SMU Data Science Review

Volume 5 | Number 2 Article 4

2021

Machine Learning Forecasting of Construction Demand in the USA as Indicated by Economic, Geographic and Demographic Cues

Jayson Barker
Southern Methodist University, jsbarker@mail.smu.edu

Emil Ramos
Southern Methodist University, emilr@smu.edu

John Rodgers
Southern Methodist University, jdrodgers@mail.smu.edu

Saqib Shahzad
Southern Methodist University, sshahzad@mail.smu.edu

Machine Learning Forecasting of Construction Demand in the USA as Indicated by Economic, Geographic and Demographic Cues

Jayson Barker, Emil Ramos, John Rodgers, Saqib Shahzad,
Peter Beckmann, Nibhrat Lohia, Martin Selzer
Master of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
jsbarker@mail.smu.edu, emilr@mail.smu.edu, sshahzad@mail.smu.edu,
jdrodgers@mail.smu.edu

Abstract. This paper aims to optimize the investment and position of a building supply company by using a predictive modeling approach that focuses on single-family housing starts in the United States. This approach aims to guide where capital investments such as truss-plants and lumber yards should be constructed before housing demand in a particular Metropolitan Statistical Area (MSA) rises. By accurately predicting these MSAs, a building supply company could save costs and improve service while establishing a competitive advantage in the market. This study leverages Multi-Linear Regressions, XGBoost Regression, and Artificial Neural Network (Multilayer Perceptron - MLP) algorithms to identify the top 5 MSAs measured in total single-family housing starts (SFH). These MSAs include (1) New York-Newark-Jersey City (2) Dallas-Fort Worth-Arlington, TX, (3) Houston-The Woodlands-Sugar Land, TX, (4) Atlanta-Sandy Springs-Alpharetta, GA, and (5) Phoenix-Mesa-Chandler, AZ. Regarding MSAs that are projected to experience growth, however, the following top MSAs were identified: 1) Knoxville, TN, 2) El Paso, TX, 3) Cleveland-Elyria, OH, 4) San Antonio-New Braunfels, TX, and 5) Tucson, AZ. Based on these results, this study suggests investments in infrastructure and capabilities should be considered in these areas primarily due to the projected growth and demand for single-family residential construction.

1 Introduction

The building supply industry is a multibillion-dollar industry with enormous potential to employ predictive analytics to improve return on investment (ROI). Construction is also at an all-time high. If building supply companies are strategic about where and when they construct lumber yards and truss manufacturing plants, they can capitalize on these skyrocketing trends. Accurately predicting where new housing demand will occur before it occurs is not easy, but doing so can be a distinct competitive advantage. Major capital investment decisions include establishing lumber yards, supply chain infrastructure, manufacturing facilities, and hiring staff. By leveraging accurate predictions of specific MSA within the United States primed for growth, a building supply company could lower costs by establishing operations closer to the building activity improving delivery times and customer service. Throughout the

development of this approach, a building supply company was involved to guide approach, business considerations, and service radius data.

There have been many approaches to predicting housing growth for a variety of reasons. There are, however, limited studies specifically targeted at informing decision-making for building supply companies on where their next capital investment should be to take advantage of housing demand targeting specific geographic regions. By leveraging data such as Gross Domestic Product (GDP), income per capita, housing permit requests, and employment data targeted at specific MSAs, a model could be constructed to identify features most critical to predicting when and which MSAs will experience growth [1] [2].

New residential single-family houses are a major source of revenue for building supply companies. Before new single-family construction can begin, a home builder must obtain an approved building permit from the county in which the home will be constructed. Plans for the building site, including any specific county requirements, are submitted to the county for review. Once approved, the building can begin; however, permits can take multiple weeks to be approved. Approval timing can be a source of considerable delay if it is not factored into a home build.

While the builder is working through the permit approval process, supplies are ordered from building supply companies. These supplies include framing lumber (2x4s, studs), structural panels (plywood, oriented strand board), manufactured products (prefabricated roof or wall/floor trusses), and a variety of other building materials needed for a new home build. These products are not all ordered at the same time, however. Certain items need to be completed before the next phase of the build can begin. A typical build includes the following steps: 1) preparing the site and pouring the foundation, 2) framing, 3) plumbing, electrical, HVAC, 4) insulation, 5) drywall, fixtures, 6) interior trim, driveway, walkways, 7) flooring, countertops, exterior grading, 8) mechanical trim, bathroom fixtures, 9) mirrors, shower doors, landscaping, and 10) final walk-through. In the earliest phases of the build (i.e., steps 1 and 2), most of the lumber supplied by a lumber yard and/or pre-manufactured trusses are consumed. Provided the project is planned with the appropriate permit lead time and supply delivery timeframe in mind, the planner can arrange to have supplies available just as soon as the specific building phase is ready. Moreover, with a lumber yard close by and manufacturing facilities already established, the builder would benefit from reduced lead times on ordered materials, reduced delivery, and shipping charges. The builder would also benefit from having an established local source of supply to support the construction process.

The new home construction process is generally the same across the United States. The same supplies are universally required to complete a build with variations depending on climate, styles, budget, and options selected by the builder and/or the buyer. Given this common thread, this study targets all states within the United States by MSA. This approach allows for the comparability of SFH predictions across a vast geography. Additionally, building supply companies are subject to macro-level construction trends and impacts. A wide geographic view would be useful in evaluating where to invest capital, especially if a particular state (or states) may not be optimal for investment in a specific time.

Three predictive modeling approaches were used to identify specific MSAs primed for investment based on their aggregate totals of predicted SFHs. The three modeling approaches used include Multi-Linear Regressions, XGBoost Regression, and Artificial Neural Network (Multilayer Perceptron). The results predict that when measured in total SFHs within the United States, (1) New York-Newark-Jersey City (2) Dallas-Fort Worth-Arlington, TX, (3) Houston-The Woodlands-Sugar Land, TX, (4) Atlanta-Sandy Springs-Alpharetta, GA, and (5) Phoenix-Mesa-Chandler, AZ are best suited for capital investment. Factoring in projected growth rates in SFHs, however, the following top MSAs were identified: 1) Knoxville, TN, 2) El Paso, TX, 3) Cleveland-Elyria, OH, 4) San Antonio-New Braunfels, TX, and 5) Tucson, AZ. A building supply company would be well-positioned to capitalize on their investments if they focused on building lumber supply and truss plants in any of these MSAs within the next three years to support the anticipated growth in demand for single-family home construction.

