MGT 6203

Instructor: Jonathan Fan

Predicting Time on the Market for New Home Listings

Group Project: Final Report

Team #21

GitHub Repo:
https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-21

Table of Contents

Overview of the Project

Team Members

1. Nildip Chaudhuri, GT Id: 903951206
2. Erik Euler, GT Id: 903848886
3. Nick Sienicki, GT Id: 903004037
4. Maryam Paknejad, GT Id: 903835757
5. Giovanni Marrero, GT Id: 903845934

Background and Problem Statement

Housing finance in the United States is a twelve trillion-dollar industry that touches every corner of the American economy, from a deep relationship with the federal funds rate and employment to historically low-risk government-backed securities for purchase by investors. Often, interacting with buying and selling a home is the largest financial transaction in an individual's life which makes getting the right home at the right time a critical decision.

There are several factors that go into which home to pick to make an offer on, but a key component remains the time that a home has been on the market. For a buyer, a home that has been on the market for a considerable time demonstrates the opportunity to offer a below-asking price due to the presumption of low demand. For a seller, the longer their home is on the market, it triggers more anxiety or re-evaluation of the asking-price, listing details, or even triggering delisting the home from the market.

With the housing market as competitive as it is, it is becoming increasingly difficult for buyers to know when to make their offer, and for sellers the same difficulty exists in timing changes in their expectations. By understanding how long a home is expected to be on the market within their metro area, buyers and sellers can act with more confidence in their decision making to remove listings from the market, identify opportunities to offer under-asking, or reevaluate the listing price.

Median time on the market provides a good barometer of the overall pace of transactions. This gives buyers and sellers a benchmark to compare their anticipated transaction against. While not a perfect measure of how long a home will take to sell, the median time on the market approach balances imperfect data availability with an easy-to-understand metric that can be used for comparison.

Approach and Initial Hypotheses

From a business perspective, this is a metric we expect to be useful in many ways. Given the size of the housing finance industry, a business insight that can drive even a component of the industry to a more efficient action is worth a great deal. Realtors often operate on qualitative knowledge of a market, and buyers and sellers typically have limited to no experience at all with buying and selling a home independently.

Creating metrics to get at a quantitative approach for evaluating time on the market for a specific listing is useful to many stakeholders. More knowledge about the time a specific home has been listed relative to the market could drive a buyer to identify a value investment or trigger a seller to lower asking-price to accelerate a sale. These small decisions often mean thousands of dollars of change in asking price. Being able to clearly define when these actions are probable or necessary to drive the desired impact using existing data, can lead to advantages in an often-asymmetric market. Macroeconomically, time on the market should serve as an effective and measurable proxy for housing demand. Our anticipation is that with more housing demand, median time on the market decreases. A cooler housing market means less demand and our prediction would be more time on the market.

In short, predicting time on the market for new home listings should continue to drive more insight into a critical portion of home buyers and sellers.

Literature Review

Given the importance of this metric to the housing market, there has been much analysis on the various components of time on the housing market over the years, taking various routes. The context and analytical first steps outlined below provided important background for analysis

In 2002, Knight studied the revision of price of a listing during the marketing period to investigate its relationship with the listing price, selling price, and selling time. He used a maximum-likelihood probit model and found evidence consistent with the theory of pricing behavior under demand uncertainty: He concluded that vacant homes and homes with high initial price, are more likely to change price during the marketing time, but homes that have any unusual features, are less likely to experience a price change. This change in price is very costly to the seller in both time and money. He stated that "homes with large percentage changes in list price take longer to sell and ultimately sell at lower prices" (Knight, 2002). This study identifies vacancies as one of the major reasons for a prolonged time on the market for a listing, and most importantly shows the financial impact of price changes. This underlines the importance of making price change decisions in a timely manner to "stop the bleeding."

