Project ID for SYDE 522 – Winter 2018

Project Title:	House Prices: Advanced Regression Techniques
Project Member(s):	Johnson Kan (20270951), Sarah Watts (20515933)
Summary of the Project:	The purpose of the study is to prove an ensemble model will perform better than any individual models for the Ames Housing dataset. The dataset will be preprocessed via cleaning, feature engineering, encoding and normalization. Six individual models will be built from the same dataset and the models will be evaluated through an inactive Kaggle competition. The six models in this project were Random Forest, Lasso Regression, Extreme Gradient Boosting, Support Vector Machine, Bagging and Bayesian Linear Regression. The best individual model was Lasso Regression with a root squared logarithmic error of 0.130. The ensemble model, which uses an average of equal weights for the six models has the best result with a root squared logarithmic error of 0.128.
Data Used:	There are 79 features to predict the sale price of homes in the Ames Housing dataset. Some of the interesting features are: type of road access, lot size, overall condition rating, type of roof, type of foundation, height of the basement, central air conditioning, full baths, half baths, pool area… etc. The size of the training and testing data is about 450 kbs each. There's 1460 entries in each file.
Source of Data Used:	The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the frequently cited Boston Housing dataset.
Results Achieved:	Ensemble provides the best results for this data set with a score of root mean squared logarithmic error of 0.128.

House Price Prediction: Ensemble

Johnson Kan, Sarah Watts
Department of Management Sciences, University of Waterloo
j2kan@edu.uwaterloo.ca, smwatts@edu.uwaterloo.ca

Abstract – The purpose of the study is to prove an ensemble model will perform better than any individual models for the Ames Housing dataset. The dataset will be preprocessed via cleaning, feature engineering, encoding and normalization. Six individual models will be built from the same dataset and the models will be evaluated through an inactive Kaggle competition. The six models in this project were Random Forest, Lasso Regression, Extreme Gradient Boosting, Support Vector Machine, Bagging and Bayesian Linear Regression. The best individual model was Lasso Regression with a root squared logarithmic error of 0.130. The ensemble model, which uses an average of equal weights for the six models has the best result with a root squared logarithmic error of 0.128.

1 Introduction

The dataset analyzed is this paper is taken from the Ames Housing dataset, which was compiled by Dean De Cock for use in data science education. This dataset is a strong alternative for the commonly cited Boston Housing dataset, which can be considered a modernized and expanded version of housing price prediction.

This dataset is accessed through Kaggle, where a training and testing set have been provided. The goal of this competition is to predict the sale price for each house. The accuracy of the model is judged based on the Root- Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. This metric ensures the price of the home (expensive or otherwise) will not bias the model.

2 Dataset Description

This dataset contains 79 features and the numeric responds variable sale price. The features contained in the data are both categorical and numeric.

The categorical features include both nominal and ordinal features. There are 23 nominal features that include information such as the type of dwelling, street access and other physical characteristics of the property (flatness, shape and lot configuration). There are also 23 ordinal features that include information such as the ratings for the property (overall condition, heating condition, condition of basement).

The numeric features include both continuous and discrete features. The 19 continuous features include information pertaining to home dimensions (lot size, square feet of the house, square feet of the basement). Additionally, there are 14 discrete variables that include information pertaining to the number of items that occur within a household (number of bathrooms, number of bedrooms).

3 Exploratory Data Analysis

3.1 Response Variable: Sales Price

The response variable Sales Price ranges from $34,900 to $755,000 with a median price of $163,000. To understand this variable, it is important to consider the median price. This is because the data has a right skewed distribution (Figure 1), which is expected because few homes sell in the higher price ranges.

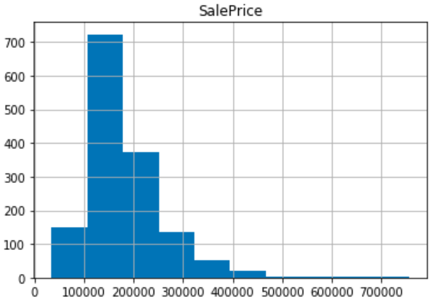

Figure 1. Sales Price Histogram

3.2 Correlations to Sales Price

The correlation between the numeric features and the response variable Sales Price was investigated. The feature with the strongest correlation to sales price (at 79%) is overall quality. Additionally, there are 10 features that have a correlation where there is a correlation that is higher than 50% (Figure 2).

