Factors Influencing Rental Investments in Paphos, Cyprus: Comparing Short and Long-Term Rental Strategies

Sam Martin , Thomas Dimopoulos * , Martha Katafygiotou

Posted Date: 9 April 2024

doi: 10.20944/preprints202404.0187.v2

Keywords: AirDNA; Airbnb; Random Forest; K-Nearest Neighbour; Multiple Linear Regression; Geographic Information System

 doi:10.20944/preprints202404.0187.v2

Article

Factors Influencing Rental Investments in Paphos, Cyprus: Comparing Short and Long-Term Rental Strategies

Sam Martin, Thomas Dimopoulos * and Martha Katafygiotou

Department of Real Estate, Neapolis University Paphos, 2 Danais Avenue, Paphos 8042, Cyprus; SamGMartin@outlook.com; m.katafygiotou@nup.ac.cy

* Correspondence: t.dimopoulos@nup.ac.cy

Abstract: Understanding the optimal strategy for a real estate investment and how performance changes based on characteristics is crucial for optimising the achievable return. This is prominent in tourist areas such as Paphos where there is no clear distinction as to whether short or long-term approaches are optimal. The study aimed to develop a model for predicting the optimal rental strategy whilst assessing which model performed best and which property attributes impacted its return the greatest. Short-term data was collected from AirDNA and long-term data was manually collected from real estate agents websites. Furthermore, Random Forest, K-Nearest Neighbour and Multiple Linear Regression models were created to predict the highest and best use for each property. Model accuracy varied between data sets with the best performing model for short-term properties being Random Forest (R-Squared: 0.843), and Distance-Based Multiple Linear Regression for long-term approach (R-Squared: 0.843). The study demonstrated that accurate models could be created to predict the optimal rental strategy with the number of bedrooms being the main driver for rental income, followed by luxury finishes and the presence of a pool. It found that locational characteristics didn't impact the returns significantly assuming that the property was located within a tourist area.

Keywords: AirDNA; AIRBNB; random forest; k-nearest neighbour; multiple linear regression; geographic information system

1. Introduction

Thanks to the natural beauty, international attractions, and Mediterranean climate, Cyprus sees a substantial amount of tourism with the peak seasons of January 2022 to August 2022, seeing visitor spending on items such as travel, accommodation, and expenses, reach €1,617 million and €2,439m annually [1]. Although significant, 2023 trends from January – May show an increase compared to the previous year of 34.2% which reaffirms Cyprus' growing global appeal [2].

The Republic of Cyprus is home to four major cities, Nicosia, Larnaca, Limassol and Paphos, and unlike larger European countries, each has its distinct characteristics. Whilst cities such as Nicosia and Limassol boast the highest populations and commercial activity, Paphos is undoubtedly the home of tourism. It is internationally recognised as such by winning numerous EU awards, the latest of which is the "2023 European Capital of Smart Tourism" [3].

Paphos' population fluctuates considerably throughout the year due to tourism, and whilst there are around 95,400 permanent residents, Paphos Airport reported over 600,000 arrivals within the first four months of 2022 [4]. As such, a complex relationship between the long and short-term rental market has been created which sees many residents being outpriced from central and tourist areas. This issue is compounded by many investors optimising for a short-term strategy which means there is insufficient housing for residents. When considering the demand curve for long-term properties, this reduction in available stock has caused a rise in rent which at some point will surpass the short-term rental returns.

 doi:10.20944/preprints202404.0187.v2

For investors, the balance between short and long-term demand is crucial as the rental strategy of their property is directly impacted by the availability of each. Although this relationship occurs across the globe, cities such as Paphos with high levels of tourism are of most interest since the demand is high relative to the overall size.

Currently, there are no publicly available studies regarding the Paphos rental market. This is concerning since as Cyprus develops and expands its global reach, there is a requirement for clear, reliable, and informative studies to better advise investors. As such, this study aims to satisfy this by informing investors of the characteristics and trends present within the Paphos market which may differ from other European areas.

Moreover, investors need to understand which rental strategy yields the highest return and which factors impact an investment's return the most. This study, therefore, identifies these factors, whilst estimating which approach, short or long-term rental, yields the highest return given a property's specific characteristics.

To achieve this, the study reviews a range of common statistical models, compares their results, and highlights any limitations they have in the Paphos market. In practice, these models act as "estimators" allowing investors to input their property's characteristics which in turn predicts the rental income for the property for each rental strategy.

Furthermore, the study will identify the model which most accurately predicts the return of rental properties when considering long and short-term rent and suggest the highest and best use (HBU) for each. The various models must be reviewed based on their relative appropriateness to Paphos as this directly impacts upon its predictive ability and as such, the accuracy of the information regarding HBU. In addition, the factors which impact each rental strategy's performance will be identified so that advice can be provided to investors on how they can best manage their property to improve their return.

From this study, investors can expect to hold the tools needed to make informed decisions on their investment portfolio and apply them accordingly to optimise their rental portfolio.

2. Literature Review

2.1. Similar Studies

A study by Shokoohyar, Sobhani, & Sobhani into the optimal rental strategy based on the rate of return (RoR) between short-term and traditional residential rental investments was conducted for the city of Philadelphia [5]. It considered a range of different factors including those relating to the property, neighbourhood and location. A range of models were examined with, K-nearest neighbour (KNN), Random Forest (RF) and Multiple Linear Regression (MLR) analysis most accurately predicting the RoR. Moreover, they were able to define specific areas for which each rental strategy would be most suitable with properties located centrally, close to historical attractions and nightlife, typically being more suited to short-term rental. Conversely, the suburbs were more suited to long-term residential rental. Whilst the study was relatively comprehensive, it only considered the area of Philadelphia which is considerably different to that of Paphos.

Another study, this time on the city of Bologna, compared the performance of both short and long-term residential properties finding that 49% of short-term properties had an economic advantage over the long-term equivalent [6]. In the study, they too used MLR for the short-term rental data however, for the long-term data the city was split into 28 areas in which the average return was taken for each using the national "Real Estate Market Observatory" due to limited public data. The comparisons made between the two were relatively fundamental and little interrogation was done regarding the factors affecting the return, nevertheless, it further supports the use of MLR and offers an alternative should data not be available for the Paphos market.

The two studies focused on tourist cities, however, neither of which were in a coastal area as is the case with Paphos. For this reason, the study by Rodríguez-Pérez de Arenaza, Ángel Hierro, & Patiño was reviewed which focused on the Andalusia area in Spain [7]. This study investigated how the rental prices of residential properties were impacted by the short-term holiday rental industry.

 doi:10.20944/preprints202404.0187.v2

Again, due to limited data on the market, they were unable to use a dynamic regression model and instead opted for a cross-sectional model. Due to being a fixed point in time, they repeated it several times to account for touristic changes and found that short-term rental sites such as Airbnb increased the average residential rental price by 13.69% creating an interesting dynamic between the two. The main limitation of the study relates to the approach taken to negate the absence of time series data which prevented the application of a more robust econometric analysis.

2.2. Data Collection

Two main collection methods which were used to collect the data are AirDNA and manual gathering. AirDNA is an analytics platform focusing on short-term rental data which it collates from a range of rental platforms such as; Airbnb, Vrbo and Booking.com [8]. This platform provides most of the data including the geographical information and proximity to POI. For long-term properties, the data was collected through manual gathering from real estate agents' websites and rental platforms including Bazaraki and Facebook Marketplace. One concern was the accuracy and availability of this data which could lead to data cleansing being required.