In 2015, Benefield and Hardin explored the ambiguity in the definition of time on the market by using Multiple Listing Service (MLS) from a medium-sized United States city. They modeled and evaluated five measures of time-on-market. Then they investigated the importance of relisted properties in defining and assessing time-on-market. They concluded that "both time-on-market definition and the handling of relisted properties substantially influence model outcomes and the statistical significance of dependent variables" (Benefield and Hardin, 2015). They also suggested using the term time-on-market needs to be reevaluated and defined for relisted properties. While this is a valid point, the definition of Days-On-Market (DOM) is still commonly defined as the number of days between a property's initial listing and the date it is either sold or taken off the market. It should be noted that many do use the technique of relisting to reset the count for the property. The concept of relisting of a home presents another important option that can be taken with a greater understanding of time-on-market.

In 2019, Castelli et al. attempted to predict days on the market to optimize real estate sales. Their study mainly focused on building an accurate predictive model for the number of days a published listing will be online, with the objective of identifying fake listings. They tested four different modeling approaches for this task: Lasso regression, Ridge regression, Elastic Net regression, and Artificial Neural Networks. The results, obtained on a vast dataset made available by the Bulgarian company Homeheed, showed the effectiveness of Lasso regression. The limitation of this study was focusing on one city only, in Bulgaria. Given the broader geographic context of our analysis, it will be interesting to see if our final modeling outcome will take the same path.

In 2020, McGreal et al. worked to study "to estimate value effects in relation to the time on the market for residential properties" within the Belfast (U.K.) metropolitan area. They hypothesized that sale price is "a function of marketing time and that properties with a shorter marketing time (TOM) are more likely to sell above list price" (McGreal et al, 2020). The results suggested that the relationship between sales price and marketing period is mixed and "although sale price is influenced by TOM, the effects are uneven" with properties selling at or above list price likely to have shorter marketing periods compared to those selling below list price. This study embarks on a useful journey and its conclusion has been a common perception lately, specifically in the post-COVID housing market.

McGreal focuses on testing the machine learning algorithm's ability to predict the Time-On-Market by combining the real state data with U.S. census data, changes in interest rates and Consumer Confidence Index (CCI). Consequently, to offer a broader perspective, the aim of this study is to better understand different variables impacting the Time-On-Market in the past 6 years. The period selected intentionally covers the unusual changes

to the housing market from pre-pandemic, to pandemic, and post-pandemic era, each with quite different landscapes; Bringing their insights to a dataset covering a broader economic period could help generate insights that cut across economic trends.

Overview of Data

Data sources

To form a well-rounded picture of each metro area and predict median days on the market for a metro area, we used a couple different data sources. Core to our analysis will be several different components of Realtor.com's data on housing market activity that includes the response variable as well as several predictors. This dataset is aggregated primarily based on regional Multiple Listing Service (MLS). MLS data is the best source of data at the individual region level for home sales, but given each MLS covers an inconsistent area of a region, Realtor.com's aggregation provides a significant amount of value for our purposes. Initially, we had planned to use Zillow.com's data set for this task, but after difficulty in working through Zillow's internal mapping of metro standard areas and the low fill rate of many of the fields, we made the switch to Realtor.com data.

With the housing specific data will be a variety of components of overall macroeconomic data, with national- and state-level data both available. Census bureau data will bring different angles on demographics that may drive market differentiation, while other macroeconomic factors will help to explain supply and demand for homes. One of the first priorities of our analysis has been stitching these disparate datasets together.

The Realtor.com dataset is an aggregated view of MLS data from each different region, which suffers from neither of the issues that the Zillow data did. Fill rate is relatively consistent and high, while the data is ready at the county level, associated with county FIPS code. Additionally, this data is made available by ZIP code, giving us an option with even more location-based granularity (Realtor.com).

There are several variables that are the center of the focus in this dataset. The first is the response variable- "Median Days on Market." For the independent variables, given the shared source and relationship to the overall housing market, we anticipate a significant amount of covariance for these variables.

The second issue is one of completeness. Despite the superior fill rate with the Realtor.com data as compared to Zillow; we still face missing values in an appreciable number of observations. We tackled this with a simple mean-based imputation approach.

Macroeconomic Data

The next area of analytical focus was the array of macroeconomic data we had sourced. There are multiple dimensions to our focus here. The first is forming an accurate picture of the overall macroeconomic climate at any one data point. As our dataset progresses through time, we will see variations in both the Zillow data and the macroeconomic climate that will tune our model.

