Real Estate Insights: Predictive Analytics for Airbnb and Zillow Listings

https://github.com/AradhyaAlva/Data_Mining-Airbnb-vs-Zillow-Data-Prediction

Aradhya Alva Rathnakar
Bhavan Kumar Basavaraju
Mahamaya Panda
Shashi Kumar Kadari Mallikarjuna

Department of Applied Data Science, San Jose State University
Data 240: Data Mining
Professor: Shayan Shams
Date: May 16, 2024

Motivation

In today's rapidly changing financial landscape, securing financial stability as individuals approach retirement is crucial. Real estate investment presents a compelling opportunity to augment traditional retirement plans such as 401(k) and IRA. By investing in real estate, individuals can gain immediate rental income and benefit from long-term property appreciation, which serves as a hedge against inflation. However, the challenge lies in identifying the best markets and properties, and in formulating effective pricing strategies. These tasks require thorough research and a deep understanding of market dynamics, making the process daunting for many potential investors.

Background

Real estate investment has gained popularity as a strategy for supplemental retirement planning. Platforms like Airbnb allow property owners to tap into short-term rental markets, while the potential for long-term capital gains remains an attractive aspect of real estate investment. Nevertheless, navigating the intricate housing market dynamics and making informed investment decisions necessitates specialized knowledge and tools. This project aims to leverage extensive public listing data from sources like Airbnb and Zillow, applying advanced machine learning and deep learning techniques to develop predictive analytics models. These models will provide investors—both individuals and real estate professionals—with refined insights into optimal markets, property attributes, and pricing strategies tailored for maximizing returns from both short-term rentals and long-term property investments. The goal is to enhance real estate-centered financial planning and improve retirement readiness, offering investors increased control over their financial futures.

Literature Review

Garlapati et al. (2021) explores Airbnb pricing in New York City using various machine learning models to predict listing prices effectively. It employs models like Linear Regression, Naive Bayes, KNN, Decision Tree, and XGBoost, among others, to analyze diverse data features over eight years. The best-performing model was the Decision Tree, which achieved an accuracy of 0.95 and an AUC score of approximately 0.95, demonstrating robust prediction capabilities. Our approach builds on this by using more advanced techniques and a broader dataset, further enhancing predictive performance and robustness.

Peng et al. (2020) explores Airbnb price prediction using advanced techniques like XGBoost, DNN, NLP, PCA, and Spectral Clustering with data from Inside Airbnb for ten cities. Despite being limited to one data source, their integration of multi-modal data achieved optimal performance with XGBoost. Our method builds on this by adding more predictive models and extending the dataset to include broader geographic and temporal variables, enhancing overall prediction accuracy and depth.

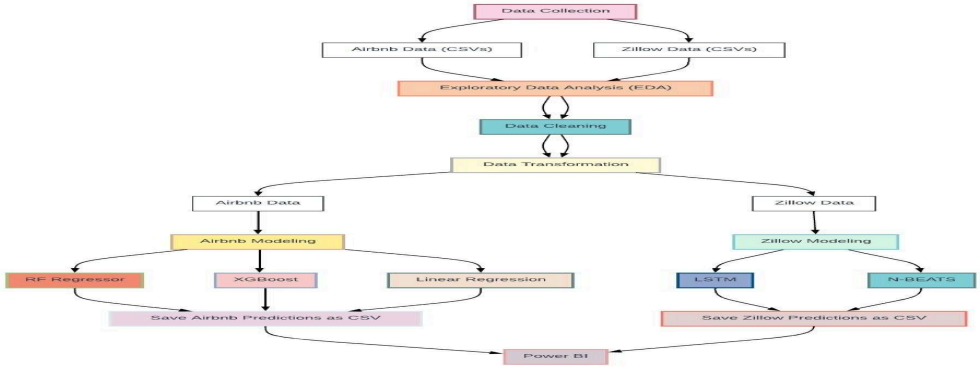
Gupta et al. (2022) explores various Time Series Prediction models like DeepAR, LSTM, NBEATS, and others across sectors such as business and weather forecasting. It evaluates their performance on metrics including MAE, MSE, RMSE, and R^2. Notably, NBEATS excels with an MAE of 1.509, MSE of 0.403, and RMSE of 0.8275, demonstrating superior accuracy and robustness in complex forecasting scenarios. Our approach enhances this by integrating hybrid models, such as those used by Zillow and Airbnb, that combine the advantages of each method across various scenarios, thereby improving accuracy and efficiency.

