ISYE 6740 Fall 2024 - Project Proposal

Team Members: Krishna Dhavala, Harshitha Thotakura

October 20, 2024

Primary Objective

The primary goal of this project is to predict Airbnb New York listing prices using an ensemble learning approach. We will leverage multiple base models—Linear Regression, Random Forest, and XGBoost—and combine their predictions through a Ridge Regression meta-model. This ensemble method aims to improve predictive accuracy and generalization by capturing both linear and non-linear relationships in the data. Additionally, we will evaluate the performance of each model individually to better understand their strengths and limitations.

Data Source

We will use the publicly available Airbnb dataset from Kaggle:
`https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata`

The dataset contains various features such as location, number of rooms, amenities, and user reviews, all of which are potentially influential in determining the price of Airbnb listings. The dataset consists of thousands of listings, providing a rich source of data for analysis and model training.

Proposed Methodology

Data Preprocessing

Data preprocessing is a crucial step that ensures the dataset is clean and ready for model training. The following steps will be applied:

- **Handling Missing Values**: Missing values can negatively impact model performance. Depending on the significance of the missing data, we will either remove or impute them using

techniques such as mean, mode, or advanced methods like cubic splines. We will analyze the missingness pattern to ensure it does not introduce bias.

- **Feature Scaling**: We will apply a MinMax scaler to normalize numerical features to the range [0,1], ensuring consistent feature scaling across all models. This is especially important for models such as Linear Regression, which can be sensitive to feature magnitudes. The transformation formula is as follows:

$$x^i_{\text{scaled}} = \frac{x^i_f - \min(x_f)}{\max(x_f) - \min(x_f)}$$

- **Categorical Encoding**: Categorical features such as neighborhood or property type will be transformed into numerical representations using OneHotEncoding. This encoding will ensure that categorical features are appropriately handled by the machine learning models.

The dataset will be split into training and testing sets (e.g., 80/20 split) for model evaluation.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) will help us understand the underlying patterns and relationships in the dataset. Some of the key EDA techniques we will use include:

- **Correlation Analysis**: Correlation coefficients will be calculated between features and the target variable (price). This will help us identify features with strong linear relationships and potential multicollinearity issues. Highly correlated features may be considered for removal.
- **Data Visualization**: We will generate scatter plots, pair plots, and heatmaps to visualize the interactions between features. These visualizations will also help identify trends and patterns that may influence the model's predictions. For example, location-based price trends or the impact of amenities on price can be identified through these plots.

Outlier Detection

Outliers can distort model training and reduce predictive performance, especially in models like Linear Regression. To detect and handle outliers, we will use the following techniques:

- **Local Outlier Detection (LOD)**: Using K-Nearest Neighbors (KNN), LOD will detect points with significantly lower local density compared to their neighbors, flagging them as potential outliers.
- **Isolation Forest**: In cases where the dataset contains clusters with varying densities, Isolation Forest will be used to isolate rare points. This technique does not rely on density estimation and is effective in high-dimensional spaces.

Outliers that are identified and deemed incorrect or irrelevant will be removed, while those that are valid will be retained.

Feature Selection

Selecting the most relevant features is essential for improving model performance by reducing noise and complexity. Our feature selection process will include:

- **Correlation Analysis**: Features with weak correlations to the target variable will be considered for removal. This step ensures that we focus on features that are predictive of Airbnb prices.
- **Feature Importance from Tree-Based Models**: Using Random Forest and XGBoost, we will calculate feature importance scores. Features with low importance will be dropped, allowing the model to focus on the most influential predictors.

This approach will help reduce dimensionality, improve computational efficiency, and enhance the interpretability of the model.

Model Selection and Implementation

We will implement three models—Linear Regression, Random Forest, and XGBoost—each serving a different purpose in the ensemble:

Linear Regression

- **Justification**: Linear Regression provides a simple and interpretable baseline model. It assumes linear relationships between features and price, making it a good starting point for understanding the general trends in the data.
- **Implementation**: We will use `scikit-learn` to implement Linear Regression. The model will be trained and evaluated using metrics such as Mean Absolute Error (MAE) and R-squared (R^2). Diagnostic plots, such as residual vs. fitted value plots, will be generated to ensure the model assumptions hold.

Random Forest

- **Justification**: Random Forest is a robust ensemble learning method that captures non-linear relationships and interactions between features. It also handles outliers and missing data effectively, making it ideal for complex datasets like Airbnb listings.

- **Implementation**: Using `scikit-learn`, we will apply RandomForestRegressor. Hyperparameters such as the number of trees (n_estimators) and the maximum depth (max_depth) will be optimized using cross-validation. We will evaluate model performance using MAE, R-squared, and Out-of-Bag (OOB) error.

XGBoost

- **Justification**: XGBoost is a powerful gradient boosting algorithm that corrects errors in sequential iterations. Its regularization techniques prevent overfitting, making it suitable for high-dimensional data and handling complex interactions between features.
- **Implementation**: XGBoost will be implemented using the `xgboost` library. We will tune hyperparameters such as learning rate, maximum tree depth, and the number of boosting rounds through cross-validation. The model will be evaluated using MAE, RMSE, and R-squared.

Model Evaluation

To assess the performance of the models, we will use the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: Emphasizes larger errors by squaring them before averaging.
- **R-squared ($\mathbf{R^2}$)**: Explains the proportion of variance in the target variable captured by the model.

Cross-validation will be used to ensure that the model generalizes well to unseen data. For Random Forest and XGBoost, feature importance will be analyzed for interpretability.

Ensemble Model Inference

To improve prediction accuracy, we will implement an ensemble learning approach. Linear Regression, Random Forest, and XGBoost will serve as base models, each capturing different aspects of the data. These base models will be combined using Ridge Regression as a meta-model.

Ridge Regression will take the outputs of the base models and learn to optimally weight them, thereby reducing individual weaknesses and improving overall performance. This stacking method allows the ensemble to leverage the strengths of each model. Out-of-sample predictions from the base models will be used to train the meta-model, ensuring that the ensemble generalizes well to new data.

Final Results

Upon completion of the project, we expect that the ensemble model will outperform the individual base models in terms of predictive accuracy. By combining the strengths of Linear Regression (which captures linear relationships), Random Forest (which handles complex interactions and non-linearities), and XGBoost (which excels at correcting residual errors through boosting), we anticipate that the Ridge Regression meta-model will effectively integrate these outputs to produce a more reliable and accurate prediction of Airbnb listing prices.

The final ensemble model is expected to reduce error rates and improve R-squared values compared to any of the individual models. Additionally, the feature importance analysis will provide insights into which features have the most significant impact on price prediction, offering further interpretability and value from the analysis.

We will report the final results in terms of key metrics such as MAE, RMSE, and R-squared, and provide a comprehensive evaluation of the model's generalization to unseen data.

References

- Wikipedia. *Ensemble Learning*. Retrieved from: https://en.wikipedia.org/wiki/Ensemble_learning
- ScienceDirect. *Ensemble Modeling*. Retrieved from: https://www.sciencedirect.com/topics/computer-science/ensemble-modeling#:~:text=Ensemble%20modeling%20is%20a%20process,the%20ensemble%20approach%20is%20used.
- Analytics Vidhya. *Comprehensive Guide for Ensemble Models*. Retrieved from: https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/
- GeeksforGeeks. *Feedforward Neural Network*. Retrieved from: https://www.geeksforgeeks.org/feedforward-neural-network/
- AFIT-R. *Feedforward Neural Networks*. Retrieved from: https://afit-r.github.io/feedforward_DNN#:~:text=The%20output%20layer%20is%20driven,the%20probability%20of%20each%20class.