

PREDICTING HOUSE PRICES USING MACHINE LEARNING



INDUPAMA HERATH

SARASIJ MAITRA

RAFATU SALIS

ERSIN SÜER



THE ERDŐS INSTITUTE

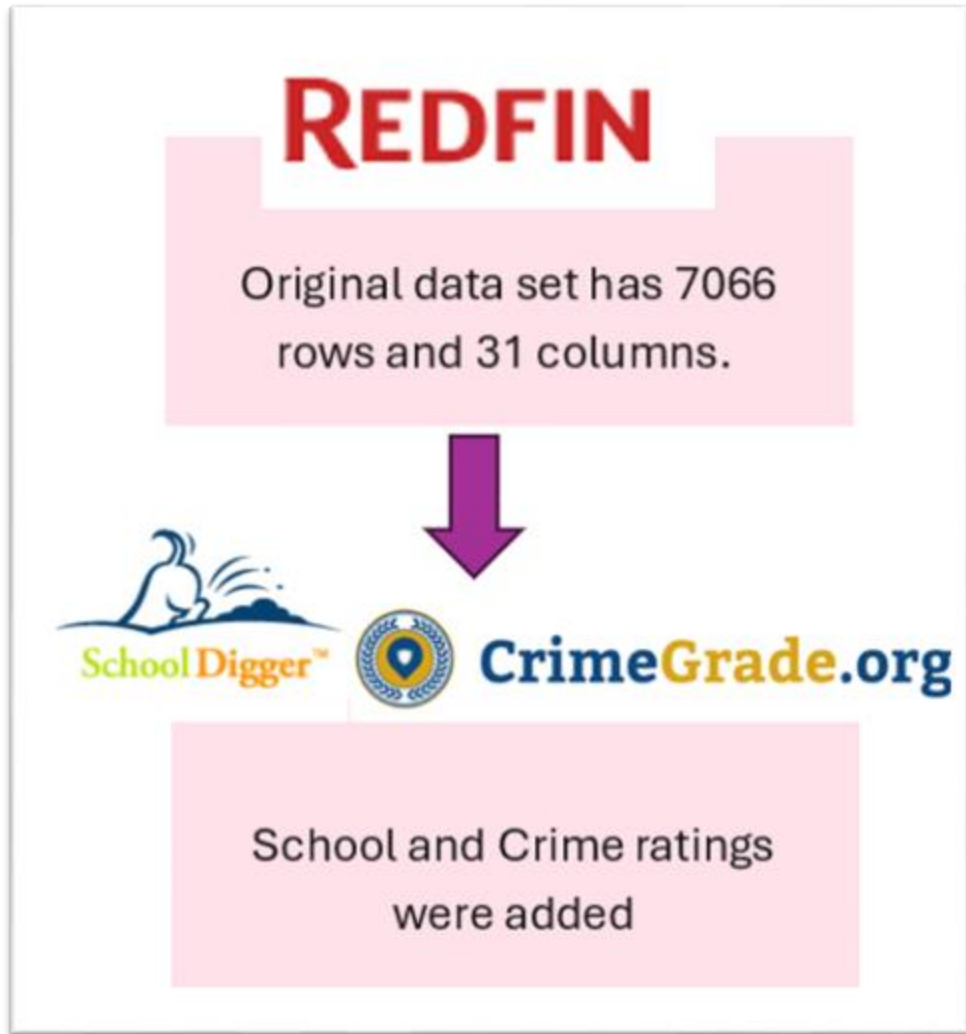
Helping PhDs get and create jobs they
love at every stage of their career.

OBJECTIVES

- Predict house prices incorporating traditional and non-traditional features using machine learning
- Build a web application to predict the house price for user input feature values



DATA SET



DATA CLEANING

- Removed unnecessary columns and rows with missing values.
- Add new columns
 - Age, Log_price

Age = 2024 – YEAR_BUILT
Log_price = Log(PRICE)

- Cleaned data set has 4700 rows and 19 columns (5 categorical and 14 numerical)