Various studies have investigated the effect of sample size on the accuracy of analysis, and depending on the quality of the data, the recommended sample ranges vary. Nguyen & Cripps found that MLR required around 500 data points to accurately create a model for the relationships whereas Benjamin, Guttery, & Sirmans found that 70 could yield good results [9,10]. Further to this, Limsombunchai, Gan, & Lee were able to achieve an R-Squared value of 0.75 with only 200 data points which demonstrates that even smaller sample sizes can be used to effectively create MLR models [11]. Interestingly, a study carried out that investigated developing markets with limited data found that MLR models outperformed both k-NN and RF regarding the mean absolute percentage errors (MAPE) [12]. The study used 318 data points from Szczecin in Poland, which is far less developed than Paphos and does not attract the same level of tourism. Based on these findings, a dataset of 200 – 300 properties was targeted for the long-term data and around 500+ for the short-term due to the methods of data collection.

2.3. Enriching Data

A study on the Cyprus market aimed to understand whether the use of Artificial Intelligence (AI) and Machine Learning (ML) could supplement the use of Mass Appraisals (MA) to achieve more accurate predictions [13]. It concluded that there were significant errors within the model used by the Department of Land and Surveys (DLS) and enriching it would help the reliability and accuracy of MA. This could be done using satellite imagery but also geographical locations relative to key value-influencing areas such as hospitals and schools. As such, locational characteristics can significantly impact the return for short and long-term rentals so for this reason, a combination approach incorporating satellite imagery may be beneficial for improving the results of the study.

For this project, the use of GIS spatial data, coordinates, and features, was invaluable as it helped enrich the analysis allowing for variables outside of just the property characteristics to be analysed. Din, Hoesli, & Bender state that even when GIS data is incomplete, it should be used where available to improve the accuracy of hedonic pricing models [14].

 doi:10.20944/preprints202404.0187.v2

3. Materials and Methods

3.1. Short-Term Rental Data

As previously mentioned, the short-term rental data was collected through AirDNA with consent from AirDNA and Axia Valuers to whom the data belongs. To ensure the data was clean and specific to the research, a thorough data cleansing was carried out to remove outliers and properties which were not within the scope of the project. From the initial 9,584 properties, only 825 satisfied all the qualities which were relevant to the study, this cleansing has been discussed as follows.

3.1.1. Locations Outside of the Paphos District

By using the longitudinal and latitudinal coordinates as the bounding box of Paphos, all the properties located outside of this area were removed with the limitation of assuming the Paphos district as a rectangle. Most properties were not located near the district borders, so the impact was deemed of low significance.

3.1.2. Duration of Operation

The research focused on one fixed point in time, so only 2022 data was considered, and the property must have been operating for the entirety to not negatively skew the equivalent monthly rate.

3.1.3. Outliers and Ghost Properties

Through analysis of the data, it was found that many properties were not active, had unrealistic prices or were so-called "Ghost" properties. These were removed as they caused significant inaccuracies as supported by Fleischer, Ert and Bar-Nahum [15]. This was done through outlier removal (q-score).

3.1.4. Missing Data

Properties with missing or partial data were removed due to the impracticality of manually imputing property characteristics.

3.1.5. Calculated Data

Due to inaccuracies in how AirDNA calculates revenue data for properties, the estimated monthly revenue was calculated using the average daily rate and the average occupancy rate of all properties within the study. Whilst this isn't necessarily the most accurate approach since it does not consider the occupancy variations between each property, the method used by AirDNA heavily skews the performance of properties which are blocked for long periods and are not easily identifiable through data cleansing, preventing deeper analysis from taking place. This is supported by Agarwal, Koch, & McNab who also found an upward bias in the metrics published by AirDNA [16]. The report assumed the properties are held for a long period thus, averaging over 12 months for all properties is a reasonable assumption, especially in heavily tourist areas such as Paphos where demand is high.

3.2. Long-Term Rental Data

Long-term data needed to be collected manually through traditional means which involved collecting data from platforms such as Bazaraki, Facebook Marketplace and the agents' website. As this data is publicly available, no consent or approvals were required, however all data which could be considered sensitive was redacted.

 doi:10.20944/preprints202404.0187.v2

While doing this ensured that all the relevant data was collected, it meant that the process was extremely time-consuming. The main benefit of this approach was that because the data was vetted as it was recorded, very little data cleansing was required, and it was almost immediately ready for analysis.

The main limitation was the fact that the prices advertised were the asking price, not the agreed rental price. Throughout the data collection, it was evident that some values were highly inflated, so they were ignored, however, others were only marginally inflated and therefore remained. Although this marginal inflation may only be a few hundred euros, since it was apparent in many properties it's likely to impact upon the final HBU decision.

Over 400 properties were reviewed with only 203 properties being used due to inflated prices or missing data. As with the short-term data, the removed properties showed no correlation so their removal should have limited impact on the model's ability to make accurate predictions. Although this is a considerably small amount compared to the short-term rentals, this figure still falls within the recommended range found during the literature review.

3.3. Geospatial Data

Using QGIS, an open-source GIS application, the road network of the Paphos district and all properties and Points of Interest (POI) were mapped. With this information, the distances across the network were interpolated for each property and POI to define the distances to each. The POI included beaches, golf courses, hospitals, the central business district and Paphos airport to name a few. This data was then reintroduced into the raw data set for use within the regression models.

Heat maps of the monthly returns of both long and short-term properties showed some interesting trends regarding their geospatial distribution. Long-term, as shown in Figure 1, shows the highest returning properties were found outside of the cities in areas known for their luxurious nature. Whereas in Paphos, there is a range of different returns but generally, it remained lower, most likely linked to the property type and size, being apartment-focused whilst the suburbs were houses/villas.

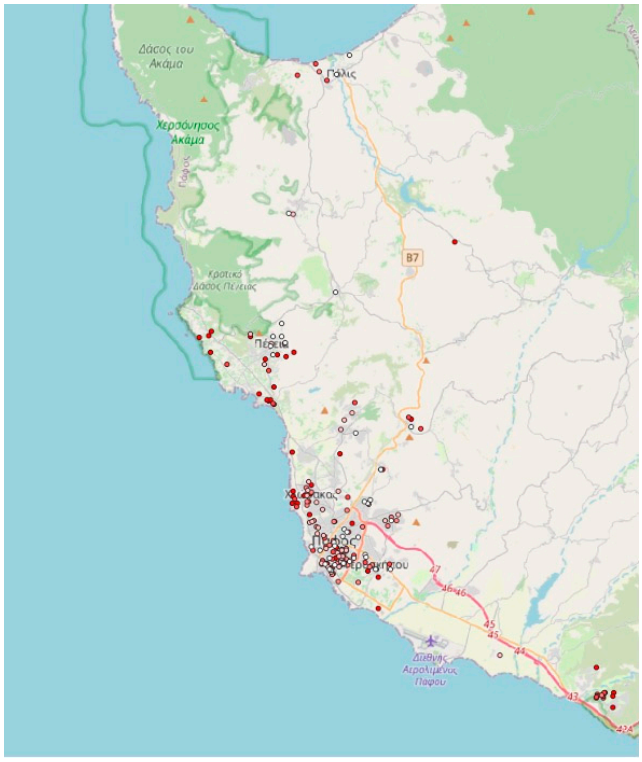

Figure 1. Price Heatmap for Long-Term Properties.

 doi:10.20944/preprints202404.0187.v2

A similar trend can be seen for short-term rentals, as per Figure 2, however, the extent is much more severe with the aforementioned areas having considerably higher values than their Paphos counterparts.

