DSC 630 Term Project

March 03, 2021

Christopher Briggs

Executive Summary

Recent home price volatility presents investment opportunities. The fluctuation in home prices has not been evenly distributed geographically. We used predictive analytics to predict areas of high price growth.

We gathered information from City-Data on each of the 3,141 counties in the United States. This collected information includes median income and its recent growth, demographic information (including breakdowns by age, sex, and familial status), poverty rate, percent urban, average commute times, and geographic location. Not all of the data was usable: it turns out that some of the home price levels were calculated instead of measured. But after discarding the bad data, we still had enough to proceed.

We collected more than 50 pieces of information on each county. We then checked which of these were associated with home price growth - either positive or negative. We tried a suite of machine learning models. Different models work better for different data sets, so we compared these models to see which would best predict the house price growth.

The best performing model was an ensemble of neural networks. This model explains about two-thirds of the change in home prices in a given county. This is a pretty good result, as we cannot expect that every factor relevant to home price change will be captured in City-Data's data sets.

The model can now be deployed to predict home price growth given anticipated changes in county or city composition, or to identify counties in which prices may be primed to increase or drop. This predictive ability has business applications to investment, underwriting, etc.

Technical Analysis

I. Abstract

The aim of this study is to employ predictive analytics and publicly available county data in order to explain and predict the long-term growth rate of home prices. The data is analyzed and fitted with various models. Bagging methods turn out to be effective, with neural network and support vector regressors being the best base regressors. These models can be used to make predictions of relevance to several industries adjacent to the housing market.

II. Intro/background of the problem

The housing market is in constant flux. Increases in home prices can be brought on by strong economies and consequent wage growth driving price competition. Monetary policy can reduce the cost of borrowing, which also drives prices higher. The labor market for homebuilding may become expensive, decreasing elasticity of demand for existing homes, or supply of existing homes may be limited for other reasons (Pharande, 2018).

In the past two decades, home prices have been on a rollercoaster. The unprecedented peak of the mid aughts was followed by the Great Recession in which the underperformance of mortgage-backed securities emerged to pose a systemic threat to the international financial ecosystem (Boyle, 2020). Unsavory lending practices were largely to blame, specifically the issuance of what has been controversially termed "predatory" loans. Borrowers were permitted to take jumbo interest-only adjustable-rate

loans on houses, with the idea that home value increase would create windfalls by reselling before the rate adjusted higher (Chen, 2020). The effect was variably pronounced across the US, and particularly in sand states such as California, Nevada, and Arizona (Sumner & Erdmann, 2020). The price collapse was commensurately dramatic in these regions.

While it is widely acknowledged that there were regional differences in home price acceleration, less attention is paid to other potential underlying factors. In this study, a wealth of underlying factors are analyzed in relation to the average annual home price percentage change during the period 2000 to 2017. The aim is to identify characteristics of a region which predict or explain differential rates in home price growth.

