Real Estate and AirBnb
Case Study in Austin (Texas)

Federico Squasselli - 0001137411
Giannicola Luciano - 0001135123

MA Business Analytics course in
Forecasting and Predictive analytics
Professor Luca Trapin - Tutor Alessandro Morico
Alma Mater Studiorum - Università di Bologna

March - August, 2024

1 Introduction and Problem Statement

In this project, our team represents "AirHomes & Co," a real estate and Airbnb agency based in Austin, Texas. Specifically, we are part of the analytics team, focusing on data-driven insights to optimize property investments and Airbnb rentals.

The agency's tasks include estimating and predicting property prices, determining optimal times for purchasing and selling properties, and maximizing profits in the resale process. Additionally, the agency considers renting out properties for short stays to minimize inefficiencies during periods when the property is not being sold. This strategic approach helps to mitigate potential losses from holding unsold inventory and ensures that the properties are used efficiently.

This project focuses on three key analysis to support AirHomes & Co's strategic decisions:

1. **Time Series Analysis of the Real Estate Market:** The aim is to identify the optimal times for purchasing or selling properties by observing historical market data. Accurate forecasting of the ZHVI enables AirHomes & Co to strategically time their transactions, reducing risk and enhancing overall efficiency in the real estate market.

2. The second analysis is focused on predicting if a property is going to be **sold within two months**. By prediciting the sale timing, we can determine which properties should be listed immediately for sale and which should be rented out.

3. **Short-term Rental Price Prediction:** With the third one we want to determine the optimal pricing for short-term rentals. Effective pricing is crucial for maximizing revenue and ensuring high occupancy rates on platforms like Airbnb.

2 Short-term Rental Price Prediction

2.1 Datasets

The first dataset used for the analysis has been retrieved from Inside Airbnb, a public platform with a vision to empower communities with data and information to understand, decide and control the role of renting residential homes to tourists. The reference year of the data is 2023, and includes only listings that are entire homes/apartments since the aim of AirHomes & Co is to short-rent unsold properties.

Table 1: Inside AirBnB Dataset variables description

Name	Description	Type
h_superhost	Host superhost status	int
h_list	Number of listings by host	int
h_list_tot	Total number of listings by host	int
h_id_ver	Host ID verified status	int
lat	Latitude coordinate	num
long	Longitude coordinate	num
prop_type	Type of property listing	chr
accommodates	Number of guests accommodated	int
bedrooms	Number of bedrooms	int
beds	Number of beds	int
price	Price per night for the listing	num
min_nights	Minimum number of nights required for booking	int
max_nights	Maximum number of nights allowed for booking	int
avlb	Availability of listing	int
avlb_30	Availability in the next 30 days	int
avlb_60	Availability in the next 60 days	int
avlb_90	Availability in the next 90 days	int
avlb_365	Availability in the next 365 days	int
tot_rev	Total number of reviews for the listing	int
rev_ltm	Number of reviews in the last twelve months	int
rev_l30d	Number of reviews in the last 30 days	int
rev_score	Review score (0-5)	num
rev_acc	Review accuracy score (0-5)	num
inst_avlb	Instant booking availability	int
monthly_rev	Average monthly revenue for the listing	num
Region	In what region the property is located	chr

2.2 Methodology

To address the challenge of predicting short-term rental prices, this study employs several regression models that incorporate different property characteristics and market dynamics. The models are designed with a log-transformation of the dependent variable, primarily to address issues related to the distribution of the data. The original price distribution showed significant right skewness and a long tail, additionally the log traformation mitigated the impact of outliers. We have considered four distinct specifications of the model. All models are linear and estimated using Ordinary Least Squares (OLS) technique.

The available data set has been split into train and test set (50% for each of them), resulting in 4318 observations for the train and the same number for the test set.

1. The first specified linear model is a full model, **Complete Model**, considering all the variables available in the dataset without transformations. This model aims to include as much information as possible and it serves as a benchmark to compare the performance of more refined models as suggested by the backward subset selection.

$$\log(\text{price}) = \beta_0 + \beta_1 \text{h\_superhost} + \beta_2 \text{h\_list} + \beta_3 \text{h\_list\_tot} + \ldots + \beta_{38} \text{regionSW} \quad (1)$$

2. The **Lasso model** performs variable selection by shrinking some coefficients to zero.

$$\log(\text{price}) = \beta_0 + \beta_1 \text{lat} + \beta_2 \text{long} + \beta_3 \text{Entire guest suite} + \ldots + \beta_{18} \text{monthly\_rev}^2 \quad (2)$$

