Real Estate Price Prediction with MLS and Redfin Data

Michello Ho, Shiyun Qiu, Jiawen Tong, Yiqi Xie

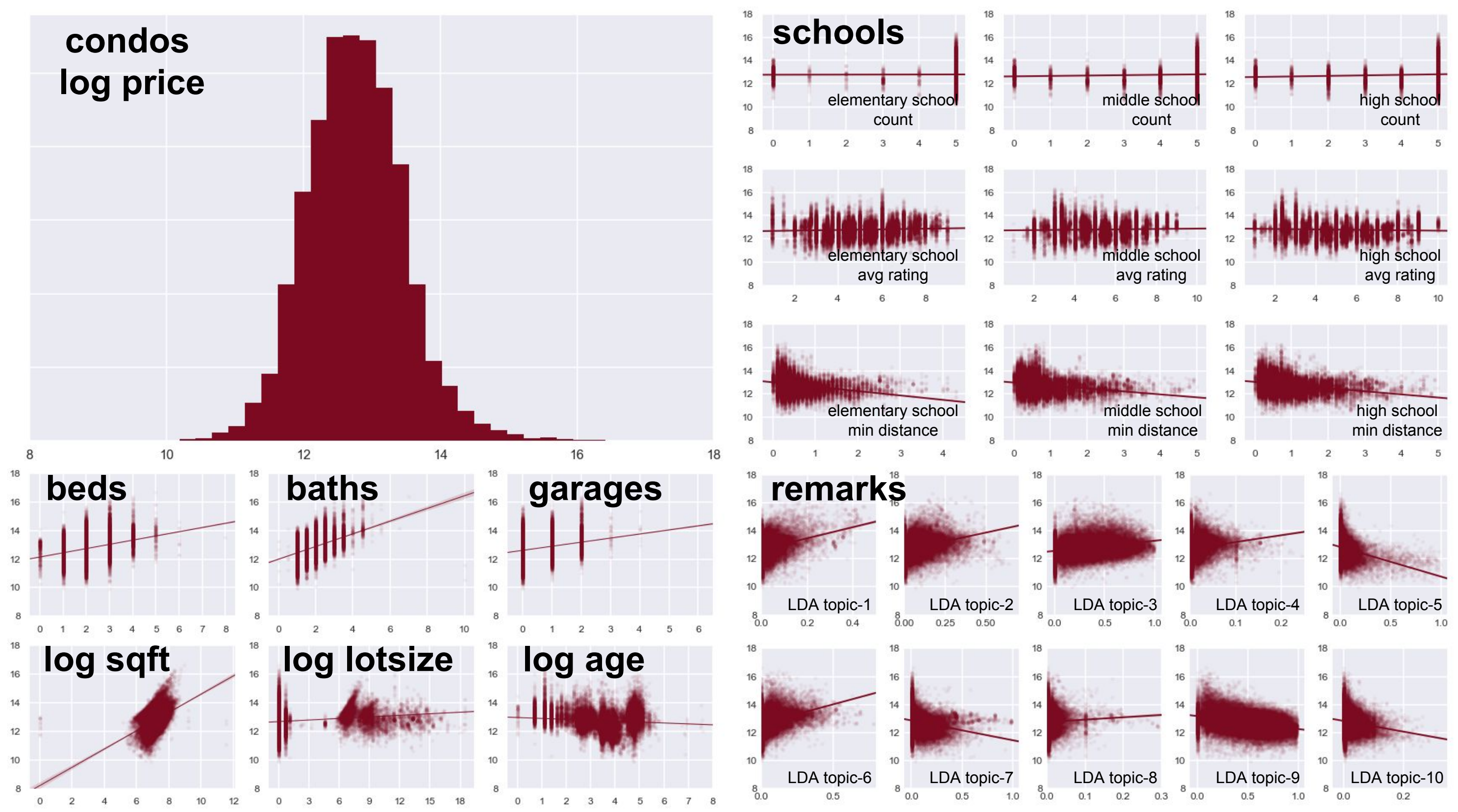
HarvardCS209B Data Science
Final Project

