MGT6203 - Data Analytics for Business

Group 59 Project Final Report

Using price determination to gauge regional rental property investment decisions.

Terence Belton Buttrick (tbuttrick3)

Salim Ben Ghorbel (sghorbel3)

Raphael Cavallari (rcavallari3)

Oleksii Rakov (orakov3)

Table of Contents

Introduction

Traditionally, a diversified portfolio will often contain a property component given its hedging ability against traditional assets like stocks and bonds [1]. This strengthens the case for the inclusion of property in an individual's investment portfolio. The argument for the inclusion of property in the form of traditional residential rental property has strengthened even more with the introduction of Airbnb. It has unlocked the potential to sweat investment properties more efficiently, giving investors the ability to earn constant returns through ease of access to customers provided through the modern technologies of the platform [2].

But how do you decide where to invest in a property with the intention to list it on AirBnB? Where should an investor look if they want to eke out a constant revenue stream as well as capital appreciation in their asset? This project attempts to address the investment decision by creating an investment model for determining the regions of a city to buy properties with the intention to rent them out on Airbnb to maximize Return on Investment (ROI). The comparative advantage envisioned in this model is to make use of the implied price discovery of Airbnb rental data inferring a property's value. At its most simple the model evaluates if the theoretical price derived from AirBnB rental prices is greater than the current listed price for a similar property in the area and, hence, is indicative of an investment opportunity that could yield excess returns for an investor be it from cash flows or capital appreciation.

In addition to identifying the undervalued and overvalued geographies, this project attempts to shed light on the factors that affect the expected attractiveness of a particular Airbnb by incorporating textual sentiment information from listing descriptions as well as broader macroeconomic conditions. Finally, the role of modelling choice is also examined by evaluating a set of statistical approaches ranging from linear regression with penalization to ensemble methods.

Methodology

Data Sources

The data sources for the model are:

Data	Description	Source
Airbnb - listing	Listing data including rooms, property type and occupancy rate inputs.	[3]
Airbnb - calendar	Each extract contains a year of rental pricing data.	[3]
Airbnb - fee	Fixed fee Airbnb charges clients - 3%.	[4]
Zillow	Property monthly data including price.	[5]
FRED	United states Macroeconomic data from FRED.	[6]
MSCI US REIT Index	Dividend Yield on US Equity REIT Index 2022-23 - 4.3%.	[7]

Calculated Fields

Annual Bookings [8]

The bookings field is an estimate of how many bookings are made on a property a year and is based on the number of reviews a property has a month. Empirical evidence has led to market usage of a rate of bookings yielding half the number of reviews. In other words:

$$Annual\ Bookings = (Reviews\ per\ month * 12)/0.5$$

However, the number of reviews must be greater than the minimum nights stay requirement set by the host, otherwise the minimum nights stay will be used. Additionally, for the calculation it is important to consider that Annual bookings must be capped to 365 as you cannot book a property out more days than there are in a year.

Annual Occupancy Rate [8]

The Occupancy rate was derived from the Annual Bookings:

$$Annual\ Occupancy\ Rate = (Annual\ Bookings/365)$$

There were considerations to the field that the Annual Availability of a property caps the Annual Bookings the property can have, and that occupancy is capped to 70% as per market convention based on empirical evidence.

Implied Price

For this analysis, the AirBnB theoretical price is derived using a discount cash flow model that is best described as similar to the Terminating Gordon Growth Model (GGM) which is usually used for company valuations based on dividends [9] but given the similar nature of cashflows in this case, it will be an effective approach. The denominator is a proxy for the necessary ROI rate for such an investment. The formula for the price looks is as follows:

$$Implied\ Price_t = \frac{(Rent_t)(Occupancy\ Rate_t)(1-T)(365)}{(Mortgage\ Rate_t)(AirBnB\ Fee_t)(REIT\ DY_t)}$$

Where:

- $Rent_t$ captures the cashflow income for the year.
- $Occupancy\ Rate_t$ captures the penalty of not having full occupancy on the property has on the income streams from the property.
- T is the effective tax rate for the period on the income stream (for simplicity, the mid tax bracket on a solo payer was selected)
- $Mortgage\ Rate_t$ captures the necessary return necessary to cover the cost of using debt to invest in a property.
- $AirBnB\ Fee_t$ is a flat rate applied to properties listed on the platform.
- $REIT\ DY_t$ is the current dividend yield on the US MSCI Equity REIT Index, and it captures the yield a property investor should be looking for on such investments net of operating costs.

