|  |
| --- |
| **Elmhurst College** |
| Classifying the Price of a House in Ames, IA |
| MSD 570 Fall A 2019 |

|  |
| --- |
| Megan Cusey  12-9-2019 |

# Overview

The following case study seeks to identify the strongest predictive variables in order to predict house sales prices in Ames, IA. A CRISP approach was applied in order to accomplish this task. The following technical documentation breaks down each component of this analysis into 1.) Business Understanding, 2.) Data Understanding, 3.) Data Preparation, 4.) Modeling, 5.) Evaluation, and 6.) Next Steps (what would typically be deployment). A compilation of all the files used in this project can be found on my GitHub repository titled [DataScienceProjects/MDS 556 – House Prices Regression/](https://github.com/megancusey/DataScienceProjects/tree/master/MDS%20556%20-%20House%20Prices%20Regression).

# Business/Data Understanding

The data set for the Ames, IA house market obtains data that describes the features of each house. The data and business problem originates from a Kaggle Competition data set that you can find [here](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview). For the purpose of this assignment, the decision was made to tweak the suggested business problem from Kaggle which was to predict the sales price of a home given a set of features to identifying houses that are in the top 25% in sales price. Identifying houses that sell in the 25% of house prices is useful for the following reasons:

1. If you are a realtor that focuses on only the most luxurious homes, this model can quickly advise you if listing a certain home may be worth his or her time.
2. If you are a real estate developer, you would want to identify homes that have a high potential for profit that may be listed at a price that under values the home.
3. As a real estate developer, knowing what features of the home to invest in can be useful to maximize profit.

In both cases, the cost of inaccurately classifying a home as a “Top Dollar” home can be significant. A real estate agent may experience an opportunity cost of focusing resources on a less profitable home as opposed to a real “Top Dollar” home. Similarly, a real estate developer may apply more labor and material into a home that will not produce the expected sale price. For these reasons, a successful model will have a high precision rate. The best model will produce results that minimizes false positives.

A [data dictionary](https://github.com/megancusey/DataScienceProjects/blob/master/MDS%20556%20-%20House%20Prices%20Regression/data_description.txt) was supplied along with the data set. While most of the features were relatively self-explanatory, there was a few that didn’t seem to be common knowledge to me, a 1-time homeowner. The following are definitions I researched in order to complete this analysis and solidify my business understanding:

* Artery Street, a highway
* Feeder Street, an important road that feeds into a highly trafficked area
* Masonry Veneer, an external layer to the home that is non-structural (example: brick or stone)

# Data Preparation

## Data Cleaning

The train.csv dataset contains 1,460 rows of data with 81 features. A combination of numeric, categorical, and ordinal data types where present in the data. The first step in data preparation was to create a method to input the raw data directly from the csv file, investigate each feature, build validations for each feature according to the data definitions provided, and ensure that no missing values were presented. The following provides general guidelines for this process:

* If the feature had an ordinal nature, convert the variable to a factor. Specify the parameter as ordered = TRUE and apply the levels in the appropriate order.
  + NA was used to specify that a feature was not present such as “No Fireplace” or “No Garage”. Missing values were mapped to the appropriate level that indicate the absence of a characteristic of a home.
* If the feature was categorical, convert the variable to a factor.
  + NA was used to specify that a feature was not present such as “No Fireplace” or “No Garage”. Missing values were mapped to the appropriate level that indicate the absence of a characteristic of a home.
* If the feature was numeric, I mostly let R default the data type to what it was when it loaded the data set.
  + In some features, a missing value indicates that the feature didn’t exist. In this case, I supplied 0 for those observations.
  + In other features, the mode was used to replace a missing value.

## Data Exploration

### About the Target

A screenshot of a cell phone

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| summary(data$SalePrice) | | | | | |
| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| 34900 | 129975 | 163000 | 180921 | 214000 | 755000 |

### About the Explanatory Variables

There are 30 numeric variables, 37 categorical/ordinal variables, and 9 binary variables.