We expected that the most critical driver of housing economic activity will be changes in the federal funds rate (FEDFUNDS in our dataset). The period we are analyzing, 2017-2023, has had great shifts in interest rates which have impacted the volume of mortgages tremendously (Rothstein). We have gone from federal funds rates effectively at 0% in February 2022 to 5% and climbing in September 2023, which has a strong direct relationship with mortgage rates. Our expectation is that there will be a strong relationship between the federal funds rate and "Median Time on the Market", as a "hot" mortgage market with low rates means more offers coming in for a property, often over the asking price and in cash (Rothstein, 2023). This provides easy decisions for sellers and requires quick action for buyers to lock in advantageous mortgage terms. So, we anticipated strong positive covariance between interest rates and "Median Time on the Market". From a usability perspective, given federal

funds rate's impact on the entire US housing markets, this was a column that applies to each MSA, making mapping straightforward. An interesting point of further exploration would be to see the differentiation between MSAs in their volume reaction to interest rates.

Clear additional areas of macroeconomic interest from the Fed were Real Personal Income, Consumer Price Index, and the Consumer Sentiment Index. Real personal income (PI) has the potential to be of importance given the impact of inflation on disposable income and housing affordability, and therefore the demand for housing. The Consumer Sentiment Index (UMCSENT) and Consumer Price Index (CPIAUSCL) on the other hand may impact consumer's propensity to spend on housing. If consumers feel an economic downturn is coming or feel squeezed by the rising costs of goods against their income, they may be less likely to make a large financial commitment.

Other, more ancillary sources of US-level macroeconomic data of interest to us were available via the Fed's API: the trade balance (BOPGSTB), retail trade index (RSXFS), and the new constructions index (HOUST). The trade balance is somewhat spurious logically but is associated with the kind of work that is being done by the United States economically, reflecting the relative strength of the US economy against its trading partners in areas like commodities and manufacturing. The Retail Trade Index is an even better indicator of a huge driver of US economic activity for much of the country, while the new construction index has the potential to point to pent-up demand (Federal Reserve).

Exploratory Data Analysis

The final dataset consists of 112,062,027 data points and 27 variables. Figure 1 shows the correlation heatmap amongst the numerical variables. With a threshold of .85 the following variables have the highest Pearson Correlation coefficient: "average_listing_price" & "median_listing_price", "total_listing_count" & "active_listing_count", "total_listing_count" and "new_listing_count", "UMCSENT_value" and "CPIAUCSL_value", "RSXFS_value" and "CPIAUCSL_value", "RSXFS_value" and "UMCSENT_value", "PI_value" and "CPIAUCSL_value", "PI_value" and "UMCSENT_value", "PI_value" and "RSXFS_value". To address the multicolinearity, the "average_listing_price" was removed from the dataset as the "median_listing_price" would be a better indicator as it's less impacted by the outliers. Similarly, the "total_listing_count" was not included as the "active_listing_count" provides more accurate insight for the purpose of this study. "RSXFS_value", "CPIAUCSL_value" and "PI_value"pairs.