Price Differential

$Price\ diff_t$ is the price differential of average prices for a region (AirBnB Price – Zillow Price) at time t. The rationale behind the model is that the price differential given a specific set of independent variables indicates whether the fair value price of the property is over- or under-valued in the market.

$$Price\ Differential_t = Implied\ Price_t - Zillow\ Price_t$$

Wrangling

Data from AirBnB was cleaned by transforming character data into the appropriate form. For example, character numbers and percentages into numbers and Boolean factors into 0 and 1. We had four data downloads of the Airbnb data (March 2022, June 2022, September 2022, and December 2022) each containing the expectations for the Airbnb rent for the coming year, these

expectations were averaged out over the year to form a panel of four observations. In this way the seasonality effect could be levelled out. Other static data were not implicitly amended.

Joining Zillow price indications required defining the geographic precision level, since Zillow has indices on state, city, and zip code level. It is worth noting that Airbnb data had neighborhood identifiers. However, after inspection, it was noticed that such data was not clean. For this reason, it was perceived that the zip code assignment to each property would be preferred. Additionally, latitude and longitude data were present for the Airbnb dataset, but we proceeded by matching the zip code using the location information. Making use of the python *uszipcode* library, which contains the centroids and limit coordinates for the area of each US zipcode, for each location we selected as zip code, among those which boundaries contained the selected location, the one with the closest centroid. With the newly added zip code information, the Airbnb data could then be easily joined with the Zillow data. Note that Zillow data was obtained per type of house and zip code. In this way a coherent price could be associated with the Airbnb information.

Finally, US Macro data was joined simply on date, as we could not obtain state-level macroeconomic indicators.

Feature Engineering

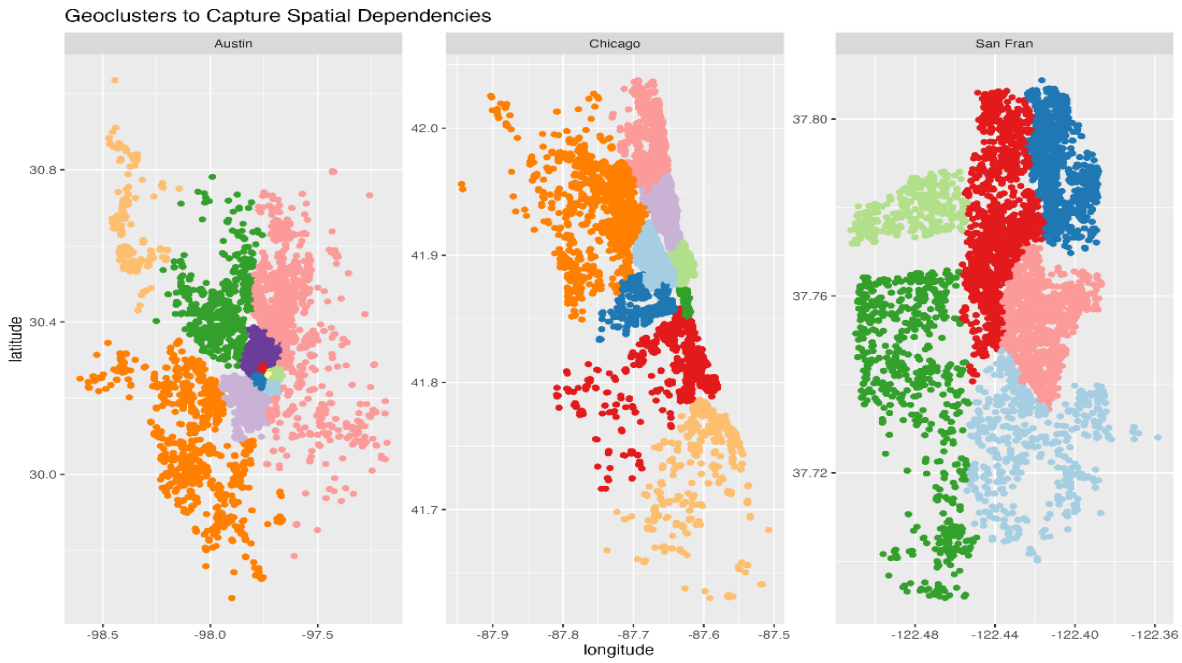
The Airbnb-Zillow dataset presented us with a set of challenges relating to incorporating textual and geospatial information as well as filling in missing data (imputation).