2 Related Work

Accurately predicting residential construction activity in regional housing markets is used in numerous banking, government, utility, and retail applications [12]. For example, retail stores use regional housing forecasts to ensure their stores have the necessary staff and inventory levels to operate. Fullerton et al., 2000 used univariate ARIMA and random-walk models to assess the accuracy of residential construction forecasts. The forecasts used in the study for single-family starts were compared with univariate time series and random-walk alternatives. The data on single-family starts used in this study were derived from quarterly forecasts between 1985 (first quarter) and 1996 (second quarter) [12]. The results from the study indicated that the accuracy of the estimates for regional single-family construction does not compare well to forecast accuracy from univariate ARIMA equations or random-walk predictions. Based on this study, the data in this paper should consider random time series specifications. The study recommended that future models should consider increasing geographic coverage. This geographic expansion could help predict MSAs within the United States primed for building supply investment.

To understand the housing demand in a particular MSA, Case et al. studied surveys from homebuyers from 2003 through 2012 to understand homebuyers' expectations [10]. The timeframe used in this study highlights the fluctuation of the housing market with consideration to the economic recession that began during the fourth quarter of 2007. Although there are many literatures on potential reasons that led to the economic recession, there is not much on the role of homebuyers' expectations during that period. This study highlights the homebuyers' thought process when purchasing a home by highlighting their perceptions, interpretations, and opinions during that period. The survey results suggest that homebuyers were aware of trends in home prices when they made their purchase and highlighted a strong correlation between the respondents' understanding of price trends and the actual movements in prices [10]. The study also highlighted that these homebuyers were aware of the short-term changes in the housing market. Still, it did not change their thought process related to purchasing a home,

indicating that homebuyers are aware of what is going on in the housing market, which can impact population migration.

Another study that predicted the housing demand in certain MSAs examined the time-series relationship between house prices in eight Southern California MSAs [26]. The evidence in the study suggests that a purchasing power parity is a link between these eight MSAs. Each coastal MSAs (Oxnard, San Diego, San Luis Obispo, Santa Ana, Santa Barbara) house price index temporally causes the Los Angeles index price. On the flip side, the Los Angeles house price index influences the house price index for these same coastal MSAs. In Mao et al. study, they predicted the housing demand in China by using the macro data on the housing market in Hangzhou during 1999-2012 to establish a forecasting model based on a backpropagation neural network of genetic algorithm optimization [28]. The model that was created in this study resulted in high precision accuracy, however, there were fluctuation years in the prediction that was caused by housing control policies in China. For example, there were new regulations issued by the State Council in 2011 to curb rising housing prices which caused the consumers to assume a wait-and-see posture to bring about low turnover [28]. The wait-and-see approach from consumers in China is similar to that consumers in the United States can relate to, which can influence the housing market in the states.

Survey data is used in Meyer's study [11], to understand the directional accuracy of housing starts in the United States. To test the accuracy of forecasts, Meyer used techniques that have been developed for understanding the Relative Operating Characteristic (ROC) curves. The insights from a ROC curve provide value since it incorporates the results for all decision criterion values, which ultimately leads to numerous individual assessments of the directional forecast accuracy [11]. The study results show that forecasts from surveys are significant when it comes to forecasting directional change in housing starts. Further investigation could deepen the understanding of available survey data, investigate the underlying forecasting models and loss functions of the surveys' panelists, and eventually combine the data with other forecasting models and ROC measures of directional accuracy to improve overall forecasting performance.

To forecast home sales, Dua, Miller, and Smythe [18] describe aspects of economics that are important in forecasting home sales, such as mortgage interest rate and current and future economic conditions. The methods used by Dua et al. are based on Vector Autoregressive (VAR) and Bayesian Autoregressive (BVAR) models. Gupta, Marco Lau, Plakandaras, and Wong [19] expand on these indicators and combine methods based on economic indicators with strategies based on housing sentiment to predict home sales. Housing sentiment represents potential home-buyers opinions related to the current conditions for buying a home. Housing sentiment can be interpreted from consumer survey data. This housing sentiment data, when combined with economic factors such as interest rates, income, and unemployment rates, can determine the demand for housing. In contrast, the new housing starts and permits are indicators of the supply of available houses. Gupta et al. (2021) utilized the housing sentiment methods defined by Bork, Møller, and Pederson [20] to create a sentiment index based on survey data collected by the University of Michigan. Bork, Møller, and Pederson developed a quarterly index beginning in 1975 and demonstrated that this sentiment analysis more closely aligned with actual market conditions in comparison methods based entirely on economic indicators such as the study by Dua et. al.

With the development of new ways of collecting real estate data, there is potential to leverage geographic dependencies for enhancing real estate appraisal. Data from real estate appraisals can also be used to better understand the housing market instead of identifying trends in the housing market, such as housing prices. In the Fu et al. study, they used a geographic method, called ClusRanking, for real estate appraisal by leveraging the mutual enforcement of ranking and clustering power [4]. Their ClusRanking method was used to predict real estate investment value with geographic information. Their problem formulation for the predictive model as follows. The estate value is mainly affected by three influential factors: $y_i = F(y_i, \rho_i\ \delta_i)$, in which (1) y_i: the geographic utility extracted from urban geography data F_{geo}; (2) ρ_i: the influence of latent business area F_{area}; (3) δ_i: the neighborhood popularity estimated from human mobility data F_{mobi}. These three factors are important in evaluating real estate investment value, because the geographic utility, the neighborhood popularity, and the influence of business areas represent three different perspectives: (1) land uses, (2) human beings, and (3) business potential, respectively [4]. These factors can be considered for this paper to accurately predict where a potential new market is for single-family homes.

Another influencing factor that impacts consumers when searching for a new home is property tax rates. In Case et al. study, they used county data from Wisconsin over twelve years. They found that property tax rate differentials harm the construction of single-family houses [5]. The results indicated that either an increase in property tax rates, the average property tax rate in a particular region, or the county's property tax rate reduces the building work on single-family housing starts. The study used a time-series approach to test the hypothesis that differences in property tax rates across multiple regions impact housing starts and developed a model that allows for inter- and intra-regional influences on housing starts [5]. The property tax rates for MSAs are a key factor to consider in a model when accurately predicting new housing builds.

