Final Paper

Housing Prediction

Members: Saniah Safat, Rachana Pandey

UTA ID: 1001863085, 1001863076

Abstract

The study of “Housing Prediction” tackles the complexities of predicting housing prices using a detailed dataset from Kaggle, which encompasses a wide range of house characteristics including size, location, and amenities. Our research employs a systematic approach involving Exploratory Data Analysis (EDA), Multiple Linear Regression, Logistic Regression, and Analysis of Variance (ANOVA) to discern the key factors influencing housing prices and to categorize homes based on their value. The project began with EDA, which was instrumental in uncovering underlying patterns and outliers, thereby shaping our modeling strategy. Multiple Linear Regression revealed that variables such as area, number of bathrooms, and amenities like air conditioning significantly affect prices. Using Logistic Regression, we successfully classified houses into "high-value" and "low-value" categories, achieving an accuracy of 87%, with commendable precision and recall measures. Additionally, ANOVA confirmed the substantial impact of features such as air conditioning and hot water heating on house valuations. The insights derived from this study not only deepen the academic understanding of real estate economics but also provides actionable information for buyers, investors, and policymakers, thus facilitating more informed decision-making in the housing market.

Introduction

In the realm of real estate, accurately predicting housing prices is crucial for a range of stakeholders, including home buyers, investors, and policymakers. We chose this project because we realised that precise price predictions enhance decision-making, optimized investment returns, and supports effective policy development when it comes to buying a house. The complexity of the real estate market, shaped by factors such as property size, location, amenities, and economic conditions, present significant challenges. As a result, we turned to sophisticated statistical techniques to analyze extensive housing data.

For this study, we utilized a dataset from Kaggle, which included detailed information on house characteristics known to influence market prices. This dataset not only covered physical attributes such as area and number of bedrooms but also included locational benefits and available amenities. Despite the rich data, extracting actionable insights that could accurately predict prices and classify houses based on value remained a formidable task. Our project aimed to first conduct exploratory data analysis (EDA) to identify key variables and understand the underlying data structure. We then developed predictive models through multiple linear regression and logistic regression to accurately forecast housing prices and categorize homes into

distinct value-based classes. Furthermore, we applied Analysis of Variance (ANOVA) to assess the impact of specific amenities on house prices.

By pursuing these objectives, we strove to contribute valuable insights to the field of real estate analytics, potentially influencing future market predictions and strategic decision-making within the housing sector.

Problem Statement as Research Questions

1) What factors most significantly influence the selling prices of houses in the dataset?

   This question seeks to identify and quantify the relationships between various house characteristics (such as area, number of bedrooms, and location) and the house prices. Understanding these factors will enable more precise price predictions and insights into market trends.

2) Can we accurately classify houses as either “high-value” or “low-value” based on a set of features such as size, location, and number of bedrooms?

   This question addresses the development of a classification model that distinguishes between high-value and low-value properties. The aim is to provide a tool for stakeholders to easily categorize homes based on their market value, enhancing decision-making processes in buying, selling, and investing.

3) Does the presence of amenities like air conditioning and hot water systems significantly increase the value of houses?

   The focus here is on determining the impact of specific amenities on house prices. By establishing whether features like air conditioning and hot water heating contribute to higher property values, stakeholders can better understand how enhancements and upgrades could affect their investments.

Proposed Methodology

The methodology for this study was designed to address the research questions effectively and involved several distinct phases, each tailored to explore various aspects of the housing data:

Exploratory Data Analysis (EDA):

Objective: To gain an initial understanding of the dataset, identify any patterns, outliers, and the underlying distribution of the variables.

Tools and Techniques: We employed visualization tools such as histograms, box plots, strip plot and violon plots. Histograms provided initial insights into the distribution of data. Violin plots were used to examine the data density and potential outliers more closely, and strip plots allowed

for the visualization of individual data points, helping to highlight any anomalies. Lastly, box plots were helped confirm the presence of outliers, ensuring a thorough understanding of data variance and extremities. We also checked for missing data and had no missing values. But we had a number of categorical variables that we had to convert to dummy variables.

Multiple Linear Regression:

Objective: To quantify the influence of multiple house features on the selling prices.

Procedure: After ensuring data normalization and handling multicollinearity, we constructed a multiple linear regression model. We selected variables based on their correlation with the house price and their practical significance in real estate valuation. The model's goodness of fit was evaluated using R-squared and Adjusted R-squared values. Further, the significance of each predictor was assessed using p-values and confidence intervals.

Logistic Regression for Classification:

Objective: To classify houses into "high-value" and "low-value" categories based on selected features.

Procedure: The dataset was divided into two categories based on a fixed value of 5 million. A logistic regression model was then developed to predict the probability of a house falling into either category. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. Validation techniques such as cross-validation were used to ensure the model's robustness and to prevent overfitting.

