Literature Survey

Title: Real estate value prediction using multivariate regression models.
Review: This article discusses various methodologies employed to predict real estate prices with an emphasis on getting accurate price predictions. We had related goals in our primary research question, so there were certain approaches in the article that inspired our own take on building viable models. To compare different models and select an optimal model, one needs a proper evaluation metric that accurately captures the problem at hand. In the article, the authors used the root mean squared error to evaluate model performance. While we considered additional evaluation metrics, such as R-squared, mean absolute error, and mean squared error, we ultimately came to the same conclusion that root mean squared error would best reflect our primary research question. Furthermore, the authors point to feature engineering as an important aspect to improving model fit. While the authors increased the powers of their predictors (e.g. squaring square feet) to improve model performance, we created new features, such as year and season of sale, to strengthen model performance instead.

Title: Investing in Elm Street: What Happens When Firms Buy Up Houses?
Review: Ever since the Great Recession of 2008, institutional investors have been on the rise. With the ever increasing number of investors, there is increased pressure in determining where to invest in real estate. Between 2000 and 2012, Miami, Tampa, and San Diego - all coastal cities - experienced the greatest increase in institutional buyers. Because this article is older, there is no newer information, such as what the trends are today. Instead of relying on investing with old information, making investing decisions with modern analytical techniques could result in an improved strategy.

Title: America's Rental Housing 2022
Review: This annual publication discusses the current state of the rental housing market in America. While our project is more geared towards the house ownership side, the cited text provided insightful context and even some pointers about what makes the market move the way it does. Geographical attributes and housing inventory are two features that are heavily discussed in the publication, and this actually informed our decision in using these variables along with others. Moreover, the geographical analysis that is showcased in the publication is quite similar to the final results we feature in this project, since the clusters we obtained after running the optimal approach correspond to the geographical locations mentioned in the publication.

Primary Research Question

The primary research question we tackled in this project is the following:

How can we build a regression model to predict the best housing investment opportunities in early 2023 for regions in the U.S.?

We believe this question accurately encapsulates the outcome we are expecting of this project. The question is rooted in our interest in using predictor variables (season, month and year, size of location, and inventory availability) and clustering techniques (geographical clustering) to predict if investing in a particular location would be a good investment. We aim to drive the investigation forward by using

regression and analytics techniques in order to develop a prediction model, and then validate it by using training and test data.

Supporting Research Questions

Our supporting research questions are the following:

- Does geography have a significant impact in the difference between sale and list prices?
- Does the available inventory of houses on sale correlate with the difference in list and sale prices?
- Is the geometry of our datacloud highly separable or linear, or is it better suited for non-parametric modeling?

Our first two supporting questions are directed towards the predicting significance of some features in the final model, while the last one is more directed towards knowing how the data behaves in parametric modeling. We believe these questions support our primary research question in a meaningful way, since they delve deeper into the intricacies of model building.

Data Source, Collection and Variables

Multiple datasets we used for this analysis. The Zillow Research Housing Data Website contains the ZVHI (Zillow Home Value Index), percent homes above list price, inventory, sales to list ratio, median sales price, and median list price information. Because clustering analysis was performed, the city/state information contained within the Zillow datasets needed to be geocoded to latitude/longitude information using the US Cities Kaggle dataset.

Because all of these factors are contained within separate datasets, each dataset needed to be transformed from 'wide' to 'long' format for ease of joining. After joining on the city, state, and date information, all factors were then contained within one dataset. Next, latitude and longitude information was joined from the 'US Cities' dataset. Lastly, the 'month' column needed to be converted from string to date format. After these steps were performed, the following schema was defined as:

- **Region ID** [int]- ID corresponding to a particular region (City/State, with the exception of the United States as a whole)
- **SizeRank** [int] - Rank of region based on population
- **RegionName** [chr] - City, State of region (with the exception to one record containing the United States as a whole)
- **RegionType** [chr] - Type of region (e.g. 'country', 'MSA' [Metropolitan Statistical Area], etc)
- **StateName** [chr] - Name of State
- **Latitude** [num] - Latitude of Region
- **Longitude** [num] - Longitude of Region
- **Month** [int] - Month / year or recorded values
- **Inventory_for_sale** [int] - number of unique listings that were active at any time in a given month
- **Median_list_price** [int] - median list price of homes
- **Median_sale_price** [int] - median sales price of homes
- **Median_StL_ratio** [num] - ratio of sales price to list price. Note this is the median of the ratio for each home, so the number is different than taking the ratio of the two columns above.

