# Real Estate Price Prediction

## with MLS and Redfin Data

**Jasmine (Jiawen) Tong**

Team: Jiawen Tong, Michello Ho, Shiyun Qiu, Yiqi Xie
Harvard CS209b Data Science Final Project

# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# To accurately predict real estate prices / DOM ...

- Augment historical real estate transaction data with an emphasis on property images
- Develop feature extraction methods and identify the key factors
- Build predictive models for sold price and number of days on market (DOM)

# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# Data Description

**MLS Data** (Provided)

- MLSNUM
- STATUS
- LISTDATE
- SOLDPRICE*
- DOM*
- ADDRESS, CITY, STATE, ZIP
- LOTSIZE
- AGE
- GARAGE
- REMARKS

**Redfin Data** (Scraped)

- MLSNUM
- beds, baths
- sqft_finished, sqft_unfinished
- year_built, year_renovated
- parking_space, garage_space
- hoa_fee
- school_ratings, school_distances
- walk_score, transit_score, bike_score
- num_photo
- photos posted on Redfin

* : response variable

# Exploratory Data Analysis I - Response Variable



Sold-Price is very right skewed
Single log-transformation makes it more symmetrical

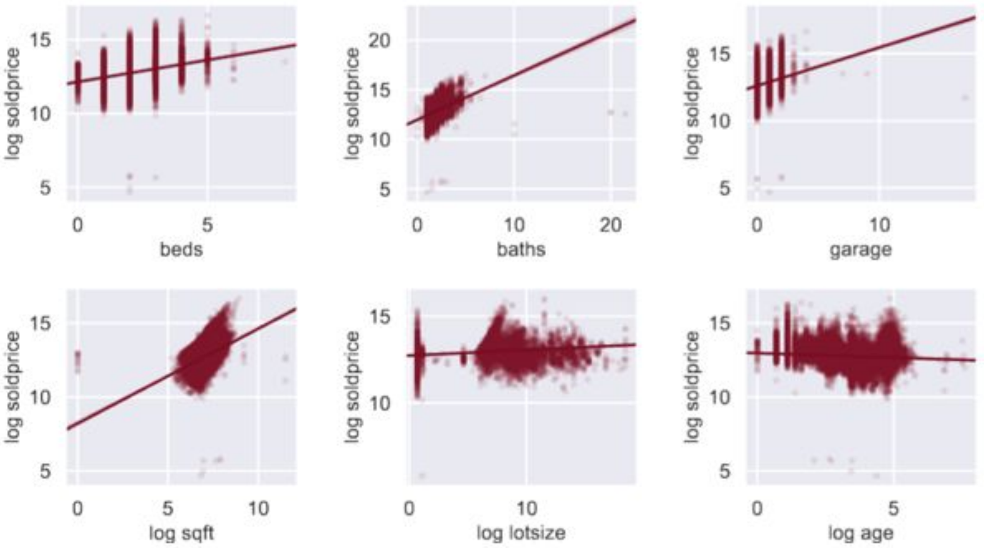
Response: Log(sold-price)

# Exploratory Data Analysis II - Selected MLS Predictors

- Beds, baths, number of parking spaces, square footage and lot size positively correlate with the response

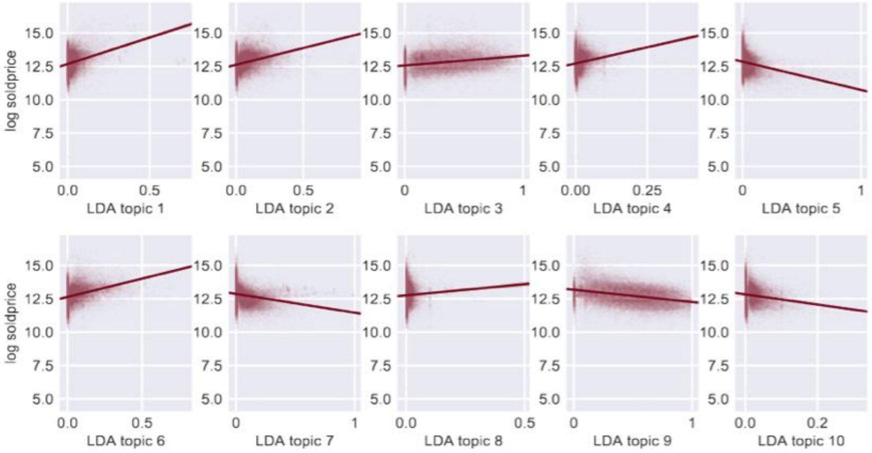
- Property age negatively correlates with the response

