CSE343 : Machine Learning Project Interim Report

"Forecasting Hospitality Costs: A Data-Driven Journey into Hotel Room Price Prediction"

Aryesh Shakya 2021238 | Aakash Agarwal 2021222 | Lakshya Kumar 2021536 | Lakshya Agrawal 2021535

March 4, 2024

Abstract

In today's dynamic hospitality industry, the accurate prediction of hotel prices has become paramount for the convenience of both travelers and hotel management. This project seeks to leverage the power of machine learning to create a robust model for predicting hotel prices. The driving force behind this initiative lies in the real-world challenges faced by travelers, who often struggle to estimate their accommodation expenses, and by hoteliers, who aim to optimize their revenue streams, so they can do there business in best possible way and to earn and grow as much as possible.

1. Introduction

Understanding property prices is essential for gauging an area's development. In the hotel industry, setting reasonable prices is vital; if they're too high, people may choose other options, and if too low, the hotel may face losses. It's also beneficial for customers to estimate a property's value in money terms.

In our project, we aimed to build a model predicting property prices or log(prices) using regression methods like Linear Regression, Lasso, Ridge, SVM, Decision tree, Random forests, XGBRegressor and MLP. We applied PCA where ever there is a need to reduce data complexity. Our project's main goal is predicting property prices based on their features.

We combined statistical analysis and data modeling to create a reliable tool for property buyers, sellers, and the real estate industry. This tool enhances the property market's transparency and efficiency, aiding decision-making and industry growth. By merging advanced statistical techniques with real estate knowledge, we strive to create a meaningful predictive model.

2. Literature Survey

2.1. Survey on Airbnb Price Prediction using Machine Learning and Sentiment Analysis

Property price prediction has been a topic of interest in both real estate and hospitality industries. Traditional methods often rely on expert knowledge and historical data to estimate property prices.

The article covers how several machine learning approaches were used to create a pricing prediction model for Airbnb rental units. Researchers have explored various machine learning techniques, feature selection methods, and data pre-processing steps to improve prediction accuracy. They have used Linear Regression, Support-Vector Regression (SVR), Random Forest Regression and Neural Networks. They also used feature selection techniques to reduce the dimensionality of input data, such as Lasso feature selection and p-value analysis.

Additionally, the researchers emphasize the significance of customer reviews in influencing the pricing of Airbnb listings. To enhance the accuracy of their predictive model, they employed sentiment analysis using the TextBlob library. This approach acknowledges the influence of guest feedback on pricing decisions and leverages it as a valuable source of information for the predictive model.

2.2. Real Estate Price Prediction with Regression and Classification.

Real estate has been an important sector almost everywhere. Both buyers and sellers need to know the estimated values of the property prices they are interested in, as it will enable them to make informed decisions.

The article discusses the prediction of house prices using machine learning models with the given data of residential houses. Yu et al. employed various formatting techniques for their data to incorporate the categorical variables. This led to an increase in the number of variables, prompting

them to use the PCA technique to counter this dimensionality problem. They divided the price values into 7 buckets and applied classification models such as Logistic Regression, SVC (linear kernel), SVC (Gaussian kernel), Random Forest Classification, and used the Naive Bayesian model as a baseline. Similarly, for regression to predict the exact value, they kept Naive Bayes as baseline and utilized regression models like Linear Regression, Lasso, Ridge Regression, SVR (Linear kernel), SVR (Gaussian kernel), and Random Forest Regression. All the models were applied both before and after applying PCA to reduce dimensionality.

Their best performing model for the classification problem is SVC with a linear kernel, and for the regression problem, it is SVR with a Gaussian kernel. They also concluded that the living area, roof material, and neighborhood hold the greatest statistical significance in predicting the property's price.

2.3. Warehouse Rental Price estimation using Machine Learning Techniques

The rental warehouse market is undergoing significant change due to growing logistics industry and digital transformation. Both supply and demand sides are facing market uncertainty leading the participants to think how to competitively price warehouses in the open market.

The research involves predicting warehouse rental price estimation using machine learning techniques. The data set is extracted from classified advertisement websites to understand and guide pricing in this dynamic landscape. Yixuan et al. have explored various machine learning techniques which include Linear Regression,Regression Tree, Random Forest Regression, Gradient Boosting Regression Trees to improve prediction accuracy. The models were trained after necessary feature extraction and preprocessing and its performance evaluated by using score metrics:RMSE and r.

