Analyzing houses in Saratoga, New York

Sanskriti Chandak

Babson College

Abstract

The housing crisis in the United States has impacted the housing market in numerous ways. In tandem with the pandemic, Upstate New York is experiencing a surge in prices and an increase in demand for housing. However, the $200 million legislation aims to bring market conditions back to pre-COVID times (Bruno, 2022). Therefore, the aim of this study is to predict the prices and probability of having central air for houses in Saratoga, New York. The data is from 2015 and includes information about 16 home features ("StatCrunch," n.d.). The models created predict how different home features affect the prices of houses and their probability of having central air. The three models created include linear regression, logistic regression, and KNN regression. Since the data set is relatively small, models were created with and without outliers to find the best possible model. The findings indicate that the linear regression model without outliers, logistic regression model with outliers, and KNN regression model without outliers can be used to predict prices and/or the probability of having central air for houses in Saratoga and are better than not using any model.

Keywords: Linear Regression, Logistic Regression, KNN Regression, RMSE, MAPE, sensitivity, specificity, p-value, outliers, error rate, benchmark

Housing Crisis

The United States is currently in the midst of a housing crisis. The country is facing a nationwide affordable housing crisis. Harvard researchers found that nearly half of renters are cost-burdened (Sisson, 2020). Home prices are rising at twice the rate of wage growth (Sisson, 2020). This issue is highly relevant because the United States has a long history of redlining, segregation, and racist housing policies. Regardless of party affiliation, Americans living in urban areas are more likely to see affordable housing availability locally as a major problem (Schaeffer, 2022). The top 5 home features buyers are looking for are price, air conditioning, number of rooms, number of bathrooms, and living area ("6 critical things," 2022). 82% of buyers cited budget as very or extremely important ("6 critical things," 2022). 79% said air condition was the second most important, and similarly, 77% said the number of rooms, 72% chose the number of bathrooms, and 69% stated that they had a preferred size or square footage ("6 critical things," 2022).

Upstate New York is one of the major regions experiencing an affordable housing crisis. The Capital Region has had the highest population growth rate in New York (Media, 2022). In recent years, more people are moving away from New York City which is increasing the demand for housing in Upstate New York. The issue has worsened due to the COVID-19 Pandemic and because of wealthy investors who are buying second homes (Media, 2022). Almost every county in upstate New York has shown an increase in median home prices. For example, the Hudson Valley region saw an increase of $49,000 (Doar, 2021). In Dutchess County, there are zero listings for any of the identified subsidized housing in any of the three towns (Doar, 2021). The consequences of the affordable housing shortage and the resulting lack of labor has the potential to be alarming . Local fire departments and EMTS are required to live within a certain radius of the towns they serve (Doar, 2021). However, as affordable housing becomes less accessible, this becomes increasingly concerning.

Lawmakers and local government leaders are calling for state funding to make living in upstate neighborhoods more affordable. Given that upstate communities are becoming more popular as a result of the pandemic, it has had a subsequent negative impact on the housing market. Thus, a $200 million budget has been proposed by the senator to increase the supply of housing in these

areas for purchase and for rent (Bruno, 2022). The legislation aims to construct new houses by partnering with non-governmental organizations that aim to build new affordable housing or restore the existing housing in a way that makes it affordable to low-income and at-risk communities (Bruno, 2022).

Saratoga has a population of 28,491 ("US Census," 2021). 88.3% of the community is White and 4.2% of the community is black/African American ("US Census," 2021). 52.3% of residents are female and 47.7% are male ("US Census," 2021). The average number of persons per household is 2.03 ("US Census," 2021). The median household income is $82,816 and median property values are $365,900 and the homeownership rate is 55.7% ("US Census," 2021). Although this data is from 2015, and COVID has increased the demand and prices for housing in Saratoga, the proposed legislation is aimed at helping bring the prices back to pre-COVID times. Therefore, the data set is relevant in predicting future house prices in Saratoga.

Data Management

The data set consists of data related to houses in Saratoga, New York from 2015. It consists of the following information: The selling price of the house in USD, the lot size in acres, whether or not it has a waterfront (0 = No, 1 = Yes), age in years, land value in USD as the assessed value of the property without the structures, whether or not it is newly constructed (0 = No, 1 = Yes), whether or not it has a central air system (0 = No, 1 = Yes), the type of fuel used (1 = Gas, 2 = Elective, 2 = Oil), the heat type used (1 = Forced Hot Air, 2 = Hot Air, 3 = Electric), sewer type (1 = None/Unknown, 2 = Private (Septic System), 3 = Commercial/Public Living), living area in square feet, Pct College which is the percentage of residents of the given zip code that attended a four-year college, the number of half-baths, the number of bedrooms, the number of fireplaces, and the number of rooms ("StatCrunch," n.d).

The raw data introduced above consists of 1,728 points. The first step involved in cleaning the data is checking for duplicates and missing values. Since there are no duplicates or missing values found in the data set, the next step in cleaning the data is removing the outliers. There are two methods to remove outliers – the box-plot method and the z-score method. The z-score method measures the deviation of data points from the mean. Using the z-score method in RStudio, the data points furthest away from the mean are removed. The removeOutliers function is run only once, since running it repeatedly changes the range that is considered when removing outliers.