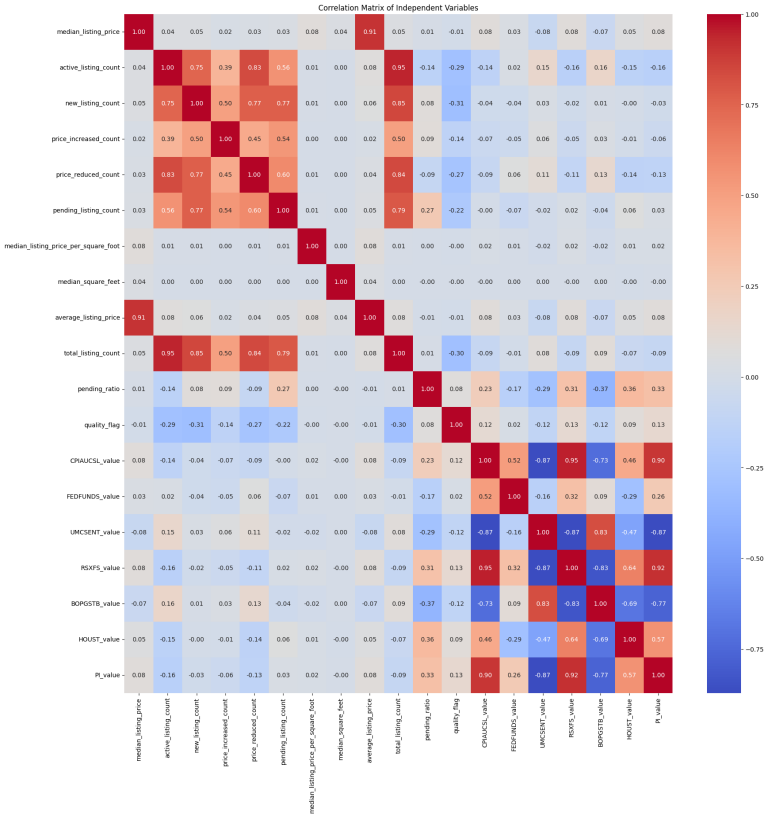

Figure 1: Correlation Matrix

Figure 2 shows the distribution of the dependent variable, 'Median Days on Market', with a KDE (Kernel Density Estimation) line which provides a clearer view of the distribution's shape, indicating the probability density of the data at different points. The distribution is skewed to the right, which means that there are a significant number of listings with a short time on the market, and fewer listings remain on the market for a longer time. The highest

peak is around 50 days, and it can be observed that the most common duration for listings on the market is around 1 to 2 months. The long tail towards the right suggests that there are outliers in the data pointing to the listings that take much longer to sell compared to the majority. This underscores both the variety of differing market conditions that can impact a deeper analysis of local housing trends and market dynamics. It also highlights the need for robust statistical models or transformations that can account for such skewness and outliers to accurately predict the time a listing will stay on the market.

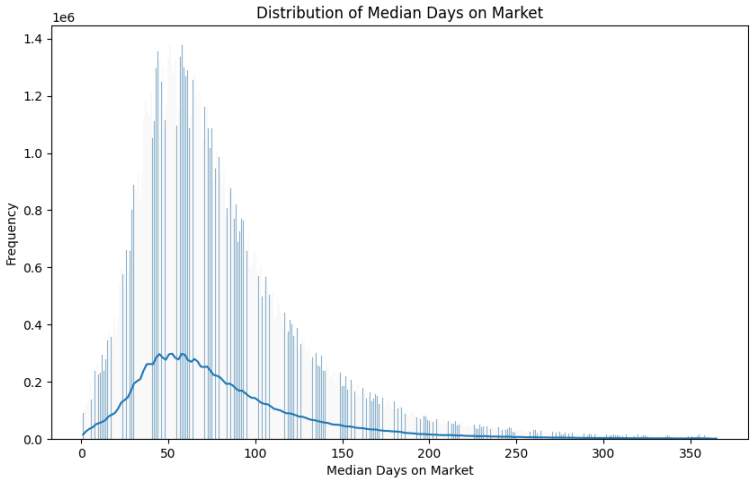

Figure 2: Distribution of Median Days on the Market

Overview of Modeling

Model Selection and Hyperparameter Tuning

For this project, we utilized three modeling approaches to predict the median time on the market of a given home. We explored models through multiple regression, random forest, and artificial neural network (ANN). We explored optimizing our hyper parameters as well, to produce the best fitting models possible for our predictive purposes. We will first create stratified training, validation, and test datasets to assure that these subsets preserve the same proportions as our original dataset. Our training data consists of 60% of our original data, while our test and validation datasets each consist of 20% of the original data. To gauge each model's effectiveness, we will select the model that provides the lowest root mean squared error (RMSE) in our test data.

Multiple Regression

Multiple linear regression is a very common modeling technique that utilizes multiple independent variables to predict the value of a dependent variable. In a linear-linear multiple regression model, the interpretation is that the coefficients associated with our independent variables are the direct relationship between independent variable and dependent variable.

In our multiple regression approach, as mentioned in the EDA section, we explored the multicollinearity of our predicting variables and elected to remove 5 variables. We also removed the "quality flag" variable which was a binary value flagging when data values are outside of their typical range as this information is not meaningful for the purpose of predictive modeling.