The findings suggest that Random Forest model, outperforms single-factor models in predicting warehouse rental prices.They also concluded that Location, distance from the city center, and local land prices were critical factors in the price prediction, though estimation errors remained notable.

3. Dataset

The main goal of this work is to use machine learning techniques to predict hotel room rental costs. To look into this, we utilise Airbnb data from Kaggle as a case study. Here, we describe the data and its essential features for predicting hotel pricing.

3.1. Dataset Description

The data set included 74,112 entries, each with 29 features. The feature include such as accommodates, number of beds, type of property, etc. Also min, max and mean of all the features are highly varied therefore scaling is required before training the models. Figure 1 and Figure 2 shows the geographic distribution of the listing prices in this dataset (see Figure 1 and Figure 2).

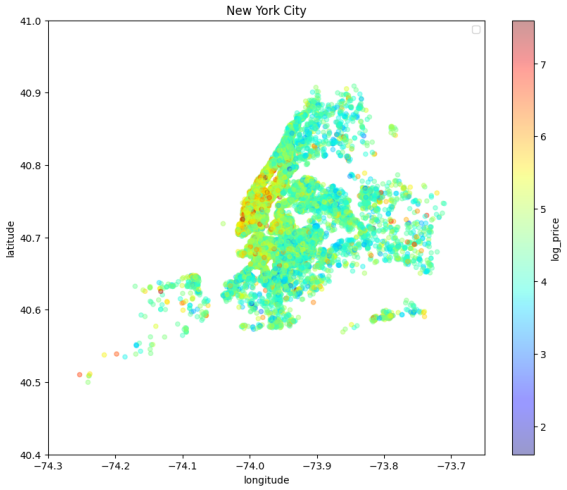

Figure 1. NYC listings price

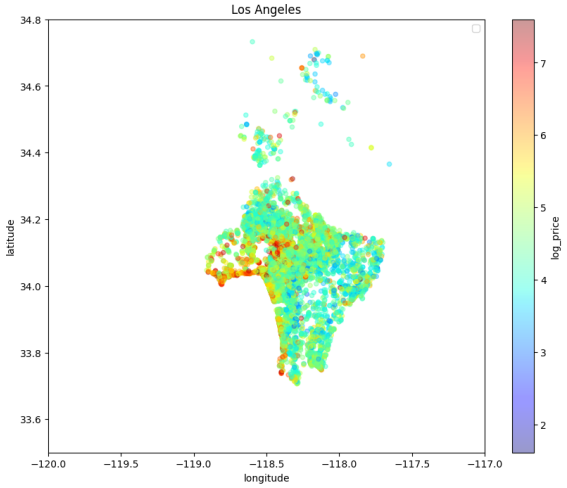

Figure 2. LA listings price

3.2. Dataset Preprocessing

It is crucial to use data preprocessing techniques to manage missing values, outliers, and maintain consistent data types since the data comprises both numerical and categorical information. Firstly, we dropped all the features that were deemed unnecessary for our analysis. Removal of these columns ensures that the dataset contains only the most pertinent information for our research.

To guarantee data consistency and integrity, duplicate records were first eliminated. To keep the data completeness, missing values in columns like 'host_response_rate', 'beds', 'bedrooms', 'bathrooms' and

'review_score_ratings', were imputed using the appropriate methods. Outliers were found and handled properly to stop them from influencing further analysis. We observed that the prices below 3 and above 6.5 were classified as outliers by box plot. Thus, records including these outliers were dropped. The percentage sign was removed from percentage columns ('host_response_rate') to convert them to a numerical format. A new feature, 'Amenities,' is derived from an existing feature to enable its use in the model.

Label encoding was used to improve the dataset's quality for categorical columns including 'host_has_profile_pic,' 'host_identity_verified,' 'instant_bookable,' and 'cleaning_fee' The incorporation into machine learning models is made simple by the conversion to numeric data.

