Shape

Spring 2024

Final Report

Shape

**Team Member Names**: David Grice, Aaron Kwan, Kevin Rice

**Project Title**: Bid Smart: Navigating Real Estate Valuations

**Problem Statement:** Helping North Carolina home buyers successfully bid the right amount on a house listed for sale in a given market.

A graph showing the growth of the stock market

Description automatically generated

Figure 1: U.S. News Housing Market Index, 2017 - 2024

Real estate market conditions are constantly changing, and the landscape of supply and demand is turbulent. Figure 1 above visualizes a measure of the health of U.S. housing markets based on three primary factors: Demand, supply, and financial health (U.S. News Housing Market Index, 2024). While the market around the country in general is increasingly difficult for home buyers to navigate, using the indicators provided by this index, North Carolina has a particularly tough supply to demand ratio of 0.54 compared to the U.S. overall (0.89). This creates a seller-friendly market with bidding wars and quick sales, where a buyer’s negotiating power and ability to predict what a home may sell for becomes almost obsolete.

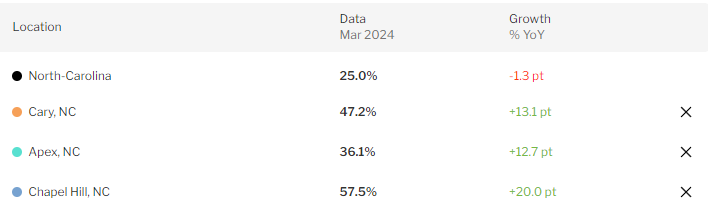
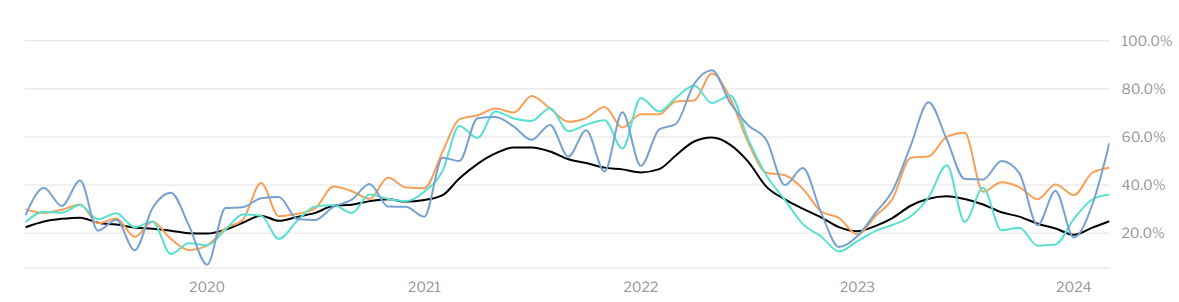


Figure 2: Redfin.com North Carolina Housing Market Metrics

Figure 2 above shows a Redfin.com (North Carolina Housing Market, 2024) analysis of the past five years of data for the percentage of homes that sold above list price for the entire state of North Carolina in comparison to the top three most competitive cities in North Carolina. Over the past five years, as much as 87.8% of homes in a given market sold for over listed price with most recent figures still hovering around 44.6%. This is a strong indication of a market that has been “hot” for a long time which further obscures the predictability of what it would take to have the most competitive offer on a sale.

Even with all this unpredictability, a home’s *value* in any given market is still relatively easy to measure in confidence. The primary resource when it comes to home valuation is an appraisal by a licensed professional which takes into consideration several key factors such as square footage, structural improvements/additions/renovations, number of bedrooms, architectural style, HVAC system age and operational efficiency, foundation, and appliances. Generally, lenders will *not* loan any funds to a buyer more than the appraised value.

The problem arises when a home’s value does not necessarily correlate with how much it will sell for, which, as we have demonstrated thus far, is common in a “hot” market where supply does not meet demand. The alternating peaks and valleys of the Redfin.com analysis should elicit sympathy for home buyer’s and their agents all across North Carolina. Is the market cooling down and can buyers go back to making “normal” offers or do they need to offer over listing price if they want a competitive chance? Additionally, *how much* (if at all) does a buyer need to offer over listing price to have a competitive bid? The answers to these questions are the ultimate goal for this project, as we aim to build a model that can help bring back some predictability to how much a home will actually sell for in a given North Carolina locality – helping buyers enter a negotiation with a realistic expectation of affordability and how to be competitive.