Yang (2024) applies machine learning to predict Airbnb prices, highlighting XGBoost as the most effective model with an MSE of 3969.53, RMSE of 62.27, and R2 of 0.506. This underscores the benefits of ensemble learning in the sharing economy. Our method expands on this by using broader data inputs and advanced techniques for enhanced predictive accuracy and model adaptability for both Zillow and Airbnb datasets.

Methodology

The project begins with data collection, where Airbnb and Zillow datasets are gathered as CSV files. This is followed by exploratory data analysis (EDA) to understand the datasets' characteristics. The process continues with data cleaning to address issues like missing values and data inconsistencies, after which data transformation is performed to ensure data is in the correct format for modeling. Separate modeling paths are followed for Airbnb and Zillow data. Airbnb data is modeled using random forest regressors, XGBoost, and linear regression, while Zillow data is analyzed using LSTM and N-BEATS models. The outputs from these models are saved as CSV files, which are then utilized in Power BI to create visualizations and further insights to analyze the best city to invest in to get the maximum return on investment.

Figure 1: *Project Methodology*



Data Collection

The dataset for the project is sourced from Inside Airbnb and Zillow Research. Inside Airbnb provides listing-level data for Airbnb properties across various cities, including details like amenities, reviews, and booking history. Zillow Research offers housing market metrics and property sales data at different geographic levels. These datasets are separately preprocessed to clean, transform, and merge the data, and the preprocessed, combined datasets are available for further analysis and model development.

Data Preprocessing

The project involves meticulous data preprocessing steps tailored to enhance the quality and reliability of the analyses. For Airbnb and Zillow datasets, the preprocessing includes handling missing data using mean and mode imputation for numerical data, and although there are currently no categorical null values in these datasets, KNN imputation is integrated to enhance future robustness should new data contain such nulls. Outliers are detected and handled using the Interquartile Range (IQR) method, replacing extreme values with NaN and subsequently filling these with the median of respective columns. Further transformations include dropping redundant columns, adding new columns, and correcting data types. Numerical features are standardized, and categorical features are one-hot encoded to prepare the data for effective model training. After cleaning and preprocessing the Airbnb and Zillow datasets, the datasets are separately prepared and then combined using concatenation techniques in a CSV file, ensuring a cohesive and unified dataset for comprehensive analysis and model development.

Airbnb Price Prediction

Model 1: Linear Regression

The Airbnb price prediction involves three models. The first model, linear regression, predicts Airbnb prices by incorporating both categorical and numerical features. The dataset is split into training (70%), validation (15%), and test (15%) sets, with preprocessing involving scaling and one-hot encoding. An initial linear regression model is evaluated, showing performance metrics such as MAE and R^2. A grid search is then conducted to find the best

hyperparameters for various models, including Ridge and Lasso, with adjustments to regularization strength (alpha). The best parameters found are Ridge with alpha = 10.0. This optimized model is used to make predictions on the test and validation sets, with updated MAE and R^2 metrics to assess improvements.

Model 2: Random Forest Regressor

The second model, the Random Forest Regressor, also predicts Airbnb prices and includes preprocessing, training, and hyperparameter tuning. Categorical features are encoded with OneHotEncoder, and numerical features are scaled using StandardScaler in a ColumnTransformer. The model is trained (70%) and evaluated on the test (15%) and validation (15%) sets. Hyperparameter tuning with GridSearchCV optimizes parameters such as n_estimators, max_features, max_depth, min_samples_split, and min_samples_leaf. The best parameters found are {'n_estimators': 750, 'max_features': 'log2', 'max_depth': None, 'min_samples_split': 2, 'min_samples_leaf': 1}. The optimized model's performance is further evaluated, and predictions are saved to a CSV file.

Model 3: XGBoost

The third model, XGBoost, predicts Airbnb prices using numerical features that are scaled and categorical features that are one-hot encoded. The dataset is split into train(70%), validation (15%), and test (15%) sets for modeling. The XGBRegressor model is fitted, and predictions are evaluated using MAE and R^2 on test and validation sets. Hyperparameter tuning with GridSearchCV optimizes parameters such as n_estimators, max_depth, learning_rate, subsample, and colsample_bytree. The best parameters found are colsample_bytree=0.7, learning_rate=0.01, max_depth=3, n_estimators=1500, and subsample=1.0. The final model, trained with these parameters, is evaluated, and predictions are saved to a CSV file for further analysis. GridSearchCV is chosen for hyperparameter tuning the Airbnb data because it systematically explores all parameter combinations, providing a thorough optimization over the search space, whereas RandomSearchCV samples a subset randomly, which might miss the optimal settings.