Processing Textual Data

Airbnb listings come with descriptions of the property and its surroundings filled out by hosts, as well as lists of provided amenities. We employed two techniques to translate this information into numeric format. First, length of all texts (excluding special characters and words with less than 3 symbols) were calculated. The underlying assumption is that more attractive Airbnb's come with richer/longer descriptions. Second, we used the "afinn" dictionary to estimate a simple length-normalized textual sentiment for property and neighborhood description. The resulting five fields (*description_sentiment, description_l, neighborhood_overview_sentiment, neighborhood_overview_l, number_of_amenities)* complemented the model.

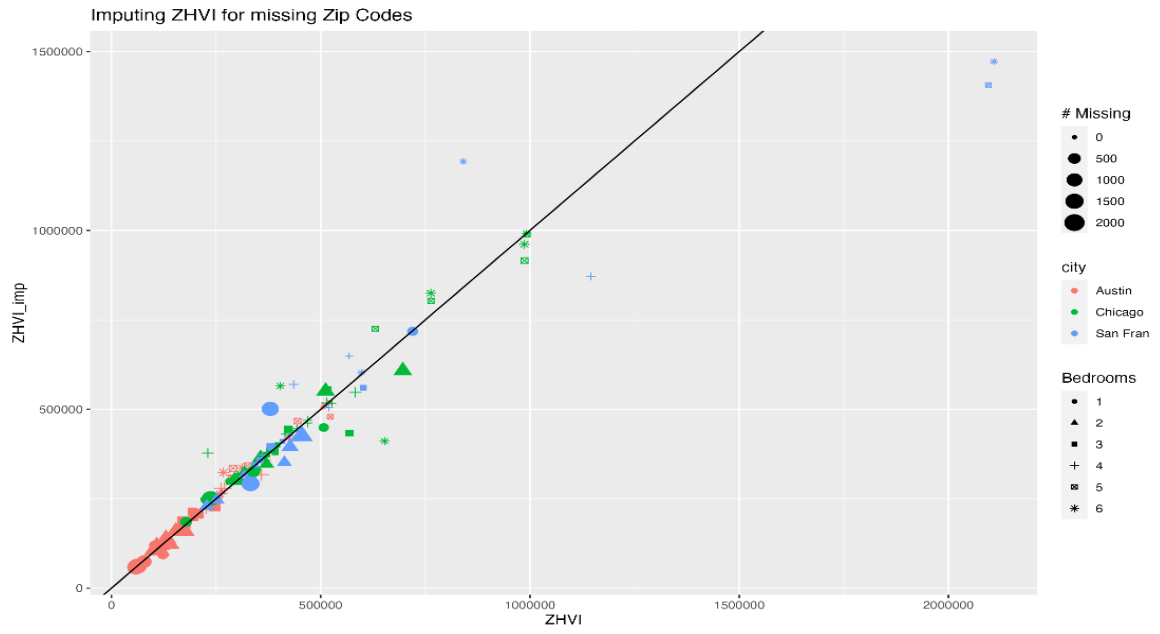
Processing Geospatial Data

Our dataset includes 3 cities and 197 unique zip codes. Including just the cities as predictors would probably underfit the spatial distributions of prices, whereas using all zip codes overfit instead and potentially lead to model estimation issues. A middle ground was needed. We solved the issue by applying spectral clustering algorithm, that provides density-based clusters. Resulting partitioning created 26 so-called geo-clusters, that are shown below.

