Chloe Adams, Alexa Martin, Katelin Meier & Rachel Redeker

Project Report, Group 6

December 8th, 2023

**0. Des Moines Housing Predictive Model Report**

The DataMiners

Chloe Adams, Alexa Martin, Katelin Meier & Rachel Redeker

## **1. Executive Summary**

This report outlines DataMiners LLC findings on the best predictive housing price model for SuperHomes Inc houses. Using a dataset provided by SuperHomes Inc, we utilized data mining techniques to create a model that effectively predicts housing prices. The provided dataset included over 30 features of houses commonly found in the Des Moines area. Being tasked with creating a model that predicts price utilizing as few features as possible, DataMiners LLC found the Gradient Boosting model produced the best results. The final model uses twelve features to predict price while maintaining model accuracy.

## **2. Problem Description**

1. ***Background***
   1. SuperHomes Inc. has been in business for over 50 years as a premier real estate company in Des Moines. The current CEO is preparing to retire, and his daughter, Sarah, is planning to step in as CEO. She has some ideas to make day-to-day operations easier and has hired us, the DataMiners LLC, to create a model that predicts housing prices. Sarah has explained that SuperHomes LLC. has few employees and many potential clients. She wants to make internal home price analysis efficient for her employees. It is particularly important to her the model uses as few features as possible while still maintaining its integrity.
2. ***Business Goal and Data Mining Goal***
   1. We are attempting to predict housing prices in Des Moines using housing characteristics. This problem is considered a regression data mining problem. We want to use the least number of features without compromising model accuracy. We plan to build models using Linear Regression, Random Forests, and Boosting models to see which models are most accurate and effective on the validation set.
      1. Given the qualities of a home, what is the predicted price?
      2. What features of a home are most important to predicting the price?

## **3. Data Description**

1. ***Data***
   1. Sarah has provided the DataMiners LLC. data from their 2014 – 2016 sales as well as numerous features collected on the homes. There are approximately 9,554 sale records and 37 features. The features provide copious amounts of details about the homes sold. We will use the data given and narrow down to the key features that predict the price of a home.
2. ***Data Dictionary***

|  |  |  |
| --- | --- | --- |
| **Feature** | **Type** | **Description** |
| Price | Numeric | The price of the house |
| Jurisdiction | Text | City in which the house is located |
| Neighborhood | Text | Neighborhood in which the house is located |
| School District | Text | School district that the house location belongs to |
| Address | Text | Address of the house location |
| Zip | Text | Zip code of the house |
| Sale Date | Date | Date in which the house sold |
| Land Acres | Numeric | Total acres of the house and its land |
| Residence Type | Text | Type of house including how many stories |
| Building Style | Text | Style of the home (ie. Ranch, Conventional) |
| Exterior Wall Type | Text | Material of the exterior |
| Roof Type | Text | Type of roof on the house |
| Roof Material | Text | Roof material of the house |
| Total Living Area | Numeric | Area of the living area not including unfinished basement |
| Foundation | Text | Type of foundation (brick, concrete, etc.) |
| Basement Area | Numeric | Total area of the basement |
| Finished Basement Area | Numeric | Area of the basement that is unfinished |
| Basement Walkout Area | Numeric | Area of basement walkout |
| Attached Garage Area | Numeric | Area of the garage |
| Open Porch Area | Numeric | Area of open porch |
| Enclosed Porch Area | Numeric | Area of enclosed porch |
| Patio Area | Numeric | Area of patio |
| Deck Area | Numeric | Area of the deck |
| Carport Area | Numeric | Area of the carport or where cars are stored |
| Bathrooms | Numeric | Number of full bathrooms in the house |
| Toilet rooms | Numeric | Number of rooms with just toilets |
| Whirlpools | Numeric | Number of whirlpools the house has |
| Hot tubs | Numeric | Number of hot tubs the house has |
| Saunas | Numeric | Number of saunas the house has |
| Fireplaces | Numeric | Number of fireplaces in the house |
| Bedrooms | Numeric | Number of bedrooms in the house |
| Rooms | Numeric | Number of rooms in the house |
| Year Built | Numeric | Year the house was built |
| Year Remodeled | Numeric | Year the house was remodeled |
| Condition | Text | Condition of the house (Excellent, Very Good, Above Normal, Normal, Below Normal, Very Poor) |
| Heating | Text | Type of heating system the house has (Gas, Electric, Etc.) |
| Air Conditioning | Numeric | Whether or not the house has air conditioning (1 for yes, 0 for no) |

