Sale Price Prediction

Alex Hunt

Abstract

Many factors are taken in to consideration when appraising a house such as the square footage, garage size, and even certain materials used. Due to these many factors, accurate house prediction becomes challenging. This report will use multiple predictive models including Simple Linear Regression, LASSO, and Principal Component Regression to accurately predict the price of homes using the Ames Housing data set. Cross-Validation will be used to minimize test errors (RMSE) and imporve the R squared statistic. Data transformations including encoding categorical variables, normalizing distributions, and removing outliers will be used to ensure accuracy of the linear models.

Keywords: Simple Multivariate Regression, Principal Component Regression, LASSO, Cross-Validation

1 Introduction

Knowing the features of a house that correlate most with sale price provides a great value to home owners, investors, and home builders. For example, when remodelling, the best materials and features can be identified to have the highest increase in home value compared to amount invested. The Simple Linear Regression model in the report will identify these features and use the top ones to predict sale price.

In addition to identifying correlated features, overall accuracy of the linear models are important for determining sale price. Unbiased home appraisals are necessary for lenders to assess credit risk and provide a mortgage[1]. One popular example of an application that provides home estimates is Zillow which uses a neural network based model[2]. LASSO and Principal Component Regression will be used in this report to optimize both test accuracy and R squared which indicates whether or not the model can accurately predict sale price.

2 Research Question

The primary research question for this project is identifying if there is a linear relationship between sale price and the predictor variables. If there is a linear relationship, what variables have the highest correlation and how accurate is the model.

3 Literature Review

One research on home price prediction uses the popular Boston housing data set[3]. Using multiple linear regression techniques, they found the best results from a LASSO model with an R-squared of 88.79%. This research uses 13 numerical variables to predict the median value of homes is various suburbs. Many of these factors are based on community information such as crime rate and low income percentage. This report expands on these topics by introducing additional data on individual houses, this time predicting the value of a specific house rather than the local median. Due to their success with using regularization (LASSO and Ridge), this report will build on that research by using a similar technique with Principal Component Regression and LASSO.

4 Methodology

4.1 Data and Data Sources

This study uses the Ames Housing data set from Kaggle[4]. It is made up of 1460 observations on 79 variables, 43 categorical and 36 numeric that describe different factors of a house. The response variable being predicted is the sale price.

4.2 Methods

This study used linear regression to predict the sale price of a house. This works by finding a line that best describes the correlation between the dependent and explanatory variables.

In mathematical notion, this can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon \tag{1}$$

where, y_i is the dependent variable, x_i are the explanatory variables, β_0, is the y intercept, and β_p is the slope coefficient. ε is the residual which represents the error (estimated vs observed). The regression line results from finding the minimized sum of squared residuals (Least squares estimation).

For an extension of the Simple Linear Regression model, LASSO and Principal Component Analysis (PCA) is used. LASSO works by adding penalties to the coefficients. In some cases, this can result in a coefficient of 0 which makes LASSO useful for feature selection. PCA is also a feature selection method which creates linear combinations of variables called components. Often times, only a few of the components explain the majority of the variance in the data. In this report, PCA will be used to optimize a regression model. The increased benefits of using these two methods include dimensional reduction, preventing collinearity, and reducing overfitting.

5 Results

In Table: 1, we have the final results of the three linear models used to predict sale price. RMSE comes from the actual errors from the testing data.

Table 1: Final Results

Model	**R Squared**	**Root Mean Squared Error**
Simple Linear Regression	.8135	0.1714
Principal Component Regression	.9054	0.1209
LASSO Regression	.8985	0.1205

5.1 Simple Linear Regression

Here is the correlation matrix of the top 10 correlated variables after adjusting for collinearity. Year, Sqft, and Quality came from a combination of variables, others were dropped.