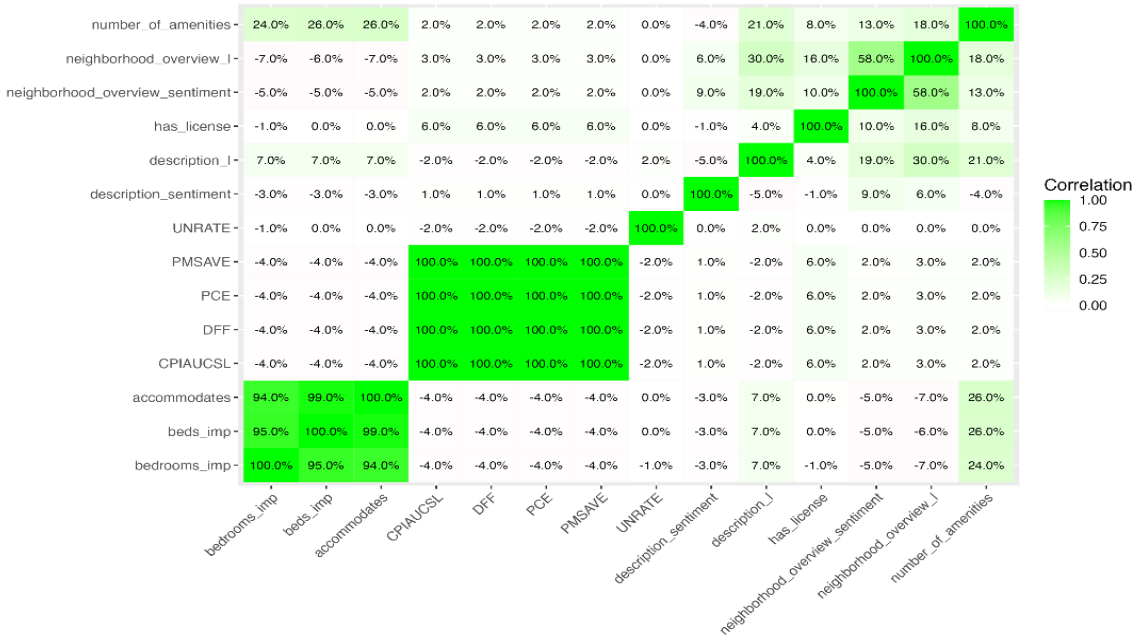


Imputation

The final challenge came from missing observations. While the Airbnb listings file was overall quite clean, there were a number of NAs in fields crucial to our modeling. First, some of the *beds* and *bedrooms* were not known. Luckily, those were straightforward to impute from each other as well as from the field *accommodates,* which encoded the maximum guest number. Overall, we used linear regression imputation with R-squared between 70% and 80%. Next, missing *reviews_per_month* were filled using the regression on available *reviews_all_time, reviews_30_days etc.* with similarly high fit quality. The biggest difficulty arose from the ZHVI (Zillow price) data, since only 117 of the zip codes had ZHVI values (the rest were either null or not listed on the ZHVI database). Linear regression would clearly not be an optimal imputation tool due to extrapolation and model misspecification risks. Instead *KNN* with k=10 was employed on scaled *longitude, latitude,* and *number of bedrooms.* The characteristics of imputed ZHVI closely match original data, apart from small clusters of large homes, as can be seen from the chart below.



Overall, 62 predictor variables entered the modelling stage. Some of them such as US Macro data and beds and bedrooms were highly correlated, as can be seen on the numeric predictor correlation matrix chart.



Modelling

Model & Feature Selection

Using Sci-kit Learn library on Python [10], we have employed a range of linear regression models as well as a Random Forest Regressor to find the best model to predict price differentiation. The dependent variable was Price Differential. For features we have employed 62 variables that can be split into the four categories:

- Property level data: *beds,* bedrooms, *accommodates, property_type, instant_bookable, has_licence,property_type.*
- Textual data: neighborhood_overview_sentiment, *description_sentiment, description_l, neighborhood_overview_l,number_of_amenities*.
- Geospatial data: cities and geoclus*ters.*
- Macroeconomic data: *CPIAUCSL, UNRATE, DFF, PCE, PMSAVE*.

All categorical variables were one-hot encoded, and all features were centered and scaled. Data were then split into 75% train and 25% test, while preserving temporal separation as is usual for panel data.

Next, we chose a representative set of predictive models and performed the standard grid search cross-validation to determine the best hyperparameters for each model. The models tested include Elastic Net, Linear Support Vector Machine (SVM), Ridge, Lasso, Bayesian Ridge, Stochastic Gradient Descent (SGD) Regressor, and Random Forest Regressor. We used the R-squared metric on test dataset to evaluate model performance.

The results of the model training and testing are as follows:

