810504947
Critical Value 10%: -2.567842671398422
Conclusion: Non-Stationary
```


```
ADF Test for Standard Rooms
ADF Statistic: -4.650429185341465
p-value: 0.00010417359492157454
Critical Value 1%: -3.4349151819757466
Critical Value 5%: -2.863556216004778
Critical Value 10%: -2.5678434198545568
Conclusion: Stationary

ADF Test for Premium Rooms
ADF Statistic: -6.091009300062941
p-value: 1.0365104637060615e-07
Critical Value 1%: -3.4349151819757466
Critical Value 5%: -2.863556216004778
Critical Value 10%: -2.5678434198545568
```

```
Conclusion: Stationary
```

- Confirms that both time series are non-stationary

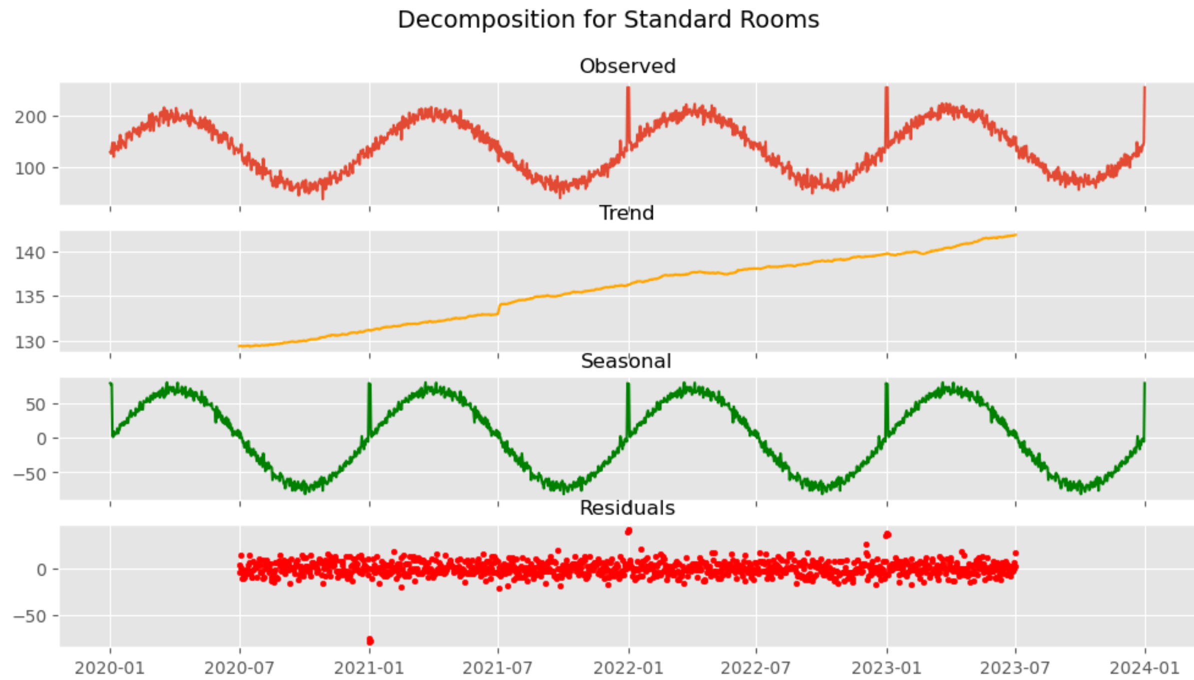
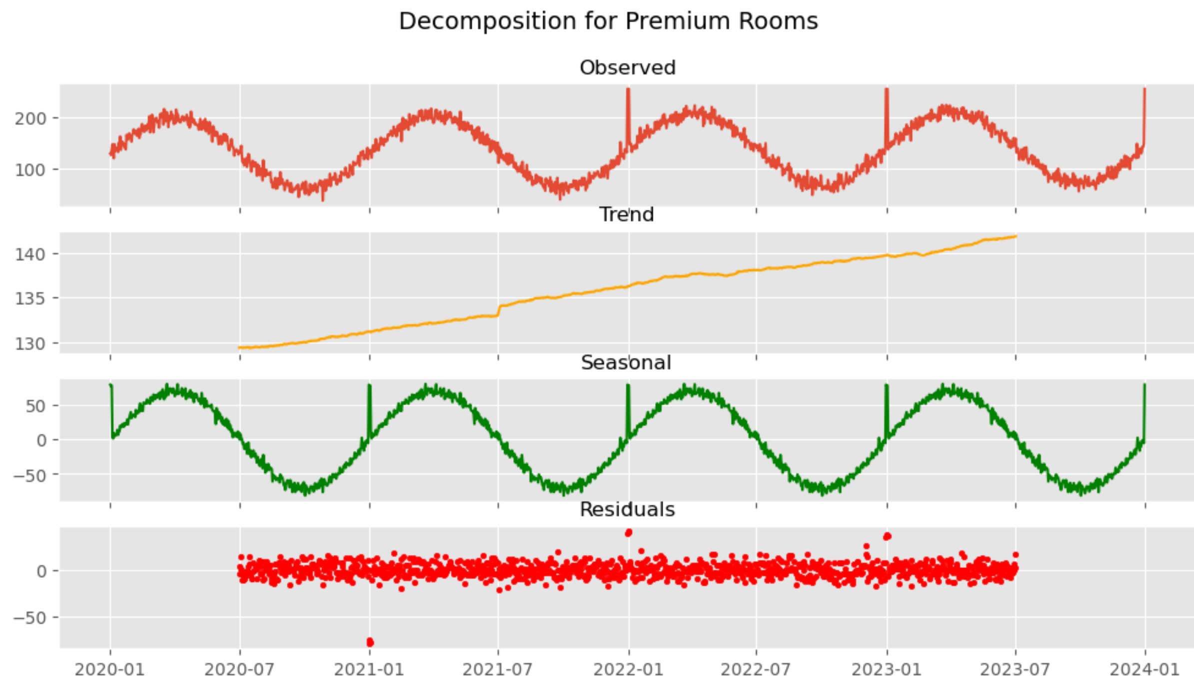


Figure A.3: Occupancy Time Series Decomposition



- Both have a small positive trend with room occupancy increasing 5 to 105% pa
- Confirms both time series are seasonal, with annual peaks and troughs
- On top of the annual seasonality, there are regular spikes leading to 100% occupancy
- Close clustering of residuals with some outliers that correspond to the seasonal spikes

A.2 Ordinary Least Squares (OLS) Linear Regression

- Unlikely to be a good model for forecasting given the strong seasonality, but examine to confirm

