**USA HOUSING LISTINGS**

**SACRED HEART UNIVERSITY**

**BUSINESS STRATEGY & ANALYSIS BUAN-680-CO**

INSTRUCTOR- SABBIR HASSAN

DATE: 06-22-2024

**GROUP-2:**

**1. UDAYA LAKSHMI PUNNAM**

**2. KEERTHI PULIPAKA**

**3. DEEPTHI POOSALA**

**4. CHANDRA SIDHARTHA KODELA**

**1. INTRODUCTION**

**1.1. CONTEXT**

The dataset originates from the real estate sector in the United States, focusing on housing listings. Real estate is a significant sector that affects the economy, urban development, and individual financial decisions. Housing listings data can provide insights into market trends, pricing strategies, and regional development, making it a valuable resource for various stakeholders in the real estate market.

**1.2. SIGNIFICANCE OF THE PROBLEM**

The dataset originates from the real estate sector in the United States, focusing on housing listings. Real estate is a significant sector that affects the economy, urban development, and individual financial decisions. Housing listings data can provide insights into market trends, pricing strategies, and regional development, making it a valuable resource for various stakeholders in the real estate market.

**2. OBJECTIVES**

The primary objective of this report is to develop and evaluate predictive models for estimating property prices based on various features of housing listings. The specific goals are:

* To preprocess and analyze the dataset to understand the key features influencing property prices.
* To build and compare different machine learning models to identify the most accurate and reliable method for price prediction.
* To provide actionable insights and recommendations based on the model's findings to enhance pricing strategies and decision-making processes in the real estate market.

The success of the analysis will be measured using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared values to evaluate the performance of the predictive models.

**3.1 DATA COLLECTION**

**3.1.1. SOURCE**

The dataset is sourced from various real estate websites and aggregated by Austin Reese, available on Kaggle. The dataset includes 384,977 rows and 22 columns, detailing housing listings across the United States.

**3.1.2. TIMEFRAME**

The specific timeframe of the data collection is not provided, but it represents a snapshot of housing listings at the time of aggregation.

**3.1.3. PREPROCESSING STEPS**

* **Handling Missing Values:** Addressing missing values, particularly in fields like "laundry\_options", “parking\_options”, “description”, “lat”, and “long”.
* **Feature Types:** Identifying categorical and numerical features for appropriate processing.
* **Multi-Collinearity:** Addressing potential multi-collinearity issues, especially between highly correlated features like "cats\_allowed" and "dogs\_allowed".

**3.1.4. LIMITATIONS AND ASSUMPTIONS**

* **Geographical Coverage:** The dataset covers listings across the United States but may not represent all regions equally.
* **Temporal Relevance:** The data represents listings at a specific time, which may not reflect current market conditions.
* **Data Quality:** Incomplete or inaccurate entries could affect the analysis.

**3.2. DATA ANALYSIS TECHNIQUES**

Several analytical techniques and tools will be used in the analysis:

* **Data Cleaning and Preprocessing:** Handling missing values, encoding categorical variables, and normalizing numerical features.
* **Exploratory Data Analysis (EDA):** Understanding data distributions, identifying outliers, and exploring relationships between features.
* **Feature Engineering:** Creating new features or transforming existing ones to enhance model performance.
* **Predictive Modeling:** Using machine learning algorithms such as Linear Regression, Random Forest, and Gradient Boosting to predict property prices.
* **Model Evaluation:** Assessing model performance using metrics like MAE, RMSE, and R-squared values.

**4.1. DATASET OVERVIEW**

The dataset comprises 384,977 rows and 22 columns, with each row representing a unique housing listing. The columns include both categorical and numerical features relevant to housing listings.

**4.1.1. KEY CHARACTERISTICS**

* **Geographical Data:** Includes latitude and longitude for geospatial analysis.
* **Price Distribution:** Varies significantly, with initial plots indicating potential outliers and skewness.
* **Feature Relationships:** Variables like "sqfeet," "beds," and "baths" have expected relationships with price.