Upon running a basic, linear-linear multiple regression model with our data, we generated a model with an adjusted R-squared value of 0.150 and an MSE of 2,256.44, which performs very poorly. This R-squared value implies that only 15% of the variability observed in our dependent variable is explained by our model.

We then ran a log-log multiple regression model, which takes the common log of both the independent and dependent variables when generating coefficients. We interpret these models as where the coefficients are the estimated percent change in your dependent variable for a percent change in your independent variable.

When running a log-log multiple regression model, we saw that our new adjusted R-squared value is 0.300, which is significantly better than our linear-linear model. We also get an MSE of 0.3713 with this model. This improved adjusted R-squared value indicates that 30% of the variability in the log of our dependent variable is now explained by the log of the independent variables, which is a considerable improvement from the 15% explained in the linear-linear model. The MSE value for log-log model shows the average squared difference between the observed and predicted values in a log scale. While it doesn't directly translate to the original scale of the data, this lower MSE signals a better fit of the model to the data when compared to the linear-linear model.

We then removed any outliers from our model as well using Cook's distance, where the threshold is 4/n (Figure 3). This gave us an improved adjusted R-squared value of 0.328. We removed just over 38,000 outliers from our dataset and ran our log-log multiple regression model again, where we obtained an improved R-squared value of .327 and an MSE of 0.3746. We then used Ridge regression & LASSO (Least Absolute Selection and Shrinkage Operator) regression as modeling methods as well.

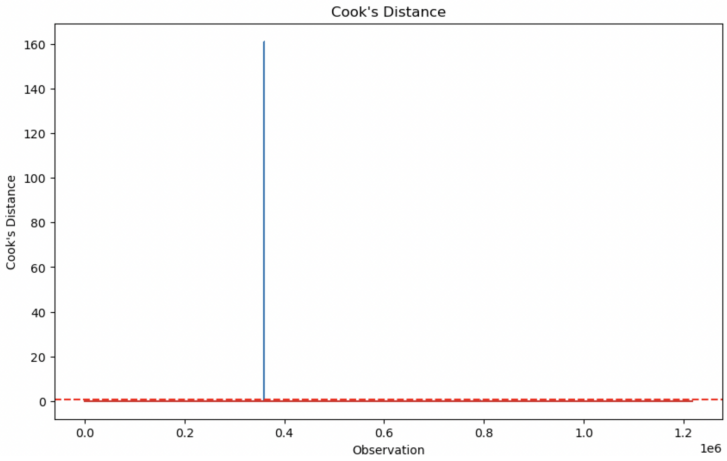

Figure 3: Removing Outliers: Cook's Distance Plot

Ridge Regression is a form of multiple linear regression that helps address the issue of multicollinearity. Ridge Regression introduces a regularization term, which adds the squared sum of the coefficient values. Ridge Regression, ultimately, tends to work better in datasets where there are many large parameters of about the same value, or when most independent variables impact the response variable.

LASSO regression is a specialized linear regression method used specifically for variable selection and regularization. It is like multiple linear regression, but with an added penalty term dependent on the absolute values of the coefficients. LASSO regression tends to work well when there are a small number of significant parameters, while the others are close to zero.

We ran our Ridge & LASSO regression models with log-log transformations without scaling our data, where we saw our Ridge regression with an optimal alpha value of 10.0 provided an R-squared value of 0.2962 and an MSE of 0.3746. Our LASSO regression model with its optimal alpha value of 0.05 provided and R-squared of 0.3964.

We then attempted to scale our log transformed data to search for improved model efficiency. This returned a Ridge regression model with an optimal alpha value of 3.5938, where we received an R-squared value of 0.2962 and an MSE of 0.3746. Our LASSO regression model with the scaled data then had an optimal lambda value of 1e-05, with an R-squared value of 0.2962 and an MSE of 0.3746, identical to the Ridge regression model. Both models performed worse than their respective counterparts in the non-scaled data models, suggesting that scaling is not productive for our model's efficiency.

