PREDICTING HOUSE PRICES WITH MACHINE LEARNING MODELS

Data Science Capstone Two Project

Mingyu (Bert) Liu Ph.D

PROBLEM

- There are many features which play roles in house price decisions. Based on historical data of house sales in different regions in the U.S., we need to predict house price based on features.

- Which features can significantly affect housing prices?

- Build a model to reflect on the relationships between price of houses and the features.

WHO MIGHT CARE?

FOR INVESTORS TO MAKE DECISIONS OF BUYING HOUSES

FOR CONSULTANT TO PREDICT HOUSE PRICES IN SPECIFIC CITIES

MAIN CONCLUSION

THE MAIN FACTORS AFFECTING HOUSING PRICES

- Locations：state，city，zip_code
- Types of houses: average_acre_lot,average_house_size
- Sizes of houses: Bedroom,bathroom,house_size,
- Status: for_sale,read_to_build
- Ages of houses: pre_sold_date

THE FINAL PREDICTION MODEL SELECTED

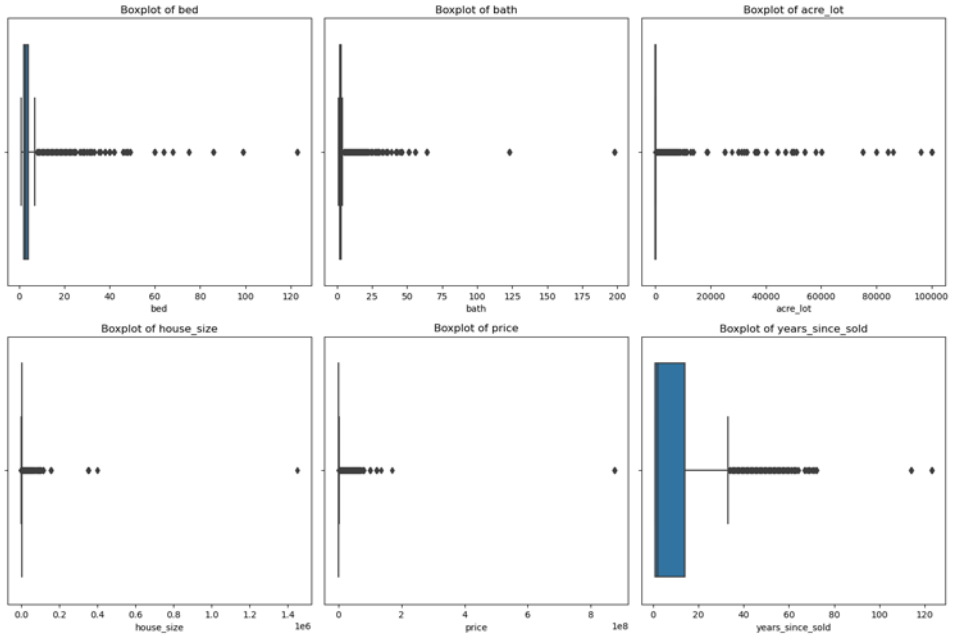
- By comparing the prediction performance of machine learning models, we finally chose the random forest model for prediction
- The random forest model achieved the prediction accuracy of 99 percent
- By mining the attribute characteristics of some houses, the prediction performance of the model will be improved，for examples: types of houses, ages of houses