**Statement of Work:** Rather than just simply estimating the *value* of a home listed for sale, we aim to feed a **Regression** model with a variety of different metrics and home features to predict what the optimal amount a home buyer would need to bid to get a fair, competitive offer in any given market in North Carolina, at any given time. We will explore variations of regression such as Random Forest, Ridge, and LASSO.

**Methodology:** The methodology to estimate a home’s *value* has largely been discussed in the problem statement. Having established the growing lack of correlation between a home’s value and what it will actually sell for, the methodology for our model needs to expand to new and/or alternative features and predictors to accurately estimate a home’s sell price (what a buyer should anticipate is needed to win the bidding war). We find it best and most useful to break the data down to the zip code level of locality to help reduce some level of noise and variance within the data. Realtor.com (*Realtor.com Real Estate Data and Market Trends for Download*, 2024) offers 56 unique metrics by zip code that offer insight into supply and demand, home features, and popularity. Zillow (*Housing Data - Zillow Research*, 2024) provides a Zillow Home Value Index (ZHVI) metric which is a smoothed and seasonally adjusted measure of typical home value in the 35th to 65th percentile range. This potentially provides a unique insight into the market, apart from medians and averages which are more frequently used in the Realtor.com metrics. We have also gathered monthly average interest rates in the U.S. (*30-Year Fixed Rate Mortgage Average in the United States*, 2024). Drawing inspiration for new perspectives from (*The hottest U.S. housing markets,* 2024) and (*Surveys of Consumers*, 2024), we have also added in our own calculated measure which is the ratio of active listings to total listings to help give a unique perspective on supply and demand.

On initial onset for a prototype model, all data from these sources was combined by zip code and month to form a workable dataset to train our model on. Due to the year-by-year variability in housing markets, especially considering the effects of the COVID-19 pandemic, we wanted to limit our training data to as recent as possible to hopefully increase predictability.

Initially we planned to train our model on monthly North Carolina transactional data from the entire year of 2022, and then validate the model on monthly North Carolina data from the entire year of 2023, to test the model to make predictions for the current year of 2024. Real estate transaction data is stored in what is called the Multiple Listing Service (MLS) but each locality and region has its own “MLS” with no real, structured way of how these regions are broken up and no centralized connection between all of the different MLS databases. Furthermore, each localized MLS has its own set of rules as to who can access this data and how. A real estate license is typically the bear minimum. Websites like Zillow.com, Redfin.com, Realtor.com, etc., have the corporate backing and resources to access every MLS all around the country from which they are able to list homes on their websites that are for sale and provide their own metrics and reports on markets. They have a firm grip on this data for proprietary analysis and profitability and only allow access to limited records.

Thus, we were limited to training our model on North Carolina transactional data from September of 2023 to February of 2024 and we split the data 70% for training and 30% for testing. Additionally, we scrubbed the dataset to contain only single-family home transactions between $250,000 and $2.5M which we determined to be a representative range of the market as a whole and eliminated the majority of outliers. We also removed records with NaN values rather than imputation, as the number of records with NaN values was relatively small and wouldn’t have a large impact on the model training. Our final dataset consisted of 22,602 records (North Carolina transactions) with 47 of the features from the various sources mentioned previously.

**Evaluation:** Starting with a basic Random Forest model we were able to achieve a ~0.60 coefficient of determination, meaning that our initial model explained about 60% of the observed variation in price. Table 1 below shows ten different examples of what this might look like in application.



Table 1: Random Forest model, predicted price vs. actual price

While some results were better than others and in the world of analytics this is often seen as acceptable results, for this particular problem the variance is unsatisfactory. Home buyers cannot afford to miss the mark by tens of thousands of dollars when they’re already being forced, in many cases, to offer more for a home than it is even worth in order to be competitive. For this model to be considered a “success” we wanted to see the majority of results being within $10,000 of actual price. This high standard makes the nature of the problem quite difficult to satisfy. Nevertheless, we attempted a few different methods to improve results, namely: Ridge, LASSO, and PCA.