**Log Sold-Price vs. Selected MLS Predictors**

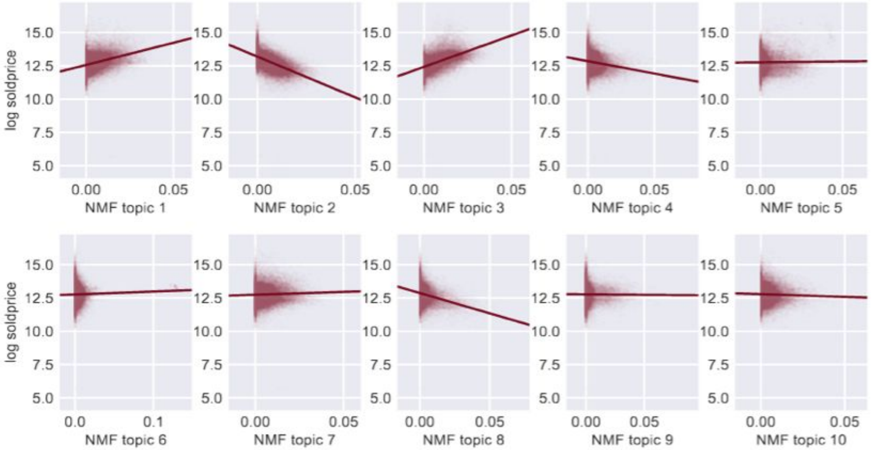


# Exploratory Data Analysis III - MLS Remarks Topics

**LDA topics**




**NMF topics**




- Apply *NLP - Topic Modeling* methods to extract topic features from remarks
- The extracted remark topic features appear to have strong correlation with the response variable

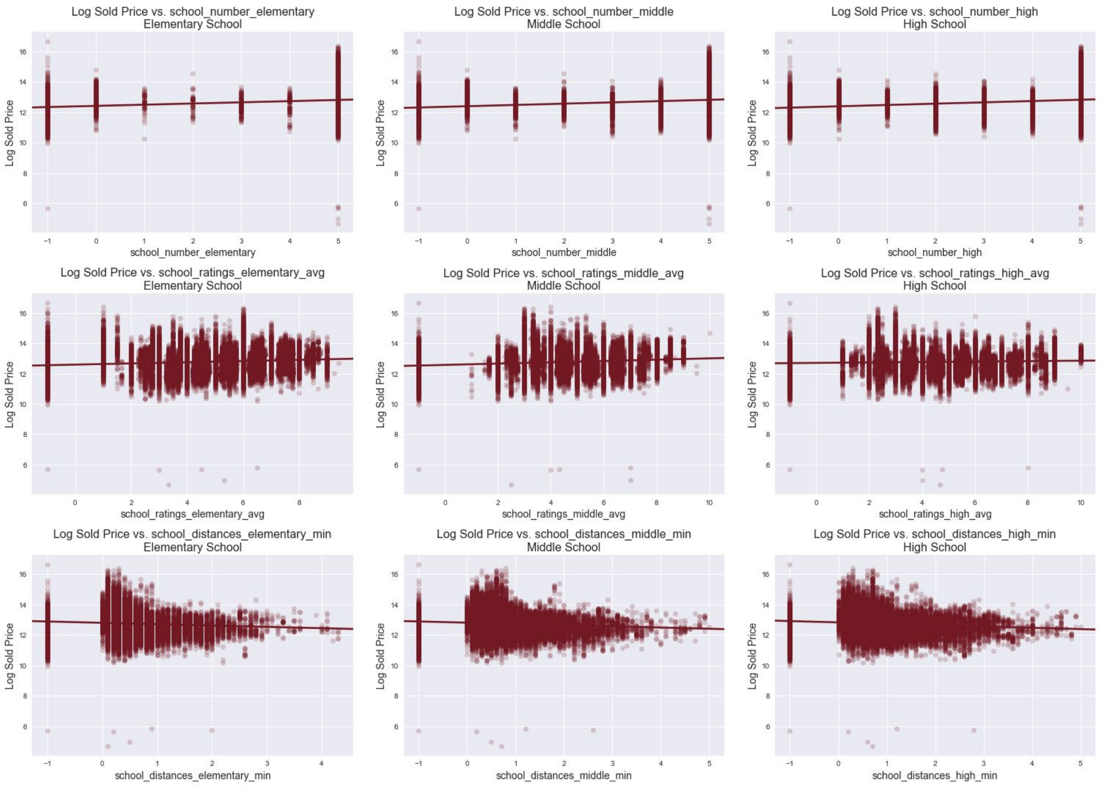
# Exploratory Data Analysis IV - Educational Resources