Below are the boxplots and histograms for the data set with and without the outliers. As observed, after removing the outliers, the data is less skewed and the histogram has a more normal distribution.

Figure 1

Boxplots of all the variables from the data set with outliers

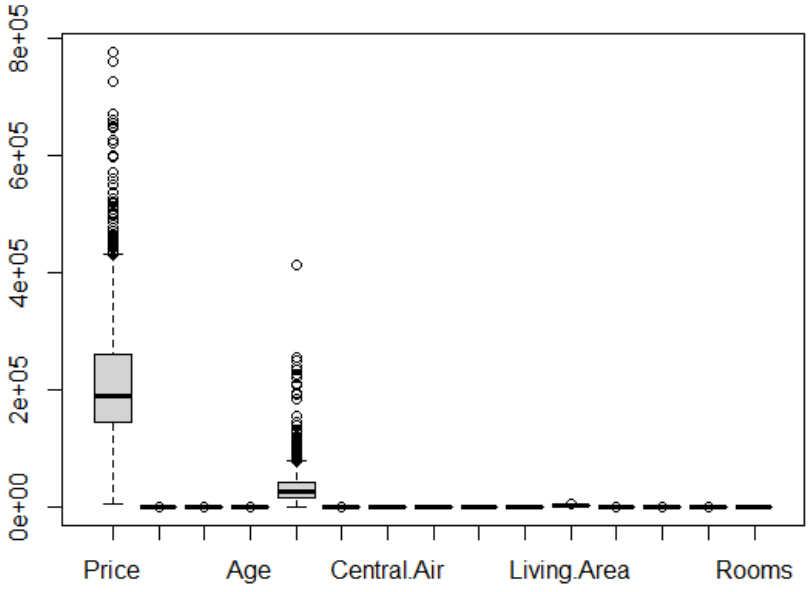


Figure 2

Boxplots of all the variables from the data set without outliers

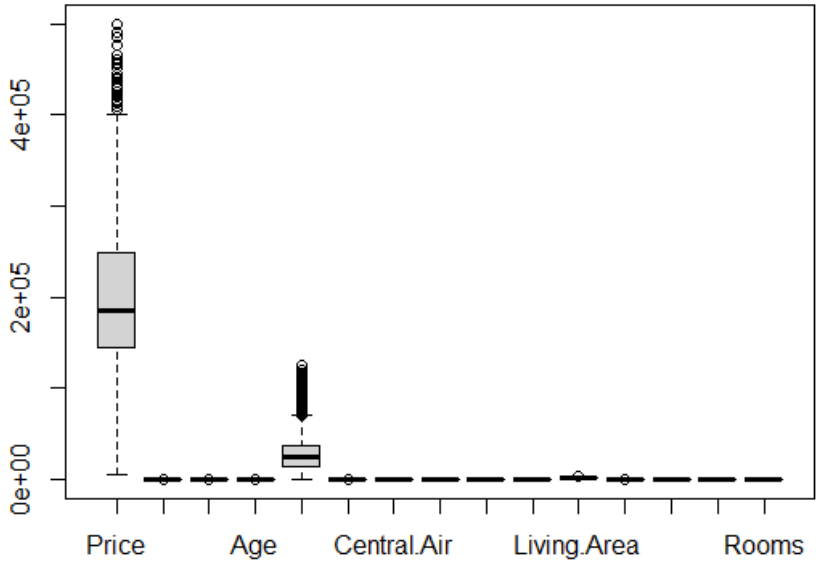


Figure 3

Histogram of the price variable from the data set with outliers

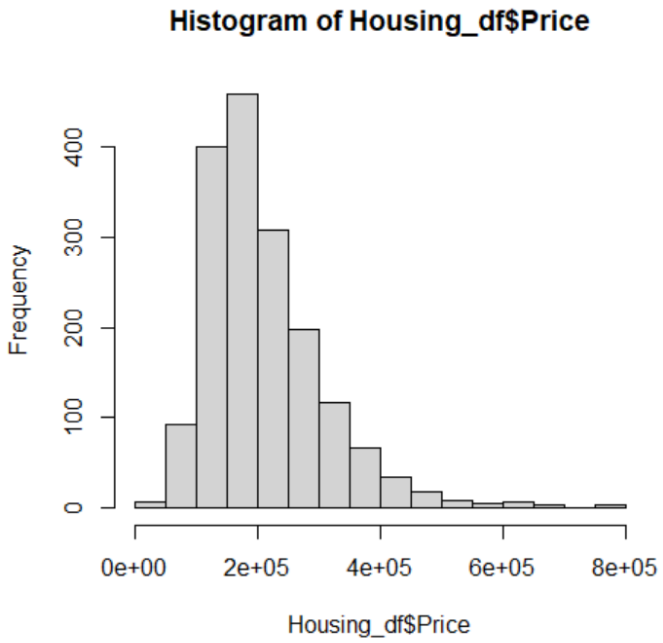
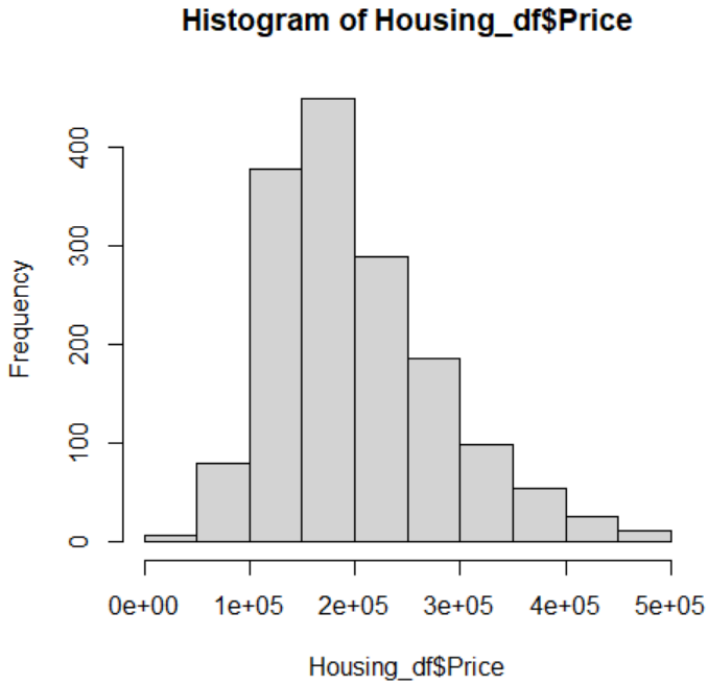


Figure 4

Histogram of the price variable from the data set without outliers



In addition to removing the outliers, the classification of certain variables was changed, too. Fuel Type and Heat Type have 3-factor levels, each. The factor levels were labeled as 2, 3, and 4, respectively, i.e., none of the variables has a Fuel Type of 1 or a Heat Type of type 1. In order to improve the comprehension and visual representation of the data, the order, and label of the factor levels were changed from 2, 3, and 4, to 1, 2, and 3, respectively, for both the Fuel Type and Heat Type variables.

Moreover, the Pct variable, which indicates the percentage of people who have attended a 4-year college in a certain zip code area, was removed. This variable was removed because the zip codes of each area are missing; thus, it is not possible to connect the percentages to their respective zip code. Secondly, since the zip codes are missing, and hypothetically more than one area can have the same value, it is not possible to manually classify the values. In addition, Pct is not one of the essential home features or variables in the data set. Therefore, it was removed from the data before creating the models.

Finally, each model uses a 70:30 data split. This implies that 70% of the data is used in the training set and 30% of the data is used in the test set. Given that the data set is relatively small, this specific data split helps avoid overfitting and uses more data in the training set to make more accurate predictions.

Model 1 – Linear Regression

Research Question: How do the home features in Saratoga, NY affect the price of houses?

The linear regression model aims to investigate the relationship between house prices in Saratoga and its home features. The goal is to predict the house prices based on the lot size, whether it is located on the waterfront, the age, the land value, whether it is newly constructed, whether it is has central air, the fuel type, heat type, sewer type, the living area, the number of bathrooms, bedrooms, fireplaces, and rooms.