- **Prct_homes_above_list** [num] - percent of homes sold above list.
- **ZHVI** [int] - Zillow Home Value Index - measures the 'typical' home value as calculated by Zillow using a variety of factors.

Data Cleaning and Preparation

We took several steps to clean and prepare the data for analysis. We looked at the variables and removed "X." and "X" because they were equivalent to the index and weren't viable as predictors. We then examined the variables and their data types and converted "RegionID" from an integer to a character since "RegionID" represents categorical data, which are numbers without meaning.

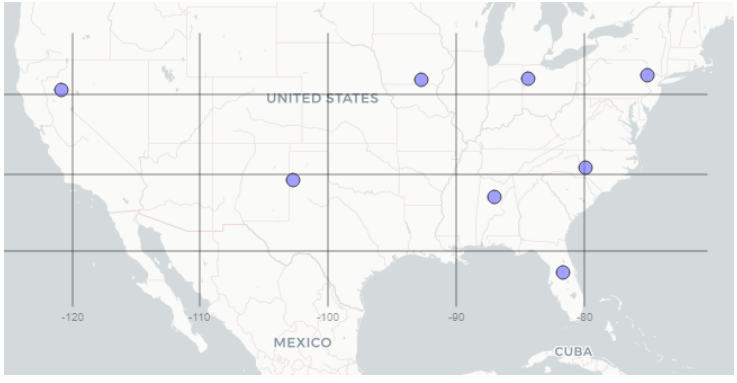
We moved onto dealing with the missing values. We noticed that the original percentage of missing values for all of the variables that had missing values was high, so we needed to devise a methodology to handle all the missing values. "StL_ratio" and "prct_above_list", which are two variables used to get the response variable, both had over 90% of missing values. After visualizing where the highest number of missing values were occurring in the timeframe, we decided to narrow the timeframe, choosing to use data from 2018 to 2022 instead of 2000 to 2022.

While we considered imputations and other approaches in dealing with the missing values, we determined that removing all the missing values after narrowing the timeframe from 2018 to 2022 was the best approach since this approach dramatically decreased the percentage of missing values for the two variables while retaining a sizable amount of data for analysis.

Initial Analysis and EDA

For our initial analysis, we first defined our dependent variable as the percent of listings sold above list price multiplied by the sale-to-list price ratio. We chose this as our final dependent variable because we want to encompass full returns in terms of both the quantity of opportunities available versus the average value of these opportunities.

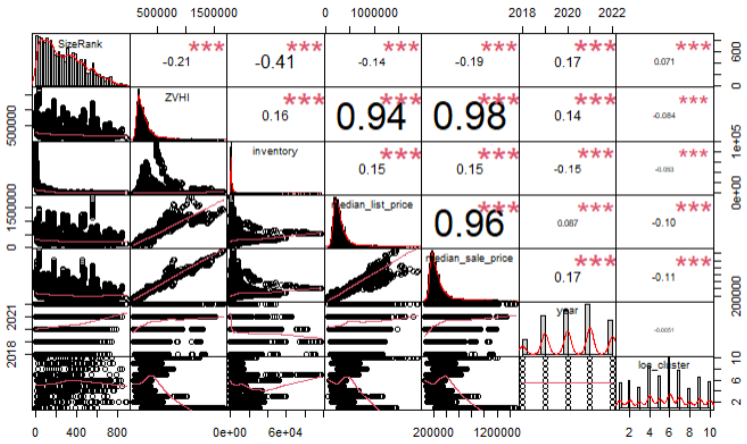
Then, we created new factors, such as year and season of sale, to assist in modeling. Since our final goal with the analysis is to identify areas of potential investment, which is inherently a geography-dependent question, we want to include as many potentially significant variables as possible to assure that geography-based coefficients in our final regression are not inflated or deflated due to omitted variable bias.