Elementary School | Middle School | High School

Number of Schools

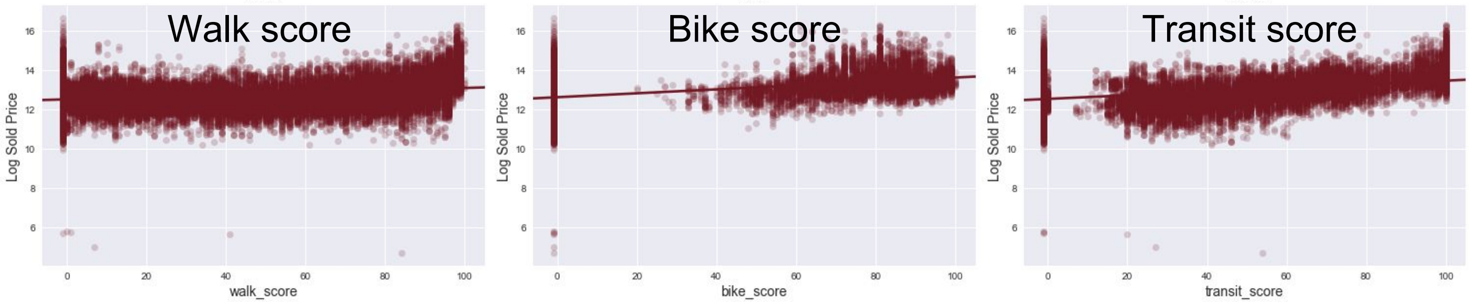
Avg School Rating

Min Distance to School



# Exploratory Data Analysis V - Geographic Location

**Log Sold-Price vs. Convenience Scores**



# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# Approach Overview

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

# Data Preprocessing / Feature Extraction

**MLS Numeric**

- Get Month from list date
- Drop non-MA rows
- Fill -1 for NA's

- **Beds**
- **Baths**
- **Sqft**
- **Age**
- **etc.**

**10 Numeric Features**

**MLS Remarks**

NLP - Topic Modeling

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

- **Topics**

**20 Numeric Features**

**Redfin Numeric**

- Get avg school ratings
- Get # closest schools
- Get min/max school distance
- Fill -1 for NA's

- **School Ratings**
- **School Distances**
- **walk/bike/transit scores**

**20 Numeric Features**

**Redfin Images**

CV – Image Modeling

- Use the output of ResNet50 last pooling layer to represent each image
- Take avg of all its image features for each house

