CFG +Masters A.I/ ML - Avril Childs

Predicting High-Occupancy Short-Term Rentals: A Machine Learning Approach for Property Investment Decisions

Real-World Scenario

A holiday lettings company is planning to expand its portfolio of serviced accommodations across Essex. The company seeks assistance in identifying the most suitable investment properties by analysing historical data on occupancy rates and revenue of similar listings. It also wishes to look at other possible influencing factors in a listing's earning potential.

Aims and Objectives

The project is structured into two main phases:

Phase 1 : Identify the features that mostly correlates with occupancy rates and annual revenue

Phase 2: Build a machine learning model that predicts which properties are likely to achieve the high occupancy rates and revenue.

Study Features (Independent Variables)

Category	Feature	Description	Data Type
Property	Bedrooms	Total bedrooms in the property	Discrete Numerical
	Bathrooms	Total bathrooms in the property	Discrete Numerical
	Max Guests	Maximum number of guests allowed	Discrete Numerical
	Location (Town)	Town in Essex where the property is situated	Categorical
Host	Ratings	Average rating given to the host	Continuous Numerical
	Number of Reviews	Total reviews received	Discrete Numerical
Amenity	Beachfront	Whether the property is located near the sea	Binary
	Hot Tub	Availability of a hot tub	Binary
	Pets Allowed	Whether pets are allowed	Binary

	Smoking Allowed	Whether smoking is permitted	Binary
Listing	Length of Stay	Min/max nights allowed per stay	Discrete Numerical
	Active Days	Days the listing is bookable	Discrete Numerical
	Blocked Days	Days the listing is blocked	Discrete Numerical
	Listed price	Nightly price of the listing	Continuous Numerical
	Average Daily Rate	Average nightly pricing based on revenue and occupancy	Continuous Numerical

PHASE 1

Target Variables

Annual Revenue Total earnings per listing over a 365-day period. This depends on price and occupancy.

Occupancy Rate Percentage of days a listing is booked versus days it is active

The target variables will be categorised into high and low for correlational study and modelling. The exploratory analysis within phase 1 will assist in determining the thresholds.

The Data

The data was retrieved from PriceLabs (https://www.pricelabs.co), a subscription-based platform used by property investors, landlords and agents. PriceLabs aggregates serviced accommodation data from platforms like Airbnb, Booking.com and Vrbo.

- Two .csv datasets were downloaded for analysis
- The file paths within the Juptyer Notebook to match user local file system in order to use the dataset correctly.

Summary of Phase 1 Methodology

This first phase analysis will focus on the following:

- Exploratory Data Analysis and Data cleaning: To manage missing data, fixing outliers, handling duplicates etc.

- Pre-processing and transformation: Merging datasets, Encoding and transforming columns as needed.
- Target vs Feature Analysis: Correlational study will be conducted according to the data type (ie. continuous/discrete numerical vs binary/categorical).

Data Cleaning and Standardisation

The following data cleaning and standardisation procedures were conducted:

Remove duplicate listings: Duplicates can occur within the dataset as one property can feature within 2 overlapping locations (towns). The data set was checked for any duplicate listings based on listing ID and discovered there were 186 duplicates. These duplicates were dropped.

Remove irrelevant columns: The columns that were irrelevant to further analysis were removed.

Missing Values: A heatmap was used to visualise the patterns of null/missing values within the dataset.

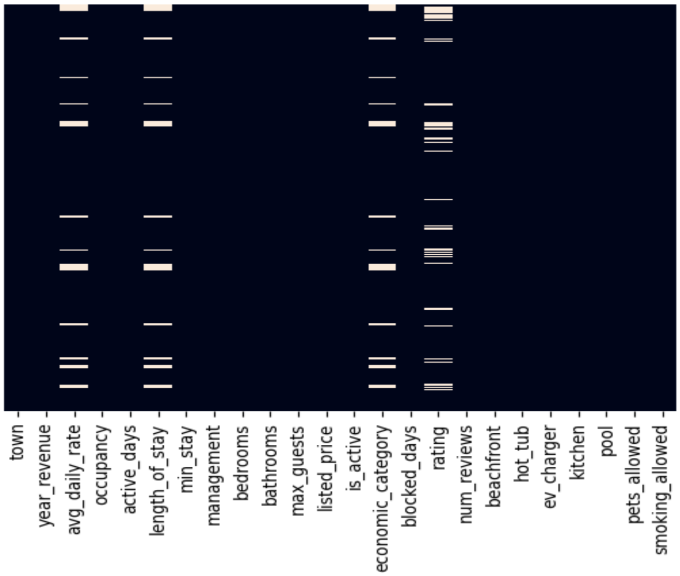


Fig.1 Heatmap of Missing values

Outlier management: One of the primary use of PriceLabs data is for agents and airbnb owners to determine listing prices. The columns (features) "booking_window" and "dynamic_pricing" are used to determine listing prices, and not our target outcome variables of revenue or occupancy. They are therefore is not relevant to our analysis. These columns/ features are excluded from our study.

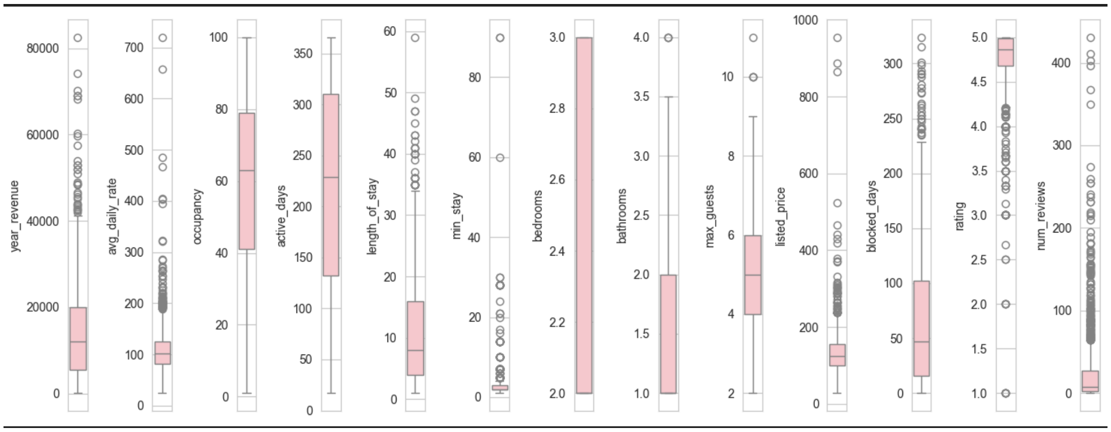


Fig. 2. Boxplots for outlier detection

Winsorization method was used to reduce impact of extreme values on the dataset and was applied to the following:

- Annual revenue
- Average Daily Rate
- Listed Price
-

Capping (Thresholds) was deemed more appropriate in addressing outliers for the following:

- Minimum length of stay
- Number of blocked days

Exploratory Data Analysis

Descriptive analysis of features were analysed using comparative tables and data visualisations such as pie charts, distribution graphs.

