**Kaggle - House Price Prediction**

**MSDS6371 Class Project**

**By Mingyang Nick Yu & Allen Miller**

# Introduction

This project is focused on the Ames Housing dataset on Kaggle. We will first explore how the sales price of a home is influenced by the home’s square footage in three different neighborhoods. We will then build different models aimed at accurately predicts a home’s sales price based off of many different features of the home.

# Data Description

The data comes from the Ames Housing dataset which describes the sale price of property in Ames, Iowa from the years 2006 to 2010. This includes 80 distinct variables, all of which help describe a property and may influence the price of a home. We looked at all the observations included in the dataset (2930) and used a mixture of these variables to build models that can aid in predicting a home’s sale price.

# Analysis 1

* 1. **Restate the problem**

The Century 21 Ames real estate company has hired us to analyze how square footage of living area of house(GrLivArea) is related to its sales price in the three neighborhoods they sell homes in (NAmes, BrkSide, and Edward neighborhoods).

* 1. **Build and Fit**
     1. Model 1
        1. Check Assumptions *(see Appendix 1.1, 1.2, 1.3, 1.4)*
           1. Linearity: *Plot 1.1* has some outliers even the general trend follows linear relationship between Square Foot of Living Area and Sale Price
           2. Normality: Q-Q *Plot(1.4)* shows skewness at either end of the plot – some outliers
           3. Equal SD: Observations 169 and 190 both have standardized residuals greater than 4 - outliers
           4. Independence: We will assume observations are independent
           5. Outliers: Observation 339 has a Cook’s Distance larger than 5.6 while observation 131 has a Cook’s Distance larger than 1.

These may be due to some unique cases

* + - 1. Decision: Since our sample is sufficiently large, removing these four outliers should not affect our results and will be removed for further analysis. (Adjusted R-squared = 0.3406)
    1. Model 2
       - 1. Model2 = model1 run without outliers
       1. Check Assumptions *(Appendix 1.5, 1.6, 1.7, 1.8)*
          1. Linearity: *Plot 1.5* shows linear relationship between square foot of living area and sales price
          2. Normality: Q-Q Plot shows normal distribution
          3. Equal SD: standardized residuals show all data within 2.5 range, so this assumption is met
          4. Independence: Assume our observations are independent.
          5. Outliers: All observations have a Cooks D less than 0.01, so no high leverage and high residual point.
       2. All assumptions are met with this model, however we are interested in adding in the neighborhood factors. So we will go on building out Model 3 and see if the Neighborhood is significant in our analysis. (Adjusted R-squared= 0.449)
    2. Model 3
       - 1. Model3 is run with neighborhood data added in.
       1. Check Assumptions *(Appendix 1.9, 1.10, 1.11, 1.12)*
          1. Linearity: As we seen from *Plot 1.9*, each neighborhood follows a linear relationship between square foot of living area and sales price
          2. Normality: QQ-Plot shows normal distribution of our model
          3. Equal SD: Standardized residual plot shows about 5% of data is beyond residual value of 2. So this assumption is met.
          4. Independence: We will assume all observations are independent.
          5. Outliers: observations have a Cook’s D less than 0.20, so no major high leverage or high residual point.
       2. Conclusion: Adjusted R-squared = 0.5165, which is higher than model 2, by adding in Neighborhoods, our model is better explained. So we will move forward to interpret Model 3.
  1. **The Analysis**
     1. Using model 3 we can generate separate models(one for each neighborhood)
        1. Overall Model:
        2. BrkSide Model:
        3. Edwards Model:
        4. NAmes Model:
     2. Analyze the plot and assumptions
        1. Linearity, Normality, Equal SD, Independence and outliers are checked with section above.
        2. There is no evidence to suggest any major outliers that will need to be accounted for as the residuals appear in a random cloud. All assumptions are met, we will move on to interpret our findings.
  2. **Conclusion** 
     1. There is sufficient evidence to suggest that Model 3 is a good fit for the data (p-value < 0.0001).
     2. We can interpret each sub-model of model 3 as follows
        1. Given that the Neighborhood is BrkSide, it is predicted that the Sales price of the house will increase by $8716.3 for every 100 square feet added to the house. We are 95% confident this increase will be between 7152.22 and 10380.29.
        2. Given that the Neighborhood is Edwards, it is predicted that the Sales Price of the house will increase by $7015.9 for every 100 square feet added to the house. We are 95% confident this increase will be between 5618.25 and 8413.43
        3. Given that the neighborhood is NAmes, it is predicted that the Sales Price of the house will increase by $4956.2 for every 10 square feet added to the house. We are 95% confident this increase will be between 4150.50 and 5761.75
     3. Scope: Because this an observational study we cannot draw any causal inference. Not knowing if this data set is randomly drawn from a bigger population from the entire sales data from 2006-2010, any inference to the population needs to remain speculative.

