Minneapolis Housing Sale Study

Predictions based on Single-Family Dwellings

Shereen Hunitie
SEIS 763
University of St. Thomas
St. Paul, MN
shereen.hunitie@stthomas.edu

Jon Sandquist
SEIS 763
University of St. Thomas
St. Paul, MN
sand8234@stthomas.edu

Youa Vang
SEIS 763
University of St. Thomas
St. Paul, MN
youa.vang@stthomas.edu

Shiao Wang
SEIS 763
University of St. Thomas
St. Paul, MN
shiao.wang@stthomas.edu

ABSTRACT

Determining the value of a home, and the correct listing price is essential in real estate processes. With the use of machine learning methods, this study was conducted on the Minneapolis single-family home housing market. Regression models were built to accurately find the sale price of single-family homes sold in 2019. The champion regression model in this study was a tie between the lasso and elastic net regression models. Both models predict the sale price with an accuracy of 83.55%. Classification models were built with the goal of determining if a homeowner should sell their home or not. This model was built to only consider selling if the sales price is higher than the assessed value plus closing costs. The champion classification model in this study is the random forest classifier using GridSearch. This model predicted if a home should be sold, or if the sale price is higher than the assessed value plus closing costs, with an accuracy of 62.19%.

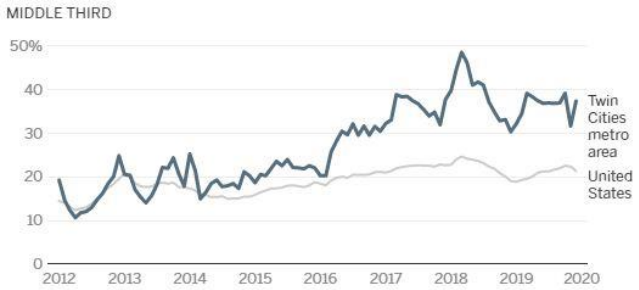
KEYWORDS

Machine Learning, Regression, Classification

1 Introduction

Intrigued by the competitive landscape of the housing market in the Minneapolis area, the goal of this study is to build a model that could accurately predict sales prices of a single-family home. With accurate predictions of the sales prices, a model could assist homeowners in deciding what to list their homes for. Likewise, a model could assist homebuyers in assessing a fair offer and could even be used as leverage in home offer discussions.

The competitive nature of the housing market in Minneapolis is discussed in a Star Tribune article[1] from March of 2020. The below graph shows the percentage of homes that are sold above listing price for the middle third range of home listings in the Twin Cities metro area, as compared to all of the United States. The percentage sold above listing price is indicated on the y-axis. At the highest peak in 2018, about 50% of homes sold in the Twin Cities metro area were sold over the listing price. During that same time, the United States saw around 25% of homes selling over the listing price, so we know that the Twin Cities metro area was far above the US average [1].



Throughout this study, relevant features for determining housing sales in Minneapolis are assessed.

1.1 Methods

There are two methods of machine learning conducted in this study: Regression, and Classification.

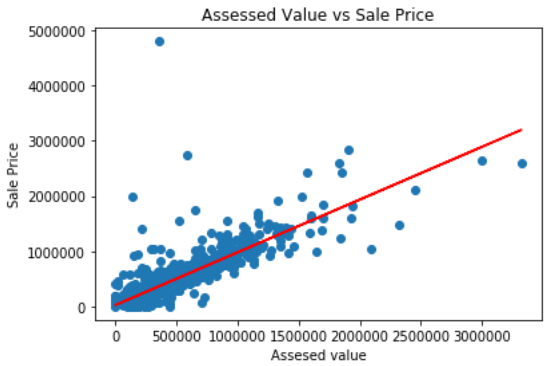
In the regression analysis, the following models were built and evaluated: linear regression, multiple regression, polynomial regression, lasso regression, ridge regression, decision tree regression, random forest regression, PCA regression, and K-PCA regression. The goal of the regression analysis and models created was to use home assessment data to try and predict the sales price of a single-family home in Minneapolis. As described above, the intent was to create a tool for homeowners to use when setting a sale price, as well as a tool for homebuyers to know if a listing is reasonable. The champion model created will be compared to other home sales price predictors, such as Zillow.com. Zillow claims that their model predicts sales prices to be within 10% of actual sales on 97% of homes listed in the Minneapolis/St. Paul region [2].