**4.2. VARIABLES AND DEFINITIONS**

**List of Variables:**

* **id:** Unique identifier for each listing.
* **url:** URL of the listing.
* **region:** Geographical region of the listing.
* **region\_url:** URL for the region's Craigslist page.
* **price:** Listed price of the rental property.
* **type:** Type of rental property (e.g., apartment, condo).
* **sqfeet:** Square footage of the property.
* **beds:** Number of bedrooms.
* **baths:** Number of bathrooms.
* **cats\_allowed:** Indicates if cats are allowed (1 if true, 0 if false).
* **dogs\_allowed:** Indicates if dogs are allowed (1 if true, 0 if false).
* **smoking\_allowed:** Indicates if smoking is allowed (1 if true, 0 if false).
* **wheelchair\_access:** Indicates if the property has wheelchair access (1 if true, 0 if false).
* **electric\_vehicle\_charge:** Indicates if the property has an electric vehicle charging station (1 if true, 0 if false).
* **comes\_furnished:** Indicates if the property comes furnished (1 if true, 0 if false).
* **laundry\_options:** Describes the laundry facilities available.
* **parking\_options:** Describes the parking options available.
* **image\_url:** URL of an image of the property.
* **description:** Text description of the property.
* **lat:** Latitude of the property.
* **long:** Longitude of the property.
* **state:** State where the property is located.

**5. IDENTIFICATION OF THE BUSINESS PROBLEM**

**5.1. PROBLEM DEFINITION**

The primary business problem identified is the accurate prediction of property prices based on various features of the listings. This problem is crucial for real estate professionals, investors, and potential homeowners who need reliable estimates to make informed decisions.

**5.2. IMPORTANCE AND IMPACT**

Accurate property price predictions are essential for:

* **Sellers:** To set competitive and realistic listing prices.
* **Buyers:** To make informed purchasing decisions and identify undervalued properties.
* **Real Estate Agents:** To provide accurate market analyses and recommendations to clients.
* **Investors:** To identify lucrative investment opportunities.

If this problem remains unsolved, stakeholders may face financial losses, market inefficiencies, and poor decision-making, leading to suboptimal investments and purchases.

**6. DATA ANALYSIS AND FINDINGS**

**6.1. ANALYSIS**

The analysis involved univariate and bivariate analyses, including histograms, scatter plots, and box plots to explore data distributions and relationships. A heatmap and VIF analysis were conducted to assess multicollinearity.

**UNIVARIATE ANALYSIS**

1. **Histograms: To understand the distribution of individual variables.**

A graph of a distribution of price

Description automatically generated

The histogram visualizes the distribution of property prices in the dataset. Here are the key points to explain the histogram:

1. **X-axis (Price)**: The horizontal axis represents the property prices, ranging from 0 to 100,000.
2. **Y-axis (Count)**: The vertical axis shows the count of properties for each price range.
3. **Distribution Shape**:

* The histogram is highly skewed to the right (positively skewed).
* Most of the property prices are concentrated towards the lower end, close to 0.
* There is a very long tail extending towards higher prices, indicating that while most properties are relatively inexpensive, there are a few very high-priced properties in the dataset.

1. **Peaks**:

* There is a significant peak at the lower end of the price range, indicating a high frequency of properties priced between 0 and around 5,000.
* Beyond 5,000, the count of properties drops dramatically.

1. **Outliers**:

* The long tail suggests the presence of outliers, which are properties with exceptionally high prices compared to the majority.

**Implications for Analysis**

* **Skewness**: The right skewness suggests that the data is not normally distributed, which is a critical consideration for choosing appropriate statistical and machine learning methods.
* **Outliers**: The presence of outliers may affect the performance of models like linear regression. Methods that are robust to outliers or preprocessing steps to handle them may be necessary.
* **Data Transformation**: To reduce skewness, a log transformation or other normalization techniques may be applied to the price data for better modelling results.

This distribution indicates that while the majority of properties are affordable, a small number are significantly more expensive, reflecting a diverse range of property values in the dataset.

1. **Countplots: For categorical variables.**

A graph of different colored bars

Description automatically generated

The bar chart shows the distribution of houses based on the number of bedrooms they have. Here's a breakdown of what the plot might be telling:

**Number of Bedrooms on the X-Axis:**

This axis categorizes the houses based on the number of bedrooms they have. It likely starts from 0 bedrooms and goes up to a certain number (possibly 7 based on your previous description).