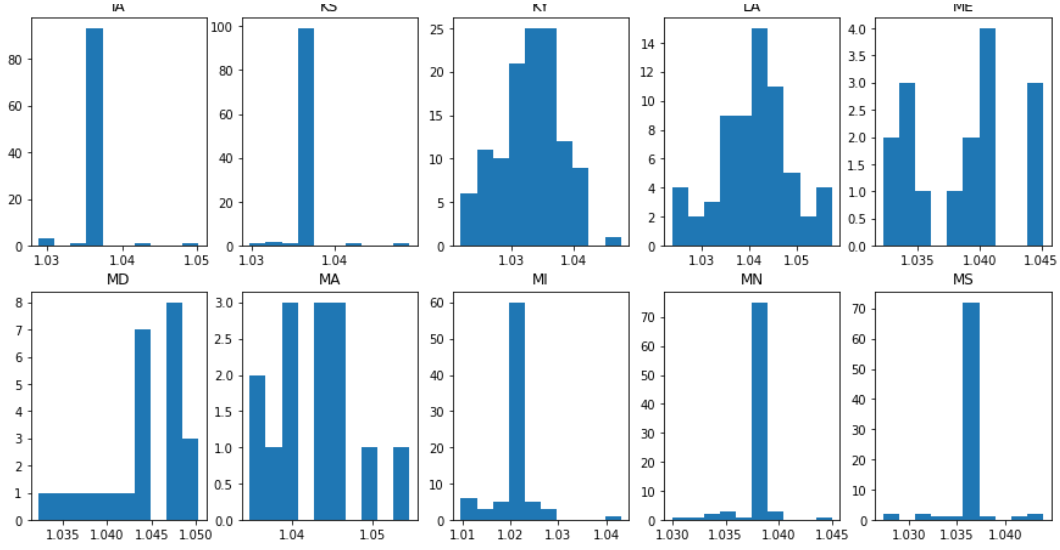
III. Methods

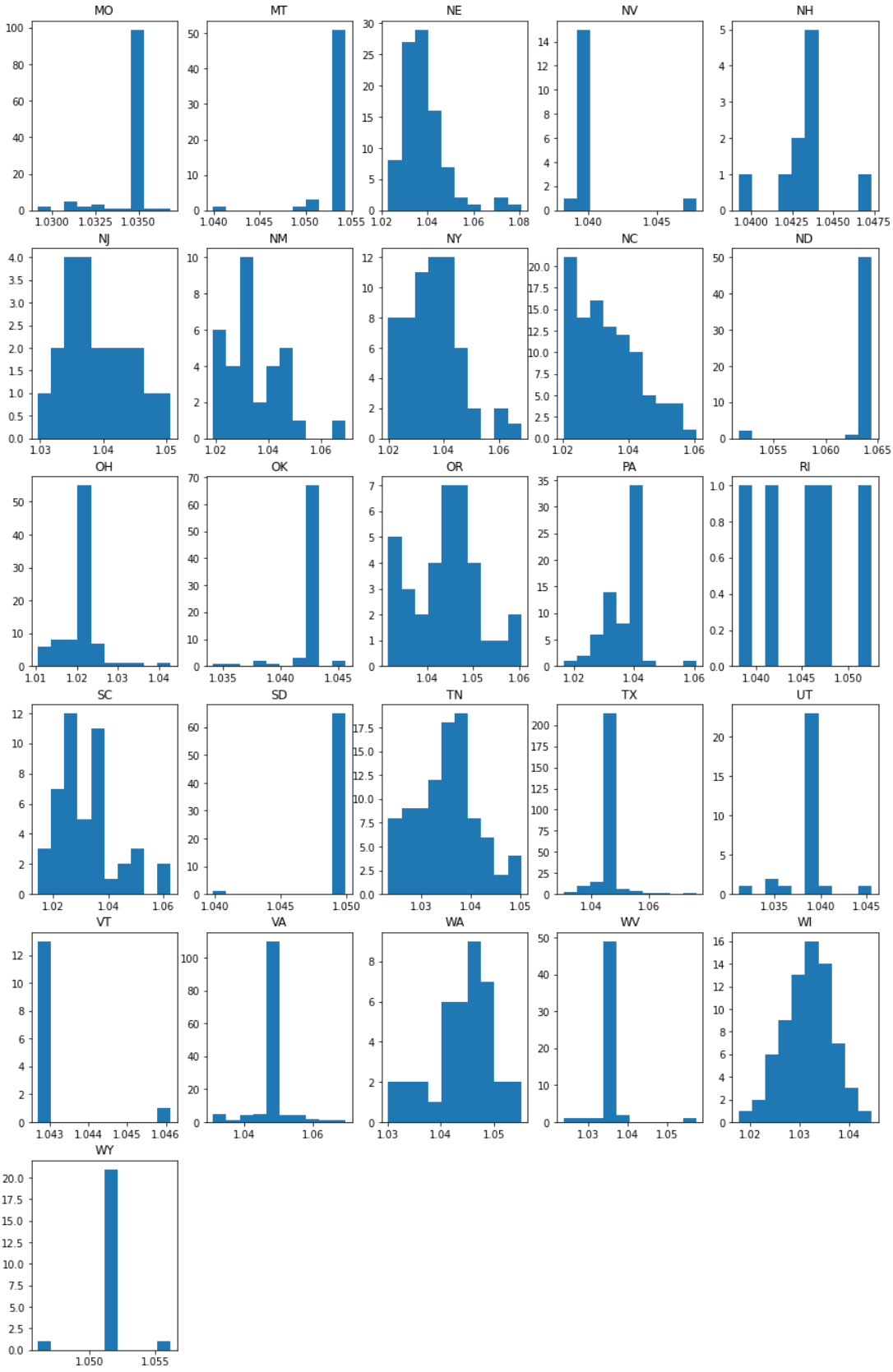
The project attempted to predict year-over-year average annual percentage change in median home values. The fundamental unit of consideration is the county. Data has been gathered on all counties in the United States. This data was programmatically collected from city-data during the Summer of 2020. Fifty features were collected. They include information on population growth, change in birth rates, median age, family composition, area measure, percentage urban and rural, farm statistics, and so on.

