Predicting High-Priced Rental Apartments in the USA

Table of contents

Milestone 1 – Group_26

Summary

This project investigates whether it is possible to predict when an apartment listing in the United States is **high-priced relative to other listings in the same state**. Using the

Apartment for Rent Classified dataset from the UCI Machine Learning Repository, we defined a binary target indicating whether each listing's price was above the state-level median. We developed a logistic regression model to classify listings as either high-priced or not. The final classifier achieved an accuracy of 0.7 on unseen data, indicating that it correctly predicts a listing as high-priced in most cases.

The notebook includes: - Data loading from the web
- Data cleaning and wrangling
- Exploratory data analysis
- Visualizations
- Classification modeling
- Results, discussion, and conclusions

Introduction

Homeownership in the United States has become somewhat less attainable for many Americans due to increasing housing prices, rising mortage rates and other increases in the cost of living. As a result, according to recent surveys, around 84% of Gen Z adults report delaying major life milestones in order to afford a home, with many turning to long-term renting instead. Nearly 75% of Gen Z say they prefer renting to owning, which is important to note as this generation represents more than 20% of the U.S. population. Growing rental demand is driving new construction as developers across the country are expected to add more than 500,000 new apartment units across the country.

These trends lead us to wonder if apartment rental prices vary dramatically across the U.S. A price considered as "expensive" in Texas may be considered "cheap" in New York. As many Gen Z adults are relocating to higher-cost states and more than 75% of them rent, understanding these price differences is increasingly important. Additionally, rental prices are influenced by a variety of listing features. For example, apartments with more bedrooms or bathrooms are generally associated with higher-prices. Other features such as whether pets are allowed or whether the apartment includes additional fees, may also signal higher quality or prices. These features will be incorporated into our model to assess their impact on whether an apartment is classified as high-priced or not. To create a meaningful comparison across regions, we evaluated whether each apartment is **high-priced relative to the median rent within its own state**.

Research Question

Can a machine learning algorithm accurately predict whether an apartment listing is high-priced relative to the median rental price in its state, using features such as square footage, number of bedrooms and bathrooms, and various listing attributes?

Dataset

We use the **Apartment for Rent Classified** dataset from the UCI Machine Learning Repository. The dataset contains: - 10,000+ apartment listings
- Structural and listing details
- Geographic information
- Price and square footage

Data Cleaning & Wrangling

We prepared the dataset by:

1. Loading the data from the UCI ML Repository
2. Selecting relevant columns

3. Removing rows with missing or invalid values

4. Computing the median rental price for each state

5. Creating a binary target variable `high_price`

	id	category	title	body
0	5668640009	housing/rent/apartment	One BR 507 & 509 Esplanade	This unit is located at 507 &
1	5668639818	housing/rent/apartment	Three BR 146 Lochview Drive	This unit is located at 146 Lo
2	5668639686	housing/rent/apartment	Three BR 3101 Morningside Drive	This unit is located at 3101 M
3	5668639659	housing/rent/apartment	Two BR 209 Aegean Way	This unit is located at 209 Ae
4	5668639374	housing/rent/apartment	One BR 4805 Marquette NE	This unit is located at 4805 M

	price	square_feet	bathrooms	bedrooms	state	pets_allowed	fee	has_photo	state_median_p
0	2195.0	542.0	1.0	1.0	CA	Cats	No	Thumbnail	2155.0
1	1250.0	1500.0	1.5	3.0	VA	Cats,Dogs	No	Thumbnail	1398.0
2	1600.0	820.0	1.0	2.0	CA	Cats,Dogs	No	Thumbnail	2155.0
3	975.0	624.0	1.0	1.0	NM	Cats,Dogs	No	Thumbnail	1012.5
4	1250.0	965.0	1.5	2.0	NM	Cats,Dogs	No	Thumbnail	1012.5

```
Target distribution check passed? True
Outliers detected in column 'price':
        price
13     3250.0
20     4950.0
27     3580.0
62     3200.0
63     3399.0
...       ...
35521  3140.0
35603  2902.0
35638  3275.0
35649  3850.0
35666  3002.0

[1713 rows x 1 columns]
----------------------------------------
Outliers detected in column 'square_feet':
       square_feet
27          2100.0
56          1850.0
63          1721.0
64          2598.0
65          2200.0
...            ...
35369       2756.0
35387       2196.0
35392       2080.0
35470       2160.0
35652       1900.0

[846 rows x 1 columns]
----------------------------------------
Outliers detected in column 'bathrooms':
       bathrooms
636          4.5
748          4.0
1014         4.0
1314         5.0
1324         4.5
1452         4.0
1547         4.0
1574         4.0
```

```
1692         4.0
1812         4.0
1835         5.0
1840         4.0
1876         4.0
2029         4.0
2130         7.0
2180         4.0
2740         4.5
2870         6.0
3064         4.0
3439         4.0
3470         4.0
3747         4.0
3789         4.0
3796         5.0
3936         4.5
3951         4.0
4150         4.0
4247         5.0
4261         4.0
4393         4.0
4466         4.0
4506         4.0
4520         4.5
9431         4.0
9836         4.0
10836        4.5
16376        5.0
17168        4.0
17977        4.0
23571        4.0
26986        4.0
29979        4.0
31518        4.0
31682        5.0
32450        4.0
34767        4.5
----------------------------------------
Outliers detected in column 'bedrooms':
       bedrooms
27          4.0
33          4.0
```

```
62          4.0
64          4.0
65          4.0
...         ...
35387       4.0
35392       4.0
35407       4.0
35436       4.0
35470       4.0

[549 rows x 1 columns]
----------------------------------------
Some checks failed:
    schema_context column                                  check  \
0  DataFrameSchema   None  Outliers detected in numeric columns.

   check_number  failure_case index
0             2         False  None

(28560, 7140)
```