3. **Transformed Complete.** The first model choosen that applies logarithmic and polynomial transformations to capture non-linear relationships in the data with the aim to improve model performance in predicting.

$$\log(\text{price}) = \beta_0 + \beta_1 \text{h\_superhost} + \beta_2 \text{h\_list} + \ldots + \beta_{47} \text{monthly\_rev}^2 \quad (3)$$

4. The **reduced transformed** linear model focuses on a subset of predictors of the complete transformed model with a p-value between 1% and 5%.

$$\log(\text{price}) = \beta_0 + \beta_1 \text{h\_list} + \beta_2 \text{h\_list\_tot} + \beta_3 \text{lat} + \ldots + \beta_{24} \text{monthly\_rev}^2 \quad (4)$$

2.3 Analysis

In this section, we assess the performance of the four different regression models using Mean Squared Error (MSE) and Bayesian Information Criterion (BIC) on the original price scale. These metrics provide a comprehensive view of the models' performance, taking into account both prediction accuracy and model complexity. The MSE was calculated on the test set, which was not used as a criterion for model selection but was only used subsequently to monitor the models' performance. The results are summarized in Table 2.

Model	MSE	BIC
Complete	78279.84	7033.591
Lasso	78304.855	6968.448
Transformed Complete	81257.33	6981.262
Transformed Reduced	7513.36	6869.300

Table 2: Evaluation Measures for the 4 Models

The Complete model, serving as a baseline, has an MSE of 78,279.84 and a BIC of 7,033.591. While the Lasso model has a slightly higher MSE of 78,304.855, it achieves a lower BIC of 6,968.448, suggesting a better balance between model complexity and fit compared to the Complete model. The Transformed Complete model shows the highest

MSE (81,257.33), indicating it is the least accurate in predicting prices on the original scale. In contrast, the Transformed Reduced model outperforms all others with the lowest MSE (7,513.36) and the lowest BIC (6,869.300). This suggests that the Transformed Reduced model has the best balance between simplicity and predictive accuracy, making it the most efficient model among those evaluated.

To further compare the predictive accuracy of the Transformed Complete and Transformed Reduced models since they had the lowest BIC, we performed the Diebold-Mariano test. The test yielded a test statistic of DM = 0.70804 and a p-value of 0.479. This result suggests no statistically significant difference in predictive accuracy between the two models.

3 Will the Property be Sold within Two Months?

The analysis evaluates three logistic regression models, each with different levels of complexity, to determine the most effective approach for predicting if the property will stay in the market for less than 60 days or not.

3.1 Dataset

The second dataset has been retrieved from OpenML, and includes data for all real estate listings in the capital. The reference year for this dataset is 2017.

Table 3: OpenML Dataset variables description

Name	Description	Type
med_price	Median property price	num
med_price_M	Monthly change in median property price	num
med_price_Y	Yearly change in median property price	num
act_count	Number of active listings	int
act_count_M	Monthly change in the number of active listings	num
act_count_Y	Yearly change in the number of active listings	num
d_market_M	Monthly median days on market	num
d_market_Y	Yearly median days on market	num
new_count_M	Monthly change in count of new listings	num
new_count_Y	Yearly change in count of new listings	num
price_incr	Number of listings with price increases	int
price_decr	Number of listings with price decreases	int
price_decr_M	Monthly change in listings with price decreases	num
price_decr_Y	Yearly change in listings with price decreases	num
pend	Number of pending listings	int
pend_M	Monthly change in pending listings	num
pend_Y	Yearly change in pending listings	num
avg_price	Average property price	num
avg_price_M	Monthly change in average property price	num
avg_price_Y	Yearly change in average property price	num

Continued on next page

Table 3

Name	Description	Type
tot_list_M	Monthly change in the total number of listings	num
tot_list_Y	Yearly change in the total number of listings	num
pend_ratio	Ratio of pending to total listings	num
pend_ratio_M	Monthly change in the ratio of pending listings	num
pend_ratio_Y	Yearly change in the ratio of pending listings	num
sold_2m	Sold within two months	int

3.2 Methodology

In this analysis, we employ logistic regression to predict whether a property will be sold within two months, denoted as the binary outcome variable *sold_ 2m*. The study employs three different logistic regression models. Each model incorporates different sets of predictors and transformations of them. Logistic regression is particularly suitable for this task as it estimates the probability of a binary outcome. The available dataset has 4611 observation that have been splitted into train and test set, 2305 for the train and 2305 for the test.