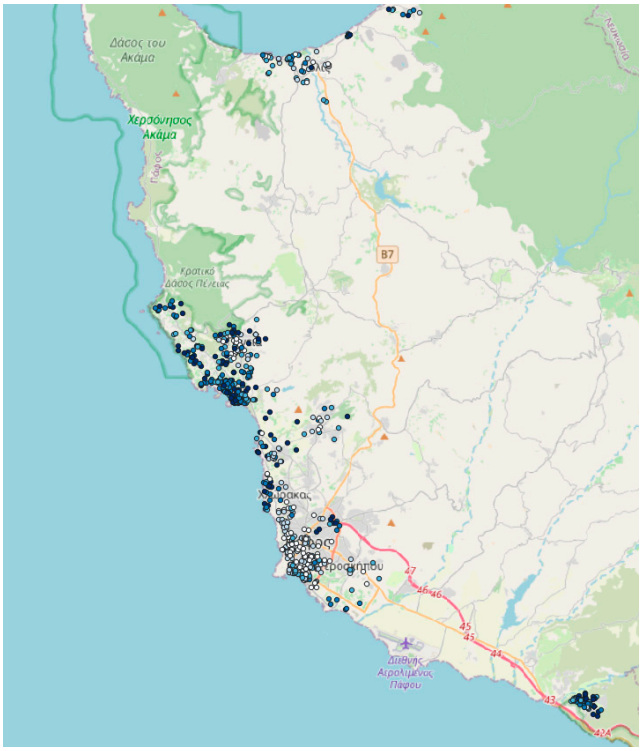

Figure 2. Price Heatmap for Short-Term Properties.

3.4. Regression Analysis

3.4.1. Data Preparation

The first approach used for the prediction of the alternative strategy was MLR, with property characteristics being a primary focus to allow investors to manage their property accordingly. Correlation between attributes, known as multicollinearity, was addressed before the regression as they severely affect the stability and interpretability of the model. This is in part due to them becoming hypersensitive to minor changes in the data, leading to large variations in the predicted results but they can also cause other variables to become insignificant in their presence. A correlation matrix was created to highlight any variables which had a high correlation which is commonly defined as follows:

$$x>0.5 \qquad x<-0.5, \tag{1}$$

Several variables were found to have a level of correlation, these can be seen in Table 1, and their relative correlation to other variables in addition to whether it remained in the analysis or not is shown.

Table 1. Regression Correlation Summary.

Independent Variable	Correlation
Property Type	Yes, high correlation with bedrooms and bathrooms.
Latitude	Regression model dependent, removed when using GIS data.
Longitude	Regression model dependent, removed when using GIS data.
Finish Quality	No correlation.
Bedrooms	No, property type removed instead.

 doi:10.20944/preprints202404.0187.v2

Bathrooms	Yes, removed due to correlation with bedrooms.
Parking	No correlation.
Pool	No correlation.
Dist. to Nearest Golf	Yes, correlation with lots of GIS variables.
Dist. to Nearest Beach	No, other variables removed due to importance to study.
Dist. to Nearest Hospital	Yes, correlation with lots of GIS variables.
Dist. to CBD	Yes, correlation with lots of GIS variables.
Dist. to Old Town	Yes, correlation with lots of GIS variables.
Dist. to Airport	Yes, correlation with lots of GIS variables.
Dist. to Harbour	Yes, correlation with lots of GIS variables.
Dist. to Sea Caves	Yes, correlation with lots of GIS variables.
Dist. to Coral Bay	No, other variables removed due to importance to study.
Dist. to Tombs of Kings	Yes, correlation with lots of GIS variables.
Dist. to Aphrodite Hills	Yes, correlation with lots of GIS variables.
Dist. to ISOP	Yes, correlation with lots of GIS variables.
Dist. to Polis	No, other variables removed due to importance to study.

When performing a regression analysis, categorical variables needed to be removed as this is another cause of multicollinearity in the model. To do this, dummy variables were created using a binary format, where 1 shows true and 0 false. In this analysis, dummy variables were taken for the finish quality, parking and pools. In addition, a technique called "dummy variable exclusion" was followed which sees one item from each category being removed, thus creating a none-perfect multicollinearity without affecting the overall fit or performance of the model. In equation form, the "dummy variable exclusions" look as follows:

$$Y = \beta 0 + \beta 1 * Dummy\_A + \beta 2 * Dummy\_B + \beta 3 * Dummy\_C, \quad (2)$$

Taking logarithms of values often improves regression analysis by reducing heteroskedasticity though, it was found to not be beneficial to the overall performance of the model.

3.4.2. Regression Models

Three regression models standard, geospatial and distance-based spatial regression were carried out on the dataset, with the variables being adjusted based on their nature and purpose, a summary of each and its variables can be seen in Tables 2–4.

Table 2. Standard Regression Descriptive Characteristics.

	Long-Term Data			**Short-Term Data**		
	Coefficients	**Std Error**	**P-value**	**Coefficients**	**Std Error**	**P-value**
Intercept	-239.5670	594.7145	0.6875	-1925.1626	231.6099	0.0000
Quality_Luxury	2772.9211	306.2989	0.0000	2753.5876	164.0873	0.0000
Quality_Good	741.5633	211.9946	0.0006	1807.0815	163.4928	0.0000
Quality_Standard	119.1536	198.8557	0.5497	950.0772	146.0243	0.0000
Bedrooms	613.5142	71.2152	0.0000	1564.2426	41.6072	0.0000
Parking_Yes	-44.8237	577.0686	0.9382	-237.3274	175.6139	0.1769
Pool_Yes	1256.7473	176.1012	0.0000	377.9773	140.8164	0.0074
Pool_Communal	138.4187	132.5471	0.2976	258.6321	195.4374	0.1861

Table 3. Geospatial Regression Descriptive Characteristics.

	Long-Term Data			**Short-Term Data**		
	Coefficients	**Std Error**		**Coefficients**	**Std Error**	
Intercept	12072.9884	57834.6681	0.8349	-201152.1098	40900.6610	0.0000
Longitude	-1210.8226	944.8119	0.2015	-1038.4441	605.9397	0.0870

 doi:10.20944/preprints202404.0187.v2

Latitude	920.7146	1080.5340	0.3952	7258.9539	814.3556	0.0000
Quality_Luxury	2713.3189	308.5177	0.0000	2845.4105	153.7311	0.0000
Quality_Good	664.6562	214.8312	0.0023	1792.2597	152.4968	0.0000
Quality_Standard	108.8621	197.8807	0.5829	859.3971	136.4362	0.0000
Bedrooms	615.4072	70.8811	0.0000	1629.2139	39.5664	0.0000
Parking_Yes	-61.9064	574.2631	0.9143	-280.3170	163.8273	0.0875
Pool_Yes	1309.2286	181.0859	0.0000	340.6442	133.3233	0.0108
Pool_Communal	110.7387	132.5807	0.4046	11.7053	183.7805	0.9492

Table 4. Distanced-Based Regression Descriptive Characteristics.