DATA INFORMATION

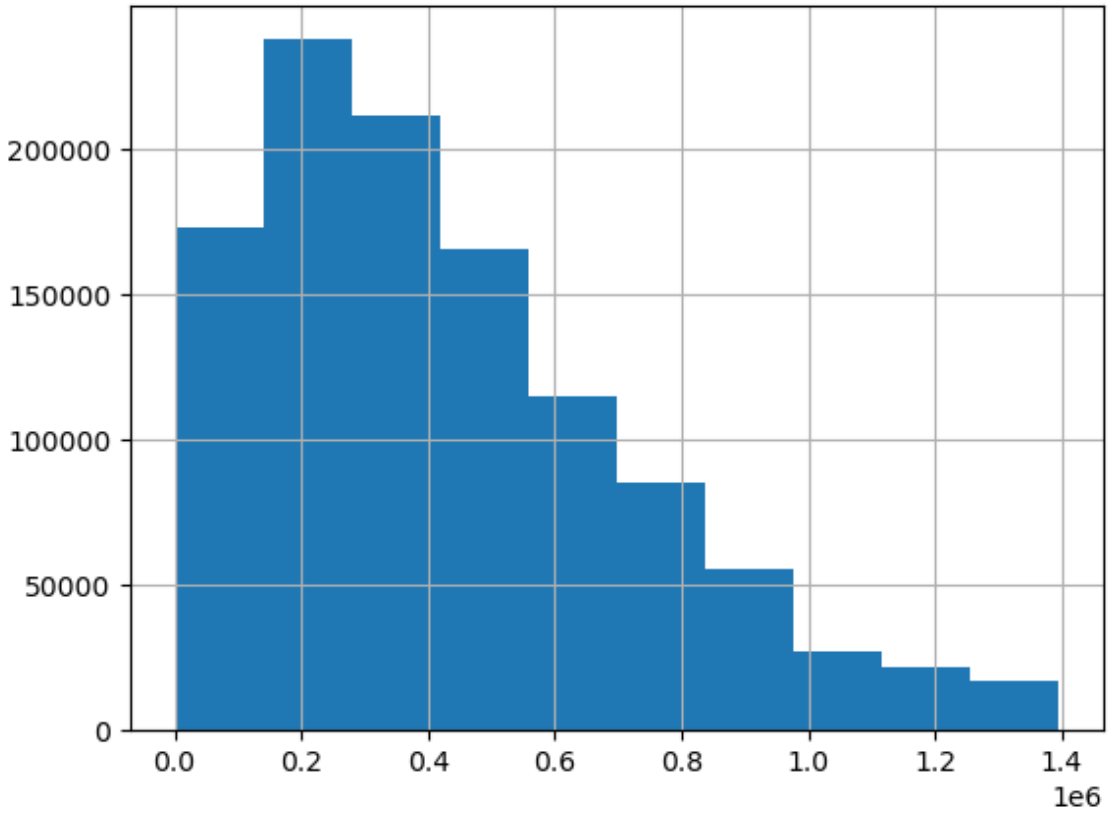
- This dataset has a total of 1,401,066 entries and 10 columns from Kaggle-USA Real Estate Dataset.
- The columns include a mix of object and float64 data types, with some missing values across several columns.
  - status: A categorical column (likely containing strings) with no missing values.
  - bed: A numerical column (floating-point) with some missing values (around 1,216,528 non-null values).
  - bath: Another numerical column with some missing values.
  - acre_lot: A numerical column representing lot sizes, with a significant number of missing values.
  - city: A categorical column with very few missing values.
  - state: A categorical column without missing values.
  - zip_code: A numerical column representing zip codes with some missing values.
  - house_size: A numerical column with substantial missing values.
  - prev_sold_date: A date column represented as an object, with a large number of missing values.
  - price: A numerical column with few missing values.