Model without Outliers

Data Management

A linear regression model requires a numerical response variable. The predictor variables that were misclassified by RStudio were converted into their correct variable types. This includes New.Construction, Waterfront, and Central.Air which were converted to logical variables. In addition, Fuel.Type, Heat.Type, and Sewer.Type were converted into factor variables. Following which, the outliers were removed using the removeOutliers command based on the z-score method.

Partitioning the Data

A seed with code 1234 is set to standardize the model results. The data is partitioned using the 70:30 split to create the training and test data set.

Build the Model

The linear regression model is built using R with Price as the response variable and the other variables as the predictor variables. The output is as follows:

Price = -3234.31 + 11399.79 Lot.Size + 140341.90 WaterfrontTRUE -270.75 Age + 0.94 Land.Value -28437.08 New.ConstructTRUE + 9416.59 Central.AirTRUE -6369.78 Fuel.Type2 - 3935.75 Fuel.Type3 -10239.52 Heat.Type2 - 1097.07 Heat.Type3 + 28328.31 Sewer.Type2 + 24207.35 Sewer.Type3 + 64.23 Living.Area - 22580.23 Bathrooms + 1367.02 Rooms

The first step to analyzing the model involves evaluating the statistical significance of the overall model. The p-value of this overall model is $p < 2.2e-16$, which indicates that this model is statistically significant. Therefore, at a significance level of alpha = 0.05, there is sufficient evidence to conclude that a linear relationship exists between Price and all the predictor variables. Next, the statistical significance of all the individual predictor variables was analyzed. Only the variables WaterfrontTRUE, Age, Land.Value, New.ConstructTRUE, Central.AirTRUE, Heat.Type2, Living.Area, and Bathrooms have at least one star at the end of the row, which indicates that these variables have p-values that are greater than 0.05 and are not statistically significant. The fit of this model is then assessed using the coefficient of multiple determination (R^2) and the standard error. The R^2 of this model is 0.6364, which means that 63.64% of the observed variation of price can be explained by the variation in the predictor variables. The standard error is 49970, which implies that the predictions of prices are within $49970 of its actual value. The low R^2 and high standard error indicate that the model does not make reliable predictions.