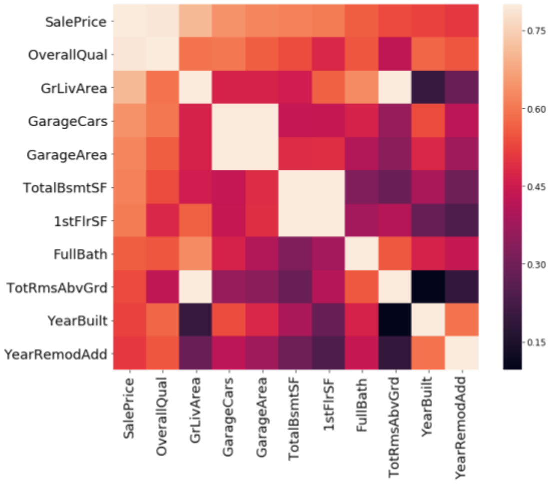

Figure 2. Correlation Matrix

4 Data Cleaning

4.1 Missing Values

In the dataset provided there are 34 features that contain missing values. Replacing these missing values was approached differently based on if the feature was categorical and numeric. As a general rule, categorical features with missing values were replaced based on taking the most commonly occurring value for that feature. Additionally, as a general rule, numeric features with missing values were replaced using the median of the feature, as many features experiences a skewed distribution.

This approach works if there a particular record only has a few missing features. For example, if the alley way type was not filled in then it would be replaced by the most common type, which is paved. However, there were a few records out of the 2,900 that missed multiple features. When these records were analyzed, it was determined that these features were typically missed entire feature groups. A group of features in this instance pertain to a section of a home, for example the basement, pool, garage etc. If all 9 basement features were missing, it was deduced that the home did not contain a basement. In this case, it made more sense the replace numeric features with the value of zero and categorical features with the value of None.

The final dataset has a value for each record and feature. These values were determined back on the number of features missing for a feature group. If a record contained most values for a feature group than median value was used as a replacement, otherwise a zero was inputted.

4.2 Numeric Encoding

To further analyze the dataset, all possible features will be numerically encoded. 23 ordinal categorical features have been encoded. These features have been assigned values based on their corresponding scale. For example, the feature pool condition is rated from “poor” to “excellent”, which have been encoded on a scale of 1 to 5 and if a property is missing a pool, it was encoded with 0.

The final features that will be converted to numeric values are the 23 nominal categorical features. They will be one-hot encoded as part of preprocessing before a model is run. For now, they will remain as categorical features to continue the data cleaning.

4.3 Removing Redundant Features

To remove redundant features, two methods were utilized. The first, is by removing features that have a low correlation with the sales price respond variable. This method works for numeric features. The second method is to graph all features using a histogram. This is helpful for understanding nominal categorical variables, where features without a distribution of values will be removed because this does not add new information into a model.

A good example of removing redundant features due to little correlation with the response variable is the rating for a rating for basement finishes. The correlation in this case is -0.002332, meaning that there is almost zero correlation between this feature and the response variable that is being predicted.

Additionally, a histogram of this feature indicates that approximately 90% of the values are concentrated at the value of 1 (Figure 3). This is a second indicated that this feature will not be beneficial in the creation of this model.

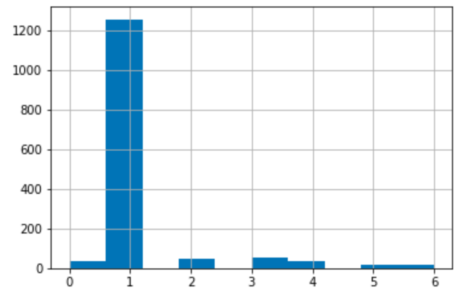

Figure 3. Histogram of Basement Finishes Feature

This approach was utilized to remove 20 redundant features from the model. These features included attributes describing the homes such as the type of pavement in the driveway or the slope of the land. These were found not to be significant indicators for where a home buyer sees value.

4.4 Removing Highly Correlated Features

Highly correlated features can be removed because they do not improve the performance and can add noise.

Based on a correlation matrix of the features, highly correlated features will be removed, where the feature with higher predictive power will be retained.

To determine which features are highly correlated, a correlation matrix has been created. For example, the features fireplaces and fireplace quality are 86% correlated. These features are 47% and 52% correlated with the response variable sales price, respectively. Therefore, the feature fireplaces will be removed from the model and the feature fireplace quality will be retained.

Table 1. Correlated Features Removed

Feature	Description
Fireplaces	Number of fireplaces (highly correlated with fireplace quality)
BsmtFinType1	Rating of basement finished area (highly correlated with basement square feet)
TotRmsAbvGrd	Total rooms above ground (highly correlated with above ground living square feet)

5 Feature Engineering

5.1 Engineering by Grouping

To determine if there were any features that could be engineered to improve their predictive effectiveness, like features have been grouped and analyzed after data cleaning. As a result, four key groups of features have been identified and will be discussed in this section.