[Data was obtained from Data Wrangling, Professor Michael Colbert]

1. ***Exploratory Data Analysis***

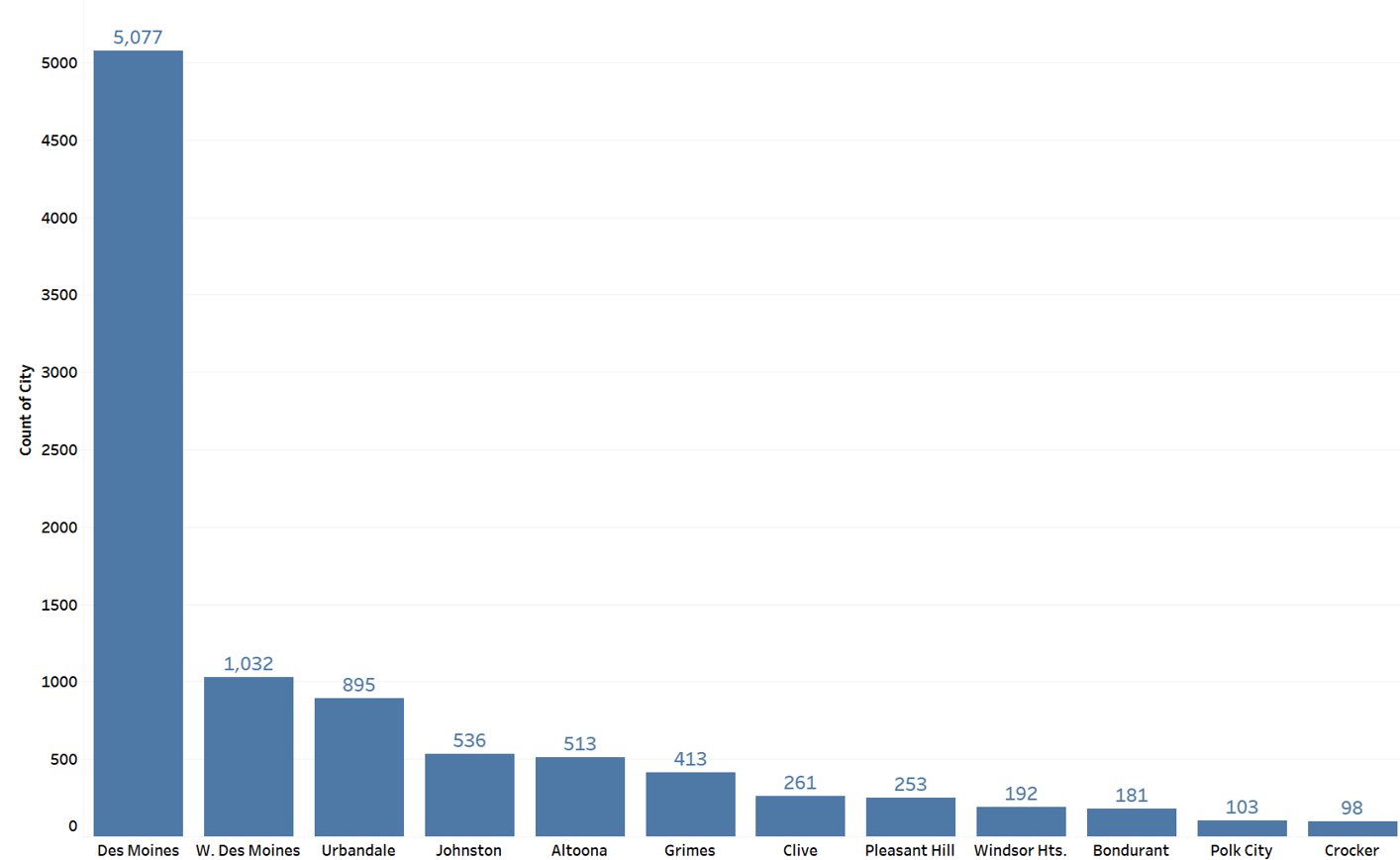
**Price is Most Highly Correlated with Total Living Area**

