Final Technical Report
Housing Prices Prediction in Iowa

Abstract

This study addresses the topic of housing prices predictions in the city of Ames, Iowa, with the purpose of using multiple regression and classification algorithms to predict these values with a reasonable accuracy, which would help individuals looking to buy a house and investors evaluate prices and make decisions. The design of this study starts with carefully studying the data, reading the data dictionary, perform the Exploratory Data Analysis, present meaningful questions whose answers can be obtained from the models, eliminating the unnecessary features before building the models and feeding the data to those models, and finally using those tuned models to answer the main question of the case study. After multiple attempts, the models have successfully given predictions with the desired reasonable accuracy. The different models accuracy ranged from 66% to 86% for regression models that aimed to predict the houses sale prices. Furthermore, multiple feature selection approached were introduced to ensure feature importance, in addition to an unsupervised clustering methods used to single out outliers.

Key Words: Regression Models; Multivariate Classification Models, Feature Selection, Machine Learning, Clustering.

Introduction

Context

This study addresses the topic of the sale of residential property in Ames, Iowa from 2006 to 2010. Upon the completion of the models, new houses sale prices can be estimated using multiple regression and classification models with a reasonable accuracy. Multiple model evaluation approaches will also be introducing to ensure the reliability of the predicted prices.

The choice of this dataset for this project was a result of multiple considerations; this dataset is a real-life problem, while a lot of people consider buying a house at some point of their lives, a large number of them don't have any experience assessing houses and the associated prices. Moreover, the comprehensive dataset makes it possible to use multiple approaches and algorithms before reaching the final version of the prediction model. The number of features helps improve the model accuracy through using multiple feature selection approaches that select the most important features to work with, while omitting others that are less important. The models were built using python 3.5 with the aid of the *Sklearn* package, and this dataset was originally compiled by *Dean De Cock*[1] for the purpose of teaching students regression models, machine learning, and data science, and was retrieved from *Kaggle.com*[2]

The Research Question

As previously mentioned, this dataset has a number of advantages; being a real life scenario that's using real and recent house market data in the United States. In order to reach a meaningful question that we try to answer through these models, we first needed to explore the data, variables, and patterns. Most of the variables focus on the quality and quantity of the house attributes. Those attributes are essential for someone who's considering buying a house to be able to assess the house overall quality, and whether or not it suits their needs. Some of those attributes are the sizes of different floors of the house, whether or not the house has a basement, the overall quality of the house as a value from 1 to 10, the age of the house and when was it last remodeled, the number of floors, rooms, and bathrooms in the house, etc.

The most important information that can be obtained from this study is the house prices. Therefore, the main research question of the project is: Can we predict the value of the house price given the associated features? A number of regression models was introduced to answer this question with a reasonable accuracy. Following that, it has been decided that this problem could also be transformed into a classification problem by mapping the houses prices into three categories: Low Price, Medium Price, and High Price, making the target a multiclass value rather than a continuous.

Methodology

The methodology of reaching an answer to the previous question was as follows: first the Data Dictionary was carefully studied to get familiar with the variables in the dataset and what they represent. Following that, an Exploratory Data Analysis was performed; this step is crucial to get an insight into what the data is representing, and what question would be stated that can be answered using the models. The Exploratory Data Analysis also includes checking for null values, plotting some of the most important variables against one another and against the price column, etc. The main research question was then stated to help orient the work flow in the right directions (i.e. choose the appropriate models). The prediction models and the feature selection approaches to be used have been decided thereafter. The successful completion of those tasks marks the readiness for the models to be built and tuned. For each case, a model was built, trained, and validated using multiple approaches.

Data Dictionary

This dataset has 2930 observations and 80 explanatory variables; 23 nominal (binary or multinomial; types and conditions of house/garage, e.g. street is gravel or paved?), 23 ordinal (binary or multinomial; rate various items, e.g. overall quality of the house), 14 discrete (related to the number of items in the house; how many kitchens, bathrooms, rooms, floors), and 20 continuous (related to area dimensions; sq. footage, basement, living area measurements.

Some of the most important features are: **MSSubClass**: The building class, **LotArea**: Lot size in square feet, Street: Type of road access, Neighborhood: Physical locations within Ames city limits, **OverallQual**: Overall material and finish quality, **YearBuilt**: Original construction date, **YearRemodAdd**: Remodel date, **BsmtQual**: Height of the basement, **BsmtCond**: General condition of the basement, **GarageType**: Garage location, **SaleCondition**: Condition of sale, and **SalePrice** - the property's sale price in dollars. This will be the target in the modelling stage.