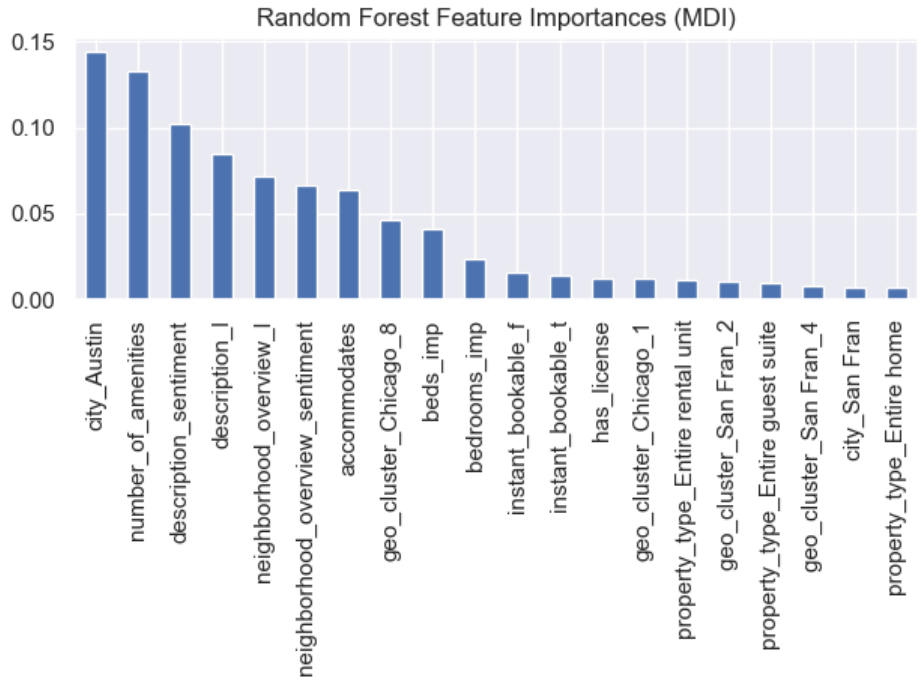
Model	Train R-squared	Test R-squared	Best Params
Elastic Net	0.1702	0.2205	Alpha = 1e-5 L1 ratio = 1
Linear SVM	0.0014	-0.0003	C = 0.001 Epsilon = 10
Ridge	0.1702	0.2419	Alpha = 0.1
Lasso	0.1702	0.2205	Alpha = 1e-5
Bayesian Ridge	0.17	0.2422	alpha_1 = 0.001 alpha_2 = 1e-06 lambda_1 = 1e-06 lambda_2 = 1e-06
SGD Regressor	-inf	-inf	alpha = 100.0 l1_ratio = 0.2 penalty = elasticnet
Random Forest Regressor	0.6815	0.5527	max_depth = 20 n_estimators = 200 min_samples_leaf = 5

The Random Forest Regressor demonstrated the best performance, with a test R-squared score of 0.55. As there is no regularization aspect in training RFR, we manually limited the hyperparameters in our search to avoid overfitting. The Bayesian Ridge model also performed well, with a test R-squared score of 0.24. The other linear models exhibited lower performance.

The improved performance of the Random Forest Regressor can be attributed to its ability to handle the large number of dummy variables and potential values in the dataset. Random Forests are particularly suited for handling high-dimensional data, as they can effectively model complex interactions between variables. Finally, Random Forest is not sensitive to multicollinearity in features, which is helpful for our dataset. Thus, the Random Forest Regressor appears to be the most suitable choice for predicting price differentials in this context.

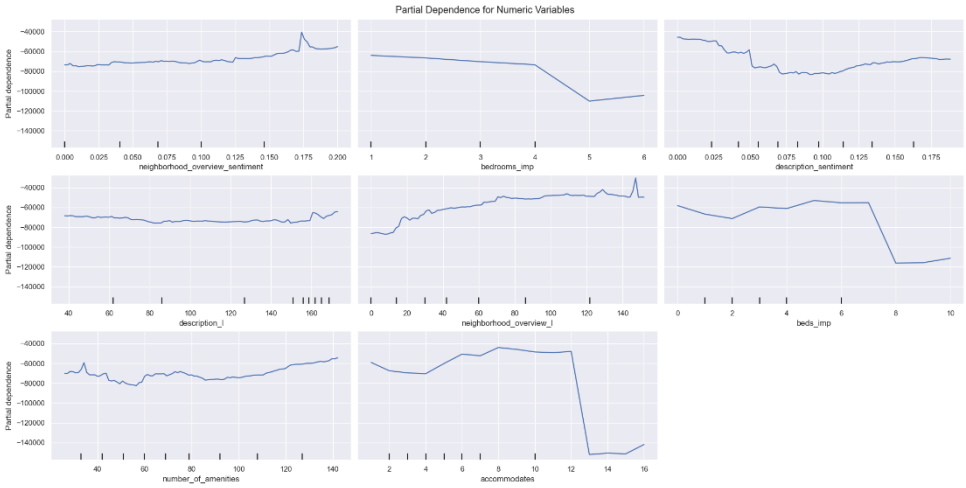
Final Model Analysis

Since Random Forest is a complicated ensemble model and not a classical regression, particular attention needs to be paid to interpret and validate model results. We start by reporting the Variable Importance using the Mean Decrease in Impurity (MDI), which is a standard measure for regression tree methods, shown below for top 20 predictors.