Applying machine learning techniques is one way to predict housing price movements across the United States. One machine learning approach is the random forest method to accommodate multiple predictors and nonlinearities [21]. Gupta et al. used random forests consisting of 50, 75, or 100 random trees to forecast results using a minimum number of observations of five per terminal node and several rounds (number of predictors/3) of randomly chosen predictors for splitting [21]. This was followed by a time-series of a ten-year rolling-estimation window and three forecast horizons: one, three, and twelve months. However, Gupta et al. suggested that future research go beyond the state-level analysis and explore the specific MSA levels to obtain the national and regional housing factors for the predictive analysis.

3 Methods

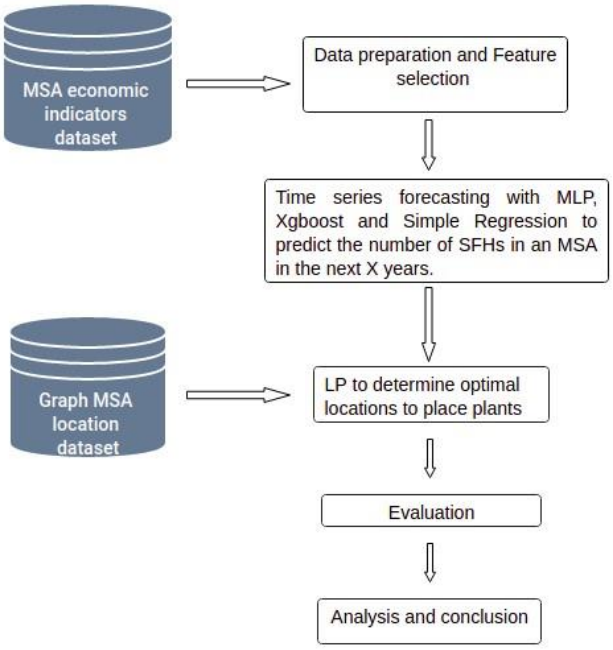

Fig. 1. ***Project flow diagram***

A. Problem Formulation

Given a set $M = \{m_1, m_2, m_3, \dots, m_n\}$ of MSAs in United States; physical distance between the MSAs; and the historical economic, demographic, and geographical performance of the MSAs; this study computes the optimal number of construction plants k a building supply company should construct, and the optimal locations (where to place the plants), such that both the number of customers benefiting from the plants and the company's profit are maximized. This approach attempts to quantify the company's profit by using a single indicator named the number of single-family homes in an MSA. Naturally, the number of SFHs reflects the extent to which a community uses construction materials, benefiting from the established construction plants, and the extent to which the building supply company sells its products, thereby yielding profit.

The problem is treated as a two-phase modeling problem. In the first phase, predictive modeling and time series forecasting are performed using historical, economic, demographic, and geographical MSA features (See Table I) to identify MSAs that are most likely to soon witness an influx of SFHs. In the second phase, MSAs are clustered based on their location and the predicted number of SFHs (from phase 1) to determine optimal locations to place the construction plants. The summary of the workflow and the modeling phases is as illustrated in figure 1.

B. Phase I: Single-Family Housing Start Forecasting

Dataset:

This approach used the United States Census Bureau dataset on construction and demographic activities. MSA and spans order the dataset from 2001 to 2019. After preprocessing the data to remove duplicates and combining features into a CSV file, 11 features remain to train a time series forecasting model. The list of selected features is as illustrated in Table I. For justification on the utility of the selected features on SFHs forecasting, please refer to Section II.

No.	Feature	Feature group
1	GDP of all industries	Economic
2	GDP of construction industry	
3	GDP per capita	
4	Unemployment income	
5	Number of employments	
6	Personal income	
7	Unemployment rate	
8	Wages salaries	
9	Number of single-family houses	Demographic
10	Population	
11	State/ Region	

Table I - Features used in the housing forecast model

Algorithm*:*

Forecasting single-family housing starts (SFHs) is, at its core, a time series problem. However, the data available from the U.S. Census Bureau is not in time series form. To ensure that this is a pure time series problem, the data was transformed as follows:

$$X12001, X22001, \ldots, XN2001 \longrightarrow SFH2002$$
$$X12002, X22002, \ldots, XN2002 \longrightarrow SFH2003$$
$$X12003, X22003, \ldots, XN2003 \longrightarrow SFH2005$$

Where $X1_{2001}$ shows the feature 1 (example, population) of an MSA in the year 2001. The above example dataset is created using a time span of 1, which means that the prediction is based on last year's statistical values.

In the final modeling approach, a time span of 5 years was used, which means the prediction is based on the previous 5th year as given:

$$X12001, X22001, \ldots, XN2001 \longrightarrow SFH2006$$
$$X12002, X22002, \ldots, XN2002 \longrightarrow SFH2007$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$X12019, X22019, \ldots, XN2019 \longrightarrow SFH2024$$

The forecasting model is based on three algorithms using train and test data sets: Multi-Linear Regressions, XGBoost Regression, and Artificial Neural Network (Multilayer Perceptron). In each case, the first 18 years of historical data were leveraged for training the model, and the year 2019 was leveraged for testing. Across the three algorithms, the parameters and model architectures that reported the best results across the three algorithms.

No.	Parameter	Value	Algorithm
1	activation function	relu	MLP
2	alpha	0.0001	
3	batch size	auto	
4	epsilon	1e-08	
5	hidden layer sizes	(500, 400, 100)	
6	maximum iteration	2000	
7	Optimizer	Adam	
1	booster	gbtree	XGBoost
2	importance type	gain	
3	Learning rate	0.01	
4	max depth	5	
5	objective	reg:squarederror	
6	n estimators	500	
1	Parameters	fit_intercept=True normalize=False	Linear Regression

Table
Table II - Machine learning models and their parameters

Among these models, the performance of MLP outperformed in terms of Root Mean Squared Error (RMSE). The performance has been evaluated on a state level as well to gauge if RMSE varies between states. For example, in states like Texas, the RMSE was much greater than in other states. To tackle this, two modifications were introduced in the methodology. First, the state was included as a feature since SFHs vary from state to state and from region to region. Second, an ensemble technique was applied (combination of all above models - taking the average of predictions) to overcome the bias behavior of the models due to varying RMSE at the state level. The performance in terms of error (difference between actual and predicted values) is shown in figure 2. In this figure, the predictive pattern of the MLP model more closely resembles the actual SFHs values indicating that it outperforms even the ensemble model.

