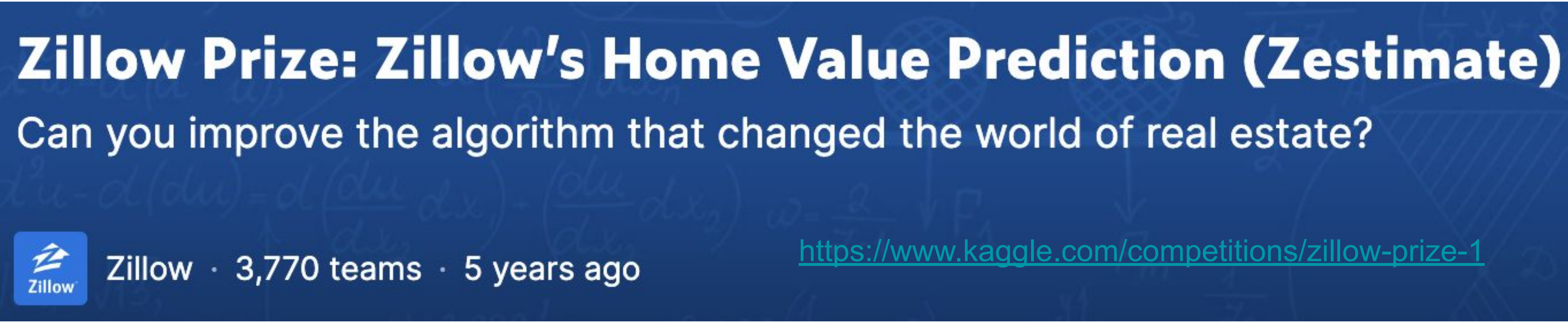
Capstone III: Zestimate

Predicting the errors of Zillow’s home price calculator



Zillow Prize: Zillow’s Home Value Prediction (Zestimate)

Can you improve the algorithm that changed the world of real estate?

Zillow · 3,770 teams · 5 years ago

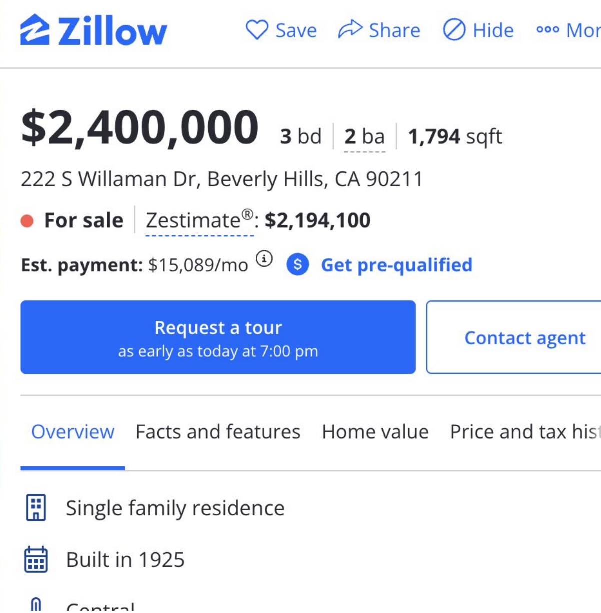
https://www.kaggle.com/competitions/zillow-prize-1

The Goal

Can we predict the error of Zillow’s “Zestimate”?

Specifically we will predict the log error of the estimate based on this calculation:

$$logerror = log(Zestimate) - log(SalePrice)$$



Zillow

Save | Share | Hide | Mor

$2,400,000 **3** bd | **2** ba | **1,794** sqft

222 S Willaman Dr, Beverly Hills, CA 90211

● **For sale** | Zestimate®: **$2,194,100**

Est. payment: $15,089/mo ⓘ **Get pre-qualified**

Request a tour
as early as today at 7:00 pm

Contact agent

Overview | Facts and features | Home value | Price and tax hist

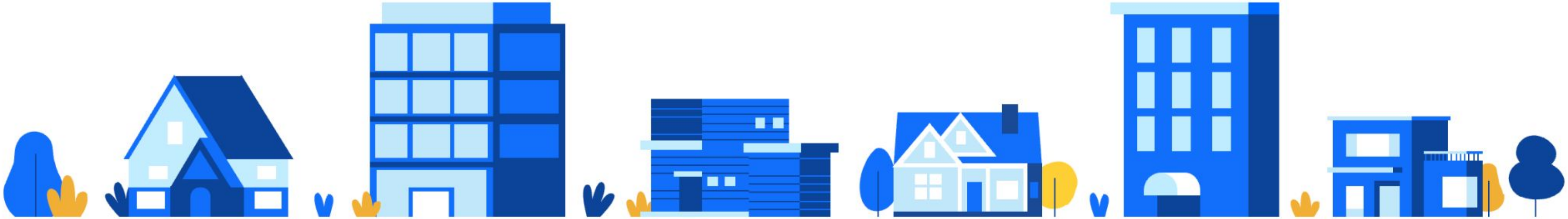
Single family residence

Built in 1925

Central

Why this matters

- Homes are typically the most expensive purchase a person makes.
- Zillow has over 200 million monthly users
- Zillow needs to keep is estimate accurate due to competition from other online home marketplaces

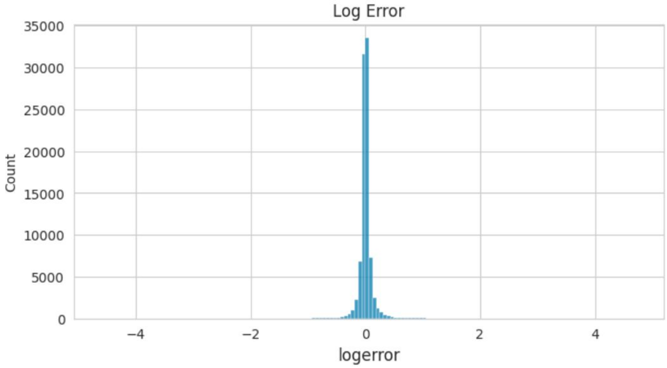


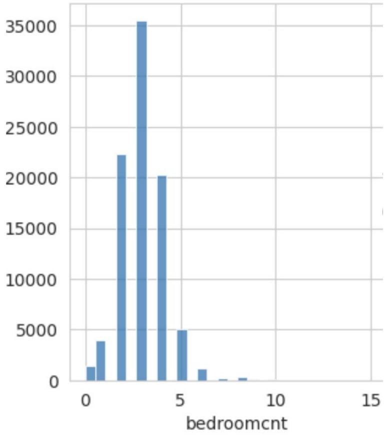
The Data

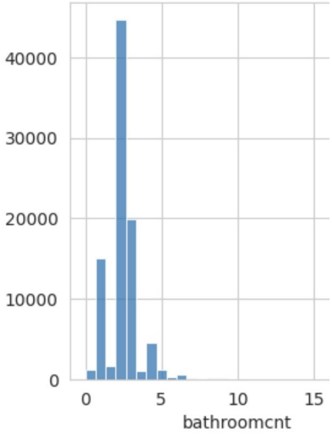
- Zillow provided data on nearly 3 million homes in LA, Orange and Ventura counties in 2016.

- Each record contained over 50 features, such as square footage, the number of rooms, the existence of a pool, etc.

- These features were merged with data from each home that was sold in 2016, including transaction date and the log error of the Zestimate.







Data Wrangling

- There was a mix of numerical and categorical data.
- Columns were renamed to be more comprehensible.
- Duplicate and overlapping features were thinned out.
- The percentages of null values ranged from 1 to 99%.
  - The median, zero, and ‘not given’ were imputed for missing values when appropriate.
- Categorical data was prepared for one-hot encoding.
  - Some data, such as zoning codes, had too many unique instances and needed to be reduced using ‘other’ categories
- After data wrangling, there were 19 numerical columns, 30 categorical columns, and over 90,000 records.

Numerical Columns

- Transaction dates
- Logerror
- Bedroom and bathroom counts
- Square footage
- Garage, fireplace counts
- Taxes

