Final Report: Predictive Analysis of AirBnb Prices in New York City

Introduction

Welcome to the comprehensive report detailing our project on predicting AirBnb rental prices in the vibrant city of New York. In this report, we will walk you through our entire project workflow, from data preprocessing and exploratory data analysis (EDA) to feature engineering, model selection, evaluation, and the ultimate price predictions.

Data Preprocessing

Handling the Initial Data

Our journey began with a dataset containing various details about AirBnb listings in New York City, including essential features like neighborhood, room type, price, and reviews. One early challenge we faced was missing values in the 'reviews' column. To address this, we decided to fill these gaps with zeros temporarily, ensuring that no valuable data was lost.

Normalization

As we progressed, we explored data normalization, a critical step in ensuring that our features were on a consistent scale. While normalization often enhances model performance, we found that it didn't yield significant improvements in our particular case.

Model Selection

Next, we ventured into model selection, considering a range of machine learning algorithms, including linear regression, XGBoost, and decision trees. Surprisingly, linear regression emerged as the top performer, even before diving deep into feature engineering. This unexpected result served as our initial foundation.

Exploratory Data Analysis (EDA)

Key Insights

During our exploratory data analysis phase, we made several intriguing discoveries:

- **Missing Values**: Our dataset had its fair share of missing values, primarily in the 'reviews' column. We initially filled these gaps with zeros, although we later recognized the potential for noise in our analysis.
- **Neighborhood Analysis**: We delved into the distribution of listings across different neighborhood groups, gaining valuable insights into the city's geography.
- **Geospatial Visualization**: We visually mapped out the listings, revealing spatial patterns and geographic trends.
- **Room Type Analysis**: Our analysis of room types helped us understand the diverse range of accommodations available.
- **Price Distribution**: We explored the distribution of prices, identifying the typical price range for AirBnb listings.
- **Neighborhood vs. Price**: A deeper analysis uncovered the significant impact of specific neighborhoods on rental prices.

Feature Engineering

To enhance model performance, we embarked on feature engineering. Here are some key techniques we employed:

- **Leveraging the 'Name' Column**: We extracted valuable information from the 'name' column, including details about the listing's name, number of bedrooms, bathrooms, and ratings.
- **Introducing 'Neighborhood Mean Price'**: A crucial addition to our features, 'neighborhood_mean_price,' captured the average price within each neighborhood. This feature allowed us to consider neighborhood-specific price trends.
- **Scikit-Learn Pipeline**: We introduced a Scikit-Learn pipeline to streamline our data preprocessing. This pipeline efficiently handled data imputation, scaling, and encoding.

Model Development and Evaluation

Implementing the Scikit-Learn Pipeline

To improve efficiency and maintain reproducibility, we integrated a Scikit-Learn pipeline. This pipeline seamlessly incorporated data imputation, scaling, encoding, and our new 'neighborhood_mean_price' feature.

Model Evaluation

We rigorously evaluated our models using cross-validation techniques, ensuring a robust assessment of their performance. Additionally, a separate test set was employed for the final evaluation.

Challenges with Handwritten Model and Pipeline

Despite our successes in model development and data preprocessing, we encountered challenges when integrating a handwritten model with the Scikit-Learn pipeline. While we aimed to create a seamless connection between our model and the preprocessing and evaluation pipeline, technical difficulties arose.

Conclusion

In conclusion, this project took us on a fascinating journey from basic data preprocessing and model selection to advanced feature engineering and pipeline integration. Our initial preprocessing steps, which included filling empty rows with zeros and data normalization, didn't yield the expected improvements.

The unexpected triumph of linear regression in the initial stages compelled us to focus on feature engineering and introduce the Scikit-Learn pipeline to streamline our workflow.

Though we faced integration challenges with our handwritten model, this project serves as a compelling example of the iterative nature of data science projects. It underscores the importance of adaptability and flexibility when dealing with unexpected findings and technical hurdles.

As we wrap up this project, we recognize the value of the 'neighborhood_mean_price' feature and see potential for future work to explore additional feature engineering techniques that could further enhance our predictive accuracy. Thank you for joining us on this data-driven journey through the vibrant world of AirBnb prices in New York City.