Exploratory Data Analysis (EDA)

We explored:

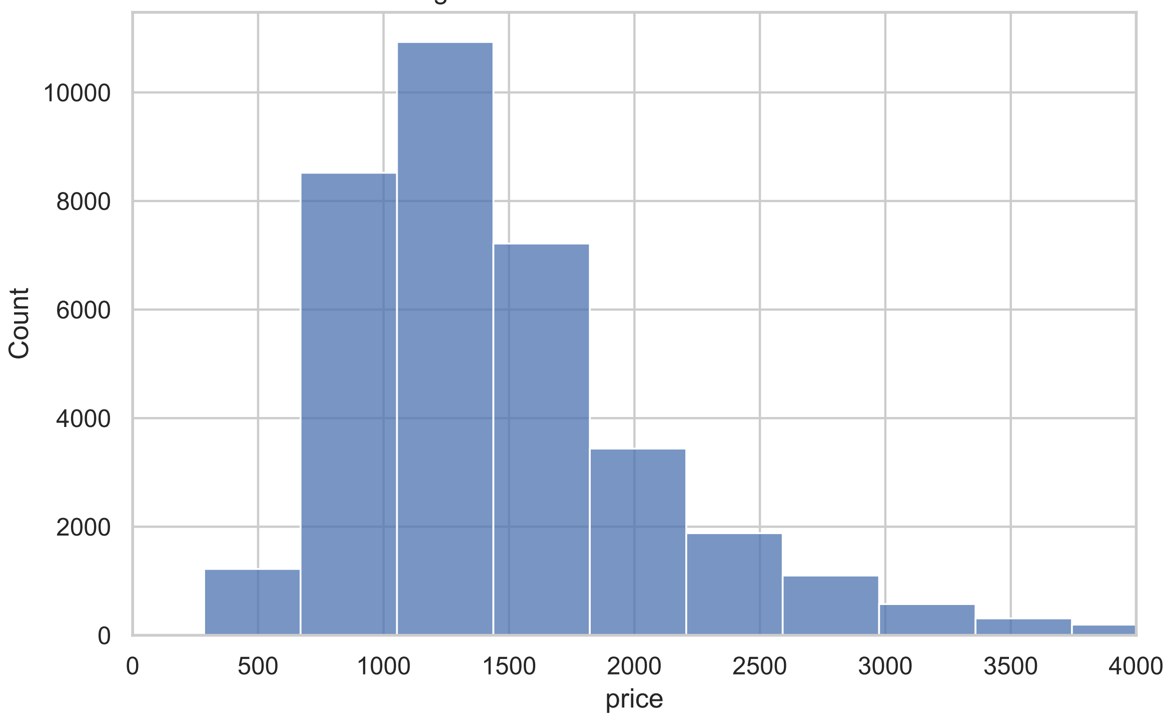
- Summary statistics
- Class balance
- Price distribution
- Relationship between size and price