In the classification analysis conducted, the following models were built and evaluated: logistic regression, SVC kernel, KNeighbors, Naïve Bayes, decision tree, random forest, PCA, K-PCA, and LDA classification models. In the classification analysis, the same dataset will be used to determine if it is in a homeowner's best interest financially to sell their home. It is understood that there are many factors when considering whether to sell a home from the homeowner's perspective. However, this study will only consider the decision from a purely financial standpoint. If a home's sale price is higher than the assessed value of the home this is classified as yes, the homeowner should sell. If a home has a higher assessed

value than sales value this is classified as no, the homeowner should not sell. To cover the additional costs of selling a home, such as closing costs, 3% was added on to the assessed value attribute. Considering data from the March Star Tribune article, it is assumed that homes in the Twin Cities sold for over asking are making a profit. The champion model created will be validated and given an accuracy rate with 2020 home sale information from Zillow.

2 Data Exploration

To begin the process of analyzing the dataset, the relationship between the assessed price and the sales price of homes was analyzed. It can be assumed that there is likely a relationship between the assessment price and the sales price of a home. For homebuyers, a bank will not lend more money than the house appraises for, so a fairly strong relationship between the appraisal price and sales price was anticipated.



Appraisal values can indicate to the seller how much they should receive in an offer. Any sort of bias analysis that may come from the influence an appraisal price has on a selling price is outside the scope of this study. We do not have access to the tools of an appraiser, nor do we have the years of experience and other data that enable an appraiser to set a value on a home. Our analysis is done on appraisal data that is publicly available through the city of Minneapolis.

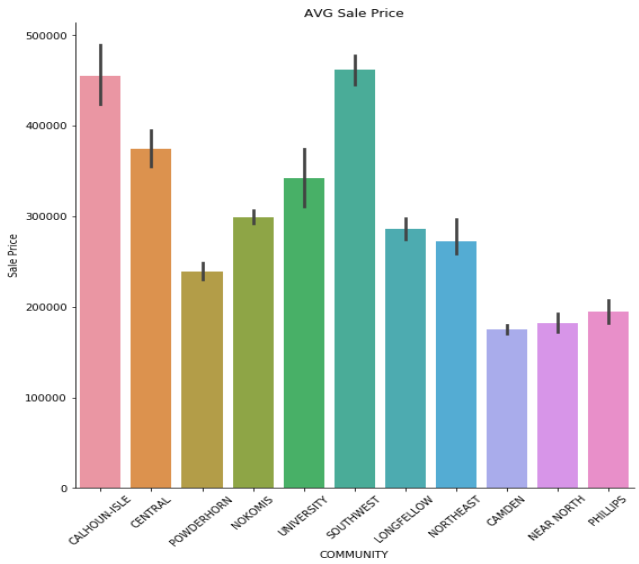
2.1 Basic Statistics

Several calculations were done to have a good understanding of the landscape of Minneapolis home sales in 2019. It was interesting to find that the lower quartile for homes sold in 2019 had a construction date of 1916 and the upper quartile had a date of 1956. This is not too surprising because Minneapolis is largely developed and there are only a few areas that have been cleared out for new homes. New construction of homes in Minneapolis is likely from tearing down older homes and rebuilding a new home in its place. This was revealed in the following chart with the YEARBUILT column for the lower quartile, labeled 25%, and for the upper quartile, labeled 75%.

	ABOVEGROUNDAREA	BELOWGROUNDAREA	PARCEL_AREA_SQFT	YEARBUILT	BATHROOMS	BEDROOMS	# STORIES	FIREPLACES	Assessed Value	Sale Price
mean	1,271	659	18,498	1,940	1.7	2.7	1.4	0.4	299,602	312,518
std	515	429	38,913	32	0.8	1.0	0.5	0.6	206,688	224,766
min	343	-	486	1,874	-	-	-	-	-	1,200
25%	956	375	5,100	1,916	1.0	2.0	1.0	-	179,000	189,858
50%	1,168	768	5,670	1,927	2.0	3.0	1.2	-	252,000	262,796
75%	1,476	936	8,581	1,956	2.0	3.0	1.7	1.0	348,500	360,371
max	6,972	3,399	315,895	2,018	8.0	8.0	4.0	5.0	3,320,000	4,800,000