We then went back to test our log-log regression model on our validation data, where we saw a remarkably similar R-squared to our original model of 0.303. When running this model on our test data, we saw an MSE of 0.3714, RMSE of 0.6094, and an R-squared value of 0.301.

Random Forest

Random forest is a machine-learning classification algorithm that consists of multiple decision trees and combines their outputs to create a single response. The strengths of random forest modeling rely on its ability to handle relatively complex datasets and avoid overfitting. When generating our random forest models, we created a model using our initial data, one using scaled independent variables, and then another with log-transformations.

We used the function RandomForestRegressor from sklearn to create a random forest model that accumulates information from many decision trees and compiles that information to build a classification model. Figure 4 shows the feature importance from our dataset, and we see that "new_listing_count" and "active_listing_count" variables are by far the most important independent variables, while "median_square_feet" is hardly important at all according to rfpimp.

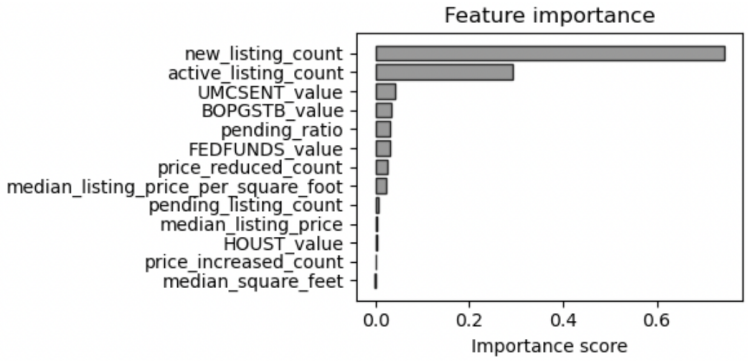

Figure 4: Feature Importance from rfpimp

When running our random forest model with our unscaled data, we obtain an R-squared value of 0.2710 and an MSE of 1,937.27, which shows the value of scaling the data and is the lowest of our unscaled model runs, but which shows clear room for improvement. We then scaled our data and ran a random forest model with that data, where we obtained an R-squared value of 0.2703 and an MSE of 1.0321. Finally, we used our log-transformed data to create a third random forest model. Without fully tuning this model, the resulting MSE was 0.3340, and the R squared value was 0.3724.

For hyper-parameter tuning of our RF model, we tested values of 50, 75 and 125 for the number of estimators and intentionally avoided higher values to avoid adding too much complexity to the model. For maximum depth and minimum samples splits we selected a wide range of values ({5, 8, 12, 15} and {2, 5, 10} respectively) to balance the potential for under-fitting and simplicity vs. over-fitting and complexity trade-offs. Similarly, the values tested for minimum sample leaf (1, 2, 3) ensures balancing Bias and Variance. The optimal parameters on the log transformed data ended up being 125 estimators, max depth of 15, minimum leaf samples of 3, and minimum split samples of 10. These optimized parameters lead to the best accuracy on the validation set, and when attempting on the testing set, lead to a MSE of 0.3174 and a R squared value of 0.4036.

Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a machine learning process that uses interconnected nodes that are helpful in solving complex problems. Information only flows forward in this type of neural network, which is beneficial for predictive analysis for our objective. When generating our vanilla ANN models, we created one on our non-scaled data, one on our scaled data, and then one with our log-transformed data.

First, we generated an ANN model with our regular unscaled data and obtained an R-squared value of –0.07 and an MSE of 2,674.09. These results were a bit confusing for us, but we then obtained superior results based on the transformations that performed better in the other modeling approaches. Our scaled data ANN model provided us with an R-squared value of 0.2766 and an MSE of 1.0232. Finally, we used our log-transformed data to generate our last ANN model, which provided us with an R-squared value of 0.3907 and an MSE of 0.3243.

We then gathered the optimal ANN hyperparameters by using the RandomSearch function from Keras for hyperparameter tuning. With the idea being that by trying random values throughout the hyperparameter space,

you save time and computing power as compared to looking at combinations one by one, while still observing nearly optimized results.

We can clearly identify that our third ANN model that utilizes the log-transformed data performs the best, consistent with our other modeling effort as well. We then ran our hyperparameter-tuned ANN model that utilizes our log-transformed data on our test data and received an R-squared value of 0.3798 and an MSE of 0.3301.

