Semester Long Project CS 487/519
What do we know about Airbnb?
Stage 5

Brian B. Marquez[1], Avery Lee[2], Calicia Perea[3]

[1]Department of Computer Science (Data Analytics), New Mexico State University-Graduate
[2]Department of Computer Science, New Mexico State University-Graduate
[3]Department of Computer Science, New Mexico State University-Undergraduate

Abstract

This report delves into the use of machine learning techniques to forecast Airbnb listing prices across the US as well as in two major US cities: New York City and Chicago. To achieve this objective, three datasets from Kaggle were utilized, all of which contained comparable features. The study evaluated several initial regressors, including Linear Regression, Support Vector Regressor, ElasticNet, Random Forest, and RANSAC. Based on its superior performance, the Random Forest Regressor was chosen for price prediction. The report also implemented data scaling techniques to enhance model efficiency and employed evaluation metrics such as Mean Squared Error (MSE) and R-squared (R2) to assess model performance. The effectiveness of the Random Forest Regressor was validated through this process. While Principal Component Analysis (PCA) was tested, it did not yield better results compared to the original model. In conclusion, this study highlights the potential of machine learning techniques in predicting Airbnb prices, particularly with the Random Forest Regressor. Future research may explore additional techniques or further feature engineering to enhance model performance.

1 Motivation

Travel and tourism have always been an important part of the world economy. Recently, this accounted for 6.1% of the global gross domestic product (Statista, 2023). This is a nice recovery from the 2019 pandemic, where travel and tourism took a huge hit. While travel and tourism have not yet returned to pre-pandemic numbers, the growth from last year has shown encouragement that it will meet and even exceed previous numbers (PricewaterhouseCoopers, n.d.). With regulations returning to normal, travelers have taken the opportunity to resume travel at a significant rate. While the hotel industry is still prominent, Airbnb has emerged as a legitimate competitor in many large markets (Weber, 2022). The role of supplemental lodging is no longer accurate as demand and revenue for Airbnb Inc. has increased each quarter following the pandemic. Currently, Airbnb is valued at over $76.73 billion USD and is trading at $121.76 USD (Airbnb Worth - Google Zoeken, n.d.). This business is a large part of travel and tourism globally which drives interest on how the company performs from consumers and business interests. We will dive into a subset of the global market and focus our attention on Airbnb data from just the United States. We will further subset to two large cities within the U.S.: New York City and Chicago.

2 Data Sets

These Airbnb data sets can be found on Kaggle. They are comparable since they originate from the Airbnb website.

NYC Airbnb Open Data

https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

This data set represents the listing activity and metrics for Airbnb data in New York City, New York. It contains information about hosts, geographical availability, and the listings themselves.

Instances: 48,896

Features: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365 (16 in total)

Usability Rating: 10.00

Last Updated: 2019

Expected Update Frequency: Annually

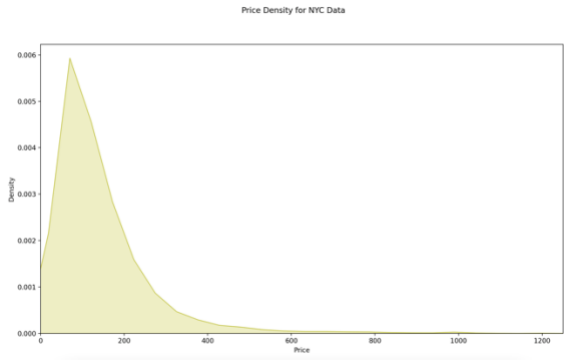

Figure 1: A price density graph to show the distribution of our target variable in the NYC dataset.

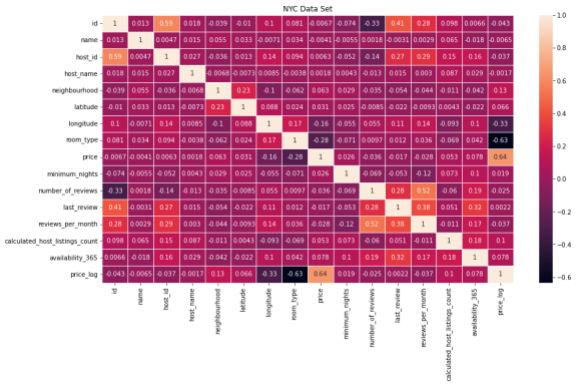

Figure 2: A heat map to show correlation between the features of the NYC dataset.

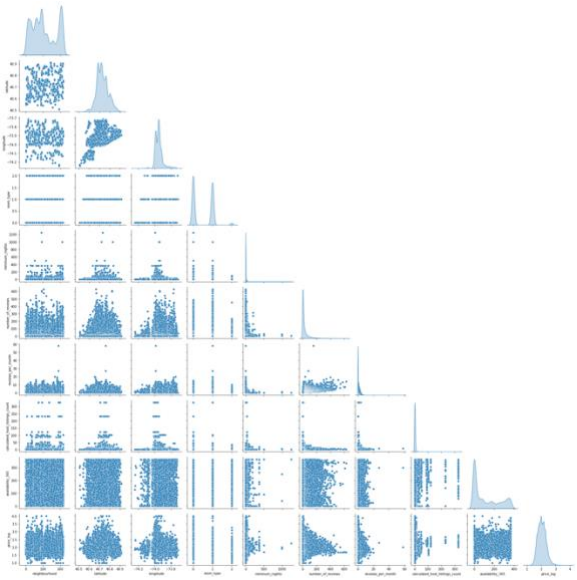

Figure 3: A pair plot to describe the relationship between the features of the NYC dataset.