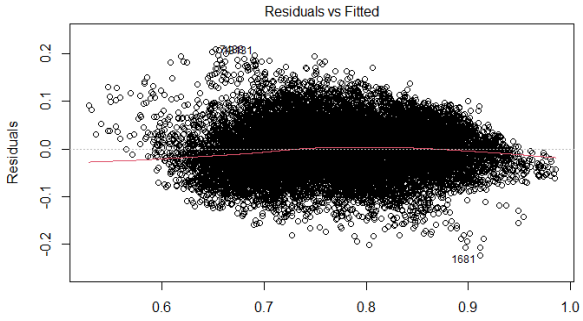


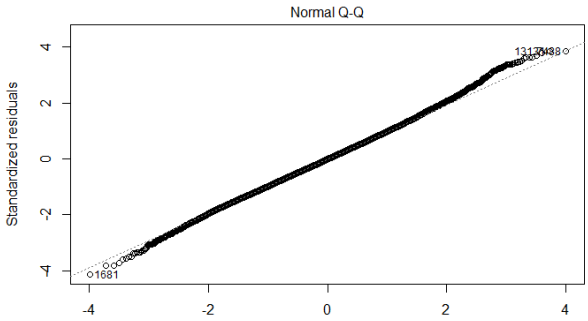
To address geographic differences, we used unsupervised learning – specifically K-means – to find clusters of listings based solely on their coordinates. In this way, cluster centers serve to divide listings into natural geographic groupings, which we hypothesize would help with modeling by representing other unseen factors affecting our dependent variable that are correlated with location.

After defining the final factors that can be extracted from our initial dataset, we analyze the factors themselves to understand their distributions, potential covariance, and their influential points, so that we can handle any such complexities when building our final model. After checking a correlation plot, we discover that most variables have very low correlation, with only three variables showing high correlation – the Zillow Value Index, mean sales price, and mean listing price, which are expected to be extremely correlated. Thus, we only include the Zillow Value Index in our final model, as well as inventory and size rank, among the factors extracted previously, such as year and location cluster.

To gain a basic understanding of the relationship among our regressors and the dependent variable, we created a base linear regression model, in which all regressors are found to be significant, with all but one being significant at p-values less than .001. Due to our base model's performance, we took the log of size rank, inventory, and ZVHI and then transformed our dependent variable by 1/y. These transformations helped solve some initial problems with heteroskedasticity, autocorrelated error values, and non-normality. After performing these transformations, as can be seen in the model's charts, the Q-Q chart shows near-normality, and the residuals vs fitted chart shows no patterns, which is ideal.







We also generated a Scale-Location, or Spread-Location plot, to see the spreading that occurs alongside different fitted values and the residuals. Our plot also showed no patterns, which is also ideal for regression. Next, some tests were run to check again for autocorrelated residuals, such as the Durbin Watson test, which showed no signs of autocorrelated residuals (due to high p-value). Finally, we

performed Cook's-Distance tests, which showed about 1,000 points to be influential points. When these points are taken out of the model, our adjusted R-squared value jumps from .6638 to .7306. However, we decided against removing these points from final modeling, as they are a large number of points, and thus must contain important information that perhaps can better serve other forms of non-linear or non-parametric modeling.

Modeling and Selecting the Optimal Model

Evaluation Metric

In order to compare different models to decide on an optimal model, we needed an evaluation metric that can assess how well a model is performing. To this end, we considered various evaluation metrics, like R-squared, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), to measure our regression models. Ultimately, we decided to use RMSE because it best measures the problem outlined in our primary research question.

RMSE measures how far off, on average, a prediction typically is from the true value, so an RMSE of 0 is ideal. The advantages of using RMSE include: it is easy to interpret, it reduces penalty for large errors, and it utilizes the same units as the response variable. The disadvantages are that the scale of the data matters and it is sensitive to outliers.

Multiple Linear Regression

One of the more common and simple models we could use for this analysis is multiple linear regression. We regress multiple factors against the added feature 'y' - which is a way of 'scoring' each datapoint:

$$y = prct\ above\ list * StL\ Ratio$$

When building the regression models, not all factors were included. There were some factors that were highly correlated with each other. For example, the cluster # was generated from the latitude and longitude data, so including all three factors would be redundant. Therefore, the factors used for the linear regression models are **SizeRank**, **ZHVI**, **Inventory**, **Year**, **Season**, and **loc_cluster**.