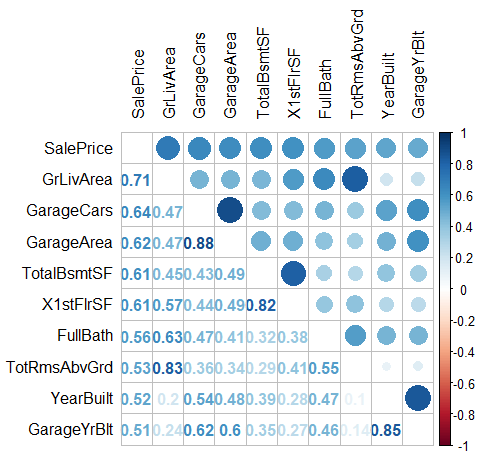
A screenshot of a cell phone

Description automatically generated

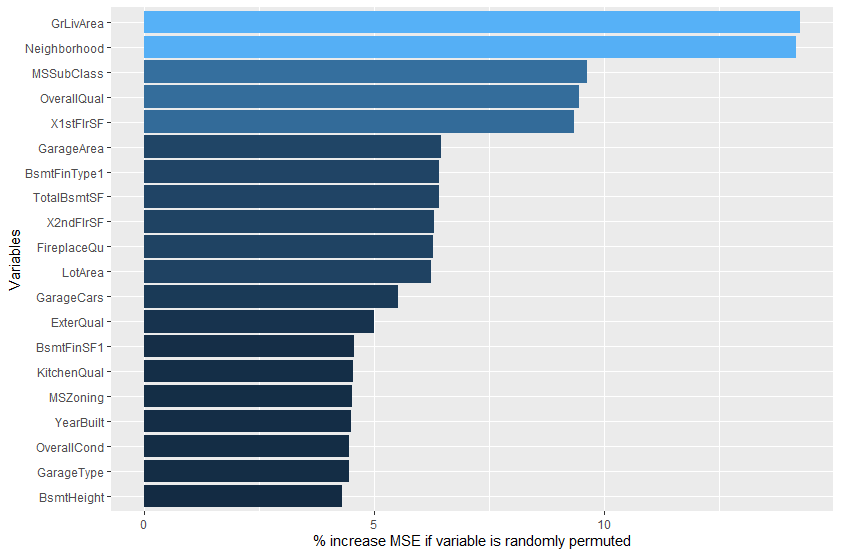
The above plot graphs the numeric features’ s distributions. I noted the following observations:

* There are many variables that have a skewed distribution. For instance, BsmtFinSF2, BsmtHalfBath, LowQualFinSF, KitchenAbvGr, WoodDeckSF, and X3SsnPorch are all features that are not that common in all houses. Therefore, the bulk of the observations are at 0.
* The GrLivArea observations center between 1000-2000 square foot. There appear to be outliers around the 4500 + square foot area.
* Most of the houses in the dataset where built around 1995+
* There appears to be some seasonality in the month sold. More homes were sold in the summer than in fall-spring months.

The below plot represents the correlations that are significant between the variables. You can see that there is some multicollinearity issues, as expected, due to the similarities between variables. For instance, GrLivArea (above ground square footage) and TotRmsAbvGrd (number of rooms above ground) are highly correlated with one another. Naturally, bigger homes have more bedrooms, so the correlation makes sense. Another expected observation of the correlations is the relationship between the sale price (SalePrice) and above ground living area (GrLivArea). Again, it is expected that the larger the home, the higher the “price tag” on the house.



The next graph represents the feature importance results after applying a quick random forest:



After an initial first glance at the weight of the features on sale price, we see that the above ground living area and the location (neighborhood) of the home may explain much of the sale price of a home.

### Feature Engineering

The following paragraphs describe the process to address issues such as multicollinearity and too many categorical variables and manipulating the features to increase their predictive power.

#### Reducing Multicollinearity

As indicated by the correlation matrix along with some of the data definitions, there were several variables that had a strong correlation to once another.

