ABSTRACT (DESCRIPTION)

The goal of this project was to predict home sale prices of the bay area in northern California through the use of regression modeling and evaluation. Using data from Redfin.com acquired via web scraping, I performed EDA to prepare the data for modeling – to determine if certain home features could accurately be used to predict sale prices of homes still available on the market.

DESIGN

The project consisted of several parts:

First, filtering for single family homes sold within the last 6 months for specific cities, I web scraped data from individual listings, from multiple search results pages on Redfin.com for home features.

The data was then cleaned and prepared for a baseline simple linear regression, to determine if any features would contribute to a model with high predictive performance and interpretability.

Then, using the processed data and significant features, to iterate thru various regression models to evaluate performance and whether additional feature engineering and transformations were required.

DATA (FEATURE AND TARGET)

The data consisted of single family homes sold within the last 6 months for the following cities:

Palo Alto, Los Altos, Mountain View, Cupertino, Sunnyvale, Santa Clara

It contained 1600+ data points with 13 features (8 of which were numerical to start, 4 of which were boolean, and 1 categorical).

'Sold Price' was selected as my target, also known as the predictor variable. After determining that some numerical features would benefit from feature engineering, it resulted in 30+ final features

ALGORITHMS

Feature Engineering

- Converting categorical features to binary dummy variables
- Combining (bucketing) particular dummies and ranges of numeric features to highlight strong signals

Modeling

- Linear Regression
- Regularization models with Cross Validation, including Ridge, Lasso, and Elastic Net

Model Evaluation and Selection

- The data set was split into 60/20/20 train, validation, test (holdout). Models were fitted on the train portion and scored and evaluated on the validation. The desired metric was Mean Absolute Error (for interpretability)

Final Model Selected: Elastic Net 5-Fold CV Scores

- R-squared 0.801
- Adjusted R-squared 0.774
- Mean Absolute Error $398667
- Root Mean Squared Error $543227
- Feature Coefficients

```
lot size : 130766.26
home size : 773471.44
school score avg : 269845.16
laundry : 3076.32
heating : -39490.85
air conditioning : -18532.17
pool : 15432.62
age of house : 0.00
beds_3.0 : 21291.99
beds_4.0 : 0.00
beds_5.0 : -31173.91
beds_6+ : -71899.62
baths_1.5 : -0.00
baths_2.0 : 0.00
baths_2.5 : -28218.86
baths_3.0 : -0.00
baths_3.5 : 0.00
baths_4.0 : -43796.74
baths_4.5 : 20622.92
baths_5.0 : 37970.59
baths_6+ : 52139.99
floors_2.0 : -36938.23
floors_3.0 : -96698.00
garage spaces_1 : 17165.13
garage spaces_2 : 42698.15
garage spaces_3+ : -37222.90
city_LOS ALTOS : 205731.06
city_MOUNTAIN VIEW : 110700.70
city_PALO ALTO : 336679.56
city_SANTA CLARA : -73019.63
city_SUNNYVALE : 29654.19
```

TOOLS

- BeautifulSoup for web scraping
- Numpy and Pandas for data cleaning, transformation, and manipulation
- Scikit-Learn for modeling, scoring and evaluation
- Matplotlib and Seaborn for data visualization and plotting

COMMUNICATION

In addition to the slides and visuals presented, all work is available on my GitHub