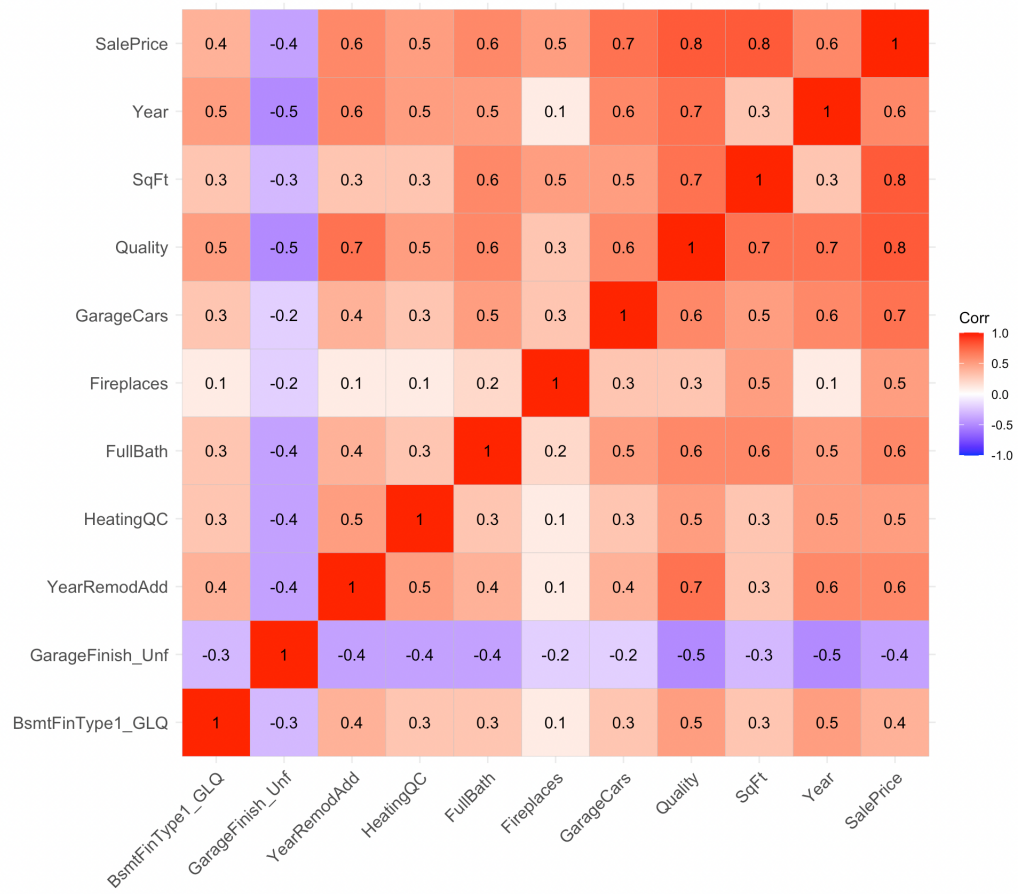

Figure 1: Correlation Matrix for House Data

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.95924 -0.08049  0.01415  0.09555  0.50238

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.642e+00  8.765e-01   7.578 7.94e-14 ***
BsmtFinType1_GLQ  3.383e-02  1.426e-02   2.372 0.017862 *
GarageFinish_Unf -2.918e-02  1.362e-02  -2.143 0.032350 *
YearRemodAdd      1.457e-03  3.850e-04   3.784 0.000163 ***
HeatingQC         1.865e-02  7.005e-03   2.662 0.007892 **
FullBath          2.769e-02  1.347e-02   2.056 0.040042 *
Fireplaces        8.916e-02  9.621e-03   9.267  < 2e-16 ***
```