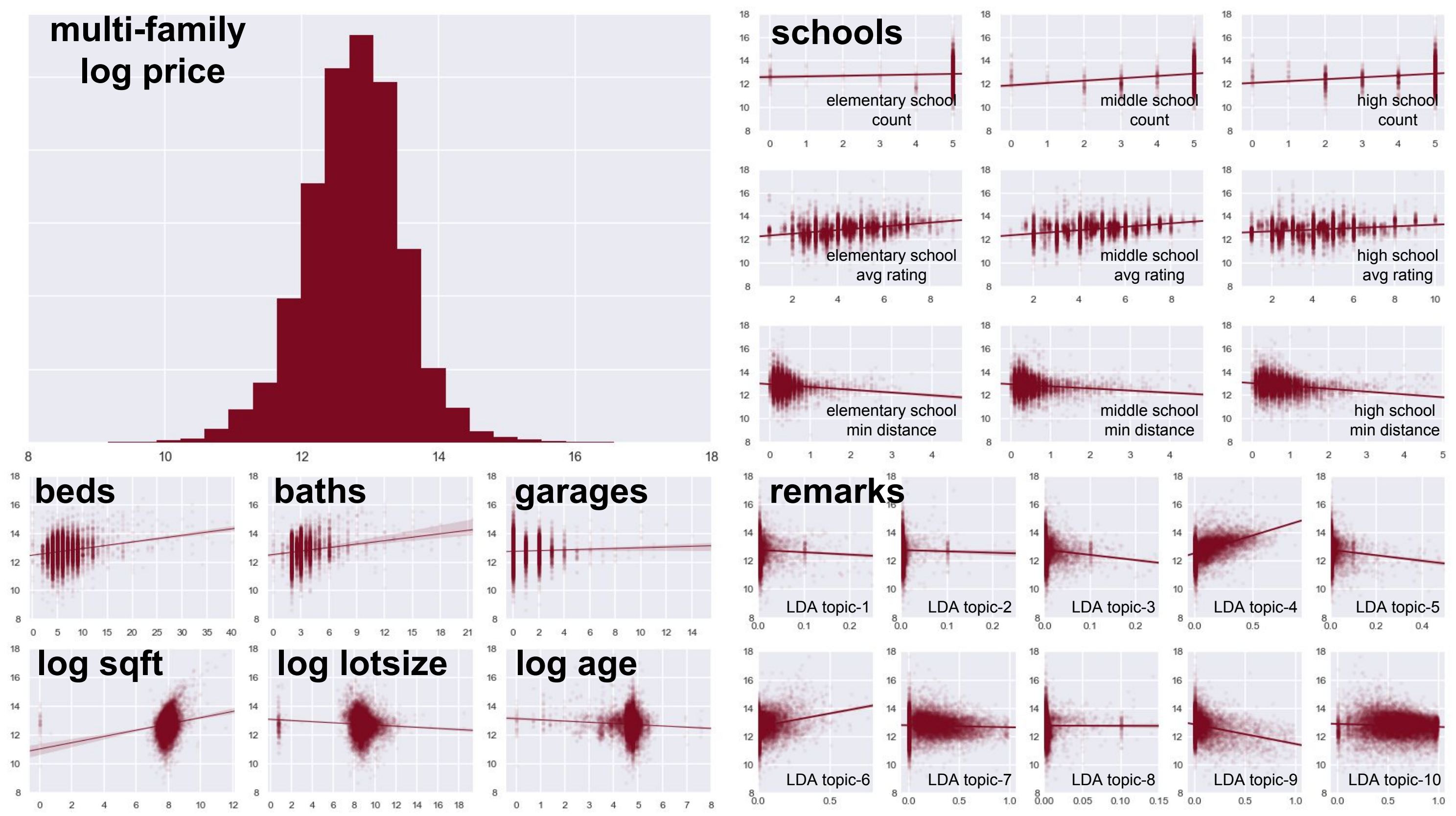
Introduction

Real estate purchases is one of the most substantial investment one can make in life, and the real estate market constitutes a significant part of the overall economy. Therefore, the ability to accurately predict real estate prices and trends is lucrative and valuable.

For this project, we aim to build a predictive model with an emphasis on property images and natural language processing to accurately forecast the **sold price** of real estate properties in Greater Boston Area.