# Analysis 2

* 1. **Restate the Problem**
     1. Select from all the variables available to us, and build a model that can accurately predict the sales price of a home in Ames, Iowa between 2006 and 2010.
     2. We will first explore the Stepwise, Forward, and Backward models and use our findings to create a more accurate model.
  2. **Clean-up and selection**
     1. We will first look at each variable and convert specific variables that are levels to factors so we can use them as categorical variable for linear regression.
     2. We plotted each continuous variable vs the sales price of the house to look for correlation and independence. From there we will begin to assemble a list of variables with strong correlation that may be good predictors in our regression model. (See *plot 2.1 – 2.12*)
     3. 31 of the 80 variables were selected for our models
        1. We used the pool area variable to us to create a new variable (poolYN) that said Yes or No for a house having a pool.
     4. After selection we decided to look closely at our selected variables that contained N/A and transform those values to useable factors such as None for quality rankings. We then replotted these variables to confirm that there was still strong correlation.
     5. After plotting the residuals, we decided to use a log transform on the sales price to help normalize or data better.
  3. **Build and Fit Models**
     1. Stepwise
        1. Using R and our selected variables we created a model using stepwise AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by internal 10-fold cross validation.
        2. Checking Assumptions
           1. Linearity: this has been checked by pair wise *plots 2.1-2.12*
           2. Normality: see Q-Q *plot 2.13*, normality is roughly met, although there are some outliers at the ends
           3. Equal SD: see Standardized Residuals *plot 2.14*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
           4. Independence: Assume all of our observations are independent
           5. Outliers: Looking at the Cook’s D *plot 2.15*, there is one data point went over 1.5, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.
        3. Conclusions: The Stepwise model has selected following predictors (see *plot 2.16*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, HouseStyle, OverallQual, YearBuilt, YearRemodAdd, ExterCond, Foundation, BsmtQual, BsmtCond, TotalBsmtSF, Heating, CentralAir, GrLivArea, FullBath, KitchenQual, Fireplaces, GarageType, GarageCars, PollYN, MoSold, YrSold, among the 31 variables we feed to the model. And it give us RMSE of 0.149 and follow performance:
           1. Final Results

Kaggle Score = 0.15372

CV Press = 0.1494

Adjusted R-squared = 0.8975

* + 1. Forward Selection
       1. Using R and our selected variables we created another model using Forward selection by AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by creating our own cross validation.
       2. Checking Assumptions:
          1. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
          2. Normality: see Q-Q *plot 2.17*, normality is roughly met, although there are some outliers at the ends
          3. Equal SD: see Standardized Residuals *plot 2.18*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
          4. Independence: Assume all of our observations are independent
          5. Outliers: Looking at the Cook’s D *plot 2.19*, there is one data point went over 1.25, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.
       3. Conclusion: The Forward Selection has chosen the following predictors(see *plots 2.20, 2.21*): OverallQual, Neighborhood, GrLivArea, MSSubClass, OverallCond, GarageCars, YearBuilt, Fireplaces, BsmtQual, MSZoning, Heating, LotArea, YearRemodAdd, CentralAir, KitchenQual, GarageType, TotalBsmtSF, PoolYN, BsmtCond, LotConfig, FullBath among the 31 variables we fed to the automatic model selection. It gives us the following performance for prediction and Cross validation.
          1. Kaggle Score = 0.15432
          2. CV Press = 0.14199
          3. Adjusted R-Squared = 0.89575
    2. Backward Selection
       1. Using R and our selected variables we created the next model using Backward selection by AIC to choose the optimal variables/model among the variables we have narrowed down above, it is further verified by creating our own cross validation.
       2. Checking Assumptions:
          1. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
          2. Normality: see Q-Q *plot 2.22*, normality is roughly met, although there are some outliers at the ends
          3. Equal SD: see Standardized Residuals *plot 2.23*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
          4. Independence: Assume all of our observations are independent
          5. Outliers: Looking at the Cook’s D *plot 2.24*, there is one data point went over 1.25, comparing to our sample size, it should not have a huge impact on our model. So we will keep this observation and move on.
       3. Conclusion: The Backward Selection has chosen the following predictors(see *plots 2.25, 2.26*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, OverallQual, YearRemodAdd, BsmtQual, TotalBsmtSF, Heating, CentralAir, GrLivArea, FullBath, KitchenQual, Fireplaces, GarageType, GarageCars, PoolYN, among the 31 variables we fed to the automatic model selection. It gives us the following performance for prediction and Cross validation.
          1. Kaggle Score = 0.15432
          2. CV Press = 0.148
          3. Adjusted R-Squared = 0.89575
    3. Custom Model:
       1. To design our optimal custom model, we used following steps:
          1. Reimport the training dataset, use predictive mean matching to fill in missing continuous variables (use best estimate and keep original distribution of each variable.)
          2. Divide all 80 variables available to us into different subgroups
          3. Run best subset selection to pick out the best predictors within each subgroup
          4. Log transformation on SalePrice and GrLivArea to increase linearity relationship.
          5. Convert Categorical variables with NA to None or Others.
          6. Use all predictors selected under each subgroup to run a stepwise AIC to choose the optimal variables/model
          7. Further verified model by internal 10-fold cross validation.
       2. Checking Assumptions:
          1. Linearity: this has been checked by pair wise *plotting 2.1-2.12*
          2. Normality: see Q-Q *plot 2.27*, normality is roughly met, although there are some outliers at the ends
          3. Equal SD: see Standardized Residuals *plot 2.28*, some outliers are outside of 2.5 range, however due to the size of our sample data, it should not cause major concern. We will assume this assumption met and move on.
          4. Independence: Assume all of our observations are independent
          5. Outliers: Looking at the Cook’s D *plot 2.29*, there are two data points went around 0.6. This should not be a problem considering our sample size.
       3. Conclusion: The Custom model has chosen the following predictors(see *plots 2.30, 2.31*): MSSubClass, MSZoning, LotArea, LotConfig, Neighborhood, Condition2, BldgType, OverallQual, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, Foundation, BsmtQual, BsmtFinType1, TotalBsmt, Heating, HeatingQC, CentralAir, Electrical, X2ndFlr, GrLivArea, BsmtFullBath, FullBath, BedroomAbvGr, KitchenAbvGr, KitchenQual, Firepplaces, GarageCars, GarageArea, PavedDrive, WoodDeckSF, ScreenPorch, MoSold, YrSold. It gives us the following performance for prediction and Cross validation.
          1. Kaggle Score = 0.14773
          2. CV Press = 0.1649
          3. Adjusted R-Squared = 0.9207
  1. **Overall Conclusion:**