EXPLORATORY DATA ANALYSIS

PRICE DISTRIBUTION

Minimum: 49000

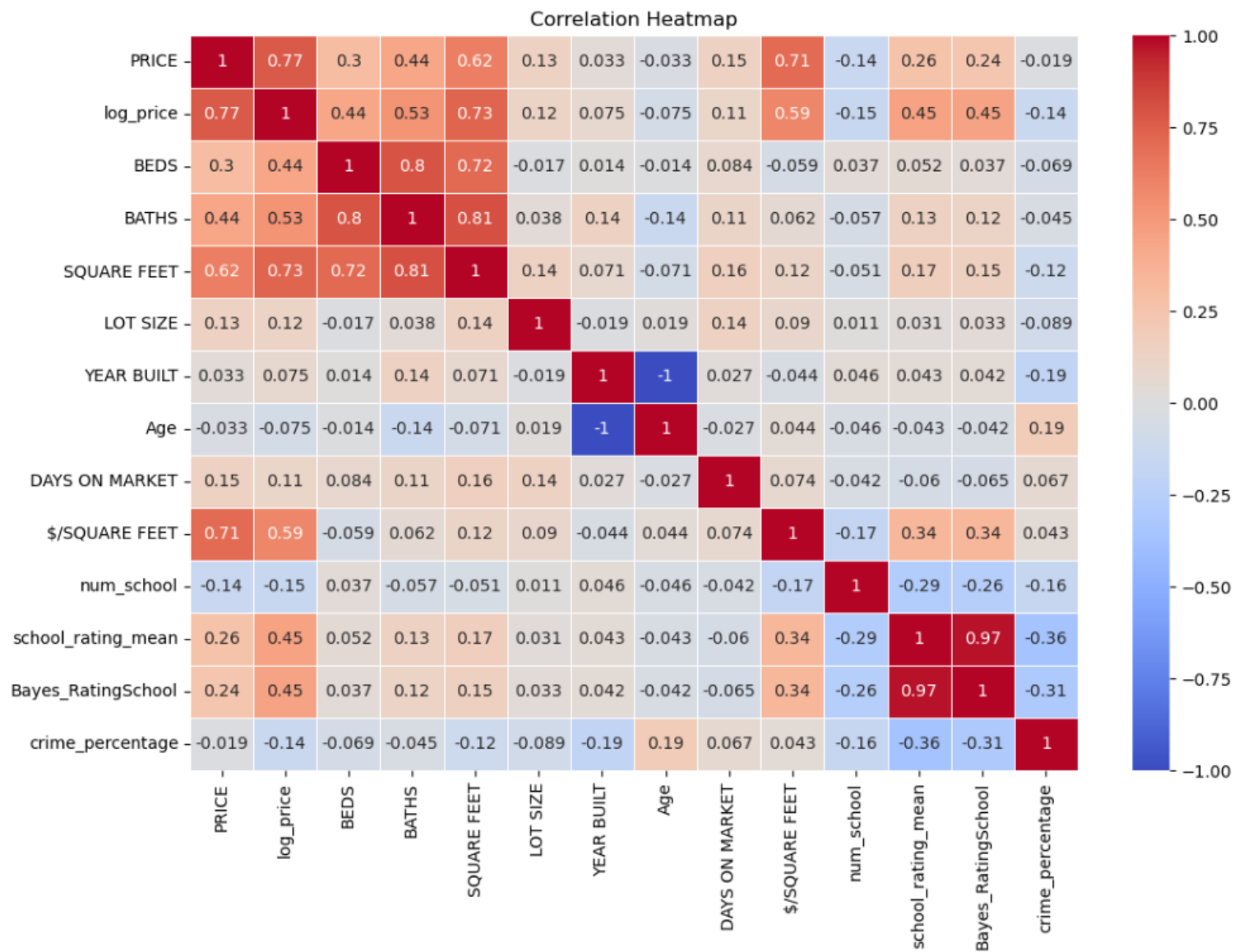
Mean: 1505699

Median: 998975

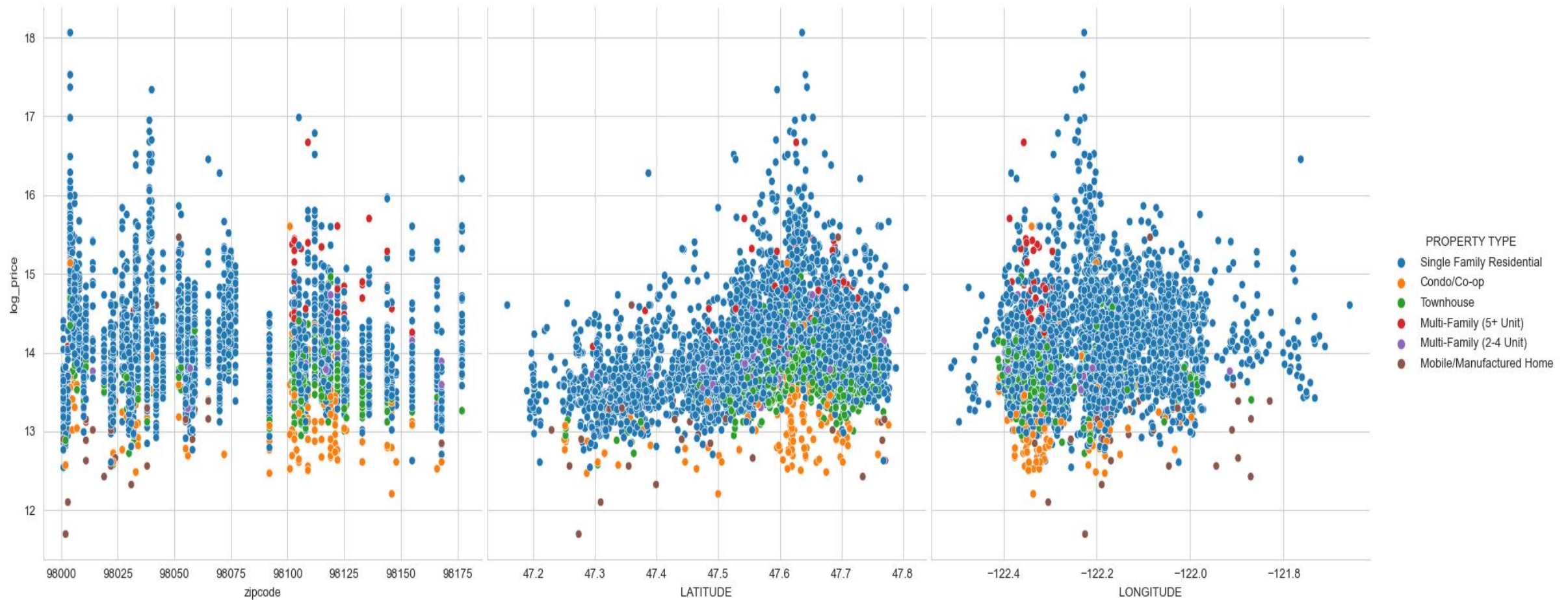