The first example of this occurrence is the relationship between the above ground living area (GrLivArea) and the sum of the 1st floor square footage (X1stFlrSF), 2nd floor square footage (X2ndFlrSF), and the low quality finished square footage (LowQualFinSF). In fact, the correlation between these two figures were a perfect correlation since the above ground living area, by definition, was the sum of the three features. As a result, I excluded the X1stFlrSF, X2ndFlrSF, and LowQualFinSF from the data set.

The same relation existed between the total basement square footage (TotalBsmtSF) and the sum of BsmtFinSF1, BsmtFinSF2, and BsmtUnfSF. As a result, I was able to drop the three individual features that characterized TotalBsmtSF.

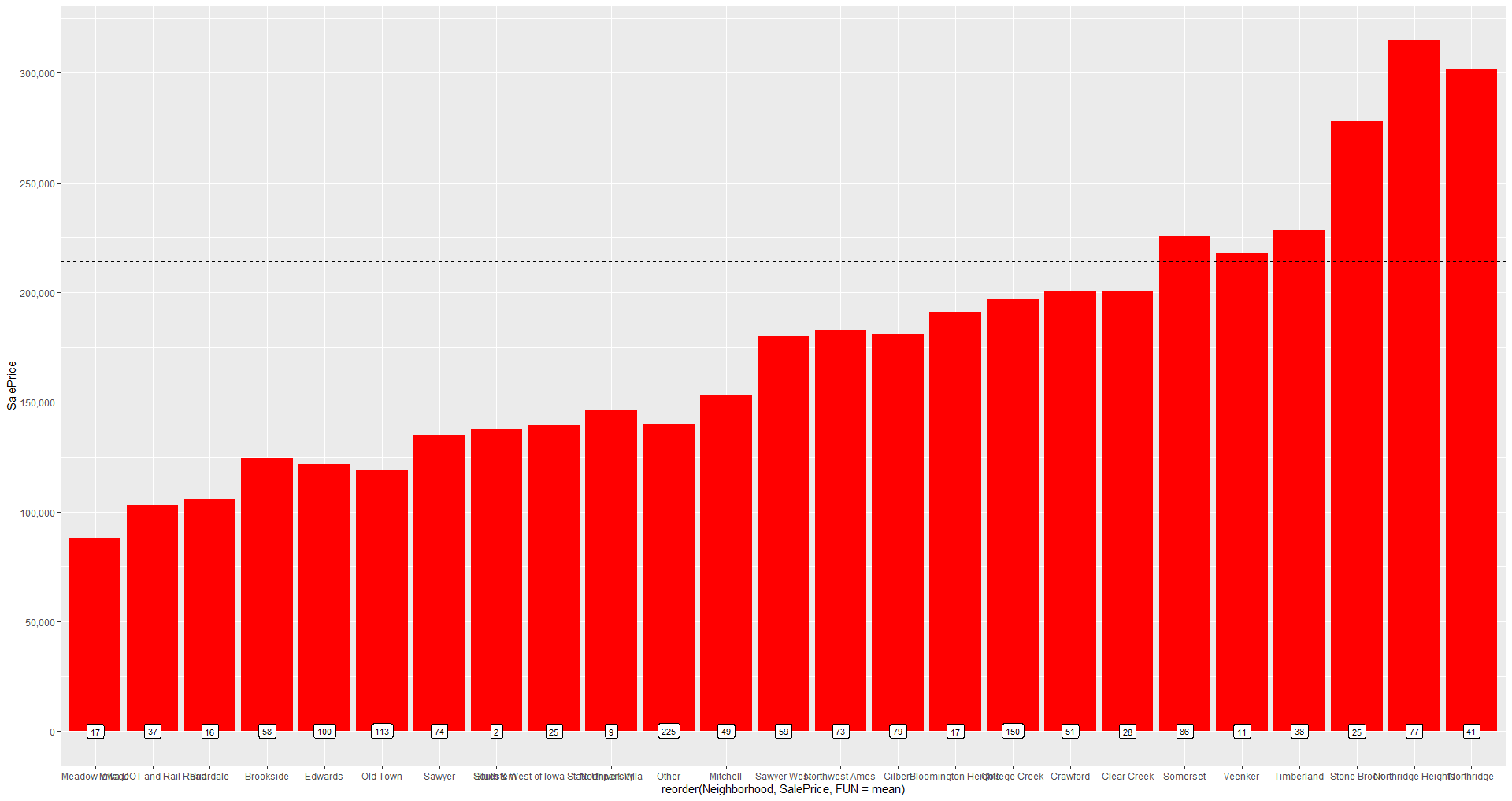
The last example of multicollinearity in this project was the relationship between GarageArea and Garage Cars (correlation: .88). Similar to number of rooms and above ground living area, the number of vehicles that can fit in a garage is really just the function of how big the garage area is. Since number of vehicles that fit into a garage is more commonly used to describe a garage, I decided to exclude the garage area feature from the analysis.