Average Price: $152.72

Minimum Price: $0

Maximum Price: $10000

Median Price: $106.0

Chicago Airbnb Open Data

https://www.kaggle.com/datasets/jinbonnie/chicago-airbnb-open-data?select=listings.csv

This dataset gives information about the hosts, locations, and homes offered by Airbnb in Chicago, Illinois.

Instances: 6,397

Features: id, name, host_id, host_name, neighborhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365 (16 in total)

Usability Rating: 10.00

Last Updated: 2021

Expected Update Frequency: Never

Figure 4: A price density graph to show the distribution of our target variable in the Chicago dataset.

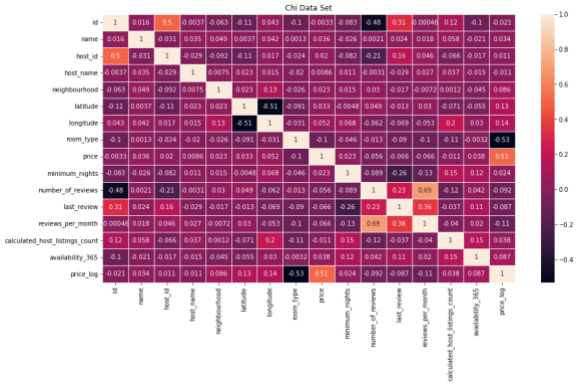

Figure 5: A heat map to show correlation between the features of the Chicago dataset.

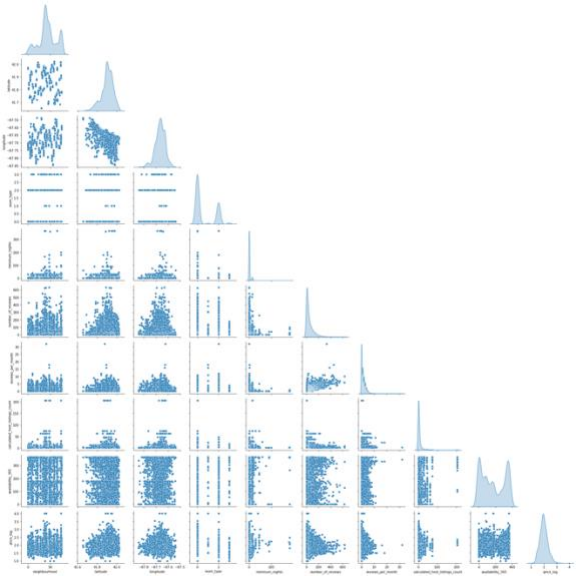

Figure 6: A pair plot to describe the relationship between the features of the Chicago dataset.

Average Price: $153.02

Minimum Price: $0

Maximum Price: $10000

Median Price: $99.0

U.S. Airbnb Open Data

https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data

This data set contains metrics and activity for Airbnb listings in 28 unique locations within the United States, including New York City, Los Angeles, and Hawaii. Its columns describe the hosts, locations, and listings.

Instances: 226,006

Features: id, name, host_id, host_name, neighbourhood_group, neighbourhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365, city (17 in total)

Usability Rating: 10.00

Last Updated: 2021

Expected Update Frequency: Annually

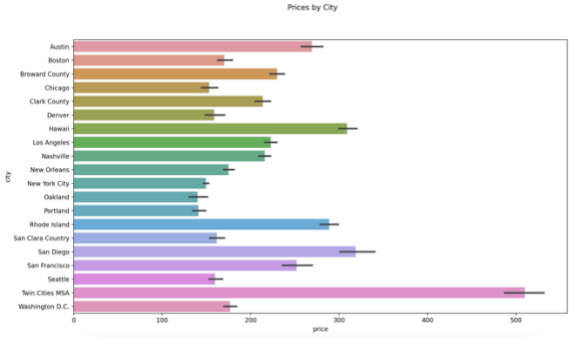

Figure 7: A bar chart detailing the average Airbnb price of the 20 most popular cities in our US dataset.

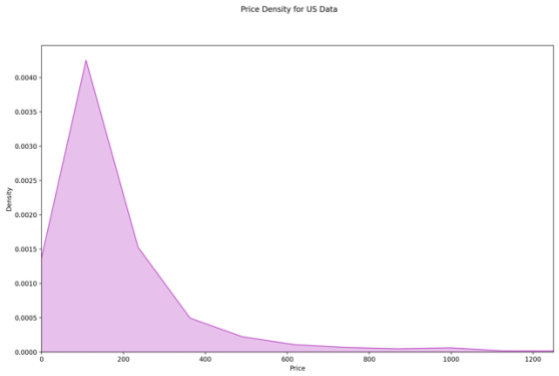

Figure 8: A price density graph to show the distribution of our target variable in the US dataset.

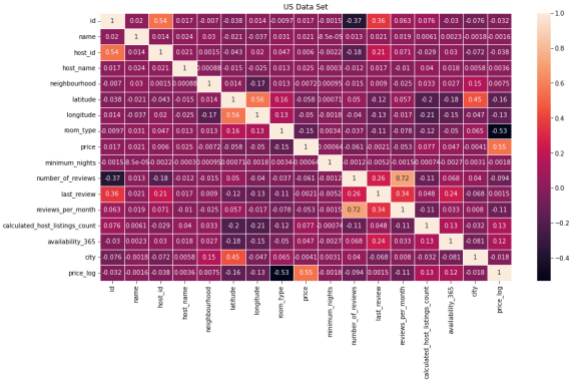

Figure 9: A heat map to show correlation between the features of the US dataset.