A white background with black dots

Description automatically generated

As we can see by the correlation widget in Orange, total living area is most correlated with sale price with a correlation of 0.831, indicating that the variables are closely related. Bathrooms and attached garage area come in next with correlations of 0.719 and 0.696 to sale price, respectively. Removing closely correlated variables reduces redundancy in our model.

**Des Moines has the Highest Number of Houses in the Dataset**



This bar chart created in Tableau shows the distribution of houses sorted by the categorical data of cities. Des Moines has the highest number of houses coming in at about 53%, or 5,077. Next is West Des Moines with about 10% of the data or 1,032. Urbandale has the third most data points with about 9 % or 895. Our model will be most successful predicting prices of homes in Des Moines and W. Des Moines given the majority of data is from Des Moines.

**Housing Locations Visualized by City**

A screen shot of a computer generated image

Description automatically generated

We decided to further visualize the locations of homes in the data set. This map created in R shows the locations of houses in our dataset. As you can see in the map Des Moines has the highest number of houses and is in the center of the region we are analyzing. Coming in next is West Des Moines which also constitutes a large part of the housing data.

**Housing Price Increases with Condition**

**A line graph with numbers and a line

Description automatically generated**

This line graph created in Tableau simplifies the conditions by average price. As expected, the price increases with improvement in condition. It is interesting to note a deviation from the linear increase with houses labeled “normal.” This could indicate that some houses labeled “normal” are mislabeled and are in better condition than stipulated. There may also be more data points referenced to find the average. We will further explore conditions in the graph below.

**Condition of Home Appears Evenly Distributed Among Cities**

A graph of a bar chart

Description automatically generated with medium confidence

Like the housing distribution bar chart above, this graph created in Tableau separates houses sold by city. However, this visualization adds an additional feature: conditions. We decided to examine conditions further to see the distribution across cities. As seen in the chart, the conditions of homes appear to be distributed evenly between cities. The majority of homes are considered “normal” condition, this makes sense with graph found above, there are more sample points that could be increasing the average. The only city that strays from distribution is Clive. Clive seems to have a higher proportion of homes labeled “very good.”

**Houses with Air Conditioning have a Higher Average Price**

A white background with numbers and lines

Description automatically generated

This boxplot created in Orange shows the price of a house split up by whether the house has air conditioning. 0 represents the house does not have air conditioning (top boxplot) and 1 represents that it does have air conditioning (bottom boxplot). As you can see in the boxplot, houses with air conditioning have a significantly higher price on average. The average price of a house without air conditioning has a mean price of $91,522.69 and those with air conditioning have an average price of $185,997.4. This distribution makes sense as individuals prefer to have air-conditioning in their homes. In the end, this variable was not included in our model as a top feature which might have to do with the fact that most of our houses do have air conditioning.

**Houses with a Basement Walkout have a Higher Average Price**

A white paper with black text

Description automatically generated with medium confidence

This boxplot created in Orange shows the price of a house split up by whether the house has a basement walkout. 0 represents the house does not have a basement walkout (top boxplot) and 1 represents that it does have a basement walkout (bottom boxplot). As you can see in the boxplot, houses with a basement walkout have a significantly higher price on average. The average price of a house without a basement walkout has a mean price of $168,323.03 and those with a basement walkout have an average price of $318,582.24. This distribution makes sense as a walkout basement is more luxurious. In the end, this feature was not included in our final model as it was not as important as other features.