Additionally, categorical columns like 'city,' 'bed_type,' 'property_type,' 'room_type,' and 'cancellation_policy,' used one-hot encoding. To provide interoperability with different machine learning algorithms, this approach generates binary columns for each category within these characteristics.

Overall, the dataset was cleaned, improved, and transformed using various data pretreatment approaches, preparing it for exploratory data analysis and further study.

3.3. Exploratory Data Analysis

To begin, The prices are distributed in a normal (Gaussian) manner, which is considered one of the best distributions for the model to learn effectively. (see Figure 3).

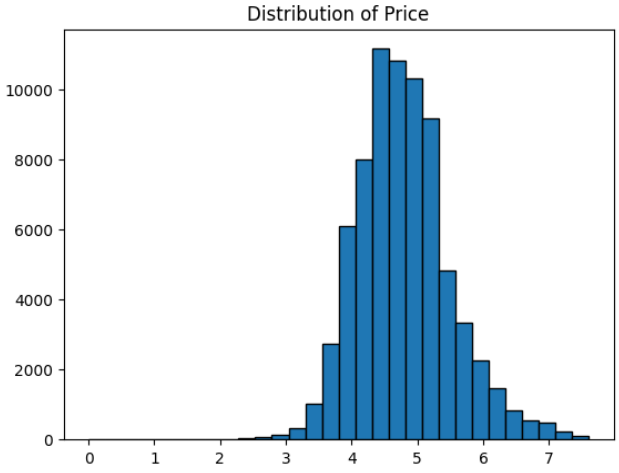

Figure 3. Top 5 property types

To visualize the relation between the type of room available and price, we plotted the graph between the average hotel room price and room type (see Figure 4). The shared room is the least expensive type of accommodation, while the entire home/apartment is the most costly property. This can be observed in the graph depicting the average price for each category.

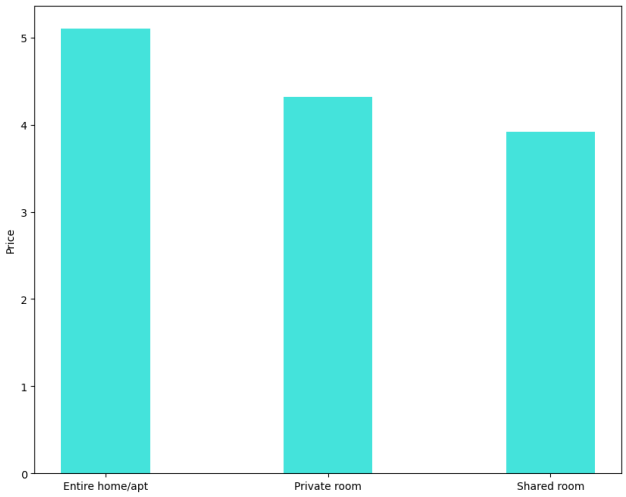

Figure 4. Average price of each room type

A bar chart presents the average log_price for each city, helping to compare the pricing trends across different locations.

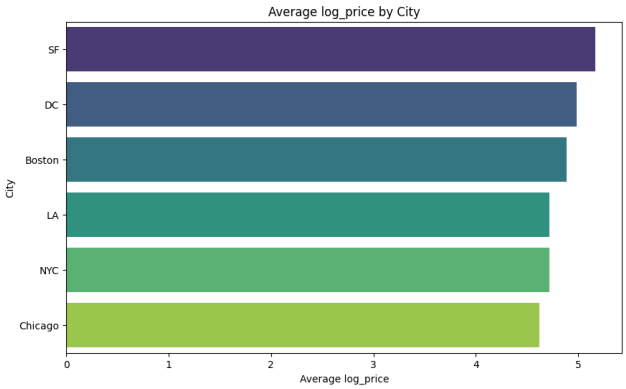

Figure 5. Average log_price for each city

Then we made geographic distribution plots for different cities, starting with an overview of all listings and then focusing on specific cities like New York City and Los Angeles. These scatter plots use latitude and longitude to map the listings' locations, with color-coding indicating the logarithm of the price. These visualizations offer insights into the spatial distribution of property prices in various cities (See Figure 1 and Figure 2).

The average price or log_price is plotted for each city (see Figure 5), It can be seen that the prices depends on the location, and SF(San Francisco) has the costliest property while Chicago has the cheapest property.