Maximum: 70000000

>

CORRELATION ANALYSIS OF HOUSING MARKET VARIABLES



Primary tool: *seaborn*

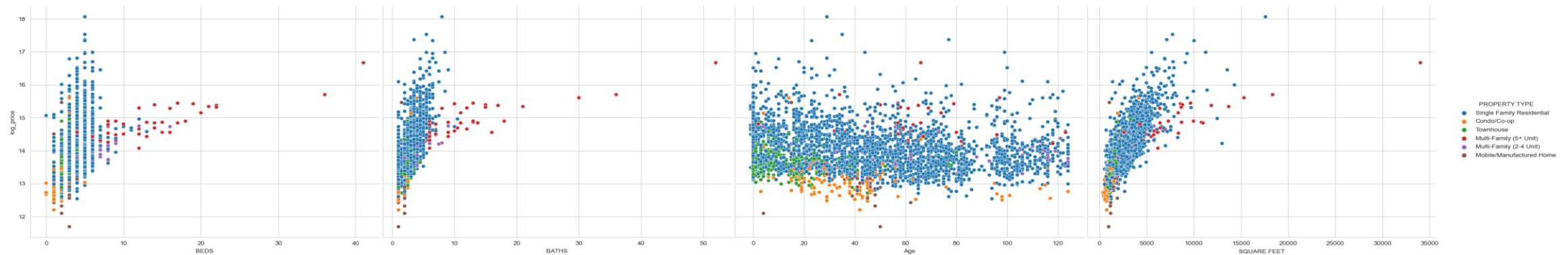
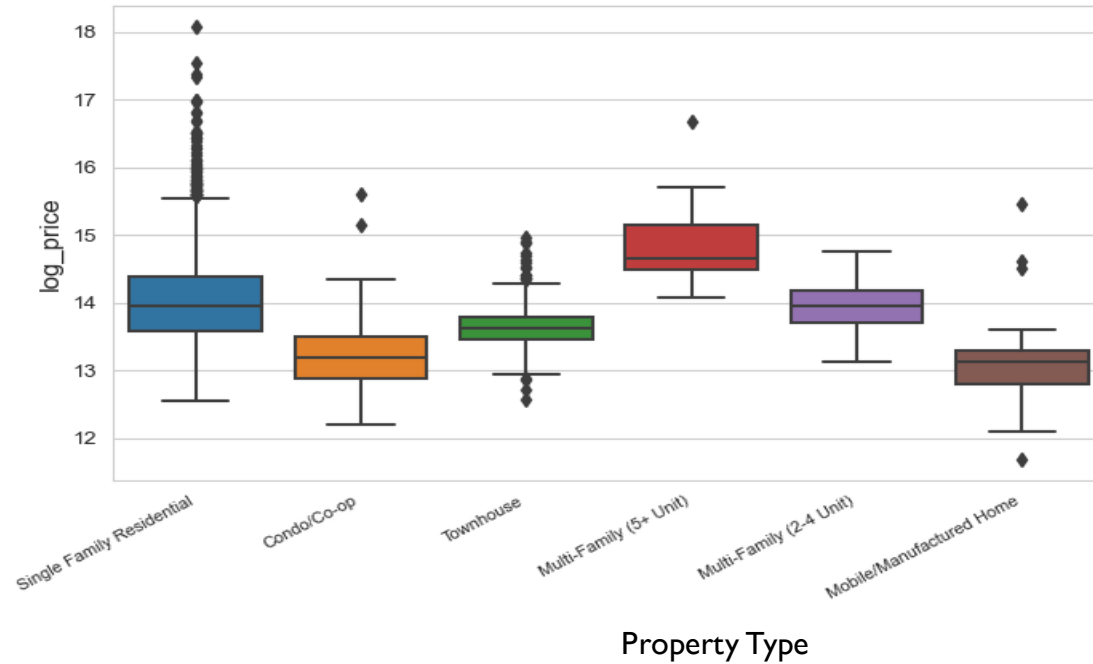