After running and comparing the four models we built, see table below, the Custom model gives us the best Adjusted R-Squared score (highest) and best Kaggle Score(Lowest). After all, we choose custom model as our best model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Predictive Models** | **Adjusted R2** | **CV PRESS** | **Kaggle Score** |
| Forward | 0.89575 | 0.14199 | 0.15432 |
| Backward | 0.89575 | 0.148 | 0.15432 |
| Stepwise | 0.8975 | 0.1494 | 0.15372 |
| CUSTOM | 0.9207 | 0.1649 | 0.14773 |

# Reference

* 1. MSDS 6371 Project Description.docx – see details at <https://github.com/nickmingyang/MSDS6371Project>
  2. Kaggle Competition – Data description, training and testing Data source, and other data details: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
  3. Project source code used for analysis and results: MSDS6371Project.Rmd under: <https://github.com/nickmingyang/MSDS6371Project>
  4. Custom Model test set prediction—custom\_model\_Miller\_YU.csv under: <https://github.com/nickmingyang/MSDS6371Project>
  5. Stepwise AIC model: <http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/>
  6. Other models result labeled csv files under: <https://github.com/nickmingyang/MSDS6371Project>

# Appendix

* 1. Analysis 1
     1. Model 1

Chart, scatter chart

Description automatically generated(1.1)

Chart

Description automatically generated(1.2)

Chart

Description automatically generated(1.3)

Chart, line chart

Description automatically generated(1.4)

* + 1. Model 2

Chart, scatter chart

Description automatically generated(1.5)

A picture containing chart

Description automatically generated(1.6)

Chart

Description automatically generated(1.7)

Chart, line chart

Description automatically generated(1.8)

* + 1. Model 3

Chart, scatter chart

Description automatically generated(1.9)

Chart, line chart

Description automatically generated(1.10)

Graphical user interface, chart

Description automatically generated(1.11)

Chart, histogram

Description automatically generated(1.12)

* 1. Analysis 2

Diagram

Description automatically generated(2.1)

Diagram

Description automatically generated(2.2)

Diagram

Description automatically generated(2.3)

Diagram

Description automatically generated(2.4)

A picture containing box and whisker chart

Description automatically generated(2.5)

Diagram

Description automatically generated(2.6)

A picture containing box and whisker chart

Description automatically generated(2.7)

A picture containing box and whisker chart

Description automatically generated(2.8)

A picture containing diagram

Description automatically generated(2.9)

A picture containing graphical user interface

Description automatically generated(2.10)

Diagram

Description automatically generated(2.11)

Diagram

Description automatically generated(2.12)

Chart, line chart

Description automatically generated(2.13)

Graphical user interface

Description automatically generated(2.14)

Chart, scatter chart

Description automatically generated(2.15)

A close up of a newspaper

Description automatically generated(2.16)

Chart, line chart

Description automatically generated(2.17)

Graphical user interface, application

Description automatically generated(2.18)

Chart, scatter chart

Description automatically generated(2.19)

Table

Description automatically generated(2.20)

Table

Description automatically generated(2.21)

Chart, line chart

Description automatically generated(2.22)

Graphical user interface

Description automatically generated(2.23)

Chart, scatter chart

Description automatically generated(2.24)

Table

Description automatically generated(2.25)

Table

Description automatically generated(2.26)

Chart, line chart

Description automatically generated(2.27)

Graphical user interface

Description automatically generated(2.28)

Chart, scatter chart

Description automatically generated(2.29)

Text

Description automatically generated(2.30)

Table

Description automatically generated(2.31)

Graphical user interface, text, application, email

Description automatically generated(2.32)