Figure 5 shows the plot of the MSEs on the training data and on the validation data as the ANN iterates through Epochs. What we are looking for here is that it is not getting too consistently close with the training data, as that could be an indication of overfitting. As well as no increasing trend in the validation set MSE values, as this would indicate that our model is somehow getting worse as its iterating through Epochs. The plot below does not strongly suggest concern in either of those two areas.

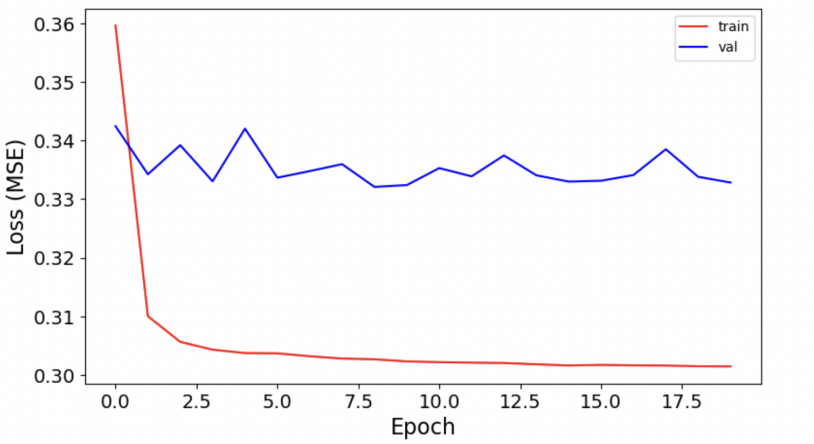

Figure 5: ANN loss function plot on log transformed data (MSE, 20 epochs)

Results and Analysis

Using MSE and R-squared as our evaluation metrics the Random Forest model comes out on top as the best performing model as compared to the other iterations of the model approaches we tried (Figure 6). However, there is still an appreciable amount of value with the other models, especially in adding interpretability to the various drivers of time on the market that the Random Forest model by its nature lacks.

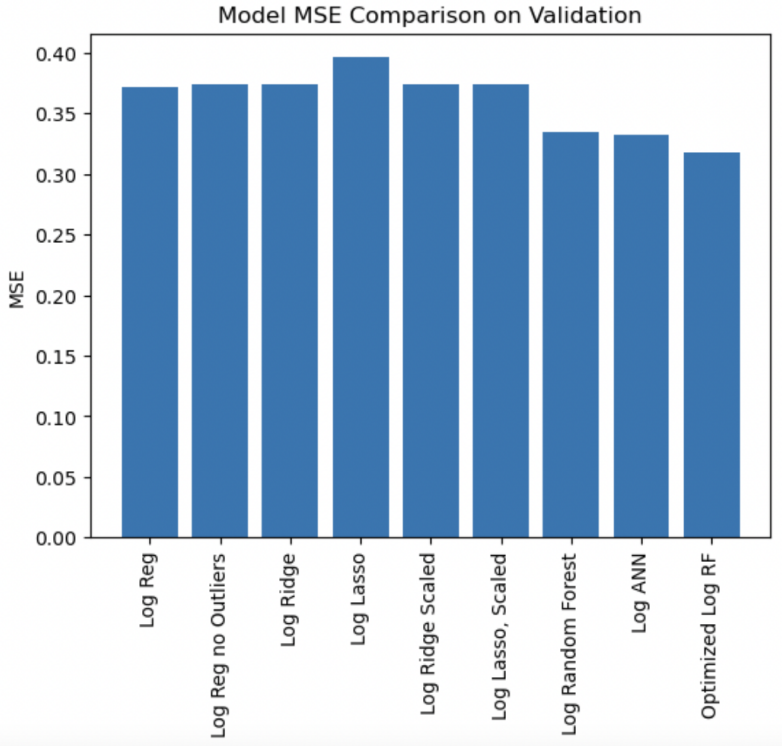

Figure 6: Model MSE Comparison on Validation

Looking at the log-log regression model, the highest coefficient values were "ln_active_listing_count "(0.4704) and "ln_new_listing_count" (-0.4386). This matches the feature importance metric coming out of the random forest model. While not perfectly statistically rigorous to draw a connection between the two outputs, the interpretability benefit coming from the regression model in understanding each of the variables adds a significant amount of context on the market forces, while still allowing for the implementation of the higher performing, less interpretable Random Forest model. Being able to give stakeholders context like for every 1% increase in the Active Listing Count we'd expect a 0.4% increase in Day to

Pending within a market is valuable, even if the optimal model takes a different and more opaque approach to modeling that relationship.