	Long-Term Data			**Short-Term Data**		
	Coefficients	**Std Error**		**Coefficients**	**Std Error**	
Intercept	-540.9797	616.8445	0.3816	-3380.4834	263.5197	0.0000
Quality_Luxury	2568.4478	306.6768	0.0000	2839.6123	156.2421	0.0000
Quality_Good	561.3841	215.1968	0.0098	1831.8435	155.3048	0.0000
Quality_Standard	47.1591	196.1454	0.8103	879.0517	138.3407	0.0000
Bedrooms	635.0916	70.4320	0.0000	1630.5508	40.5600	0.0000
Parking_Yes	4.8095	566.3701	0.9932	-283.5373	165.8014	0.0876
Pool_Yes	1320.2718	175.8417	0.0000	412.1615	137.3391	0.0028
Pool_Communal	60.9690	131.8932	0.6444	60.9744	190.0521	0.7484
Nearest_Beach	-0.0527	0.0225	0.0203	0.0904	0.0387	0.0197
Dist_Coral_Bay	0.0132	0.0090	0.1434	0.0286	0.0055	0.0000
Dist_Polis	0.0105	0.0078	0.1798	0.0311	0.0047	0.0000

It is important to note that these results were achieved after numerous optimisations involving data cleansing, logarithm tests and adjustments in the chosen independent variables. Since long and short-term rentals were to be compared as part of the study, the variables needed to be matched to ensure that they could be applied to the opposing dataset without impacting the integrity of the analysis.

3.5. Random Forest Models

Machine learning can be done through various techniques, for this study Python was used along with the scikit-learn (sklearn). Due to the characteristics of RF, the entire dataset was used so the model was able to capitalise on all descriptive characteristics which strengthened its prediction of returns.

As with MLR, the dummy variables were created since categorical ones are not handled well by RF, other than that, the data was not modified as logarithms and scaling did not impact the performance of the model.

Unlike MLR, because the entire dataset could be used for both long and short-term rentals, only two models were created, one for each rental strategy. These datasets were then separated into a "training" and "testing" split automatically using the sklearn module with the test size equalling 20% of the total values. In addition, various forest sizes were tested, and it was found that 100 trees performed the best in both situations.

Due to its black-box nature, it's not possible to define an equation as is the case with MLR however, Figures 3 and 4 show the relative feature importance for both long and short-term rentals.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0187.v2

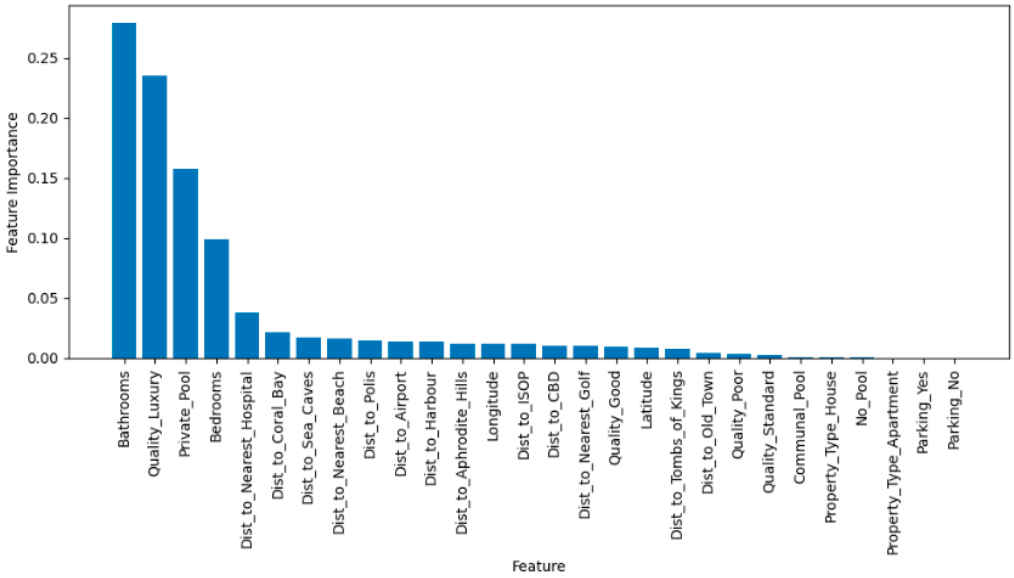

Figure 3. Random Forest Feature Importance (Long-Term Predictor).

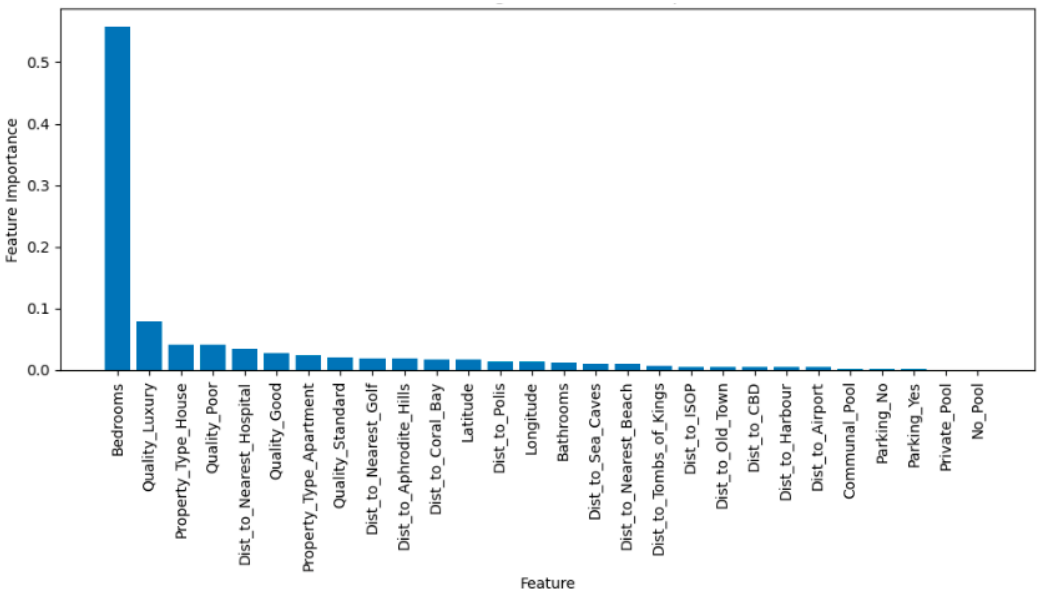

Figure 4. Random Forest Feature Importance (Short-Term Predictor).

3.6. K-Nearest Neighbour

Like RF, the KNN approach was carried out using sklearn on the Python platform and data was pre-processed similarly. There were several differences for KNN, the first being the conversion of longitude and latitude values to Cartesian coordinates which is due to the geometry of the Earth's surface being curved and not flat. Secondly, due to the model making predictions based on the distance between points, the data needed to be scaled so that larger scales did not disproportionately affect the results and helped to ensure that each feature was treated with equal importance.