Effect of Location on Log_Price colored by Property Type

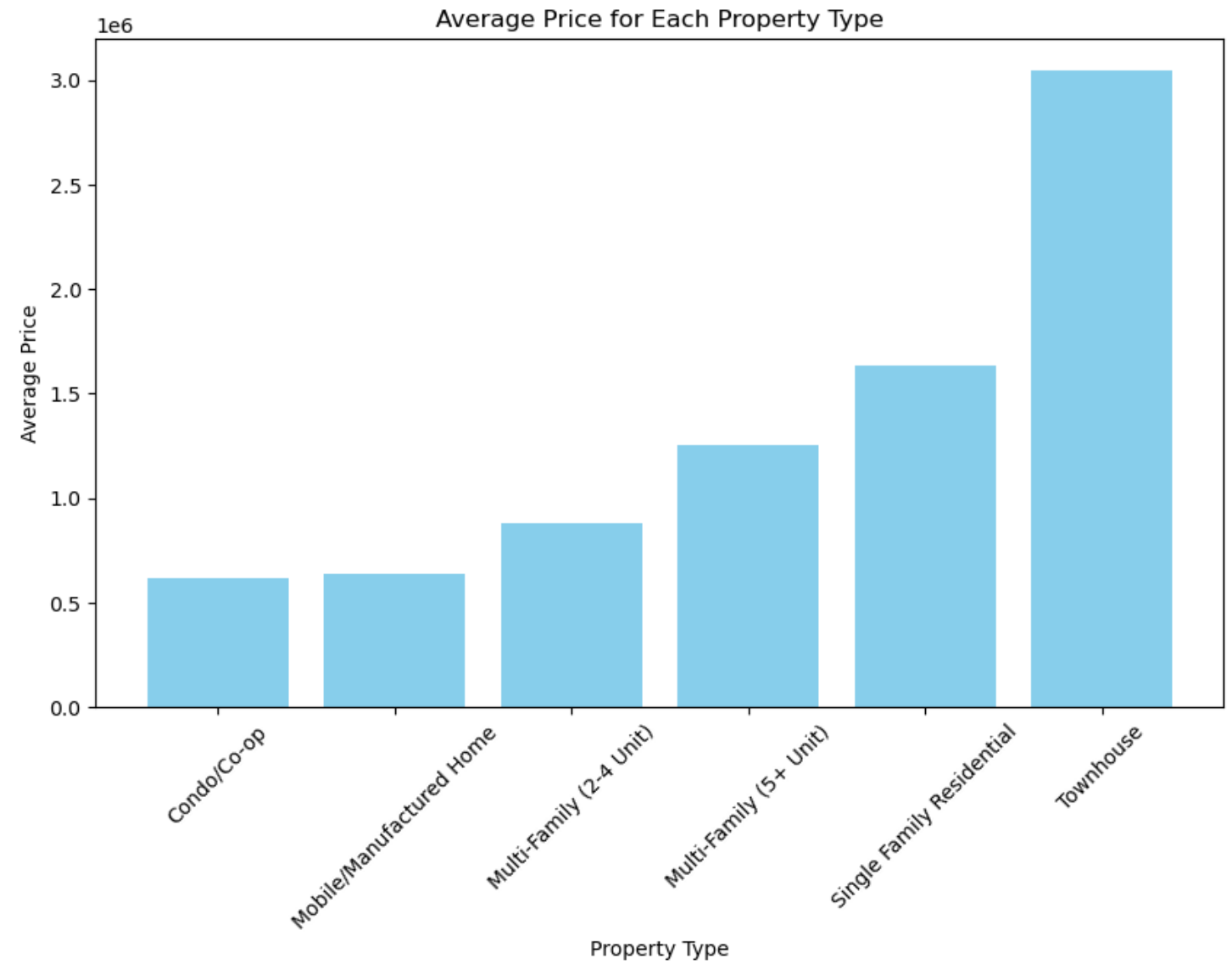
MORE PLOTS

Primary tool: seaborn

Boxplot on the right shows the distribution of *log_price* for the categorical variable *property type*.



MORE PLOTS



FEATURE ENGINEERING

- Delete outliers based on price.
- One hot encoding on PROPERTY TYPE column

PROPERTY TYPE

Single Family Residential	3030
Townhouse	430
Condo/Co-op	160
Multi-Family (2-4 Unit)	63
Multi-Family (5+ Unit)	45
Mobile/Manufactured Home	32

FINAL DATA SET

Data columns (total 16 columns):

#	Column	Non-Null Count
---	-----	-----
0	BEDS	3759 non-null
1	BATHS	3759 non-null
2	SQUARE FEET	3759 non-null
3	LOT SIZE	3759 non-null
4	zipcode	3759 non-null
5	LATITUDE	3759 non-null
6	LONGITUDE	3759 non-null
7	Bayes_RatingSchool	3759 non-null
8	crime_percentage	3759 non-null
9	Age	3759 non-null
10	Single Family	3759 non-null
11	Townhouse	3759 non-null
12	Condo	3759 non-null
13	Multi_Family4	3759 non-null
14	Multi_Family5	3759 non-null
15	log_price	3759 non-null

15 features to predict House
Price (in log scale)