Conclusion and Key Takeaways

As demand for housing continues to rise in the United States, understanding the components of a mortgage transaction is central for several stakeholders, but most especially homebuyers and sellers. The ability to take a data-driven view to a critical factor like time on the market adds rigor to a metric that is at times clouded by judgement calls on how long is too long within a given market at a specific point in macroeconomic time. Defining how market and macroeconomic conditions impact time on the market represents a step toward clearing that cloud.

Outside of creating a workable model, there were a few implications of the research that jumped out. For one, our hypothesis was that many of our variables would covary with the federal funds rate. A “hot” market with a low federal funds rate would have more price increases, more active listings, and a higher listing price. Looking at the correlation matrix, however, we saw limited covariance between those factors. Each of the realtor.com-sourced factors generated were highly correlated with one another as we expected, and some of the macroeconomic variables were correlated with one another as well, but the very bare quadrants representing the two groups" overlap was striking despite most of the factors significantly impacting the median time to close.

From a model performance perspective, we saw the best success for the optimized Random Forest model running on our log transformed, scaled data. An interesting observation was the differing performance of the scaled data between models contrasted with the similarly superior results for the log transformed data.

There are several different areas of exploration that could be built on this analysis. Given the relative “stickiness” of the home market in the sense of time it takes to prepare to list and remove listings from available inventory, adding lagged versions of the macroeconomic variables to varying degrees could generate interesting results, and perhaps generate more of an association with median time on the market and our realtor.com-sourced data. Adding more factors associated with specific zip codes that we reviewed, including diving deeper into Census Bureau, may help to identify zip code specific variation that may have been lost with the broad brush with which we painted the macroeconomic climate. Furthermore, due to the time constraint only a limited number of algorithms and smaller range of their hyperparameters were explored in this study which leaves plenty of opportunity for future research.

Overall, the relationships we did identify between our data sources with median time on the market explained a good deal of its variation. This paper provides a great first step for interested stakeholders to begin to model this datapoint in their business decisioning within the housing market.

Works Cited

Rothstein, R. “Mortgage Rate Forecast for 2023.” Forbes, October 31, 2023. https://www.forbes.com/advisor/mortgages/mortgage-interest-rates-forecast/.

Benefield, J.D., Hardin, W.G. Does Time-on-Market Measurement Matter? J Real Estate Finan Econ 50, 52–73 (2015). https://doi.org/10.1007/s11146-013-9450-z

Knight, J.R., Listing Price, Time on Market, and Ultimate Selling Price: Causes and Effects of Listing Price Changes. Real Estate Economics, 30: 213-237. https://doi.org/10.1111/1540-6229.00038

Castelli, M., Dobreva, M., Henriques, R., Vanneschi, L., "Predicting Days on Market to Optimize Real Estate Sales Strategy", Complexity, vol. 2020, Article ID 4603190, 22 pages, 2020. https://doi.org/10.1155/2020/4603190

McGreal, S., Adair, A., Brown, L., Webb, J. “Pricing and Time on the Market for Residential Properties in a Major U.K. City”, Journal of Real Estate Research, 31:2, 209-234, DOI: 10.1080/10835547.2009.12091248 https://doi.org/10.1080/10835547.2009.12091248

Realtor.com. “Realtor.com Real Estate Data and Market Trends.” Realtor.com Economic Research, www.realtor.com/research/data/. Accessed 2 Oct. 2023.

Federal Reserve Bank of St. Louis. “FRED Economic Data.” Stlouisfed.org, Federal Reserve Bank of St. Louis, 2023, fred.stlouisfed.org/.