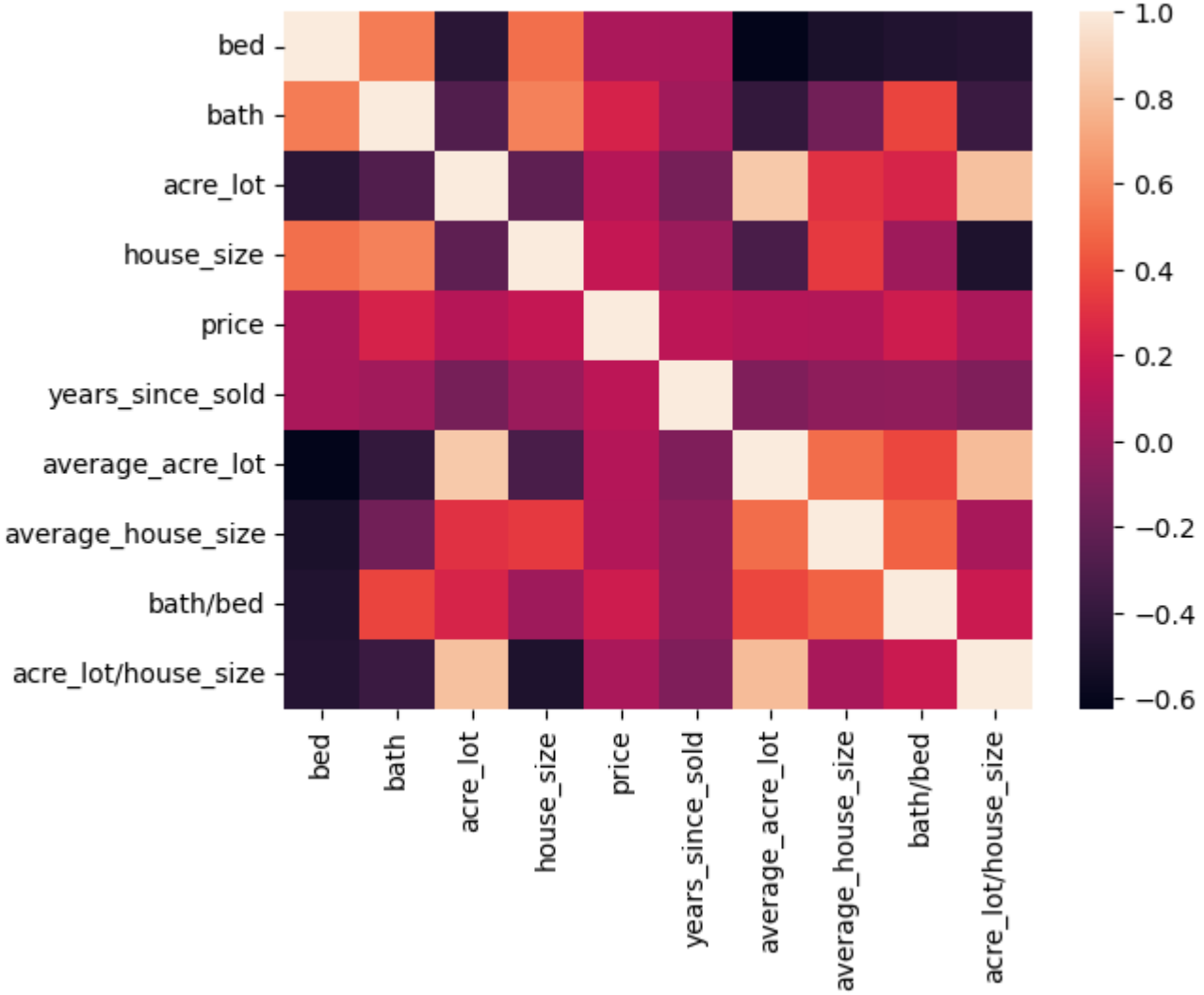
DATA CLEANING

- Handle Missing Values: Decide on strategies to fill missing values in numerical and categorical columns.
- ➢ Convert Data Types: Convert prev_sold_date from an object to a datetime type.
- ➢ Outlier Detection: Identify and potentially remove or adjust outliers in numerical columns like price, acre_lot, or house_size.