Zillow Home Value Prediction

Model 1: LSTM

The first model, LSTM, is utilized for Zillow home value time-series-based prediction and includes preprocessing steps to scale numerical features with StandardScaler and encode categorical features using OneHotEncoder. The bidirectional LSTM architecture has three LSTM layers with units set to 100 for the first two layers and 50 for the third, all with L2 kernel regularization (0.01), dropout (0.2), and batch normalization. The model's output layer is a dense layer with one unit. Compiled with the RMSprop optimizer and mean squared error (MSE) as the loss function, the model is trained for 100 epochs. Hyperparameter tuning via Keras Tuner optimizes the number of units (30 to 500) and learning rate (0.01, 0.001, 0.0001) using the Adam optimizer. The dataset is split into train(70%), validation (15%), and test (15%) sets for modeling. The tuning process identifies the best hyperparameters, to get the best model performance. Evaluation using MAE and R^2 metrics confirms the model's effectiveness.

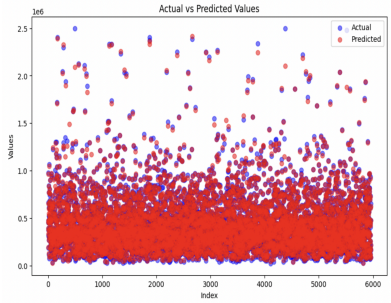
Model 2: N-BEATS

For Zillow home value prediction, the N-BEATS model is used with preprocessing steps including StandardScaler for numerical features and OneHotEncoder for categorical ones. The model architecture features two hidden layers with 256-unit Dense layers and ReLU activation functions, followed by an output layer. Data is divided into training (70%), validation (15%), and test sets (15%), with dynamic preprocessing to compute input shape and feature transformation count, all combined into a pipeline with the N-BEATS model. Hyperparameter tuning uses a HyperModel subclass, adjusting parameters like hidden layer units (128 to 512) and learning rates (1e-2, 1e-3, 1e-4) through Keras Tuner's RandomSearch over 25 trials for 100 epochs, optimizing with the Adam optimizer. The best model is evaluated on test and validation sets, updating MAE and R^2 metrics to improve forecasting performance.

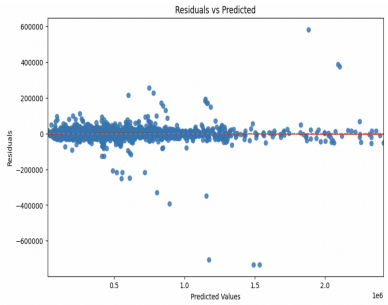
Experiments and Results

In our experiments and results, we have performed a comprehensive analysis of the predictive model's performance, through a scatter plot, residual plot, and a histogram to compare actual and predicted values and to assess residuals. Initially, it visualizes the actual vs. predicted values, indicating the model's performance at each index. It then checks and reshapes the arrays if necessary before calculating residuals, which highlight the differences between actual and predicted values. The residuals are further analyzed with a residual plot to examine the relationship between predicted values and residuals, and a histogram to inspect their distribution. In this context, Mean Absolute Error (MAE) is preferred over Mean Squared Error (MSE) as a metric because it provides a straightforward average of the absolute errors, making it more interpretable and less sensitive to outliers. MAE gives an easy-to-understand measure of average prediction error in the same units as the target variable, which is particularly useful for practical applications where clear and direct insights into prediction accuracy are essential.