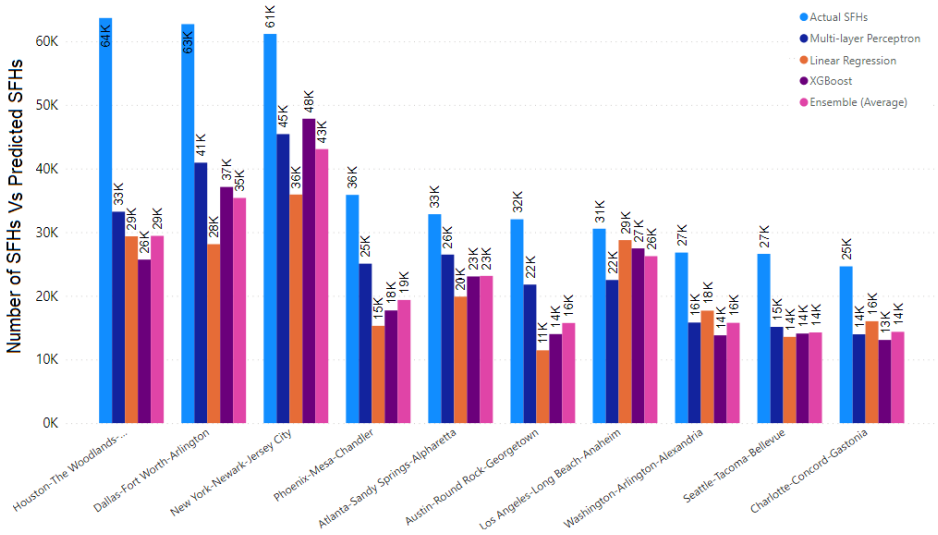

Fig. 2. Performance of algorithms

The RMSE is one technique that can be used to evaluate how well the dataset fits the model. It provides a measure of the distance (average) between the actual values from the dataset and the predicted values from the model. Generally, a lower RMSE means that the model and data fit better. Another technique is the Mean Absolute Error (MAE). This measure averages the error across the dataset, and a lower MAE is preferred. Both measures were calculated for all models, are used to evaluate model performance, and are shown in Table III:

No.	ML Model	RMSE	MAE
1	Multi-layer Perceptron	3238.69	1395.94
2	XGBoost	3634.69	1414.46
3	Linear Regression	4120.28	1781.21
4	Ensemble (Average)	3572.64	1415.05

Table III - RMSE and MAE for each machine learning algorithm

Table III shows that MLP outperforms the base model XGBoost and Linear Regression as indicated by the lowest RMSE and MAE scores. This can be attributed to the ability of the MLP to model to extract complex patterns from the data. The significance of this finding is that SFHs are a function of time and that the future number of SFHs is significantly affected by the geo-economic features and the past SFHs values.

***Feature Importance**:*

SFHs forecasting is a complex problem that depends on many features, including the features selected for this approach and other inherent features like climate change, available land, etc. All these features play a significant role, but all features are not equally important. In this study, feature importance was evaluated using the XGBoost technique. Feature 001 (i.e., GDP of all industries) was identified as the most important feature, and the GDP construction (i.e., the GDP of construction companies) is the second most important feature. Employment opportunities are the third most important feature, as shown in Figure 3.

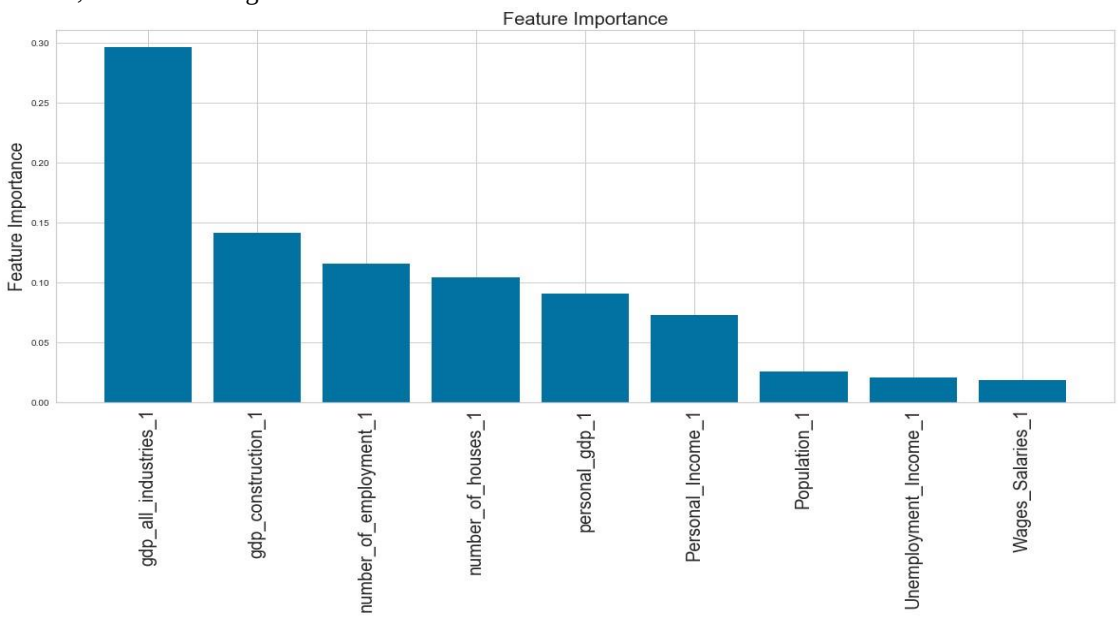

Fig. 3. Feature importance plot

C. Phase II: Centroid/ Plant Location Identification

After forecasting SFHs for 2024, the second problem is to identify the optimal locations to build plants. These locations should be optimized to serve a maximum number of regions using minimum resources. We used the criteria that the average distance from a plant to its closest regions should be less than 100 miles. The 100-mile constraint was requested by the business to limit shipping costs.

KMeans Clustering Algorithm:

The longitude and latitude information was gathered for the MSAs in the United States. The resulted geolocation dataset was used to position MSAs relative to one other in a complete graph. Using these locations, KMeans clustering is applied to find centroids of these MSAs. In this study, the centroids are the optimal locations where the company should build a plant to manage resources optimally.

The KMeans clustering algorithm is used to partition a given set of observations into a predefined amount of k clusters. In this study, clusters are the set of regions served

by k plants. The KMeans algorithm starts with a random set of k centroids (μ). During each update step, all observations are assigned to their nearest centroid. In the standard algorithm, only one assignment to one center is possible. If there is a case where multiple centers have the same distance to the observation, a random one would be chosen from given centroids.