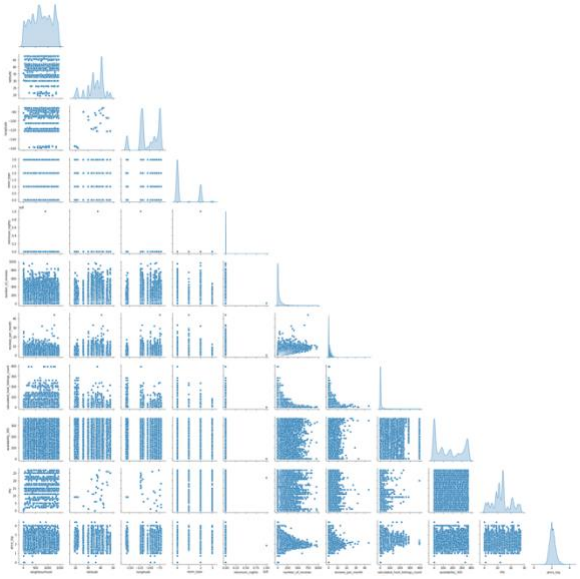

Figure 10: A pair plot to describe the relationship between the features of the US dataset.

Average Price: $219.71

Minimum Price: $0

Maximum Price: $24999

Median Price: $121.0

3 Data Set Justification

All three datasets should be compatible for our use case because they contain the same features (aside from the additional 'city' feature in the U.S. data). Between the three sets, we have 281,299 instances, satisfying the 10K instance requirement outlined in the project description. These datasets also obtain a usability rating of 10.0 on the Kaggle website; this means they score well on completeness, reliability, and compatibility. Thus, we feel confident we can run several models successfully.

4 Related Work

Here is some related work on price prediction.

https://www.kaggle.com/code/swapnilnarwade/airbnb-sales-prediction-rmse-482-126

https://www.kaggle.com/code/ellage/housing-price-analysis-in-nyc

5 Problem Definition

Our team utilized these datasets to predict the price of each house in different regions. We explored several machine learning options to achieve this goal, then narrowed in on those techniques that proved successful to produce our final result.

6 Tasks

We started this project by performing data visualization to create graphs and explore our data. Particularly, we delved into the price feature and observed the correlation between other features. This helped us gain an understanding of the datasets before we moved on to our price prediction.

Next, we performed cleaning on the data. We dropped the neighbourhood_group feature to make the sets more consistent with each other and transformed the price column to be the log of the price column to aid with our predictions. We also dropped rows containing null values. Finally, we utilized sklearn's StandardScaler library to conduct data scaling to improve the efficiency of our models.

We implemented five different initial regressors for this project: Linear Regression, Support Vector Regressor (SVR), ElasticNet, Random Forest, and RANSAC. We used 70% of the original US data for our training, then tested on the remaining 30% as well as the whole original NYC and Chicago datasets. Our results are detailed in the experiments section. We were able to compare the performance of these five models to determine the direction to take for the remainder of the project.

Our team implemented several evaluation metrics to judge the performance of each model. For the normal US data, on each model, we reported fitting time, training R2 score and mean squared error (MSE), testing R2 score and MSE, and created a residual plot. Because we initially used the NYC and Chicago datasets for testing only, we only reported the testing R2 score and MSE. Finally, we graphed a comparison of each model's fitting time and R2 score. See figure 16 for this comparison.

We also ran an experiment to test each of these initial models on US data that had undergone Principal Component Analysis, detailed in the second part of our experiments section. However, these results were no better than the original; in fact, in many cases, this extra step actually lowered our R2 scores. As a result, we decided to omit PCA in our final model. For the US data that had undergone PCA, we reported the fitting time, training R2 score and MSE, testing R2 score and MSE, and the associated residual plot. Please reference figure 22 for a time and performance comparison plot on the models using PCA data.

Due to its high performance in the initial model testing phase, we decided to utilize the Random Forest Regressor to move forward with our price prediction. We also changed our strategy to define a new Random Tree for each dataset, rather than just using the NYC and Chicago for testing. We increased our iterations from 1,000 to 10,000 and

evaluated each tree's performance on its respective dataset.

To evaluate this final set of models, we reported fitting time, training R2 score and MSE, testing R2 score and MSE, and created a residual plot for each dataset. These results are detailed in our solution section below.

7 Experiments

The first part of this section shows our results when running the five initial regressors on our data. The second part of this section explains each model on our data after it had undergone PCA.

Normal Results:

Linear Regression

- US Data:
  - Fitting Time: 0.072 seconds
  - Training R2: 0.01208
  - Training MSE: 188086.48
  - Testing R2: 0.01424
  - Testing MSE: 171377.92

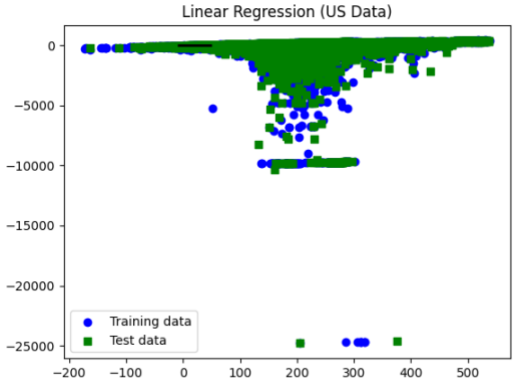

Figure 11: A residual plot to display the results of the Linear Regression model on our US dataset.