Data Transformation and Dummy Variables

In order to fit the linear regression model, the assumption of normal distribution has to hold. In this data set, the distribution of **SalePrice** is skewed to the left. After the log transformation, however, the normal distribution is satisfied. In addition, the standardizer and normalizer from sklearn package were applied to the continuous features.

Exploratory Data Analysis

In this part, the relationship between multiple variables were obtained, primarily the price, age of house, year of remodeling (if any), the overall quality of the house, the living area above ground in square feet, the area of garage, and the total number of rooms.

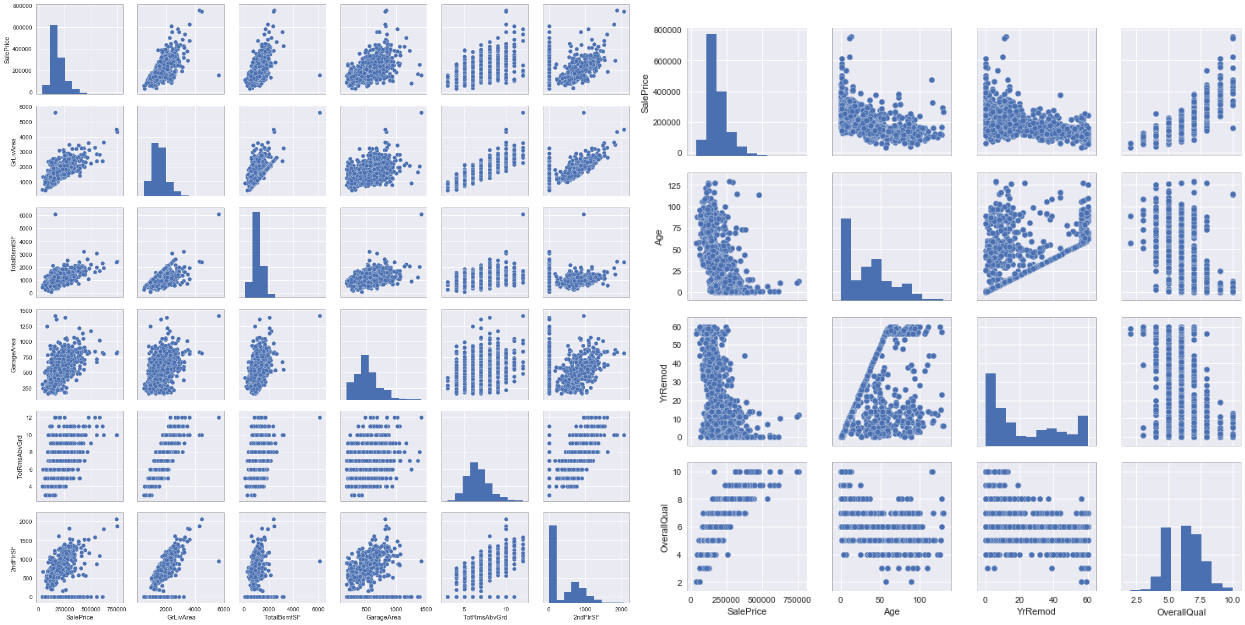


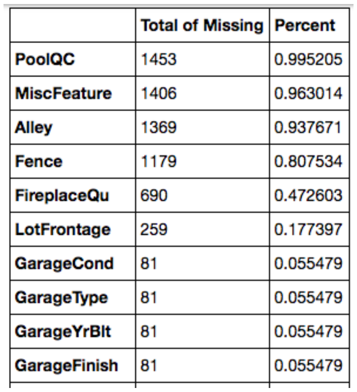
Figure 1: Exploratory data analysis

From *figure 2*, it can be observed that:

1- The Sale Price increases with the increase of:
   - Sq. footage of house
   - Basement footage
   - Garage area
   - Number of rooms
   - Second floor area
2- The bigger the living area, the bigger the basement
3- No or little correlation between living area and garage area
4- No or little correlation between house price and age & year of remodeling!!
5- The better the overall house quality, the higher the selling price

In addition, it was important to know how many empty values are there in the dataset, as well as the percent of the total (for example, in the PoolQC, 99.5% of the records don't have a value). This is helpful as it gives us an idea early on regarding what variables are less important than other (figure 2)

Finally, a correlation plot was plotted to see how different variables are correlated with each other. Correlation plots are useful as they help reduce data redundancy if two variables carry the same information, they can be replaced with only one, thereby reducing the number of dimensions.