#### Binning Categorical Features

In some instances, there were two columns to describe the same feature. For instance, variables Exterior1 and Exterior2 described material used to on the outside of the house. While some houses had a 2nd Exterior material, most do not. As a result, I combined the two columns into one. Similarly,

Condition1 and Condition2 described a potential characteristic of the house such as near a railroad, highway, high trafficked area, or park. I decided to combine these conditions into binary variables (4 explanatory variables) and drop Condition1 and Condition2.

Throughout the dataset are ordinal features that qualify a condition of a feature in the home into categories such as “poor”,” fair”,” average”,” good”,” very good”, or “excellent”. For these features, I investigated the distribution of the categories. Some features had little observations in the extreme categories such as “poor” or “excellent”. As a result, I combined the data into more general categories such as “Above Average”, “Average”, or “Below Average”.

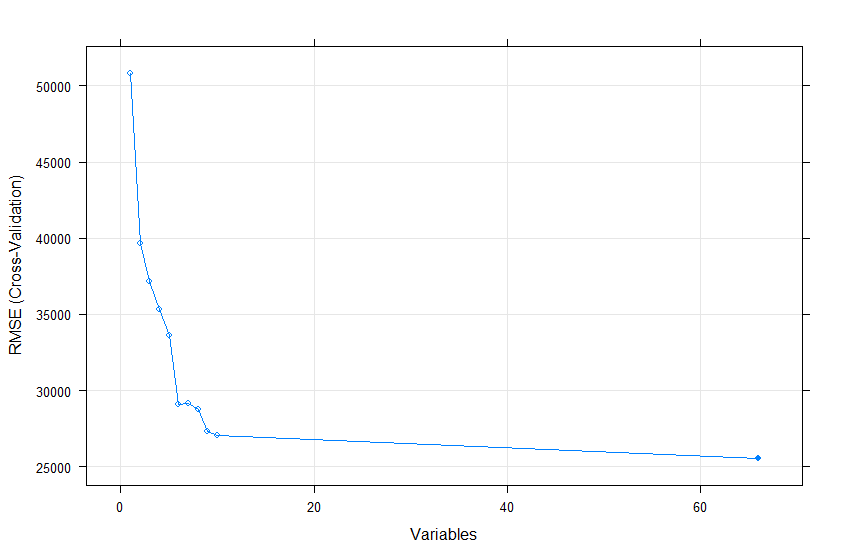


The above image is a plot of the median of the sale price according to the neighborhood of the observations. The neighborhoods are sorted by the mean of the sale price. The dotted line is the 3rd quartile of the dataset. The image depicts a clear ordinal nature of neighborhoods against sale price. As a result, I decided to create an ordered factor for this feature and classify the three highest neighborhoods as “Very Rich”, the next three highest above the 3rd quartile as “Rich”, and the rest as “Other”.