Additionally, the continuous variables were scaled and centered using the scale() function. To compute the RMSE value of the regression models, the data was split into a 70/30 test/train split.

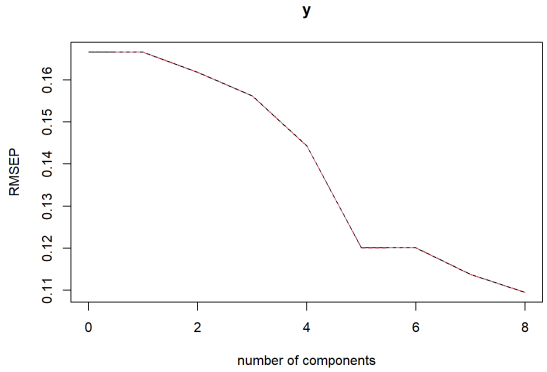
Simple Multiple Linear Regression

The simple multiple linear regression model was built with the factors mentioned above and the RMSE value was calculated to be **0.1109242**. A log transformation was also applied, resulting in a log-linear model with an RMSE value of **0.1232917**.

Multiple Linear Regression with PCA

Principal Component Analysis (PCA) is a common technique for reducing dimensionality of a dataset. As a result, the number of correlated features is reduced, and overfitting is less common. Additionally, PCA

models typically perform faster. However, the added complexity makes these models more difficult to interpret and draw conclusions around relationships between variables.



A validation plot was created, and 5 principal components were found to be the optimal number for our model. After the components were generated, regression was run across the components and a RMSE value of **0.12222** was calculated.

LASSO Regression

Another common way of performing feature selection is to use LASSO. This method allows for quick and easy feature selection, while reducing multicollinearity and overfitting. There is a hyperparameter λ, also referred to as the 'penalty' term, that must be optimized for the LASSO model. The optimal value was found to be 0.0003 - very small. Therefore, the effect of LASSO was minimal, resulting in an RMSE value of **0.1155**

Stepwise Regression

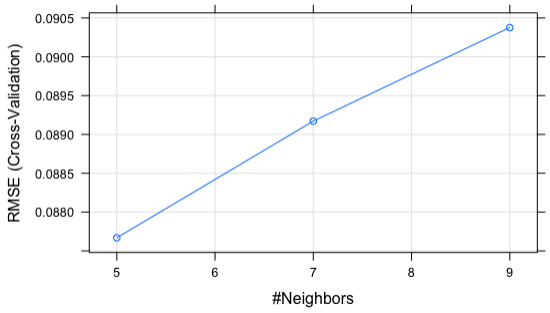
One very easy way to perform feature selection is to run stepwise regression. Backwards and forwards stepwise regression was run across our factors; however, this resulted in the removal of 0 factors in both cases. Thus, stepwise regression is unnecessary and results in an RMSE value of **0.1109 -** equal to the simple multiple linear regression model.

K-Nearest Neighbor Regression

Another model we built was one using the k-nearest neighbor (KNN) regression approach. The advantages to using KNN include it having great flexibility and a lazy learner, so it outright does not build a relationship between the predictors and outputs. However, this also comes with disadvantages, the main one being that it does not play well with datasets with high dimensionality and thus suffers from performance issues.

First, we have to prepare our data, meaning that we have to create dummy variables for our categorical values. KNN relies on distance computation, so this is a necessary measure. Also necessary to get the best results is scaling and centering. Next, we have to build the model. This is a very straightforward thing to do, since KNN is very flexible and doesn't require much setup, except from the preparations described above. To ensure we are getting the best model, we ran a cross-validation method against our model. In