Optimisation of the model was primarily carried out by adjusting the number of neighbours used for the predictions. To find this optimal value, a balance between variance and bias needed to be chosen, this was aided using the MSE, along with sklearn's integrated cross-valuation analysis. The MSE for training and validation for both long and short-term rentals can be seen in Figures 5 and 6 respectively.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0187.v2

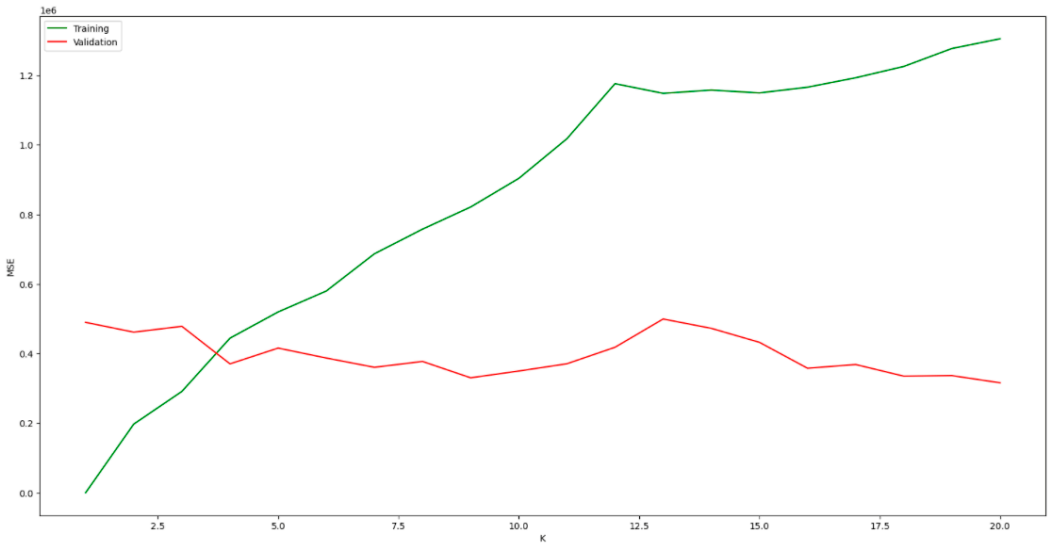

Figure 5. KNN Long-Term MSE Variation.

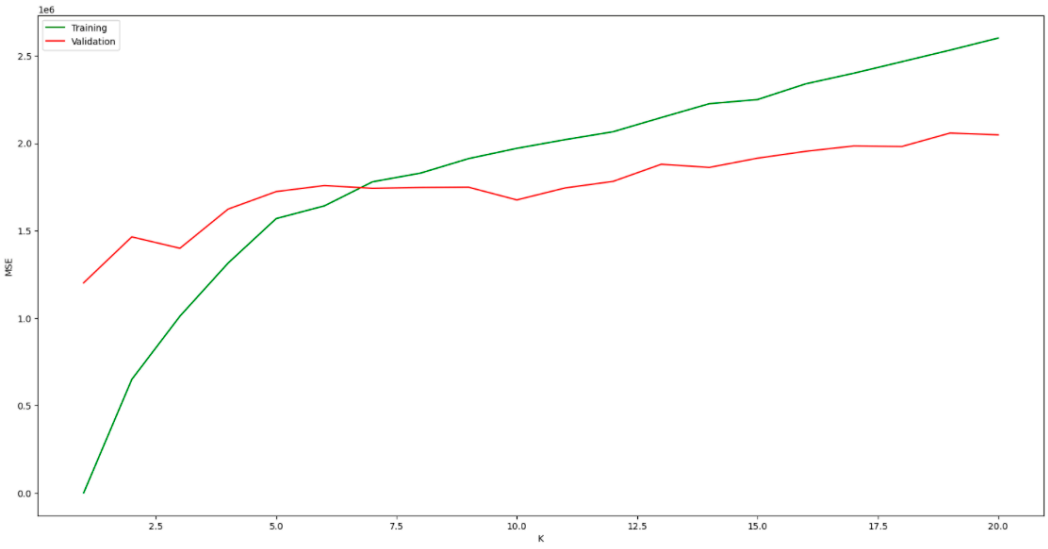

Figure 6. KNN Short-Term MSE Variation.

From this information, a "k" value of 5 was chosen for long-term rentals and a value of 7 for short-term. Due to the methodology of KNN, the weighting of each attribute is not defined and instead, the average of the nearest neighbours is used to predict the estimated returns.

4. Results

4.1. Optimal Model Selection

With multiple approaches being taken, identifying the best-fitting model was crucial to ensure that the predicted returns for the property were accurate. To do this, the R-squared value, also known as the coefficient of determination, was found, and compared for each model. The R-squared value is a measure of "goodness-to-fit" where 0 doesn't explain any variability and 1 explains the model perfectly. The R-squared value for each method can be seen in Table 5.

Table 5. Comparison of Model Performance (R-Squared).

	Long Term Predictor	Short Term Predictor
Random Forest	0.560	**0.843**

 doi:10.20944/preprints202404.0187.v2

KNN	0.518	0.752
MLR (Distance-Based)	**0.803**	0.739
MLR (with Lat/Long)	0.796	0.745
MLR (no Geo)	0.791	0.706

From the data, for long-term predictions, the Distance-Based MLR was chosen and for the short-term the RF model since they returned the highest R-Squared values of 0.803 and 0.843 respectively. These values show that in both cases over 80% of the variation in the rental incomes can be explained by the property characteristics used in the analysis. Based on these models, investigations into the alternative strategy were made based on the estimates they provided.

4.2. Descriptive Characteristics

For each property, the HBU was identified based on the predictions and a summary of their characteristics can be seen in Table 6.

Table 6. Characteristics of Predicted Data.

Short Term Data (Predicted)		**Long Term Data (Predicted)**	
Mean	€ 4,136	Mean	€ 3,082
Median	€ 3,333	Median	€ 3,077
Mode	€ 1,010	Mode	€ 5,108
Std. Dev	€ 3,032	Std. Dev	€ 1,239
Range	€ 20,666	Range	€ 5,630
Minimum	€ 628	Minimum	€ 869
Maximum	€ 21,294	Maximum	€ 6,500
Count	601	Count	427

From the total collected data points, around 58.5% of properties we found to achieve a higher return if they were rented short term compared to long-term. Moreover, on average the return for these properties was around 25.5% higher than that of their long-term counterparts however the standard deviation between them was over 59.1% higher.

4.3. Incorrect Rental Strategy

Hypothetically, it's possible for property owners to incorrectly market their property for the opposing strategy thus yielding lower returns. The cost of this can be seen in Figure 7.

 doi:10.20944/preprints202404.0187.v2

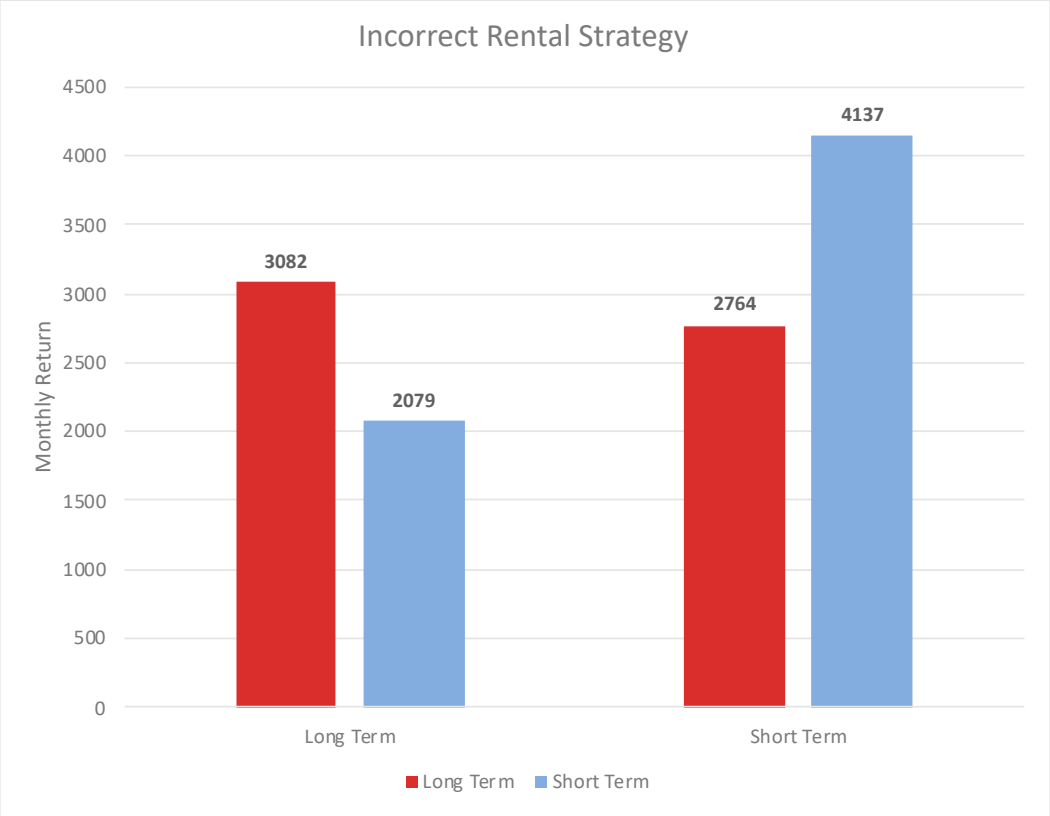

Figure 7. Cost of Incorrect Rental Strategy.