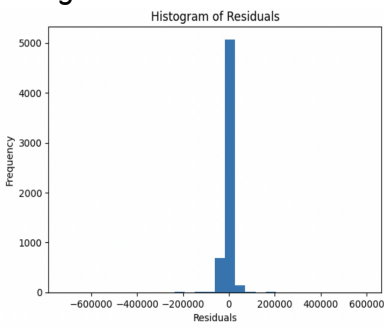
Scatterplot: Comparison of Actual and Predicted Values



Residual Plot: Residuals vs. Predicted Values



Histogram of Residuals



The performance of the models in predicting Airbnb data is evaluated using MAE ($\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\text{Actual}_i - \text{Predicted}_i|$) and R-squared ($R^2 = 1 - \frac{\sum_{i=1}^{n}(\text{Actual}_i - \text{Predicted}_i)^2}{\sum_{i=1}^{n}(\text{Actual}_i - \overline{\text{Actual}})^2}$) metrics, revealing that the Random Forest Regressor stands out as the best performer after hyper-tuning. Before hyper-tuning, the Random Forest Regressor has an MAE of 102.965 on the test set, which improves to 101.609, and its R^2 value increases significantly from 0.055 to 0.333, indicating a better fit to the data. The Linear Regression model also shows improvements in both MAE and R^2 after hyper-tuning, though the gains are modest. Interestingly, the XGBoost model's performance deteriorates post-hyper-tuning, with increased MAE values and only marginal improvements in R^2. Thus, hyper-tuning has the most positive impact on the Random Forest Regressor, enhancing its predictive accuracy and model fit.

Figure 2: *Airbnb: Model Comparison with Evaluation Metrics*

AIRBNB			
Model	**Metric**	**Before Hypertuning**	**After Hypertuning**
Linear Regression	MAE (Test)	148.277	146.636
Linear Regression	MAE (Valid)	152.625	146.636
Linear Regression	R2 (Test)	0.121	0.129
Linear Regression	R2 (Valid)	0.09	0.107
Random Forest Regressor	MAE (Test)	102.965	101.609
Random Forest Regressor	MAE (Valid)	108.843	106.459
Random Forest Regressor	R2 (Test)	0.055	0.333
Random Forest Regressor	R2 (Valid)	0.012	0.186
XGBoost	MAE (Test)	114.708	126.667
XGBoost	MAE (Valid)	118.032	132.166
XGBoost	R2 (Test)	0.105	0.223
XGBoost	R2 (Valid)	0.025	0.151

The performance of the models in predicting Zillow data is evaluated using MAE and R-squared metrics, revealing that the N-BEATS model performs the best after hyper-tuning.

Initially, the N-BEATS model has a high MAE of 165570.15 on the test set, which improves to 24378.49 after hypertuning. Its R^2 value also increases significantly from 0.452 to 0.978, indicating a much better fit to the data. In comparison, the LSTM model shows only slight changes after hypertuning, with MAE values remaining relatively stable and R^2 values showing modest improvement. Thus, hyper-tuning has a substantial positive impact on the N-BEATS model, significantly enhancing its predictive accuracy and model fit.

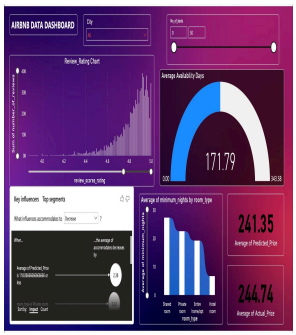
Figure 3: *Zillow: Model Comparison with Evaluation Metrics*

ZILLOW			
Model	**Metric**	**Before Hypertuning**	**After Hypertuning**
N-BEATS	MAE (Test)	165570.15	24378.49
N-BEATS	MAE (Valid)	163999.041	24148.658
N-BEATS	R2 (Test)	0.452	0.978
N-BEATS	R2 (Valid)	0.414	0.966
LSTM	MAE (Test)	437771.81	435140.62
LSTM	MAE (Valid)	430513.728	430513.728
LSTM	R2 (Test)	0.274	0.247
LSTM	R2 (Valid)	0.245	0.372

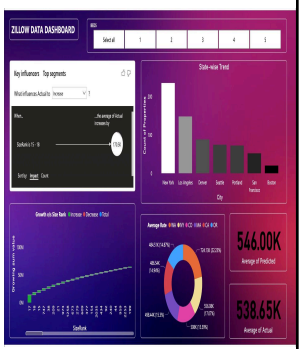
An In-Depth Comparative Study: Analyzing and Predicting Trends

