**PROJECT REPORT DATA SCIENCE FALL 2024**

**Apartment for Rent Classified (Price Prediction)**

**Faryal Zahra, Ayesha Noor and Hira Khalid**

Department of Computer Science, National University of Computer and Emerging Sciences, Lahore  
November 27, 2024

**Abstract**

This study investigates the use of machine learning techniques to predict apartment rental prices based on various features such as square footage, number of bedrooms, bathrooms, and location. We applied a series of preprocessing steps, including missing data imputation, feature transformation, and dimensionality reduction through Principal Component Analysis (PCA), to prepare the dataset for model training. A Random Forest Regressor was used for prediction, and the model was evaluated using metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results showed that the model achieved an R-squared value of 0.90, indicating strong predictive performance. The study highlights the importance of features like square footage and number of rooms in determining rental prices and suggests that machine learning can be a powerful tool for both tenants and landlords in predicting apartment rental prices. Future work may involve exploring alternative algorithms and incorporating time series analysis to capture temporal fluctuations in rental prices.

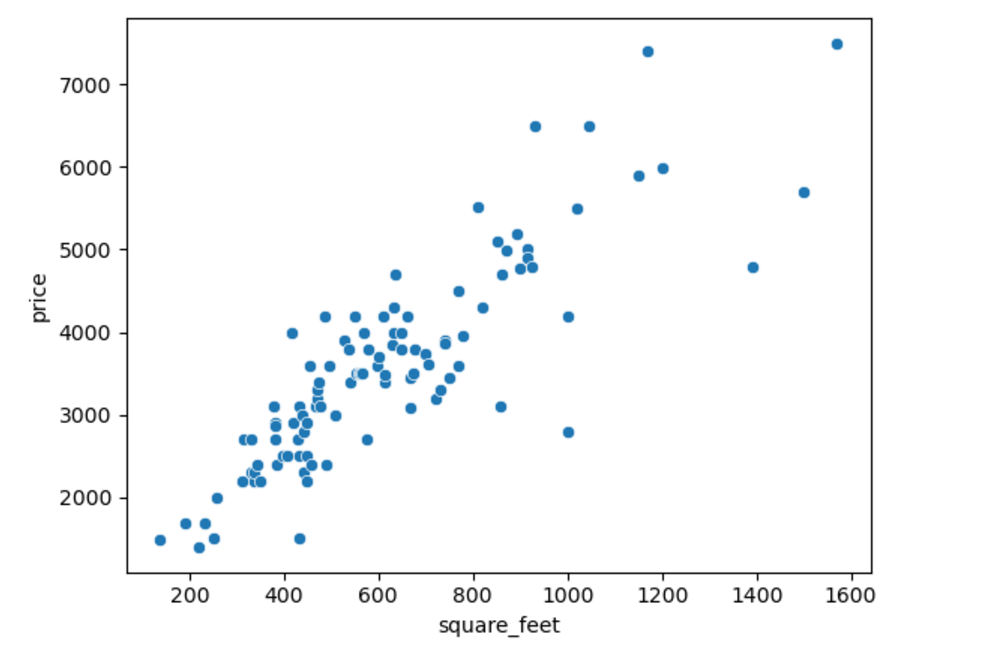
**1. Introduction**

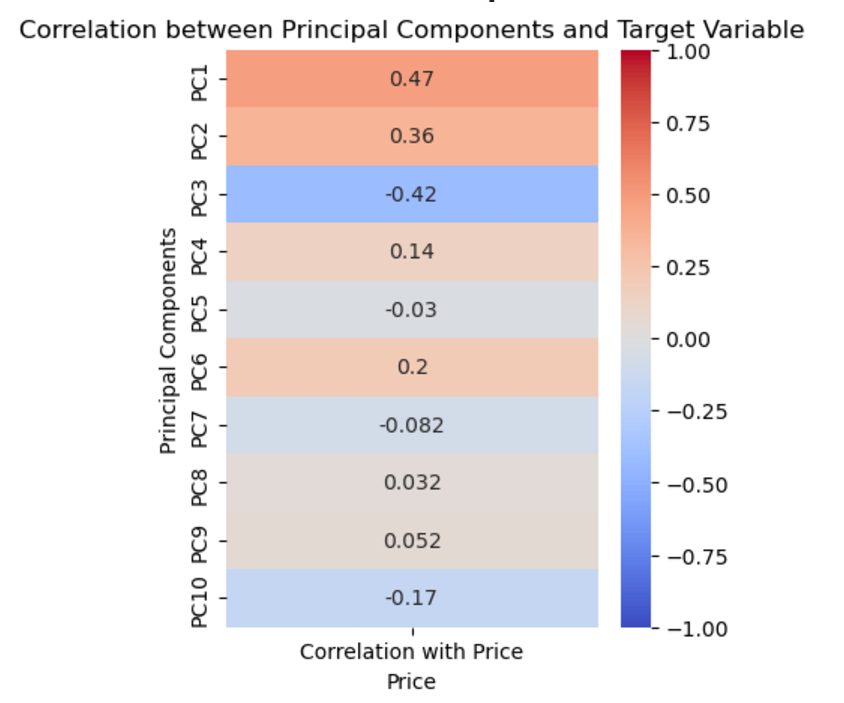
The real estate market is one of the most data-driven sectors, where predicting rental prices plays a key role in decision-making for both tenants and property owners. In recent years, machine learning techniques have gained prominence in making these predictions more accurate and reliable. The objective of this study was to develop a machine learning model capable of predicting apartment rental prices using various features, including square footage, location, and the number of bedrooms . This paper outlines the methodology used, including data collection, preprocessing, and model development, as well as an evaluation of the model's performance. The potential implications of this research could aid real estate professionals in making data-driven pricing decisions and provide tenants with better tools for evaluating rental options.