Analysis of Variance (ANOVA):

Objective: To examine if amenities such as air conditioning and hot water heating significantly affect the house prices.

Procedure: ANOVA was conducted to compare the mean prices of houses with and without specific amenities. We checked for the assumptions of ANOVA, including homogeneity of variances and normality of residuals.

Analysis and Results

Multiple Linear Regression Findings:

The regression analysis revealed that the model's R-squared value is 68.2%, indicating that approximately 68.2% of the variation in house prices can be explained by the included variables. Key predictors significantly impacting house prices include area (positive effect), bathrooms, stories, main road access, presence of a guestroom, basement, air conditioning, and parking facilities. Preferred location and furnishing status (furnished and semi-furnished) also substantially increase prices. Notably, the number of bedrooms did not significantly affect the pricing, contradicting common real estate assumptions.

Logistic Regression Performance:

The classification model achieved an accuracy of 83%, with a precision of 85% for low-value homes and 79% for high-value homes. Recall rates stood at 85% for low-value and 79% for high-value homes, respectively. The F1-scores were 0.85 for low-value and 0.79 for high-value homes, indicating a balanced precision-recall for both classes. These metrics demonstrate the model's effectiveness in reliably classifying houses based on their market value.

ANOVA on Amenities:

The ANOVA test highlighted the significant impact of amenities on house prices. Specifically, air conditioning significantly increased property values with a sum of squares at 16.695583, an F-value of 155.867454, and a p-value nearing zero (1.284855e-31). Hot water heating also showed a notable effect, though less pronounced than air conditioning, with a sum of squares at 1.682191, an F-value of 15.704680, and a p-value of 8.394876e-05. This analysis confirms the hypothesis that certain amenities like air conditioning and hot water heating are crucial for increasing house values.

These findings directly contribute to our understanding of the housing market by confirming which features and amenities significantly affect house prices and how properties can be segmented into value-based categories. This provides stakeholders with valuable insights into factors influencing housing investments and assists in decision-making processes related to property sales and purchases.

Conclusions

The study successfully identified and quantified the factors that significantly influence house prices and provided a reliable method for classifying houses based on their market value. Our findings reveal that area, number of bathrooms, amenities such as air conditioning and hot water heating, and the presence of features like basements and guestrooms are pivotal in determining housing prices. Notably, the number of bedrooms was found to have no significant impact on price, challenging conventional real estate wisdom. The logistic regression model proved effective in classifying homes into "high-value" and "low-value" categories with considerable accuracy and balance between precision and recall. ANOVA tests further confirmed the significant value addition of amenities like air conditioning to property prices. This comprehensive analysis not only advances our understanding of the real estate market dynamics but also offers practical guidance for investors, policymakers, and buyers in making informed decisions in the housing sector.

Lessons that we have learned:

Throughout this project and the data mining class, we gained invaluable insights into the complexities of data analysis. We learned the importance of thorough exploratory data analysis in

identifying key variables, missing data and outliers, which are crucial for building accurate predictive models. The project highlighted the significance of choosing the right statistical methods, like multiple linear regression and logistic regression, to address specific research questions effectively. We also realized the impact of data quality on model outcomes and learned various techniques for data cleaning and preparation to ensure robust analysis. Additionally, the project emphasized the practical implications of data-driven insights in real-world decision-making, underscoring the critical role of analytical skills in shaping strategic business outcomes. This experience has increased our analytical thinking and problem-solving skills, preparing us to tackle complex data challenges in our future careers.

Bibliography

Badole, M. (2024, January 17). *Multiple linear regression: Definition , example and applications*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

Bevans, R. (2023a, June 22). *Multiple linear regression: A quick guide (examples)*. Scribbr. https://www.scribbr.com/statistics/multiple-linear-regression/

Bevans, R. (2023b, June 22). *One-way ANOVA: When and how to use it (with examples)*. Scribbr. https://www.scribbr.com/statistics/one-way-anova/

KUMARdatalab, H. (2023, July 7). *Housing price prediction*. Kaggle. https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction

Sklearn.linear_model.logisticregression. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Appendix

Output Results:

EDA:

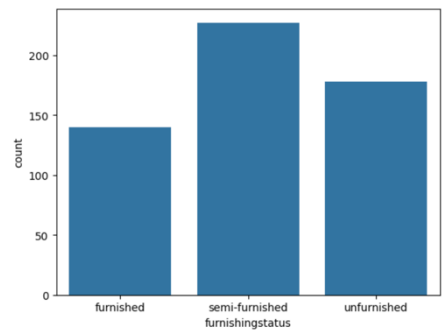
	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