Another interesting finding was that the mean selling price on a home in Minneapolis in 2019 was $312,518. However, this dataset has a large standard deviation: $224,756. This high standard deviation in sales prices will prompt us to look at potential outliers in the dataset and is something to consider when building models. Additionally, the upper and lower quartiles for the Sales Price column are $189,858 and $360,37 respectively. This is not a large difference in selling price considering the standard deviation is $224,756, again indicating there may be outliers in this dataset. The minimum and maximum selling prices are $1,200 and $4.8M respectively. This difference would certainly contribute to the high standard deviation. The maximum sale price for 2019 does seem to be a realistic maximum value. The minimum sale price, however, does not seem realistic and may be worth investigating. Perhaps, this property is considered a shed or a barn and is simply zoned incorrectly, or was sold or gifted to a family member. Here is a side by side view of the highest sale (right image) [4] and the lowest sale (left image) [3] in 2019.

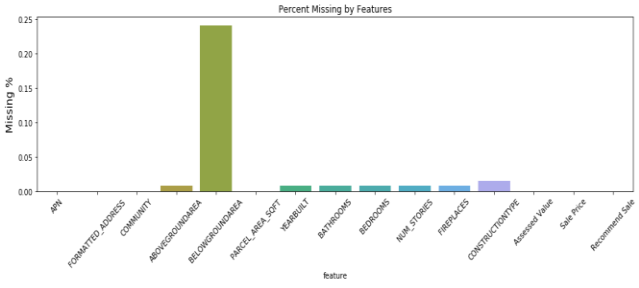
Another area of interest was the average home selling price in the different communities within Minneapolis. The average sales prices of homes in the 11 communities or neighborhoods in the dataset do differ significantly. It was found that the neighborhoods Camden, Near North, and Phillips have homes with the lowest average sales prices, and neighborhoods Calhoun-Isle and Southwest had the highest average sales prices. The graph below illustrates these findings and includes the confidence intervals in black associated with each community.



2.2 Data Preparation

The city of Minneapolis has made housing sale records and appraisal data available on the website opendata.minneapolismn.gov. After exploring the property sales data, the scope was narrowed to only single-family home properties that were sold in Minneapolis in 2019. This resulted in a list of 5,768 properties. An additional file was extracted that itemized assessment values from the year 2019. This file was extracted from the same source, opendata.minneapolismn.gov. By combining sales records and assessment records, all attributes desired for regression and classification analyses were obtained. The two data files were then merged to create the master data file. Because the files were from the same source, all 5,768 single-family homes sold in 2019 were available for analysis and included in the master data file.

After the merge was completed, additional data exploration was conducted in preparation for creating models. The graph below was created for assessing missing values in the dataset.



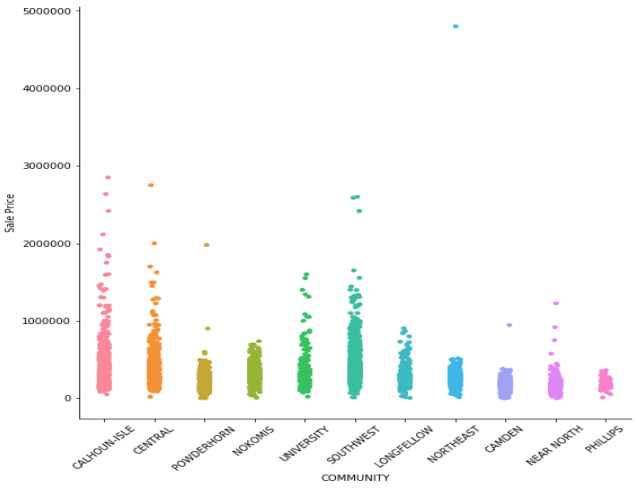
The attribute 'Below-Ground Square Footage' was identified as problematic because of the high volume of missing values: 1,390 properties. There was no explanation in the data or any other published text that accompanied the data. It would be reasonable to assume if the below-ground area value is blank, this could mean there is either no basement in the house or the house does not have what would constitute a finished basement. Because this information could mean no basement, missing values were replaced with the value zero. The remaining numerical attributes: above ground square footage, parcel area square footage, bathrooms, bedrooms, and number of stories, were then imputed using mean imputation. The variable for year was not included in mean imputation, because the variable was not missing any values.

Community and Construction Type fields are both categorical fields. Community consisted of 11 unique values with no blank values. Construction type consisted of 7 unique values and had 47 rows of blank data. When looking into the 47 missing values, it was found that many other attributes were missing values in the same rows, these rows were removed from the dataset. After listwise deletion was complete, 5,721 properties were left for analysis.