**Number of Houses on the Y-Axis:**

This axis shows the frequency or count of houses that fall under each bedroom category. The height of each bar represents the number of houses with that specific number of bedrooms.

**Observations from the Plot:**

**Most Common Bedrooms:** The bars for 1 and 2 bedrooms are likely the tallest, indicating that these are the most common number of bedrooms found in the dataset.

**Less Frequent Bedrooms:** As the number of bedrooms increases (moving to the right on the x-axis), the bars become shorter. This suggests that houses with a higher number of bedrooms (possibly 3 or more) are less frequent.

**Presence of Less Common Layouts:** Even though less frequent, there might be some houses with a very high number of bedrooms (up to 7 in your example). These bars would be the shortest on the right side of the plot.

**Overall Significance:**

This bar chart quickly summarizes the prevalence of different bedroom configurations within the housing data. It helps identify the typical number of bedrooms in the listings and provides insights into the distribution of houses across various bedroom categories.

**BIVARIATE ANALYSIS**

1. **Scatter Plots: To explore relationships between variables.**

A graph with blue dots

Description automatically generated

The above image appears to be a scatter plot of square footage and price, which means it shows the relationship between these two variables for individual houses. Each point on the plot represents a house listing. The x-axis shows the square footage of the house, and the y-axis shows the corresponding price of the house.

There is a positive correlation between square footage and price in this dataset. This means that as the square footage increases, the price of the house also tends to increase. This is likely because larger houses generally offer more space and amenities, which are desirable features that buyers are willing to pay more for.

Overall, this plot suggests that square footage is a significant factor that can influence housing prices. However, it is important to note that it is not the only factor. Other factors, such as the location, amenities, and condition of the house, will also play a role in determining the price.

1. **Box Plots: To identify outliers and understand variable distributions.**

A graph of a number of beds

Description automatically generated

The boxplot shows the variation in price by number of beds in listings. The x-axis shows the number of beds in a rental unit. The y-axis shows the price of the rental unit.

Here are some of the key points can learned from the boxplot:

* The middle line in the box is the median price. This means that half of the rentals with that number of beds are more expensive than the median price and half are less expensive.
* The box shows the interquartile range (IQR). This is the range that contains the middle 50% of the prices. For example, in the boxplot for 0 bed rentals, the IQR is between $1200 and $1800. This means that half of the 0-bedroom rentals cost between $1200 and $1800.
* The lines extending from the box are called whiskers. The whiskers show the lowest and highest prices that are within 1.5 times the IQR from the median. Prices that fall outside the whiskers are considered outliers.

The boxplot shows that the price of rentals generally increases as the number of bedrooms increases. There is also more variation in price for rentals with more bedrooms. For example, the IQR for 3-bedroom rentals is much larger than the IQR for 0-bedroom rentals. This means that there is a wider range of prices for 3-bedroom rentals than for 0-bedroom rentals.

**CORRELATION ANALYSIS**

1. **Pairplot**

A graph of data on a white background

Description automatically generated

The pair plot visualizes a collection of scatter plots that show the relationship between each numeric variable in a dataset. This can be a useful tool for exploring relationships between features in the data.

Let's take a look at a few of the panels:

* **price vs sqfeet:** This scatter plot shows a positive correlation between the square footage and the price of the rental unit. This means that as the square footage increases, the price tends to increase as well.
* **price vs beds:** This scatter plot shows a positive correlation between the number of bedrooms and the price of the rental unit. This means that listings with more bedrooms tend to be more expensive than listings with fewer bedrooms.
* **baths vs sqfeet:** This scatter plot also shows a positive correlation, which means that listings with more bathrooms tend to have a larger square footage.

It is important to note that correlation does not necessarily equal causation. There may be other factors that influence the price of a rental unit besides square footage and number of bedrooms.

1. **Heatmap**

A screenshot of a graph

Description automatically generated

A correlation heatmap was created to identify relationships between variables, followed by calculating the Variance Inflation Factor (VIF) to assess multicollinearity.