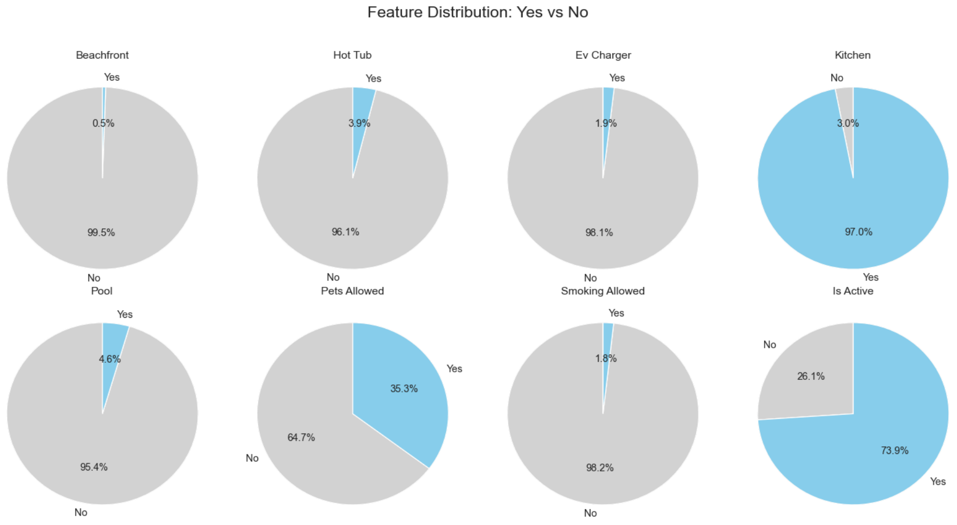


Fig 3. Pie chart demonstrating proportion of yes/no values in binary features

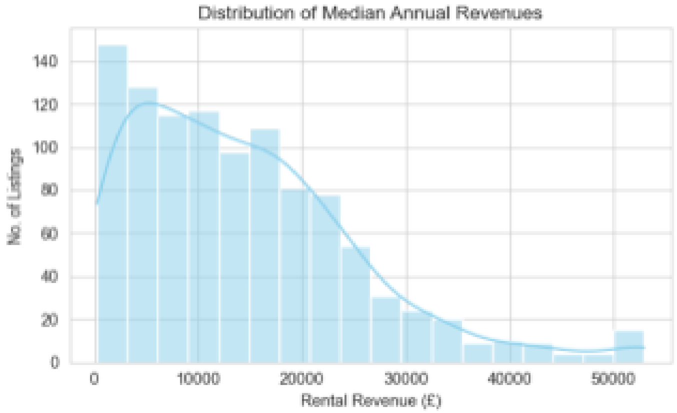


Fig. 4 Distribution plot for annual rental revenue

Both revenue and occupancy data and their percentile values were analysed to determine threshold values of “High” and “Low” which is used for correlational/comparative study and modelling. These values are selected from median values:

High Revenue (£): ‘year_revenue’ > = 12500
Low Revenue (£): ‘year_revenue’ < 12500

High Occupancy (%): ‘occupancy’ > = 63
Low Occupancy (%): ‘occupancy’ < 63

Comparative Analysis of features and target variables were dealt with depending on whether the feature is a continuous/discrete numerical or categorical/ binary.

Continuous/ discrete numerical data were visualised using violinplots.

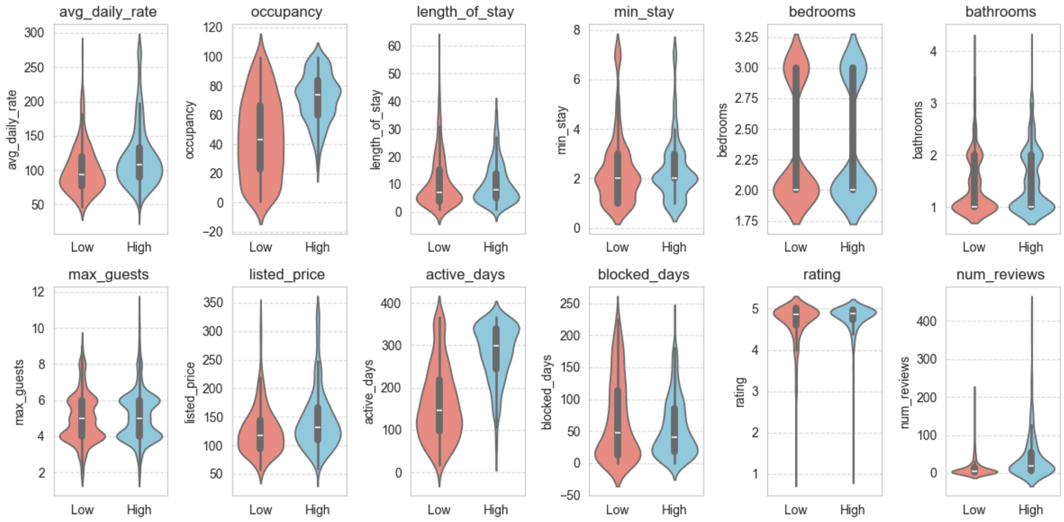


Fig.5 Numerical Features vs Revenue Violinplots using High vs Low Revenue Markers

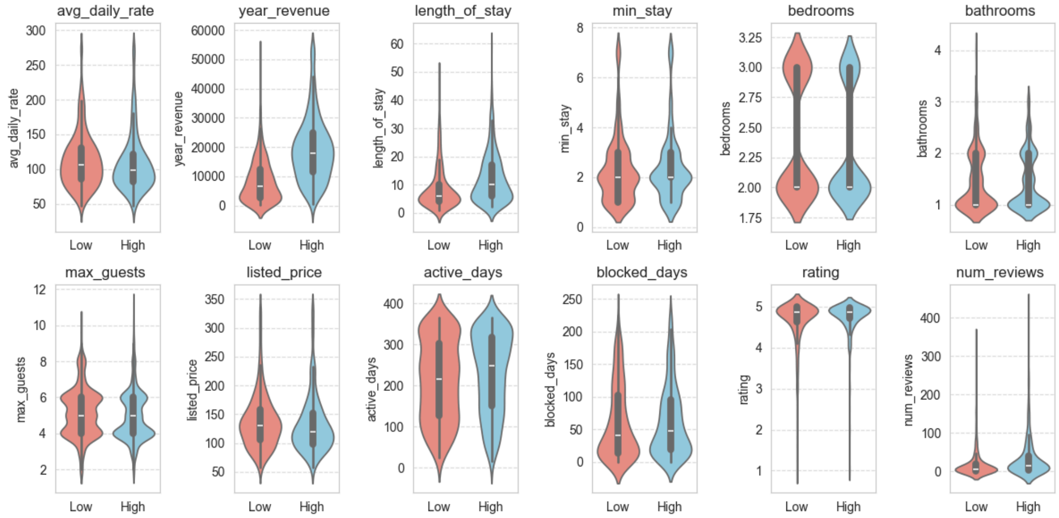


Fig. 6 Numerical Features vs Occupancy Violinplots using High vs Low Occupancy marker