In each county, the median house value was measured in 2000 and in 2017. The target variable is defined by the change in price between these years divided by the price at 2000, then raised to the 1/17th power. The target variable represents the average annual percentage change in home prices in the county during the 17 year

period from 2000 to 2017. In addition to the target variable, I computed some features that I expected should be of relevance to the target variable. These features are population density, the difference in median age between males and females, the income growth rate, the percentage of the population consisting of married couples with children, the average commute time divided by the square root of the county area, the change in birth rates (measured in the 90s and the aughts), and the percent of the population which is foreign born. A data dictionary is included with the analysis.

During preliminary analysis I discovered that the data for median home values was computed rather than directly measured in many counties. What this means: a uniform growth rate was assumed for many counties in some states, and the 2017 average home value was calculated based on the 2000 value rather than measured. This phenomenon is clearly visible in the histograms of home price appreciation rate by county and state.




MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY

These data points must be discarded as they obscure rather than reveal the underlying processes affecting home price changes. I identified the states in which this occurred by plotting histograms of the percentage change by county; states which computed values would have the histogram dominated by the mode, in most cases comprising well over 50% of the values.

I programmatically identified states in which the target variable state mode was dominant in this way. There is a clear division between states at the level of 25% of counties matching the mode. I assume that in each state in which at least 25% of counties have the mode growth rate, the mode growth rate is invalid for each county having it, and these data points were discarded.

Next I pared down the features. There were more than 50 features; this overwhelms the number of records at around 1200. Available computing resources are also insufficient to properly train neural networks on so many features.

I used R to generate a correlation matrix, retaining the ten most correlated (directly or inversely) with the target variable.

```
1 cor(dplyr::select_if(df, is.numeric))
```

A matrix: 11 × 11 of type dbl

	median_house_income_2017	median_house_value_2017	cost_of_living_usd	adult_obes_rate	rent_3br_usd	pct_farms_fam_op	pop_growth_rate	income_growth_rate	foreign_born_pct	longitude	house_appreciation_pct
median_house_income_2017	1.00000000	0.7313496	0.69433680	-0.4985991	0.7589516	-0.2525027	0.53026901	0.7871637	0.4475764	-0.02205084	0.2444308
median_house_value_2017	0.73134956	1.0000000	0.74671174	-0.6669307	0.8449243	-0.4305947	0.48236431	0.5898110	0.6267177	-0.30178757	0.5051657
cost_of_living_usd	0.69433680	0.7467117	1.00000000	-0.4845616	0.8166910	-0.4242793	0.44831339	0.4985441	0.5787950	0.03558599	0.3208905
adult_obes_rate	-0.49859906	-0.6669307	-0.48456160	1.0000000	-0.5956000	0.3567675	-0.39740655	-0.4186651	-0.4864396	0.25834098	-0.3869259
rent_3br_usd	0.75895155	0.8449243	0.81669100	-0.5956000	1.0000000	-0.4408794	0.55480406	0.5487476	0.6432804	-0.17951678	0.3543771
pct_farms_fam_op	-0.25250268	-0.4305947	-0.42427932	0.3567675	-0.4408794	1.0000000	-0.21765057	-0.2110680	-0.5772794	0.06584030	-0.2425913
pop_growth_rate	0.53026901	0.4823643	0.44831339	-0.3974065	0.5548041	-0.2176506	1.00000000	0.3719077	0.5455913	-0.06017255	0.2771201
income_growth_rate	0.78716374	0.5898110	0.49854412	-0.4186651	0.5487476	-0.2110680	0.37190765	1.0000000	0.3232947	-0.13795792	0.4462096
foreign_born_pct	0.44757643	0.6267177	0.57879496	-0.4864396	0.6432804	-0.5772794	0.54559131	0.3232947	1.0000000	-0.18849332	0.3595692
longitude	-0.02205084	-0.3017876	0.03558599	0.2583410	-0.1795168	0.0658403	-0.06017255	-0.1379579	-0.1884933	1.00000000	-0.3326852
house_appreciation_pct	0.24443076	0.5051657	0.32089050	-0.3869259	0.3543771	-0.2425913	0.27712013	0.4462096	0.3595692	-0.33268521	1.0000000