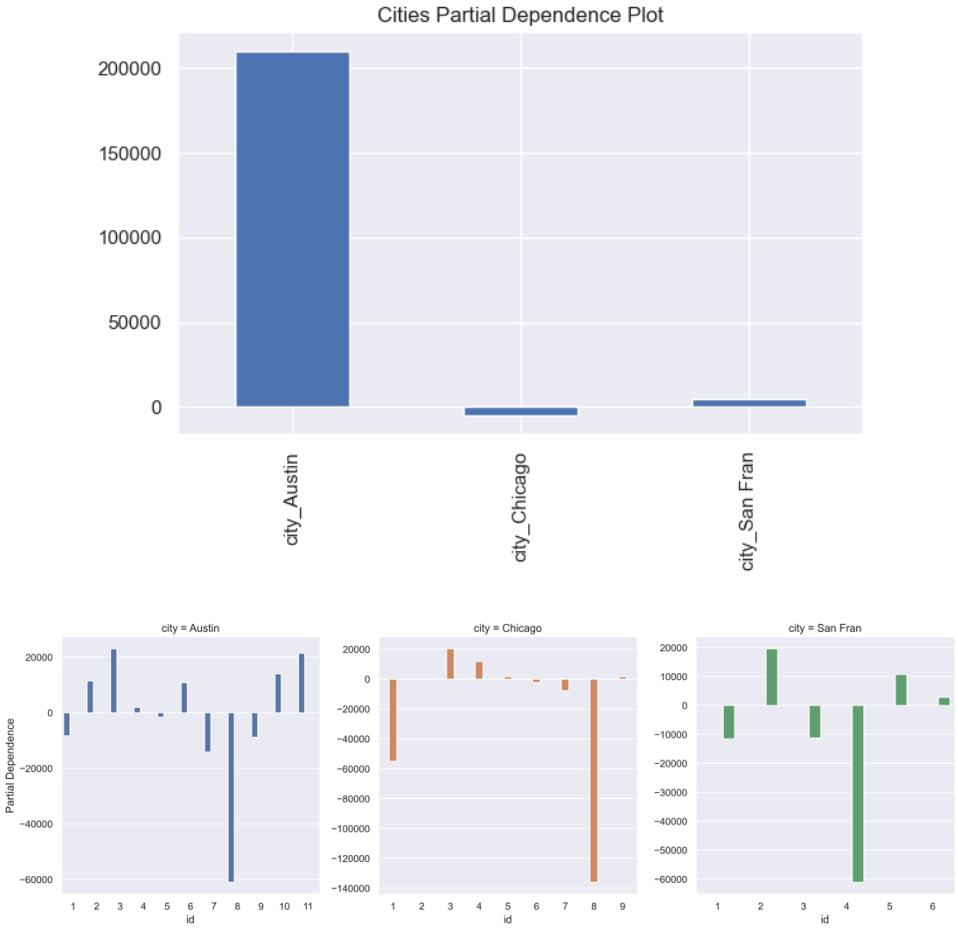


Our model incorporates both geospatial and textual information. Location in Austin as well as several individual geoclusters have high variable importance. Furthermore, all textual indicators contribute to substantial decrease in inpurity. Finally, as expected, hard data on a property such as number of guests, beds and bedrooms seems to matter. Overall, the model prediction mostly depends on geospatial, textual data, as well as property size.

Next, we analyse the impact of predictors on the price differentials through the lens of partial dependence plots. These aggregate prediction levels for the range of observed predictor values. In case of dummy variables, the interpretation is thus same as for the usual regression dummies, whereas for numeric predictors, response shape can be highly nonlinear. Graph below shows that larger Airbnb's (4 and above bedrooms, beds and many accommodates) are overall overvalued from buy-to-let perspective. On the other hand, more amenities and better neighbourhood sentiment lead to increased price differential. Finally, relationship between property descriptions and price differential is nonlinear and not straightforward to interpret.



Geospatial variables are one-hot encoded in our analysis, and thus can be interpreted as dummy variables (without level dropping). Charts of cities and geoclusters show that according to our model Austin is highly attractive to investors with mean differential of 200k USD. Otherwise, a few of the individual geoclusters exhibit significant differences in mean price differential.