```
Durbin-Watson statistic: 0.07265854622866856
Durbin-Watson statistic: 0.24391749000237792
```

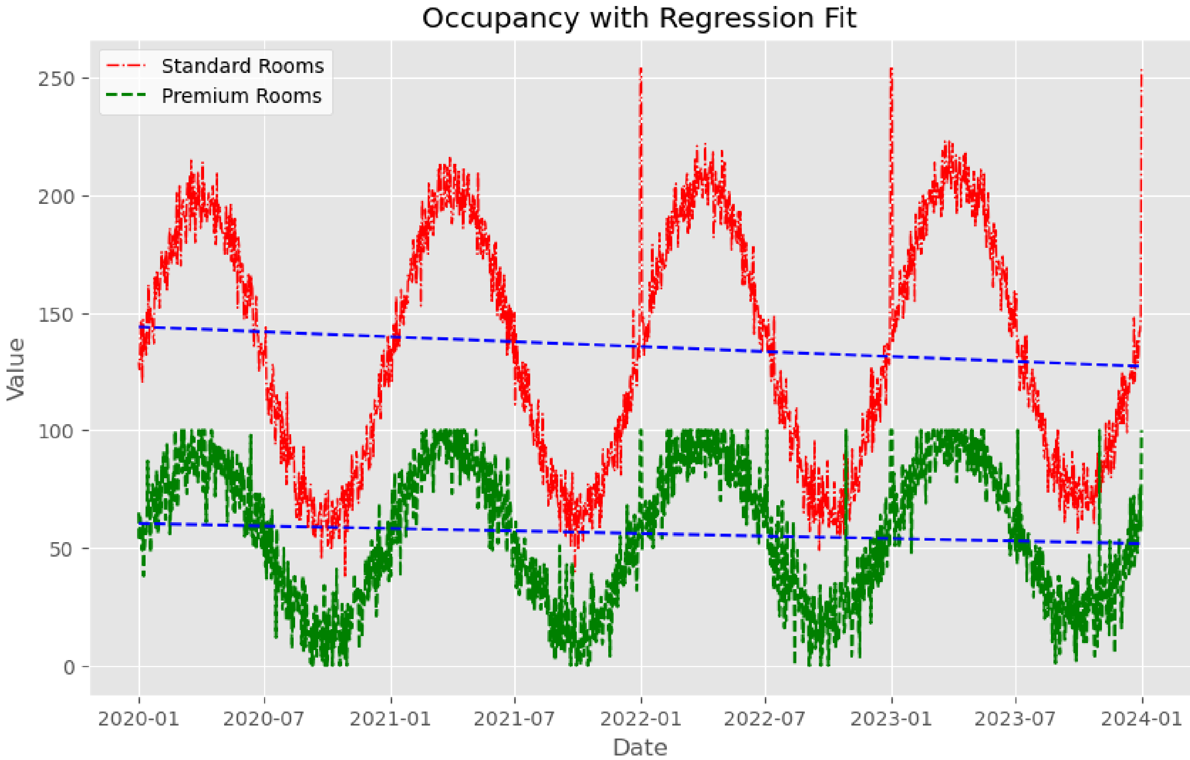


Figure A.4: OLS Linear Regression

- The two fitted lines do not capture any seasonality
- Also show a downward trend line, which is not consistent with the decomposition trend line
- The Durban Watson statistics for both time series are less than 1.5 which confirms evidence of autocorrelation

A.3 SARIMA Model Creation & Forecasting

- The strong autocorrelation for the room occupancy time series, suggest an autoregressive model such as ARMA or ARIMA or SARIMA
- The occupancy history is non-stationary and there is an upward trend, so the integrated component of ARIMA will automatically perform the differencing needed to transform the data into stationary data. So preferable to ARMA
- The strong seasonality suggests SARIMA would be most appropriate as it can process seasonal patterns. So preferable to ARIMA

SARIMA

- nb to be clear SARIMA stands for Seasonal Autoregressive Integrated Moving Average
- S: seasonal component, here handle the annual occupancy seasonality
- AR: autoregressive component …. p
- I: integrated, here handle the positive historical trend through differencing to make it stationary …. d
- MA: moving average component …… q
- SARIMA(p,d,q)(P,D,Q),m
- p,d,q non-seasonal
- P,D,Q seasonal
- m the seasonality period, length of the seasonal cycle
- !! But this is daily, with an annual cycle so would be m=365 which is too large ??

Residuals …. ?? - Residuals plot - ACF plot of residuals - Durbin-Watson of residuals …. ?? be close to 2

Approach - Confirm/Assess the autoregressive order (AR, p) and moving average order (MA, q). Using PACF plot and ACF plot respectively [repeat autocorrelation findings from previously?] - Use the SARIMA(p,d,q)(P,D,Q,m) model - Identify the best parameters using Auto Arima and using AIC (Akaike Information Criterion) to compare

- ?? daily 365d is too much

- ?? First evaluate using residuals ….

- Split data and use 3 years history to the train the model, then 1 year to evaluate its accuracy

- ??? Use accuracy measures MAE, RMSE, MAPE ….