We can more easily visualize the magnitudes of the correlations in a heat map.

	median_house_income_2017	median_house_value_2017	cost_of_living_usd	adult_obes_rate	rent_3br_usd	pct_farms_fam_op	pop_growth_rate	income_growth_rate	foreign_born_pct	longitude	house_appreciation_pct
median_house_income_2017	1.000000	0.731350	0.694337	-0.498599	0.758952	-0.252503	0.530269	0.787164	0.447576	-0.022051	0.244431
median_house_value_2017	0.731350	1.000000	0.746712	-0.666931	0.844924	-0.430595	0.482364	0.589811	0.626718	-0.301788	0.505166
cost_of_living_usd	0.694337	0.746712	1.000000	-0.484562	0.816691	-0.424279	0.448313	0.498544	0.578795	0.035586	0.320891
adult_obes_rate	-0.498599	-0.666931	-0.484562	1.000000	-0.595600	0.356768	-0.397407	-0.418665	-0.486440	0.258341	-0.386926
rent_3br_usd	0.758952	0.844924	0.816691	-0.595600	1.000000	-0.440879	0.554804	0.548748	0.643280	-0.179517	0.354377
pct_farms_fam_op	-0.252503	-0.430595	-0.424279	0.356768	-0.440879	1.000000	-0.217651	-0.211068	-0.577279	0.065840	-0.242591
pop_growth_rate	0.530269	0.482364	0.448313	-0.397407	0.554804	-0.217651	1.000000	0.371908	0.545591	-0.060173	0.277120
income_growth_rate	0.787164	0.589811	0.498544	-0.418665	0.548748	-0.211068	0.371908	1.000000	0.323295	-0.137958	0.446210
foreign_born_pct	0.447576	0.626718	0.578795	-0.486440	0.643280	-0.577279	0.545591	0.323295	1.000000	-0.188493	0.359569
longitude	-0.022051	-0.301788	0.035586	0.258341	-0.179517	0.065840	-0.060173	-0.137958	-0.188493	1.000000	-0.332685
house_appreciation_pct	0.244431	0.505166	0.320891	-0.386926	0.354377	-0.242591	0.277120	0.446210	0.359569	-0.332685	1.000000

I employed principal component analysis before training a linear regression model on the extracted feature set. I then performed feature selection for a random forest model and fitted the random forest model. I also trained a neural network, a support vector regressor, and an xgboost regressor.

After testing the individual models on the validation data set, I turned to bagging methods. I trained bagging models using four base regressors: decision trees, neural networks, support vector regressors, and xgboost regressors. For each of these kinds, I optimized the number of regressors by searching over a linear space for diminishing returns in prediction accuracy. I then tested each of the validation data set.

IV. Results

The bagging neural network model performed the best on the validation data, explaining close to 63% of the variation in the target variable (R^2=0.634). The bagging support vector regressor model was next, explaining about 60% of the variation in the target

variable (R^2=0.607). Unsurprisingly, the linear regression model had the worst performance, explaining 25% of the variation in the target variable.

