**440 Final Report**

**Younghoon Kim**

**DS 440**

**Abstract**

This research looks into the sophisticated task of predicting Airbnb prices across 33 U.S. cities (excluding Hawaii and Alaska) using data from Airbnb listings. The challenge lies in the fact that traditional machine learning models often struggle to differentiate between properties with similar features yet substantial price variations. To address this, we employ the XGBoost model enhanced by conformal prediction. This integration allows us to model the prediction error, capturing information that is typically lost during the boosting process. Our initial experiments with models like Random Forest, SVM, and Gradient Boosting led us to the selection of XGBoost, which initially achieved an R² of 54%, the highest among other models. Furthermore, we explore two types of conformal prediction: marginal and conditional. While marginal coverage reached an average of 90%, it was inconsistent across locations. In contrast, conditional coverage, which groups properties into 100 clusters using KNN based on longitude and latitude, successfully achieved consistent 90% prediction coverage in nearly all cities. These findings are significant as they pave the way for more accurate and reliable Airbnb price predictions, benefiting both hosts and guests.

**Introduction**

Predicting Airbnb prices accurately is crucial for both hosts and guests to ensure fair pricing and market efficiency. Traditional machine learning models, however, often fall short when they encounter properties with nearly identical features but different pricing due to subtle factors unaccounted for by the models. This study aims to overcome these limitations by applying an advanced machine learning technique, XGBoost, combined with conformal prediction methods to enhance accuracy and reliability across diverse U.S. cities.

The research initially experimented with several machine learning models, including Random Forest, SVM, and Gradient Boosting, but XGBoost was selected for its superior performance in preliminary tests, achieving an R² of 54%. Conformal prediction was applied to address the issue of information loss during the boosting process. This method allowed for modeling the prediction error, providing a more robust framework for uncertainty in price predictions.

Two approaches of conformal prediction were explored: marginal and conditional. While achieving an average coverage of 90%, the marginal approach showed variability in performance across different regions. The conditional approach was employed to improve geographic consistency, using KNN to cluster properties based on longitude and latitude. This significantly enhanced the model’s ability to achieve consistent 90% coverage across US cities.

**Literature Review**

**Introduction to Conformal Prediction**

Conformal prediction provides a robust method for assessing the reliability of predictions made by machine learning models. It offers distribution-free uncertainty quantification, particularly valuable in applications requiring high confidence in prediction accuracy, such as medical diagnostics. This technique generates prediction sets reliably containing the true outcome with a user-specified probability, regardless of the underlying data distribution.

**Methodology for Continuous Variables**

Conformal prediction handles continuous variables effectively through techniques such as conformalized quantile regression. This approach leverages the properties of quantile regression to form prediction intervals that adaptively adjust to the uncertainty inherent in the data points. The intervals are statistically valid even when the assumptions about the underlying data distributions are relaxed. A similar method could also be applied to any predictive model.

**Marginal and Conditional Coverage**

Marginal coverage ensures that the prediction intervals include the true value across all test points with a specified probability. However, this does not account for variations in coverage probability across different subsets of the data. Conditional coverage addresses this by ensuring that the prediction intervals provide reliable coverage for every possible value of the input features, thus providing a more nuanced and equitable measure of uncertainty across diverse conditions.

**Feature-Stratified Coverage**

Feature-stratified coverage extends the principles of conditional coverage by stratifying the test inputs based on specific features, such as geographic location or demographic groups. This method aims to maintain uniform coverage across different strata, thus ensuring that the model's predictions are accurate on average and fair and reliable across different subgroups of the population.

**Conclusion on Conformal Prediction**

Conformal prediction is a powerful tool for enhancing the reliability of machine learning predictions, especially in settings where accurate uncertainty quantification is critical. By adapting the methodology to handle continuous variables and emphasizing both marginal and conditional coverage, conformal prediction provides a comprehensive framework for addressing the challenges of prediction in complex, real-world datasets.

**Data Collection**

The data for this study was sourced from an open-access platform called "Inside Airbnb." This resource provides publicly available data concerning Airbnb listings, categorized by city globally. The data includes comprehensive details about each listing, such as location, pricing, reviews, and host information. By utilizing "Inside Airbnb," researchers can access and analyze data across various urban settings. This is crucial for a study that aims to predict Airbnb prices based on specific property features and regional market trends. The advantage of using this data source is its transparency and accessibility, which ensures that the analysis can be replicated and validated by others in the research community. Additionally, the extensive nature of the dataset allows for a robust analysis of the factors influencing Airbnb pricing dynamics in different cities.

**Methodology**

**Data Preprocessing**

The dataset includes various attributes of Airbnb listings, including location coordinates, pricing, amenities, number of reviews, and host details. Prior to analysis, the data underwent a comprehensive preprocessing stage. This included removing outliers, filling in missing values, and encoding categorical variables to numerical formats to prepare the dataset for machine learning models. Special attention was given to ensuring that the preprocessing steps preserved the integrity of the data while making it suitable for predictive modeling.