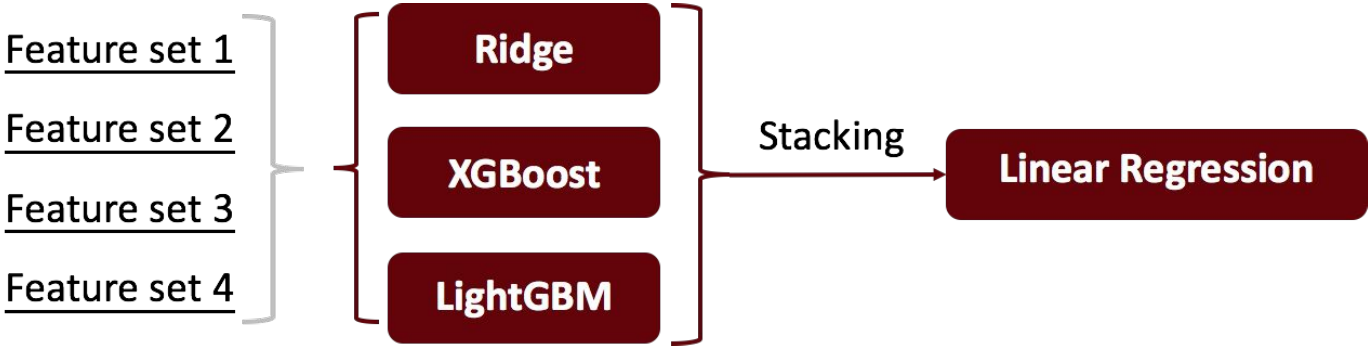
- **Image Features**

**2048 Numeric Features**

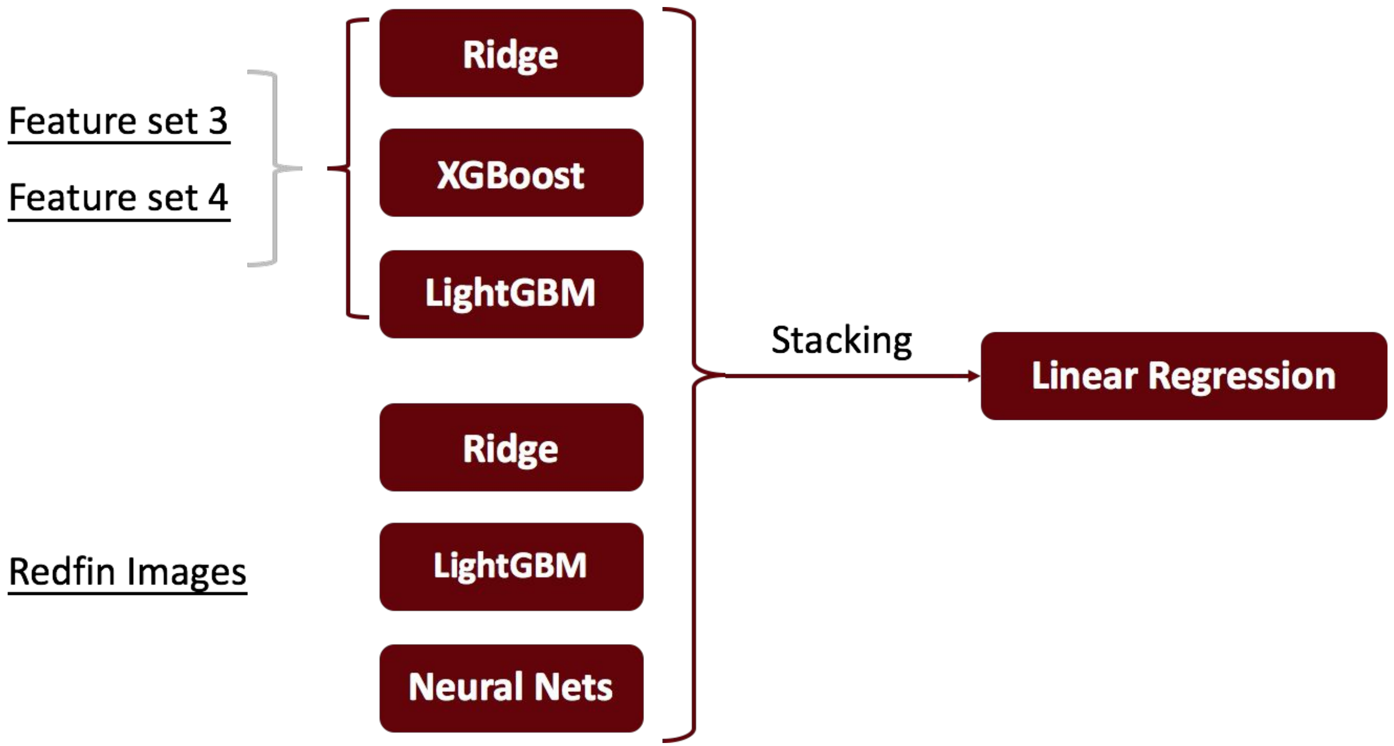
# Feature Sets

| Type | Source | Feature Set | | | | | |
|---|---|---|---|---|---|---|---|
| | | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 |
| Non-image | MLS numerical | X | X | X | X | X | X |
| | MLS remarks | | X | | X | | X |
| | Redfin numerical | | | X | X | X | X |
| Image | Redfin images | | | | | X | X |
| **Total # of features** | | 10 | 30 | 30 | 50 | 2078 | 2098 |

# Modeling Approach I



# Modeling Approach II



# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# Results I - Condo

*Apr 2016 – Dec 2017*
**Train: ~ 36000**
**Validation: ~ 4000**

*Jan 2018 – Mar 2018*
**Test: ~ 1300**

**Condo: Price $R^2$** (Ensemble Model)

| | | Training | Validation | Test |
|---|---|---|---|---|
| Feature set 1 | MLS | 0.977 | 0.936 | 0.884 |
| Feature set 2 | MLS + Remarks | 0.990 | 0.942 | 0.899 |
| Feature set 3 | MLS + Redfin | 0.987 | 0.947 | 0.893 |
| Feature set 4 | MLS + Redfin + Remarks | 0.989 | 0.947 | 0.907 |
| Feature set 5 | MLS + Redfin + Images | 0.988 | 0.949 | 0.898 |
| Feature set 6 | MLS + Redfin + Remarks + Images | 0.990 | 0.948 | 0.911 |

# Results II - Multi-Family

*Apr 2016 – Dec 2017*
**Train: ~ 12000**
**Validation: ~ 1300**

*Jan 2018 – Mar 2018*
**Test: ~ 1300**

**Multi-family: Price** $R^2$ (Ensemble Model)

| | | Training | Validation | Test |