In this study, k=60 was chosen because it was the optimal number of centroids given the current business climate. However, k=60 is dynamic and will allow modifications in the future as business needs change. For the value of k, 60 optimal locations for plants were identified, as shown in Figure 4. These locations were in the form of latitude and longitude. Reverse geocoding is applied to get the names of locations. Based on the output from the algorithm, some of the top locations where plants can be built to get maximum profit from SFHs are California, Texas, New Jersey, and Atlanta.

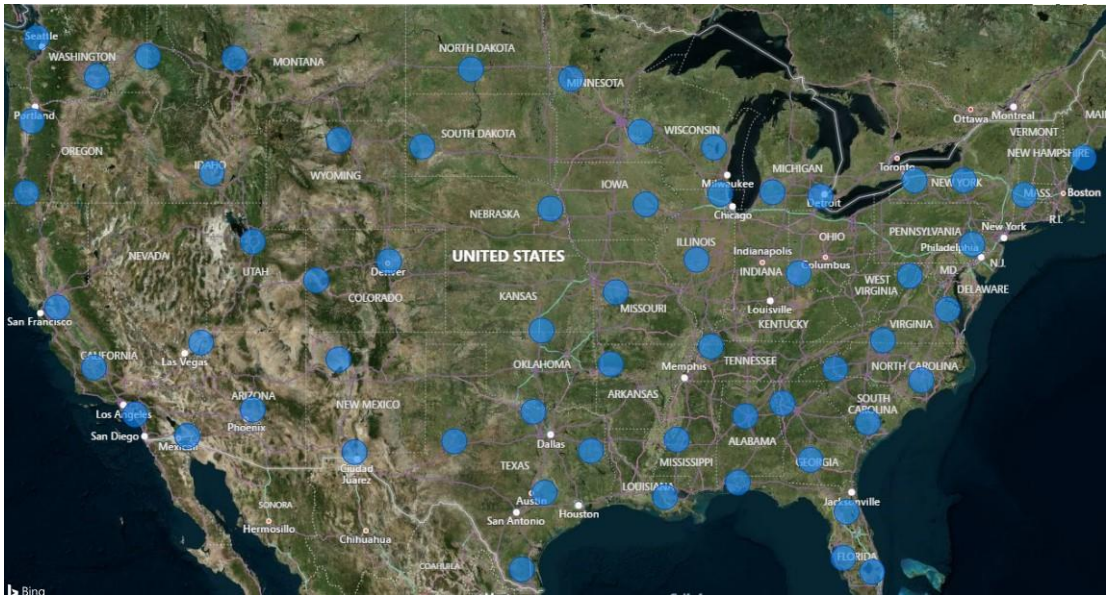

Fig. 4. 60 Optimal points / locations for constructing Plants

4 Results

After applying the supervised techniques (Multi-Linear Regressions, XGBoost Regression, and Artificial Neural Network) and an unsupervised technique (K-Means), a set of plant location recommendations were generated. This included the average number of SFHs forecasted through the year 2024. The 'number of houses in 2024' is the target variable, and the 'number of houses in 2019' is the predictor. The top MSAs and regions where the SFHs will be high (in aggregate) in 2024 are given in Table IV below. The results show that MSA's will have numerous houses built in 2024 enabling

the business to decide where to make future plants to service these MSA's. It is essential to clarify that though we used the year 2024 for our forecasting, it can easily be any year that the business wants, given that each plant takes about a year to be planned and built and be fully functional.

No.	Geo Name	State	SFHs in 2024
1	New York-Newark-Jersey City	NY-NJ-PA	41650
2	Dallas-Fort Worth-Arlington	TX	35698
3	Houston-The Woodlands-Sugar Land	TX	33081
4	Austin-Round Rock-Georgetown	TX	27656
5	Atlanta-Sandy Springs-Alpharetta	GA	24200
6	San Antonio-New Braunfels	TX	24088
7	Phoenix-Mesa-Chandler	AZ	19710
8	Los Angeles-Long Beach-Anaheim	CA	18011
9	Washington-Arlington-Alexandria	WA	16114
10	Las Vegas-Henderson-Paradise	NV	14264

Table IV – Top regions / MSAs for investment via predicted SFHs

To validate the model's performance, the existing company's ranking logic was used as validation data (a non-statistical approach) vs. the forecasting model's output. The top MSAs (in order of the company's existing ranking algorithm) are displayed in Table V below.

No.	Geo Name	State
1	Riverside-San Bernardino-Ontario	CA
2	Los Angeles-Long Beach-Anaheim	CA
3	Austin-Round Rock	TX
4	Seattle-Tacoma-Bellevue	WA
5	Denver-Aurora-Lakewood	CO
6	Phoenix-Mesa-Scottsdale	AZ
7	New York-Newark-Jersey City	NY-NJ-PA
8	Washington-Arlington-Alexandria	DC-VA-MD-WV
9	Dallas-Fort Worth-Arlington	TX

Table V – Company ranked MSAs for future investment

One of the advantages of using Machine Learning is that it can extrapolate complex patterns out of data that are not possible by using ranking or simple scoring functions. Because of this, the results between the company's approach and the forecasting model's approach differ in terms of houses construction prediction, the optimal position of plants and plants, regions served by the plant. However, there is some overlap in MSAs such as Los Angeles, Dallas, and New York. The forecasting model comes complete with scoring metrics to quantify the error

and accuracy, whereas the legacy scoring model does not. Therefore, beyond comparing the MSAs identified for overlap, a qualitative analysis would be the next step to identify further any additional influences that may change the listings and validate the predictive model.

Along with predicting the absolute number of SFHs, providing relativity to the absolute number is highly beneficial. An example is evaluating the last known SFH start value (i.e., the actual value) and comparing the percent change to the predicted SFH value in 2024. This creates a growth metric that effectively normalizes each MSA and allows for comparison. In the example in Table VI below, Wheeling, WV was selected as the top MSA projected to experience the largest growth in terms of SFHs, followed by Johnson City, TN, and Battle Creek, MI. These results illustrate the percent change from 2019 to the predicted 2024 values. Even so, these MSAs all have a relatively small number of starts in 2019.