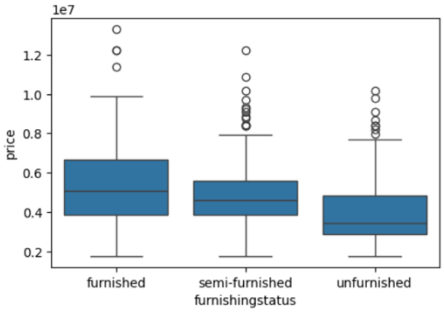
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   price             545 non-null    int64 
 1   area              545 non-null    int64 
 2   bedrooms          545 non-null    int64 
 3   bathrooms         545 non-null    int64 
 4   stories           545 non-null    int64 
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64 
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
```

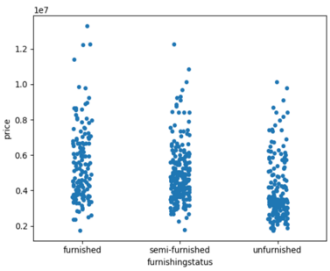
```
price               0
area                0
bedrooms            0
bathrooms           0
stories             0
mainroad            0
guestroom           0
basement            0
hotwaterheating     0
airconditioning     0
parking             0
prefarea            0
furnishingstatus    0
dtype: int64
```

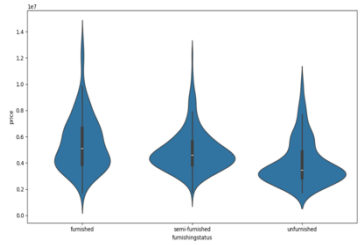
```
#Checking for duplicates in our data
data = df_housing.drop_duplicates(subset ="furnishingstatus",)
data
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
7	10150000	16200	5	3	2	yes	no	no	no	no	0	no	unfurnished









Multiple Linear Regression Output:

ıroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus_furnished	furnishingstatus_semi-furnished	furnishingstatus_unfurnished
1	0	0	0	1	2	1	1	0	0
1	0	0	0	1	3	0	1	0	0
1	0	1	0	0	2	1	0	1	0
1	0	1	0	1	3	1	1	0	0
1	1	1	0	1	2	0	1	0	0

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.682
Model:                            OLS   Adj. R-squared:                  0.674
Method:                 Least Squares   F-statistic:                     87.52
Date:                Mon, 22 Apr 2024   Prob (F-statistic):          9.07e-123
Time:                        15:46:22   Log-Likelihood:                -3311.9
No. Observations:                 545   AIC:                             6652.
Df Residuals:                     531   BIC:                             6712.
Df Model:                          13
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -36.8463     23.781     -1.549      0.122     -83.562       9.869
area                             0.0244      0.002     10.052      0.000       0.020       0.029
bedrooms                        11.4788      7.260      1.581      0.114      -2.783      25.740
bathrooms                       98.7668     10.336      9.555      0.000      78.462     119.072
stories                         45.0848      6.417      7.026      0.000      32.479      57.690
mainroad                        42.1273     14.222      2.962      0.003      14.188      70.066
guestroom                       30.0526     13.171      2.282      0.023       4.179      55.926
basement                        35.0107     11.028      3.175      0.002      13.346      56.675
hotwaterheating                 85.5447     22.315      3.833      0.000      41.708     129.382
airconditioning                 86.4958     10.835      7.983      0.000      65.210     107.781
parking                         27.7107      5.853      4.735      0.000      16.214      39.208
prefarea                        65.1544     11.568      5.632      0.000      42.429      87.879
furnishingstatus_furnished      41.1234     12.621      3.258      0.001      16.330      65.917
furnishingstatus_semi-furnished 36.4890     10.919      3.342      0.001      15.040      57.938
==============================================================================
Omnibus:                       97.909   Durbin-Watson:                   1.234
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              258.281
Skew:                           0.895   Prob(JB):                     8.22e-57
Kurtosis:                       5.859   Cond. No.                     3.09e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.09e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Logistic Regression:

```
df_housing['Value'] = df_housing['price'].apply(lambda x: 1 if x >= 5000000 else 0)
df_housing
```

guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus_furnished	furnishingstatus_semi-furnished	furnishingstatus_unfurnished	Value
0	0	0	1	2	1	1	0	0	1
0	0	0	1	3	0	1	0	0	1
0	1	0	0	2	1	0	1	0	1
0	1	0	1	3	1	1	0	0	1
1	1	0	1	2	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...
0	1	0	0	2	0	0	0	1	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0

	precision	recall	f1-score	support
0	0.85	0.85	0.85	129
1	0.79	0.79	0.79	89
accuracy			0.83	218
macro avg	0.82	0.82	0.82	218
weighted avg	0.83	0.83	0.83	218

ANOVA

	sum_sq	df	F	PR(>F)
C(airconditioning)	16.695583	1.0	155.867454	1.284855e-31
C(hotwaterheating)	1.682191	1.0	15.704680	8.394876e-05
Residual	58.055775	542.0	NaN	NaN