Since there are variables that are not statistically significant, and to avoid potential overfitting, the model is improved using the stepwise regression methods. All three stepwise methods – forward, backward, and both – are performed and analyzed to obtain the best final model. All three stepwise regression methods generated the same output. Thus, the forward stepwise regression model is used. The forward stepwise method starts with a model consisting only of the intercept and then adds one variable at a time until the model stops improving. The equation generated using this model is:

Price = 21888.30 + 66.06 Living.Area + 0.93 Land.Value +139148.45 WaterfrontTRUE + 22049.36 Bathrooms + 9177.05 Central.AirTRUE -282.02 Age - 27920.37 New.ConstructTRUE + 12623.45 Lot.Size - 10759.38 Heat.Type2 - 7170.77 Heat.Type3

The model uses the following predictor variables: Living.Area, Land.Value, WaterfrontTRUE,Bathrooms, Central.Air, Age, New.ConstructTURE, Lot.Size, and Heat.Type

The summary output shows that the overall model is statistically significant with a p-value of < 2.2e-16. All of the predictor variables have a p-value less than 0.05, except for Heat.Type3 which has a p-value of 0.09.

Predictions and Evaluation

The forward stepwise regression model is used to make predictions on the test data set. Based on these predictions, the model is evaluated using the relevant KPIs.

To assess the model, the accuracy of the predictions made using the model is compared with the actual values in the test data set. This is computed using the residuals, which is the difference between the actual price and the predicted price. Following this, the KPIs are generated – RMSE and MAPE. The RMSE for this model is $45207.82, which means that on average, using the given seed, the expected predictions are within $45207.82 of the actual price of the house. The MAPE is 20.5%, which indicates that on average, given the current seed, the predictions are within 20.5% away from the actual price of the house.

Benchmarking

In order to evaluate the model KPIs, benchmark KPIs are created. Comparing the model KPIs to the benchmark KPIs assesses the model created in comparison to having no model at all. The mean of the response variable, Price, from the training data set is used to compute the benchmark, following which the relevant residuals are calculated. The residuals of the benchmark are the difference between the actual house prices in the testing data and the benchmark. The benchmark RMSE is $74883.25, and the benchmark MAPE is 36.8%. These benchmark numbers indicate that, on average, the predictions made using the mean of all the house prices would be $74883.25 away from the actual price of the house. In addition, predictions made with the benchmark are expected to be within 36.8% away from the actual house price. The model KPIs are lower than the benchmark KPIs indicating the model without outliers is better than not using any model.

Model with Outliers

Outliers can influence the result and predictive power of a model. However, simply removing all the outliers can also lead to a biased sample. In order to analyze the effects of outliers on the data set, a linear regression model is created with the data set including outliers.

Build the Model

Following the procedure as the model without outliers, except for removing the outliers, the linear regression model generated the following output based on the forward stepwise regression method:

Price = 17164.12 + 74.11 Living.Area + 0.88 Land.Value + 23158.62 Bathrooms + 111877.37 WaterfrontTRUE - 43125.92 New.ConstructTRUE - 12734.90 Heat.Type2 - 10033.28 Heat.Type3 - 9082.46 Bedrooms + 2873.11 Rooms + 5047.88 Lot.Size + 8613.31 Central.AirTRUE

The model uses the following predictor variables: Living.Area, Land.Value, Bathrooms, WaterfrontTRUE, New.Construct, Heat.Type, Bedrooms, Rooms, Lot.Size, and Central.AirTURE.

The model is first analyzed based on the statistical significance of the overall model as well as all the predictor variables. At a significance level of alpha = 0.05, there is sufficient evidence to conclude that there is a linear relationship between price and the predictor variables. In addition, all the predictor variables in the model are statistically significant since they all have p-values less than 0.05. The model is also evaluated using the R^2 and standard error. The R^2 of this model is 0.6392, which means that 63.92% of the observed variation in price can be explained by the observed variation in the predictor variables. The standard error implies that the predictions have an error of $58440, on average.

Predictions and Evaluations

The model is evaluated using RMSE and MAPE. The RMSE is $58248.44, and the MAPE is 30.4%. This suggests that the predictions made using the model with outliers will be, on average, within $58248.44 from the actual house prices. The MAPE of 30.4% indicates that, on average, predictions are about 30.4% away from the actual prices of the house.

Benchmarking

Similar to the previous model, the benchmark KPIs are generated to compare the model KPIs. The benchmark RMSE is $102059.64, and the benchmark MAPE is 49.4%. Both the model KPIs are lower than the benchmark KPIs indicating that the model with outliers is better than not using any model.