No.	Geo Name	State	2019	2024	% Growth
1	Wheeling	WV-OH	12	738	6048.68
2	Johnson City	TN	54	2044	3686.64
3	Battle Creek	MI	56	1044	1764.30
4	Fairbanks	AK	22	407	1748.31
5	Lawton	OK	67	970	1348.31
6	Bay City	MI	83	1170	1310.11
7	Niles	MI	206	1479	617.83
8	Grand Forks	ND-MN	185	1211	520.82
9	Great Falls	MT	134	828	518.345
10	Pittsfield	MA	276	1595	477.91

Table VI – Top 10 regions / MSAs identified for investment based on SFH growth rate

To reduce the number of MSAs with a relatively small number of SFHs in 2019, an additional filter was applied that returned the MSAs with a greater number of SFHs in 2019 than the cohort average. This produced a more useful set of results as it identifies MSAs with a larger number of SFHs in 2019 that are projected to experience growth into 2024. Table VII displays the MSAs and results below, identifying Knoxville, TN as the leading MSA, followed by El Paso, TX and Cleveland-Elyria, OH.

No.	Geo Name	State	2019	2024	% Growth
1	Knoxville	TN	3947	6875	74.18
2	El Paso	TX	3066	5007	63.29
3	Cleveland-Elyria	OH	2032	4680	54.35
4	San Antonio-New Braunfels	TX	15895	24088	51.54
5	Tucson	AZ	4313	6216	44.13

6	Spokane-Spokane Valley	WA	3300	4412	33.70
7	Little Rock-North Little Rock-Conway	AR	3005	3936	30.98
8	Fort Collins	CO	2490	3080	23.68
9	Savannah	GA	2591	3026	16.80
10	North Port-Sarasota-Bradenton	FL	3553	4026	13.33

Table VII – Top 10 regions / MSAs identified for investment based on SFH growth rate – above average cohort

The company provided a list of their current plant locations across the United States. This list was used to validate the MSAs identified by the model. The identified MSAs were plotted against the company's existing locations visually to demonstrate whether the model accurately identified MSAs that were good choices for lumber yards and/or truss plant construction. Figure 6 shows green dots as the predicted MSAs determined by the model and blue dots for the company's existing facilities. A striking overlap between the two indicates that the model correctly identified areas the company has already considered an essential and worthwhile investment. The red dots demonstrate the future value the business can acquire from this model by adding commercial railways that can be used to efficiently supply lumber yards, as railways are one of the cheapest forms of delivery.

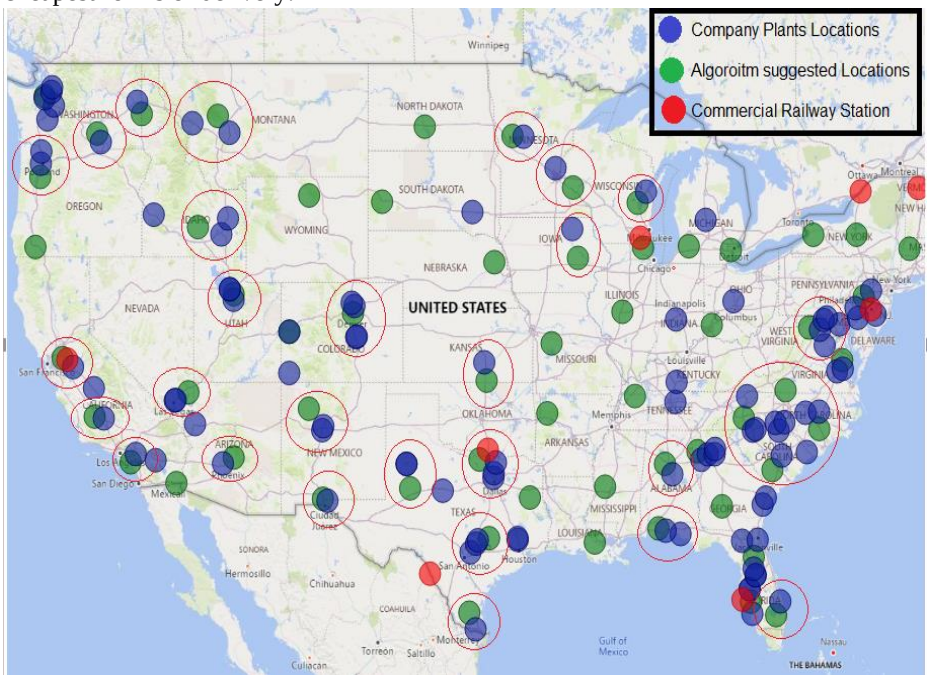

Fig. 6. Algorithm suggested company plant locations (green) vs. existing company locations (blue)

5 Discussion

Of the top 5 MSAs (in aggregate SFHs) predicted by the model, Texas and Arizona correlate well with the mainstream news of large migrations to these areas. Texas and Arizona are routinely identified as prime moving locations for many Americans in 2021 due to the lower cost of housing. In fact, four of the top 10 MSAs identified by the model are located within Texas, which shares some overlap with the company's ranking model. Surprisingly, the New York-Newark-Jersey City MSA is the top MSA in terms of SFHs. It is known that New York is experiencing a large outflow of population to other states. Thus, ranking as the top MSA given the outflow of the population is unexpected. This could be attributed to the larger population that resides in this MSA and, therefore, a larger number of SFHs relative to the rest of the nation. When evaluating the growth rate of SFHs (comparing the latest actual to the prediction), this MSA does lose ground relative to the rest of the country, which is more in line with expectations.

This study is bound by datasets that are prepared at the MSA level. That limits the number of datasets that could be included substantially. Moreover, the existing datasets that were used required extensive transformation to enable the analysis. While additional data sets could be constructed, they would need to be aggregated and mapped to an MSA before using this model. If available, however, other data sources such as weather conditions, education, healthcare, property tax, and grocery store / fast food facility construction by MSA would prove helpful in improving the accuracy of the model. Identifying ideal weather conditions by MSA would act as a draw for transplants searching for a better climate. New school and/or university construction could serve as a leading indicator of expected population growth relocating and, thus, new home construction. From the Case et al. study, property taxes have a negative impact on SFH growth [5]. Incorporating that data into the model would further help eliminate MSAs that may not be optimal. Grocery store and/or fast-food facility construction would also indicate expected population growth. However, this data would need to be collected and synthesized at the MSA level for use in this approach.

Expanding this analysis beyond the construction industry is an evolution for this model. The health care industry could benefit from this approach with minor adjustments. With the introduction of insurance company-sponsored retail health clinics popping up across the United States, this model could be used to help identify the optimal areas to establish those health clinics. Different data sources may be needed – such as existing hospitals, clinics, and insured coverage and demographics. Incorporating those sources into this model would prove useful in predicting where the next series of retail health locations should be placed.