Where the HBU is a long-term rental, if a short-term rental is chosen then the owner would lose 1,003 EUR/month, earning only 67.5% of its potential. Whereas if the HBU is short-term and the property was rented out long-term, the owner would lose 1,373 EUR/month and around 66.8% of its potential.

4.4. Locational Trends

By plotting the HBU for each property in QGIS, the individual points for each approach, red for the long-term and blue for the short could be displayed, highlighting any potential trends. Shown in Figure 8 the distribution can be seen as relatively consistent across the city with short-term properties tending to be closer to the sea/beach due to a higher proportion of touristic amenities such as restaurants and shops. Interestingly, it showed that a short-term approach outperforms long-term in some villages that are located away from tourist activity.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2024 doi:10.20944/preprints202404.0187.v2

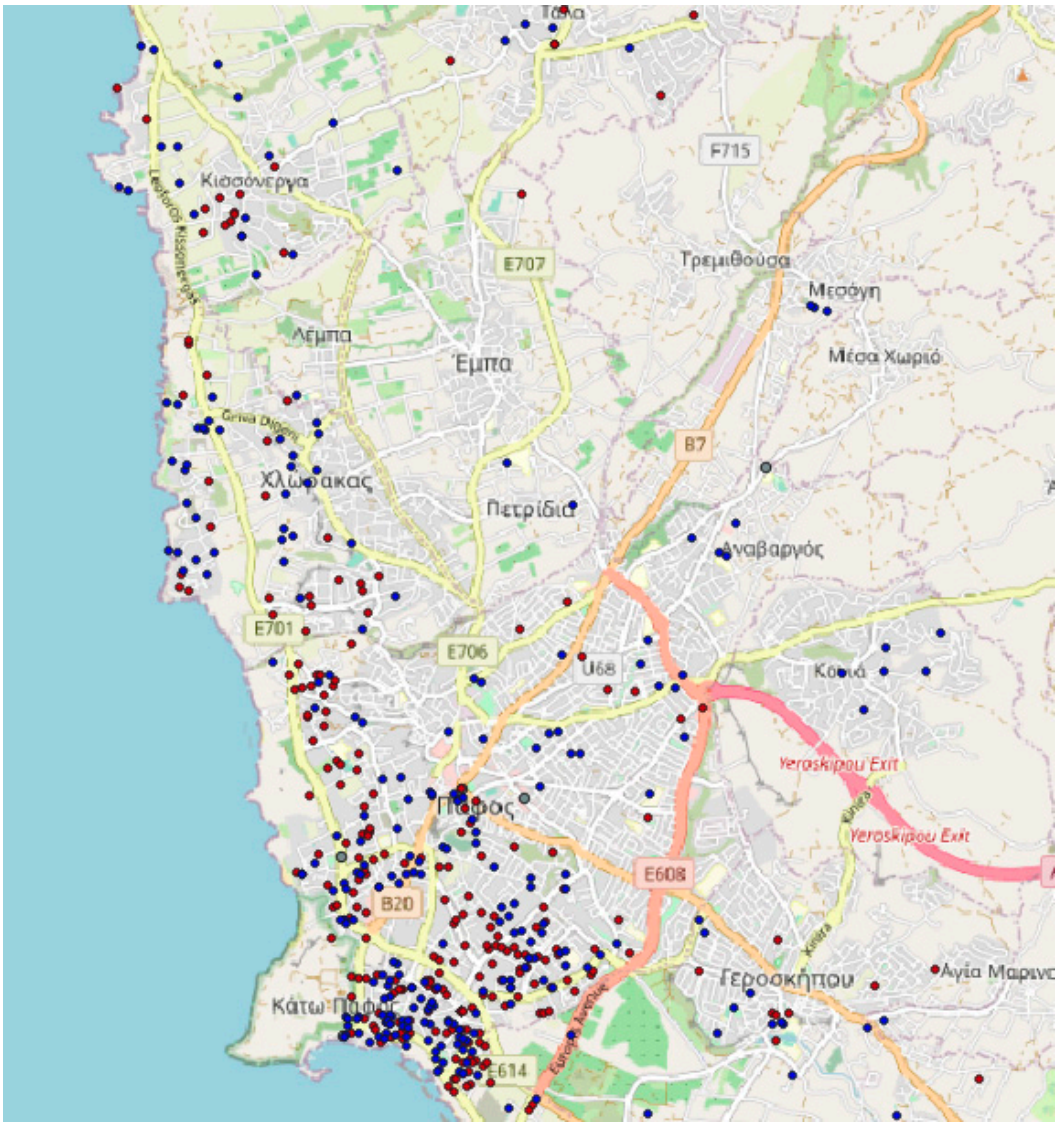

Figure 8. Long and Short-Term HBU Properties (Paphos).

4.5. Variable Weighting

Figures 3 and 4 highlight the relative importance of different variables regarding a specific rental approach. As such, it was found that for short-term predictions, the number of bedrooms was by far the most significant feature with the RF predicting an importance of 0.526, with the next most important being luxury finish with only 0.098. Thus, making the number of bedrooms over 5 times more important when predicting short-term returns.

For long-term predictions, it was a luxury finish which was found to be most important having an MLR coefficient of 2568, this was followed by a private pool with 1320 and finally bedrooms at 635. It's important to note that the first two features are only applied singularly, whereas the bedroom importance is applied multiple times depending on the number of bedrooms a property holds. For all data points recorded, the average bedrooms per property were 2.47, so the importance of bedrooms averages at around 1568, putting it behind luxury finish as the highest importance feature for long-term rentals.

4.6. Sensitivity Analysis

Since assumptions were made on the occupancy rate of short-term properties, two additional scenarios were briefly reviewed to assess how they would impact the strategy choice. These two

 doi:10.20944/preprints202404.0187.v2

scenarios took the original occupancy rate (approximately 65%) then increased and reduced it by 5%, as shown in Table 7.

Table 7. Sensitivity Analysis.

Short Term Data (-5% Occupancy)		Short Term Data (+5% Occupancy)	
Mean	€ 3,929	Mean	€ 4,323
Median	€ 3,283	Median	€ 3,452
Mode	€ 1,216	Mode	€ 1,441
Std. Dev	€ 2,898	Std. Dev	€ 3,229
Range	€ 19,077	Range	€ 22,257
Minimum	€ 579	Minimum	€ 676
Maximum	€ 19,657	Maximum	€ 22,933

For the pessimistic scenario (-5% occupancy), 534 properties were better rented short-term compared to 494 long-term. Whereas for the optimistic scenario (+5% occupancy), 669 properties were better rented short-term compared to 359 long-term. When compared to the original scenario, the number of properties where the HBU was higher for short-term rental was reduced by 67 in the pessimistic scenario but increased by 68 in the optimistic. As a percentage change, this was -10.9% and 11.1% respectively. Concerning the average rental income, the pessimistic scenario saw a reduction of €207 and around -5.01%, whereas the optimistic saw an increase of €187 and 4.53%. The rest of the characteristics varied by similar degrees.