The heatmap appears to show the correlation between various features in a dataset of rental listings. Each square in the heatmap represents the correlation between two features. A darker shade of red indicates a stronger positive correlation, and a darker shade of blue indicates a stronger negative correlation. A white square indicates that there is no correlation between the two features.

Here are some of the interesting correlations that can be seen in the heatmap:

* There is a positive correlation between the price of a rental unit and the square footage, number of bedrooms, and number of bathrooms. This means that listings with a larger square footage, more bedrooms, and more bathrooms tend to be more expensive.
* There is a positive correlation between the square footage and the number of bedrooms and bathrooms. This means that listings with larger square footage tend to also have more bedrooms and bathrooms.
* There is a positive correlation between beds and baths. This means that listings with more bedrooms tend to also have more bathrooms.
* There is a weak positive correlation between price and laundry options. This means that listings that mention laundry options in the description tend to be slightly more expensive.

It is important to note that correlation does not necessarily equal causation. There may be other factors that influence the price of a rental unit besides square footage and number of bedrooms.

**6.2. FINDINGS**

* The Random Forest model outperformed Linear Regression but was surpassed by the Gradient Boosting model, which showed the best performance with the lowest MAE and MSE and the highest R-squared value.
* Significant variations in property prices across different regions.
* Strong correlations between price and features like square footage, number of bedrooms, and number of bathrooms.
* Identification of key drivers of property prices.

**7.1. SOLUTION APPROACH**

To solve the business problem of predicting property prices accurately, we utilized various machine learning models and evaluated their performance. The models used include Linear Regression, Random Forest, and Gradient Boosting. Here’s a breakdown of the approach and underlying theory:

1. **Linear Regression:** Basic predictive model.
2. **Random Forest:** Ensemble method improving prediction accuracy.
3. **Gradient Boosting:** Advanced ensemble technique yielding the best performance.

**Model Evaluation Metrics:**

1. **Linear Regression:** MAE: 515.45, MSE: 1008516.36, R-squared: -3.66

This model performs very poorly, with high MAE and MSE values and a negative R². It suggests that a linear relationship is not suitable for this dataset.

1. **Random Forest:** MAE: 12.30, MSE: 35929.29, R-squared: 0.83

This model performs much better, with a relatively low MAE and MSE and a high R² value, indicating a good fit and reasonable prediction accuracy.

1. **Gradient Boosting:** MAE: 6.35, MSE: 261.15, R-squared: 0.99

This model performs the best among the three, with the lowest MAE and MSE and an R² value very close to 1. It suggests that Gradient Boosting is highly effective for this dataset and provides very accurate predictions.

In conclusion, the Gradient Boosting model demonstrated superior performance with the lowest MAE and MSE, and the highest R² value, indicating it is the most accurate and reliable model.

**Actionable Recommendations**: Based on the analysis, the Gradient Boosting model was selected for its exceptional performance. This model was further tested and validated to ensure robustness and reliability.

**7.2. RECOMMENDED ACTIONS**

**Implement Gradient Boosting Model**:

* Deploy the Gradient Boosting model for property price prediction.
* Integrate the model into the organization's decision-making processes for buying, selling, and investing in properties.
* Develop a user-friendly interface to make the model accessible to stakeholders such as real estate professionals, investors, and homeowners.

**Regular Model Updates**:

* Continuously update and retrain the model with new data to maintain accuracy.
* Monitor the model's performance regularly and make necessary adjustments.

**Data-Driven Decision Making**:

* Use the model's predictions to inform pricing strategies, investment decisions, and market analysis.
* Leverage insights from the model to identify market trends and opportunities.

**7.3. POTENTIAL BENEFITS AND RISKS**

**Potential Benefits**:

* **Improved Accuracy**: The Gradient Boosting model provides highly accurate property price predictions, enhancing decision-making capabilities.
* **Informed Strategies**: Stakeholders can make better-informed decisions regarding property investments, sales, and purchases.
* **Competitive Advantage**: Access to precise property price predictions can give the organization a competitive edge in the real estate market.