Boolean Columns

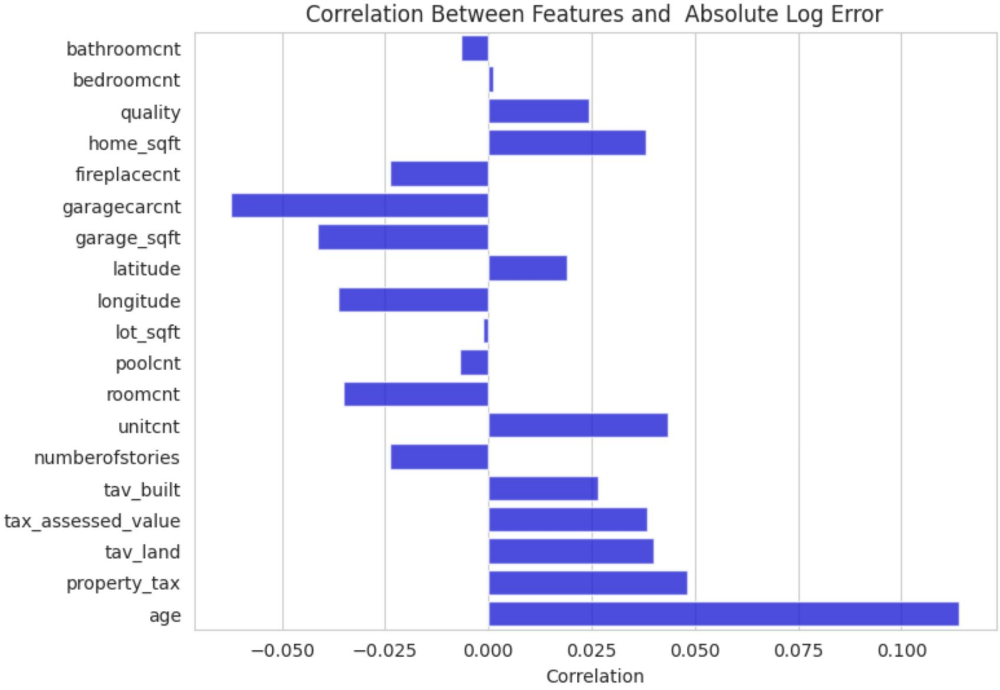
- 'has_spa',
- 'pool_with_spa',
- 'pool_without_spa',
- 'fireplaceflag',
- 'taxdeliquencyflag'

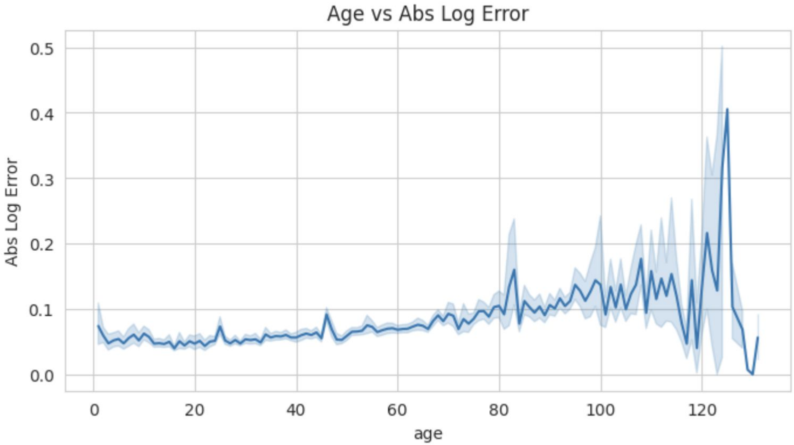
Categorical Columns

- 'aircon'
- 'architecture'
- 'framing'
- ‘quality'
- 'deck'
- 'heating'
- 'material'
- 'storytypeid'
- 'latitude'
- 'longitude'
- 'censustractandblock'
- 'fips'
- 'city’
- 'county'
- 'neighborhood'
- Zipcode'
- 'unitcnt'
- 'Numberofstories'

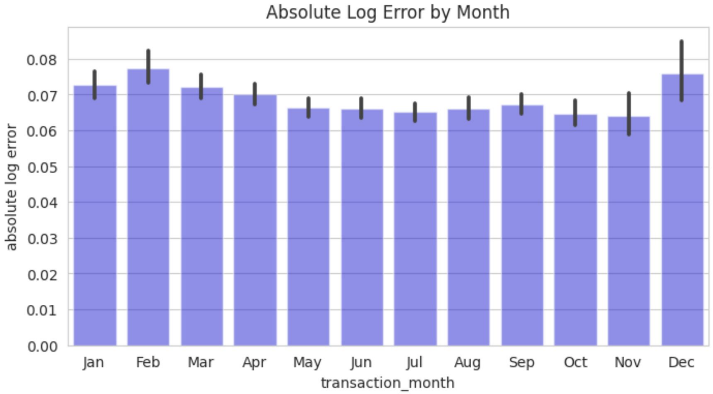
Exploratory Data Analysis

- Examined links between features and log error or the absolute value of log error



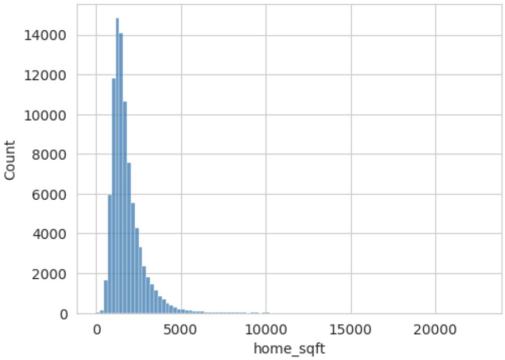


The sale price of older homes tended to be more difficult to predict.