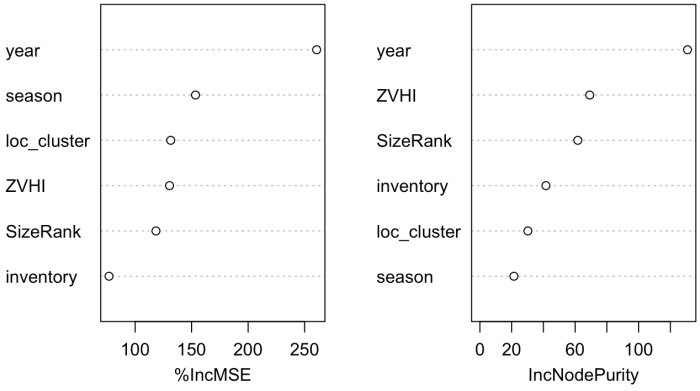
fact, this is a 10-fold cross-validation method. After this, we get a RMSE of **0.0875** and a R-squared value of **0.7244**. This is the best result we can get, using the k value - the hyperparameter - of **5**.



Random Forest Regression

We also built a random forest regression model to predict our response variable. Generally, the advantage of using a random forest model is that it tends to perform well, while its disadvantage is that it can be difficult to understand as it is considered a black box model. We maintained many of the same procedures as mentioned above, such as using SizeRank, ZVHI, inventory, year, season, and loc_cluster as the predictors and splitting our training and testing data in a 70/30 ratio. We did not scale the data since scaling is not necessary for random forest models. To optimize the model, we performed a grid search for the optimal "mtry" value, which was determined to be 6. The parameter "mtry" represents the number of variables sampled at each split and is often a good parameter to tune for random forest models.

The RMSE of the optimal random forest model was **0.056**, which was lower than that of many of the other models. The metrics "%IncMSE" and "IncNodePurity", which are two metrics involved in assessing feature importance, showed that year, season, and ZVHI were the most important features in the random forest model.



Support Vector Machine Regression

Due to some of the complications found in linear regression, such as having a high number of outliers, we wanted to test an SVM regression model (furthermore referred to as an SVR model), as SVR models generally perform better with outliers than a linear regression model, due to an SVR's use of boundaries,

making them more generalizable. Furthermore, with an SVR model, we can test different kernels to see if the geometry of our datacloud perhaps has a unique shape that can be exploited by SVR. The main drawback of using this type of model over linear regression is its poor performance with categorical data, due to its dependency in measuring distances between data points and the decision boundaries. However, as our data is composed of half continuous and half categoricals variables, we felt it still worth testing.

After scaling the continuous variables we used training/validation subsets of the data to train on the optimal kernel of the SVR model, and found a radial kernel to be optimal, highlighting potential shapes in the geometry of our datacloud. The final model performed better than our regression models, showing that our assumptions might have been true, in terms of an SVR model being more generalizable and less sensitive to outliers, of which there are many in our dataset. However, the model did not perform the best, perhaps limited by the need of the model to strictly depend on the geometry of the datacloud, unlike model types like random forest and KNN.

Optimal Model

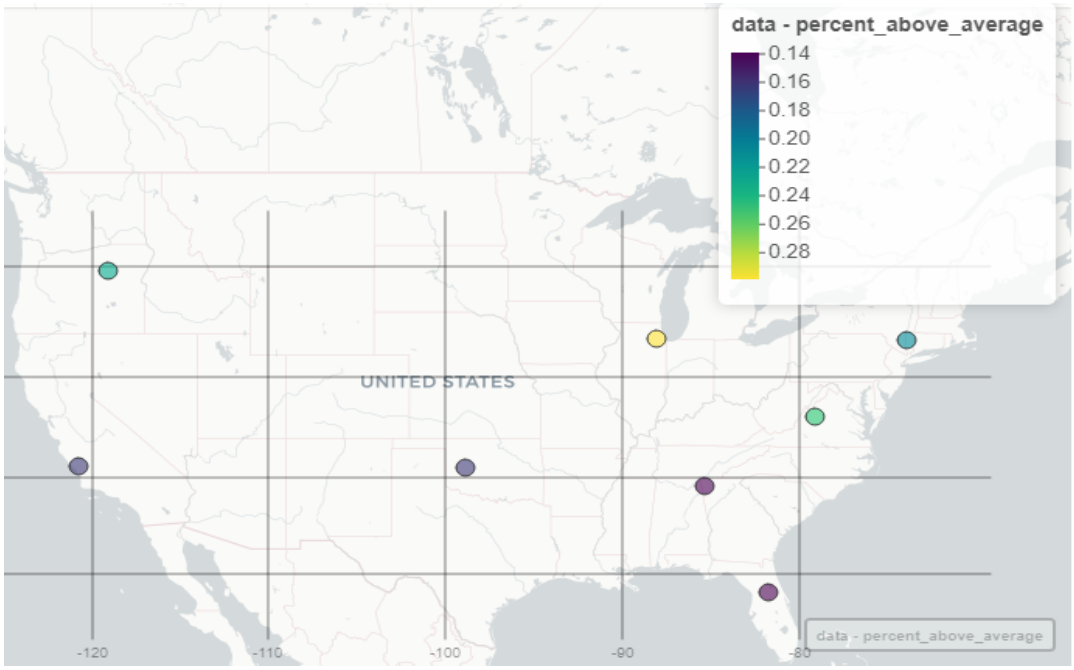
Model	RMSE
Simple multiple linear regression	0.1109
Multiple linear regression w/ PCA	0.1222
LASSO	0.1155
Stepwise	0.1109
KNN	0.0875
Random forest	**0.0562**
SVM	0.0857