**Risks and Mitigation Strategies**:

* **Model Overfitting**: There is a risk of the model overfitting the training data. This can be mitigated by using cross-validation techniques and regularizing the model.
* **Market Changes**: The model relies on historical data, which may not always account for future market changes. Continuous updates and monitoring can help mitigate this risk.
* **Data Quality**: The accuracy of predictions is dependent on the quality of data. Ensuring high-quality, up-to-date data can mitigate issues related to data quality.
* **Implementation Challenges**: Integrating the model into existing systems may present challenges. Providing adequate training and resources can help overcome these challenges.

By implementing the Gradient Boosting model and following the recommended actions, the organization can significantly enhance its property price prediction capabilities, leading to better decision-making and improved business outcomes.

**8. RESOURCE ALLOCATION**

To implement the proposed solution for property price prediction using the Gradient Boosting model, the following resources are required:

**Personnel**:

* **Data Scientists**: Experts to fine-tune the model, handle data preprocessing, and ensure the model's accuracy and robustness.
* **Data Engineers**: Professionals to manage data collection, cleaning, and integration from various sources.
* **Software Developers**: Developers to create and maintain the user interface and integrate the model into existing systems.
* **Project Manager**: A manager to oversee the implementation process, coordinate between teams, and ensure timely delivery of the project.
* **IT Support**: IT personnel to provide infrastructure support and ensure the deployment environment is stable and secure.

**Technology**:

* **Computational Resources**: High-performance servers or cloud services (e.g., AWS, Google Cloud, Azure) to handle model training and deployment.
* **Software Tools**:
  + Development environment (e.g., Jupyter, PyCharm).
  + Machine learning libraries (e.g., scikit-learn, XGBoost, TensorFlow).
  + Data processing tools (e.g., Pandas, NumPy).
* **Data Storage**: Databases to store historical and new data (e.g., SQL, NoSQL databases).

**Budget**:

* **Personnel Costs**: Salaries for the project team (data scientists, data engineers, software developers, project manager, IT support).
* **Technology Costs**: Costs for computational resources, software licenses, cloud services, and data storage.
* **Operational Costs**: Expenses related to ongoing maintenance, updates, and support.
* **Training and Development**: Budget for training personnel to use new tools and technologies effectively.

**External Support and Collaborations**:

* **Consulting Services**: Collaboration with consulting firms or experts for advanced model tuning and validation.
* **Data Providers**: Partnerships with real estate data providers for access to high-quality and up-to-date data.
* **Academic Institutions**: Collaboration with universities for research and development support, and to stay updated with the latest advancements in machine learning and data science.

Implement the Gradient Boosting model for predicting housing prices due to its superior accuracy. Regularly update the model with new data to maintain its performance.

**9. CONCLUSION**

In this report, we explored various machine learning models to predict property prices accurately. The key findings are as follows:

* **Model Performance**:
  + **Linear Regression** performed poorly with high errors and a negative R² value.
  + **Random Forest** showed better performance with significantly lower errors and a high R² value.
  + **Gradient Boosting** outperformed both models, with the lowest errors and an R² value close to 1, indicating a nearly perfect fit.
* **Proposed Solution**:
  + Based on the performance metrics, the Gradient Boosting model was selected as the most accurate and reliable model for predicting property prices.
  + Actionable recommendations include deploying the Gradient Boosting model, regularly updating it with new data, and integrating it into decision-making processes.
* **Potential Benefits**:
  + **Improved Accuracy**: The Gradient Boosting model provides highly accurate predictions, enhancing decision-making capabilities.
  + **Informed Strategies**: Stakeholders can make better-informed decisions regarding property investments, sales, and purchases.
  + **Competitive Advantage**: Accurate property price predictions give the organization a competitive edge in the real estate market.

This analysis contributes significantly to addressing the business problem of property price prediction. By leveraging advanced machine learning techniques, the organization can enhance its predictive capabilities, leading to better business outcomes and a stronger market position.

**10. REFERENCES**

<https://www.kaggle.com/datasets/austinreese/usa-housing-listings/data>

**11. CONTRIBUTION**

1. Udaya Lakshmi Punnam: Data preprocessing, initial analysis, report writing.
2. Keerthi Pulipaka: Model development, evaluation, report writing.
3. Deepthi Poosala: Data visualization, report writing.
4. Chandra Sidhartha Kodela: Model validation, report writing.