- NYC Data:
  - Testing R2: -0.00205
  - Testing MSE: 38885.56
- CH Data:
  - Testing R2: 0.00119
  - Testing MSE: 134698.10

The results indicate that the model did not perform well in predicting the prices of Airbnb listings in the US, NYC, and CH datasets. The testing R2 score ranges from -0.00205 to 0.01424—a very low score.

SVR

- US Data:
  - Fitting Time: 640 seconds
  - Training R2: -0.02429
  - Training MSE: 195013.02
  - Testing R2: -0.02669
  - Testing MSE: 178493.70

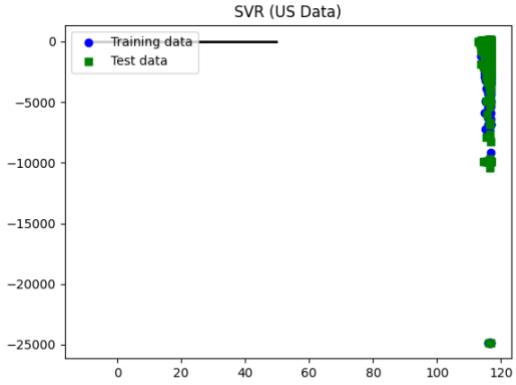

Figure 12: A residual plot to display the results of the SVR model on our US dataset.

- NYC Data:
  - Testing R2: -0.01785
  - Testing MSE: 39498.89
- CH Data:
  - Testing R2: -0.00620
  - Testing MSE: 135694.51

The negative R2 scores indicate that the model did not fit the data well, and the high MSE values suggest that the model's predictions were far from the actual prices of the listings. These scores indicate that the SVR did not perform well in predicting the prices of Airbnb listings in any of the datasets.

ElasticNet

- US Data:
  - Fitting Time: 2.5 seconds
  - Training R2: 0.02793
  - Training MSE: 185069.26
  - Testing R2: 0.03294
  - Testing MSE: 168126.14

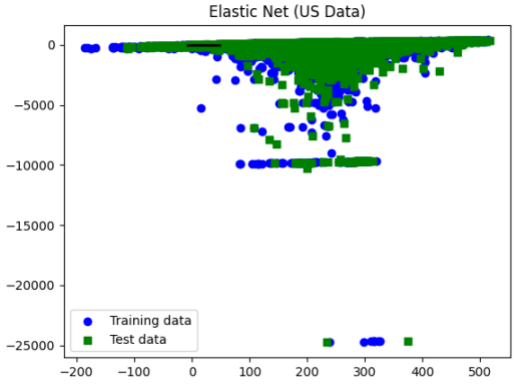

Figure 13: A residual plot to display the results of the Elastic Net model on our US dataset.

- NYC Data:
  - Testing R2: 0.04188
  - Testing MSE: 37180.77
- CH Data:
  - Testing R2: 0.01481
  - Testing MSE: 132859.91

ElasticNet is one of the initial regressors evaluated in this project. The model was evaluated on the US, NYC, and CH datasets. These findings suggest that the ElasticNet may be a suitable model for predicting Airbnb listing prices, especially in the US dataset.

Random Forest

- US Data:
  - Fitting Time: 1,200 seconds
  - Training R2: 0.93830
  - Training MSE: 11746.91
  - Testing R2: 0.47562
  - Testing MSE: 91164.91

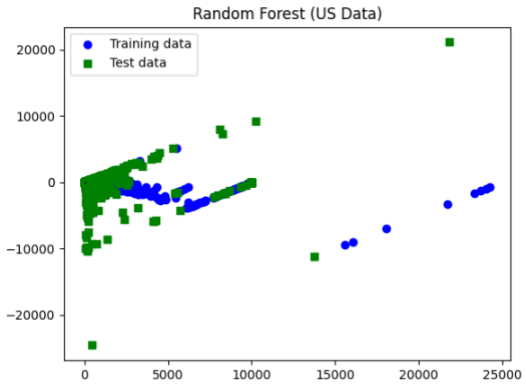

Figure 14: A residual plot to display the results of the Random Forest model on our US dataset.

- NYC Data:
  - Testing R2: -1.52757
  - Testing MSE: 98084.94
- CH Data:
  - Testing R2: -0.42414
  - Testing MSE: 192056.46

The Random Forest Regressor outperformed the other models in predicting Airbnb listing prices for both New York City and Chicago. It had a relatively low error rate. Our results suggest that the Random Forest Regressor may be a suitable model for predicting Airbnb listing prices in some cases but may not perform well in all datasets.