Forecasting With model - Make month forecast and demonstrate its use in the revenue/occupancy spreadsheet

A.3.1 Data Load & SARIMA Model Factors

- Load time series data
- Assess the autoregressive order (AR, p) - Using PACF plot
- Assess moving average order (MA, q) - Using ACF plot

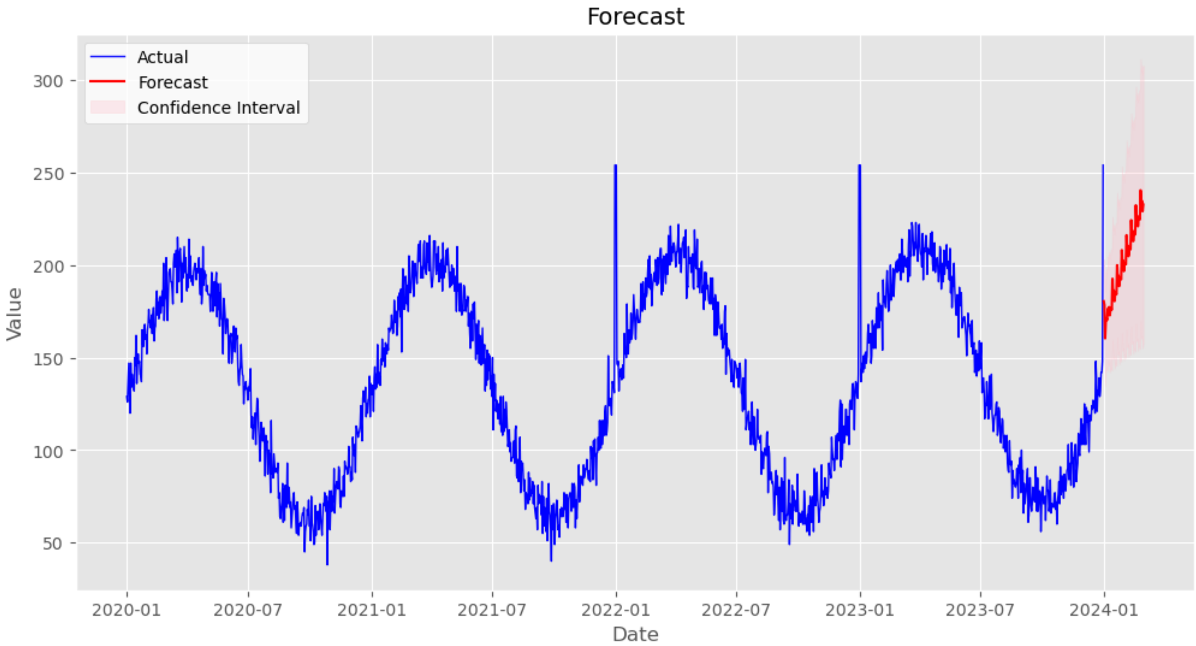
PACF plot suggests AR order, p of approximately 3

ACF plot suggests MA order q of approximately 60

A.3.2 Create the SARIMA Model

- Save the training dataset to be used to train a SARIMA model, in a separate notebook
- Load the model in here for forecasting

A.3.3 Forecast Using the SARIMA Model



B Customer Satisfaction - Jupyter Notebook Output

B.1 TO DO:

- SHAP Values use? why only 5 interactions?
- Classification report, confusion matrix interpretation
- Other evaluation metrics?
- Add comments and/or tables to main report?
- Combine customer comments from other file .. for Sentiment analysis
- ?? use importance ranking on page 357 as a starting point for analysis .. refine later by each hotel's management team
- Evaluation of model : https://www.datacamp.com/blog/classification-machine-learning
- SHAP: https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

B.2 Data Load & Characteristics

- Load data and look at its characteristics

B.2.1 Libraries & Functions

B.2.2 Characteristics

	Count	Missing	Empty	Unique	Type	String	Int	Float	List
id	103904	0	0	103904	int64	0	103904	0	0
gender	103904	0	0	2	category	103904	0	0	0
age	103904	0	0	75	int64	0	103904	0	0
purpose_of_travel	103904	0	0	5	category	103904	0	0	0
type_of_travel	103904	0	0	2	category	103904	0	0	0
type_of_booking	103904	0	0	3	category	103904	0	0	0
score_wifi	103904	0	0	5	int64	0	103904	0	0
score_transport	103904	0	0	5	int64	0	103904	0	0
score_booking	103904	0	0	5	int64	0	103904	0	0
score_location	103904	0	0	5	int64	0	103904	0	0
score_restaurant	103904	0	0	5	int64	0	103904	0	0

	Count	Missing	Empty	Unique	Type	String	Int	Float	List
score_staff	103904	0	0	5	int64	0	103904	0	0
score_parking	103904	0	0	5	int64	0	103904	0	0
score_checkin	103904	0	0	5	int64	0	103904	0	0
score_local_sites	103904	0	0	5	int64	0	103904	0	0
score_housekeeping	103904	0	0	5	int64	0	103904	0	0
score_overall	103904	0	0	5	int64	0	103904	0	0