The table shown above summarizes all the models we built and their respective RMSE scores. Looking at this table, it's clear that the random forest regression model has the lowest RMSE value of 0.0562. An RMSE that gets as close to 0 as possible is ideal. Thus, our optimal model is the random forest regression model.

Results

With our final model decided, we then sought to use the model to determine the location cluster with the highest overall expected returns, to decide where to move our investment firm in 2023. To do so, we first had to create a data frame of imitation data representing potential types of investment opportunities in 2023, balanced across every variable that is being used in prediction as a regressor. The theory behind this is that, in 2023, we cannot predict what types of properties, or how many properties will be available in any given area, so we must test all possible combinations of properties, value indices, and inventory, among the other factors included in our model; per location. Then, once predicting on all of these hypothetical data points, we can group by location to see which location has the highest expected returns, given all else being equal.

To determine the overall expected returns per location, we used the mean of our outcome variable of our entire real dataset as a baseline, and took the percent per region of predicted values from the hypothetical dataset that were 1 or more standard deviations above average. This percent we considered the expected percent of homes to have higher-than-average returns, and then plotted the final value across each region, to see which region is predicted to have higher overall expected returns.



As can be seen, the Midwest is expected to have the highest overall returns, with an expected 28% of 2023 listings to have an above-average return. Southern regions had lower predictions overall than northern regions, with the Southeast having the lowest of all expected returns, at just 14% of 2023 listings having above-average returns.

Conclusion

Over the course of the project, we identified some important steps that needed to be fulfilled in order to get the best results, and they all boil down to being prepared to work with the correct tools and approaches. From selecting the correct accuracy metric to identify how to transform the dataset we are going to work on, it is important to have everyone at the team aligned to understand how those decisions affect the way we interact with the dataset and our desired outcome.

It is also important to point out that the approach we deemed as the optimal one, the Random Forest Regression, holds an enormous advantage against the other approaches when comparing their RMSE values. This is due to the way the data is laid out, since it has certain complexity due to the various categorical and continuous variables, as well as the clusters introduced beforehand.
Finally, some next steps that we want to pursue in a next iteration of the project would include using other external factors related to the housing market, such as inflation, to further enrich our analysis. Also interesting would be to incorporate time series analysis into the project, since seasonality and forecasting parameters can be of great help.

Sources

Lambie-Hanson, L., Li, W., & Slonkosky, M. (2018). Investing in Elm Street: What Happens When Firms Buy Up Houses?. *Economic Insights, Federal Reserve Bank of Philadelphia*, 9-14.

Manjula, R., Jain, S., Srivastava, S., & Kher, P. R. (2017). Real estate value prediction using multivariate regression models. *IOP Conference Series: Material Science and Engineering*. doi: 10.1088/1757-899X/263/4/042098

Airgood-Obrycki, W. et al. (2022). America's Rental Housing 2022. *Joint Center for Housing Studies of Harvard University*.