RANSAC

- US Data:
  - Fitting Time: 1.5 seconds
  - Training R2: -161412.95
  - Training MSE: 30731172475.88
  - Testing R2: -0.03618
  - Testing MSE: 180143.78

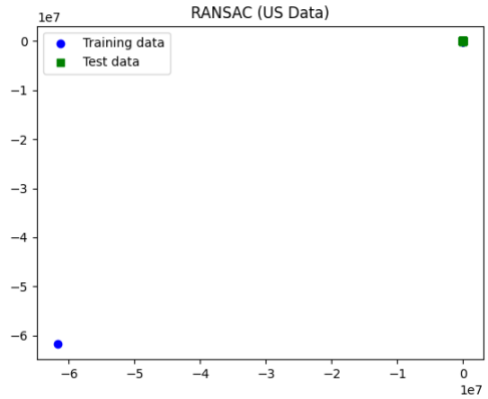

Figure 15: A residual plot to display the results of the RANSAC model on our US dataset.

- NYC Data:
  - Testing R2: -0.07917
  - Testing MSE: 41878.47
- CH Data:
  - Testing R2: -0.01207
  - Testing MSE: 136486.56

The negative R2 scores for all three datasets indicate that the RANSAC model did not fit the data well. The high training MSE and the relatively high testing MSE suggest that the model's predictions

were far from the actual prices of the listings. The RANSAC model did not perform well in predicting the prices of Airbnb listings in any of the datasets. Based on these findings, the RANSAC model is not recommended for predicting Airbnb listing prices.

Comparison

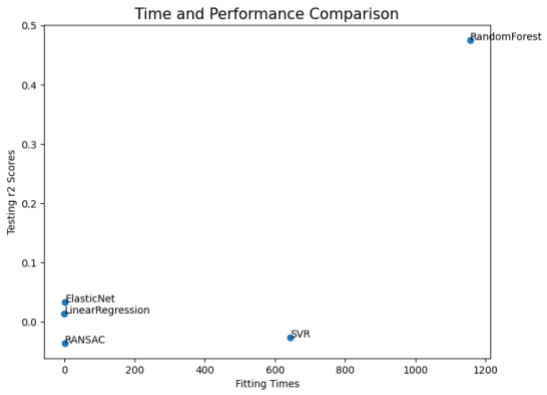

Figure 16: Comparison of time and performance of each initial model on the US dataset.

In the normal results, the team evaluated five initial regressors for predicting the price of Airbnb listings in the US, NYC, and Chicago datasets. The ElasticNet and Random Forest Regressor models outperformed the other models in predicting Airbnb listing prices for both New York City and Chicago. The Random Forest Regressor had a relatively low error rate with a testing R2 score of 0.47562 for the normal US data. The R2 score indicates that the model's predictions explain a significant portion of the variance in the target variable, and the low MSE values suggest that the model's predictions were close to the actual prices of the listings. However, the testing R2 score was lower than the training R2 score, which could indicate overfitting.

From the results detailed above, it's clear to see that the Random Forest Regressor model easily outperforms its competitors. Although it takes more time to complete training, the tradeoff for the high R2 score made it our choice moving forward with the project.

PCA Results:

Linear Regression

- Fitting Time: 0.032 seconds
- Training R2: 0.01321
- Training MSE: 187872.01
- Testing R2: 0.01602
- Testing MSE: 171068.25

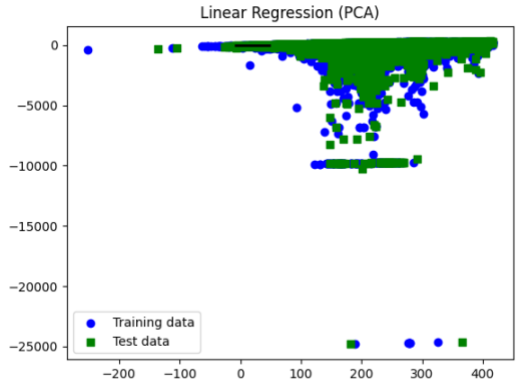

Figure 17: A residual plot to display the results of the Linear Regression model on the US PCA dataset.

The R2 score indicates that the model's predictions explain only a small portion of the variance in the target variable. The relatively low MSE values suggest that the model's predictions were close to the actual prices of the listings. However, it is important to note that these results are not as good as those obtained by the Random Forest Regressor in the initial testing phase.

SVR

- Fitting Time: 540 seconds
- Training R2: -0.00807
- Training MSE: 191925.41
- Testing R2: -0.00857
- Testing MSE: 175344.72

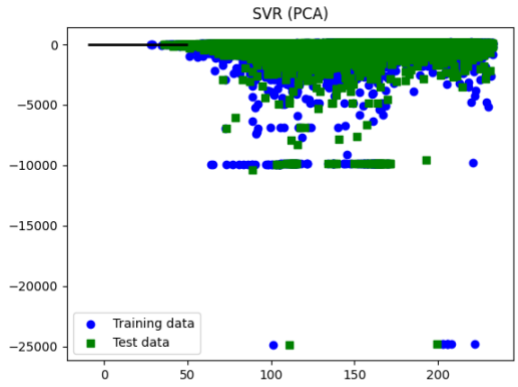

Figure 18: A residual plot to display the results of the SVR model on the US PCA dataset.

These results indicate that the SVR did not perform well in predicting the prices of Airbnb listings in any of the datasets. The negative R2 scores suggest that the model did not fit the data well, and the high MSE values suggest that the model's predictions were far from the actual prices of the listings. Overall, the SVR is not recommended as a suitable model for predicting Airbnb listing prices based on these findings.