After backwards elimination was completed it was found that the variables Community, AboveGroundArea, ParcelArea, Bathrooms, Num_Stories, ConstrutionType, FirePlaces, and Assessed Price were all statistically significant. The variables that were not statistically significant are Bedrooms, BelowGroundArea, and YearBuilt. Any values eliminated for the one-hot encoding of the categorical variables for community and construction type were added back into the model. It was hypothesized that the assessed value was going to be a strong factor, which was indeed significant for the model. Additionally, after looking at the average sales price for each community, it was correctly hypothesized that community would also be a significant factor.

2.3 Outliers

As mentioned, there was concern over the standard deviation being so large, it was important to acknowledge the potential impact of outliers. The chart below exhibits each home sales price by neighborhood. There is a visible clumping of homes priced within the sub $1M sales price, with a few communities representing a handful of homes between the $1M and $3M range. There is a single property in Northeast that is feeding the larger standard deviation, at a sales price of $4.8M.



The property in question is on a quarter-acre plot with a home constructed in 1900. The assessed price on the property was $362,200, which lead us to believe some additional variables were not present in the dataset that may have inflated this sales price. After further research on Google and Zillow, it was found that this was likely a typo in the dataset, as the home sale price was $480,000 and not $4.8M [4].

2.4 Data Assumptions

To model the sales price of single-family homes in Minneapolis, as well as whether a homeowner should or should not sell their home, assumptions need to be made. These are very complex models and it is not possible to successfully consider every factor. Therefore, many potentially relevant factors were not used in the creation of this study's models. For example, nearby school ratings, walkability, highways, nearby shopping locations, public transportation, nearby gyms, and many other factors could have been included but were not. It was assumed that the variables used in these models were the most relevant for the models created.

In the classification model created, it was assumed that the homeowner's decision whether to sell or not sell is solely based on potential profit from the sale. There are certainly many other factors that go into deciding whether to sell a home or not. It was also assumed that the sales price would carry a 3% fee with other closing costs associated with the sale of the home. For this reason, the assessment price on the classification model was increased by 3%.

As mentioned, for the value below-ground area, it was also assumed if the value was blank, this indicates that there is either no basement in the house or the house does not have what would constitute a finished basement.

3 Regression

Eight models were created for regression analysis. Five of the eight models returned a model accuracy of 83.55%. This accuracy score was the highest. The tie-breaker between these top five models came down to the ten-thousandth decimal place. The Lasso and Elastic Net regression models are the champion regression models and had the same model accuracy, narrowly beating the multiple regression, polynomial regression, and PCA regression models.

MultipleRegression: 0.8355390116130667

Poly degree (1): 0.8355390116130667

PCA comp=None: 0.8355390116130667

Lasso: 0.8355412101431431

ElasticNet: 0.8355412101431431

See model accuracies and items of interest for each model used in regression analysis.

3.1 **Multiple Regression**

Model Accuracy: 83.55%

3.2 **Principal Component Analysis (PCA)**

Model Accuracy:

Component 2 = 69.82%

Component 1 = 36.01%

Component None = 83.55%

PCA regression model returned the highest accuracy with components equal to none.

3.3 **Kernel Principal Component Analysis (K-PCA)**

Model Accuracy:

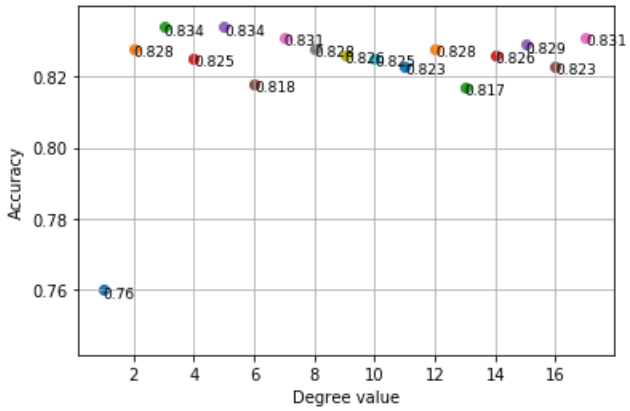
Component 2 = 23.24%

Component 1 = 19.52%

Component None = -160641595031%

3.4 **Random Forest Regressor**

The following plot shows that the Random Forest Regressor model performed best with 3 or 5 trees.