MSSubClass and HouseStyle have very similar data definitions. In fact, I removed MSSubClass all together to see what features would be considered important without MSSubClass and HouseStyle took its place. It’s clear that the number of floors and style of the house has an impact on price. However, each feature has quite a few categories and there is a relationship between the two variables. The two variables also describe whether the home is finished or unfinished and the age of house. Both of these factors are addressed elsewhere in the data set. As a result, I decided to create separate variables from MSSubClass and HouseStyle. In the end, I decided to capture number of floors (NumOfFloors) and SplitHome (binary operator) from the information in the two original features. When I ran the top-8 variables that included the MSSubClass, I got an adjusted R-squared of .8537. When I excluded MSSubClass all together and HouseStyle was in the top-8 variables, the adjusted R-squared was around .82. Which the new variables, adjusted R-squared is .8496.

### Feature Selection

To identify the features that should be included in the modeling processed, I used recursive backwards elimination. As a result of applying this method, I received the following plot:



This plot described the diminishing value of additional variables on root mean squared error. It appears between the 8-10 variable range; additional variables stop having a major impact in reducing RMSE. The top 8 variables are as follows:

> predictors(results)

[1] "TotalSF" "OverallQual2" "YearBuilt" "LotArea" "GarageCars"

[6] "KitchenQual" "FireplaceQu" "GarageType"

# Modeling

Before diving into model, I first need to do a few steps:

1. Divide the Sale Price Feature into a binary variable. Since the 3rd quartile (75%) of the houses are less than $214,000, I will assign each house in the test set above this price as TRUE (or 1). If the observation represents a house below this amount, I will set it to false.
2. Calculate the base rate. If the model guessed FALSE every time the accuracy of the model will be 75.3% accurate.

My method for modeling seeks to improve upon an assignment previously applied to the dataset prior to using feature engineering and feature selection approaches.

## Decision Tree

A decision tree model seeks to achieve purity by separating features that optimizes information gain. Decision trees are at risk for overfitting the training data because the patterns in the training data can vary between data sets. That is, one sample data set can create an entirely different decision tree if it doesn’t represent the entire population well. This causes an increase variance between the train and test datasets.

Prior to applying feature engineering and feature selection approaches, the follow accuracy measurements were produced:

|  |  |
| --- | --- |
|  | ACCURACY |
| TRAIN | .9085851 |
| TEST | .8861386 |
| VARIANCE: | .0224465 |

After applying feature engineering and feature selection approaches, the accuracy measures are:

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACCURACY | PRECISION | RECALL |
| TRAIN | .9203822 | 0.9189744 | 0.9770992 |
| TEST | .9257426 | 0.9491525 | 0.9655172 |
| VARIANCE: | .0053604 |  |  |

As you can see, the variance and accuracy results were both improved by the feature engineering and feature selection techniques.

Confusion Matrix based on classifying a Top 25 house when the probability is greater than 75%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | No | Yes |
| No | 896 | 21 |
| Yes | 79 | 19 |

Decision Tree:

|  |
| --- |
| 1) root 1256 247.502400 0.26990450  2) TotalSF< 2929.5 876 49.793380 0.06050228  4) TotalSF< 2613 699 8.884120 0.01287554 \*  5) TotalSF>=2613 177 33.062150 0.24858760  10) KitchenQual=Poor,Fair,Average 72 4.652778 0.06944444 \*  11) KitchenQual=Good,Excellent 105 24.514290 0.37142860 \*  3) TotalSF>=2929.5 380 70.747370 0.75263160  6) KitchenQual=Poor,Fair,Average 84 15.750000 0.25000000 \*  7) KitchenQual=Good,Excellent 296 27.753380 0.89527030  14) OverallQual2=Below Average,Average 116 21.241380 0.75862070  28) LotArea< 8604 15 3.333333 0.33333330 \*  29) LotArea>=8604 101 14.792080 0.82178220 \*  15) OverallQual2=Above Average 180 2.950000 0.98333330 \* |

## Decision Tree with Bagging

Bagging can be applied to a decision tree model in order to limit the risk of overfitting. Bagging involves gathering bootstrap samples from the training data and applying a decision tree model to each sample. The final decision tree model is the average of all the bootstrapped sample decision trees compiled during this process.