5. Discussion

5.1. Optimal Strategy

During the study, ten predictive models for rental return were created, five for long and five for short-term rentals. When comparing these models to find the most suitable, the R-squared values, were used as the determination factor for identifying the best fitting, as was the case for several similar studies such as Shokoohyar, Sobhani, & Sobhani's [5].

In doing so, some clear differences between the long and short-term models were found which led to two different approaches being required for predicting the alternative strategy. Firstly, both ML methods for long-term rental were found to be significantly lower than that of the MLR approaches returning R-Squared values of 0.560 and 0.518 for RF and KNN respectively. Whilst ML models generally perform better than MLR since they can identify complex relationships and do not suffer from overfitting or multicollinearity, where data is limited, they can be outperformed simply due to insufficient data to make precise predictions. These results are also supported by Gnat & Doszyn's findings which were discussed previously [12]. Due to the size of the Paphos market, and the accuracy provided regarding long-term rental properties, acquiring sufficient data to achieve higher-performing ML models would be difficult since through extensive market research almost 50% of the properties advertised were not suitable. Therefore, the prediction of long-term returns was done through an MLR approach.

Conversely, where more data was available, both ML models outperformed the equivalent MLR methods as can be seen by the short-term rentals. It's also interesting to note that even with 825 properties, the ML models were able to achieve a maximum R-Squared value of 0.843 which when compared to similar studies, shows they perform well considering the amount of data. Moreover, it suggests that even with a data scraping approach to collecting data, a model can still return good results when extensive data cleansing takes place.

When comparing the MLR between the long and short-term predictors, strangely the long-term model outperforms the short-term in every approach, even when it contained significantly less data. Most likely this is due to manual data collection versus data scraping which results in the quality of data being much higher and inaccuracies such as outliers being more effectively removed. The drawback is the considerable time investment required to manually collect data, with only around

 doi:10.20944/preprints202404.0187.v2

20 properties being collected per hour, and the risk of human error involved. This also highlights the potential drawbacks of AirDNA data, be that the data collection approach or the processing they do to publish the data concisely. Regardless, these results show that both data quality and volume are crucial factors when deciding whether MLR or ML methods are used.

Finally, when comparing the three MLR methods, every model containing geographical data, be that geospatial or distance-based, outperformed the equivalent nongeographical model albeit to varying degrees. This supports the findings of Din, Hoesli, & Bender whereby whenever geographical data is available, it should be used to improve model performance [14]. In this study, the long-term regression was improved by 1.5% and the short-term by 5.2% just by introducing geographical data. For short-term prediction models, the extent to which the geographical was considerably more beneficial which may be in part due to the models generally performing worse than the long-term equivalents, but most probably due to the quality of the data. As such, this suggests that where data quality is poorer, geographical data can more significantly increase its accuracy compared to an already well-performing model. This finding also supports the conclusion that AirDNA's data may have issues regarding its accuracy.

5.2. Variable Weighting

For long-term properties, it was found that the most significant factor to consider was the overall finish quality of the property, with luxury holding the highest weight of all property attributes. Improving finish quality can be done with varying degrees of ease depending on the specific property, be that its age or structure, so targeting this must be carefully assessed taking into consideration ROI.

Pools also considerably increased the achievable returns especially when a private pool was available. Consequently, investors should consider the practicality of installing these to their properties where possible as it will increase the return and add value to the property should they wish to sell. As with renovation, the construction of a pool is not simple and has associated issues including acquiring permits, losing income during construction and additional costs for landscaping before/after the work.

Adding bedrooms to a property also increases returns but would most likely require an extension, however, conversion of spare rooms within the property could be an option. Most studies also support this, including those by Limsombunchai, Gan & Lee and Shokoohyar, Sobhani, & Sobhani, which is unsurprising since the number of bedrooms correlates strongly with property size [11,5].

For all the variables, conversion of existing properties is most likely not cost-effective unless the property requires it. For new investors, when looking for a property to invest in, they should consider each of these and how they can ensure that they are maximised to increase returns. The purchase price for acquiring a property which maximises these characteristics will be higher than those without, so the ROI should be calculated to identify the best property accordingly.

The variable weight for short-term properties on the other hand was significantly weighted towards the number of bedrooms compared to all other attributes including a luxury finish. This is understandable since the main driving factor for short-term renters is the number of people whom it can sleep and less about the finish quality. In addition, AirBnB is not aimed at luxury travellers so there are much fewer luxury properties compared to the normal quality expected from short-term rentals. Therefore, for short-term rentals, the focus should be on sleeping as many people as possible whilst maintaining a quality which still ensures it is marketable and attracts guests.

5.3. Locational Importance

It was found that the location of the property had a relatively low importance when compared to the other features such as the number of bedrooms and finish quality according to the data. This differs from many other studies' findings including Kim, Kwon, & Choi's which found that rental income was directly related to the proximity to favourable and less favourable areas [17]. A potential reason for this finding is that when looking at the location of properties in the study, they are

 doi:10.20944/preprints202404.0187.v2

predominantly located within tourist areas of Paphos, close to the coast, due to the nature of the city. Interestingly, it also shows that regardless of the location, both long and short-term approaches can be viable assuming that the property is in or near tourist areas and even suburban locations can see higher returns for short-term rentals. When looking at Paphos this can be attributed to the large volume of tourists, with those seeking larger cheaper properties opting for the suburbs compared to the centre when on a tighter budget. This range in demand means there is a large area of influence for short-term rentals, so assuming the property has the desired characteristics, the location may have less of an impact.

5.4. Incorrect Strategy

Assuming the model is accurately able to predict the optimal rental strategy for a property, it shows that many investors should consider adjusting their approach from short to long-term. This is evident by 224 short-term properties finding higher returns if they were rented as long-term (of course this may be slightly more since some long-term properties may also have the incorrect strategy). This stands as a stark reminder to investors to continually review their investment strategy as economic trends such as rising interest rates and reduced construction, can significantly impact the dynamic of the market.

It was also found that renting a property out using the incorrect strategy would result in losing between 66.8% and 67.5% of income depending on which strategy returned the highest. As an investor, this difference is significant and can make or break an investment which clearly shows the importance of thorough portfolio management.

5.5. Sensitivity Analysis

When comparing the optimistic and pessimistic scenarios, the overall trends were consistent with the scale of the variation. This is evident from the average returns increasing and decreasing by similar magnitudes compared to the occupancy rate, which is expected since these variables were used to define the models. Nevertheless, it's noteworthy that even with a 5% change in occupancy rate, a minimum of a 10% shift between the HBU strategy was seen in favour of the alternative strategy. This reaffirms the importance of the occupancy rate for high returns but also shows that in uncertain tourist conditions, a more reliable approach, in the form of long-term, could be most advantageous.

6. Conclusions

In summary, it was found that with a larger dataset of 825 properties, ML methods such as RF and KNN significantly outperformed standard MLR approaches. Where larger datasets cannot be found, the opposite occurs whereby the MLR models outperform ML. In these cases, the incorporation of geographical information in the form of longitude and latitude or spacial distances improves the performance of the models, and their incorporation should be used whenever possible.