	Total of Missing	Percent
PoolQC	1453	0.995205
MiscFeature	1406	0.963014
Alley	1369	0.937671
Fence	1179	0.807534
FireplaceQu	690	0.472603
LotFrontage	259	0.177397
GarageCond	81	0.055479
GarageType	81	0.055479
GarageYrBlt	81	0.055479
GarageFinish	81	0.055479

Figure 2: Missing values in different columns

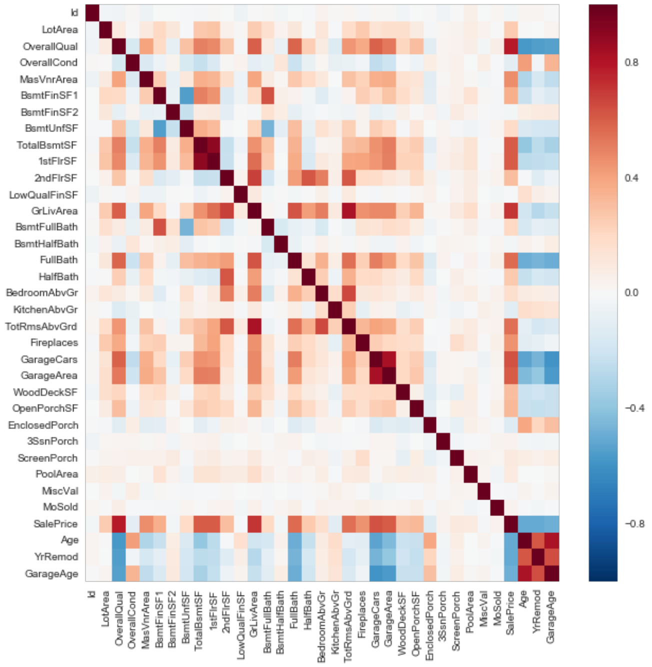


Figure 3: Correlation Plot

Features and Targets

As mentioned before, this dataset has over 80 variables describing different attributes of the sold house. For the regression models, those attributes are all considered features, except for the house price, which will be the target that we try to predict. However, since this study includes a classification problem where the target (sale price value) was replaced into categorical variables, the old sale price variable was dropped from the dataset, and Sale Class was used as the target instead (value to predict). The criteria used for creating this new multi-class target was using the 25th and 75th percentiles as the threshold for defining the end of one price class and the beginning of the other;

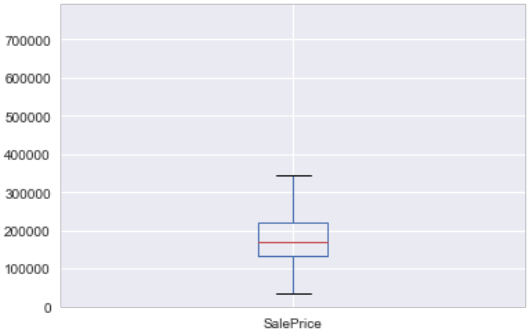


Figure 4: Box plot showing target ranges

- Low =< 135000
- 135000 < Medium < 22000
- High >= 220000

Feature Selection

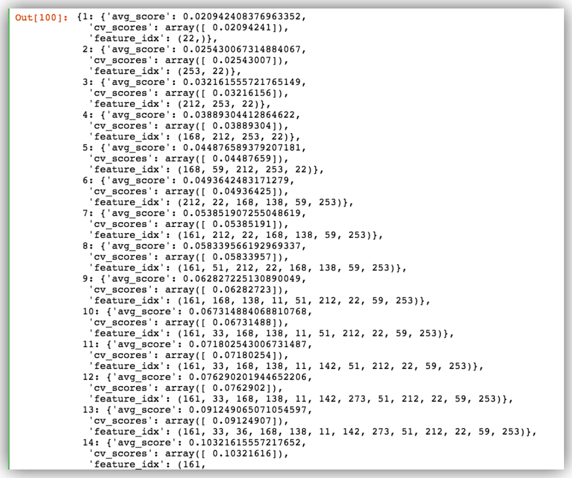
1- Mutual Information

Deciding on which features to work with in regression models is often necessary to get the desired accuracy and reliable predictions. With 79 features between categorical and continuous variables representing different attributes of the house, using a feature selection algorithm was crucial to reduce that number to a reasonable number that would give a good accuracy. The Mutual Information approach was chosen for this study with the results of the most important features confirmed by the findings of the Random Forest Algorithm. The Mutual Information approach uses a measure of the mutual dependence between two variables; in this case each feature, including dummy variables, and the house sales price (target). The output of this approach is a score ranging between 1 (highest score indicating most important variables) and 0 (lowest score indicating least important variables).

