Project Kojak

Will Hetfield

Executive Summary

Project is home price prediction on a new web scraped data set and incorporating text features from home listing description

Redfin for three geographic regions (LA, San Diego, Seattle) was the data source for approximately 25,000 transactions over three years ending in first quarter of 2018

Used Seleniuim with multiprocessing to scrape Redfin URLs to get data

Non-Text Features in four broad categories:

1. Basic Home Data (Beds, Baths, Square Footage, etc.)
2. Geographic (Latitude, Longitude, Flags for 3 regions)
3. Zip Code Level Demographic Info (Education, Income, Public Schools)
4. Lagged Market Comparables

NMF Topic Modeling performed on scraped home descriptions. Each topic description vector was normalized (divided by sum of topic values in the vector) so all the topic features for a particular home summed to 1

Grid Search for hyperparameter tuning was done for Random Forest and Gradient Boosted Regressor using 64% of the data. With selected hyper parameter values, models incorporating text features were trained on same 64% and evaluated on 16% of data. A nominal and log of the home price were used as targets with the evaluation criteria being the median absolute error of the home price.

Gradient Boosted Regressor with 10-Topic NMF text features trained on the log of the home price was the best performing model in validation.

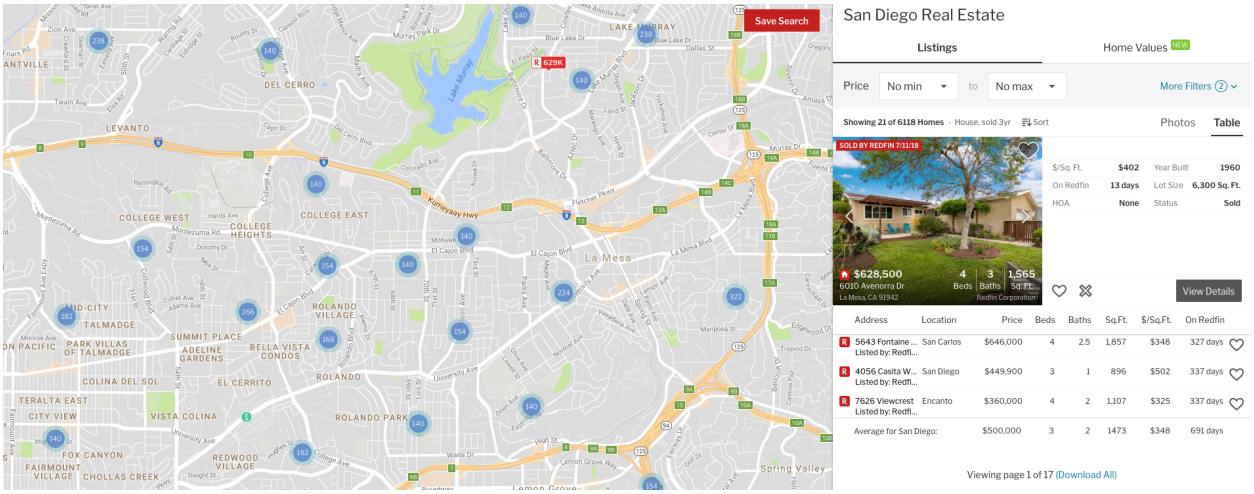
Median absolute error on the full holdout set of the most recent 20% of the transactions was 7.9%

Background / Motivation

I wanted to build a home price prediction model with two new components from what was available online: 1) Create a new data set from recent transactions rather than what is available online from an old Kaggle competition, for instance 2) See if text features from a home price listing could add any predictive power to a model

Data_Acquisition

Redfin does allow CSV data downloads of up to 350 transactions in the map view a user scrolls to. However, the user cannot control which homes are downloaded if there more than 350 homes in the visible area of the map



So while in the above map, there are perhaps thousands of transaction which we would like data on, the download all button on in the bottom right hand side of the picture would only give data for 350 of the transactions. By clicking on each of the blue cluster buttons in the map, we would zoom in on the map and see fewer homes, and thus the limit of 350 homes would represent a larger portion of the homes in the visible area on the map.

We could continually zoom into and out of different areas to get CSVs of 350 transactions. This process can be automated through Selenium. An automated web browser can be launched which will progressively click through the different clusters and

hit the download csv file. I provided approximately 50 high level URLs from which hundreds of csv files were downloaded through automating this process.

Each csv file contained basic info such as address, sale price, zip code, beds, baths, latitude and longitude and square footage. Additionally, it has the unique Redfin url for the property page for the house. This url has the text description from the listing that was used to generate text features.

This was an additional automated workflow with Selenium, having the browser go to each of the 25,000 property urls and scraping some more information, most importantly the text property description the agent used.