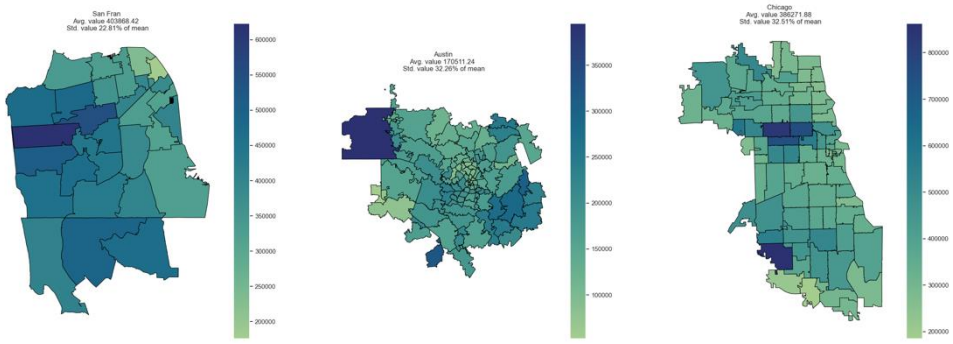
Finally, macroeconomic indicators play close to no role in our analysis. This is probably due to the limitations of our panel dataset spanning only 4 quarters, as well as lack of access to state or county level economic indicators. Improvements in this area would enrich our analysis.

In summary, our Random Forest regressor provides an interpretable and detailed model incorporating geospatial as well as property level information while delivering a high degree of cross validation and test fit.

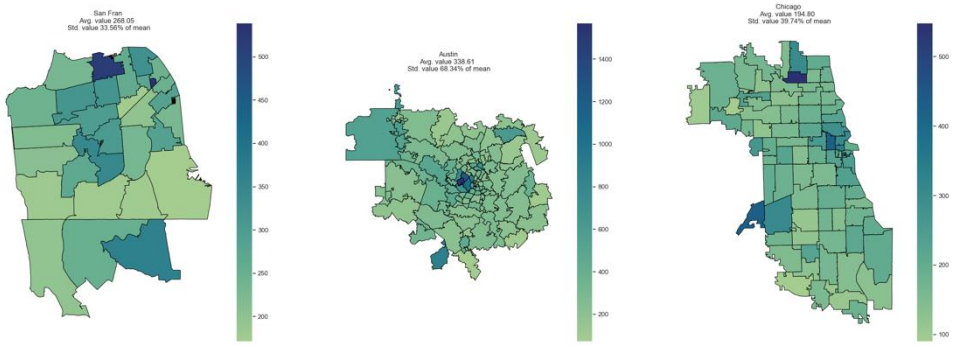
Insights

The goal of this project is to help inform investors in short term rental properties on attractive opportunities. Since real estate market is illiquid and heterogenic, mispricings can arise. This can be easily observed from contrasting the two main variables going into the return-on-investment calculation of such an asset: cash flow (proxied by Airbnb rental price) and investment cost (proxied by Zillow fair value estimate). We plot the data for our dataset on zip code level to observe little to no correlation between the two.

Average Zilow ZHVI property price per zip code by city



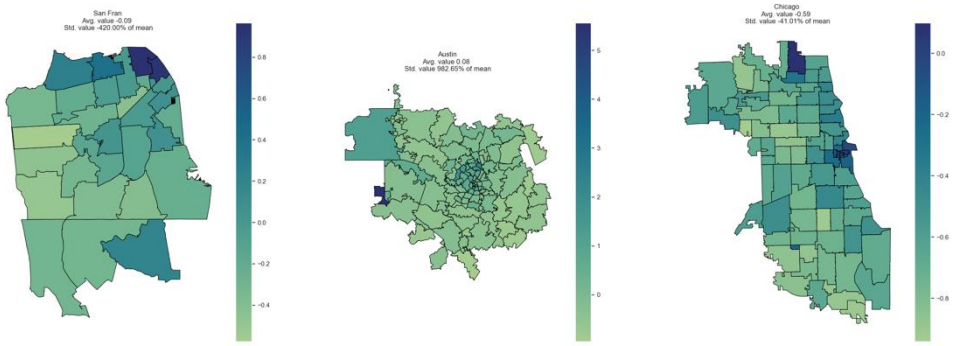
Average AirBnB listing price per zip code by city



Although this information alone is interesting, low listing price and ZHVI correlation might be due to other factors such as occupancy rates, property sizes and attractiveness.

That is where the added value of a modelling approach like our lies. By transforming occupancy and listing price into a fair value estimate and parametrising it through property and geospatial characteristics we can produce predicted pricing differentials (mispricings) on a zip code level.