1. The first logistic regression model is the **complete model**, which incorporates all available variables from the dataset.

$$\mathrm{P}(\mathrm{sold\_2m}) = \beta_0 + \beta_1 \mathrm{med\_price} + \beta_2 \mathrm{med\_price\_M} + \cdots + \beta_{25} \mathrm{pend\_ratio} \quad (5)$$

2. The **complete transformed model** extends the previous model by applying various transformations to the predictor variables.This approach helps address potential non-linearitys, enhancing the model's ability to capture complex relationships within the data.

$$\mathrm{P}(\mathrm{sold\_2m}) = \beta_0 + \beta_1 \log(\mathrm{med\_price}) + \cdots + \beta_{28} \mathrm{pend\_ratio\_Y} \quad (6)$$

3. The second model with variable transformations, the **reduced transformed**, has a reduced set of predictors compared to the previous one. We selected only significant variables with p-values between 1% and 5%. This model aims to simplify the previous one.

$$\mathrm{P}(\mathrm{sold\_2m}) = \beta_0 + \beta_1 \log(\mathrm{med\_price}) + \cdots + \beta_{16} \mathrm{pend\_ratio\_Y} \quad (7)$$

3.3 Analysis

In this section, we evaluate the performance of the three logistic models using MSE and Accuracy. The results of these evaluation measures are summarized in Table 4.

The probability threshold for classifying a property as sold within two months is set at 0.5. This means that if the predicted probability is greater than or equal to 0.5, the property is classified as sold within the specified period (`sold_2m` = 1), `sold_2m` = 0 if the probability is lower than 0.5.

The Transformed Complete model shows the best performance with the lowest MSE (0.0767) and highest accuracy (0.8959), suggesting it effectively captures complex patterns. The Reduced Transformed model also performs well, slightly behind in MSE (0.0769) and accuracy (0.8933). The Complete model, serving as a baseline, has the highest MSE (0.0981) and lowest accuracy (0.8673).

	MSE	Accuracy	AUC
Complete	0.0981	0.8673	0.9363
Transformed	0.0767	0.8959	0.9595
Reduced Transformed	0.0769	0.8933	0.9595

Table 4: Evaluation Measures for the three models

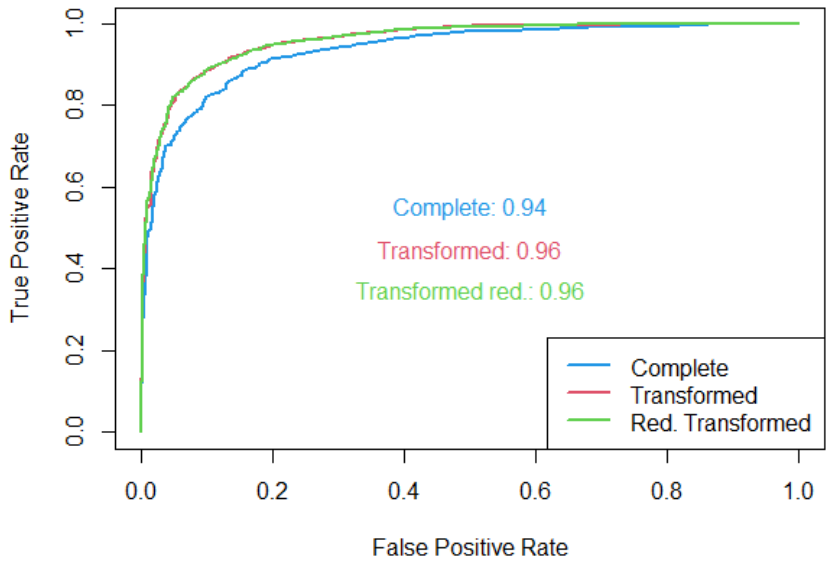


Figure 1: Roc curve for the 3 models

To assess discriminative power, we use the Area Under the Curve (AUC) metric from ROC curve analysis. The Transformed and Reduced Transformed models both achieve high AUC values of 0.9595, outperforming the Complete model's AUC of 0.9363. This indicates that the transformed models are better at distinguishing between properties that will be sold within two months and those that will not. Overall, the Transformed models demonstrate superior predictive capabilities across all metrics. The Reduced Transformed model, in particular, offers a good balance of simplicity and performance.

4 Time series Analysis. Forecasting ZHVI

4.1 Dataset

The dataset used for the time series analysis has been retrieved from Zillow Home Values, a comprehensive platform that provides data on property values. This dataset includes monthly observations from 1996 to the present and is essential for forecasting the Zillow Home Value Index (ZHVI) specific to Austin, Texas.