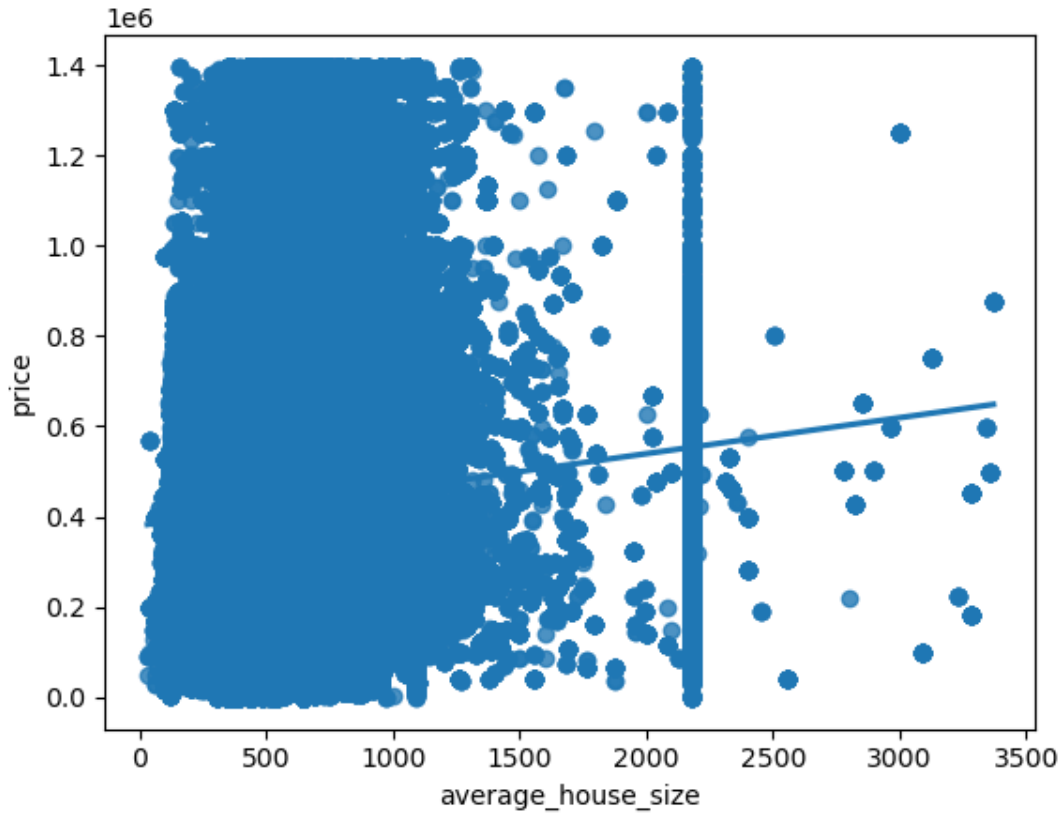


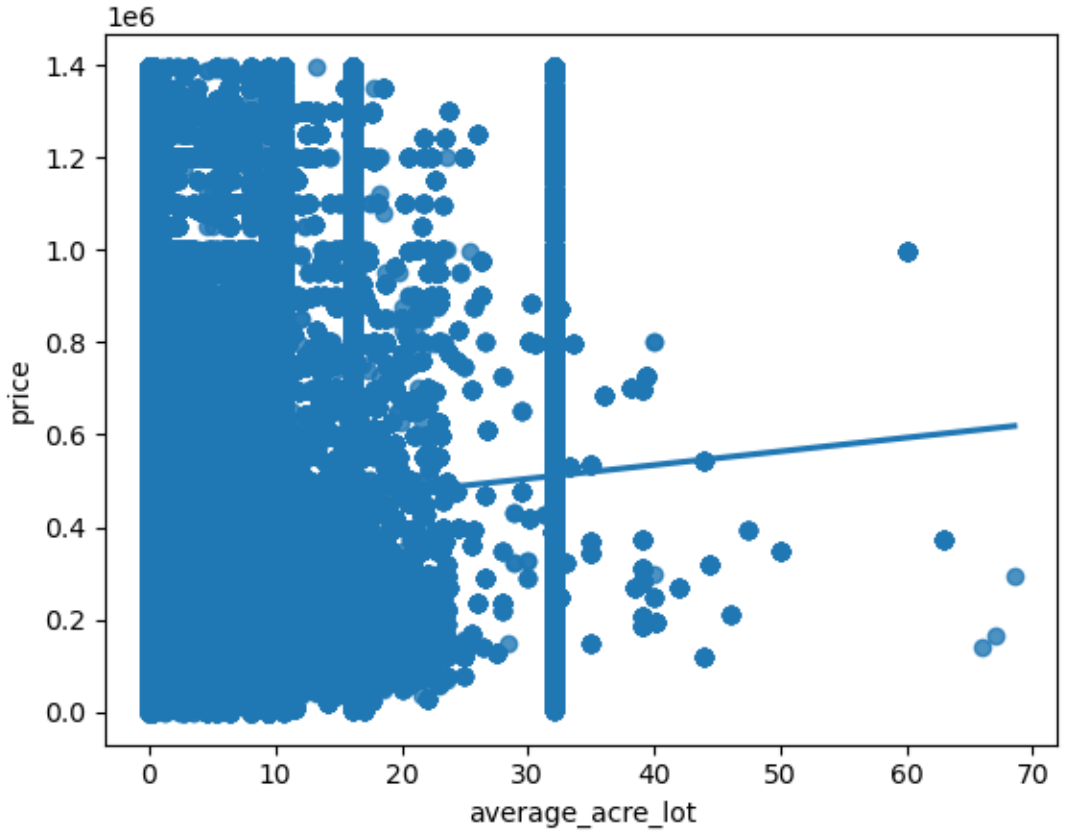
EDA(EXPLORATORY DATA ANALYSIS)





CORRELATIONSHIP





FEATURE ENGINEERING

- Create New Features: Use existing columns to create new features, like 'average_acre_lot','average_house_size',etc.
- Encode Categorical Variables: Convert status, city, zip_code, and state into numerical formats for machine learning tasks.