Bar charts were used for categorical/ binary features

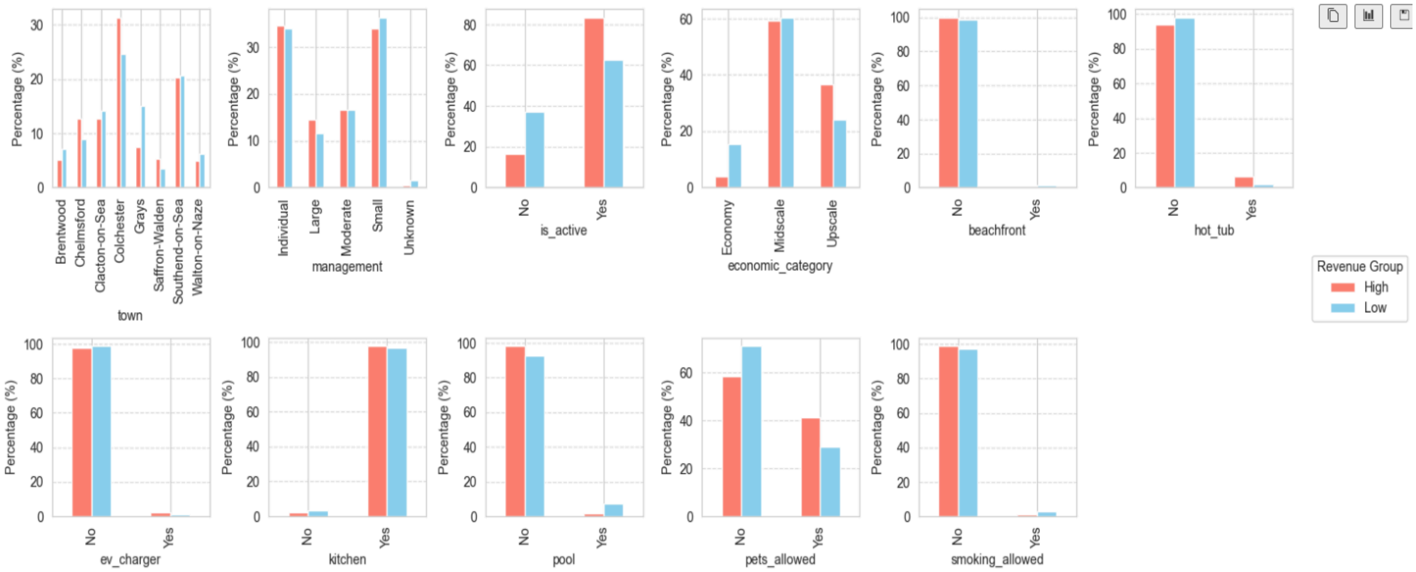


Fig 7. Bar chart feature analysis for annual revenue

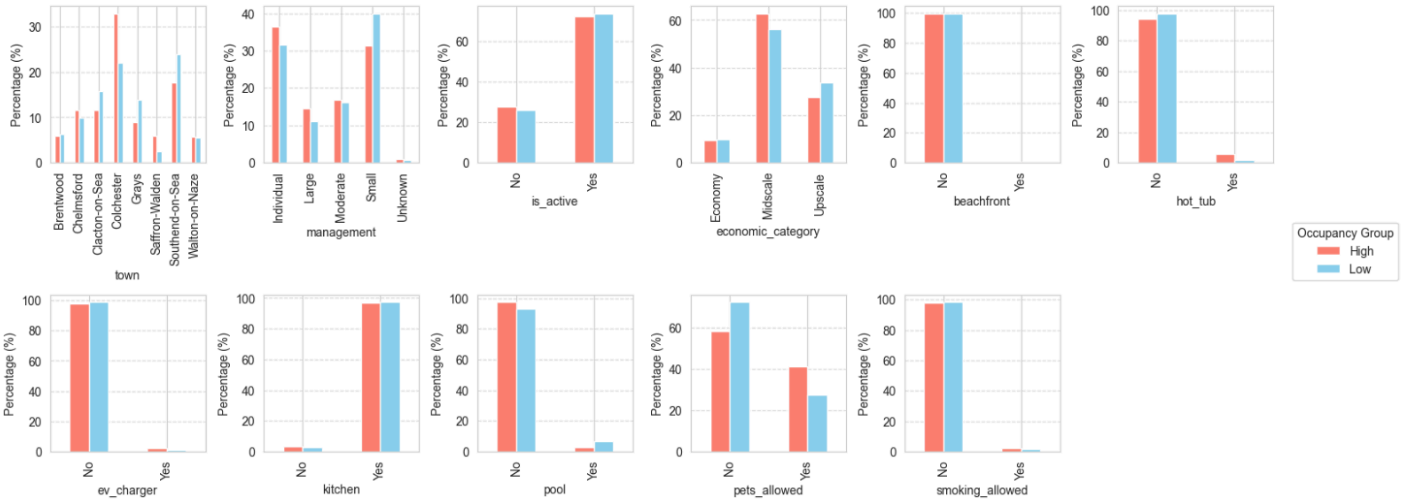


Fig 8. Bar chart feature analysis for Occupancy

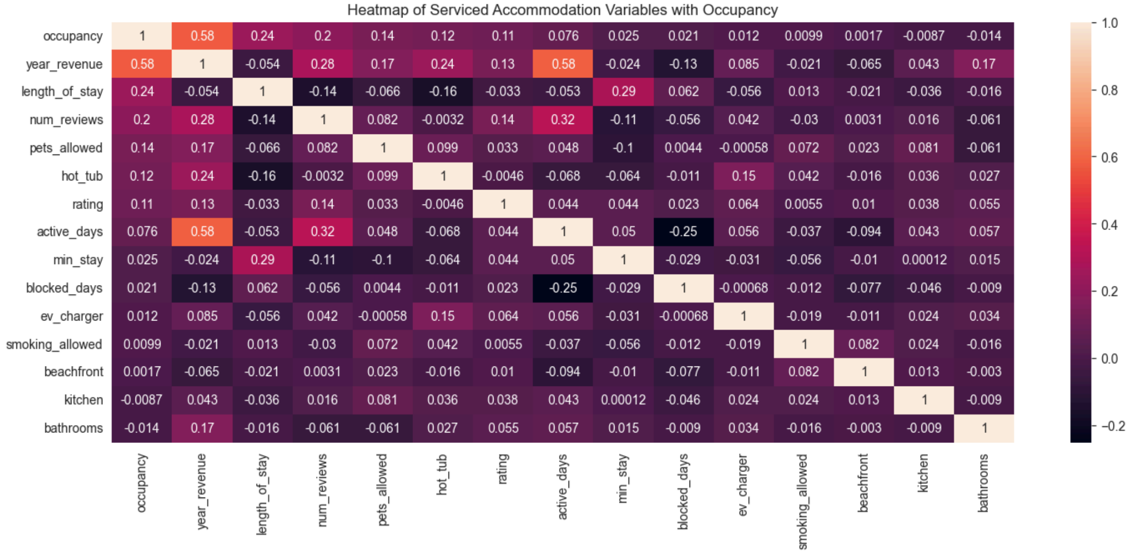
Data Pre-Processing

Data pre-processing was also conducted to transform categorical/binary features into numerical values using Scikit Learn module ‘preprocessing, LabelEncoder.

Column values transformed were ‘town’, ‘management’, ‘is_active’, ‘economic category’, ‘beachfront’, ‘hot_tub’, ‘ev_charger’, ‘kitchen’, ‘pool’, ‘pets_allowed’ and ‘smoking_allowed.

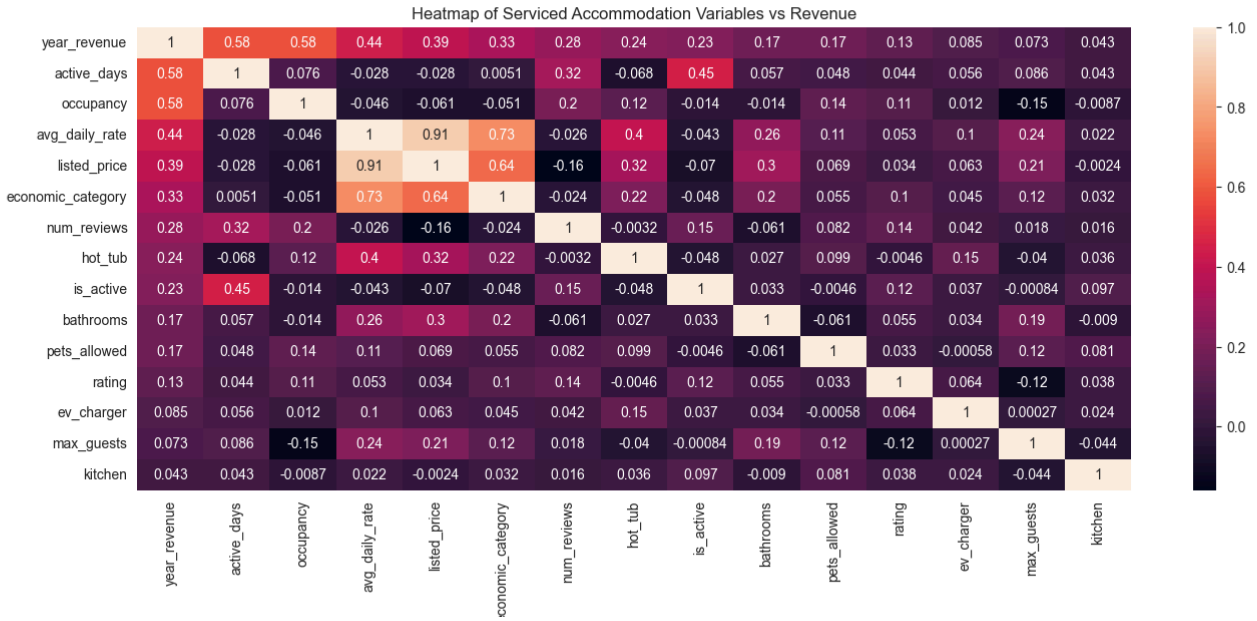
Correlational analysis

Aside from the bar charts and boxplots mentioned the section before, correlational analysis was conducted using heatmaps to determine which features show the strongest relationship between the target variables. For increased readability, this was limited to the top 15 features.