This data includes two variables:

Name	Description	Type
Date	The month and year of the observation	Date
ZHVI	The Zillow Home Value Index value for that month	Numeric

Table 5: Time series Dataset structure

4.2 Methodology

This part of the report focuses on forecasting. Firstly, it is really important to understand on which type of data we are working, identifying whether the ts that we are dealing with is a stationary one ready for the work, or if, having a non-stationary ts, we have to integrate data in some ways to obtain stationarity. Once we have identified the type of ts, we tried to add some trend and seasonal variables on the integrated data. Then we split the data in a training set and a test set in order to choose the best model(s) to make predictions based on our company's goals. The candidates models are Arima and Sarima models with different parameters for the autoregressive and the moving average order. The first ones (Arima models) are suitable for non-seasonal data where random changes can happen, while Sarima models are best with seasonal data that could follow some kind of pattern.

4.3 Analysis

First thing first, we turn our data into a ts object to obtain some information about it. Analysing the ts data and visualizing it using decomposition (trend, seasonality and noise components), it is possible to notice an increasing trend and a slight seasonality during spring and summer of each year.

The results coming from the Augmented Dickey-Fuller test and the Phillips-Perron Test are indicative of a non-stationary ts, as the obtained p-value is equal to 0.99. For this reason it is useful to integrate data through differentiation:

$$y'_t = y_t - y_{t-1}$$

finally obtaining a stationary ts, as shown by a 0.01 p-value. At this point we can build some linear models for our data to see if trend and seasonal variables can be useful in the explanation, by resulting significant, or not. The trend variable is the only significant one, while seasonal variables seem to add nothing in terms of explanation, never resulting signficant at all. The Adjusted R-squared of nearly 0.75 every time, is maybe simply reflecting the autocorrelation in the data rather than the true explanatory power of the model. It seemed obvious that working with differentiated data could result in better and more accurate insights, so we tried an AR1 on differentiated ts:

Estimate	**p-value**	**Adj R-squared**
0.61715	<2e-16	0.3697

Table 6: AR1 on differentiated data

that could be surely improved as spikes in the Acf are out of the confidence interval. Similar results are obtained adding the deterministic trend to the previous model:

Estimate	**p-value**	**Adj R-squared**
0.97005	0.0832	0.3736

Table 7: Deterministic trend coefficients in AR1 on differentiated data

We can now split the total of 324 observation into a train set made of 259 observations (80% of total observations) and a test set made of 65 observations (20% of total observations). Here are some of the results that we obtained with different models' BIC:

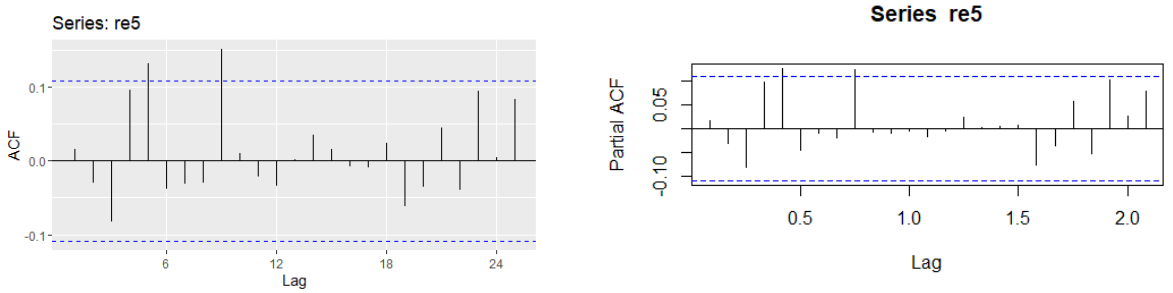
Model	**Arima110**	**Arima111**	**Arima211**	**Arima212**	**Sarima211**	**Sarima212**
BIC	4213.053	4213.675	4195.016	4190.793	3925.238	3927.656

Table 8: Models' BIC

As we can see from Table 8, Sarima(2,1,1) and Sarima(2,1,2) have the lowest BIC. Among all Arima models, Arima(2,1,2) has the lowest BIC, followed by Arima(2,1,1).