**Model Selection and Development**

The initial phase of modeling involved experimenting with several traditional machine learning algorithms, including Random Forest, SVM, and Gradient Boosting, to establish a baseline performance. After preliminary testing, XGBoost was selected due to its superior performance in handling sparse data and its effectiveness in managing different data scales. The XGBoost model was then tuned for parameters such as learning rate, max depth, and n\_estimators through a grid search method coupled with cross-validation to optimize performance.

**Implementation of Conformal Prediction**

Conformal prediction was integrated into the methodology to enhance the predictions' reliability and address the information loss issue during model training. This technique was applied to calibrate the uncertainty in the predictive models, ensuring that the prediction intervals would cover the true Airbnb prices with a specified probability of 90%. Initially, the model utilized marginal coverage to gauge the general applicability across different cities. Observing some inconsistencies in coverage across diverse geographic regions, the method transitioned to conditional coverage, where the model accuracy was enhanced by stratifying the data using location-based clustering.

**Clustering for Geospatial Analysis**

To further refine the model, properties were grouped into clusters based on their geographical coordinates using the K-Nearest Neighbors (KNN) algorithm. This approach created 100 distinct clusters, facilitating a more location-based analysis of price determinants and improving the model’s ability to achieve consistent 90% coverage across different cities. This geospatial clustering optimized the model's accuracy and provided more profound insights into regional pricing trends crucial for potential investors and hosts in the Airbnb market.

**Conclusion**

This research successfully implemented an advanced machine learning approach using XGBoost coupled with conformal prediction techniques to predict Airbnb prices across 33 major U.S. cities, excluding Hawaii and Alaska. The study leveraged a comprehensive dataset from "Inside Airbnb," meticulously preprocessed to ensure data quality and relevance for predictive modeling. The adoption of XGBoost as the primary algorithm was justified by its robust performance in preliminary tests, demonstrating strong capability in handling diverse data characteristics inherent in Airbnb listings.

Integrating conformal prediction added a layer of reliability to the model predictions, enabling the generation of prediction intervals with a specified coverage probability. This was particularly beneficial in addressing the uncertainties commonly associated with predictive modeling in dynamic market environments like Airbnb. Through geographical clustering, the research introduced a novel approach to handling regional variations in Airbnb pricing, enhancing the model’s accuracy and applicability. The conditional coverage method, adjusted for geographical specifics using KNN clustering, proved effective in achieving consistent coverage across different cities, illustrating the model's adaptability and precision.

The maps presented offer a detailed geospatial evaluation of the predictive model's capability in forecasting Airbnb prices with 90% confidence intervals. The presence of blue and green markers across the landscapes in the maps signifies regions where the model's predictions attained the intended 90% coverage, denoting effective performance. Conversely, the emergence of orange markers points to localities where the model's performance fell short of the anticipated coverage threshold. Notably, the map depicting the model with conditional coverage (on the left) displays no orange markers, suggesting a uniform coverage achievement across various urban areas. On the other hand, the map illustrating the model with marginal coverage (on the right) reveals several orange markers, indicating a lack of uniformity in coverage in several city vicinities, which underscores areas for potential model enhancement.

A map with many points

Description automatically generatedA map of a city

Description automatically generated

[Conditional Coverage (Left), Marginal Coverage (Right)]

The research findings contribute valuable insights into the factors influencing Airbnb pricing, offering a helpful tool for hosts, investors, and analysts in the real estate and hospitality sectors. Additionally, the methodologies employed can be adapted for broader applications in predictive analytics across various domains, requiring a nuanced understanding of geospatial and market-driven dynamics. Overall, this study not only advances the field of predictive modeling in real estate pricing but also sets a precedent for using machine learning techniques enhanced by conformal prediction to manage and interpret complex datasets, thereby providing a robust framework for future research in similar arenas.

**Future Work**

**Optimization of Prediction Intervals**

The first area of future work involves optimizing the model to minimize the size of the prediction intervals without compromising the coverage guarantee. This refinement aims to provide more precise price estimates, thereby increasing the utility for users who rely on narrow price ranges for decision-making. Techniques such as feature engineering to identify variables that could narrow these intervals and advanced regression methods that better capture the complexities of the market could prove valuable.

**Exploration of New Models and Features**

Secondly, exploring alternative modeling approaches and integrating additional predictive features hold promise for enhancing model performance. Investigating models that better capture the non-linear relationships within the data, such as neural networks or ensemble methods, could offer substantial improvements. In parallel, the addition of novel features derived from emerging market trends or user-generated content could provide deeper insights and more accurate predictions. To ensure their efficacy, continuous evaluation of model performance with these enhancements would be essential.

**Incorporation of Temporal and Seasonal Factors**

Lastly, incorporating temporal and seasonal factors into the model's predictive dynamics is a crucial next step. Airbnb pricing is inherently time-sensitive, fluctuating with changes in demand patterns, local events, and seasonal trends. Developing a model that can adapt to these temporal dynamics using time-series analysis or recurrent neural networks would make the predictions more relevant for real-time applications. This temporal modeling could also extend to predicting long-term price trends, offering users insights into the future of the Airbnb market.