Summary of Non-Text Features

Basic Home Date

Beds, Baths, Square Footage, Lot Size, Age of House

Geographic

Latitude, Longitude, Flags (1 or 0) for either LA, San Diego, Seattle areas

Zip Code Level Demographics

Population, Education (% Bachelors, % Graduate), Income, Public School Rating

Lagged Market Comparables

Price based on $ per Square Foot in Zip Code over Prior 6 Months
Case Shiller National Price Index

The basic home data was acquired and geography was acquired in the scraped CSV files from Redfin. Zip Code Level Demographics were scraped from city-data.com. For lagged market comparables, an estimated price calculated as the average price per square foot in the home's zip code in the six months prior to listing times the listed home's square footage was feature engineered. Additionally, the Case Shiller National Price Index was included to capture some of the general trend in home prices during the

time period. So the feature for a home listed in July 2016 would be the level of the home price index for May 2016 (most recent data available at the time of listing).

Text Features

Because Redfin is a brokerage, its site provides access to the Multiple Listing Service and listing descriptions used, even after the home is sold. Below is an example:

VR: $375,000-$399,000! Large sun drenched living space perfect for entertaining! The living room has a gorgeous exposed brick fireplace making a beautiful accent wall in the room! Newly installed hardwood flooring! An additional living area with plenty of room perfect for kicking back and relaxing! Large kitchen with combined dining and updated appliances! The kitchen also provides backyard access! The backyard is perfect for soaking up the sun and summer BBQ's! Master bedroom with en-suite!

In order to analyze descriptions, we need to vectorize them / make them numeric. The most basic model is the bag of words model, where we turn each individual document (the term for an observation in the NLP context) into a vector based on word counts in the document. Each column can be a word in the document, and we simply register a 1 or 0 if the particular word is in the observation / document. For this analysis, instead of the bag of words representation, I used the Term Frequency - Inverse Document Frequency (TF-IDF) which takes the word count method described above but also adjusts for the frequency the words appear in the overall corpus of documents, thus placing more importance on words that are unique.

This results in a very large, sparse matrix of mostly zeroes as only very few of the words will be any given document / observation. I utilized a few standard techniques for what “counts” as a term to reduce the size of the document term matrix and give it more meaningful features .

Stopwords: Very common words which (a,an,the) which can be eliminated from every document as they don’t provide any meaning

Stemming: Transforming variations of the same word into a single root , ie, “running”, “ran” , “run” all become “run”

Minimum document frequency: Require a minimum number occurrences within all documents in order for a term to be included in all documents.

The resultant matrix can then be used to generate features for analysis. Nonnegative Matrix Factorization (“NMF”) is a linear algebra numeric method which can be applied to a Term Frequency Inverse Document Frequency Matrix. It finds two matrices whose product is approximately equal to the original matrix, with the restriction that no terms in any of the three matrices can be negative. In the document-term matrix context, the TF-IDF matrix is factored into two matrices, a topics matrix of documents (rows) by topics (columns), where the number of topics is specified for the algorithm, and a topic by terms matrix, which can be viewed as the weights of each term in the topic space. There are different cost functions / optimization constraints which can be used to calculate the “best” two matrices. In this, case the Frobenius norm was used, which can be thought of as minimizing the squared difference in a matrix space (norm) between the original document-term matrix and the product of the two factors from NMF.

```
Topic 1:
room live room live famili dine famili room dine room larg area fireplac
formal bedroom formal dine patio kitchen room fireplac spaciou formal din
e room laundri open door pool laundri room lead garag dine area bathroom
breakfast featur car
Topic 2:
new paint new kitchen brand new brand floor new remodel carpet new roof r
oof new carpet interior new floor paint new new paint new window exterior
window new fixtur quartz kitchen new complet bathroom new bathroom applia
nc new cabinet counter new door roof new electr
Topic 3:
thi home thi home ha home ha thi home ha make beauti come perfect bedroom
offer make thi love home thi readi home featur look miss charm welcom thi
beauti featur great space just bathroom thi home featur enjoy entertain
Topic 4:
level updat main bath fenc basement hardwood lower rm lower level ga park
deck fulli floor street charm hardwood floor garden fenc yard storag main
floor bed fulli fenc garag craftsman yard space finish bdrm
Topic 5:
view deck open design lake modern plan enjoy floor plan entertain light p
rivat mountain home outdoor thi space floor stun live natur open floor op
en floor plan citi downtown washington offer hill high expans
Topic 6:
stainless steel stainless steel granit counter tile steel applianc stainl
ess steel applianc floor upgrad applianc granit counter dual newer window
remodel pane cabinet light wood pane window dual pane custom dual pane wi
ndow kitchen recess beauti recess light featur tile floor
Topic 7:
locat school shop freeway close park great access restaur neighborhood co
nveni home easi distanc center walk distanc nice bedroom quiet near locat
locat conveni locat central car walk singl veri easi access shop center g
arag
Topic 8:
properti buyer lot sq ft sq ft hous sold verifi opportun thi properti inf
orm potenti seller permit need squar build great condit thi investor zone
agent properti sold bath ft lot properti ha citi befor
Topic 9:
master suit walk closet walk closet master suit bath bedroom upstair tub
custom island shower master bath master bedroom built ceil main featur pa
ntri spa larg vault downstair ga vault ceil luxuri includ kitchen open
Topic 10:
continu read read continu fixer need larg rv park solar lot cover potenti
san diego diego great height famili rv park huge mesa cul cul sac backyar
d sac sold home loft readi prior stori
```

Some of the stronger topics intuitively appear to be 2 (new furnishings / remodeling), 5 (dealing with unique views / location) and 8 (targeted at investors, ie, references to make an offer, verification, square footage, permits)

Data Split and Model Tuning

64% of the data used for performing grid search and selecting hyperparameter values. No text features were included when selecting the best parameters and the models were trained on nominal price.