If we take a look at the behaviour of residuals for each of the six models, we can notice that not all of the spikes are inside the confidence interval for all of the Arima models, suggesting that the actual values are not adequately explained by the models. One possible explanation is that the models might not be capturing all relevant features such as seasonality, trends, or autoregressive patterns, and that they could be too simple or that some omitted variables could improve the fit. So, from this point on, we will focus on Sarima models as they resulted as being the best in terms of BIC and have almost

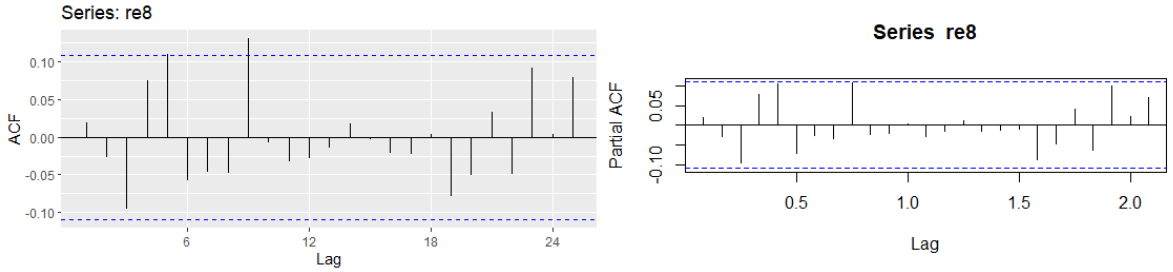
all of their spikes inside the confidence interval. While the Acf (Autocorrelation function) includes both direct and indirect effects, the Pacf (Partial Autocorrelation function) isolates the direct effect of a particular lag on the time series, without the influence of intermediate lags. Our two Sarima models only have few spikes that are significantly different from zero and go over the limits of the 95% confidence interval. Let's have a look at results from Sarima(2,1,1):



(a) Acf (b) Pacf

Figure 2: Acf and Pacf for Sarima(2,1,1)

At this point we wondered: what if we consider a differentiation order of two? Following this path we created two new models, Sarima(2,2,1) and Sarima(2,2,2). The first one doesn't give back any better results than the previous ones. Sarima(2,2,2), instead, not only has a BIC similar to the other Sarima models chosen (3926.441) but also has a Pacf where none of the spikes go outside of the confidence interval.



(a) Acf (b) Pacf

Figure 3: Acf and Pacf for Sarima(2,2,2)

We can now compute the MSE on the test set for each model, which can not be used as a selection criterion for both models and Arima's parameters.

Model	**Arima211**	**Sarima211**	**Sarima212**	**Sarima222**
MSE	18065521	10829967	12162221	9975069

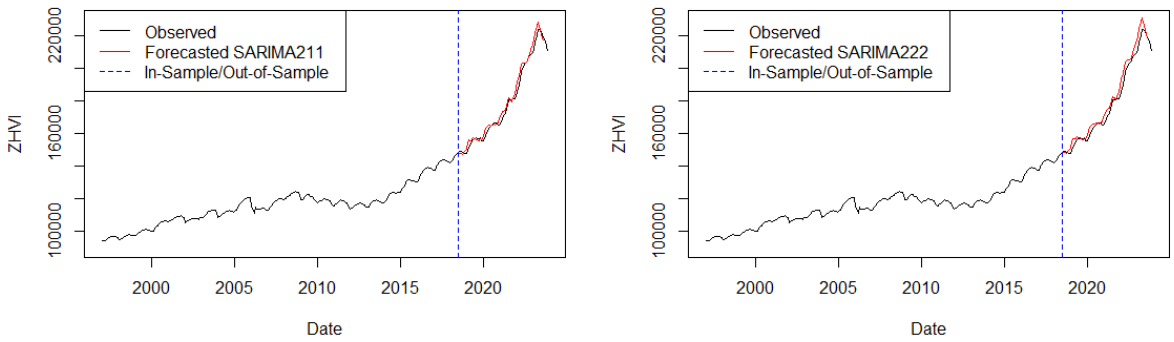
Table 9: Models' MSE

Sarima models even present lower values of the MSE, which can be seen as a subsequent evaluation of the model's performance in the 'future'.

4.4 Three-step ahead forecast of the models on out-of-sample data

Our agency decide whether to buy or sell apartments after a period of three months, so our forecast will be oriented in three months from the moment it will be made.

So, what will happen in the next three months?



(a) Sarima(2,1,1) (b) Sarima(2,2,2)

Figure 4: Three-step ahead forecast

Looks like the house-prices, as pointed out by the ZHVI indicator, will rise a little bit during the next months. Our forecast looks always close to the observed data and a strong increase of the ZHVI is evident from the analysis that we have conducted.

References

Openml, 2022. URL https://www.openml.org/search?type=data&status=active&id=43631&sort=run.

Zillow.com, 2024. URL https://www.zillow.com/home-values/10221/austin-tx/.

Inside airbnb, 2024. URL http://insideairbnb.com/get-the-data.