The results of this approach suggested that the overall quality of the house is the most important variable (have the largest amount of mutual information with the Sale Price target), followed by the total area of the house's above ground living area, the basement area, the garage area, and the age of the house.

2- Sequential Feature Selection – Logistic Regression

The next step was to confirm those findings with another dimensionality reduction approach, namely the Stepwise Feature Selection Logistic Regression model. In this mode, the algorithm starts with an empty dataset, adding one feature every trial that would optimize the performance of the model. Like in the Mutual Information, the first 14 features were obtained using the forward propagation, with no floating, and cv of 0 (indicating that no cross validation will be performed). However, those 14 features were different than those resulting from the Mutual Information Method. It was also noted that while the Mutual Information output was very logical (Overall Quality of the house is the most important, followed by the Living Area of the house, Garage Area), the importance of the resulting 14 features from the Stepwise feature selection was less logical. Therefore, it was decided to confirm the results using a third dimensionality reduction approach using Random Forest.



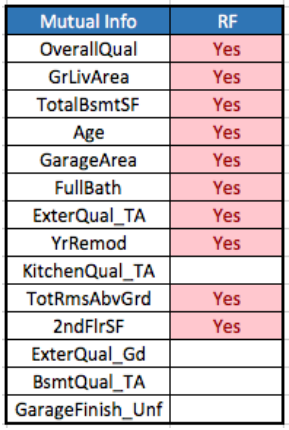
```
Out[100]: {1: {'avg_score': 0.020942408376963352,
            'cv_scores': array([ 0.02094241]),
            'feature_idx': (22,)},
           2: {'avg_score': 0.025430067314884067,
            'cv_scores': array([ 0.02543007]),
            'feature_idx': (253, 22)},
           3: {'avg_score': 0.032161555721765149,
            'cv_scores': array([ 0.03216156]),
            'feature_idx': (212, 253, 22)},
           4: {'avg_score': 0.038893044128646222,
            'cv_scores': array([ 0.03889304]),
            'feature_idx': (168, 212, 253, 22)},
           5: {'avg_score': 0.044876589379207181,
            'cv_scores': array([ 0.04487659]),
            'feature_idx': (168, 59, 212, 253, 22)},
           6: {'avg_score': 0.049364248317127 9,
            'cv_scores': array([ 0.04936425]),
            'feature_idx': (212, 22, 168, 138, 59, 253)},
           7: {'avg_score': 0.053851907255048619,
            'cv_scores': array([ 0.05385191]),
            'feature_idx': (161, 212, 22, 168, 138, 59, 253)},
           8: {'avg_score': 0.058339566192969337,
            'cv_scores': array([ 0.05833957]),
            'feature_idx': (161, 51, 212, 22, 168, 138, 59, 253)},
           9: {'avg_score': 0.062827225130890049,
            'cv_scores': array([ 0.06282723]),
            'feature_idx': (161, 168, 138, 11, 51, 212, 22, 59, 253)},
           10: {'avg_score': 0.067314884068810768,
            'cv_scores': array([ 0.06731488]),
            'feature_idx': (161, 33, 168, 138, 11, 51, 212, 22, 59, 253)},
           11: {'avg_score': 0.071802543006731487,
            'cv_scores': array([ 0.07180254]),
            'feature_idx': (161, 33, 168, 138, 11, 142, 51, 212, 22, 59, 253)},
           12: {'avg_score': 0.076290201944652206,
            'cv_scores': array([ 0.0762902]),
            'feature_idx': (161, 33, 168, 138, 11, 142, 273, 51, 212, 22, 59, 253)},
           13: {'avg_score': 0.091249065071054597,
            'cv_scores': array([ 0.09124907]),
            'feature_idx': (161, 33, 36, 168, 138, 11, 142, 273, 51, 212, 22, 59, 253)},
           14: {'avg_score': 0.10321615557217652,
            'cv_scores': array([ 0.10321616]),
            'feature_idx': (161,
```

Figure 5: Sequential Feature Selection output using 14 features

3- Random Forest Feature Importance

One of the outputs that the Random Forest Algorithm gives is the dataset Features Importance, which is a list of all the features in the dataset (including the dummy variables) with the associated importance indicating the influence on the target variable; the house price. It can be noted that out of the most important 14 features in the Mutual Information, 10 are in common with the random forest 14 most important features, which implies a good accuracy of the Mutual Information approach.