```
GarageCars        1.003e-01  1.031e-02   9.727  < 2e-16 ***
Quality           4.107e-02  3.586e-03  11.454  < 2e-16 ***
SqFt              1.324e-04  1.013e-05  13.068  < 2e-16 ***
Year              5.688e-04  3.426e-04   1.660 0.097154 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1723 on 1011 degrees of freedom
Multiple R-squared:  0.8135,Adjusted R-squared:  0.8116
F-statistic: 440.9 on 10 and 1011 DF,  p-value: < 2.2e-16
```

The text above shows the R summary output of the Simple Linear Regression model. All variables but Year were statistically significance. This could be happening due to the collinearity between Year and Quality (0.7).

5.2 Principal Component Regression

The cross-validation results for optimal number of components to minimize test error was 94. The percent of variance explained by these components was 90.54% and the RMSE of the testing set was 0.1209.

5.3 LASSO Regression

The cross-validation results for finding the best LASSO model was a lambda value of 0.003364. Of the 111 variables passed into the model, 53 were shrunk to zero resulting in 58 variables explaining 89.85% of the variance. The LASSO model saw the best test errors with an RMSE of .1205

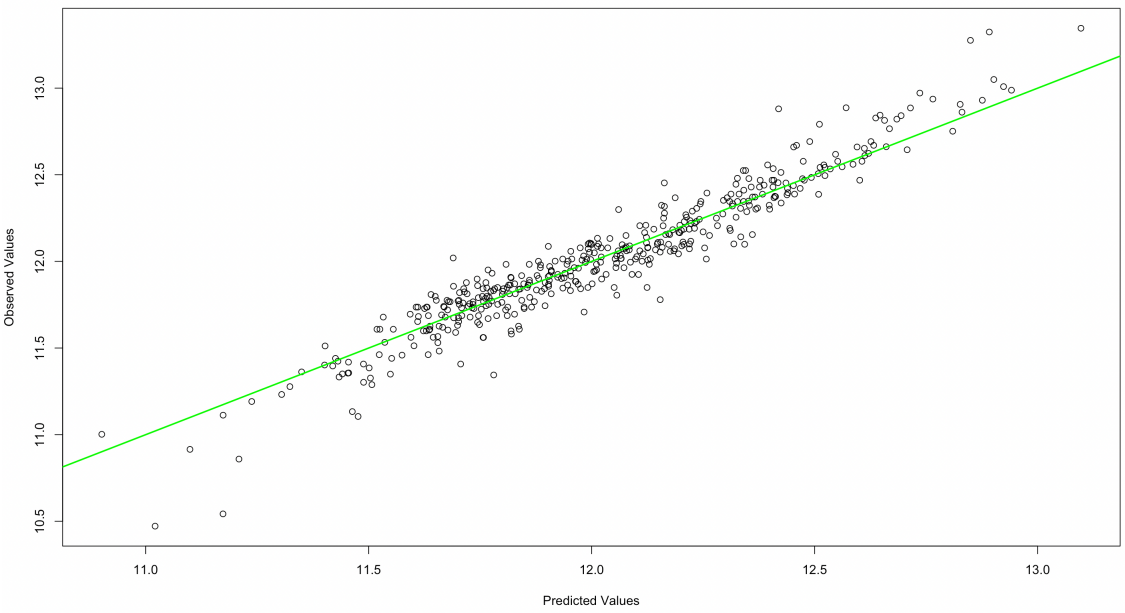

Figure 2: LASSO Prediction vs Actual Values

6 Discussion

The LASSO and PCR models had much better results over the Simple Linear Regression model (SLR). This is due to the added benefits of dimensional reduction which helps prevent over fitting and the removal of collinearity. When looking at the correlation matrix of the top 20 variables, half of them were collinear such as garage area and garage cars. Although steps were taken to remove this, some variables such as year and quality did show significant (0.7) correlation. These results are similar to the research in the literature review with SLR being the worse performing model and LASSO being the best.

7 Conclusion

Overall, the PCR and LASSO regression models were very similar with the PCR providing a better R squared (.9054) and LASSO with a slightly improved test error (0.1205), only a 0.004 difference. All models answered the research question of there being a significant linear relationship between some combination of predictor variables and sale price. Variables related to quality and square footage had the highest correlation with sale price. Similar to how Zillow uses

market data such as comparable homes, and current trends[2], this research could be improved with additional data that goes past home features such as mortgage rates and school districts.

8 Codes