Bagging ensemble models were effective for this data set. I used individual regressors and a bagging approach with three different kinds of base regressors: neural networks, support vector regressors, and xgboost regressors. I also trained a bagging model using decision trees, but I did not train an individual decision tree for base comparison.

In each of these three instances, the bagging model provided an accuracy improvement over the base regressor model. Bagging improved the neural network accuracy (measured as R^2) from 0.501 to 0.634, an improvement of more than 25%. For the xgboost model, bagging improved the accuracy from 0.511 to 0.566, an improvement of about 10%. The support vector regressor was not improved much by bagging: to 0.607 from 0.602.

The most accurate model was the neural network, and this is also the model that benefited most from the bagging approach. This may suggest, as an area of potential improvement, that the individual neural network model was not as well optimized as the other models. As noted in the analysis, a limitation of computational resources restricted the domain for the parameter optimization search (using GridSearchCV). Given more computational resources, this is the first area I would target for improvements to the model. I would expect both the individual neural network and the bagging approach to see accuracy gains, with perhaps a narrowing between the two.

V. Discussion/conclusion

The bagging neural network model had the best performance of the nine models attempted. The bagging neural network model explained close to 2/3 of the variation of the target variable. This is excellent performance, considering that there are factors contributing to the variation in the target variable not captured in the data set, not the least of which is chance. The worst performing model is linear regression. This should not be surprising, as the problem is a priori not a linear problem: we have no good reason to expect the target variable to vary linearly with any of the relevant input variables. Nevertheless, the linear model explains 25% of the variation in the target variable, and is therefore a useful model to use as a baseline for predictive analytics approaches.

The performance of the model has implications for businesses. While the study does not demonstrate causality (i.e. the home prices did not necessarily increase as a result of the measured features), the hypothetical real estate investor would do well to closely scrutinize counties predicted by the model to undergo larger increases than have been actualized. Such areas may currently have latent conditions which prime the local values to rise. This is similar to the logic of item recommendation based on affinity analysis ("Affinity analysis," 2020). Investors might likewise seek securities backed by mortgages in areas with higher scores in the model and avoid those with lower scores. Home loan underwriters might more closely scrutinize loan applications in areas which have experienced higher growth than the model predicts: inverse to the reasoning about conditions existing for unactualized growth, some counties may be found to have undergone home price growth unsupported by the conditions analyzed here. This may

justify higher interest rates, higher down payments for a given level of private mortgage insurance, and other measures to protect the financial integrity of the loans which are approved.

Demand for the homebuilding industry is nonlinear in that demand for homebuilding soars once existing home valuations exceed the cost of new construction. Homebuilding businesses might then seek to expand operations or relocate capabilities to regions predicted by the model to experience higher than average growth. They might employ closer analysis to compare regional construction costs to growth rate predicted by the model, and identify regions likely to soon experience a large increase in homebuilding demand.

VI. Acknowledgments

The author gratefully acknowledges the support and instruction from Dr Werner of Bellevue University and the helpful feedback from peers in DSC 640. The author also gratefully acknowledges City-Data for gathering and hosting the county data which informed the present study.

VII. References

1. Pharande, A. (2018, Nov 26). Top 5 factors that make property prices appreciate. Retrieved from https://housing.com/news/top-5-factors-make-property-prices-appreciate/

2. Boyle, M. (2020, Oct 23). The Great Recession. Retrieved from https://www.investopedia.com/terms/g/great-recession.asp#:~:text=The%20Great%20Recession%20refers%20to,Great%20Depression%20of%20the%201930s.
3. Chen, J. (2020, Oct 20). Interest-Only ARM. Retrieved from https://www.investopedia.com/terms/i/interestonlyarm.asp
4. Sumner, S. and Erdmann, K. (2020). Housing Policy, Monetary Policy, and the Great Recession. Retrieved from https://www.mercatus.org/system/files/sumner-housing-recession-mercatus-research-v2.pdf
5. Affinity analysis (n.d.). Retrieved Jan 10, 2021 from https://en.wikipedia.org/wiki/Affinity_analysis