The predictive model used for this paper can also be used in other industries. For example, the logistics industry can use this model to identify the best locations to build new distribution facilities to improve the supply chain and move goods from their raw state through production and customers. When companies launch new products, they need to create new or relocate existing logistics networks to deliver the goods

efficiently to the customer. Some of the data sources they can use in their predictive model could be the labor wages in MSAs that help identify if hiring new labor for these new distribution facilities will be difficult. Another data source that can be useful is transportation congestion because it helps determine if the new distribution facilities will be in areas with low or high traffic congestion. This is important because MSAs with high traffic congestion translate to longer wait times for customers to receive their goods from the distribution facilities.

The real estate industry can also use this predictive model to identify MSAs where people would relocate. This can help people who choose to invest in real estate by picking suitable properties in a particular MSA to maximize profits. Investors can use data sources such as property value, rental income, and occupancy rate data sources to help them become familiar with a neighborhood. For example, if investors knew they only could afford a small home in a particular MSA, the investors can use a visual map that highlights the MSA they should focus on.

5.1 Ethics

Ethics can evolve. When ethics evolve, the understanding of ethics should evolve too. For instance, during the COVID-19 pandemic, companies allowed their employees to work remotely. This enabled employees to assess their current living situation and identify new states to relocate such as California, Florida, and Texas. This change impacted employees and businesses such as Charles Schwab, which relocated its headquarters from California to the Dallas area. When looking to relocate a company to another state, there are several things to consider. One consideration when relocating business is legal concerns depending on the type of business and the laws for certain states, such as costs, employee impact, community impact, customer impact, and future growth capabilities. This paper aims to optimize the company investment and position a building supply business into single-family residential areas in the United States. By predicting where a company should invest its capital in certain MSAs primed for growth, there are potential ethical issues to consider, such as shareholder transparency and conflicts of interest in related party transactions.

Ethical issues may include construction companies that circumvent governmental regulations and start building without filing for specific permits. By doing so, construction companies can omit items such as leveraging products that do not conform with the project build specifications, potentially putting workers' (and the eventual buyer's) safety at risk. Transparency can also be challenging for leaders when they face adversity, but it is necessary to build an ethical culture. If a company is to move forward with the predictive modeling approach and invest in certain MSAs, it is important to be transparent about why those investments were made in those MSAs.

A conflict of interest occurs when private interests interfere, or even appear to interfere, with the company's interests. This applies to the objective of this paper since it can impact not only the company but employees creating the predictive model. A conflict can occur if the employees have interests that make it challenging to perform the company work objectively and effectively. The employee should conduct the company's business honestly and ethically, including handling actual, apparent and potential conflicts of interest between personal and business relationships. It is also a

conflict of interest for an employee to receive an improper personal benefit because of successfully predicting the company's capital investment.

Another potential ethical issue is having personal monetary interests in other businesses that benefit from a capital investment decision. Suppose the company decides to move forward with the predictive approach and invest in the identified MSAs. In that case, caution must be exercised that those close to decision-makers do not have a financial interest affected by those choices. Additionally, employees with privileged access to this model must refrain from disclosing the results to the competition. While the overall modeling approach is publicly available here, a company may modify the model and apply specific qualitative business rationale to the results. Disclosing those results without the authorization of the company is a violation of ethical standards.

6 Conclusion

This paper presents a framework to determine where a building supply company should invest its capital in maximizing material supply (and consequently maximize profit) using machine learning techniques. This model considers all the states within the United States. Data from the U.S. Census Bureau is used to predict the demand for SFHs in the future based on economic and demographic features. That data is transformed into a combined data set conducive to time series modeling. Algorithms such as MLP, XGBoost, KMeans, and Linear Regression are used to forecast the SFHs demand and find the optimal locations for a lumber yard and/or truss plant. The latitude and longitude were converted to human-readable names using geocoding and reverse geocoding. Results produced a significantly low RMSE for the MLP approach, which outperformed all other algorithms. Ensemble techniques were also attempted but did not outperform MLP. The results were validated using existing company MSA ranking data. Furthermore, the ML forecasting also highlighted numerous insights and potential for scalability that the existing methodology of the company simply is unable to decipher given the inherent limitations of ranking-based decision making.

The New York-Newark-Jersey City was ranked as the top MSA in total SFHs, followed by a large representation of MSAs within Texas. Factoring in current migration patterns within the United States, it is expected that Texas, which is experiencing a large influx of new residents, is represented strongly in the model's predicted results. After evaluating growth percentages of SFHs, New York falls from its top spot as would be expected due to large outflow of migration from that MSA. Replacing it is Knoxville, TN, in the top spot, and two Texas MSAs (El Paso and San Antonio).

Using these results, a building supply company has the basis for narrowing the list of potential MSAs to invest in using both the absolute total of SFHs and the projected SFH growth rates. Ruling out the MSAs where facilities already exist, a building supply company could operationalize this model quickly; leveraging the list of predicted MSAs, a layer of qualitative business-specific knowledge could be applied to further narrow down the list to a handful of solid candidates. Qualitative considerations such

as what the business knows about that MSA, competitors in the area, existing vendor and builder relationships, and additional insider knowledge would be needed to operationalize this model fully. Once done, however, the data-driven results would provide a reliable means to identify where to invest and give the company a competitive advantage compared to competitors still using manual ranking methods.