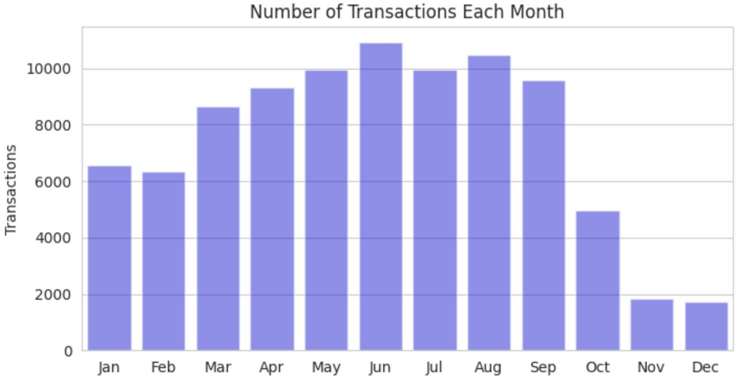


Exploratory Data Analysis

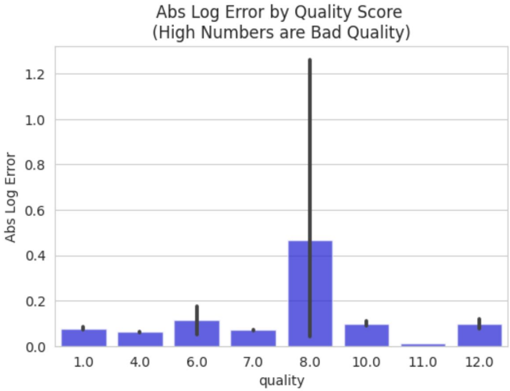
- Examined the distributions of specific features



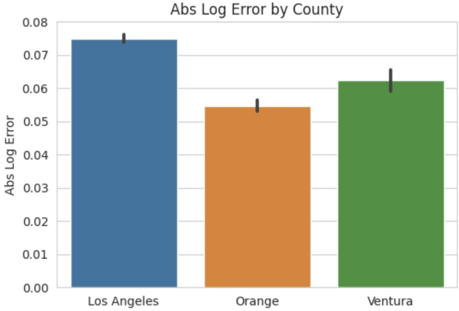
Many features skewed right.



The winter months saw the fewest transactions.



“Quality” was an assessment of the condition of the building. Eight was by far the most common score, and the Zestimate seemed to struggle with that particular value.



Some codes led to more information

The ‘rawcensustractandblock’ column held information about state, county, and blocks, which could then be linked to demographic data from the census.

Raw Census Tract and Block: 61110010.011023

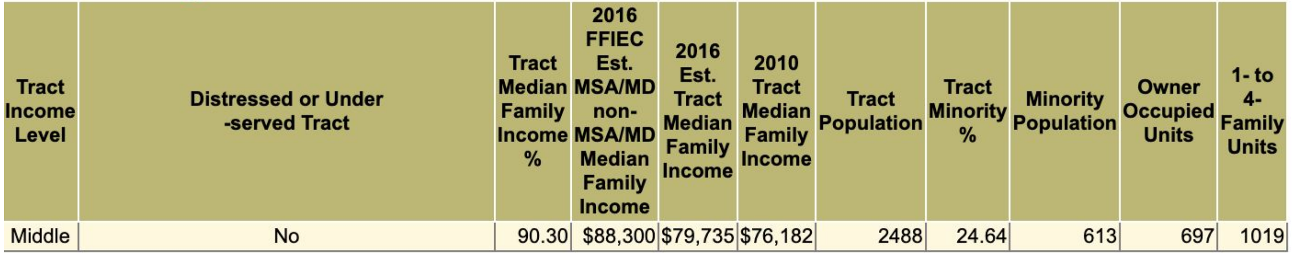
California 6

Ventura County 111

Tract 0010.01

Block 10123

a. https://www.ffiec.gov/census/report.aspx?year=2016&county=111&tract=0010.01&state=06&report=demographic



Tract Income Level	Distressed or Under -served Tract	Tract Median Family Income %	2016 FFIEC Est. MSA/MD non-MSA/MD Median Family Income	2016 Est. Tract Median Family Income	2010 Tract Median Family Income	Tract Population	Tract Minority %	Minority Population	Owner Occupied Units	1- to 4- Family Units
Middle	No	90.30	$88,300	$79,735	$76,182	2488	24.64	613	697	1019

Ultimately, I decided not to parse this data since there were other columns such as zip code, county, latitude and longitude.