Data Exploration

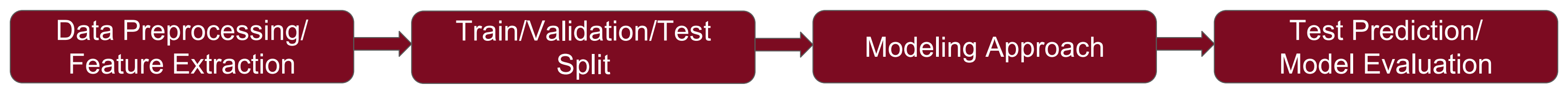




- ★ Price roughly follows log-normal distribution
- ★ The same set of features have different relationships with the price for condos and multi-family properties, so separate models for different property types are appropriate
- ★ Larger properties usually indicate higher sold price, and newer houses with good schools nearby are more popular on the boston market
- ★ Strong relationships are observed between sold price and some of the remark topics.

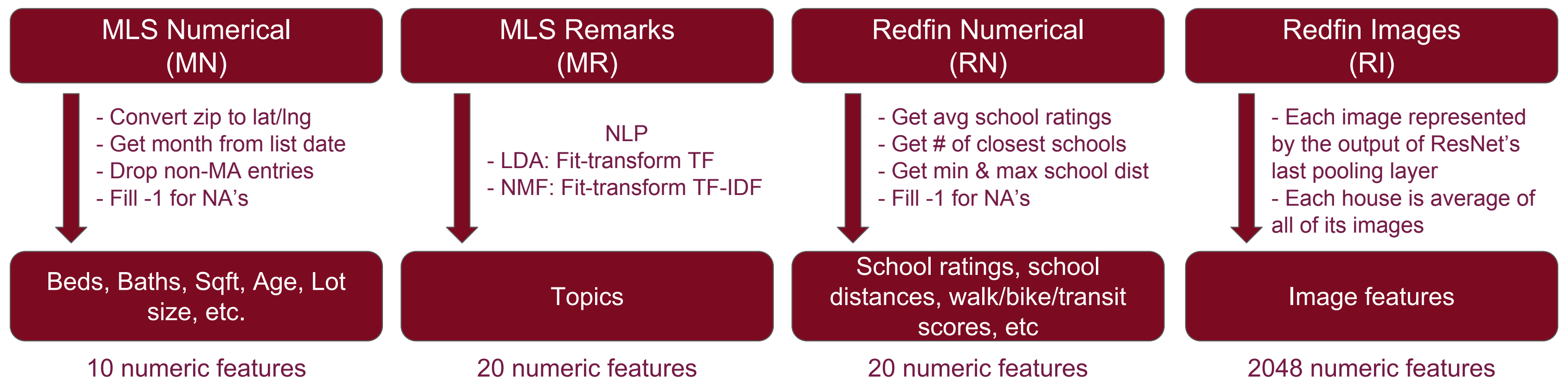
Our Approach

Overall Workflow



Data Preprocessing/ Feature Extraction → Train/Validation/Test Split → Modeling Approach → Test Prediction/ Model Evaluation

Data Preprocessing/Feature Extraction



MLS Numerical (MN)
- Convert zip to lat/lng
- Get month from list date
- Drop non-MA entries
- Fill -1 for NA's

Beds, Baths, Sqft, Age, Lot size, etc.

10 numeric features

MLS Remarks (MR)

NLP
- LDA: Fit-transform TF
- NMF: Fit-transform TF-IDF

Topics

20 numeric features

Redfin Numerical (RN)
- Get avg school ratings
- Get # of closest schools
- Get min & max school dist
- Fill -1 for NA's