```
       id  gender  age purpose_of_travel   type_of_travel type_of_booking  \
0   70172    Male   13          aviation  Personal Travel     Not defined
1    5047    Male   25           tourism     Group Travel  Group bookings
2  110028  Female   26           tourism     Group Travel  Group bookings

   score_wifi  score_transport  score_booking  score_location  \
0           2                3              2               1
1           2                2              2               2
2           2                2              2               2

   score_restaurant  score_staff  score_parking  score_checkin  \
0                 4            4              4              3
1                 1            1              1              1
2                 4            4              4              3

   score_local_sites  score_housekeeping  score_overall
0                  4                   4              1
1                  3                   1              0
2                  3                   4              3
```

```
                  id            age     score_wifi  score_transport  \
count  103904.000000  103904.000000  103904.000000    103904.000000
mean    64924.210502      39.379706       2.179839         2.425912
std     37463.812252      15.114964       0.943846         1.131421
min         1.000000       7.000000       0.000000         0.000000
25%     32533.750000      27.000000       2.000000         2.000000
50%     64856.500000      40.000000       2.000000         2.000000
75%     97368.250000      51.000000       3.000000         3.000000
max    129880.000000      85.000000       4.000000         4.000000

       score_booking  score_location  score_restaurant    score_staff  \
count  103904.000000   103904.000000     103904.000000  103904.000000
mean        2.199935        2.333192          2.538324       2.698991
std         1.011443        0.909634          0.965042       0.975912
min         0.000000        0.000000          0.000000       0.000000
25%         2.000000        2.000000          2.000000       2.000000
50%         2.000000        2.000000          2.000000       3.000000
75%         3.000000        3.000000          3.000000       4.000000
```

```
max         4.000000        4.000000          4.000000       4.000000

       score_parking  score_checkin  score_local_sites  score_housekeeping  \
count  103904.000000  103904.000000      103904.000000       103904.000000
mean        2.648127       2.552443           2.818900            2.569901
std         0.976920       0.944624           0.898211            0.968967
min         0.000000       0.000000           0.000000            0.000000
25%         2.000000       2.000000           2.000000            2.000000
50%         3.000000       2.000000           3.000000            2.000000
75%         3.000000       3.000000           4.000000            3.000000
max         4.000000       4.000000           4.000000            4.000000

       score_overall
count  103904.000000
mean        2.211590
std         1.296201
min         0.000000
25%         1.000000
50%         2.000000
75%         3.000000
max         4.000000
```

B.3 XGBoost Model Creation & Forecasting

B.3.1 Initial XGBoost Model Creation

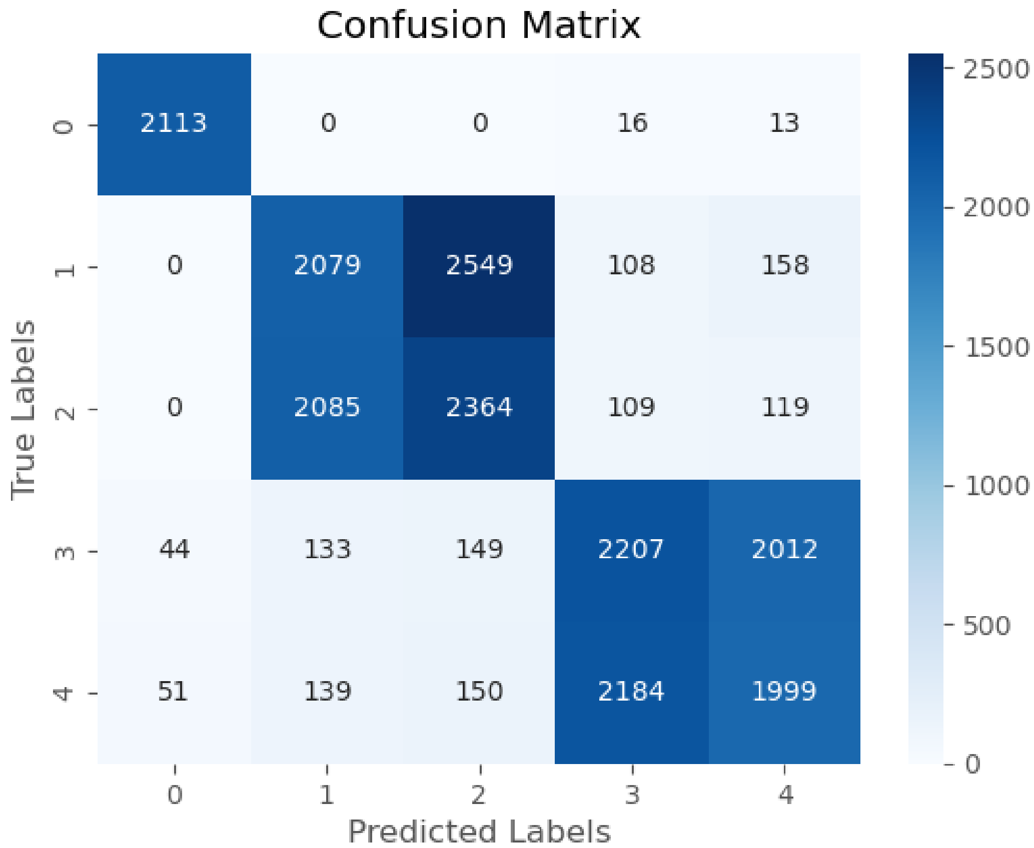
```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=True, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, objective='multi:softprob', ...)
```

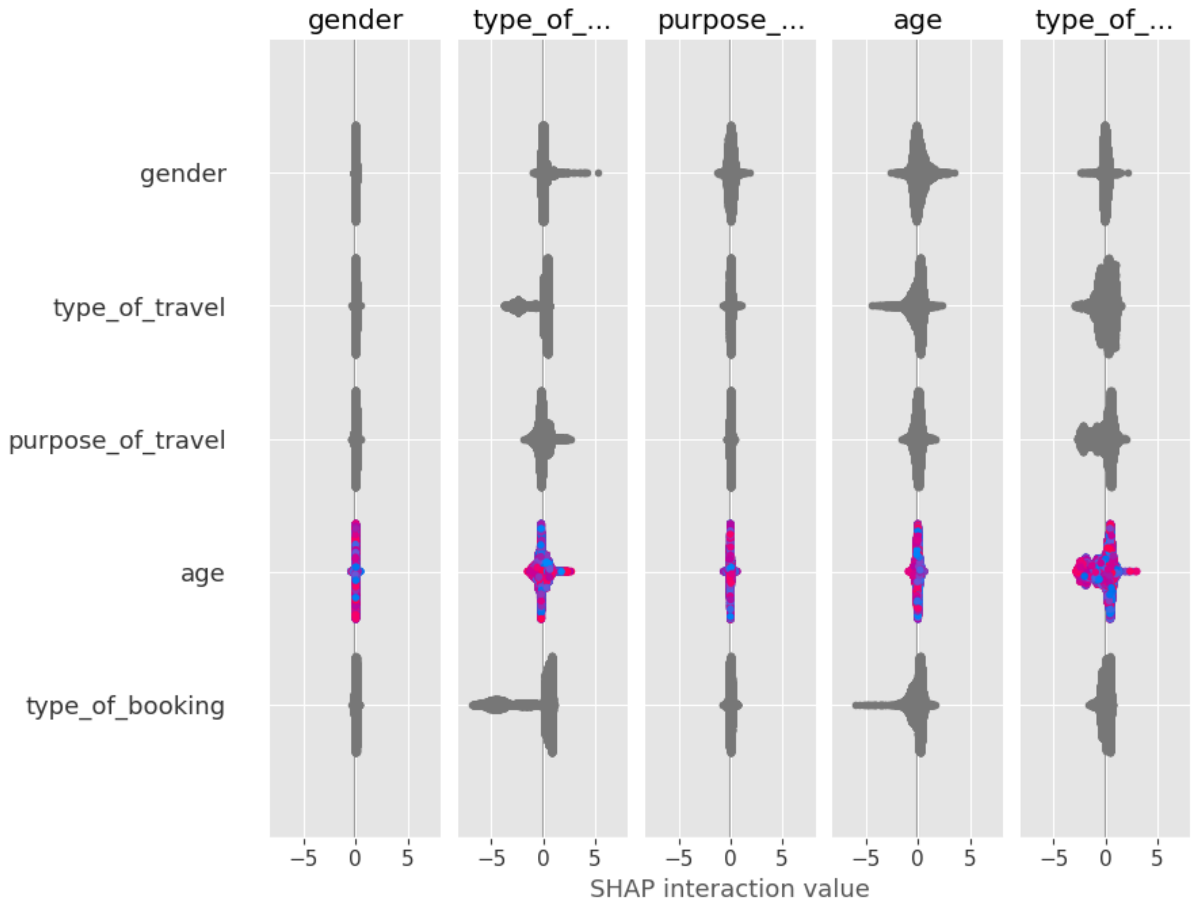
B.3.2 Evaluate The XGBoost Model