Model Accuracy: 83.41%

3.5 **Polynomial Regression**

Model Accuracy: 83.55% (degree= 1)

Degree=0: -0.05%

Degree=1: 83.55%

Degree=2: -988.89%

3.6 **Ridge Regression**

Model Accuracy: 83.54%

3.7 **Lasso Regression**

Model Accuracy: 83.55%

3.8 **Elastic Net Regression**

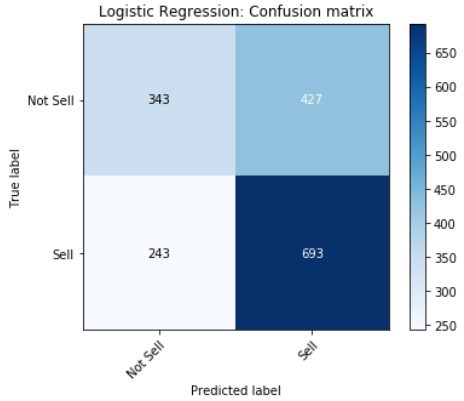
Model Accuracy: 83.55%

4 Classification

Nine models were created for classification analysis before the GridSearch method was implemented. The champion model without GridSearch is the KPCA model where components are set to none. The model has an accuracy of 62.13%. The random forest classifier model is the second most accurate with an accuracy of 61.84%. See model accuracies and items of interest for each model used in classification analysis.

4.1 **Logistic Regression**

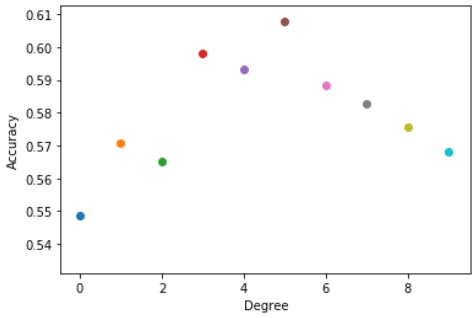
The following is a confusion matrix for the logistic regression model.



Model Accuracy: 60.73%

4.2 **Support Vector Classification (SVC)**

The following plot shows that the SVC model with kernel set to poly performed the best when the degree was set to five.



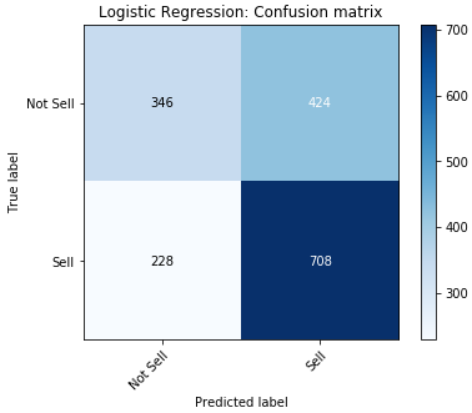
The SVC model performed best when kernel was set to rbf.

Model Accuracy of 61.78 (kernel = rbf)

Model Accuracy: 59.55% (kernel = poly, degree = 5)

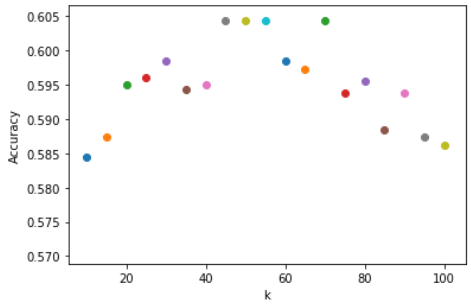
Model Accuracy: 57.79% (kernel = linear)

The following is a confusion matrix for the SVC model with kernel set to rbf.



4.3 **KNeighbor**

The following plot shows that the KNeighbor model performed the best when the hyperparameter n_neighbor was set to 45, 50, 55or 70.



Model Accuracy: 60.43% (n_neighbor = 45)

4.4 **Naïve Bayes**

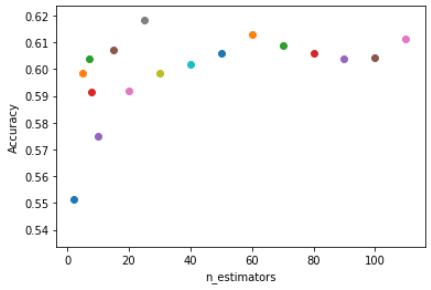
Model Accuracy: 56.92%

4.5 **Decision Tree**

Model Accuracy: 56.92%

4.6 **Random Forest Classifier**

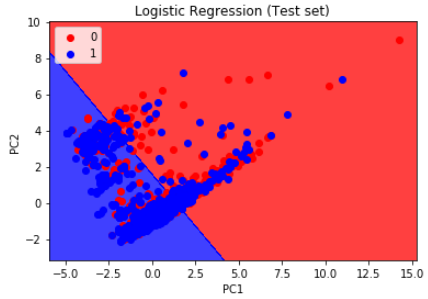
The following plot shows that the random forest classifier model performed the best with n_estimators set to 25.