```
data = read.csv("train.csv")
library("dplyr")
library(tidyverse)
library("ggplot2")
library("ggcorrplot")
library("caret")
library("data.table")
library(mltools)
library(pls)
library(glmnet)
library(coefplot)
library(Hmisc)
library(rcompanion)

set.seed(1)
#colSums(is.na(data))[colSums(is.na(data)) > 0]

garage_cols <- c('GarageType', 'GarageFinish', 'GarageQual', 'GarageCond')
bsmt_cols <- c('BsmtExposure', 'BsmtFinType2', 'BsmtQual', 'BsmtCond', 'BsmtFinType1')

data$MiscFeature[is.na(data$MiscFeature)] = 'NA'
data$Alley[is.na(data$Alley)] = 'NA'
data$PoolQC[is.na(data$PoolQC)] = 'NA'
data$Fence[is.na(data$Fence)] = 'NA'
data[garage_cols][is.na(data$GarageType),] = 'NA'
data$Fence[is.na(data$Fence)] = 'NA'
data$GarageYrBlt <- ifelse(is.na(data$GarageYrBlt), data$YearBuilt, data$GarageYrBlt) #Set t
data$FireplaceQu[is.na(data$FireplaceQu)] = 'NA'
data$LotFrontage[is.na(data$LotFrontage)] = 0
data[bsmt_cols][is.na(data$BsmtFinType2),] = 'NA'
data[bsmt_cols][is.na(data$BsmtExposure),] = 'No' #Most common value
data$MasVnrType[is.na(data$MasVnrType)] = 'NA'
data$MasVnrArea[is.na(data$MasVnrArea)] = 0
data$Electrical[is.na(data$Electrical)] = 'SBrkr' #House built in 2006, newer electicity typ

#colSums(is.na(train_df))[colSums(is.na(train_df)) > 0]

categorical <- dplyr::select(data, where(is.character) | Id) #Get the categorical data + id
ordinal <- subset(data,select=c('GarageQual','GarageCond','PoolQC','FireplaceQu',
```

```
                                    'KitchenQual','HeatingQC','BsmtExposure','BsmtCond',
                                    'BsmtQual','ExterCond','ExterQual'))

#Assign a number scale to the ordinal variables
y = colnames(ordinal)
data[y][data[y] == 'NA'] = 0
data[y][data[y] == 'Po'] = 1
data[y][data[y] == 'No'] = 1
data[y][data[y] == 'Mn'] = 2
data[y][data[y] == 'Av'] = 3
data[y][data[y] == 'Fa'] = 2
data[y][data[y] == 'TA'] = 3
data[y][data[y] == 'Gd'] = 4
data[y][data[y] == 'Ex'] = 5
data[y] <- sapply(data[y], as.numeric)

numeric <- select_if(data, is.numeric)

#Get the rest of the categorical data and make them factors
nominal = categorical[,!names(categorical) %in% names(ordinal)]
nominal[sapply(nominal, is.character)] <- lapply(nominal[sapply(nominal, is.character)],
                                                 as.factor)
#One hot encoding of nominal variables. Note: Dropping one factor won't be required for the
newdata <- one_hot(as.data.table(nominal))
newdata = as.data.frame(newdata)

#Combine the two data tables
All <- merge(newdata,numeric,by="Id")
#Get columns with near zero variance, provides no value
zero = nearZeroVar(All, names = TRUE)
All = All[,!names(All) %in% zero]

par(mfrow=c(1,2))
hist(All$SalePrice)
hist(log(All$SalePrice))

set.seed(1)
Transformations = c("YearBuilt", "GarageYrBlt", "YearRemodAdd", "GrLivArea", "TotalBsmtSF",
                    "X1stFlrSF", "LotArea", "LotFrontage", "BsmtFinSF1", "BsmtUnfSF", "MasVn
for (x in Transformations){
  All[,x] = transformTukey(All[,x])
}
All$SalePrice = log(All$SalePrice)