```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97      2142
           1       0.47      0.42      0.45      4894
           2       0.45      0.51      0.48      4677
           3       0.48      0.49      0.48      4545
           4       0.46      0.44      0.45      4523
```

accuracy			0.52	20781
macro avg	0.56	0.57	0.57	20781
weighted avg	0.52	0.52	0.52	20781

Classification Error: 0.48





B.3.3 XGBoost Model - Features Ranking

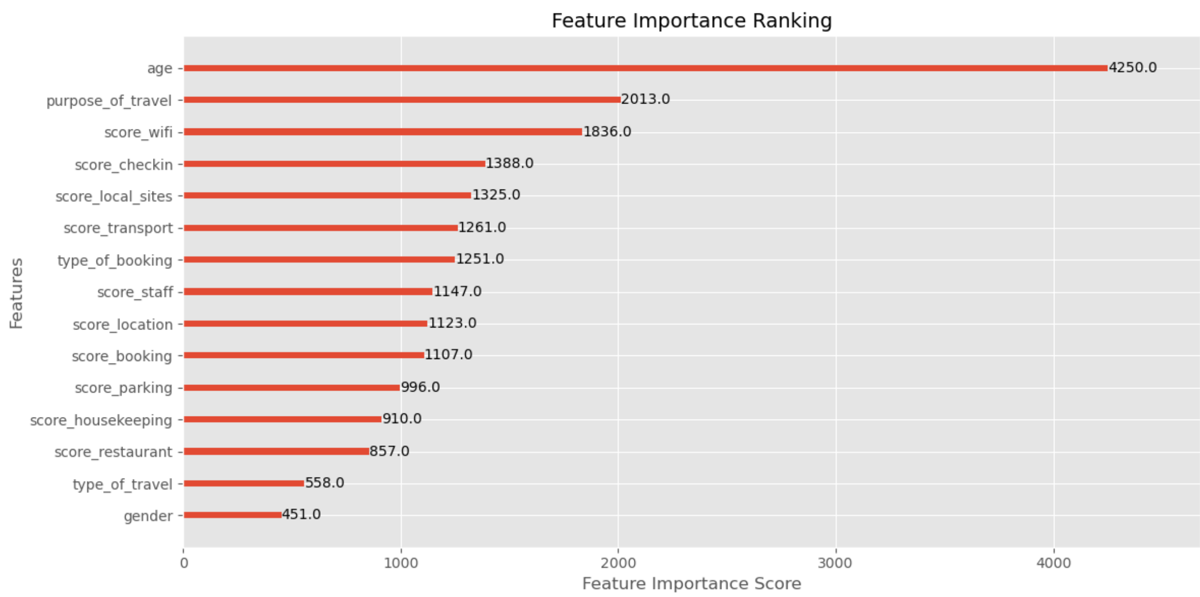
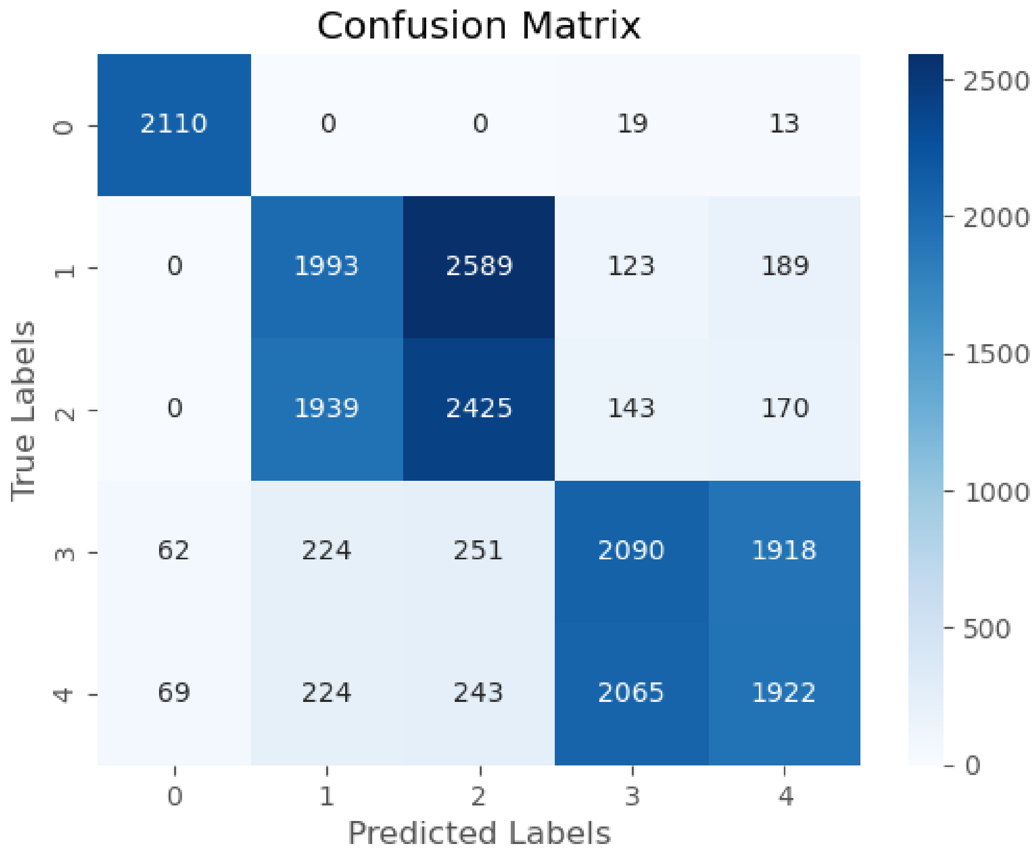
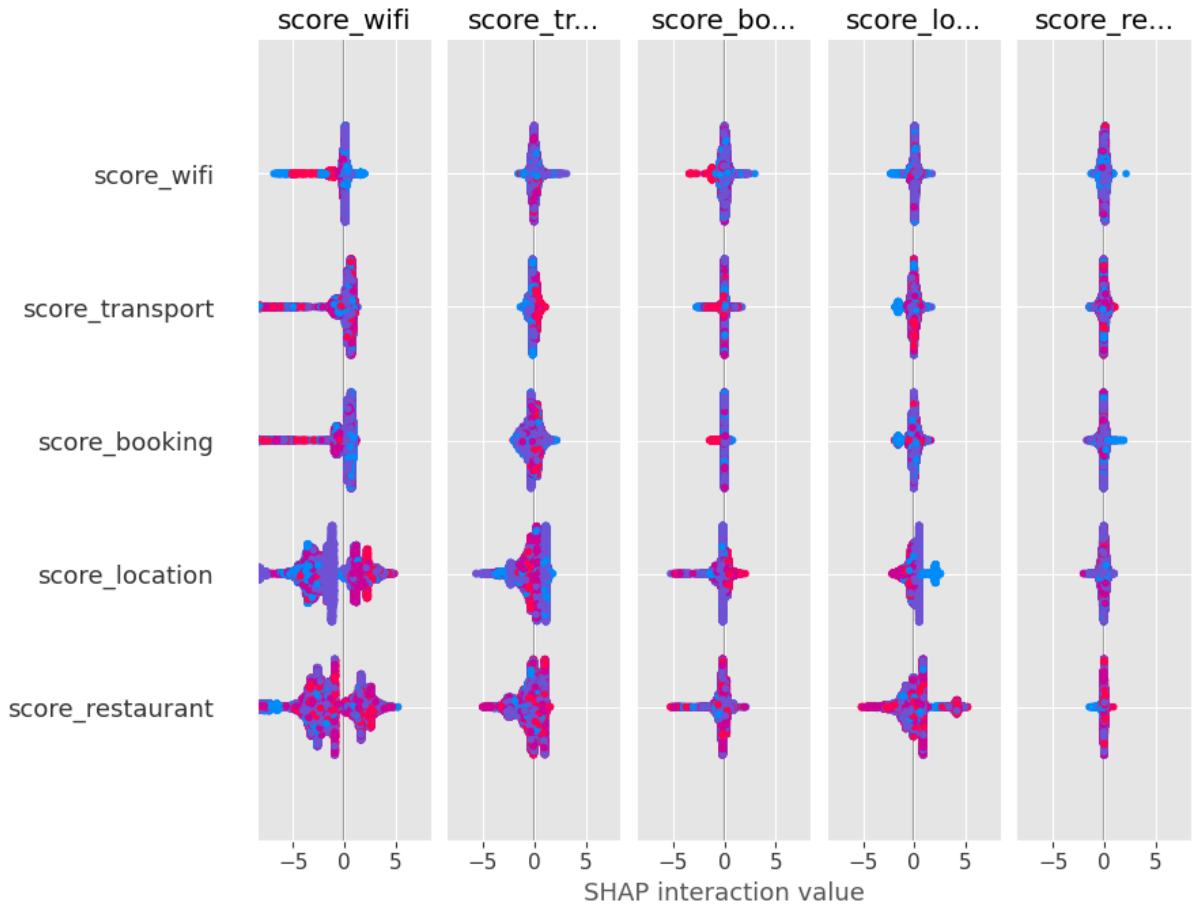


Figure B.1: XGBoost Feature Importance Results - All Features

B.3.4 XGBoost Model - Features Subset

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=True, eval_metric=None, feature_types=None,
              gamma=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=None, n_jobs=None,
              num_parallel_tree=None, objective='multi:softprob', ...)

              precision    recall  f1-score   support

           0       0.94      0.99      0.96      2142
           1       0.46      0.41      0.43      4894
           2       0.44      0.52      0.48      4677
           3       0.47      0.46      0.47      4545
           4       0.46      0.42      0.44      4523

    accuracy                           0.51     20781
   macro avg       0.55      0.56      0.55     20781
weighted avg       0.51      0.51      0.51     20781

Classification Error: 0.49
```