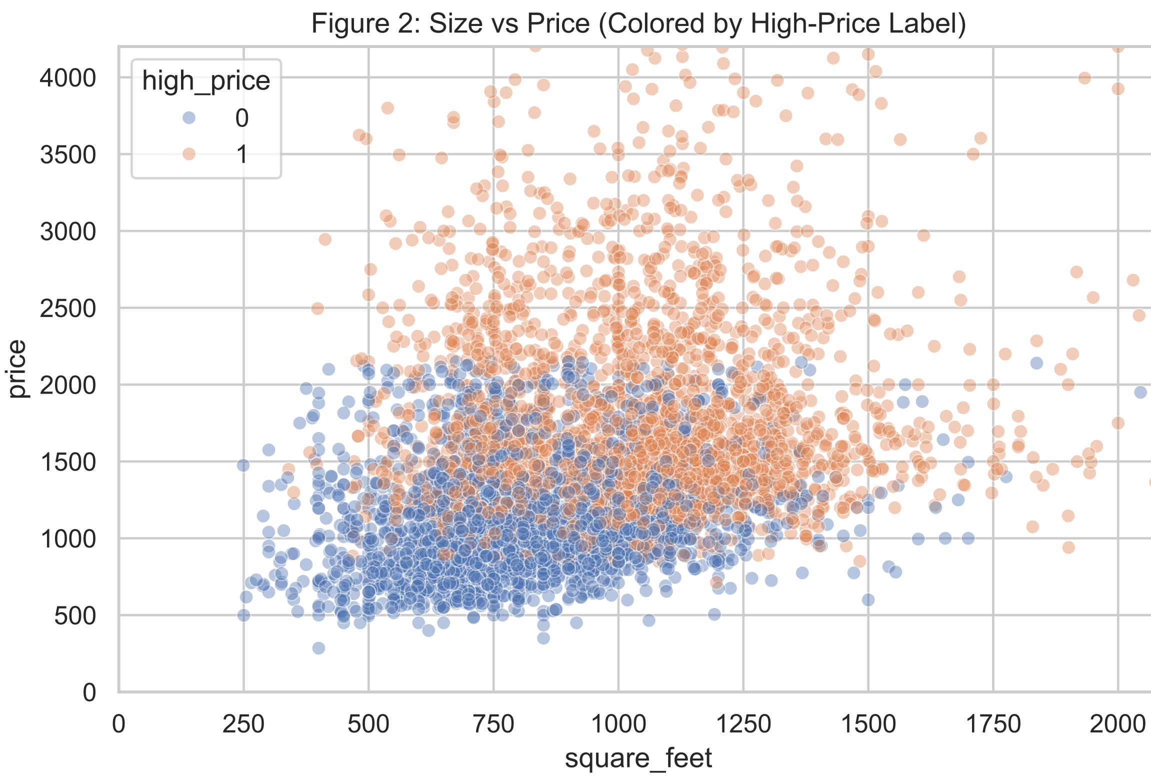
	price	square_feet	bathrooms
count	35700.000000	35700.000000	35700.000000
mean	1487.928235	940.632157	1.440294
std	722.414581	338.539094	0.533467
min	285.000000	200.000000	1.000000
25%	1025.000000	720.750000	1.000000
50%	1327.000000	897.000000	1.000000

	price	square_feet	bathrooms
75%	1745.000000	1105.000000	2.000000
max	19500.000000	12000.000000	7.000000

```
high_price
1    0.510728
0    0.489272
Name: proportion, dtype: float64
```

Figure 1: Distribution of Rental Prices





Modeling: Logistic Regression Classifier

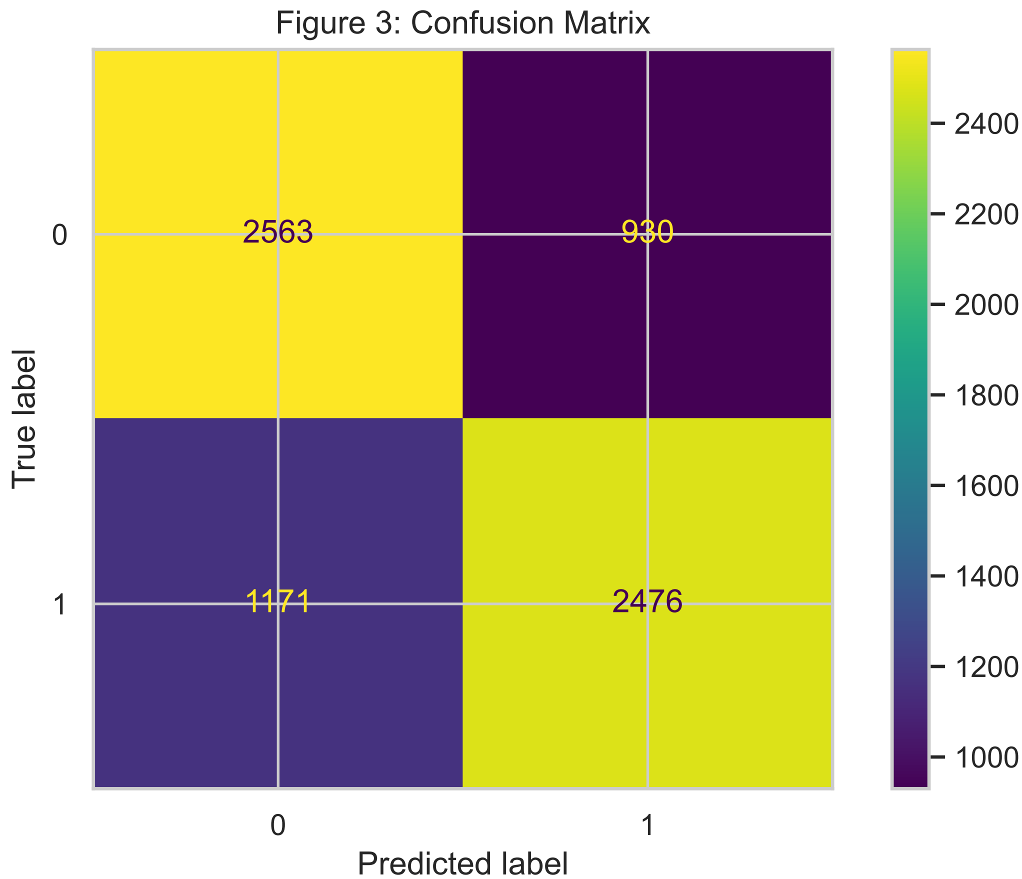
We predict `high_price` using:

Numeric features - square_feet
- bathrooms

Categorical features - bedrooms
- state
- pets_allowed
- fee
- has_photo

We use a train-test split and a preprocessing pipeline.

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.705742	0.726952	0.678914	0.702113



Discussion

Summary of Findings

The logistic regression classifier performed reasonably well in predicting whether an apartment is high-priced relative to its state's median rental price. Our logistic regression model achieved an accuracy of approximately 0.7, indicating that the model correctly classified 70% of apartment listings as either high-priced or not relative to their state's median rent. The model has a precision score of 0.715 which means the model was correct about 71% of the time when predicting an apartment to be high-priced. The recall score of approximately 0.660 shows the model was able to correctly identify about 66% of actually high-priced listings. Combined, these values resulted in a F1-score of 0.686.

- Overall, these metrics suggest that the model provides a reasonably balanced performance, though there is room for improvement in correctly identifying all high-priced apartments.

Analysis

The model performed well when identifying low-priced listings as indicated by the large number of true negatives (2,893). However, it was slightly less successful at identifying high-priced listings misclassifying 1,323 of them as not high-priced when they in fact were. These results align reasonably well with our expectations. We anticipated that apartment prices would vary across the country and the model's high precision of 0.72 suggests that it is usually correct when predicting a listing as high-priced, supporting the idea that certain features contribute meaningfully to the classification. The F1-score of ~0.69 indicates that the model is not overfitting to one class and maintains a balanced trade-off between identifying high-priced listings and avoiding false positives.

Overall, the results suggest that rental price classification at the state level is predictable to a moderate degree, but not with perfect accuracy. This is likely due to the variability of housing markets across the country. The model performs as expected for a logistic regression approach and provides a useful baseline for predicting whether a listing is high-priced relative to its state's median rent.

Limitations and Future Work

A key limitation of this analysis is that logistic regression may be too simple to capture the full complexity of rental markets, which vary widely across and within states. The model's moderate accuracy (0.70) and recall (0.66) indicate that many high priced listings remain misclassified, suggesting that important predictive features were not included.

Future work could expand this analysis by incorporating more advanced models such as random forests, as well as extracting additional insights through NLP features derived from apartment descriptions. It may also be valuable to explore state by state differences in greater depth and to perform regression modelling to predict exact prices rather than broader categories. Further improvements could come from adding location specific features like the neighbourhood characteristics, building age, or available amenities which may contribute more meaningfully to overall model performance.

References

1. Investopedia. (2023). Gen Z is having more trouble affording a home — How some are achieving homeownership. https://www.investopedia.com/gen-z-is-having-more-trouble-affording-a-home-how-some-are-achieving-homeownership-11826137

2. Newsweek. (2023). Gen Z is renting, not buying: What it means for the country's future. https://www.newsweek.com/gen-z-renting-not-buying-what-means-country-future-2120726

3. PR Newswire. (2023). Top 10 states to which Gen Zers are moving and the states they are leaving. https://www.prnewswire.com/news-releases/top-10-states-to-which-gen-zers-are-moving-and-the-states-they-are-leaving-302058380.html

4. Starmer, J. (n.d.). Classification metrics educational videos [YouTube channel]. StatQuest. https://www.youtube.com/user/joshstarmer

5. Scikit-Learn. (n.d.). Logistic regression & preprocessing. https://scikit-learn.org/stable/modules/linear_r
regression

6. UCI Machine Learning Repository. (n.d.). Apartment for rent classified dataset. https://archive.ics.uci.edu/ml/datasets/Apartment+for+Rent+Classified