5.11 Quality/Condition Groupings

Many areas of a home contain both a quality and a condition rating. These ratings are not always correlated, additionally they individually may not have a strong correlation with the response variable sales price.

As a result, these numeric features have been added together to provide an overall rating for each location for a home. This has been applied to the garage condition and quality, the basement condition and quality and the overall condition and quality. This combined rating feature provides more information than the individual quality & condition score and has replaced each pair of features.

5.12 Square feet Grouping

The square feet of a home is another example of a beneficial grouping of features. There are two features that describe the square feet of a home, the total area above ground and the basement square footage. Both features are in the same units and combined form the square feet of the home. This feature better describes the useable square feet in a home and are a better indicator of the value a home buyer is willing to pay for.

5.13 Bathroom Grouping

There are four features that describe the number of bathrooms in a home. There is a feature for full baths and half baths above ground, as well as full baths and half baths below ground. Individually, these features are varying levels of importance to the response variable. Combined, these features tell a more variable information about the size of a home and the number of baths a buyer is looking to pay for.

To ensure that this aggregated feature is a proper reflection of the number of baths, all full bathrooms were multiplied by 1 and all half baths were multiplied by 0.5 before being added to the aggregate. Using this new aggregated value, there is a 63% correlation between the response variable and the number of bathrooms.

5.14 Porch Grouping

There are five different features that are used to describe the porch/deck areas of a home. These include the square feet of a wood deck, open porch, enclosed porch, three season porch and screened in porch. When analyzing the correlation between these variables and the sales price of a home, there is a large discrepancy between the 52% correlation of the wooden deck and the >5% correlation of the remaining porch features. As such, it appears that the wooden deck is a meaningful metric for home buyers, whereas the porch metrics are not.

To create a more meaningful feature, the square feet of all porch areas has been consolidated. This allows for a more meaningful description of the porch area. The original four porch area features have been removed from the dataset, and only the wooden deck and total porch area features remain.

5.2 Engineering Home Age

Four features in the dataset point to the age of a home, without a feature that explicitly defines it. Understanding the age of a home is very important when understanding what drove the appeal of the home for buyers. Naturally, a newer home would be expected to sell at a higher price point than an older home. The four features are the year the home was built, the year the home was remodelled, the year the home was sold and the age of the garage.

After examination of the dataset, an age of the home can be determined by taking the year the home was sold and subtracting the year the home was remodelled. Homes without a remodel have the same value for both the feature of year the home was built and year of remodel.

To determine which features to retain after the addition of the house age feature, two key ideas were considered. First, there had to be an introduction of a variable for homes with and without remodelling. If a home was partial remodelled, this could not be considered an entirely "new" home. Therefore, the feature is remodelled has been introduced. Second, all features that will not provide new information to a model must be removed. Therefore, the features home age and is remodelled will be retained and the four original home age features (year the home was built, the year the home was remodelled, the year the home was sold and the age of the garage) will be removed.

7 Preprocessing

7.1 One hot encoding

Multiple categorical variables were processed with one hot encoding. One hot encoding is a process in which categorical variables are converted into a form that will allow machine learning algorithms to perform better. For example, a categorical variable of gender will be encoded by replacing the gender column with two columns of isMale and isFemale. All one hot encoded columns have a binary representation with 1s and 0s.

7.2 Normalize Features

All numeric features were preprocessed with feature scaling, a method to standardize the range of features in the dataset. This is a necessary step for many machine learning algorithms because their objective functions depend on the features have the same range of values. For example, Euclidean distance is a popular mechanism for machine learning and in this dataset without normalization the total square footage ranges hundreds to thousands whereas the number of rooms will have ranges in the teens causing the two features to be skewed in results.

7.3 Adjust for Skewness

After data exploration, it was determined that the sales price experienced a right skewed distribution. Analyzing the skewness of the dataset, at 1.877 and the Q-Q plot (Figure 4) confirm this hypothesis.