LASSO provided the best results, however the best coefficient of determination achieved was ~0.50, which is actually worse than our original Random Forest model. The dimensionality reduction with LASSO eliminated 41 of the 47 features. Table 2 below provides the results for the same ten records evaluated in Table 1. This also provides interesting contrasts and commonalities to the feature importance selection from the Random Forest model as shown in Figures 3 & 4, suggesting that deeper variable selection may be of particular interest for future study.



Table 2: LASSO model, predicted price vs. actual price

A graph with many words

Description automatically generated with medium confidence

Figure 3: Random Forest Feature Importance

A graph with black text

Description automatically generated

Figure 4: LASSO Variable Coefficients

Overall, both models were able to make some extremely accurate predictions but were also way off the mark at other times. Comparisons below in Table 3. Root mean squared error of the Random Forest model was 184,826 and of the LASSO model was 206,953.



Table 3: Random Forest predictions vs. LASSO predictions

Figures 5 & 6 below show the actual values from the dataset plotted against the predicted values obtained from the LASSO and Random Forest regression models, respectively. The dashed line represents perfect predictions, where actual values equal predicted values. Points close to the dashed line indicate accurate predictions, while points far from the line indicate errors in prediction. For both models, most predictions lie relatively close on both sides of the dashed line – meaning the models are not consistently over or under predicting.

A graph with a blue line and black dotted line

Description automatically generated

Figure 5: Random Forest Actual vs. Predicted values

A graph showing a line graph

Description automatically generated with medium confidence

Figure 6: LASSO Actual vs. Predicted values

Observing the residuals in Figure 7 we confirm that heteroscedasticity is not a major concern, though some transformation or interaction of variables could provide some improvement to the variance of residuals.

A red line with black lines

Description automatically generated with medium confidence

Figure 7: Residuals Plot

**Conclusions:** While our optimal model has the ability to make a prediction within a few dollars, variance is too great to have the model be considered effective for a home buyer. What we have observed, however, is that our methodology of selecting 47 different market and home-specific metrics from various sources may be worth considering in greater detail for future research. Current valuations from buyer’s agents typically only take into consideration neighborhood comparables of similar homes and their market knowledge and expertise. Our project at least shows that using machine learning to model on dozens of supply and demand features, market features, property features, etc., has promise. Limitations of time, resources, and availability of data hindered the model’s ability to make more accurate predictions. We would have been able to delve deeper into the larger errors and determine cause for refinement. Additionally, we recognize that the real estate industry can be seasonal in nature and since we were only able to gather data for a few months, that seasonality was not captured in the dataset. We would have also liked to get market-specific insight from local experts. However, we hope the insights gained from this project spur the industry into realizing how valuable a centralized, governed database of real estate data could be. Not only would we be able to train a model on a larger and more inclusive dataset, but we would also have access to dozens of new metrics and features we could refine the model with and could explore additional methods of machine learning such as deep learning. If we were able to achieve the results desired it would be a game changer for the real estate industry and could dramatically have a permanent effect on the entire home buying negotiation process.

**Data Source(s):**

*Realtor.com real estate data and market trends for download*. (2024, March 5). Realtor.com Economic Research. <https://www.realtor.com/research/data/>

*Housing data - Zillow Research*. (2024, February 1). Zillow. <https://www.zillow.com/research/data/>

*30-Year fixed rate mortgage average in the United States*. (2024, March 5). <https://fred.stlouisfed.org/series/MORTGAGE30US>

*The hottest U.S. housing markets* (D. Bradbury & L. Rajala, Eds.). (2024, February 12). U.S. News.  <https://realestate.usnews.com/real-estate/housing-market-index/articles/the-hottest-housing-markets-in-the-us>

*Surveys of Consumers - Demographic Subgroups*. (n.d.). <https://data.sca.isr.umich.edu/subset/subset.php>

U.S. News Housing Market Index. (2024, March 21). U.S. News. <https://realestate.usnews.com/housing-market-index/interface#performance-indicator>

North Carolina Housing Market. (2024, March 21). Redfin.com. <https://www.redfin.com/state/North-Carolina/housing-market>