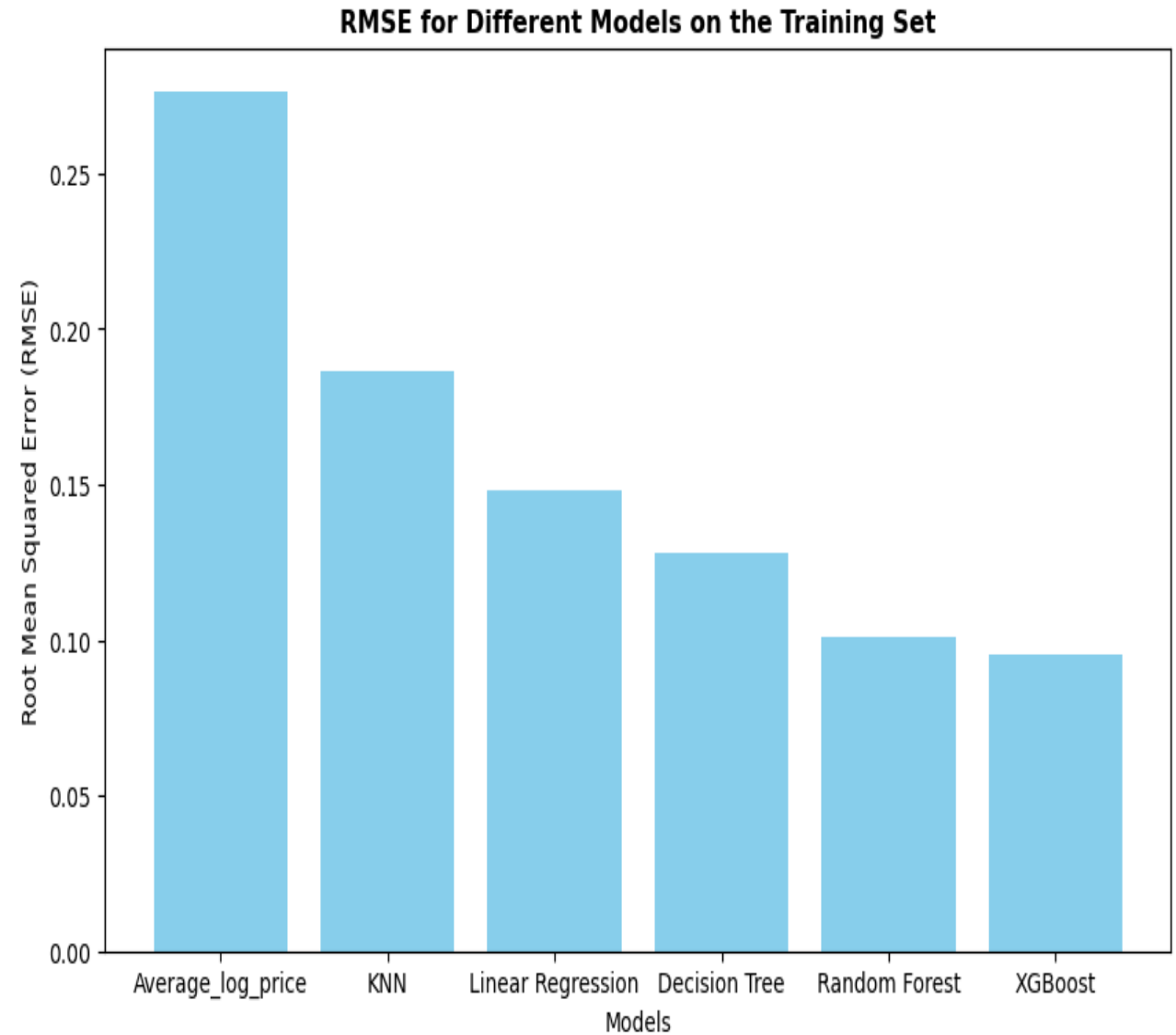
MODELING

- 1.The baseline model – Average of the House Prices (log scale)
- 2.K – Nearest Neighbors
- 3.Multiple Linear Regression
- 4.Decision Tree
- 5.Random Forest
- 6.XGBoost