Linear Regression Summary

	Model Without Outliers	Model With Outliers
RMSE ($)	45207.82	58248.44
Benchmark RMSE ($)	74883.25	102059.64
MAPE	20.5%	30.4%
Benchmark MAPE	36.8%	49.4%

The linear regression model without outliers has a lower RMSE and MAPE than the linear regression model with outliers. The model created without outliers has an RMSE of $45207.82 and a MAPE of 20.5%. This indicates that the predictions made using the model without outliers are closer to the actual house prices and are better than the model with outliers. Furthermore, the average error for the model without outliers is lower than the average error for the model with outliers. This means that the predictions made using the model without outliers are better and more accurate than having no model at all. In conclusion, the model without outliers is useful in predicting the prices of houses in Saratoga, New York.

Model 2 – Logistic Regression

Research Question: To what extent can the probability of a house having central air be predicted using the other home features in Saratoga, NY?

The logistic regression model aims to investigate the relationship between central air and other home features. The goal is to predict the probability of a house having central air based on the price, lot size, whether it is located on the waterfront, the age, the land value, whether it is newly constructed, the fuel type, heat type, sewer type, the living area, the number of bathrooms, bedrooms, fireplaces, and rooms.

Model without Outliers

Data Management

Logistic regression requires a logical variable as its response variable. Thus, the response variable – central air – is converted to a logical variable. In addition, the predictor variables with binary or factor levels must be converted to factors. This includes, Sewer.Type, Fuel.Type, Heat.Type, Waterfront, and Newly Constructed. Following which, the outliers were removed using the removeOutliers command based on the z-score method.

Partition the Data

A seed with code 1234 is set to standardize the model results. The data is partitioned using the 70:30 split to create the training and test data set.

Build the Model

The logistic regression model is built using R with Central Air as the response variable and the other variables as the predictor variables. The output is as follows:

Log(P(New.Construct/1-P(New.Construct)) = -15.067761584810 +0.000003241731 Price -1.173404269351Lot.Size -1.446681163697 waterfornt1 -0.015248758774 Age +0.000013613153 Land.Value -1.554191756494 New.Construct 0.670303684255 Fuel.Type2 -0.200728920196 Fuel.Type3 -1.897968243421 Heat.Type2 -1.547889594694 Heat.Type3 +

13.742864093732 Sewer.Type2 + 13.939368747856 Sewer.Type3 +0.000407733923 Living.Area - 0.385811298057 Bedrooms -0.783986218028 Fireplaces + 0.648424692743 Bathrooms -0.064040714055 Rooms

Given that many of the predictor variables are not statistically significant, the model is improved using the stepwise regression method. The backward stepwise regression is conducted to eliminate predictor variables that are not statistically significant and obtain a model with the lowest AIC value.

The equation generated using this model is:

Log(P(Central.Air/1-P(Central.Air)) = -1.211038615806 +0.000003770712 Price-1.376033335893 Lot.Size -1.616448175516 waterfront1 -0.016748925255 Age +0.000015123563 Land.Value -1.513974212663 New.Construct1 -1.850417474184 Heat.Type2 -0.860718047932 Heat.Type3 -0.372845905209 Bedrooms +0.829865094334 Fireplaces +0.725971623473 Bathrooms

The backward stepwise regression removed Fuel Type, Sewer Type, Living Area, and Rooms. The odds are depicted as the probability of a house having central air compared to the probability of a house not having central air. Taking the logarithm of the odds results in the logit function.

Predictions and Evaluations

The predictions are generated using a threshold of 0.5. As a result, when the value of the prediction is greater than 50%, the output is TRUE.

The model is first evaluated using the error rate. The model has an error rate of 25.4%. The benchmark error rate is 38.7%. Since the model error rate is lower than the benchmark error rate, it suggests that the model without outliers is better than not having any model.

		Actuals	
		FALSE	TRUE
PredTF	FALSE	252	82
	TRUE	38	101

Sensitivity	55.20%
Specificity	86.90%

Figure 5

Confusion Matrix (without outliers), Sensitivity and Specificity table (without outliers)

The above table shows the number of false positives and false negatives. A false positive is a when house is predicted to have a central air but in reality, it does not. There are 38 false positives. A false negative is when a house is predicted to not have central air when in reality, it does. There are 82 false negatives.

The sensitivity of the model is 55.2%, which indicates that the model correctly identifies houses with central air 55.2% of the time. The model is averagely sensitive to houses with central air. The specificity is 86.9%, which means that the model is able to identify a house without a central air 86.9% of the time. This means that the model can specifically distinguish between houses with and without the central air.

The logistic regression model can also be evaluated using the ROC and Lift chart. Below is the ROC chart for the model without outliers. The ROC chart reflects the trade-off between sensitivity and specificity. The AUC is 0.8 which is close to the ideal value of 1. Thus, the AUC of the model is closer to the idea case than the benchmark.

