How long will a house stay on the market?

Purpose: Based on historical data for single-family residential property sales in SF Bay Area, predict house long a house would stay on the market once listed

Sergey Ermolin

Acknowledgements: dataset - MLS Listings, Inc.
Advisors: Hao Sheng, Stanford; - Pavel Berkhin, Microsoft

Dataset – courtesy of MLS Listings, Inc.

3 months of Single Family Residential sales data in five local counties (Santa Clara, Alameda, Marin, San Mateo, Santa Cruz). 4998 records total. Proprietary data provided by MLS Listings, inc, validated and regularized. Data can not be shared or distributed.

DOM	age	zip	Pool	Fire PLace	city	ListPrice	SalesPrice	Bdrm	Bath	SqFt	ElemSchool DIstrict	HIghSchool DIstrict	Elementary School	HighSchool	Latitude	Longitude
32	76	95002	0	0	ALVISO	450000	470000	2	1	584	Santa Clara Unified	Santa Clara Unified	George Mayne	Adrian Wilcox High	37.42815	-121.973
43	59	95008	1	1	CAMPBELL	1329000	1329000	4	3	2925	Cambrian Elementary	Campbell Union High	Bagby Elementary	Branham High	37.27322	-121.939
36	33	95008	0	1	CAMPBELL	699000	703000	2	2	1346	Cambrian Elementary	Campbell Union High	Bagby Elementary	Branham High	37.28507	-121.936
9	38	95008	1	0	CAMPBELL	410000	425000	1	1	704	Cambrian Elementary	Campbell Union High	Bagby Elementary	Branham High	37.28472	-121.936
44	58	95008	0	1	CAMPBELL	1285000	1230000	4	2	1997	Cambrian Elementary	Campbell Union High	Bagby Elementary	Branham High	37.28293	-121.93

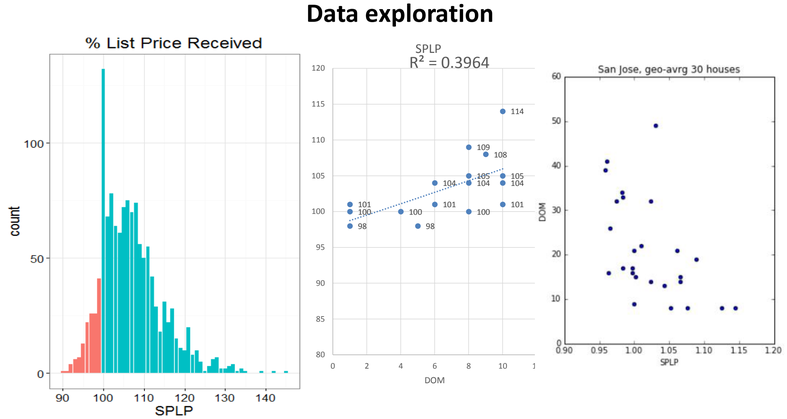
Motivation and a typical use-case

- For a new house about to be put up for sale, use real-estate agent's best judgement (or Zillow's Zestimate) as a proxy for the final sale price of the house. The listing price may be different from this value, depending on sales strategy and owner's cost of keeping the house on the market. Enter house data (location, size, list price, predicted sales price) into the model. Run the model to predict and predict Days-on-Market.
- If the house does not sell after a certain time period, re-evaluate the situation, adjust list/sales price and come up with new predicted DOM.
- Real estate market is season-dependent, so predicting DOM to within a week is considered "good enough".

Data augmentation and pre-processing

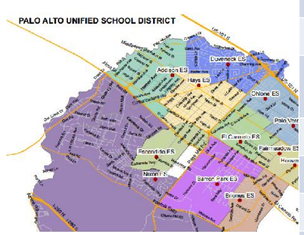
- Augmented data with School Ratings for elementary and high schools
- Converted categorical data to quantitative (eg. "Santa Clara Unified" -> 1, Cambrian Elementary -> 2"
- Removed outliers (Days on Market > 60, SalesPrice/ListPrice < 0.9,)
- Added meaningful statistics: SalesPrice/ListPrice (SPLP), SalesPrice/Sq.Ft.

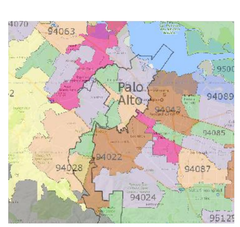
Data exploration



Data exploration – cont

- Real-estate data are very geo-dependent. Thus, a model trained on San Jose data will not necessarily be applicable to predicting Days-on-Market for Palo Alto.
- Initial hope of incorporating Lattitude/Longtitude by grouping N-nearest geo sales into one training dataset proved to be futile. City and school boundaries are jagged. N <15 was not a large enough a sample. N > 15 spilled over into neighboring schools and even cities polluting the data.





Feature	R_sqrd
'DOM ~ age'	0.003
DOM ~ C(zip)	0.028
DOM ~ C(city) * not relevant – only Palo Alto was selected	0.000
DOM ~ LP	0.014
DOM ~ SP	0.005
DOM ~ SPLP	**0.142**
DOM ~ SPSqFt	0.013
DOM ~ C(ElemSchoolDIstrict)	0.022
DOM ~ C(HIghSchoolDIstrict)	0.008
DOM ~ C(ElementarySchool)	**0.183**
DOM ~ C(HighSchool)	**0.053**
DOM ~ SPLP + C(HighSchool) + C(ElementarySchool)	0.320

Selected Learning Models and Analysis

- We tried three different model for regression: Multivariate Linear Regression, Random Forest and a simple DecisionTreeRegressor. All three yielded similar results with DecisionTreeRegressor given a slightly better estimate. Prediction error was measured as
  Error = sqrt(sum(DOM_actual – DOM_predicted)^2)/N)
- The fitting was done on geographically-specific data from the same city.
- Cross-validation: withholding zipcodes one-at-a time.

Model	Error	Notes
Multi-Variate Linear Regr	13.5	Variables: SPLP, HighSchool, ElemSchool
DecisionTree (depth 10, min_samples 6)	11.85	Zip, city, LP, SP, SPLP, SqFt, Lat, Long.
RandomForest n_estimators=100, max_leaf_nodes = 10, max_features = 10	14.2	

Conclusion and future work

Predicting DOM turned out to be much harder than predicting sales price (in prior work, we were able to do it with 80% accuracy). Still, we were able to fulfil our prediction goal.
In the remaining days, we will continue refining the models and attempting new ones in hopes of predicting DOM to within 1 week.