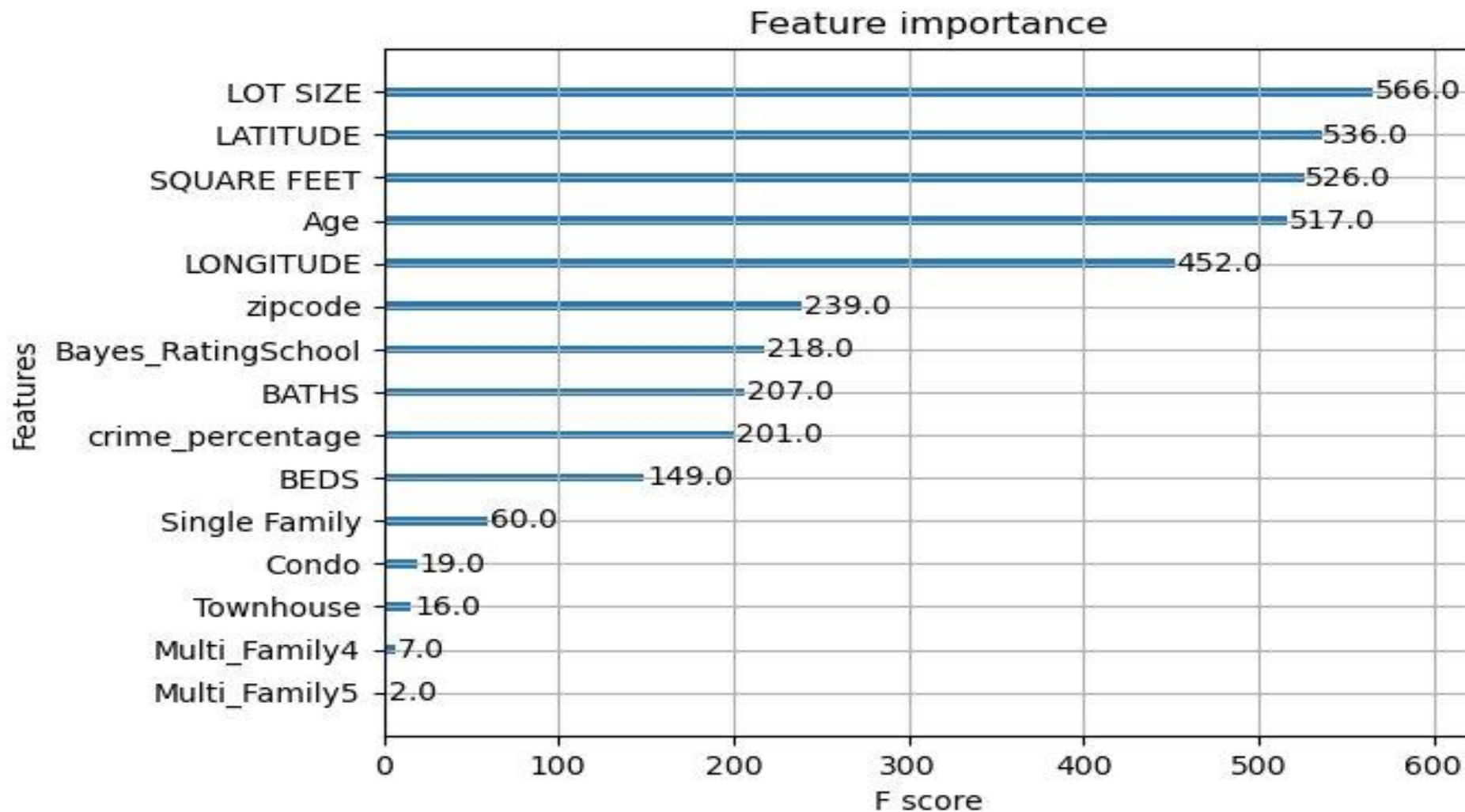
Used 5 – fold cross validation on the training set and RMSEs were computed for the predicted price in log scale.

MODEL TUNING

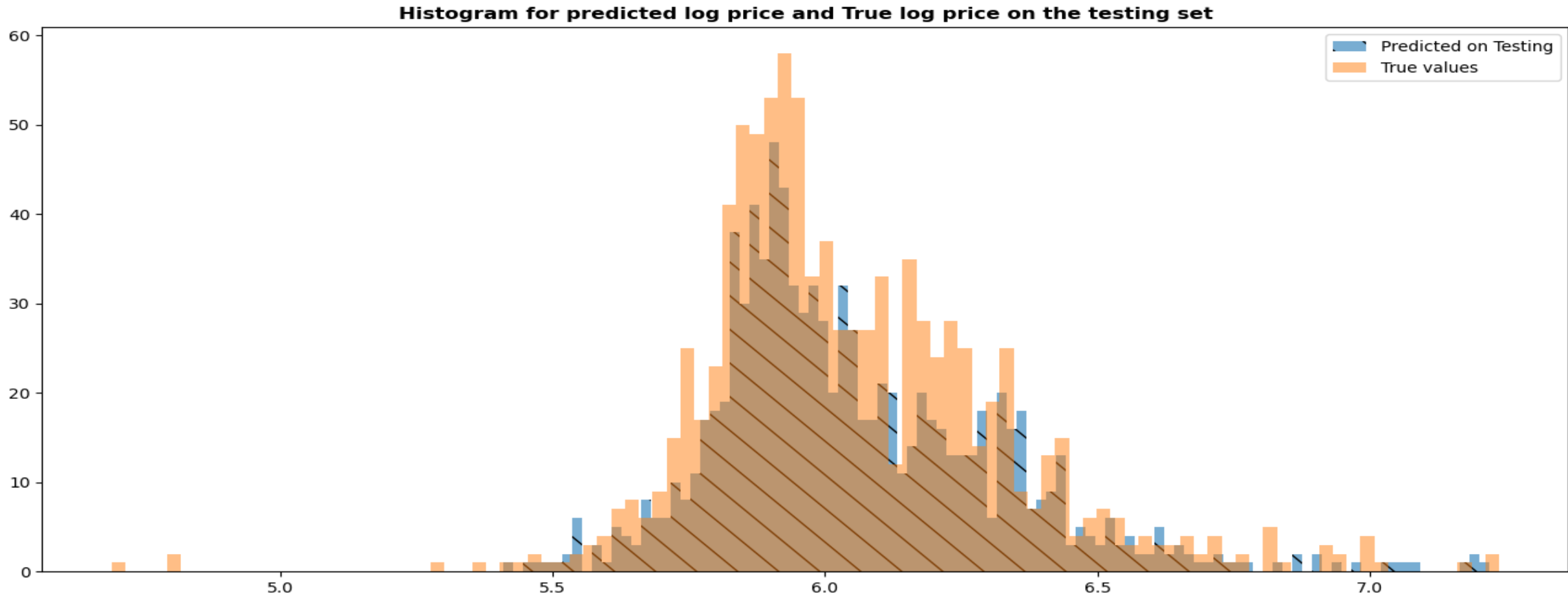
- Used a grid search-based approach to find the best set of parameters for KNN, Decision Tree, Random Forest, and XGBoost.
- Best model, XGBoost gives RMSEs,
 - **0.0954 on training set**
 - **0.1003 on testing set**