Figure 6

ROC chart

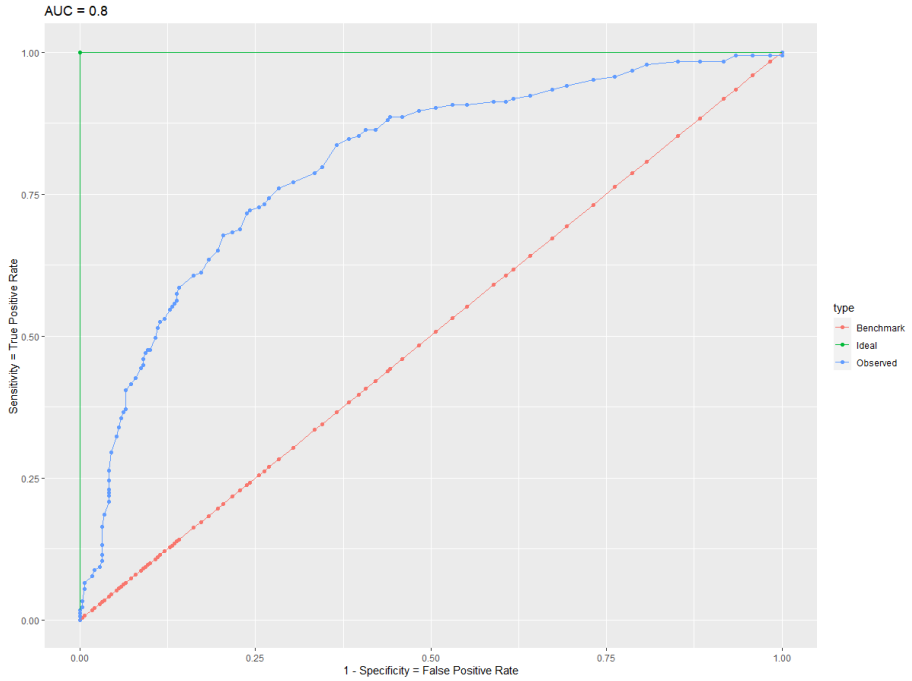
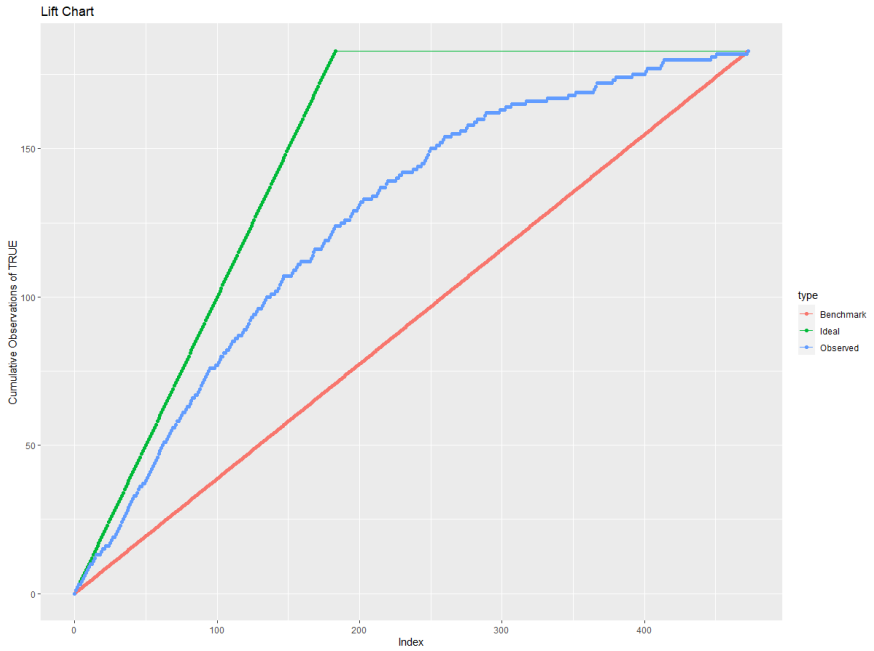


Figure 7

Lift Chart



In addition, the observed line for the lift chart is between the ideal line and the benchmark line. Thus, the observed line is lifted away from the benchmark line indicating that the model is a good fit for the data.

Model with Outliers

Similar to linear regression, a model with outliers is created for logistic regression.

Build the Model

Following the same procedure as the model without outliers, except for removing the outliers, the logistic regression model generated the following output based on the backward stepwise regression method:

Log(P(New.Construct/1-P(New.Construct)) =-0.144571166575 + 0.000006526178 Price - 0.031494624647Age -0.780185422296 New.Construct1 -0.256901804277 Bedrooms - 0.910708293908 Fuel.Type2 -1.108148588870 Fuel.Type3

The backward stepwise regression removed all the predictor variables except Price, Age, New Construct, Bedrooms, Fuel.Type2, and Fuel.Type3.

Predictions and Evaluation

The predictions are generated using a threshold of 0.5.

The model is first evaluated using the error rate. The model has an error rate of 25.5%. The benchmark error rate is 38.8%. Since the model error rate is lower than the benchmark error rate, it suggests that the model without outliers is better than not having any model.

		Actuals	
		FALSE	TRUE
PredTF	FALSE	275	90
	TRUE	42	111

Sensitivity	55.20%
Specificity	86.80%

Figure 8

Confusion Matrix (with outliers), Sensitivity and Specificity table (with outliers)

The above table shows the number of false positives and false negatives. A false positive is a when house is predicted to have a central air but in reality, it does not. There are 42 false positives. A false negative is when a house is predicted to not have central air when in reality, it does. There are 90 false negatives.