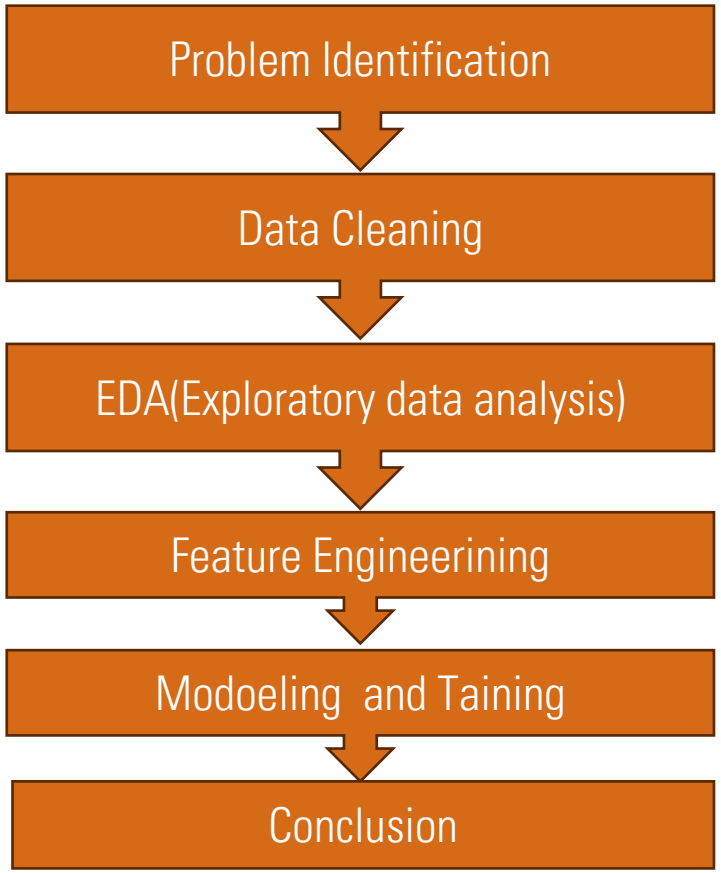
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1110643 entries, 0 to 1110642
Data columns (total 14 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   bed                   1110643 non-null  float64
 1   bath                  1110643 non-null  float64
 2   acre_lot              1110643 non-null  float64
 3   house_size            1110643 non-null  float64
 4   price                 1110643 non-null  float64
 5   years_since_sold      1110643 non-null  float64
 6   average_acre_lot      1110643 non-null  float64
 7   average_house_size    1110643 non-null  float64
 8   bath/bed              1110643 non-null  float64
 9   acre_lot/house_size   1110643 non-null  float64
 10  zip_code_encoded      1110643 non-null  int64
 11  city_encoded          1110643 non-null  int64
 12  state_encoded         1110643 non-null  int64
 13  status_encoded        1110643 non-null  int64
dtypes: float64(10), int64(4)
memory usage: 118.6 MB
```

SELECT A MACHINE LEARNING MODEL

- Machine Learning Models
  - Regression Models: Predict price based on other features using linear regression.
    - Model Performance:
    - Mean Squared Error: 0.8443409875971409
    - R^2 Score: 0.15328624829493187
  - Random Forest Model:
    - Model Performance:
    - Mean Squared Error: 0.014193229245588218
    - R^2 Score: 0.9857668849909297
  - Neural Networks Model:
    - Model Performance:
    - Mean Squared Error: 0.3350867183075398
    - R^2 Score: 0.6639716221616279

- Comparing the prediction performance of the three models
- Random forest has the best prediction effect, with a prediction accuracy of 99%
- We recommend choosing the random forest model as the prediction model

STEP BY STEP PROCESS



Split the data into training and testing sets
test_size=0.2

Chose the random forest model for prediction
The prediction accuracy of 99 percent

THANK YOU FOR LISTENING