ElasticNet

- Fitting Time: 0.028 seconds
- Training R2: 0.01257
- Training MSE: 187992.86
- Testing R2: 0.01496
- Testing MSE: 171251.78

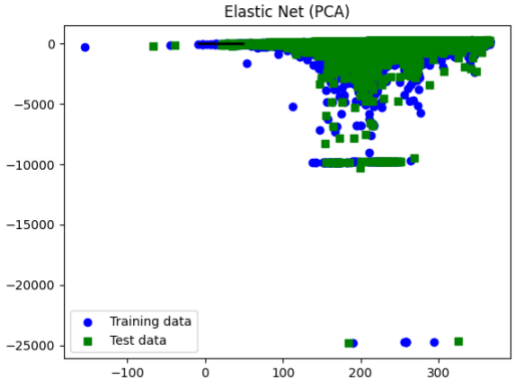

Figure 19: A residual plot to display the results of the Elastic Net model on the US PCA dataset.

Our findings suggest that the ElasticNet may be a suitable model for predicting Airbnb listing prices, especially in the US dataset. An R2 score of 0.2793 indicates that the ElasticNet model's predictions explain about 28% of the variance in the target variable. The relatively low MSE values suggest that the model's predictions were close to the actual prices of the listings. However, the ElasticNet model did not perform as well as the Random Forest Regressor in the normal US data. The ElasticNet model is recommended as an alternative to the Random Forest Regressor if a simpler model is desired or if there are concerns about overfitting the data.

Random Forest

- Fitting Time: 400 seconds
- Training R2: 0.87708
- Training MSE: 23401.11
- Testing R2: 0.07726
- Testing MSE: 160420.34

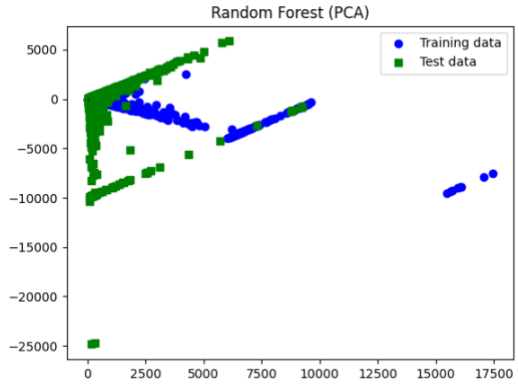

Figure 20: A residual plot to display the results of the Random Forest model on the US PCA dataset.

The results show that the Random Forest Regressor performed well in predicting the prices of Airbnb listings in the US. The high R2 scores indicate that the model's predictions explain a large portion of the variance in the target variable, and the low MSE values suggest that the model's predictions were close to the actual prices of the listings. However, the testing R2 score was lower than the training R2 score, which could indicate overfitting.

RANSAC

- Fitting Time: 3.1 seconds
- Training R2: -0.02410
- Training MSE: 194976.70
- Testing R2: -0.02552
- Testing MSE: 178290.30

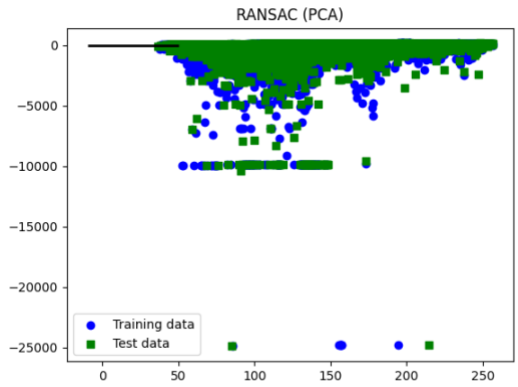

Figure 21: A residual plot to display the results of the RANSAC model on the US PCA dataset.

The negative R2 scores for all three datasets indicate that the RANSAC model did not fit the data well. The high training MSE and the relatively high testing MSE suggest that the model's predictions were far from the actual prices of the listings. Based on these findings, the RANSAC model is not recommended for predicting Airbnb listing prices.

Comparison

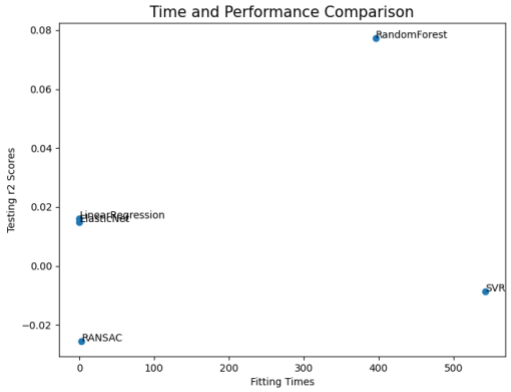

Figure 22: Comparison of time and performance of each initial model on the PCA US dataset.

The ElasticNet and Random Forest Regressor models performed better than other models in predicting Airbnb listing prices for both New York City and Chicago. The ElasticNet model is a good model for predicting Airbnb listing prices, especially in the US dataset, with an R2 score of 0.2793 explaining about 28% of the variance in the target variable. The low MSE values suggest that the model's predictions were close to the actual listing prices. However, the ElasticNet model did not perform as well as the Random Forest Regressor in the normal US data. The Random Forest Regressor model is recommended if a simpler model is desired or if there are concerns about overfitting the data. Further analysis could be conducted to investigate if this model performs similarly well in other countries or regions. The RANSAC model is also not recommended for predicting Airbnb listing prices as it did not fit the data well.

The PCA transformation itself proved ineffective for our five initial regression models; the difference in performance was negligible. However, we can still see that the Random Forest Regressor is an easy frontrunner for this task, thus fortifying our decision to move forward with this model.