Average Relative Price Differencial between valuation model & ZHVI price per zip code by city



Using the predictions of the model, investors can build geographically diversified buy-to-rent portfolios. For instance, our results indicate that downtown Austin, north-eastern San Francisco and eastern Chicago are under-priced compared to Airbnb income and thus represent attractive investment opportunities.

The estimated price differentials are substantial, up to an average of 200k USD for downtown Austin. We speculate, that this apparent arbitrage arises from the real estate market segmentation. Higher prices on Zillow correspond to residential areas, whereas busy Airbnb's are often located in downtown areas, that are less preferred for living. Additional analysis beside the scope of this projects would be needed to elaborate on this issue.

Potential future enhancement

The analysis in this project can be enhanced in several ways such as expanding the limited dataset, improving the fair price methodology, and exploring additional modelling choices.

First, a major limitation of our present research are the geographic and time series constraints. We focus on only 3 out of many US cities. Widening the cross section would both allow finding more mispriced areas but also improving the statistical power and predictive accuracy. Furthermore, due to the noticeably brief time period, we were not able to adequately incorporate dependencies on macroeconomic conditions, that undoubtedly influence the attractiveness of long-term investments such as rental properties not to mention trends, cycles and seasonality in variables.

Next, we adopted simplistic approaches to the costs of rental business such as utilities and internet, as well as complicated tax brackets and strategies in our analysis. Commercial Airbnb analytics providers rely on much more sophisticated assumptions regarding the cost basis. Such improvements could give a more nuanced price differential estimate.

Finally, choices of models in this analysis were driven in part by technical constraints. State of the art deep learning neural networks or proven ensemble techniques such as XGBoost were simply too computationally expensive for the scope of the project.

Overall, wider and longer panels with more precise fair value calculation fed into more modern machine learning techniques would likely produce a richer and more useful model.

Conclusion

In this project, we have created an investing model which helps us determine which cities are best for buying homes to advertise on Airbnb. We have shown how property attributes, location, and economic factors affect investment returns by utilizing the comparative advantage of Airbnb rental data. Insights regarding potential mispricing and investment opportunities in numerous places, including downtown Austin, the north-eastern part of San Francisco, and eastern Chicago, have been revealed by our random forest model. The model has limits because of its geographic focus, time series restrictions, and fair value estimation methodologies, but it nevertheless offers investors looking to build buy-to-rent portfolios with a variety of geographic exposures a useful tool. To increase the model's accuracy and relevance in assisting property investment decisions, future developments can include extending the dataset, improving the fair price estimation approach, and investigating alternative modelling methodologies.

Bibliography

[1] C. Tan, “The role of real estate in a diversified portfolio,” 07 01 2019. [Online]. Available : https://www.bankofsingapore.com/research/the-role-of-real-estate-in-a-diversified-portfolio.html.

[2] H. R. Koster, J. Van Ommeren and N. Volkhausen, “Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles,” *Journal of Urban Economics,* vol. 124, p. 103356, 2021.

[3] Inside Airbnb, “Get the data,” [Online]. Available: insideairbnb.com/get-the-data. [Accessed 16 Apr 2023].

[4] Airbnb, “Airbnb service fees - Airbnb help center,” [Online]. Available : https://www.airbnb.co.za/help/article/1857. [Accessed 16 Apr 2023].

[5] P. Mooney, “Zillow House Price Data,” Kaggle, 08 Dec 2020. [Online]. Available : https://www.kaggle.com/datasets/paultimothymooney/zillow-house-price-data. [Accessed 16 Apr 2023].

[6] ALFRED, “Archival fred,” [Online]. Available: https://alfred.stlouisfed.org/.. [Accessed 16 Apr 2023].

[7] MSCI, “MSCI US REIT index,” [Online]. Available : https://www.msci.com/documents/10199/08f87379-0d69-442a-b26d-46f749bb459b. [Accessed 16 Apr 2023].

[8] Inside Airbnb, “Data assumptions,” Inside Airbnb, [Online]. Available : http://insideairbnb.com/data-assumptions. [Accessed 16 Apr 2023].

[9] A. Damodaran, “I. the stable growth DDM: Gordon Growth Model - New York University.,” [Online]. Available: https://pages.stern.nyu.edu/~adamodar/pdfiles/ddm.pdf. [Accessed 16 Apr 2023].

[10] Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *JMLR,* vol. 12, pp. 2825-2830, 2011.