The Power BI analysis is conducted after the model comparison to offer comprehensive insights and actionable data, enabling comparative analysis and trend predictions across various real estate and hospitality metrics for Airbnb and Zillow. The saved CSV files for Zillow and Airbnb are separately analyzed in the dashboard and then compared against each other to provide a thorough comparative analysis. The Airbnb and Zillow data are correlated in terms of analyzing the top market to get the best return over investment in real estate in terms of making the least investment to get the maximum returns from it. For Airbnb, increasing beds and review ratings correlate with fewer listings, with review scores significantly affecting availability, which jumps from 171 days for a 4.0 rating to over 343 days for a 5.0 rating. Hotel rooms on Airbnb have the shortest minimum stay requirement at 6.68 nights, whereas shared rooms require the longest at 27.29 nights. Zillow's data reveals that as Market Position increases from 15 to 18, property sales prices surge by over $170,000, particularly in markets like New York, Los Angeles, and Denver. Comparative analysis indicates Boston has the highest average costs and property values on both platforms, while Portland ranks lowest. The Prediction Dashboard suggests the highest property value in Los Angeles could reach $2.13M, with Boston at the lower end around $0.6M, and New York showing the highest average Airbnb costs. The dashboards can be seen below.

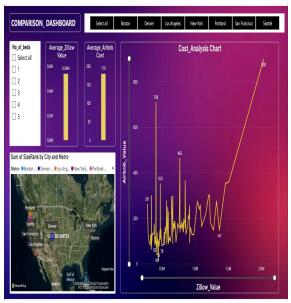
Airbnb Dashboard



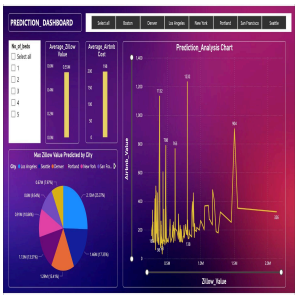
Zillow Dashboard



Comparison Dashboard



Prediction Dashboard



Discussion

In this project, various models are implemented to address regression problems for predicting Airbnb rental prices and Zillow home values, utilizing advanced machine learning techniques after thorough data preprocessing. For Airbnb, the Random Forest Regressor emerged as the best model, achieving a Mean Absolute Error (MAE) of 101.609 and an R2 score of 0.333 after hyperparameter tuning, indicating its effectiveness in capturing the variability in rental prices. Similarly, for Zillow home value predictions, the N-Beats model excelled with an R2 score of 0.978 and a reduced MAE of 24,378.49, showcasing its superior predictive performance in real estate value estimation. These models significantly enhanced the project's ability to deliver accurate and reliable real estate price predictions, underlining the potential of machine learning in real estate analytics.

Future Improvements

The future scope of the project includes further enrichment of predictive models by integrating additional data sources such as crime rates, school district information, and transportation statistics. This integration is aimed at enhancing the contextual understanding of various factors that influence daily life and decision-making. Building on this enriched data framework, a personalized investment recommendation system will be developed, catering to individual preferences and risk tolerance levels, thereby refining the accuracy and relevance of the financial advice provided. Also, the focus will be on developing scalable deployment strategies for these predictive models and associated dashboards in a production environment. This approach will ensure the robustness, reliability, efficiency, and effectiveness of the models in delivering real-time insights and recommendations across various platforms.

References

About. (n.d.). https://insideairbnb.com/about/

Garlapati, A., Garlapati, K., Malisetty, N., Krishna, D. R., & Narayana, G. (2021). Price listing predictions and forthcoming analysis of Airbnb. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt51525.2021.9579773

Gupta, K., Tayal, D. K., & Jain, A. (2022). An experimental analysis of state-of-the-art Time Series Prediction models. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. https://doi.org/10.1109/icacite53722.2022.9823455

Housing data - Zillow Research. (2024, April 22). Zillow. https://www.zillow.com/research/data/

Lektorov, A., Abdelfattah, E., & Joshi, S. (2023). Airbnb Rental Price Prediction Using Machine Learning Models. *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*. https://doi.org/10.1109/ccwc57344.2023.10099266

Peng, N., Li, K., & Qin, Y. (2020). Leveraging Multi-Modality Data to Airbnb price prediction. *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*. https://doi.org/10.1109/icemme51517.2020.00215

Sinthong, P., & Carey, M. J. (2021). Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data. *2021 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata52589.2021.9671603

Yang, Y. (2024). Predicting US Airbnb listing prices by machine learning models. *Highlights in Business, Economics and Management*, *24*, 1408–1417. https://doi.org/10.54097/m187nw17