School ratings, school distances, walk/bike/transit scores, etc

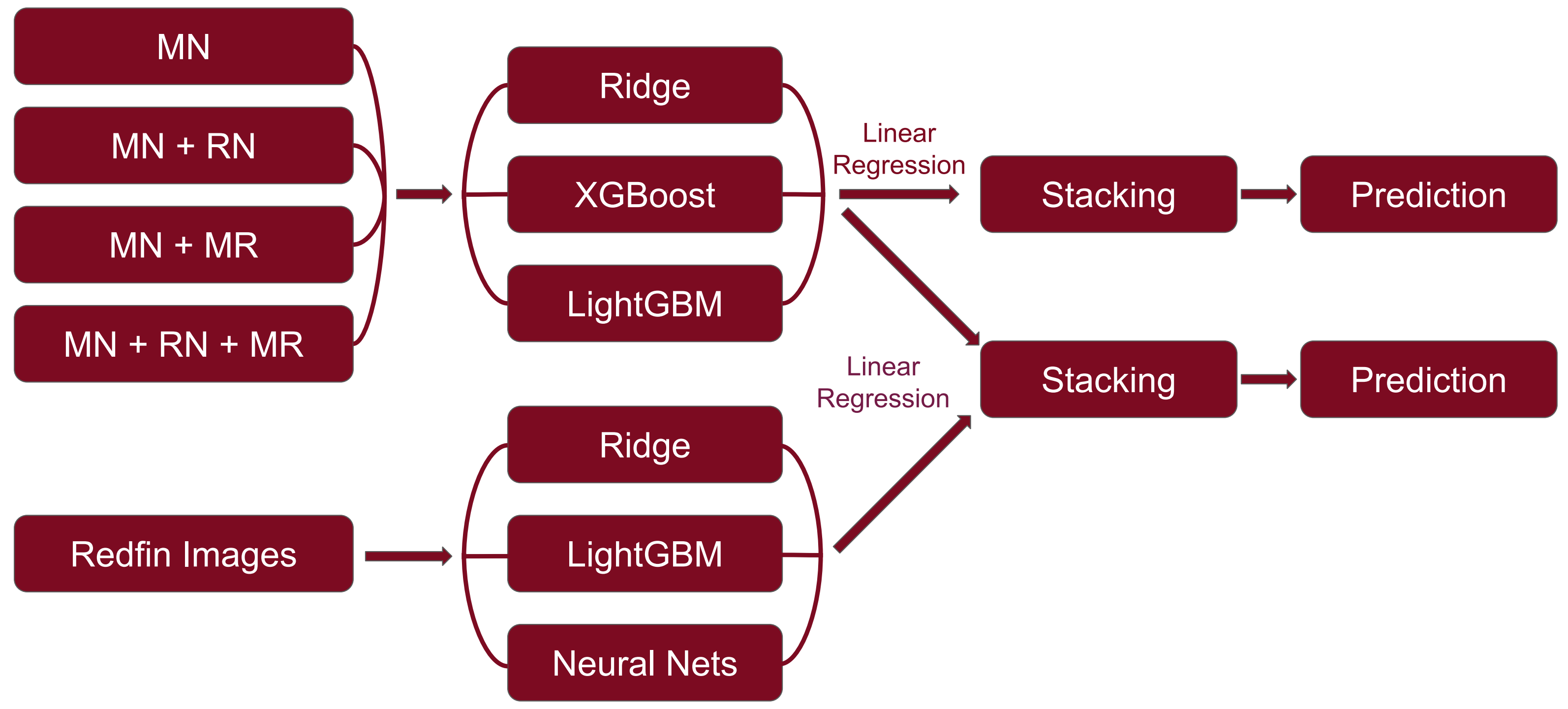
20 numeric features

Redfin Images (RI)
- Each image represented by the output of ResNet's last pooling layer
- Each house is average of all of its images