Model Accuracy: 61.84% (25 n_estimators)

4.7 **Principal Component Analysis (PCA)**

The PCA model performed best with components set equal to none. The following visualization is the PCA model with components set equal to two.



Model Accuracy: 57.68% (components = 2)

Model Accuracy: 58.09% (components = 1)

Model Accuracy: 60.73% (components = none)

4.8 **Linear Discriminant Analysis (LDA)**

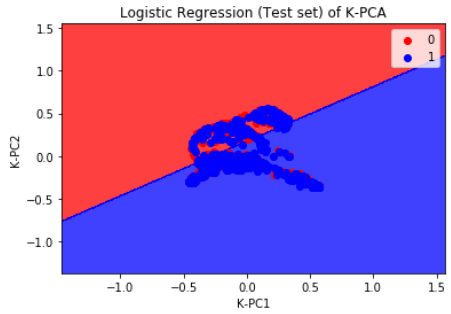
Like the PCA model, the LDA model performed best with components set equal to none.

Model Accuracy: 60.08% (components = 1)

Model Accuracy: 60.08% (components = none)

4.9 **K-Principal Component Analysis (K-PCA)**

Like the PCA and LDA models, the K-PCA model performed best with components set equal to none. The following visualization is the K-PCA model with components set equal to two.



Model Accuracy: 56.09% (components = 2)

Model Accuracy: 54.98% (components = 1)

Model Accuracy: 62.13% (components = none)

4.1 Classification with GridSearch

The performance of classification models with GridSearch returned a mixed result. The decision tree and random forest classifier models performed better with GridSearch, whereas the SVC and KNeighbor models did not. The champion classification model using GridSearch is the random forest classifier model. This model is also the champion classification model overall. See model accuracies and items of interest for each model used in classification analysis with GridSearch.

4.1.1 **Decision Tree Classifier**

Without the use of GridSearch, the accuracy of the decision tree classifier model is much lower. The accuracy increases from 56.92% to 60.73%.

GridSeach enabled the use of parameters 'Max Depth' and 'Criterion.' The decision tree classifier model returned the best results with 'entropy' as the criterion, and with Max Depth set to 6.

Model Accuracy: 60.73% (criterion = entropy, max depth = 6)

4.1.2 **Random Forest Classifier**

Without the use of GridSearch, the accuracy of the decision tree classifier model is lower. The accuracy increases from 61.84% to 62.19%.

With GridSearch, the random forest classifier model performed best with the criterion set to entropy, n_estimator set to 190, and bootstrap set to True. The following is a visualization of the random forest classifier model results with the parameters discussed.

Model Accuracy: 62.19% (criterion = entropy, n_estimator=190, bootstrap = True)

4.1.3 **SVC**

Without the use of GridSearch, there was a higher accuracy for SVC with kernel set to rbf, 61.78%.

Model Accuracy: 60.48% (degree = 2, kernel = rbf)

4.1.4 **KNeighbor**

Without the use of GridSearch, there was a higher accuracy with the KNeighbor model, 60.43%.

Model Accuracy: 58.25% (algorithm = auto, n_neighbor = 3)

5 MODEL vs. ZILLOW

As mentioned, Zillow claims that 97% of homes sold are within 10% of their Zillow score for the Twin Cities area. All of the Twin Cities housing market would be outside the scope of this study since only the Minneapolis single-family housing market was analyzed. However, for the purpose of testing models from this study, houses in the Nokomis and Calhoun neighborhoods were used to compare models. The homes selected were all sold in the year 2020. The models created in this study were built from 2019 assessment data.

There were two champion regression models, elastic net and lasso. Each model had the same accuracy score, so elastic net was chosen for analysis. The champion classification model, random forest

classifier with GridSearch, as well as the elastic net regression model were both applied to the following 2020 home sales.