References

1. Linné, M., & Cirincionne, J. (2010). Integrating geographic information and valuation modeling for real estate. The Appraisal Journal, 78(4), 370-378.
2. Krylovas, A., Kosareva, N., & Laura Gudelytė. (2011). Construction of social indicators using information measuring principles. case study of real estate prices simulation model. Lietuvos Matematikos Rinkinys, 52 doi:10.15388/LMR.2011.mt03
3. Kontsevaya, N. V. (2016). Modeling real estate market: Forecasting the price of a square. Statistika i Èkonomika, (4), 31-34. doi:10.21686/2500-3925-2016-4-31-34
4. Fu, Y., Xiong, H., Ge, Y., Zheng, Y., Yao, Z., & Zhou, Z. (2016). Modeling of geographic dependencies for real estate ranking. ACM Transactions on Knowledge Discovery from Data (TKDD), 11(1), 1-27. doi:10.1145/2934692
5. McGibany, J. M. (1991). The Effect of Property Tax Rate Differentials on Single-Family Housing Starts in Wisconsin, 1978-1989. Journal of Regional Science, 31(3), 347. https://doi-org.proxy.libraries.smu.edu/10.1111/j.1467-9787.1991.tb00152.x
6. Andrade-Núñez, M. J., & Aide, T. M. (2020). The Socio-Economic and Environmental Variables Associated with Hotspots of Infrastructure Expansion in South America. Remote Sensing, 12(1), 116. https://doi-org.proxy.libraries.smu.edu/10.3390/rs12010116
7. Tong, M., Yan, Z., & Chao, L. (2020). Research on a Grey Prediction Model of Population Growth Based on a Logistic Approach. Discrete Dynamics in Nature & Society, 1–14. https://doi-org.proxy.libraries.smu.edu/10.1155/2020/2416840
8. Stephen Ezennia, I., & Hoskara, S. O. (2019). Methodological weaknesses in the measurement approaches and concept of housing affordability used in housing research: A qualitative study. PLoS ONE, 14(8), 1–27. https://doi-org.proxy.libraries.smu.edu/10.1371/journal.pone.0221246
9. Hua, G. B. (1996). Residential construction demand forecasting using economic indicators: a comparative study of artificial neural networks and multiple regression. Construction Management and Economics, 14(1), 25-34.
10. Case, K. E., Shiller, R. J., & Thompson, A. K. (2012). What have they been thinking? Homebuyer behavior in hot and cold markets. Brookings Papers on Economic Activity, 265+. https://link.gale.com/apps/doc/A327585906/AONE?u=txshracd2548&sid=AONE&xid=8344e923
11. Meyer, T. (2019). On the directional accuracy of united states housing starts forecasts: Evidence from survey data. Journal of Real Estate Finance and Economics, 58(3), 457-488. doi:http://dx.doi.org.proxy.libraries.smu.edu/10.1007/s11146-017-9637-9
12. Fullerton, T., Luevano, J., & West, C. (2000). Accuracy of Regional Single-Family Housing Start Forecasts. Journal of Housing Research, 11(1), 109-120. Retrieved January 31, 2021, from http://www.jstor.org/stable/24833749

13. Cohen, J.P., Coughlin, C.C. & Clapp, J.M. Local Polynomial Regressions versus OLS for Generating Location Value Estimates. J Real Estate Finan Econ 54, 365–385 (2017). https://doi.org/10.1007/s11146-016-9570-3
14. Ewing, Bradley & Wang, Yongsheng. (2005). Single housing starts and macroeconomic activity: An application of generalized impulse response analysis. Applied Economics Letters. 12. 187-190. 10.1080/1350485052000337806.
15. Burge, G., & Ihlanfeldt, K. (2006). Impact fees and single-family home construction. Journal of Urban Economics, 60(2), 284-306. doi:https://doi-org.proxy.libraries.smu.edu/10.1016/j.jue.2006.03.002
16. Author, S. (2019). Texas' economy: The 9 industries driving GDP growth. New York: Newstex. Retrieved from http://proxy.libraries.smu.edu/login?url=https://www-proquest-com.proxy.libraries.smu.edu/blogs,-podcasts,-websites/texas-economy-9-industries-driving-gdp-growth/docview/2251478678/se-2?accountid=6667
17. Gupta, R., & Miller, S. M. (2012). The time-series properties of house prices: A case study of the southern california market. Journal of Real Estate Finance and Economics, 44(3), 339-361. doi:http://dx.doi.org.proxy.libraries.smu.edu/10.1007/s11146-010-9234-7
18. Dua, P., Miller, S. M., & Smyth, D. J. (1999). Using leading indicators to forecast U.S. home sales in a bayesian vector autoregressive framework. Journal of Real Estate Finance and Economics, 18(2), 191-205. Retrieved from http://proxy.libraries.smu.edu/login?url=https://www-proquest-com.proxy.libraries.smu.edu/scholarly-journals/using-leading-indicators-forecast-u-s-home-sales/docview/203144143/se-2?accountid=6667
19. Gupta, R., Marco Lau, C. K., Plakandaras, V., & Wong, W. K. (2019). The role of housing sentiment in forecasting U.S. home sales growth: evidence from a Bayesian compressed vector autoregressive model. Economic research-Ekonomska istraživanja, 32(1), 2554-2567.
20. Bork, L., Møller, Stig V, Pedersen, Thomas Q. (2019). A New Index of Housing Sentiment. Management Science 66(4). https://doi-org.proxy.libraries.smu.edu/10.1287/mnsc.2018.3258
21. Gupta, R., Marfatia, H.A., Pierdzioch, C. et al. Machine Learning Predictions of Housing Market Synchronization across U.S. States: The Role of Uncertainty. J Real Estate Finan Econ (2021). https://doi-org.proxy.libraries.smu.edu/10.1007/s11146-020-09813-1
22. Murdoch, B. L. (2007). A study of housing policy and the economy (Order No. MR33545). Available from ProQuest Dissertations & Theses Global. (304844242). Retrieved from http://proxy.libraries.smu.edu/login?url=https://www-proquest-com.proxy.libraries.smu.edu/dissertations-theses/study-housing-policy-economy/docview/304844242/se-2?accountid=6667
23. Andersson, K. (2005). Housing investments and economic growth (Order No. 10774345). Available from ProQuest Dissertations & Theses Global. (2001909895). Retrieved from http://proxy.libraries.smu.edu/login?url=https://www-proquest-com.proxy.libraries.smu.edu/dissertations-theses/housing-investments-economic-growth/docview/2001909895/se-2?accountid=6667
24. Beckett, A. C. (2018). New-build housing, mobility and the life course : A study of housing-driven economic growth strategy in doncaster (Order No. 27776088).
25. Florida: An Economic Overview. (2020). Office of Economic & Demographic Research. http://edr.state.fl.us/Content/index.cfm
26. Gupta, R., Gupta, R., Miller, S., & Miller, S. (2012). The Time-Series Properties of House Prices: A Case Study of the Southern California Market. The Journal of Real

Estate Finance and Economics, 44(3), 339–361. https://doi.org/10.1007/s11146-010-9234-7

27. Seman, M., & Carroll, M. (2017). The Creative Economies of Texas Metropolitan Regions: A Comparative Analysis Before, During, and After the Recession. Growth and Change, 48(4), 831–852. https://doi.org/10.1111/grow.12198
28. Mao, Y., Yao, N., & Zhang, M. (2014). Hangzhou Housing Demand Forecasting Model Based on B.P. Neural Network of Genetic Algorithm Optimization. Applied Mechanics and Materials, 587-589, 37–41. https://doi.org/10.4028/www.scientific.net/AMM.587-589.37