Image features

2048 numeric features

Modeling Approach



MN

MN + RN

MN + MR

MN + RN + MR

Ridge

XGBoost

LightGBM

Linear Regression

Stacking → Prediction

Linear Regression

Stacking → Prediction

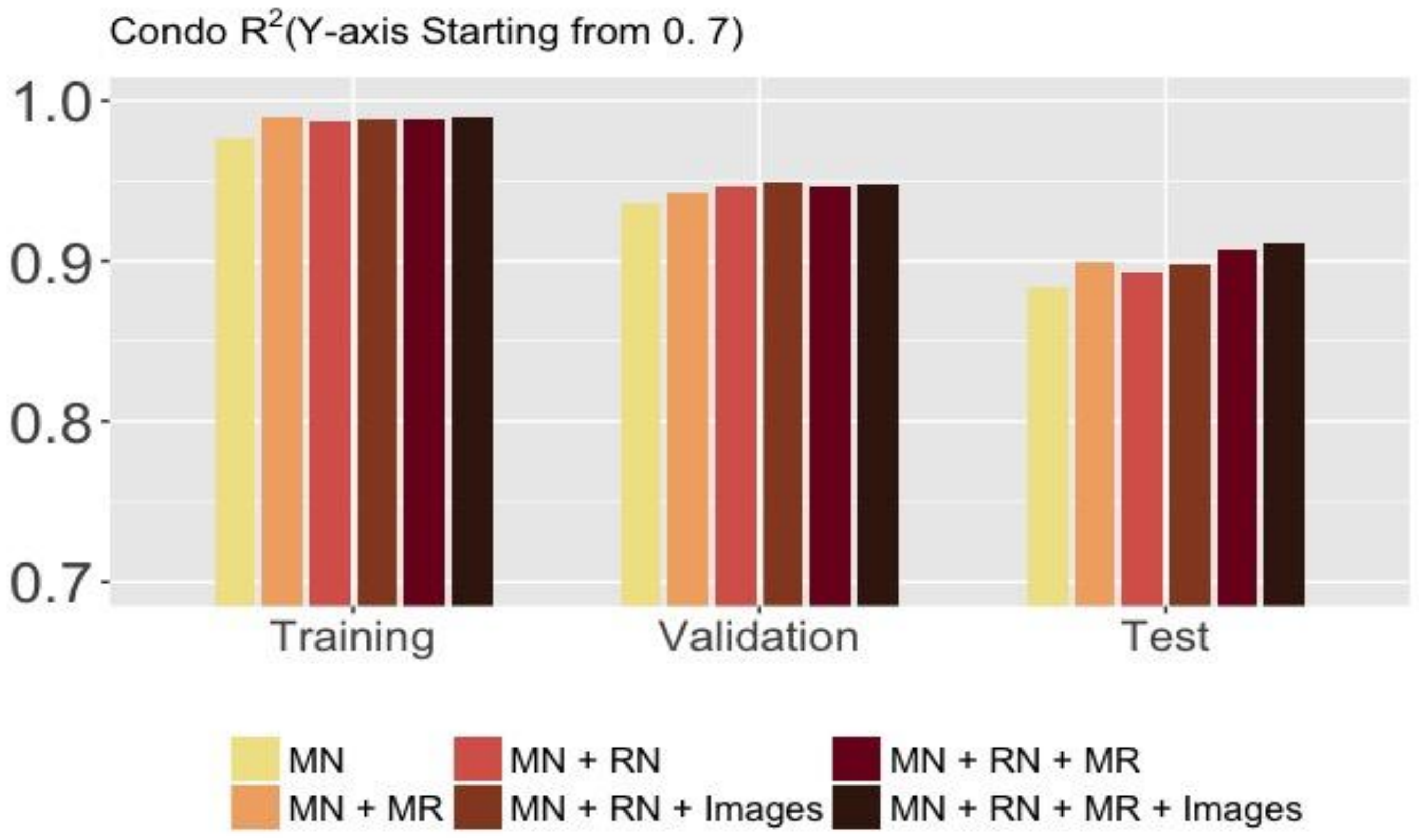
Redfin Images

Ridge

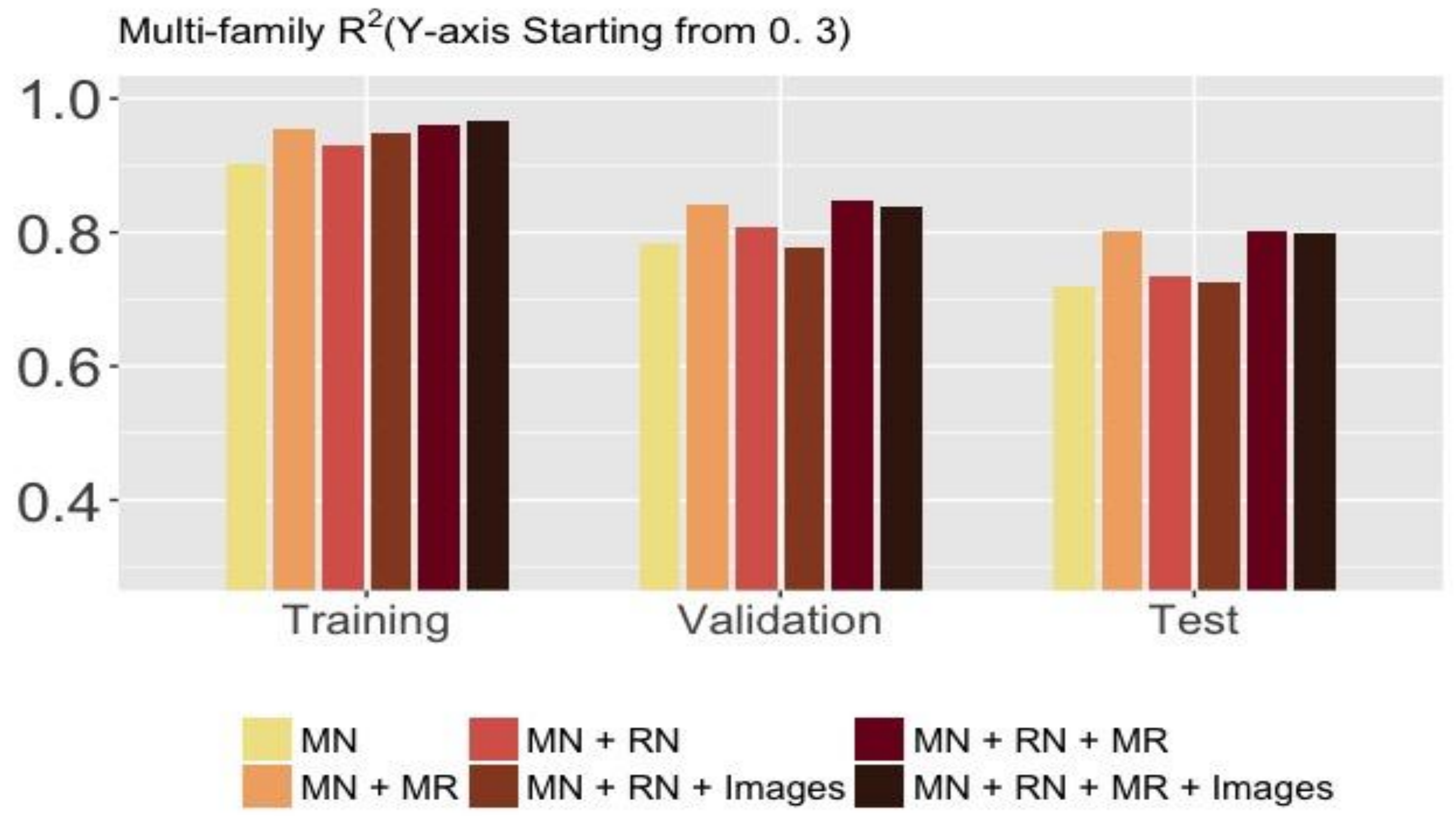
LightGBM

Neural Nets

Results



- ★ Using all features, i.e. MLS and Redfin data combined with images and remarks, produces the best performance for Condo properties
- ★ Redfin data greatly improves R^2 scores for training, validation and test sets, suggesting that it is useful in capturing data variance
- ★ Adding remarks helps to enhance R^2 score, showing that the content of remarks is closely related to property prices
- ★ Images increases the predictive power for modeling condo prices, but only by a small amount



- ★ MLS data combined with Redfin data plus remarks yields the best performance for Multi-Family houses
- ★ Adding images improves training performance but both the validation and the test R^2 scores decrease, suggesting over-fitting
- ★ Better R^2 scores for Condo properties in general due to the high data variance and smaller data size of Multi-family properties
- ★ Slightly lower test performance is expected since our predictions are extrapolations in time, and the overall economy condition is different from year to year

Conclusions and Future Work

- ★ We developed methods:
  1) to curate/scrape information and images from Redfin;
  2) to extract image features from curated property images for prediction;
  3) to extract language features from MLS remarks for prediction.
- ★ We found that features scraped from Redfin, such as transit score, walk score, and school ratings, are highly predictive of property prices.
- ★ We found that language features extracted from remarks using LDA and TF-IDF are predictive of property prices.
- ★ Adding image features extracted using the last pooling layer of ResNet only inconsistently and marginally improved model performance. This suggests that current implementation is sub-optimal. Possible future work is to build a Convolutional Neural Network from scratch using images as input to predict the response variable directly.

References

[1]. Q.You, et al. "Image-Based Appraisal of Real Estate Properties." *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2751-2759, 2017.

IACS

VE RI TAS