Home 1: 5508 29th Ave S (Nokomis) [5]

Home 2: 5200 39th Ave S (Nokomis) [6]

Home 3: 5833 11th Ave S (Nokomis) [7]

Home 4: 3333 Irving Ave S (Calhoun) [8]

The elastic net regression model was then measured against the Zillow results to test accuracy compared with Zillow's model. The elastic net regression model from this study created regression predictions within less than 10% of the actual sales price on three of the four homes randomly tested. Zillow was within 1% of the actual sales price on three of the four homes tested, performing much better than the model from this study. See the results from the comparison in the following chart.

FORMATTED_ADDRESS	ASSESSED VAL	SALES PRICE	ZILLOW EST.	OUR MODEL	DIFFERENCE	% OUR DIFF	% ZILLOW DIFF
5508 29th Ave	265,000.00	330,000.00	333,326.00	298,388.00	(31,612.00)	-9.58%	1.01%
5200 39th ave S	269,000.00	315,000.00	315,522.00	267,522.00	(47,478.00)	-15.07%	0.17%
5833 11th Ave S	426,500.00	430,000.00	430,605.00	467,916.00	37,916.00	8.82%	0.14%
3333 Irving Ave S	611,500.00	625,000.00	631,150.00	591,674.00	(33,326.00)	-5.33%	0.98%

The classification champion model, random forest classifier with GridSearch, was then tested against the same homes listed above. The results were not as impressive as the results from the regression model. The random forest classifier model performed with a 50% accuracy. All four homes were predicted to 'Not Sell,' which means the predicted assessed value - plus the 3% closing cost - is less than the sales price. Actual results should have been 'Not Sell' for two of the homes, not four of the homes, indicating a 50% model accuracy. This is, however, a very small sample size of four homes. It would be interesting to test a larger random sample. See the results from the comparison in the following chart.

FORMATTED_ADDRESS	ASSESSED VAL	ASSESSED VAL + FEE	SALES PRICE	DIFFERENCE	ACTUAL	OUR MODEL
5508 29th Ave	265,000.00	272,950.00	330,000.00	57,050.00	Sell	Not Sell
5200 39th ave S	269,000.00	277,070.00	315,000.00	37,930.00	Sell	Not Sell
5833 11th Ave S	426,500.00	439,295.00	430,000.00	(9,295.00)	Not Sell	Not Sell
3333 Irving Ave S	611,500.00	629,845.00	625,000.00	(4,845.00)	Not Sell	Not Sell

REFERENCES

[1] Jim Buchta, 2020. How competitive is the Twin Cities housing market? Only 3 in U.S. are tougher' Star Tribune (March 2, 2020). https://www.startribune.com/how-competitive-is-the-twin-cities-housing-market-only-3-in-u-s-are-tougher/568401182.

[2] Zillow.com, website, https://www.zillow.com/zestimate/

[3] Google Maps. 2020. 2210 N James Ave. Retrieved from: https://www.google.com/maps/place/2210+N+James+Ave,+Minneapolis,+MN+55411/@45.0017601,-93.3042771,3a,75y,79.91h,90t/data=!3m6!1e1!3m4!1sON-7II81JAWb8mDYkB-smQ!2e0!7i16384!8i8192!4m5!3m4!1s0x52b3324cd182170d:0x2937f8f1b9a3e805!8m2!3d45.0017837!4d-93.3039609

[4] Google Maps. 2020. 1024 NE Main St. Retrieved from: https://www.google.com/maps/place/1024+NE+Main+St,+Minneapolis,+MN+55413/@44.9983333,-93.2671148,3a,75y,262.47h,90t/data=!3m6!1e1!3m4!1sF_b8SCuggEoNpxrZZjQEUQ!2e0!7i16384!8i8192!4m5!3m4!1s0x52b3327997788293:0xd50306875c009e87!8m2!3d44.998308!4d-93.2673901

[5] Zillow.com, website, https://www.zillow.com/homedetails/5508-29th-Ave-S-Minneapolis-MN-55417/1919955_zpid/

[6] Zillow.com, website, https://www.zillow.com/homedetails/5200-39th-Ave-S-Minneapolis-MN-55417/1861715_zpid/

[7] Zillow.com, website, https://www.zillow.com/homedetails/5833-11th-Ave-S-Minneapolis-MN-55417/1912042_zpid/

[8] Zillow.com, website, https://www.zillow.com/homedetails/3333-Irving-Ave-S-Minneapolis-MN-55408/1717970_zpid//