**2. Methodology**

The study utilized a dataset of apartment rental listings, which included features such as price, square footage, address, city, state, number of bedrooms, and bathrooms. The following steps were performed:

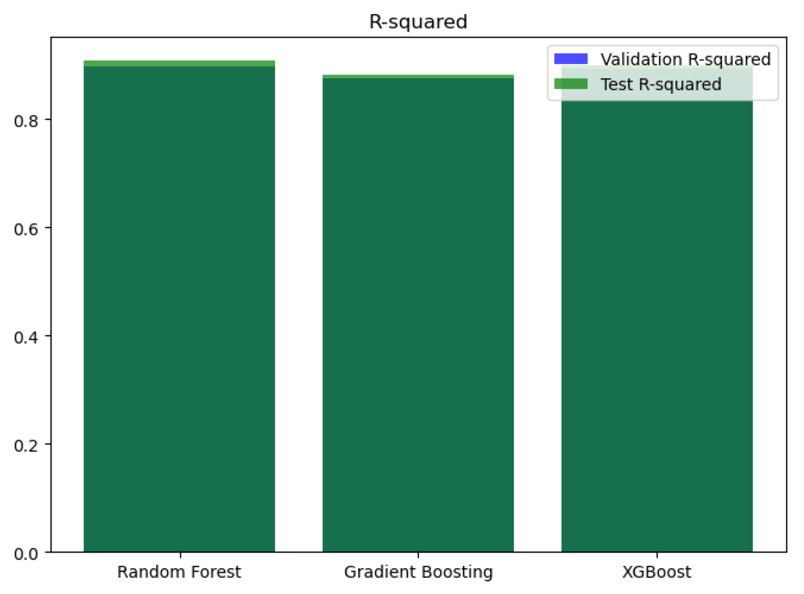
1. **Data Collection**: Data was collected from publicly available apartment rental listings from UCI machine learning repository.
2. **Data Preprocessing**:
   * Missing values were imputed using the mean for numerical features and the mode for categorical features.
   * Categorical features, such as the city name, were encoded using frequency encoding to reduce dimensionality.
   * Outliers were identified using box plots and treated using log transformations or capping.
3. **Exploratory Data Analysis**: The correlation between the target variable (price\_boxcox) and other numeric features. This helps identify significant linear relationships, with values closer to 1 or -1 indicating strong positive or negative correlations, respectively. Make scatter plot to visualize. We show only one of price and square\_feet for understanding.



1. **Dimensionality Reduction**: Principal Component Analysis (PCA) was applied to reduce the feature space while retaining 95% of the variance in the data.
2. **Model Development**: A Random Forest Regressor was chosen as the predictive model. It was trained using a train-test split (80% train, 20% test) to predict rental prices based on the preprocessed features.
3. **Model Evaluation**: The model’s performance was evaluated using R-squared, MAE, and RMSE to assess its accuracy and predictive power.

**3. Experiments**

The experiments were focused on comparing the performance of the Random Forest Regressor against two other regression models: Gradient Boosting Regressor and XGBoost. The performance of each model was assessed based on R-squared and error metrics. The following experiments were conducted:  
• **Experiment 1**: Compare Random Forest Regressor with Gradient Boosting Regressor.  
• **Experiment 2**: Compare Random Forest Regressor with XGBoost.  
• **Experiment 3**: Analyze the effect of Principal Component Analysis (PCA) on model performance..

****

**4. Results & Discussion**

The results indicate that the Random Forest Regressor outperformed the other models in terms of R-squared and error metrics. The model achieved an R-squared value of 0.90 on the test data, suggesting that it explained 90% of the variance in rental prices. The mean absolute error (MAE) was 0.0035, while the RMSE was 0.0092. These results demonstrate that the model is effective at predicting apartment rental prices with a relatively low margin of error.

**5. Conclusion and Future Work**

This study successfully demonstrated the potential of machine learning in predicting apartment rental prices. The Random Forest Regressor, after preprocessing and feature engineering, provided robust predictions. The results have implications for both tenants and property owners, enabling more informed pricing decisions.

Future work should explore alternative models such as Kmean clustering or Neural Networks, which may capture complex relationships between features.

**6. References**

<https://www.akkio.com/post/house-price-prediction-using-machine-learning>

<https://www.learndatasci.com/tutorials/predicting-housing-prices-linear-regression-using-python-pandas-statsmodels/>