The sensitivity of the model is 55.2%, which indicates that the model correctly identifies houses with central air 55.2% of the time. The model is moderately sensitive to houses with central air. The specificity is 86.8%, which means that the model is able to identify a house without a central air 86.8% of the time. This means that the model can specifically distinguish between houses with and without the central air.

The logistic regression model can also be evaluated using the ROC and Lift chart. Below is the ROC chart for the model without outliers. The AUC is 0.82 which is close to the ideal value of 1. Thus, the AUC of the model is closer to the idea case than the benchmark.

Figure 9

ROC chart

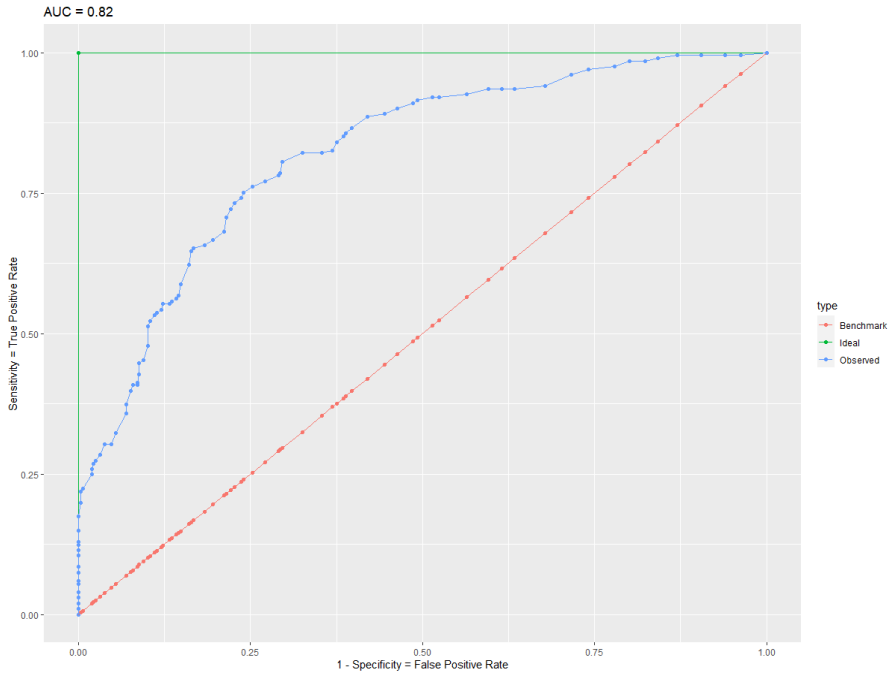


Figure 10

Lift chart

In addition, the observed line for the lift chart is between the ideal line and the benchmark line. Thus, the observed line is lifted away from the benchmark line indicating that the model is a good fit for the data.

Logistic Regression Summary

	Model with outliers	Model without outliers
Model's error rate	25.5%	25.4%
Benchmark error rate	38.8%	38.7%
AUC	0.82	0.80
Lift chart	Similar graph	

The model is evaluated using the error rate, sensitivity, specificity, ROC and Lift chart. There is minimal difference in the error rates of the model with outliers and the model without outliers. Similarly, there is minimal difference between the benchmark error rates of the model with outliers and the model without outliers. Regardless, the error rates for the models are better than the benchmark error rates. The AUC of both the models are also very close to the ideal case and the lift charts are very similar. Thus, the model with outliers is better since it has a slightly higher AUC, and is more accurate than having no model at all. In conclusion, the model with outliers is useful in predicting the probability of having central air for houses in Saratoga, New York.

Model 3 – KNN Regression

Research Question: To what extent can central air, number of rooms, number of bathrooms, and living area be used to predict houses with similar prices in Saratoga, NY?

The response variable is price and the predictor variables are central air, number of rooms, number of bathrooms, and living area since these are the highly valued home features according to buyers.

Model without Outliers

Data Management

KNN regression requires a numerical response variable, Price. In addition to converting the predictor variables to their correct types, the KNN regression model requires additional data management. Since the predictor variables are measured on different scales, one or more predictor variables could overpower the others when calculating the proximity. Therefore, all the predictor variables are standardized. To standardize them, they are first converted to integers so that they are accepted as valid input by the preProcessor function. The predictor variables are standardized instead of normalized in order to include the factor and logical variables from the data set. Standardization takes each data point, subtracts the mean, and then divides by the standard deviation, resulting in z-scores. All the variables except for the response variables, Price, are standardized.

Partition the Data

A seed with code 1234 is set to standardize the model results. The data is partitioned using the 70:30 split to create the training and test data set. A k-value of 3 is used since the data set is relatively small.

Build the Model

The model is built in R using the KNNReg function with Price as the response variable, and Central Air, Rooms, Bathrooms, and Living Areas as the predictor variables. The training data is used with a k-value of 3.

Predictions and Evaluation

The KNN regression model is evaluated using the RMSE and MAPE. The RMSE for the model is $60176.91. It implies that there is an average error of $60176.91 in the prediction of prices for houses in Saratoga. In addition, the MAPE is 25.3%. This indicates that the average percentage error in the predictions for housing prices using this model is 25.3%.