#split data into a test and training set
```

```
train <- All %>% dplyr::sample_frac(0.70)
test  <- dplyr::anti_join(All, train, by='Id')

test = test[,-1]
train = train[,-1]

correlation = cor(train, train$SalePrice, method = "pearson")
names = rownames(correlation)
abs_cor = abs(correlation) #absolute value of the correlation coefficients
corr_data = data.frame(variable = names, abs_cor = abs_cor,cor = correlation)
corr_data = corr_data[order(corr_data$abs_cor,decreasing = TRUE),]
head(corr_data, 21)
cols = corr_data[c(1:21),]$variable

#Keep only the top 20 correlated values
train_1 = train[,names(train) %in% cols]
test_1 = test[,names(test) %in% cols]

#Removing Collinearity (>= 0.7)

#1. Combine correlated quality variables
train_1$Quality = train_1$OverallQual + train_1$ExterQual + train_1$KitchenQual + train_1$Bs
#2. Combine interior square footage variables
train_1$SqFt = train_1$GrLivArea + train_1$TotalBsmtSF
#3. Combine year
train_1$Year = (train_1$YearBuilt + train_1$GarageYrBlt)/2
#4. Drop columns
high_corr = c("GarageArea", "FireplaceQu", "TotalBsmtSF", "Foundation_PConc", "GarageYrBlt",
train_1 = train_1 %>% dplyr::select(-SalePrice,SalePrice)
train_1 = train_1[,!names(train_1) %in% high_corr]

test_1$Quality = test_1$OverallQual + test_1$ExterQual + test_1$KitchenQual + test_1$BsmtQua
test_1$SqFt = test_1$GrLivArea + test_1$TotalBsmtSF
test_1$Year = (test_1$YearBuilt + test_1$GarageYrBlt)/2
test_1 = test_1 %>% dplyr::select(-SalePrice,SalePrice)
test_1 = test_1[,!names(test_1) %in% high_corr]

corr <- round(cor(train_1), 1)
ggcorrplot(corr, lab = T, type = "full")


lm.fit = lm(SalePrice ~ ., data=train_1)
summary(lm.fit)
plot(lm.fit)

rmse = mean((test_1$SalePrice - predict(lm.fit, test_1[,-11]))^2) %>% sqrt()
```

```
rmse

pcr_model <- pcr(SalePrice ~ ., data=train, scale=TRUE, validation="CV")
comps <- RMSEP(pcr_model)$val[1,,]
best <- which.min(comps) - 1
pred <- predict(pcr_model, test[,-111], ncomp=best)
summary(pcr_model)

#validationplot(pcr_model, val.type = "R2")
rmse <- mean((test$SalePrice - pred)^2) %>% sqrt()
rmse

y = data.matrix(train$SalePrice)
x = data.matrix(train[,-111])
x2 = data.matrix(test[,-111])

cv_model <- cv.glmnet(x, y, alpha = 1) #cross validation to find lowest test error
best_lambda <- cv_model$lambda.min
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
best_model
pred<- predict(best_model, s = best_lambda, newx = x2)
rmse <- mean((test$SalePrice - pred)^2) %>% sqrt()
rmse

par(mfrow=c(1,1))
plot(pred,test$SalePrice,
     xlab = "Predicted Values",
     ylab = "Observed Values")
abline(a = 0, b = 1, lwd=2,
       col = "green")
```

References

[1] A. Bogin and J. Shui, "Appraisal accuracy and automated valuation models in rural areas," 08 2020, pp. 60,40–52.

[2] "Zillow zestimates," https://www.zillow.com/z/zestimate/, accessed: 2022-12-11.

[3] S. Sanyal, S. Biswas, D. Das, M. Chakraborty, and B. Purkayastha, "Boston house price prediction using regression models," 08 2022.

[4] Kaggle data, "Ames housing dataset," 2022, data retrieved from kaggle, https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.