The correlation heatmap illustrates the relationship between accommodations and their average prices. The plot reveals a logarithmic correlation between price and accommodates, indicating how the number of occupants influences pricing. This insight provides valuable information

for understanding the dynamics of pricing based on accommodation capacity. (see Figure 6).

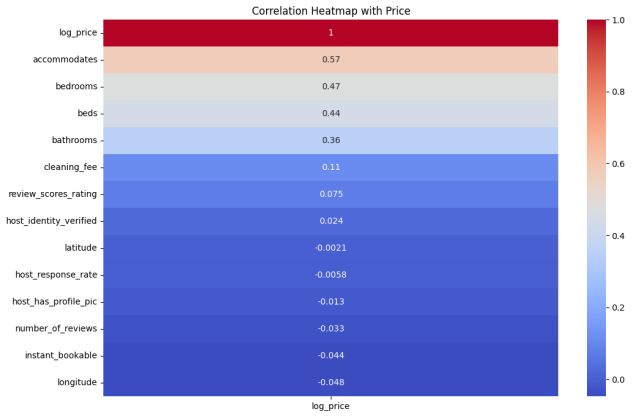

Figure 6. Correlation Heatmap

A correlation heatmap was also created to investigate the connections between different attributes and the 'log_price' of postings. This heatmap assisted in locating possible correlations that could affect the price choices made by Airbnb hosts (see Figure 7). It can be seen that price is mainly dependent on accommodates, bedrooms, beds, bathrooms and cleaning fees as the features.

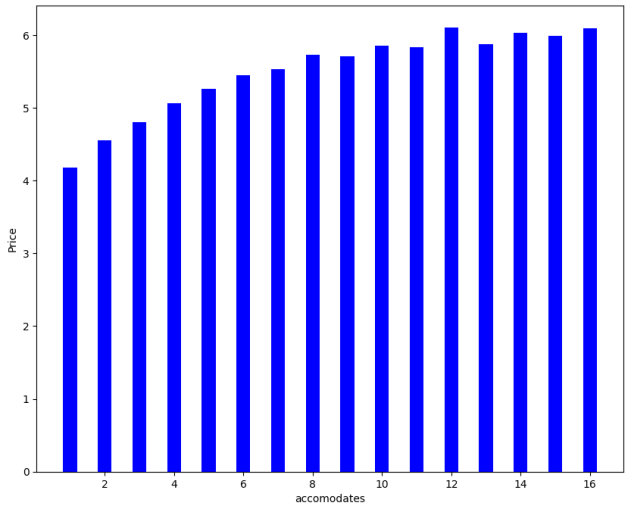

Figure 7. Avg price vs accomadates

4. Methodology

- In our analysis, we aimed to predict the log prices of rental properties using different machine learning regression techniques while incorporating feature engineering, feature scaling, and dimensionality reduction techniques.
- As part of feature engineering, data was clustered using KMeans. The cluster labels obtained were added to the dataset as additional features as these labels would capture the underlying patterns present in the data that are useful for regression.
- Initially, we applied StandardScaler to standardize numerical features before applying any type of encoding to improve the performance of certain algorithms. We label encoded and one-hot encoded categorical features in the dataframe
- We then applied a basic linear regression model. The data was divided into training and testing sets and it's performance was evaluated using Mean Square Error (MSE), Root Mean Square Error (RMSE), R2 Score, and Mean Absolute Error (MAE)
- We then applied regularized regression algorithms including Lasso and Ridge regression with the same preprocessing steps and further evaluated the performance of these regularized regression models using the same evaluation metrics.
- To prepare the data for Support Vector Machine (SVM) regression, we applied PCA to reduce dimensionality of the dataset. We then trained an SVM regression model with kernels including linear and rbf, and evaluated its performance on the testing data.
- We then applied Tree-Based and Ensemble methods. A Decision Tree Regressor was trained on the dataset to predict the log prices of rental properties. To optimize the model's performance, hyperparameter tuning was performed using Grid Search on parameters like the maximum depth of the tree, minimum samples required to split an internal node, minimum samples required at a leaf node, and maximum features considered for the best split. The performance was then evaluated on the model having the best parameters(higher predictive power).
- Random Forest Regressor was also employed to predict the log prices. The model uses an ensemble of decision trees, each trained on a random subset of features. Subsequently, predictions were made on the test dataset to evaluate the model's performance.
- XGBRegressor (Extreme Gradient Boosting Regressor), a gradient boosting algorithm, was also employed for the regression problem. It is an ensemble method that combines multiple decision trees sequentially to create a strong predictive model. Grid search was applied on parameters: max depth, subsample count and learning rate to find the best predictive model. Evaluation was done using the same metrics as used in previous models.