Random Forest Parameter Grid

```
param_grid =  { "n_estimators" : [100, 500, 1000],
                "max_depth" : [10,30,50],
                "max_features": [None, "sqrt"],
                "min_samples_leaf" : [1, 5, 10]}
```

Random Forest Best Parameters

```
{'max_depth': 50,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'n_estimators': 1000}
```

Gradient Boosting Regressor Parameter Grid

```
param_grid = {'learning_rate': [0.1, 0.05, 0.02, 0.01],
              'max_depth': [1, 3, 5],
              'n_estimators': [100, 500, 1000],
              'max_features': [None ,'sqrt'] }
```

Gradient Boosting Regressor Best Parameters

```
{'learning_rate': 0.1,
 'max_depth': 5,
 'max_features': 'sqrt',
 'n_estimators': 1000}
```

Using hyperparameter values from grid search, models incorporating text features and a were trained on target of nominal home sale price and the log of the home sale price

By doing this, essentially treat the inclusion of text features and whether to log the home sale price target as hyper parameters which are tuned.

Training was done on the same 64% of the data and models were evaluated on 16% of the data. The evaluation metric was the median absolute error as the percentage of the home price, which is the typical evaluation criteria in online portal home price prediction estimates

Model	Target	Text Features	Median Absolute Error
Zip Code Comps	Nominal Price	None	14.6%
Linear Regression	Nominal Price	None	14.0%
Random Forest	Nominal Price	None	7.3%
Random Forest	Log Price	None	7.1%
Random Forest	Log Price	20-Topic NMF	7.8%
Random Forest	Log Price	10-Topic NMF	7.3%
Gradient Boosting Reg	Nominal Price	None	7.1%
Gradient Boosting Reg	Log Price	None	7.0%
Gradient Boosting Reg	Nominal Price	10-Topic NMF	7.0%
Gradient Boosting Reg	Log Price	10-Topic NMF	6.7%
Gradient Boosting Reg	Log Price	20-Topic NMF	6.9%

Gradient Boosting Regressor with 10-Topic NMF text features trained on the log of the home price was selected as the best model through this validation process. The remaining 20% of data, representing the most recent observations, was then used to evaluate this model's performance on a true out of sample data. The median absolute error as a percentage of the home price in the full holdout was 7.9%.

Feature Importance Scores in Holdout Set

Feature importance scores measure which features of the trees in the gradient boosted regressor contribute most to the reduction in the loss function. These scores are normalized so they sum to 1. We can see that there is no feature that overwhelmingly dominates. Three text features are in the top 10 of feature importances. Two of these features deal with what can be described in extremes in quality. Topic 5 had to deal with exceptional views and locations, and Topic 8 had to deal with properties targeted at investors. Some evidence suggesting that utilizing text can be a useful way to capture these extremes in house quality.

```
feature_importances = gbr.feature_importances_
sorted(zip(feature_importances,features), reverse=True)
```

```
[(0.08585142622823085, 'LONGITUDE'),
 (0.07417378654828236, 'LATITUDE'),
 (0.06685291195546261, 'Price_Based_on_Comps'),
 (0.0535386535493304, 'SQUARE FEET'),
 (0.05036713923757738, 'Topic5_10_Adj'),
 (0.0492790728255577, 'LOT SIZE'),
 (0.04841480370834834, 'Age_of_House'),
 (0.047014198785786956, 'Home_sqft_per_lot_sqft'),
 (0.039472680370878595, 'Topic8_10_Adj'),
 (0.03685517129094154, 'Topic4_10_Adj'),
 (0.03400665757711701, 'Lagged_Case_Shiller_Index'),
 (0.032548194238468726, 'Topic3_10_Adj'),
 (0.029794509313772585, 'School_Rating'),
 (0.029102033949144363, 'Topic7_10_Adj'),
 (0.028923964940674226, 'Topic9_10_Adj'),
 (0.02755412828760185, 'Topic1_10_Adj'),
 (0.027048906938919415, 'Grad_%'),
 (0.026873457439823743, 'Photos'),
 (0.026696713424588236, 'Married_%'),
 (0.025590894058159966, 'BA_%'),
 (0.024820001076151166, 'Topic10_10_Adj'),
 (0.02425900150035176, 'Median_Income'),
 (0.023656477078706196, 'Topic2_10_Adj'),
 (0.022947996442407987, 'BATHS'),
 (0.02266282286309374, 'Population'),
 (0.021087798240653842, 'Topic6_10_Adj'),
 (0.014984090731767376, 'BEDS'),
 (0.0022312770408642953, 'Seattle_Area'),
 (0.001862844063744332, 'LA_Area'),
 (0.0015283862935924653, 'San_Diego_Area')]
```