Prior to applying feature engineering and feature selection approaches, the follow accuracy measurements were produced:

|  |  |
| --- | --- |
|  | ACCURACY |
| TRAIN | .9340223 |
| TEST | .8712871 |
| VARIANCE: | .0627352 |

After applying feature engineering and feature selection approaches, the accuracy measures are:

|  |  |  |  |
| --- | --- | --- | --- |
|  | ACCURACY | PRECISION | RECALL |
| TRAIN | .9306931 | .9516729 | .7551622 |
| TEST | .933121 | .7826087 | .6428571 |
| VARIANCE: | .0024279 |  |  |

As you can see, the feature engineering and feature selection approach significantly improved the accuracy of the model along with the variance between the test and train sets.

Confusion Matrix based on classifying a Top 25 house when the probability is greater than 75%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | No | Yes |
| No | 904 | 13 |
| Yes | 83 | 256 |

Decision Tree:

1) root 1256 235.5000000 0.250000000

2) TotalSF< 2922.5 881 35.4460800 0.041997730

4) YearBuilt< 2007.5 869 28.0322200 0.033371690

8) TotalSF< 2539.5 671 0.9985097 0.001490313 \*

9) TotalSF>=2539.5 198 24.0404000 0.141414100

18) LotArea< 18680.5 187 18.6417100 0.112299500 \*

19) LotArea>=18680.5 11 2.5454550 0.636363600 \*

5) YearBuilt>=2007.5 12 2.6666670 0.666666700 \*

3) TotalSF>=2922.5 375 72.3893300 0.738666700

6) KitchenQual=Poor,Fair,Average 85 15.8117600 0.247058800

12) LotArea< 11959.5 42 0.9761905 0.023809520 \*

13) LotArea>=11959.5 43 10.6976700 0.465116300

26) FireplaceQu=No Fireplace,Poor,Fair,Average 23 2.6086960 0.130434800 \*

27) FireplaceQu=Good,Excellent 20 2.5500000 0.850000000 \*

7) KitchenQual=Good,Excellent 290 30.0137900 0.882758600

14) LotArea< 8587.5 31 7.7419350 0.516129000

28) LotArea>=7916 13 1.6923080 0.153846200 \*

29) LotArea< 7916 18 3.1111110 0.777777800 \*

15) LotArea>=8587.5 259 17.6061800 0.926640900 \*

## Feature Importance with Random Forest

A random forest model uses subsets of the features in order to determine feature importance. If the inclusion/exclusion of the features has a considerable impact on the outcome of the decision tree, the feature is considered to be important.

Here we can see that Total Square Footage is the more important feature due. When this feature was excluded from the decision tree model, the mean decrease in accuracy was about 25%.

A screenshot of a cell phone

Description automatically generated

# Evaluation

Based on the accuracy, precision, and recall rate, I would select the decision tree with bagging approach over just the decision tree approach. There are a couple reasons for this:

1. Though the decision tree without bagging had a higher precision rate, I am suspicious of this model. Decision trees tend to overfit the training data and have a hard time generalizing to new data. The bagging method seeks to improve on this behavior for the decision tree.
2. A lower precision rate for the decision tree with the bagging method could simply be a more realistic statistic due to generalizing patterns in the entire Ames, IA housing market instead of just the training set.
3. The accuracy between the training and test sets for the decision tree with bagging have very little variance. This makes me believe that we would see similar results when applied to new data.

The next steps I would take to improve on this model would be:

1. Optimizing parameters in the classification to maximize precision and recall rate
2. Reviewing the decision tree in more detail for overfitting
3. Perhaps find another way to factor in Neighborhood. It lost it feature importance when put the different neighborhoods into bins.
4. Fireplace Quality seems to be an odd feature with high importance since many houses do not have a fireplace. I think this feature is correlated with another feature does not present in this analysis. Some additional research may uncover a better predictive feature that explains this correlation.