	occupancy	year_revenue	length_of_stay	num_reviews	pets_allowed	hot_tub	rating	active_days	min_stay	blocked_days	ev_charger	smoking_allowed	beachfront	kitchen	bathrooms
occupancy	1	0.58	0.24	0.2	0.14	0.12	0.11	0.076	0.025	0.021	0.012	0.0099	0.0017	-0.0087	-0.014
year_revenue	0.58	1	-0.054	0.28	0.17	0.24	0.13	0.58	-0.024	-0.13	0.085	-0.021	-0.065	0.043	0.17
length_of_stay	0.24	-0.054	1	-0.14	-0.066	-0.16	-0.033	-0.053	0.29	0.062	-0.056	0.013	-0.021	-0.036	-0.016
num_reviews	0.2	0.28	-0.14	1	0.082	-0.0032	0.14	0.32	-0.11	-0.056	0.042	-0.03	0.0031	0.016	-0.061
pets_allowed	0.14	0.17	-0.066	0.082	1	0.099	0.033	0.048	-0.1	0.0044	-0.00058	0.072	0.023	0.081	-0.061
hot_tub	0.12	0.24	-0.16	-0.0032	0.099	1	-0.0046	-0.068	-0.064	-0.011	0.15	0.042	-0.016	0.036	0.027
rating	0.11	0.13	-0.033	0.14	0.033	-0.0046	1	0.044	0.044	0.023	0.064	0.0055	0.01	0.038	0.055
active_days	0.076	0.58	-0.053	0.32	0.048	-0.068	0.044	1	0.05	-0.25	0.056	-0.037	-0.094	0.043	0.057
min_stay	0.025	-0.024	0.29	-0.11	-0.1	-0.064	0.044	0.05	1	-0.029	-0.031	-0.056	-0.01	0.00012	0.015
blocked_days	0.021	-0.13	0.062	-0.056	0.0044	-0.011	0.023	-0.25	-0.029	1	-0.00068	-0.012	-0.077	-0.046	-0.009
ev_charger	0.012	0.085	-0.056	0.042	-0.00058	0.15	0.064	0.056	-0.031	-0.00068	1	-0.019	-0.011	0.024	0.034
smoking_allowed	0.0099	-0.021	0.013	-0.03	0.072	0.042	0.0055	-0.037	-0.056	-0.012	-0.019	1	0.082	0.024	-0.016
beachfront	0.0017	-0.065	-0.021	0.0031	0.023	-0.016	0.01	-0.094	-0.01	-0.077	-0.011	0.082	1	0.013	-0.003
kitchen	-0.0087	0.043	-0.036	0.016	0.081	0.036	0.038	0.043	0.00012	-0.046	0.024	0.024	0.013	1	-0.009
bathrooms	-0.014	0.17	-0.016	-0.061	-0.061	0.027	0.055	0.057	0.015	-0.009	0.034	-0.016	-0.003	-0.009	1

Fig 9. Heatmap showing correlationship between features and Occupancy



	year_revenue	active_days	occupancy	avg_daily_rate	listed_price	conomic_category	num_reviews	hot_tub	is_active	bathrooms	pets_allowed	rating	ev_charger	max_guests	kitchen
year_revenue	1	0.58	0.58	0.44	0.39	0.33	0.28	0.24	0.23	0.17	0.17	0.13	0.085	0.073	0.043
active_days	0.58	1	0.076	-0.028	-0.028	0.0051	0.32	-0.068	0.45	0.057	0.048	0.044	0.056	0.086	0.043
occupancy	0.58	0.076	1	-0.046	-0.061	-0.051	0.2	0.12	-0.014	-0.014	0.14	0.11	0.012	-0.15	-0.0087
avg_daily_rate	0.44	-0.028	-0.046	1	0.91	0.73	-0.026	0.4	-0.043	0.26	0.11	0.053	0.1	0.24	0.022
listed_price	0.39	-0.028	-0.061	0.91	1	0.64	-0.16	0.32	-0.07	0.3	0.069	0.034	0.063	0.21	-0.0024
economic_category	0.33	0.0051	-0.051	0.73	0.64	1	-0.024	0.22	-0.048	0.2	0.055	0.1	0.045	0.12	0.032
num_reviews	0.28	0.32	0.2	-0.026	-0.16	-0.024	1	-0.0032	0.15	-0.061	0.082	0.14	0.042	0.018	0.016
hot_tub	0.24	-0.068	0.12	0.4	0.32	0.22	-0.0032	1	-0.048	0.027	0.099	-0.0046	0.15	-0.04	0.036
is_active	0.23	0.45	-0.014	-0.043	-0.07	-0.048	0.15	-0.048	1	0.033	-0.0046	0.12	0.037	-0.00084	0.097
bathrooms	0.17	0.057	-0.014	0.26	0.3	0.2	-0.061	0.027	0.033	1	-0.061	0.055	0.034	0.19	-0.009
pets_allowed	0.17	0.048	0.14	0.11	0.069	0.055	0.082	0.099	-0.0046	-0.061	1	0.033	-0.00058	0.12	0.081
rating	0.13	0.044	0.11	0.053	0.034	0.1	0.14	-0.0046	0.12	0.055	0.033	1	0.064	-0.12	0.038
ev_charger	0.085	0.056	0.012	0.1	0.063	0.045	0.042	0.15	0.037	0.034	-0.00058	0.064	1	0.00027	0.024
max_guests	0.073	0.086	-0.15	0.24	0.21	0.12	0.018	-0.04	-0.00084	0.19	0.12	-0.12	0.00027	1	-0.044
kitchen	0.043	0.043	-0.0087	0.022	-0.0024	0.032	0.016	0.036	0.097	-0.009	0.081	0.038	0.024	-0.044	1

Fig 10. Heatmap showing relationship between features and Annual Revenue

Phase 1 Summary of Findings and Conclusions

Whilst calculating correlation coefficients have not shown strong correlations between property features, occupancy rates and annual revenue, some patterns have emerged.

Property Features

Chelmsford and Colchester appear to be the most likely areas of higher occupancy rates and higher annual return.

There is no evidence to show that more bedrooms, beachfront properties or availability of a pool or EV charger on the properties will yield better earnings.

There is some weak evidence that the availability of a hot-tub, properties considered within a higher economic bracket, more bathrooms and are pet-friendly may be slightly more profitable.

Listing and Host Features

Listed price and average daily rate of a property has the strongest relationship with revenue but does not impact occupancy rates. For both target variables, the listing has had a positive relationship with occupancy and revenue, suggesting that customers are more likely to book a property if that property has had a record of reviews. However, rating has a weak correlation between both groups.

PHASE 2 - Machine Learning

In Phase 2 of this project, we focus on the second core objective: to train, tune, and evaluate predictive models that will help the client identify high-performing properties for their serviced accommodation business. The aim is to provide a data-driven approach to support property selection and investment decisions.