To conclude this part, the conformity between the Random Forest Feature Importance and that of the Mutual Information has been considered while choosing the final 14 features to work with while the findings of the Stepwise Feature Selection were ruled out.



Mutual Info	RF
OverallQual	Yes
GrLivArea	Yes
TotalBsmtSF	Yes
Age	Yes
GarageArea	Yes
FullBath	Yes
ExterQual_TA	Yes
YrRemod	Yes
KitchenQual_TA	
TotRmsAbvGrd	Yes
2ndFlrSF	Yes
ExterQual_Gd	
BsmtQual_TA	
GarageFinish_Unf	

Figure 6: Top 14 features in common using MI and RF

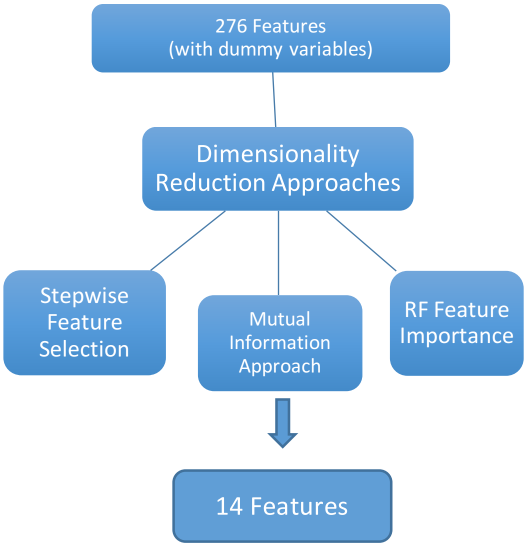

276 Features
(with dummy variables)
Dimensionality Reduction Approaches
Stepwise Feature Selection
Mutual Information Approach
RF Feature Importance
14 Features

Figure 7: Feature Selection Diagrams

Regularization

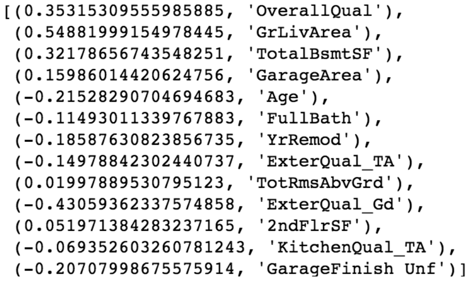
The LASSO and the Elastic Net were unutilized to test the number of selected features (n = 14) from Mutual Information approach, also the cross-validation was applied. The result from two methods were similar: the mean-square-error (MSE) on test set from LASSO is 0.0285, and R^2 is 78.78%; in Elastic Net, the MSE is 0.0284 and the R^2 is 78.89%. Moreover, in Elastic Net, the coefficient of one feature **BsmtQual_TA** was shrinking to 0, so this feature was deleted before model building. In sum, 13 features (**OverallQual, GrLivArea, TotalBsmtSF, GarageArea, Age, FullBath, YrRemod, ExterQual_TA, TotRmsAbvGrd, ExterQual_Gd, 2ndFlrSF, KitchenQual_TA,** and **GarageFinish_Unf**) were selected for regression modeling.

Prediction Models

Linear Regression

The first introduced regression model in this study is linear regression. Using **SalePrice** as the target, and 13 features as the predictors, the model was trained and tested.