- We also applied an MLP Regressor, a neural network based model, to predict rental property prices. Through Grid Search, we explored various hyperparameters like hidden layer sizes, activation functions, and solvers, aiming to minimize mean squared error (MSE).

5. Results and analysis

Our analysis yielded the following results:

1. **Linear Models**:
   Linear Regression, Lasso, Ridge: These traditional linear models were tested for predicting Airbnb prices. Among these:
   Linear Regression: Served as the baseline model, displaying moderate predictive ability with an R^2 score of 0.56 and an MSE of 0.44.
   Lasso Regression: Presented slightly lower performance, achieving an R^2 score of 0.54 and an MSE of 0.46.
   Ridge Regression: Demonstrated competitive results, attaining an R^2 score of 0.56 and an MSE of 0.44.

2. **Support Vector Machine (SVM)**:
   Showed promising performance with an R^2 score of 0.58 and an MSE of 0.42. Notably, it outperformed Linear, Ridge, and Lasso Regression models, showcasing a different approach to prediction.

3. **Tree Based and Ensemble Methods**:
   Decision Tree Regressor: Showed reasonable performance with an R^2 score of 0.63 and an MSE of 0.37.
   Random Forest Regressor: Displayed improved performance, achieving an R^2 score of 0.69 and an MSE of 0.31.
   XGBRegressor: Demonstrated the best performance among these models, attaining an R^2 score of 0.70 and an MSE of 0.30.

4. **Multi Layer Perceptron (MLP)**:
   MLP displayed moderate performance with an R^2 score of 0.57 and an MSE of 0.42, indicating reasonable predictive ability but slightly lower than ensemble methods.

While traditional linear models showed moderate predictive abilities, they slightly trailed behind ensemble methods and SVM. Particularly, Support Vector Machine (SVM) displayed distinct performance, outperforming linear models with an R^2 score of 0.58 and an MSE of 0.42, highlighting its unique predictive power. Ensemble methods such as Random Forest and XGBRegressor demonstrated superior performance, showcasing their capability in capturing intricate data patterns. Remarkably, XGBRegressor emerged as the top-performing model, boasting an impressive R^2 score of 0.70 and the lowest MSE of 0.30, signifying its robust predictive accuracy.

	LR	Lasso	Ridge	SVM
MSE	0.44	0.46	0.44	0.42
RMSE	0.66	0.67	0.66	0.65
R2	0.56	0.54	0.56	0.58
MAE	0.51	0.52	0.51	0.50

	DT	RF	XGBoost	MLP
MSE	0.37	0.31	0.30	0.42
RMSE	0.46	0.41	0.41	0.65
R2	0.63	0.69	0.70	0.57
MAE	0.61	0.56	0.55	0.50

6. Conclusion

Traditional linear models like Linear Regression, Lasso, and Ridge performed moderately, with Lasso slightly lagging behind. Surprisingly, Support Vector Machine (SVM) outperformed these linear models. Ensemble methods like Decision Trees and Random Forests showed better performance, while XGBRegressor stood out with the highest accuracy. Multi-Layer Perceptron (MLP) demonstrated moderate predictive ability. Overall, ensemble methods, notably XGBRegressor, displayed superior accuracy in predicting Airbnb prices, indicating their effectiveness in capturing intricate data patterns and offering more precise forecasts compared to linear models and MLP.

References

1. Rezazadeh Pouya, et al (2019). Airbnb Price Prediction Using Machine Learning. ResearchGate.

2. Yixuan Ma,et al(2018). Estimating Warehouse Rental Price using Machine Learning Techniques. ResearchGate.

3. Yu Hujia, et al (2016). Real Estate Price Prediction with Regression and Classification.

4. DATASET