Phase 2 Methodology

Data Preprocessing

Before training, the cleaned dataset from phase 1 will be prepared and refined to ensure model readiness. This includes:

- Defining the target variable and selecting relevant features
- Removing multicollinear or redundant features
- Encoding categorical variables using appropriate techniques (e.g., one-hot or label encoding)
- Scaling numerical features where required, especially for models sensitive to magnitude (e.g., SVM, Neural Networks)

The dataset is split as follows:

- **80% Training set** - for fitting models and tuning hyperparameters
- **10% Validation set** - for interim performance evaluation and model selection
- **10% Test set** - for final performance assessment on unseen data

Models Selected for Training

A range of classification models were selected *(See Appendix 1 for further information)*:

- **Logistic Regression**- as a baseline model
- **Random Forest Classifier** - for non-linear pattern detection and feature importance
- **Support Vector Machine (SVM)** - for high-dimensional classification boundaries
- **K-Nearest Neighbor** - a non-parametric model that classifies based on local neighbourhoods. Effective for small datasets and intuitive to interpret.
- **Neural Network (Keras Sequential)** – for capturing complex, non-linear interactions

Model Optimisation & Tuning

- *RandomizedSearchCV* will be used for hyperparameter tuning across models, allowing for efficient exploration of large parameter spaces
- The best-performing model will be validated using K-Fold Cross-Validation to assess generalizability and stability across data subsets

Model Evaluation Metrics

Models will be evaluated and compared based on the following performance metrics:

- *Accuracy* - overall correctness of predictions
- *Precision*- how many predicted positives are actually positive
- *Recall* - how many actual positives were correctly identified
- *F1-score*- harmonic mean of precision and recall
- *ROC AUC* - ability to discriminate between classes
- Confusion Matrix - to visualise true/false predictions per class

Selection of Target Variable and Features

Target Variable Selection

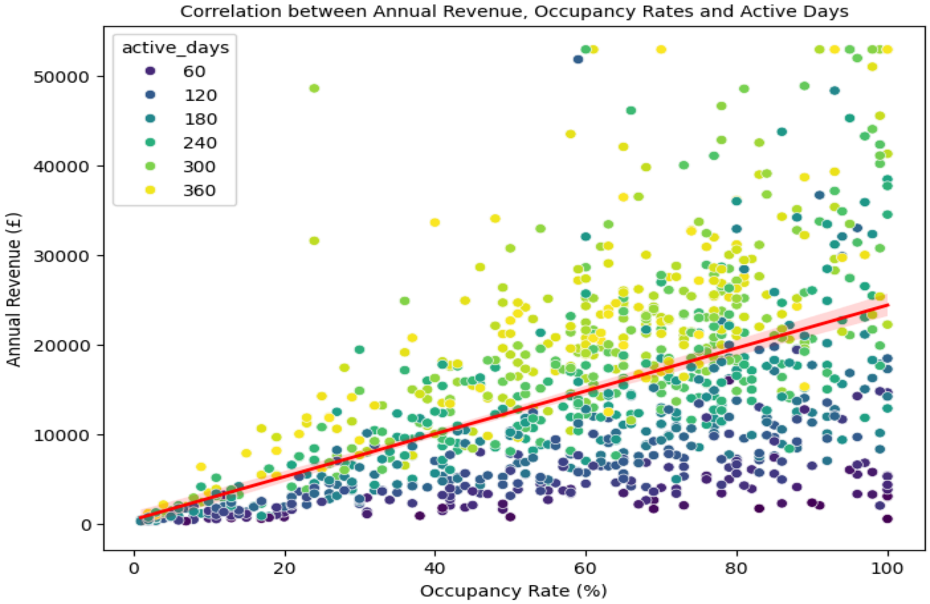


Fig.11. Revenue vs Occupancy vs Active Days

As can be seen in this scatterplot, the higher the active days (i.e the lighter colours), the higher the annual revenue. To simplify model selection and reduce the effect of multicollinearity, it is decided that the target focus will be on **Occupancy** as the target variable as revenue outcomes are directly influenced by occupancy rate and it is less affected by the number of active days the property is listed (corr = 0.07).

Target Variable:
`occupancy_group` (binary classification >= 63% High, <63% Low)

Feature Selection

Following the initial exploratory analysis in phase 1 and target variable assessment, several features have been excluded from the predictive modelling process for the reasons outlined below:

1. `active_days`, along with closely associated variables, has been removed:
    - `is_active`: Indicates whether the property was active at the time of data collection.
    - `blocked_days`: Reflects the number of days a property was made unavailable by the host.

These features primarily reflect listing availability rather than intrinsic property characteristics. Availability-based variables fall outside the scope of influence and interest. Their removal does not impact the integrity of the analysis.

2. Exclusion of `length_of_stay`

This variable represents guest behavior and is inherently **outcome-driven**, closely related to the target variable `occupancy`. Including it would introduce **data leakage**, as it reflects post-booking behavior rather than a property trait.

3. Exclusion of `management`

The type of management (e.g., agency-managed vs. self-managed) is a static business decision that is unlikely to be influenced by property characteristics or easily changeable. As such, it is excluded to ensure the focus remains on actionable or investable features.

4. Exclusion of `beachfront`

There were only 6 "yes" cases for this feature, so cannot provide a statistically significant contribution to modelling and may increase noise.

Balance of Classes

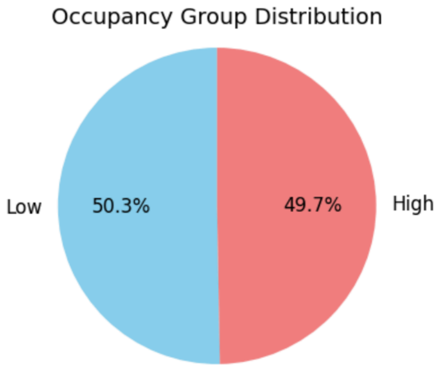


Fig.12. Proportion of classes in Occupancy Group

As the median was used to determine thresholds of occupancy rate, the target variable is balanced.

Model Training Outcomes

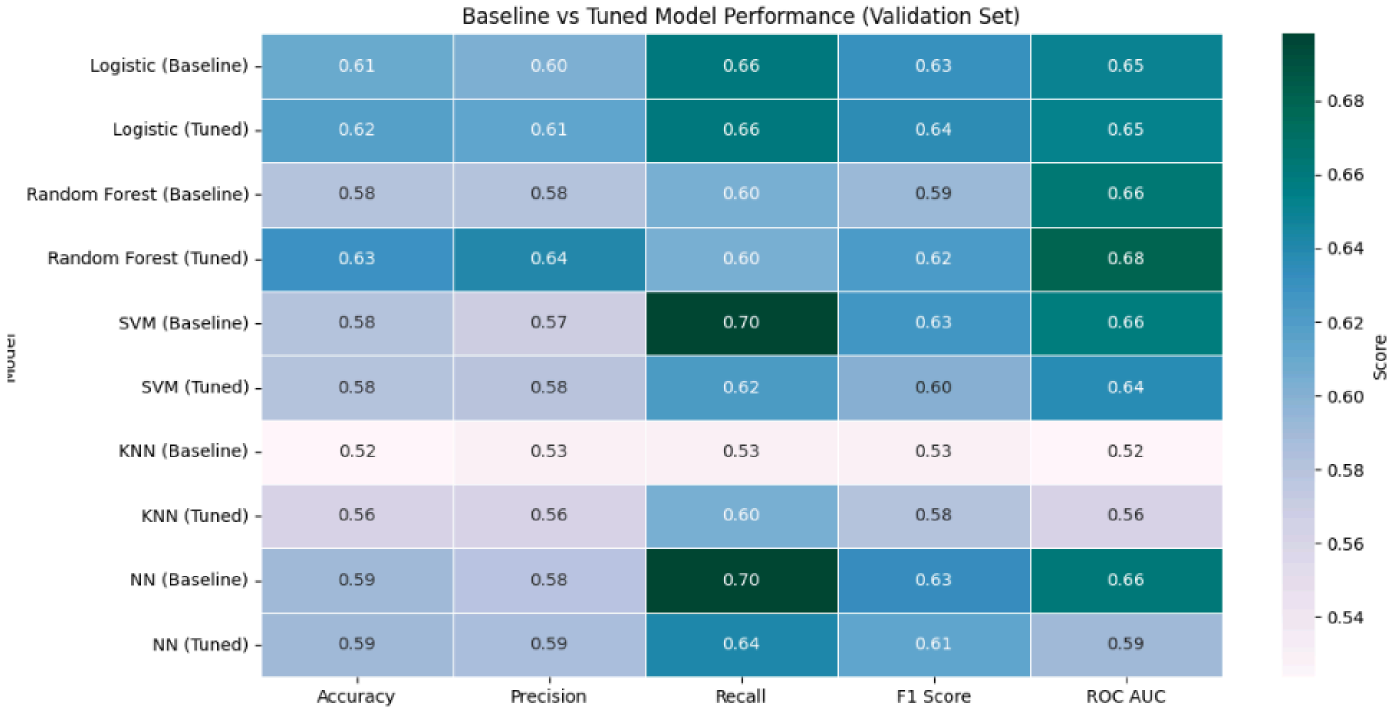


Fig.13. Heatmap of Accuracy, Precision, Recall, F1 Score and ROC AUC for

The heatmap compares baseline and tuned versions of five classification algorithms — Logistic Regression, Random Forest, SVM, KNN, and Neural Networks — across key performance metrics: **Accuracy**, **Precision**, **Recall**, **F1 Score**, and **ROC AUC**. Findings are:

- **Best overall performance**: Random Forest (Tuned) achieved the top scores in **accuracy**, **precision**, and **ROC AUC**, making it the most balanced and effective model across all metrics.
- **Robust generalisation**: The improvements from baseline to tuned version suggest the model generalises well on unseen data.
- **Flexibility**: Handles mixed data types (numeric and categorical) and is robust to outliers and multicollinearity.
- **Business use**: Prioritises **correct identification of high-occupancy properties**, which is crucial for informing **investment decisions**, **revenue projections**, and **listing strategy**.

CONFUSION MATRIX

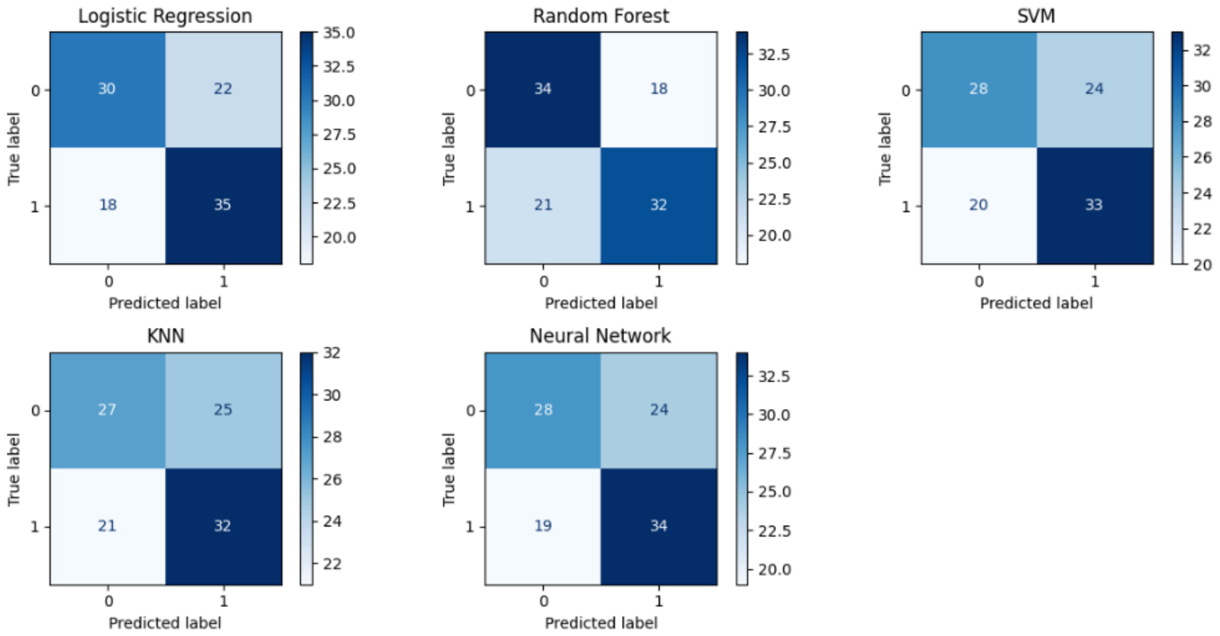


Fig 14. Confusion Matrices of the different tuned trained models

- **Random Forest**: Achieved the best balance with the **fewest false highs (18)** while maintaining strong **true highs (32)**. This means it reliably detects both high- and low-occupancy listings with fewer risky misclassifications.

- **Logistic Regression**: Had the **highest number of true highs (35)**, indicating strong detection of valuable listings. However, it also suffered from **relatively high false highs (22)**, potentially suggesting properties are profitable when they are not, which can have serious financial implications.

- **Neural Network**: Closely matches Logistic Regression, with a slightly more balanced trade-off between false and true highs/lows.

- **SVM and KNN**: Both models had the **highest number of false highs and false lows**, reflecting **less reliable predictions** and a riskier output pattern for business application.

Precision- Recall and ROC Curves

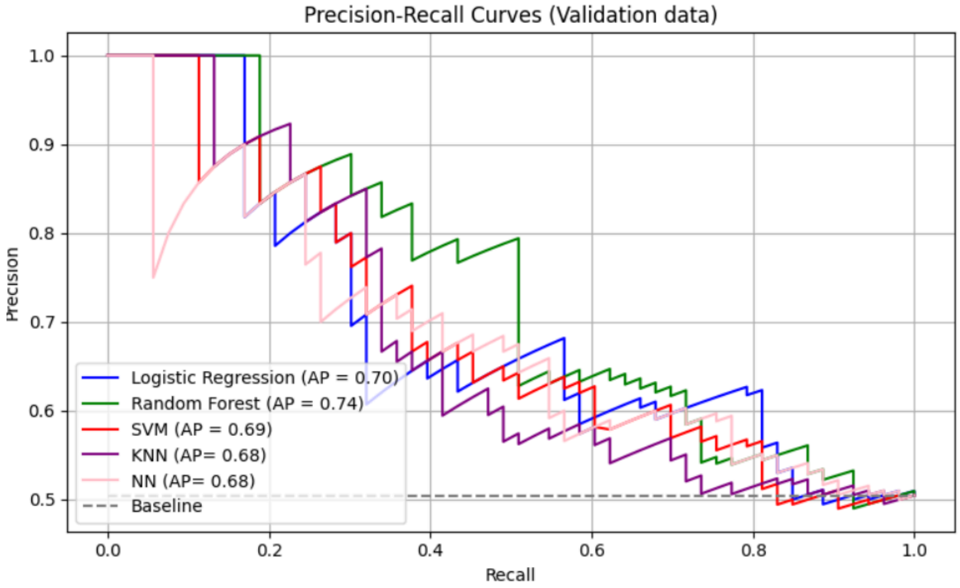


Fig. 15 Precision- Recall Curve with Average Precision

The Precision-Recall (PR) curve visualizes the trade-off between precision and recall for different classification thresholds. It is particularly useful in imbalanced classification tasks, such as occupancy classification, where the cost of misclassification is high.