The **OverallQual, GrLivArea, TotalBsmtSF and GarageArea** are having the positive effect on the houses prices (target), while **Age** and **FullBath** show negative effect, which means that with the increase of these features, the lower the house price goes.



```
[(0.35315309555985885, 'OverallQual'),
 (0.54881999154978445, 'GrLivArea'),
 (0.32178656743548251, 'TotalBsmtSF'),
 (0.15986014420624756, 'GarageArea'),
 (-0.21528290704694683, 'Age'),
 (-0.11493011339767883, 'FullBath'),
 (-0.18587630823856735, 'YrRemod'),
 (-0.14978842302440737, 'ExterQual_TA'),
 (0.01997889530795123, 'TotRmsAbvGrd'),
 (-0.43059362337574858, 'ExterQual_Gd'),
 (0.05197138428323716, '2ndFlrSF'),
 (-0.069352603260781243, 'KitchenQual_TA'),
 (-0.20707998675575914, 'GarageFinish Unf')]
```

Figure 8: Features Coefficients in Linear Regression

The overall R^2 is 0.787 and the MSE on test set is 0.0286. The cross-validation was used to test the over-fitting, with a Mean Square Error of 0.031.

Next, the plot of the actual and the predicted value, and residual plot were applied for the model diagnosis.

From the plot above, the relationship of the actual and the predicted price are linearly related.

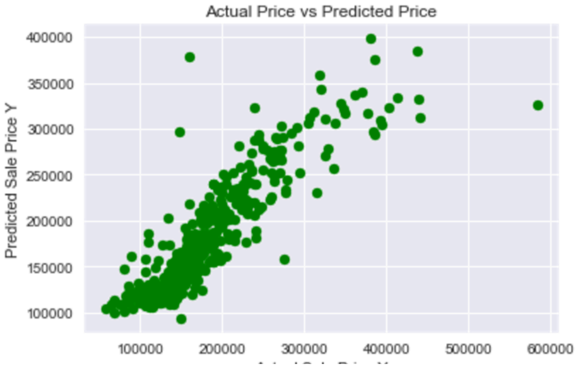


Figure 9: Actual vs predicted Sale Prices - Linear Regression

Furthermore, the residual plot is randomly scattered, which indicates that the linear model is a good-of-fit. However, there might be some outliers with higher residual value in the data set. As a result, outlier detection may be necessary.

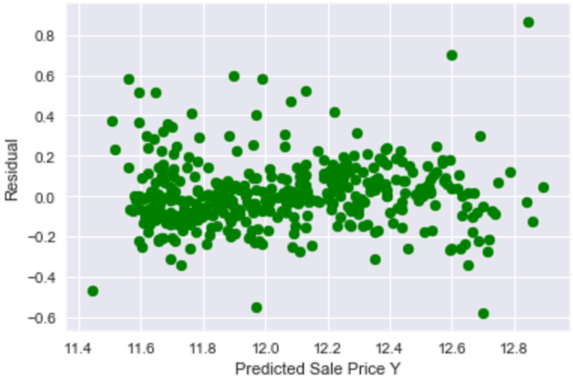


Figure 10: Residual Plot – Linear Regression

Stochastic Gradient Descent

The stochastic gradient descent regressor was then built and used. Incidentally, the results are not as promising as the linear model (R^2 is 0.69 and MSE of 0.041). The reason of this decrease has to do with the number of samples in the data set. While this dataset contains more than 2900 examples, it's recommended that gradient decent be used for a number of examples of 10,000 or more. Therefore, data points in this dataset may not be sufficient for the gradient descent training.

As we can observe in *figure 11*, the pattern residuals are more sparse than the .

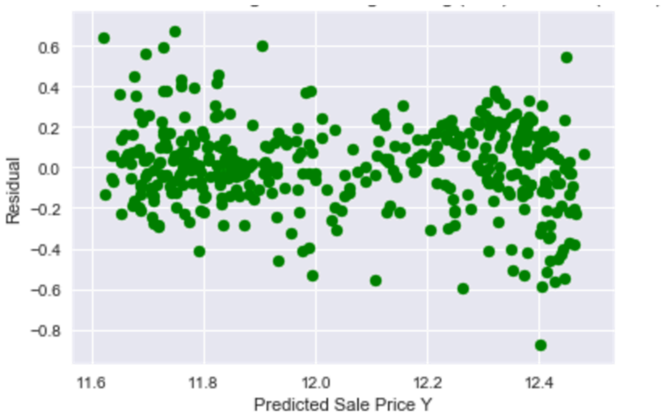


Figure 11: Residual Plot - Gradient Decent

Unsupervised Learning for Outlier Detection

Having outliers in a dataset often times causes a significant impact on the model prediction accuracy. Here, two approaches were used for clustering similar houses in order to be able to detect any outliers, if any. The MeanShift, and the K-means clustering methods were used to investigate this matter.