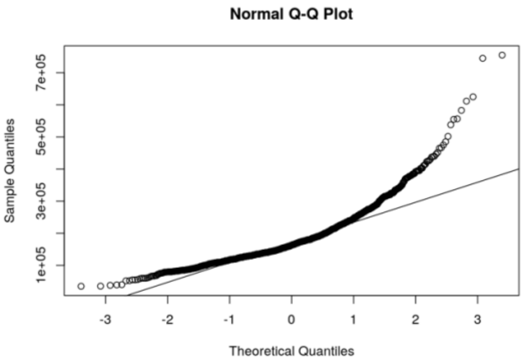

Figure 4. Skewed Sale Price Q-Q Plot

To remove bias associated with the skew of the response variable, the natural log of the sale price was taken. The resulting skewness is 0.12 and the Q-Q plot indicates a near normal distribution (Figure 5).

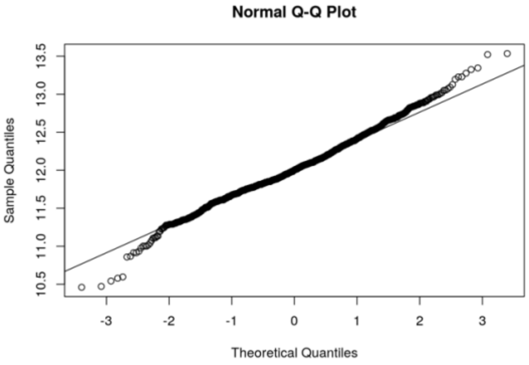

Figure 5. Normalized Sale Price Q-Q Plot

8 Models

8.1 Random Forest

Random forest uses many trees instead of growing a singular tree and each tree. The trees are likely to be different from each other as the tree is grown by sampling the training data with replacement. For this regression problem of predicting the sale price, the predicted value is a weighted average of the value predicted from each tree.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.147.

8.2 Lasso Regression

Lasso, which stands for least absolute shrinkage and selection operator, is a regression method that performs L1 regularization. The Lasso method will set features that are not relevant to zero. Therefore, the model is simpler with less features being used.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.130.

8.3 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) has an objective function that minimizes training loss and regularization. XGBoost is an implementation of gradient boosted decision trees with run time performance in mind, which is a primary reason for why people use XGBoost. Although XGBoost uses trees, it is very different from Random Forest. XGBoost will sequentially builds models and punishes more heavily on observations mistaken by previous models, whereas Random Forest will build ensemble models in parallel.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.139.

8.4 Support Vector Machine

Support vector machine for regression implemented used a non-linear kernel function that transforms the data into higher dimensional feature space. Non-linear kernel makes it possible to have a linear separation in support vector machine.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.133.

8.5 Bagging

Bagging, also known as bootstrap aggregating, is designed to improve stability, reduce variance and help avoid overfitting. Bagging works by resampling a training set to create multiple training sets. These individual training sets are then used to independently build a tree. The average outputs of these trees are used in regression.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.183.

8.6 Bayesian Linear Regression

Bayesian linear is a variation of linear regression with the statistical analysis in the context of Bayesian inference. This means data is supplemented with more information in the form of a prior probability distribution.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.133.

8.7 Ensemble

Ensemble is a method that uses multiple learning algorithms in one. The result is usually a better predictive model than any one individual model. Ensemble tends to perform better when there is a significant diversity among models. Therefore, the following models were used to build the ensemble models: Random Forest, Lasso Regression, Extreme Gradient Boosting, Support Vector Machine, Bagging and Bayesian Linear Regression.

The model was used to predict the testing dataset provided by Kaggle. The root mean squared logarithmic error of this model is 0.128, which is the best accuracy score observed across all models.

9 Conclusion

The hypothesis for this project was to prove an ensemble model will perform better than any one individual model. To prepare the data dataset for the models many steps occurred. First, the exploratory data analysis was conducted to understand each of the 79 features and the response variable, sale price. This involved plotting the features and analyzing properties such as distribution type, skewness etc. Next, data cleaning occurred to add values for the 35 features with missing values, to convert all possible features to numeric values and to remove redundant and highly correlated features. Third, feature engineering occurred to create more meaningful features from the dataset. For example, adding the age of a home as a feature. Finally, preprocessing occurred which was used to encode and normalize features.

To give the ensemble model the best chance of success, diverse machine learning models must be used. To test and measure the accuracy of the models, each model was to predict the sale price of the testing data and then submitted to Kaggle for a root mean squared logarithmic score, the lower the better. The best score from the six individual models was 0.130 via Lasso Regression and the worst score was 0.183 via Bagging. The combination of the six models: Random Forest, Lasso Regression, Extreme Gradient Boosting, Support Vector Machine, Bagging and Bayesian Linear Regression yielded a score of 0.128. Therefore, for this Kaggle competition, the 7th model, an ensemble model which is takes the average sale price predictions of the six-individual model performed better than any of the individual models.