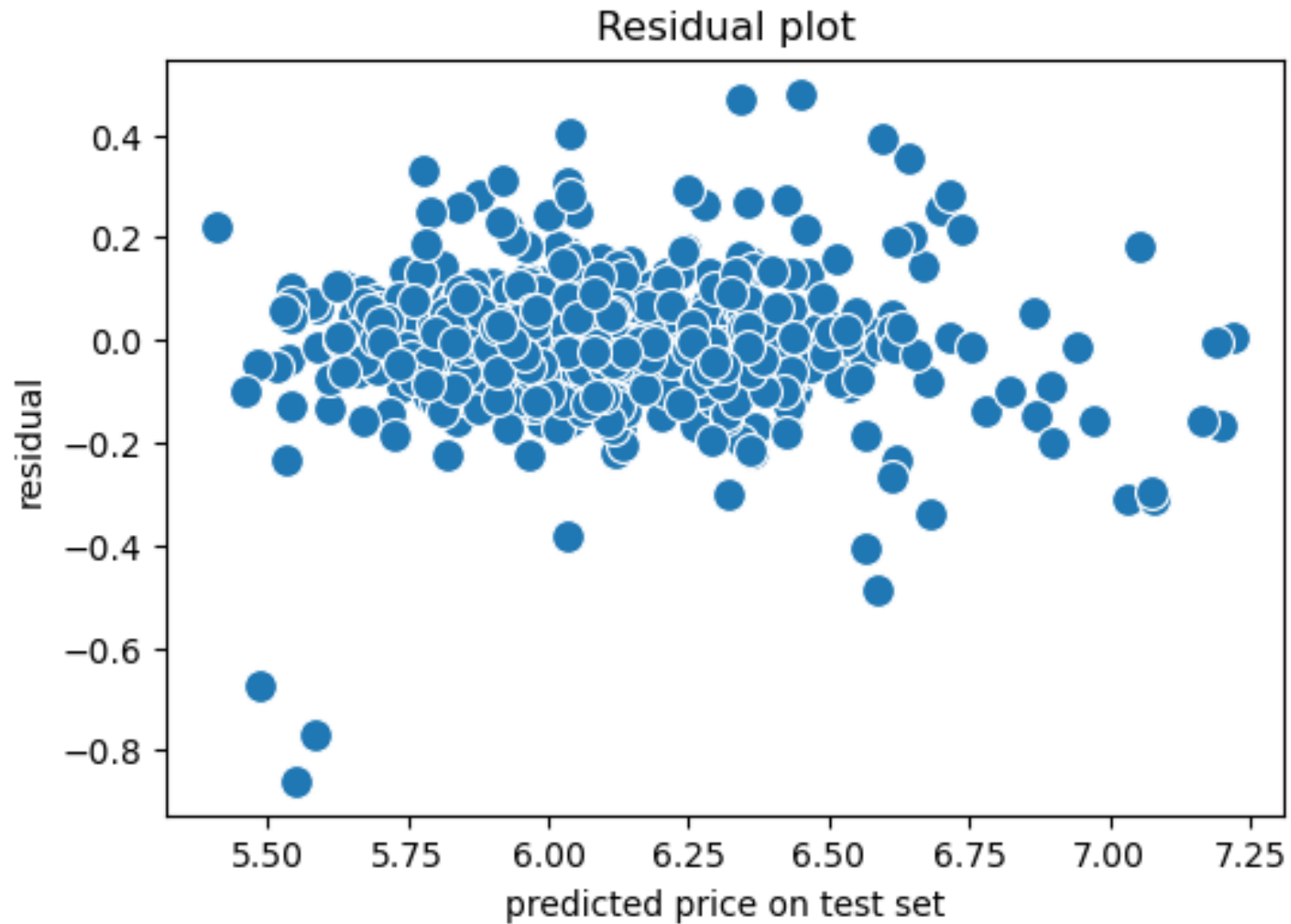
IMPORTANT FEATURES OF XGBOOST



HISTOGRAMS FOR PREDICTED VS TRUE LOG PRICE

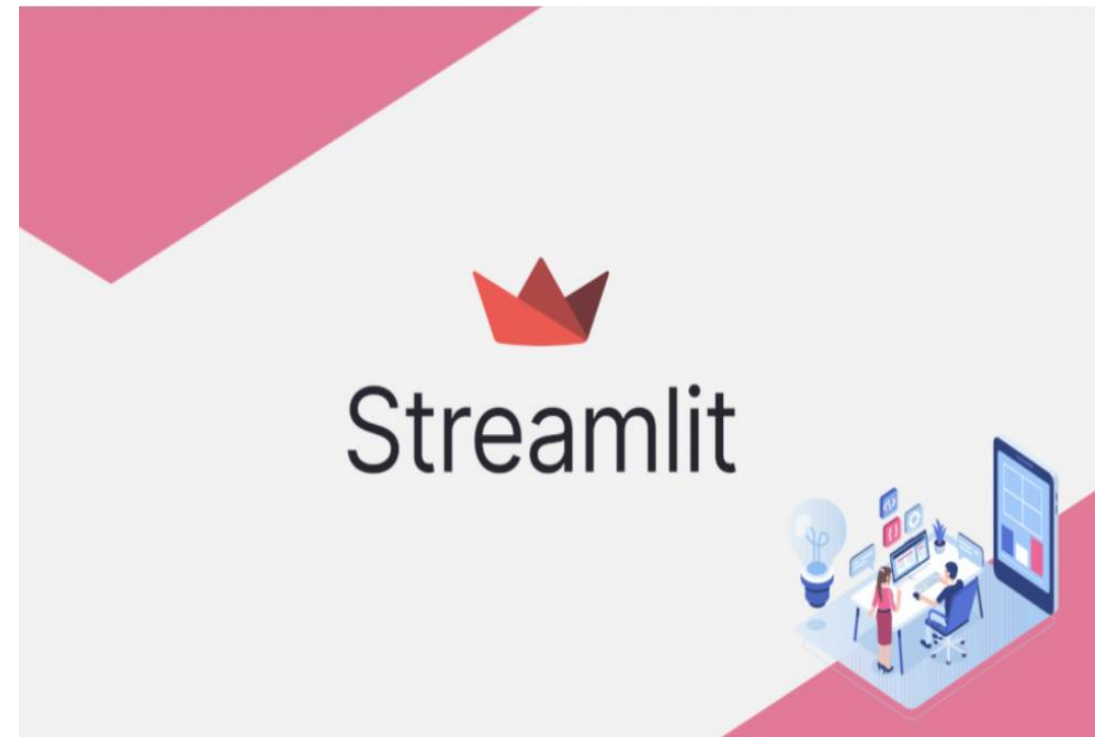


RESIDUAL PLOT



WEB APPLICATION

- Using our final model, we built a simple web application on Streamlit that takes in user inputs (relevant to our model) and predicts the house price.
- The app is publicly available at <https://erdos-datascience-may2024-realestate-project-nevrbzjn2sh2zgsrc.streamlit.app/>



- Visualizations using streamlit_folium
- clearer view of the surrounding big cities, airports, etc. , to make informed choices

Enter desired Property Type

Condo/Co-op

Property type chosen by user: Condo/Co-op

Choose your desired zipcode

98023

98001

98002

98003

98004

98005

98006

98007

98008

Enter Latitude

47.2906325

47.2735581

47.3534573

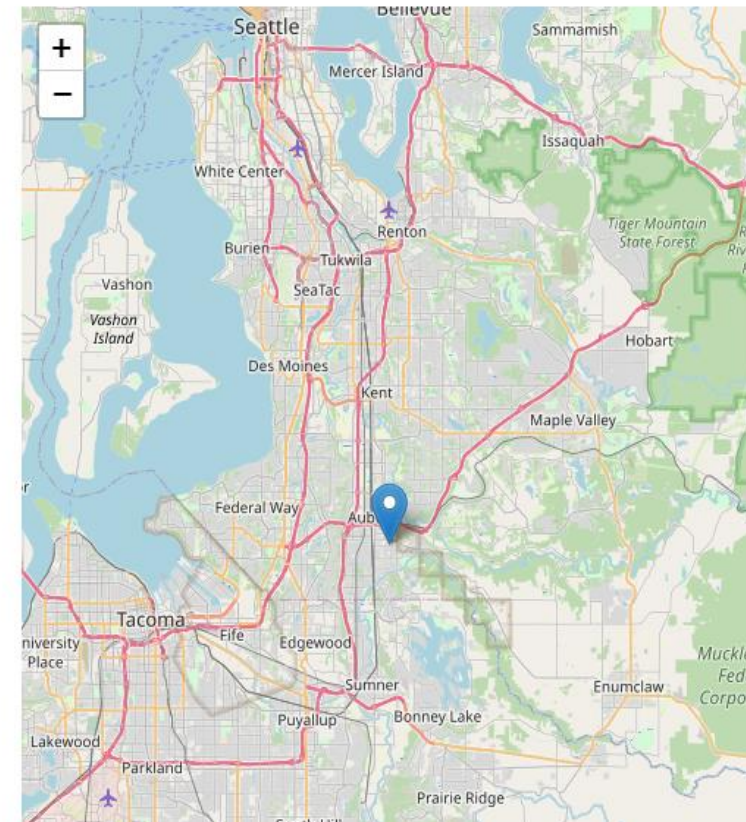
Enter Latitude

-122.2162074

-122.2286486

-122.1982048

Chosen Latitude and Longitude (47.2906325, -122.2162074)



CONCLUSIONS

- Overall, all models did improve from the baseline model and did well on predicting the price.
- Contribution of non-traditional features are similar to some of the traditional features.

FUTURE WORK

- Extend the study for other states.
- Incorporate more relevant features such as
 - whether the house has experienced flooding,
 - has mold issues,
 - the quality of construction materials,
 - the floor plan, and
 - whether fixtures and appliances have been recently updated.

ACKNOWLEDGEMENTS

Our team is very grateful for all the guidance and advice provided by our instructor Steven Gubkin, our mentor Alec Traaseth and Alec Clott throughout the course of this project. We also deeply thank Roman Holowinsky and The Erdős Institute for providing us the platform and opportunity to work on this project.



THE ERDŐS INSTITUTE

Helping PhDs get and create jobs they
love at every stage of their career.