Model Selection

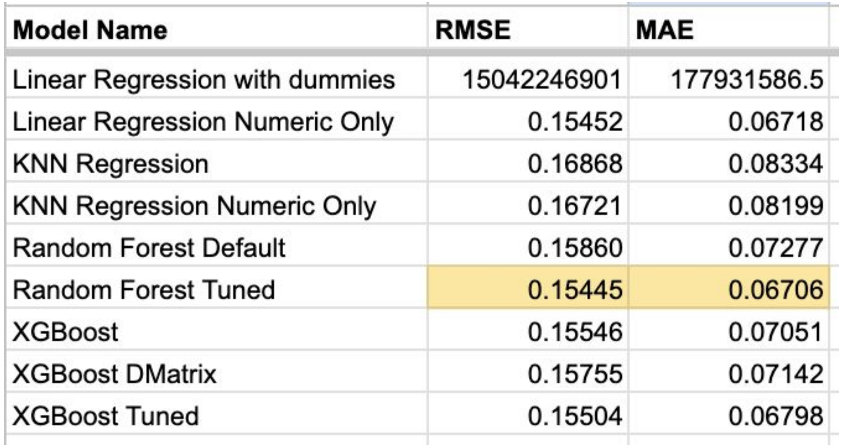
I experimented with the following regression models:

- Linear Regression
- K-Nearest Neighbors
- Random Forest
- XGBoost

I made adjustments to the hyperparameters and the number of features used.

Overall, a Random Forest Regressor performed the best of all models after hyperparameter tuning.

However, the basic linear regression model performed nearly as well using only the numerical features.



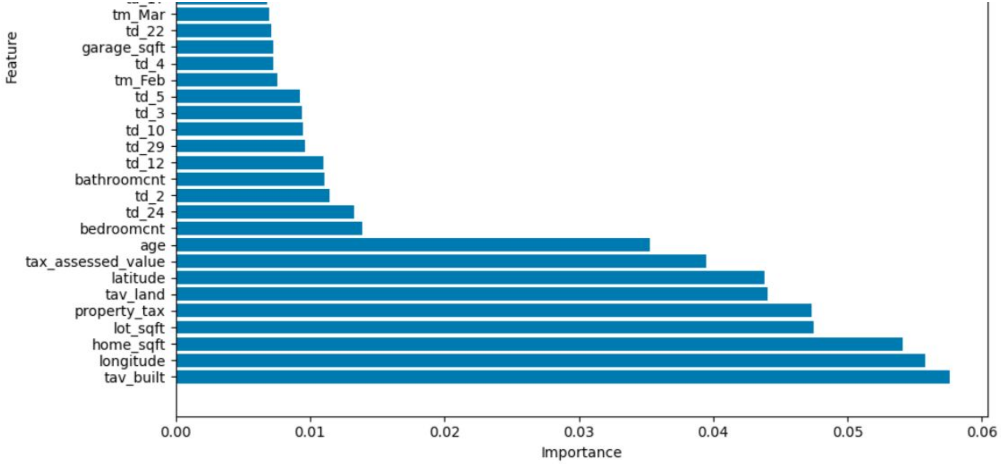
Model Name	RMSE	MAE
Linear Regression with dummies	15042246901	177931586.5
Linear Regression Numeric Only	0.15452	0.06718
KNN Regression	0.16868	0.08334
KNN Regression Numeric Only	0.16721	0.08199
Random Forest Default	0.15860	0.07277
Random Forest Tuned	0.15445	0.06706
XGBoost	0.15546	0.07051
XGBoost DMatrix	0.15755	0.07142
XGBoost Tuned	0.15504	0.06798

Takeaways and Future Research

- Although the Random Forest Regressor had the lowest Mean Absolute Error, the Linear Regression might be a better option due to its simplicity and speed.

- The numerical columns were the most important features.
- Further testing could be done using only the most important features.

- There is opportunity to bring in more data using background information on zoning codes or census data.


Feature
tm_Mar
td_22
garage_sqft
td_4
tm_Feb
td_5
td_3
td_10
td_29
td_12
bathroomcnt
td_2
td_24
bedroomcnt
age
tax_assessed_value
latitude
tav_land
property_tax
lot_sqft
home_sqft
longitude
tav_built
Importance