- **Random Forest** demonstrates the **highest average precision (0.74)**, outperforming all other models across the full range of recall values. This indicates that it consistently maintains a strong ability to identify **true high occupancy listings**, while minimizing false positives.
- **Logistic Regression** also performs well with an AP of 0.70, suggesting a reliable ability to balance precision and recall. However, it slightly trails Random Forest in higher recall ranges.
- **SVM, KNN, and Neural Network** exhibit lower AP scores (0.68–0.69), with performance dropping faster as recall increases. This indicates that these models may struggle to maintain precision when aiming to capture more high-occupancy listings.
- The **baseline (0.5)** represents random guessing — all models significantly outperform this threshold, validating the robustness of the predictive pipeline.

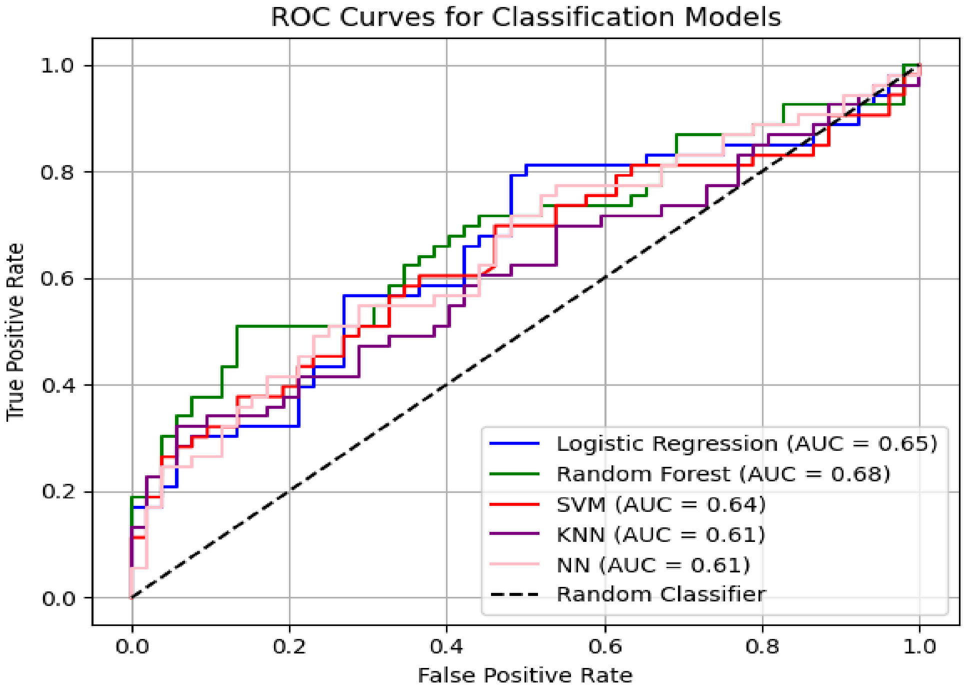


Fig. 16. ROC Curve Comparison

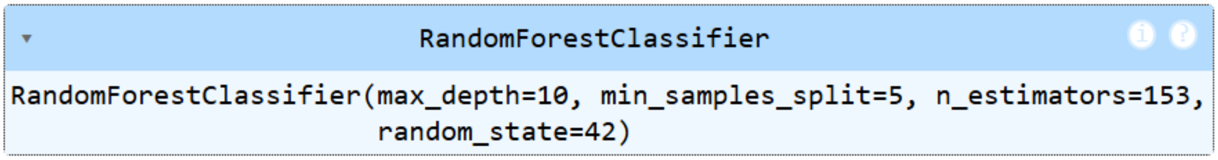
Overall the Random Forest performed the best overall with the **highest AUC of 0.68.**

KNN underperformed in comparison with the curve closest to the baseline, indicating poorer discrimination. Both KNN and Neural Networks had the lowest AUC score at 0.61.

Random Forest model (tuned) gives the highest confidence in correctly separating high vs. low occupancy properties.

Model Selection and Testing

Based on the above analyses, **Random Forest (tuned)** was deemed the best fit model, with the following parameters:



```
RandomForestClassifier

RandomForestClassifier(max_depth=10, min_samples_split=5, n_estimators=153,
                       random_state=42)
```

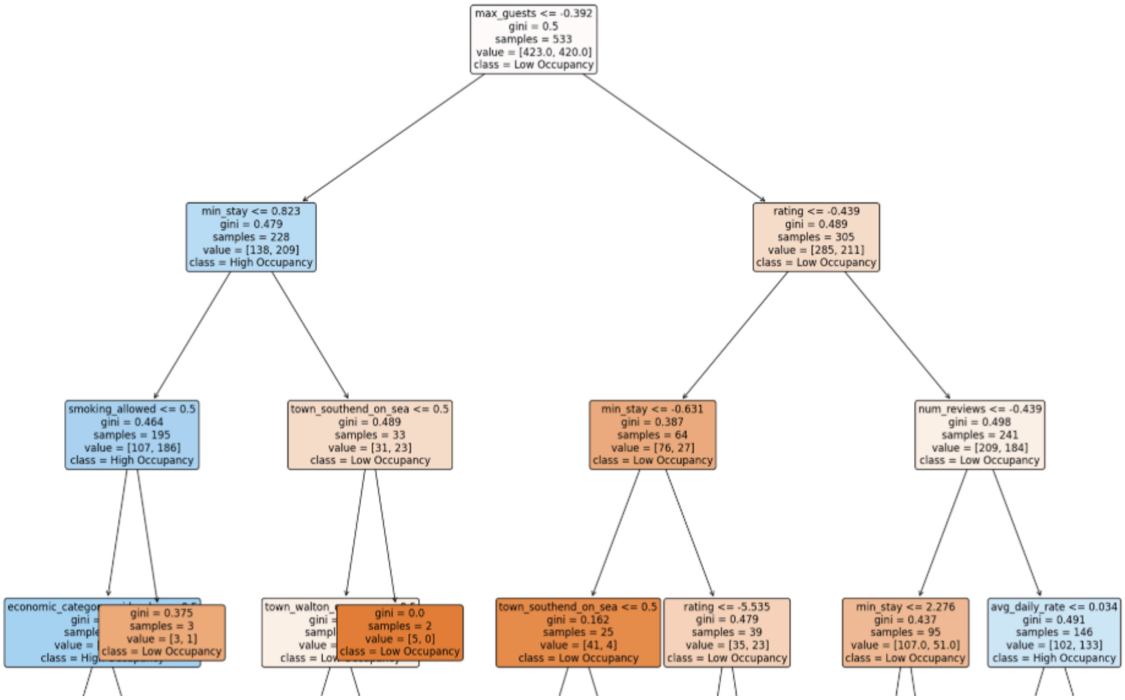
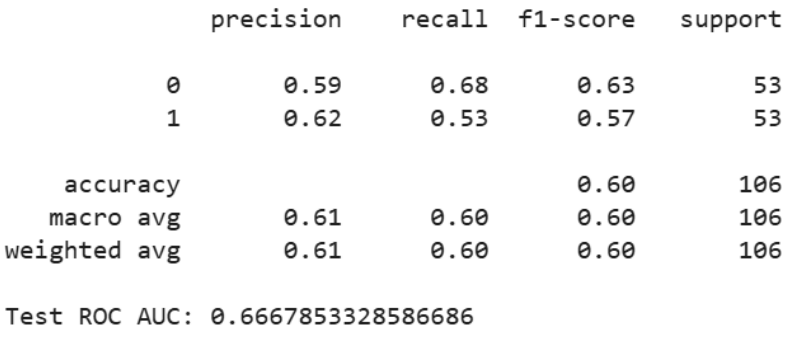


Fig. 17. Decision Tree using Random Forest

The model was trained and tested on the (unseen) test set with the following outcomes:



```
              precision    recall  f1-score   support

           0       0.59      0.68      0.63        53
           1       0.62      0.53      0.57        53

    accuracy                           0.60       106
   macro avg       0.61      0.60      0.60       106
weighted avg       0.61      0.60      0.60       106

Test ROC AUC: 0.6667853328586686
```

Interpretation:

- The model performs slightly better in identifying **low-occupancy properties** (class 0) with higher recall.