The data collection methods used had both benefits and drawbacks, be that the high achievable volume from data scraping or the accuracy of manual collection. It is, therefore, important to assess the needs in future studies and apply them accordingly. In general, where lower amounts of data are required, a manual approach should be taken as it ensures that the data collected is accurate and reduces the risks of misinterpretation which can be common for data scraping. Where large amounts of data are required or data over a period of time, data scraping, when cleaned well, can still result in good results especially when used in combination with ML methods.

Several key variables were found to impact the returns of properties, regardless of the strategy, these included the number of bedrooms, finish quality and for long-term properties, whether the property had a pool. These variables are difficult to change once the property is constructed without significant financial and time investments, so as part of the purchasing process investors need to pay special attention to these factors so that they can capitalise on their investment.

 doi:10.20944/preprints202404.0187.v2

The study also found that many investors had the incorrect rental strategy which significantly impacted their returns. This emphasises the importance of good portfolio management, be that understanding economic changes or flexibility to change to capitalise on current demands. Moreover, investors should always review their situation so they can plan accordingly through more beneficial lease contracts and marketing material to enable seamless changes.

Limitations and Future Research

Due to limited data, the volume of long-term rental data was relatively small meaning that the ML models were not suitable. Throughout the literature review, it found that more accurate results could be achieved through these methods so it can be assumed that with more data the results would be more accurate. To achieve this, additional research needs to be done to create a tool which can automatically carry out data scraping on the market through various sources to capture large volumes of data at one time. Of course, the data quality must remain high, or the benefit would be offset by the inaccuracies caused by poor data.

Another limitation was the fact that AirDNA data had large amounts of "ghost" properties and properties which were clearly overpriced or incorrectly marketed. The main reason for that was that owners did not input the data properly to the platform. As a result, significant time was spent cleaning the data, but without manually checking each data point, it's hard to reliably say that all outliers were removed. Although studies into the accuracies found that it could generally be relied upon, they also found that AirDNA tended to overestimate the potential returns [15]. Within this study, this overestimation was not considered, nor was the fact that the long-term rentals were advertised and not agreed prices, so there is an inherent uncertainty as to the extent to which these impacted the predicted returns. For this reason, future research tracking the variations of these advertised values compared to the agreed would be interesting and would paint a clearer picture for investors on how to market their properties. It could also be applied retrospectively to the study assuming economic conditions haven't changed too significantly.

Another figure which was not considered was the operational costs associated with both short and long-term rental. Whilst these figures vary greatly depending on the property, in general, short-term operational costs are much higher than long-term which will therefore impact profitability. Additional research should be done in this area, and again, the results can be applied retrospectively.

As discussed previously, the same occupancy rate was assumed for all short-term rental properties based on the average occupancy rates of all properties in the study. This assumption will have skewed the data for certain properties, but due to the unreliability of the AirDNA data regarding the revenue it was the only option for estimations of the HBU. That being said, additional research into the short-term rentals in the Paphos area could shed light on the true occupancy rate whilst additional analysis of AirDNA revenue calculations may help better define potential revenues for the Paphos region.

Finally, as all real estate professionals know, the age of a property significantly impacts the returns which you can expect, regardless of the strategy. Unfortunately, this data is not published on short-term rental platforms, so was left from the analysis.

Author Contributions: Conceptualization, S.M. and T.D.; methodology, S.M.; software, S.M.; validation, S.M.; formal analysis, S.M.; investigation, S.M.; resources, S.M.; data curation, S.M.; writing—original draft preparation, S.M.; writing—review and editing, S.M.; visualization, S.M.; supervision, T.D and M.K.; project administration, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Due to privacy reasons, the dataset and generated data are unavailable to the public.

Acknowledgments: We would like to express our gratitude to all those who have contributed to the completion of this work, including members of Neapolis University Paphos for providing the necessary resources and fostering an environment conducive to academic growth. We want to extend our thanks to Axia Valuers for

supplying the AirDNA data, which has enabled us to carry out this research and AirDNA for allowing the information to be shared.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cystat. Available online: www.cystat.gov.cy/en/Announcement?id=65400 (accessed on 31 October 2023).
2. Cystat. Available online: www.cystat.gov.cy/en/PressRelease?id=67917 (accessed on 31 July 2023).
3. European Commission. Available online: www.smart-tourism-capital.ec.europa.eu/pafos-winner-2023-competition_en (accessed on 9 August 2023).
4. Cyprus Mail. Available online: www.cyprus-mail.com/2022/05/18/paphos-airport-sees-600000-passengers-in-first-four-months-of-2022 (accessed on 18 May 2022).
5. Shokoohyar, S.; Sobhani, A.; Sobhani, A. determinants of rental strategy: short-term vs long-term rental strategy. *International Journal of Contemporary Hospitality Management* **2020**, *32*, 3873-3894. https://doi.org/10.1108/IJCHM-03-2020-0185
6. Manganelli, B.; Tataranna, S.; De Paola, P. A Comparison of Short-Term and Long-Term Rental Market in an Italian City. *Computational Science and Its Applications* **2020**, 12251, 884-899.
7. Rodríguez-Pérez de Arenaza, D.; Ángel Hierro, L.; Patiño, D. Airbnb, sun-and-beach tourism and residential rental prices. *Current Issues in Tourism* **2019**, 25, 3261-3278. https://doi.org/10.1080/13683500.2019.1705768
8. AirDNA. Available online: https://www.airdna.co/vacation-rental-data (accessed on 8 July 2023).
9. Nguyen, N.; Cripps, A. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *Journal of Real Estate Research* **2020**, 22, 313-336. https://doi.org/10.1080/10835547.2001.12091068
10. Benjamin, J. D.; Guttery, R. S.; Sirmans, C. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation. *Journal of Real Estate Practice and Education* **2020**, 7, 65-77. https://doi.org/10.1080/10835547.2004.12091602
11. Limsombunchai, V.; Gan, C.; Lee, M. House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences* **2004**, 1, 193-201. https://doi.org/10.3844/ajassp.2004.193.201
12. Gnat, S.; Doszyń, M. Parametric and Non-parametric Methods in Mass Appraisal on Poorly Developed Real Estate Markets. *European Research Studies Journal* **2020**, 23, 1230-1245. https://doi.org/10.35808/ersj/1740
13. Dimopoulos, T.; Bakas, N. Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. *Remote Sensing in Applications of Geoinformation* **2019**, 11, 3047-3062. https://doi.org/10.3390/rs11243047
14. Din, A.; Hoesli, M.; Bender, A. Environmental Variables and Real Estate Prices. *Urban Studies* **2001**, 38, 1989-2000. https://doi.org/10.1080/00420980120080899
15. Fleischer, A.; Ert, E.; Bar-Nahum, Z. The Role of Trust Indicators in a Digital Platform: A Differentiated Goods Approach in an Airbnb Market. *Journal of Travel Research* **2021**, 61, 1-14. https://doi.org/10.1177/00472875211021660
16. Agarwal, V.; Koch, J. V.; McNab, R. M. Differing Views of Lodging Reality: Airdna, STR, and Airbnb. *Cornell Hospitality Quarterly* **2018**, 60, 193-199. https://doi.org/10.1177/1938965518777218
17. Kim, H.; Kwon, Y.; Choi, Y. Assessing the Impact of Public Rental Housing on the Housing Prices in Proximity: Based on the Regional and Local Level of Price Prediction Models Using Long Short-Term Memory (LSTM). *Sustainability* **2020**, 12, 7520. https://doi.org/10.3390/su12187520.