Benchmarking

The model KPIs are assessed by comparing them to the benchmark KPIs. The RMSE is $74883.25 and the MAPE is 36.8%. This implies that there is an average error of $74883.25 in the prediction of prices for houses in Saratoga. In addition, the average percentage error the predictions for housing prices using this model is 36.8% The model KPIs are better than the benchmark KPIs since they are lower. Thus, indicating that the model without outliers is better than not using any model.

Model with Outliers

Similar to linear and logistic regression, a model with outliers is created for KNN regression as well.

Build the Model

Following the same procedure as the model without outliers, except for removing the outliers, the KNN regression model is created using the same function in R.

Predictions and Evaluations

The model without outliers has an RMSE of $77788.38 and a MAPE of 36.7% The RMSE for the model implies that there is an average error of $77788.38 in the prediction of prices for houses in Saratoga. In addition, the average percentage error in our predictions for housing prices using this model is 36.7%.

Benchmarking

The model KPIs are assessed by comparing them to the benchmark KPIs. The RMSE is $102059.64 and the MAPE is 49.4%. This implies that there is an average error of $102059.64 in the prediction of prices for houses in Saratoga. In addition, the average percentage error of the predictions for housing prices using this model is 49.4% The model KPIs are better than the benchmark KPIs since they are lower. Thus, indicating that the model with outliers is better than not using any model.

KNN Regression Summary

	Model Without Outliers	Model With Outliers
RMSE ($)	60176.91	77788.38
Benchmark RMSE ($)	74883.25	102059.64
MAPE	25.3%	36.7
Benchmark MAPE	36.8%	49.4%

The model KPIs are lower than the benchmark KPIs for the model with and without outliers. Overall, the model without outliers has a lower RMSE and MAPE than the model with outliers. Therefore, removing the outliers improved the model and the model without the outliers since it has a significantly lower RMSE and MAPE compared to the benchmark. Thus, it is better than not using any model to predict houses with similar prices of houses in Saratoga, NY based on the central air, number of bedrooms, number of bathrooms, and living area.

Conclusion

In light of the above analysis, it can be concluded that both the linear regression and KNN regression models without outliers perform better than the models with outliers; i.e., those models are better at predicting housing prices than the models with outliers.

On the other hand, the logistics regression model with outliers that predicts whether a house has or does not have a central air, performed better than the model using the data without outliers. As a result, this model has the potential to be used to predict whether a house has central air using other data sets with outliers as well.

In summary, the linear regression and KNN regression models could possibly be implemented to predict the prices of houses in other housing areas similar to Saratoga (NY), such as Chapel Hill (NC), Ithaca (NY), and Athens (GA). Furthermore, as mentioned above, the logistics regression model has the aptitude to predict whether a house has central air using data sets with outliers.

Finally, housing is a topic that everyone should have some knowledge of, whether they plan on buying a house or not. In a world where the housing boom happened less than fifteen years ago, and there is yet another housing shortage and price surge today, a basic understanding of what is going on and what is to be expected in the housing market is important, if not crucial, for everyone to know.

References

6 critical things to look for when buying a house: Zillow. Home Buyers Guide. (2022, April 1). Retrieved April 15, 2022, from https://www.zillow.com/home-buying-guide/what-to-look-for-when-buying-a-house/

Bruno, G. (2022, March 10). *Lawmaker: $200m needed in NY budget for Upstate Housing Crisis*. NEWS10 ABC. Retrieved April 15, 2022, from https://www.news10.com/news/lawmaker-200m-needed-in-ny-budget-for-upstate-housing-crisis/

Doar, K. (2021, June 20). *Upstate New York's twin plagues: Housing crisis and labor shortage*. New York Daily News. Retrieved April 15, 2022, from https://www.nydailynews.com/opinion/ny-oped-upstate-new-yorks-housing-crisis-20210620-aasirmxghjh63hpydmrzoeh6xu-story.html

Media, N. V.-H. C.-G. (2022, March 25). *Migration from NYC lifts greene population*. HudsonValley360. Retrieved April 15, 2022, from https://www.hudsonvalley360.com/news/columbiacounty/migration-from-nyc-lifts-greene-population/article_d01c9254-8a63-5346-b813-a3ad8d55070a.html

Schaeffer, K. (2022, February 1). *A growing share of Americans say affordable housing is a major problem where they live*. Pew Research Center. Retrieved April 16, 2022, from https://www.pewresearch.org/fact-tank/2022/01/18/a-growing-share-of-americans-say-affordable-housing-is-a-major-problem-where-they-live/

Sisson, P., Andrews, J., & Bazeley, A. (2020, March 2). *The U.S. has an affordable housing crisis. here's why*. Curbed. Retrieved April 15, 2022, from https://archive.curbed.com/2019/5/15/18617763/affordable-housing-policy-rent-real-estate-apartment

StatCrunch. (n.d.). Retrieved March 25, 2022, from https://www.statcrunch.com/app/index.html?dataid=599280

U.S. Census Bureau QuickFacts. United States Census Bureau. (2021, July 21). Retrieved April 15, 2021, from https://www.census.gov/quickfacts/fact/table/US/PST045221