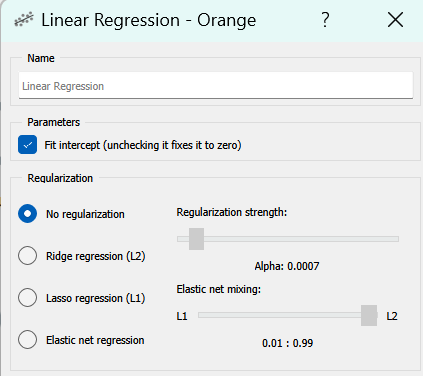
1. ***Data Preprocessing***
   1. Feature Engineering
      1. We did some feature engineering to minimize variables while still incorporating most of the provided data:
         1. Created **Outdoor Living Area** by adding open porch area + patio area + deck
         2. Created **Extra Fixtures** by adding whirlpools + saunas + hot tubs
         3. Created **Age** of the house from subtracting sale date from year built.
         4. Created **Parking** **capacity** by looking at parking area and using an if then analysis.
            1. If square feet are less than 150, then 0 cars
            2. If square feet are less than 360, then 1 car
            3. If square feet are less than 600, then 2 cars
            4. If square feet are greater than or equal to 600, then 3 cars
         5. Created **Bathrooms** by adding bathrooms + 0.5\* toilet rooms.
         6. Created **Max Occupants** by multiplying bedrooms by 2.
         7. Created **Price Per Square Foot** by dividing price/total living area.
         8. Changed **Porch**, **Basement Walkout** and **Air Conditioning** to categorical binary (0, 1) values.
   2. Feature Preprocessing
      1. Categorical Features
         1. We used the preprocess widget to continuize discrete variables using most frequent as base to convert categorical features to binary values.
         2. A screenshot of a computer

            Description automatically generated
      2. Numeric Features
         1. We used the preprocess widget to normalize our numeric features from the interval [0,1] to ensure larger values did not dominate smaller values like housing price with age of the house.
         2. A screenshot of a computer

            Description automatically generated

## **4. Data Mining Solution**

1. ***Models Created***
   1. **Model 1: Extreme Gradient Boosting**
      1. Used 12 out of the 37 features
         1. Acres, living area, basement area, garage area, fireplaces, rooms, parking, baths, year built, extra fixtures, condition, and city.
      2. A screenshot of a computer

         Description automatically generatedHyperparameters
         1. Method: Extreme Gradient Boosting (xgboost)
         2. Number of trees: 96
         3. Learning Rate: 0.098
         4. Lambda: 1
         5. Limit Depth: 2
   2. **Model 2: Linear Regression**
      1. Used 12 out of the 37 features
         1. Acres, living area, basement area, garage area, fireplaces, rooms, parking, baths, year built, extra fixtures, condition, and city.
      2. Hyperparameters
         1. Parameters: Fit intercept
         2. Regularization: No regularization
   3. **Model 3: Random Forest**
      1. Used 12 out of the 37 features
         1. A screenshot of a computer

            Description automatically generatedAcres, living area, basement area, garage area, fireplaces, rooms, parking, baths, year built, extra fixtures, condition, and city.
      2. Hyperparameters
         1. Number of trees: 400
         2. Number of attributes at each split: 5
         3. Limit depth: 5
         4. Do not split subsets smaller than: 4
2. ***Performance Evaluation***
   1. Model 1 Performance (Boosting)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MSE** | **RMSE** | **MAE** | **R^2** |
| **Train Data** | 1,298,797,269.73 | 36,038.83 | 23,713.48 | 0.889 |
| **Test Data** | 1.443,250,975.39 | 37,990.14 | 24,186.74 | 0.877 |
| **Cross Validation** | 1,681,329,185.66 | 41,004.01 | 25,066.98 | 0.857 |