Confusion Matrix
True Labels
Predicted Labels
0
1
2
3
4
2110
0
0
19
13
0
1993
2589
123
189
0
1939
2425
143
170
62
224
251
2090
1918
69
224
243
2065
1922
2500
2000
1500
1000
500
0



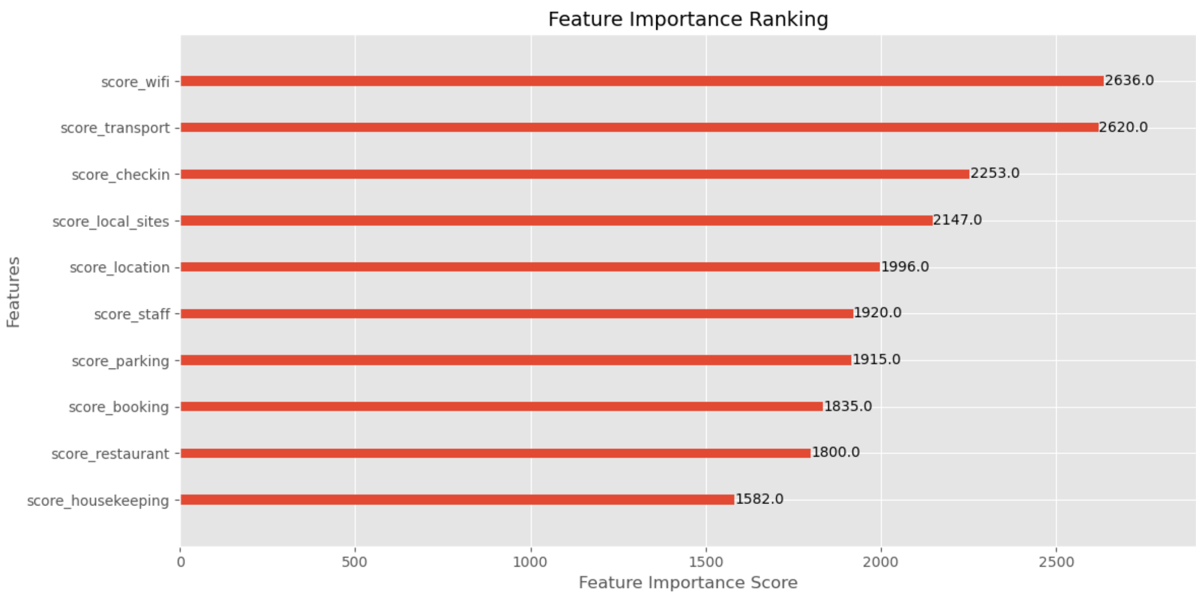
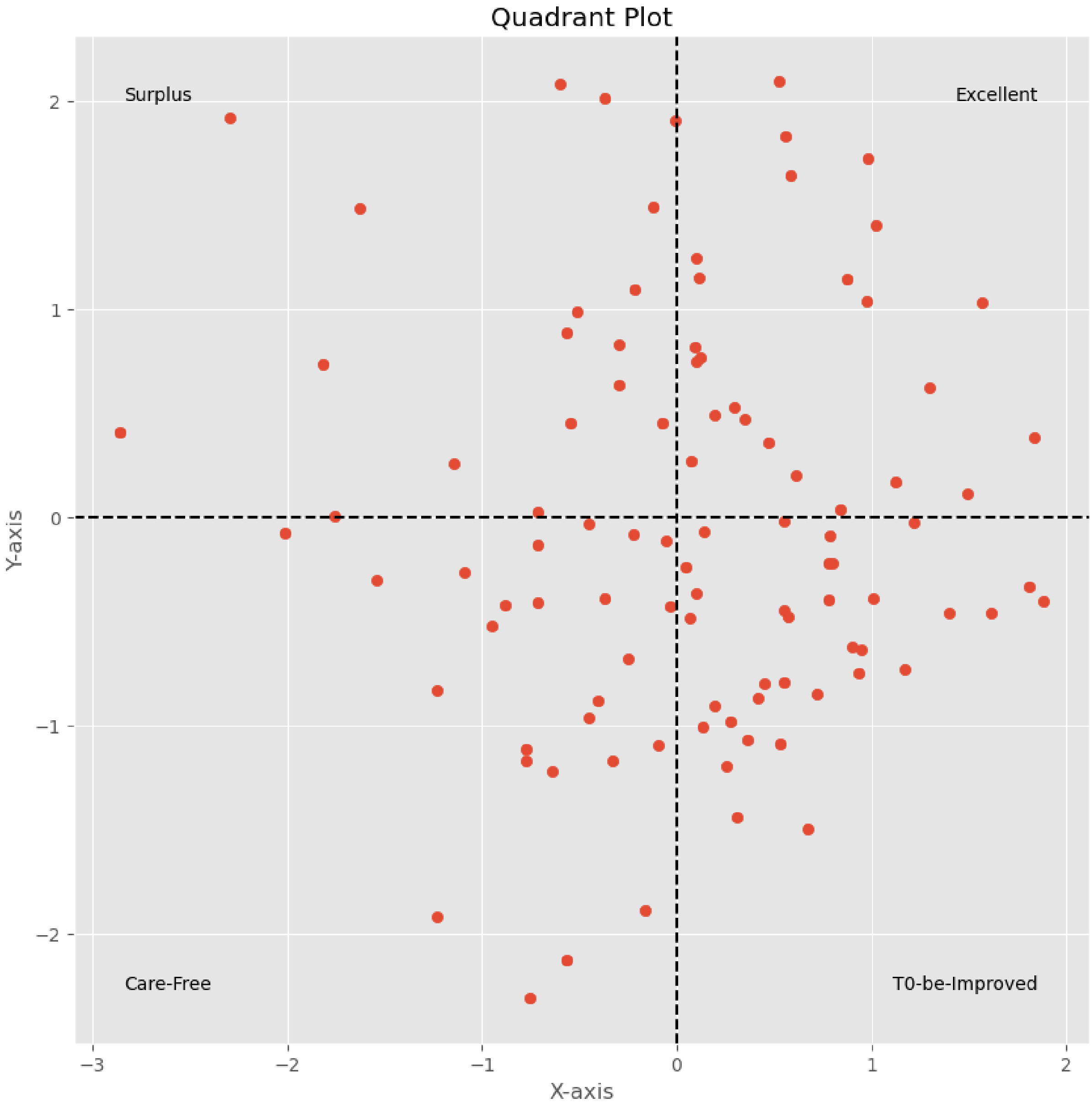