- Precision for high-occupancy listings (class 1) is slightly better than recall, indicating some **false positives**—properties predicted as high occupancy that are actually low.

- The **balanced F1-scores** for both classes (0.63 and 0.57) suggest a modest ability to handle class imbalance and generalise across classes.

- An **ROC AUC of 0.667** indicates that the model has moderate discriminative power—better than random guessing (0.5), but not highly robust.

Feature Importance

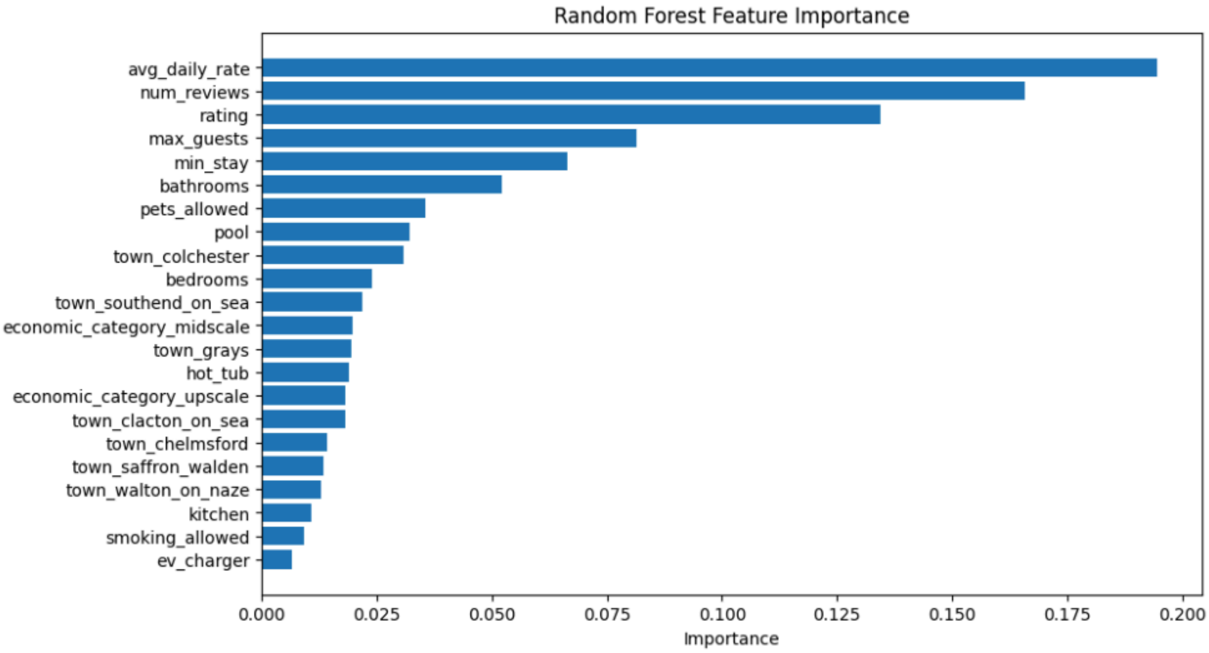


Fig. 17. Feature importance using Random Forest

This chart shows the relative importance of each feature in predicting occupancy levels (e.g., High vs Low) using a Random Forest model.

Top 5 Most Influential Features

1. `avg_daily_rate`: Strongly affects occupancy. Pricing strategy plays a key role.
2. `num_reviews`: A higher number of reviews often signals popularity and trust.
3. `rating`: Better-rated listings are more likely to be booked.
4. `max_guests`: Accommodations for more guests appeal to larger groups.
5. `min_stay`: Minimum stay restrictions influence booking flexibility.

CONCLUSIONS

Summary of Key Findings

- Random Forest (Tuned) emerged as the most effective model
- The model showed moderate discriminatory power, performing better than random guessing, and exhibited consistent results across cross-validation folds.
- Confusion matrix analysis revealed that the model is conservative in predicting high occupancy (fewer false positives), which is beneficial in a business context to avoid overestimating property potential.
- The Precision-Recall curve further confirmed Random Forest's strength in identifying high-occupancy properties while minimizing misclassification risk.

Business Recommendations

1. Use the predictive model as a decision-support tool, not a sole authority. Its predictions should complement expert judgment, especially when making high-stakes investment decisions.
2. Focus on high-confidence predictions: Properties flagged with high probability of occupancy should be prioritised for further manual due diligence and site analysis.
3. Monitor model performance regularly as new data becomes available, ensuring that it continues to reflect market dynamics and business needs.

Future Analysis & Modelling Recommendations

1. Expand Dataset and Enhance Feature set:
   - Add location-specific data (e.g., proximity to transport, attractions)
   - Incorporate seasonality or temporal booking trends
   - Expand dataset (increase sampling) by incorporate data from additional time periods and other geographic regions beyond Essex for broader generalisability and more robust predictions.

2. Explore Advanced Modelling:
    - Gradient boosting (e.g., XGBoost)
    - Deep learning with engineered features
3. Deploy Model in a Live System:
    - Integrate into a dashboard or decision-support app using FastAPI, Streamlit, or a BI tool.

Appendix 1: Model Selection and Rationale

In this study, the aim is to build a classification model capable of predicting occupancy performance (e.g., occupancy_group as "high" or "low") based on a diverse set of property and listing features. To address this problem, a selection of classification algorithms was chosen based on their theoretical suitability, empirical success in similar tasks, and ability to capture different types of relationships in the data. Each model offers unique strengths in terms of interpretability, complexity handling, and performance. The following models were selected:

1. Logistic Regression

Logistic Regression was chosen as a baseline model due to its simplicity, interpretability, and widespread use in binary classification tasks. It assumes a linear relationship between the independent variables and the log-odds of the dependent binary variable.

Although it may not capture complex patterns or interactions, it serves as a strong reference point to compare more advanced models. Furthermore, its coefficients offer direct insights into feature importance, making it valuable for initial analysis and explainability.

2. Random Forest

Random Forest is a robust ensemble method based on decision trees. It was selected for its ability to model non-linear relationships and interactions between features without requiring feature scaling or extensive preprocessing.

Random Forest is particularly effective in datasets with a mix of numerical and categorical variables, as is the case in this project. Additionally, its built-in feature importance scores provide a useful way to evaluate which variables contribute most to model performance. It is less prone to overfitting than individual decision trees due to its use of bagging and random feature selection.

3. Support Vector Machines (SVM)

Support Vector Machines were chosen due to their strength in high-dimensional spaces and ability to define complex decision boundaries using kernel functions.

In particular, the use of a radial basis function (RBF) kernel can capture non-linear relationships between features and the target class.

SVMs are known for their generalisation ability and robustness, especially in cases where the data is not linearly separable. While SVMs can be computationally intensive on large datasets, the size of the dataset used in this study is moderate (n=1054), making them a feasible and promising option.

4. K-Nearest Neighbours (KNN)

K-Nearest Neighbours was included as a non-parametric model that makes minimal assumptions about the data distribution. It classifies a data point based on the majority class of its nearest neighbours in the feature space.

KNN is intuitive and can perform well in cases where class boundaries are complex and local patterns dominate. However, its performance can be sensitive to the choice of k and the scaling of features, which will be addressed through hyperparameter tuning. Despite its simplicity, KNN provides a useful benchmark and an alternative perspective on the classification task.

5. Neural Networks

Neural Networks (NN) were selected to capture non-linear interactions and complex feature combinations that may not be easily learned by other models. Neural networks are capable of approximating any continuous function given sufficient capacity and training data. In this context, they are well-suited for uncovering subtle relationships within the data, especially with a mix of numeric and binary categorical inputs. While neural networks require more tuning and computational resources, they offer high flexibility and have shown state-of-the-art performance in many classification problems.