While K-mean is a very common approach for clustering, it has a very important shortcoming; it takes the number of clusters (*k*) as an input. This means that one has to know (or infer) the number of clusters and use it as an input for the algorithm. To overcome this shortage, the MeanShift algrithm was introduced, which gives the number of clusters as an output, which is more reliable and useful for real-life problems. Meanshift gave a number of clusters of 4, and detected an outlier of the lowest price category.

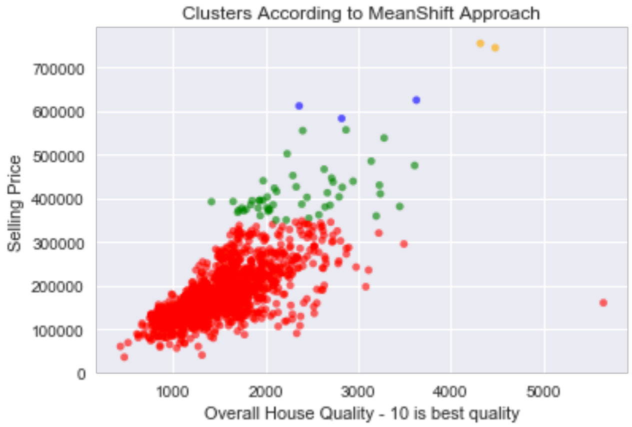


Figure 12: Houses Clusters - MeanShift Approach

To confirm these findings, the K-mean algorithm was run using the number of clusters suggested by the MeanShift algorithm, i.e. k=4.

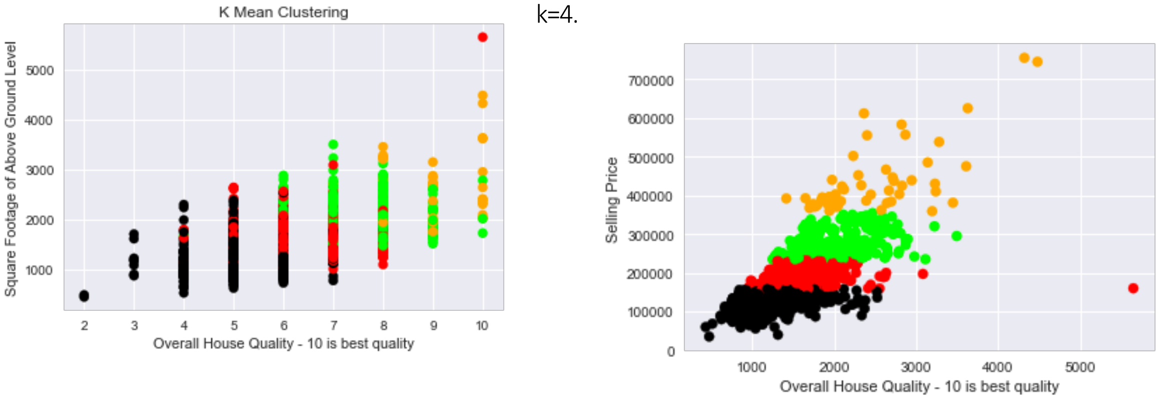


Figure 13: Houses Clusters – K-mean Approach

As for the left plot, as the overall quality rate (x) and living area (y) increased, data points were clustering into 3 groups. But there is one data point was considered as cluster 1 while having the highest House Living Area and Overall Quality. Furthermore, when plotted the relationship of living area and price on the right, it turned out that this particular house was sold for a very low price while having a large living area. Thus, this house was considered an outlier.

After deleting the data point, there is no significant change in the accuracy from both LASSO and Elastic Net models. While K-means is very effective on outlier detection in this case, the outlier has no strong impact in the model.

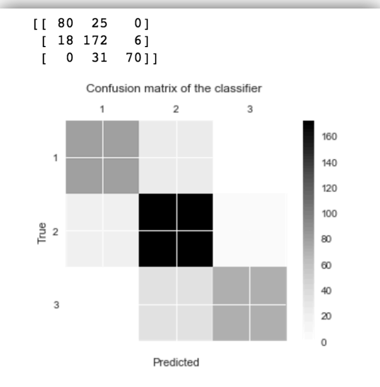
Random Forest for Regression

1- Model Building and Parameters

Similar to the previous models, the random forest regression model is designed to predict the sale price of a house given the features after training the model. Random Forests are very powerful regression or classification prediction models whose concept is to aggregate the results of an ensemble of decision trees which are classification models based on the splitting that data at any node following the greedy approach. Random Forests use bagging (Bootstrap sampling and aggregation) and random feature sampling.