8 Solution

For our final solution, our team utilized the Random Forest Regressor from the scikit learn library to create three different models—one for each of our datasets. For each model, we trained using about 70% of the dataset and withheld 30% for testing. We also increased the number of estimators in our Random Forests from 1,000 to 10,000. We reported the fitting time, training R2 and MSE, testing R2 and MSE, plotted the residuals, and created a new visual to display the feature importance for each model. Our results are shown below.

US Dataset Model

- Fitting Time: 8,700 seconds
- Training R2: 0.94262
- Training MSE: 0.00649
- Testing R2: 0.57007
- Testing MSE: 0.04869

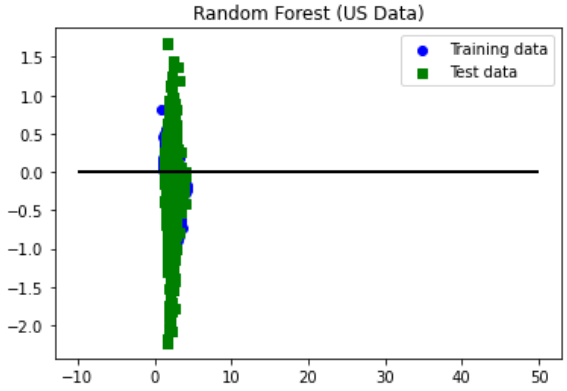

Figure 23: A residual plot to display the results of the Random Forest model on the US dataset.

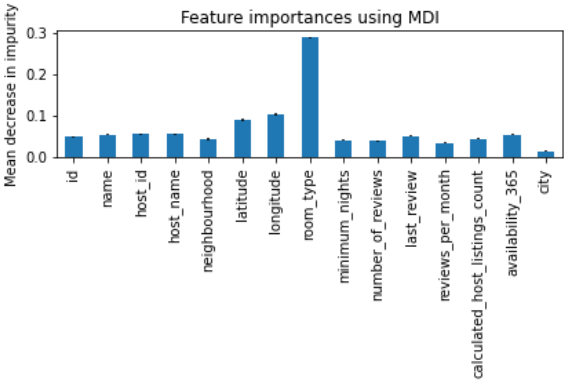

Figure 24: A feature importance bar plot to display how different attributes affect the prediction.

We can see that this model performed far better than any of our initial models with a testing R2 score of 0.57. However, this model achieves an even better training R2 of 0.94, indicating that the model may be overfitting. Furthermore, from the feature importance graph, we can see that the Airbnb price in this dataset is highly affected by the room type. Intuitively, this makes sense. Overall, we feel that this model performs well on the task at hand.

NYC Dataset Model

- Fitting Time: 1,400 seconds
- Training R2: 0.95054
- Training MSE: 0.00414
- Testing R2: 0.63286
- Testing MSE: 0.02988

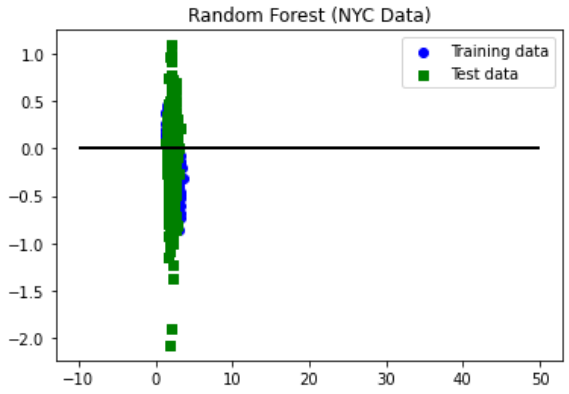

Figure 25: A residual plot to display the results of the Random Forest model on the NYC dataset.

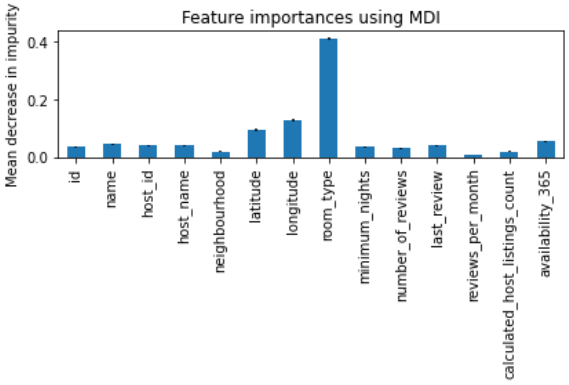

Figure 24: A feature importance bar plot to display how different attributes affect the prediction.

This model performs well on its dataset. The decision to train a separate model on this smaller dataset seems to have paid off, as it achieves a far better score than any of the previous models. This random forest may be overfitting the data, as the training results are better than the testing results. However, we are still satisfied with the performance. From the feature importance graph, we can see that the price is most affected by the latitude, longitude, and room type of the listing.

Chicago Dataset Model

- Fitting Time: 150 seconds
- Training R2: 0.93684
- Training MSE: 0.00647
- Testing R2: 0.53915
- Testing MSE: 0.04881

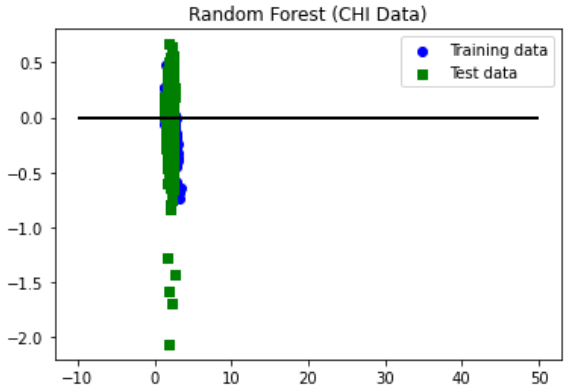

Figure 26: A residual plot to display the results of the Random Forest model on the Chicago dataset.

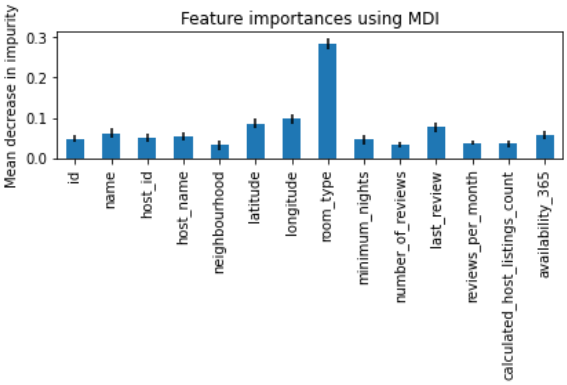

Figure 27: A feature importance bar plot to display how different attributes affect the prediction.

Like the two previously listed models, this technique achieved high performance metrics, but is also possibly overfitting. However, it is also the best model for the Chicago dataset that we have achieved thus far. We can see from the feature importance graph that the latitude, longitude, and room type have a high impact on the results.

Comparison

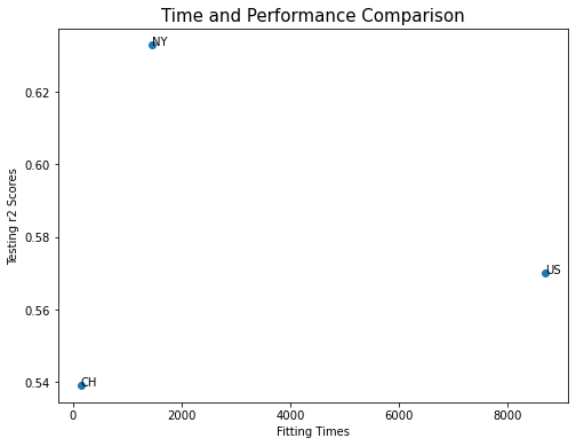

Figure 28: Comparison of time and performance of each model on its respective dataset.

From the graph above, we can draw some interesting conclusions. First, the NYC dataset easily achieved the highest testing performance, followed by the US and then Chicago. By looking at the times, we can see that they go in order of dataset size, from Chicago being the smallest and lowest training time to the US being the largest and highest training time.

9 Future Work

There are several ways that the findings of this report can be expanded upon to create a more accurate, robust price prediction model. First, one could experiment with the parameters of the Random Forest regressor to reduce the overfitting that we saw in our own experiments. Next, one could remove outliers from the datasets in order to feed the model a more accurate representation of the data. Lastly, exploring datasets from other countries may provide a more well-rounded price prediction. These points were out of the scope of our project.

10 Conclusion

The study aimed to leverage machine learning techniques for the purpose of forecasting Airbnb listing prices in two major cities, New York City and Chicago. To achieve this, the study employed three datasets that contained comparable features and evaluated several initial regressors. Following a thorough analysis, the Random Forest Regressor was ultimately selected for its high performance and used for price prediction.

In addition to the selection of the Random Forest Regressor, the report also implemented data scaling to enhance model efficiency and employed evaluation metrics to assess model performance. Specifically, the study utilized metrics such as mean squared error (MSE) and R-squared (R2) to evaluate the accuracy of the model predictions. Through this process, the study was able to validate the effectiveness of the Random Forest Regressor in accurately forecasting Airbnb prices.

While Principal Component Analysis (PCA) was also tested as a potential technique to improve model performance, the results did not show any significant improvement compared to the original model. However, this does not detract from the overall conclusion of the study, which is that machine learning techniques, specifically the Random Forest Regressor, can be highly effective in predicting Airbnb prices.

In conclusion, the findings of this study highlight the potential of machine learning techniques in the field of short-term rental price forecasting. The Random Forest Regressor was identified as a suitable choice for this task, and the study's implementation of data scaling and evaluation metrics further enhance the model's performance. Future research may build on this foundation by exploring additional machine learning techniques or further feature engineering to improve the accuracy of price predictions in this domain.

11 References

air bnb worth - Google Zoeken. (n.d.). https://www.google.com/search?q=air+bnb+worth

New York City Airbnb Open Data. (2019, August 12). Kaggle. https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

PricewaterhouseCoopers. (n.d.). *US Hospitality Directions: November 2022*. PwC. https://www.pwc.com/us/en/industries/consumer-markets/hospitality-leisure/us-hospitality-directions.html

Statista. (2023, January 20). *Travel and tourism: share of global GDP 2000-2021*. https://www.statista.com/statistics/1099933/travel-and-tourism-share-of-gdp/

U.S. Airbnb Open Data. (2020, October 25). Kaggle. https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data

Weber, E. (2022, November 15). *How Airbnb has Disrupted Hotel Management*. Verdant®. https://www.verdant.co/how-airbnb-has-disrupted-hotel-management/