|---|---|---|---|---|
| Feature set 1 | MLS | 0.903 | 0.782 | 0.718 |
| Feature set 2 | MLS + Remarks | 0.956 | 0.841 | 0.801 |
| Feature set 3 | MLS + Redfin | 0.930 | 0.807 | 0.736 |
| Feature set 4 | MLS + Redfin + Remarks | 0.961 | 0.849 | 0.803 |
| Feature set 5 | MLS + Redfin + Images | 0.947 | 0.777 | 0.724 |
| Feature set 6 | MLS + Redfin + Remarks + Images | 0.967 | 0.837 | 0.800 |

# AGENDA

- Problem Statement & Motivation
- Data Description
- Exploratory Data Analysis
- Approach
- Results
- Conclusions
- Future Work

# Conclusions

- Developed:
  - topic feature extraction methods using NMF and LDA
  - a method to scrape property data and images from Redfin
  - a method to extract visual features from property images (the average 2048-dimensional ResNet final average pooling layer output)
- Found:
  - that both transformed remark topic features and information from Redfin are useful features for predicting the sold price
  - that our current method of extracting images is likely sub-optimal

# Future Work

- Gather more multi-family observations to reduce overfitting and improve model generalizability.

- Develop a better approach to incorporate zip code information:
    - Join open census data

- Gather additional features from external sources to try to capture market temperature and the overall economy.

- Develop a better method of incorporating image features:
    - (image_feature) X (numeric_feature)