Figure B.2: XGBoost Feature Importance Results - Service Features Only

	feature	importance	satisfaction_mean
0	score_wifi	66	3.179839
1	score_transport	66	3.425912

	feature	importance	satisfaction_mean
2	score_booking	46	3.199935
3	score_location	50	3.333192
4	score_restaurant	45	3.538324
5	score_staff	48	3.698991
6	score_parking	48	3.648127
7	score_checkin	56	3.552443
8	score_local_sites	54	3.818900
9	score_housekeeping	40	3.569901



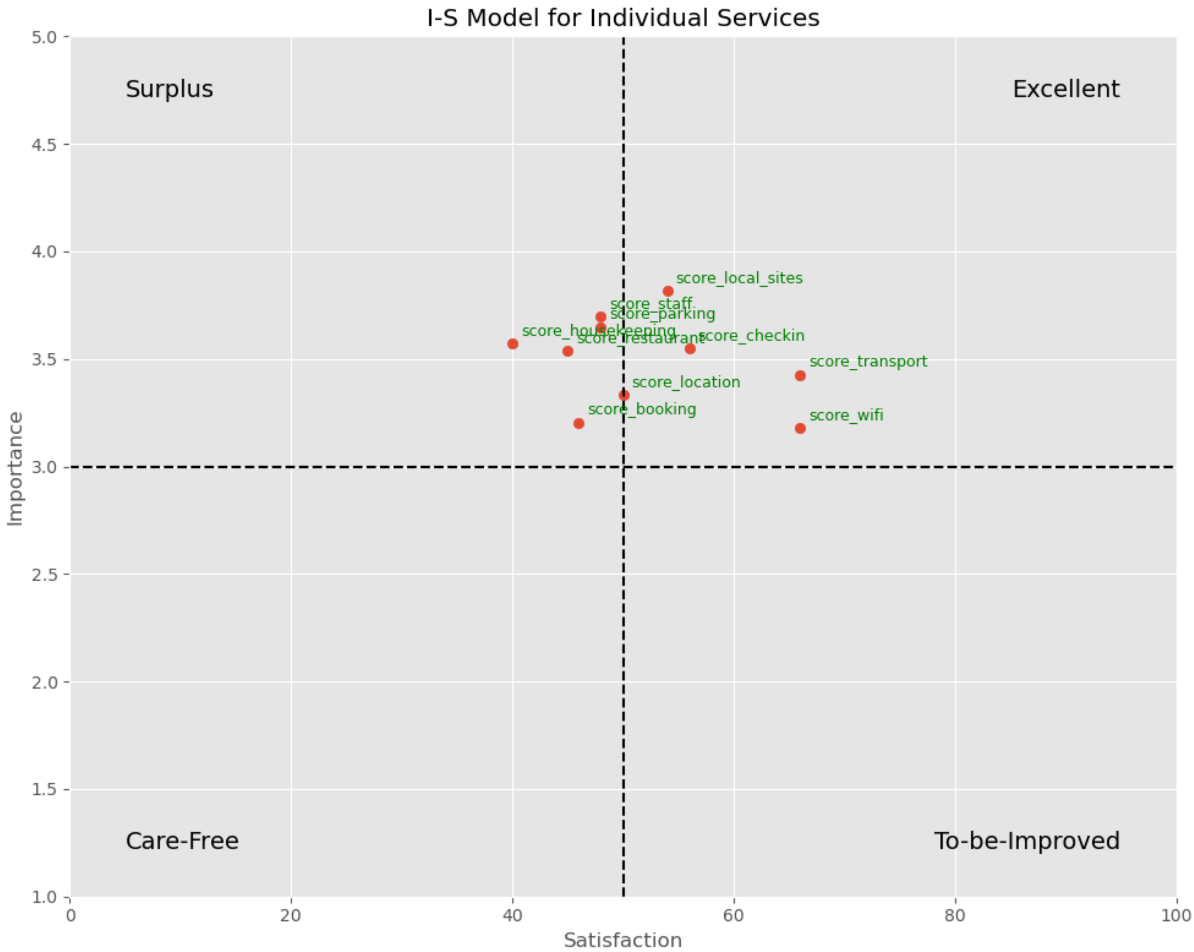


Figure B.3: Importance-Significance Predictions