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	bo
0	5668640009	housing/rent/apartment	One BR 507 & 509 Esplanade	Th
1	5668639818	housing/rent/apartment	Three BR 146 Lochview Drive	Th
2	5668639686	housing/rent/apartment	Three BR 3101 Morningside Drive	Th
3	5668639659	housing/rent/apartment	Two BR 209 Aegean Way	Th

	id	category	title	bo
4	5668639374	housing/rent/apartment	One BR 4805 Marquette NE	Th
...	...	...	...	...
99821	5121219946	housing/rent/apartment	Houston - superb Apartment nearby fine dining	Re
99822	5121219696	housing/rent/apartment	The Best of the Best in the City of Jacksonvil...	Co
99823	5121219420	housing/rent/apartment	A great & large One BR apartment. Pet OK!	Fu
99824	5121218935	housing/rent/apartment	The Crest offers studio, 1, 2 & Three BR homes...	An
99825	5121218844	housing/rent/apartment	Large Remodeled Two BR 1. Five BA Apartment Home	Th

Table 2: Cleaned and feature-engineered apartment listings used for modeling. The tab includes selected numerical and categorical predictors (price, square footage, batl rooms, bedrooms, state, pet policy, fee status, and photo availability) after removin missing, non-positive, and duplicate records. A state-level median price is compute and used to derive the binary target variable high_price, indicating whether a listin is priced above the median for its state.

	price	square_feet	bathrooms	bedrooms	state	pets_allowed	fee	has_photo	state_media
0	2195.0	542.0	1.0	1.0	CA	Cats	No	Thumbnail	2195.0
1	1250.0	1500.0	1.5	3.0	VA	Cats,Dogs	No	Thumbnail	1420.0
2	1600.0	820.0	1.0	2.0	CA	Cats,Dogs	No	Thumbnail	2195.0
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
...	...	...	...	...	...	...	...	...	...
35695	1314.0	1000.0	2.0	2.0	NC	Cats,Dogs	No	Yes	1126.5
35696	685.0	625.0	1.0	1.0	TX	Cats,Dogs	No	Yes	1149.0
35697	798.0	650.0	1.0	1.0	FL	Cats,Dogs	No	Yes	1350.0
35698	1325.0	650.0	1.0	1.0	CA	Cats,Dogs	No	Yes	2195.0
35699	931.0	701.0	1.0	1.0	NC	Cats,Dogs	No	Yes	1126.5

Table 3: Training feature matrix (X_train) obtained after performing the train–test split. This dataset contains only the predictor variables used to train the model, with all preprocessing steps applied.

	square_feet	bathrooms	bedrooms	state	pets_allowed	fee	has_photo
0	1350.0	1.0	3.0	MI	Cats,Dogs	No	Thumbnail
1	900.0	1.0	2.0	LA	Cats,Dogs	No	Yes
2	781.0	1.0	1.0	MA	Cats,Dogs	No	Thumbnail
3	738.0	1.0	1.0	TX	Cats,Dogs	No	Thumbnail
4	1068.0	2.0	2.0	NC	Cats,Dogs	No	Thumbnail
...	...	...	...	...	...	...	...
28555	862.0	1.0	1.0	CO	Cats,Dogs	No	Thumbnail
28556	995.0	2.0	2.0	TX	Cats,Dogs	No	Yes
28557	1030.0	2.0	2.0	WA	Cats,Dogs	No	Thumbnail
28558	723.0	1.0	1.0	GA	Cats,Dogs	No	Yes
28559	522.0	1.0	3.0	WA	Cats,Dogs	No	Thumbnail

Table 4: Training target vector (y_train) obtained after the train–test split. This table contains the binary response variable indicating whether an apartment listing is classified as high-priced based on the state-level median price.

	high_price
0	0
1	0
2	0
3	0
4	1
...	...
28555	1
28556	1
28557	0
28558	1
28559	0

Table 5: Test feature matrix (X_test) obtained after performing the train–test split. This dataset contains the predictor variables used to evaluate the trained model, with all preprocessing steps applied.

	square_feet	bathrooms	bedrooms	state	pets_allowed	fee	has_photo
0	594.0	1.0	1.0	IA	Cats,Dogs	No	Thumbnail
1	874.0	2.0	2.0	TX	Cats,Dogs	No	Thumbnail
2	647.0	1.0	1.0	CA	Cats,Dogs	No	Thumbnail
3	1030.0	2.0	3.0	CO	Cats,Dogs	No	No
4	1180.0	1.0	2.0	MD	Cats,Dogs	No	Yes
...	...	...	...	...	...	...	...
7135	1061.0	2.0	2.0	VA	Cats,Dogs	No	Yes
7136	785.0	1.0	1.0	NC	Cats,Dogs	No	Yes
7137	796.0	1.0	1.0	UT	Cats,Dogs	No	Yes
7138	800.0	1.0	1.0	NE	Cats,Dogs	No	Thumbnail
7139	630.0	1.0	1.0	TX	Cats,Dogs	No	Thumbnail

Table 6: Test target vector (y_test) obtained after the train–test split. This table contains the true binary labels used to evaluate the model's predictive performance.

	high_price
0	0
1	1
2	1
3	1
4	1
...	...
7135	1
7136	0
7137	1
7138	0
7139	1

Exploratory Data Analysis (EDA)

We explored:

- Summary statistics

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

Table 7: Summary statistics for apartment listings, including price, square footage, and number of bathrooms. The table reports the count, central tendency (mean and median), dispersion (standard deviation and interquartile range), and extrema, providing an overview of the scale and variability of key features in the dataset.

	Unnamed: 0	price	square_feet	bathrooms
0	count	35700.000000	35700.000000	35700.000000
1	mean	1487.928235	940.632157	1.440294
2	std	722.414581	338.539094	0.533467
3	min	285.000000	200.000000	1.000000
4	25%	1025.000000	720.750000	1.000000
5	50%	1327.000000	897.000000	1.000000
6	75%	1745.000000	1105.000000	2.000000
7	max	19500.000000	12000.000000	7.000000

Table 8: Class distribution of the binary target variable (high_price). Approximately 51.1% of listings are labeled as high-priced (1), while 48.9% are labeled as not high-priced (0), indicating a nearly balanced classification problem.

	high_price	proportion
0	0	0.501765
1	1	0.498235

Modeling: Logistic Regression Classifier

We predict `high_price` using:

Numeric features.

- square_feet
- bathrooms

Categorical features.

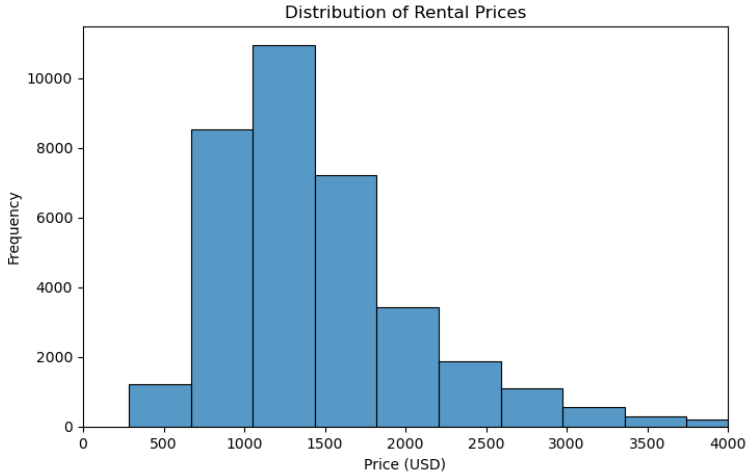

Figure 1: The histogram shows a right-skewed distribution of apartment rental prices, with most listings concentrated between approximately $800 and $1,500 and a long upper tail.

Figure 2: Relationship between apartment size and rental price, with points colored by the high-price label (1 = high-priced, 0 = not high-priced). Larger units generally command higher prices, though substantial overlap between the two classes highlights variability in pricing beyond square footage alone.

- bedrooms
- state
- pets_allowed
- fee
- has_photo

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

Table 9: Classification performance of the logistic regression model evaluated on the held-out test set using a train–test split and a preprocessing pipeline. Reported metrics include accuracy, precision, recall, and F1-score, summarizing the model's overall predictive effectiveness and balance between false positives and false negatives.

	Unnamed: 0	Accuracy	Precision	Recall	F1-score
0	Logistic Regression	0.695938	0.707983	0.663199	0.68486

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.

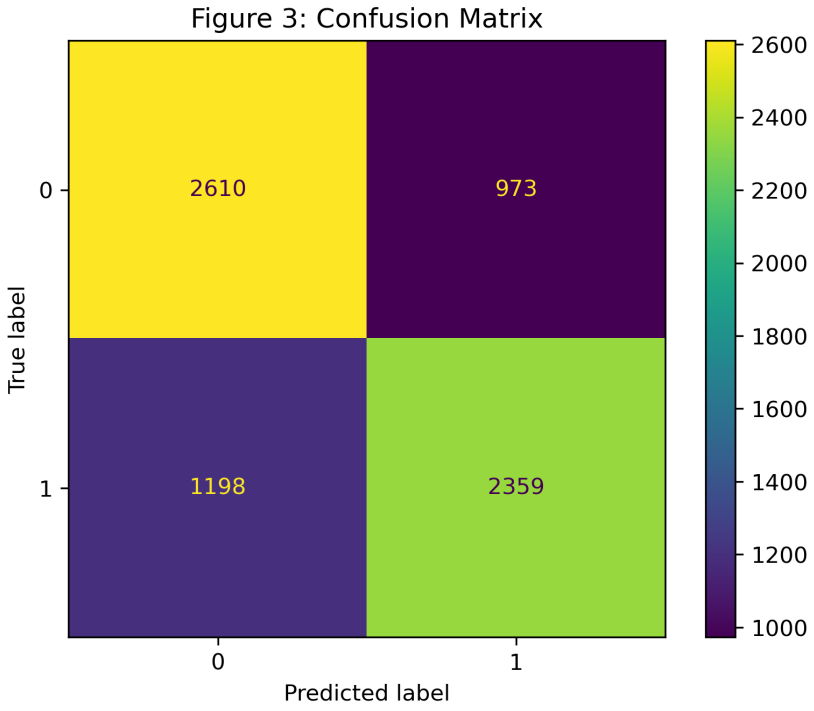

Figure 3: Confusion matrix summarizing the binary classification performance for predicting high-priced apartments. The model correctly identifies 2,610 low-priced and 2,359 high-priced units, while misclassifying 973 low-priced units as high-priced and 1,198 high-priced units as low-priced, highlighting a moderate trade-off between false positives and false negatives.

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_model.html#logistic-regression

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