* + 1. Our performance on this model seems to be doing well. Our cross-validation R^2 shows that around 85.7% of our variability in this model can be explained by the variables we are using. Our model has similar mean squared errors among the training set and the testing set suggesting that our model fits the data well and is not overfitting very much. There may be a little overfitting but that is expected with more complex models like Gradient Boosting.
  1. Model 2 Performance (Linear Regression)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MSE** | **RMSE** | **MAE** | **R^2** |
| **Train Data** | 1,848,608,691.97 | 42,995.45 | 27,035.22 | 0.843 |
| **Test Data** | 1,878,420,503.52 | 43,340.75 | 26,812.01 | 0.840 |
| **Cross Validation** | 1,887,785,389.09 | 43,448.65 | 27,206.66 | 0.839 |

* + 1. Our performance on this model seems to be slightly underfitting. As you can see through the error metrics, our training data error and testing data error are very similar. However, compared to our first model the results are lower, suggesting that the model is not complex enough.
  1. Model 3 Performance (Random Forest)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MSE** | **RMSE** | **MAE** | **R^2** |
| **Train Data** | 1,747,383,792.50 | 41,801.72 | 27,336.65 | 0.851 |
| **Test Data** | 1,995,821,877.34 | 44,674.62 | 27,972.80 | 0.830 |
| **Cross Validation** | 2,075,360,426.04 | 45,556.12 | 28,277.33 | 0.823 |

* + 1. Our performance on this model is not nearly as good as our first model. This is because it has some overfitting due to the use of a more complex model and still is not performing very well on either the training or testing set.

## **5. Conclusion**

1. ***Gradient Boosting Model Recommended***
   1. Data Miners LLC concluded the Gradient Boosting model is the best model to estimate the price of homes for SuperHomes Inc. The model includes only twelve features: acres, living area, basement area, garage area, fireplaces, rooms, parking, baths, year built, extra fixtures, condition, and city. Our Gradient Boosting model allows the most accurate prediction with fewer features while still maintaining the integrity of the data without overfitting or underfitting.
   2. The most influential features impacting the model include living area, garage area, and basement area. These features align with common assumptions about homes. Much of the price of a home is determined by how large the house is, larger homes cost more while smaller homes are cheaper. The chart below visualizes the features contributing to the model most.

**Living Area Contributes Most to Boosting Model**

A screenshot of a graph

Description automatically generated

* 1. We wanted to provide further visualization surrounding the top contributing feature: living area. We created a scatterplot showing the relationship between living area and housing prices. Through the chart, you can see that these two variables are highly related as there is a correlation of 0.83 and the price increases as living area increases. It is also important to note that most of our houses fall in the lower price range and living area as there is a cluster in the bottom left corner of the graph.

**Living Area and Price have a Positive Relationship**

A diagram of a graph

Description automatically generated with medium confidence

1. ***Limitations to Current Model***
   1. Over half of the provided data is strictly from the city of Des Moines, if SuperHomes wanted to expand to other areas, such as Altoona or Pleasant Hill, the created model would produce less accurate results. DataMiners would need to obtain more data instances to produce an updated model.
   2. This model only functions using twelve features we have selected. All twelve features for a new house or potential property must be known for this model to predict the pricing accurately.
   3. The data used to create the model only included SuperHomes sales from 2014-2016. The real estate market changes often, to create a more accurate model, DataMiners would need to obtain more recent sales data.
2. ***Future Work & Going Forward*** 
   1. SuperHomes should implement the new Gradient Boosting model into their analysis. DataMiners will continue to update the model using new data provided by SuperHomes every year.
   2. We, the DataMiners, would be interested in collaborating with other companies: Zillow or Realtor.com to find more information that would better assist SuperHomes in price prediction.