In this random forest, all of the features were fed to the models, using the *RandomForestRegressor*, a number of estimators = 1000, which is the number of trees in the forest, a *min_samples_split* value of 2 (minimum number of samples required to split on any given node), and a *min_samples_leaf* value of 1 (minimum number of samples required for the node to be a leaf node).

Out-Of-Bag score is an evaluation method that uses any sample that has been left out by a certain number of trees due to random sampling as the test set. This score is also set to true.



```
[[ 80  25   0]
 [ 18 172   6]
 [  0  31  70]]
```

Figure 14: Confusion Matrix - Random Forest Classification

2- Results and Analysis

As mentioned before, the random forest model's feature importance method was used to quantify the features for the purpose of dimensionality reduction. The most important feature was the overall quality of the house, followed by the above ground area, and the basement area.

The model's accuracy was of 86% using Cross Validation, and 85.8% using the Out-Of-Bag Score.

Random Forest for Classification

1- Model Building and Parameters

Similar to the Random Forest Regression model, the random forest classification model was run after changing the dataset's target from continuous variables into price classes (3 categorical variables): Low =< 135000, Medium between 135000 & 22000, High >= 220000.

As a classification problem, the dataset was split into two sets; a training set (70% of the data) and a test set (30% of the data). The model was trained using the training set while the test set was left out of the model. The model was then used to predict the Prices Classes of the test set providing the test set features (all variables but the target). To evaluate the performance, the predicted Prices Classes of the test set were compared to the true Prices Classes of that same set.

2- Results and Analysis

For classification models, one of the evaluation methods is using the confusion matrix, which is a comparison between the true and predicted values of a certain set of data (in our case the test set), showing the number of misclassifications. The confusion matrix of this model shows a number of 80 misclassifications (18 houses of Medium prices incorrectly classified as Low prices, 25 Low prices houses incorrectly classified as Medium price houses, 31 High prices houses incorrectly classified as Medium price houses, and 6 Medium price houses incorrectly classified as High price houses.

Models Comparison

Regression Model	Linear Regression	LASSO Regression	Elastic Net Regression	Gradient Descent	Random Forest Regression
Performance	R^2 = 0.7868	R^2 = 0.7878	R^2 = 0.7889	R^2 = 0.690	R^2 = 0.8581

As per the table shown, the Random Forest Regressor has the highest accuracy, while the Gradient Descent performed poorly in this case.

It's also worth mentioning that the regularization methods undertaken in this study have optimized the Linear Regressor by increasing its accuracy and lowering the associated error.

For the classification model, the Random Forest Classifier was used, with the result being fairly good (80 misclassifications as opposed to 322 correct classifications). More classification methods can be applied for further analysis and model comparisons.

Conclusion and Moving Forward

In sum, the prediction models presented in this study could be very beneficial to both house buyers and house market investors. For buyers, they could use it as a reference during price negotiations, while investors can benefit from those models by gaining some insights for developing the price strategy and costumer preference analysis. In general, size and the quality of the house would more or less reflect in the house price, which is also reflected in our model, the overall house quality, the above ground living area, and the size of garage and basement are all very influential attributes controlling the house price to a large extent. Moreover, while the parametric regressors (linear regression, LASSO, Elastic Net) perform fairly good in the prediction, the nonparametric regressor provide better accuracy. The reason is possibly that there is no distribution assumption that holds in the nonparametric model, so it is closer to the actual value. However, as a trade-off, nonparametric model is more complex than the parametric.

Unsupervised Algorithms could be very helpful grouping similar data points together making it very useful for Exploratory Data Analysis, outliers detection, and deciding on models to use. K-mean while simple and popular, it needs the number of clusters to be known prior to running the algorithm, which is somewhat arbitrary. MeanShift approach, however, gives the number of cluster, which is more realistic and useful.

For further research, since we have tried the random forest classification method, other classification techniques are recommended for further analysis and comparisons, such as SVM. Meanwhile, as we known most of the problems are Unsurprising Machine Learning, we could apply or implement some of those algorithms in features selection and model building for more robust and stable prediction.

References

1. *http://ww2.amstat.org/publications/jse/v19n3/decock.pdf*
2. *https://www.kaggle.com/c/house-prices-advanced-regression-techniques*
3. *http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf*