Modelling House Prices

Rhys Kilian

September, 2017

Executive Summary

This report outlines the approaches used to develop an analytical model to predict property prices for houses within the target city. Initial data processing, exploratory data analysis and feature engineering ensure the data was suitable for model building. Subsequently several different approaches to model building were investigated, including ordinary least squares (OLS), variable selection (forward selection), regularization methods (LASSO, ridge regression and elastic net) and dimension reduction methods (principle component analysis (PCA) and partial least squares (PLS)). The models developed were trained on the training data issued by the client, and model selection occurred using a cross-validation approach. Subsequently, these models were used to predict values in the client's test data (50% of the total test data), resulting in a validation score being obtained. The objective of this report was to build a model (based on the training data) with the best predictive accuracy in the full test data.

It was found that the best performing models were the elastic net (EN) and LASSO models, which both had the top two cross-validation scores in the model selection phase and the lowest MAE values on 50% of the test data in the model evaluation phase. Having the ability to select variables, both approaches produce simpler models, reducing the variance, expected as a consequence of the bias-variance trade-off. Reduced model complexity also means that the models selected are at less risk in over-fitting to the training data. Hence, the two proposed models will allow councilors to make more informed decisions regarding property prices, and will provide a systematic method of calculating taxes, providing policy information and negotiating with property developers.

Contents

1 Introduction

Property prices are used by city governments to not only determine taxes (including property taxes and stamp duties), but to also inform city planning including making public policy decisions and to negotiating with property developers. However, it is often the case that current prices are not available for most properties, as each year only a small percentage of properties will be placed on the market. In order to estimate the current price of all properties, a predictive model can be developed using data from recent house sales.

The dataset used to develop this model contains highly detailed information about recent house sales. The predictors include 22 nominal variables (categorical data), 23 ordinal variables (ordered categorical data), 20 continuous variables (infinite possible values) and 14 discrete variables (particular range of real values, eg. integers). Further details of these predictors can be found in Appendix A.

2 Data Processing

Prior to performing variable selection and exploratory data analysis (EDA), data in both the train and test sets must be 'cleaned up' to remove any outliers, handle missing values and ensure correct processing for model building. Feature creation is discussed later, in Section 4. Feature Engineering. It is important that the same processing is applied to both the training and test data to ensure the same model can be applied to both.

As with any large dataset, it is important to consider missing data, specifically the prevalence of missing data and if this missing data is random or follows a pattern. Missing data can cause issues for modeling due to a reduction in the sample size. Moreover missing data, from a substantive perspective, can hide key facts about the data, resulting in bias.

Table 1: Largest Missing Values for Predictors

	Total	Percent
PoolQC	801	0.996
MiscFeature	781	0.971
Alley	749	0.932
Fence	644	0.801
FireplaceQu	386	0.480
LotFrontage	157	0.195
GarageType	36	0.045
GarageFinish	36	0.045
GarageQual	36	0.045
GarageCond	36	0.045
GarageYrBlt	36	0.045
BsmtExposure	22	0.027
BsmtCond	21	0.026
BsmtFinType2	21	0.026
BsmtFinType1	21	0.026
BsmtQual	21	0.026
MasVnrArea	4	0.005
MasVnrType	4	0.005

As seen in Table 1, there are several variables with significant percentage of observations missing (up to 99.6%). For a number of these variables, the missing values represent null values (i.e. not having a particular feature). These include the ordinal variables 'PoolQC', 'Fence', 'FireplaceQu', 'GarageFinish', 'GarageQual', 'GarageCond', 'BsmtExposure', 'BsmtCond', 'BsmtFinType1', 'BsmtFinType2' and 'BsmtQual', as well as the nominal variables 'MiscFeatures', 'Alley', 'MasVnrType' and 'GarageType'. The disrete variable 'GarageYrBlt' and continuous variable 'MasVnrArea' likewise also correspond to not having a particular feature. The 'LotFrontage' variable however, as a continuous variable, can be assumed to be missing data, and can be replaced by using the median LotFrontage of that particular neighborhood. In this way, all the missing data values can be filled in with appropriate responses.

Furthermore, the variables 'MSSubClass' and 'MoSold' are ordinal variables and not continuous. These were appropriately transformed from numerical values into ordinal strings. Additionally, several categorical features were encoded as ordered numbers when there is information in the order. These include variables such as 'BsmtCond' and 'BsmtExposure'.

3 Exploratory Data Analysis

In order to build a good predictive model it is important to understand the key features of the data including the behavior of key variables, the relationships between variables, and anomalies in the data (such as outliers). The exploratory data analysis here is conducted on the training data only in order to reserve the test data for model evaluation.

3.1 Univariate Analysis of Response (Sale Price)

Given the model aims to predict the property price, gaining an understanding the distribution of the response variable, that is the sale price of the data given, will be critical to building an accurate model.

The property prices in the dataset range in value from \$35,000 to \$615,000, with a mean of \$175,324.47 and standard deviation of \$70,035.49. However the distribution of the sales prices as seen in Figure 1 does not follow a normal distribution, with significant positive skewness (skew of 1.581) and signs of peakedness (kurtosis of 4.277).

Figure 1: Histogram for Sale Price

For the sale prices of properties, the year in which the property is sold has the potential to influence the overall sale price. However looking at the relationship between 'YrSold', the year in which the property was sold and the sale price, as seen in Figure 2, the average sale price and distribution remains fairly constant for each year of the sample data.

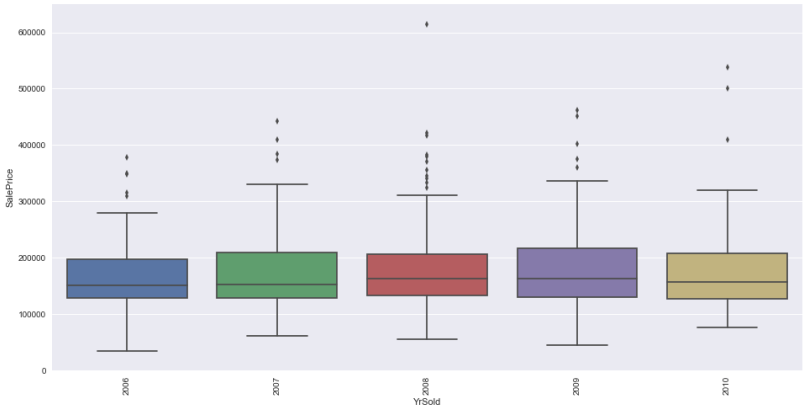

Figure 2: Box Plot of Year Sold and Sale Price

3.2 Covariance of Predictors

Given the large number of predictors in the data, before the bivariate relationships between the response and predictors are analyzed, it is important to first gain an understanding of the relevance of predictors. Specifically, high correlation between predictors and the response suggests that the variable is significant for the prediction of the response. High correlation between predictors however, could indicate that the same prediction information is captured by multiple predictors (and thus is worth considering excluding one to prevent collinearity issues). Figure 13 in the Appendix B shows the correlation matrix for the full selection of variables.

Using this figure, it is evident that the pairs 'TotalBsmSF' and '1stFlrSF', 'GarageCars' and GarageArea', and finally 'GarageYrBlt' and 'YearBuilt' are highly correlated with each other. These three pairs of variables all suggest a possible cases of multicollinearity. An understanding of what these variables represent in these cases leads to the realization that all three pairs roughly predict the same information. For instance, the number of cars which fit in a garage will be a direct reflection of the area of the garage.

Consideration of both the meaning behind these variables as well as the correlation with the sale price, will help determine which variables should be included. Figure 3 shows the correlation between predictors and the response sale price, for the seven predictors most correlated with sale price. From this graph is can be seen that the 'OverallQual' (overall house quality), 'GrLivArea' (above ground living area) and 'GarageArea'/'GarageCars' (garage size) are all strongly correlated with the sale price.

Significantly, the variable pairs previously discussed with strong elements of multicollinearity, all have similar correlation, further suggesting that including only one of these two variables is necessary. 'GarageCars' will be included as this is usually a consequence of the garage area. Similarly, '1stFlrSF' should be included over 'TotalBsmtSF' as not all properties contain a basement, and the inclusion of a basement can be captured by other ordinal variables such as 'BsmtQual' (basement height) and 'BsmtCond'

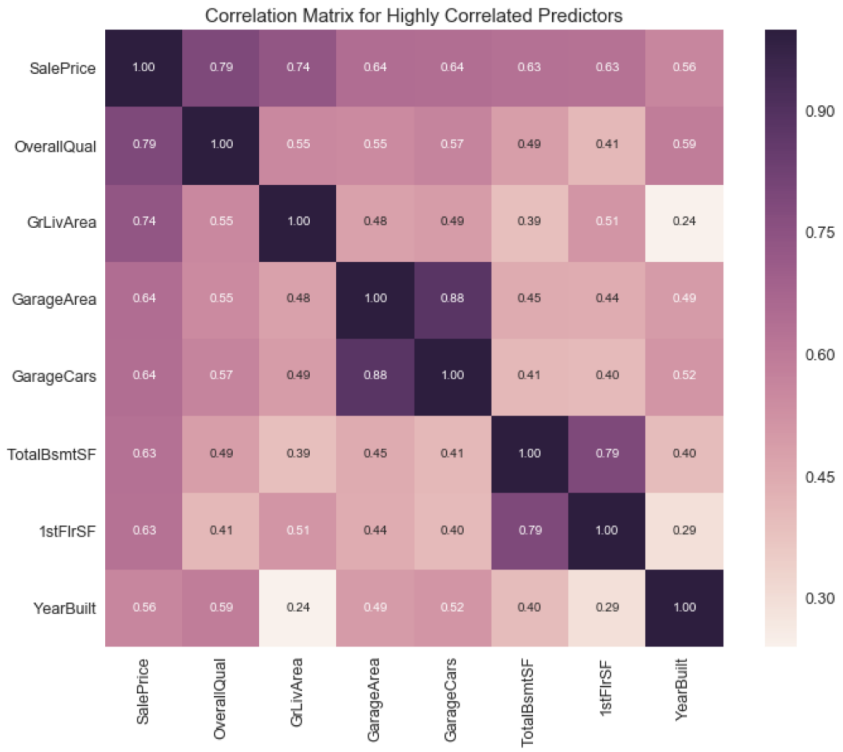

Figure 3: Correlation Matrix for Predictors Strongly Correlated with Sale Price

(condition of the basement). Finally, the 'YearBuilt' if more favorable to include than the 'GarageYrBuilt' due to the house usually being built first, if not at the same time, as well as its stronger correlation with the sale price (0.561 vs. 0.515). These variables were dropped from the dataset in Section 4 Feature Engineering before the models were fitted.

Looking into the bivariate relationships between these significant variables in Figure 14 in the Appendix indicates some interesting trends. Interestingly, the plot between 'GrLivingArea' the above ground living area and '1stFlrSF' the 1st floor living area has two distinct trends. The first is an almost, a linear line, which would correspond to houses of a single story. The rest of the data lies below this line, following a somewhat linear trend, representing multi-story houses, where the 1st floor area will always be smaller than the above ground living area.

Another interesting trend from Figure 14 is the relationship between 'SalePrice' and 'YearBuilt'. The relationship between these two variables appears to almost be exponential in nature. Furthermore, the older houses appear to have a narrower and lower range of prices, which newer houses have a much wider spread of prices, with the lowest price seemingly increasing for newer houses.

The variation in bivariate relationships between significant parameters prompts the investigation of bivariate relationships further. As has been established, certain predictors seem to have linear relationships with the response (like 'OverallQual', 'GrLivArea'), whilst others have more complex relationships (like 'YearBuilt'). Furthermore a degree of collinearity between variables has been established, especially in predictors with related measurements (such as square footage of different areas).

3.3 Bivariate Analysis

Understanding the bivariate relationship between key predictors and the response variable allows for a greater understanding of the dataset and potential models to fit to the data.

Given the large number of predictors, only the predictors highly correlated with the data will be investigated here. Further relationships can be found in Appendix B.

The predictor with the highest correlation to sale price was 'OverallQual' (correlation of 0.790), which is an ordinal variable measuring the overall material and finish of the house. Figure 4 shows the variation between sale prices and the overall quality, indicating an upwards linear trend. Interestingly, the spread of possible sale prices increases for higher quality houses. There are also a number of outliers present which have higher house prices than those commonly found for that quality.

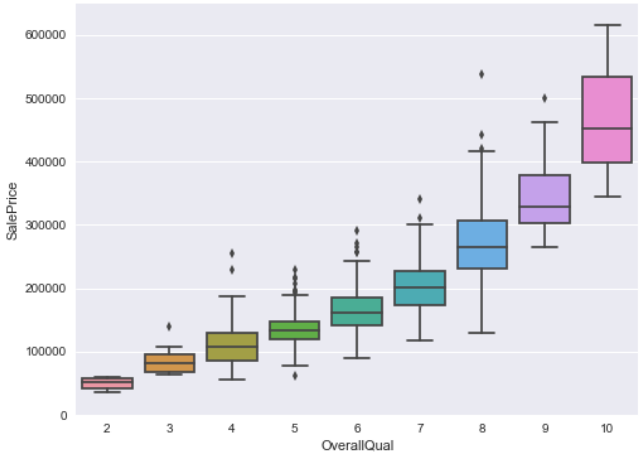

Figure 4: Box Plot for 'OverallQual' and 'SalePrice'

The next highest correlated predictor was 'GrLivArea' (correlation of 0.738), which is a continuous variable measuring the above group living area in square feet. As shown in Figure 5, there appears to be a linear relationship between above ground living area and the sale price. Notably as 'GrLivArea' increases, the spread of sale prices also increases, indicating increased variance in the data.

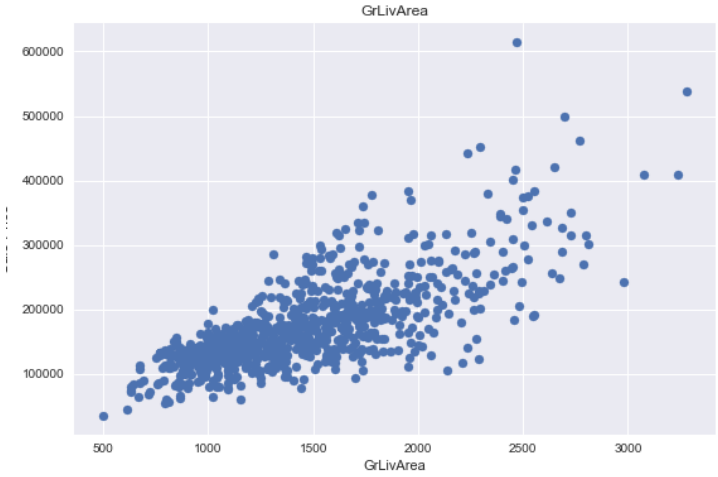

Figure 5: Scatter Plot for 'GrLivArea' and 'SalePrice'

The year in which the house was built 'YearBuilt' is another predictor with significant correlation with the sale price (correlation of 0.561). The year built is a discrete vari-

able, mapping the original construction date. The relationship between 'YearBuilt' and 'SalePrice' is not as strong as other predictors, however overall there is an increase in sale price for more recently built properties. This trend does see significant variability with certain years having larger spreads of data, and/or significant variation from the upwards trend.

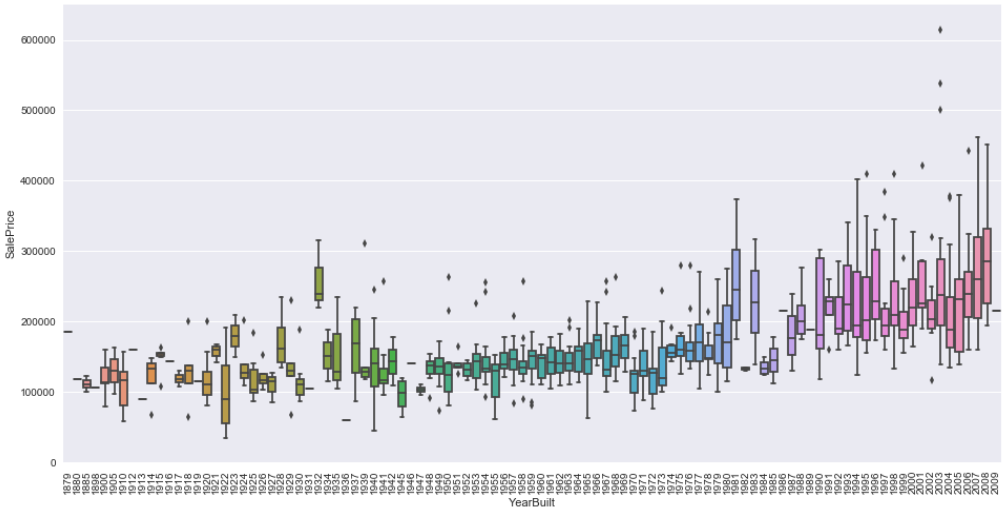

Figure 6: Box Plot for 'YearBuilt' and 'SalePrice'

3.4 Outliers

In a similar manner, the presence of outliers in the data not only has the potential to markedly affect the models produced, but to also provide valuable information about specific behaviors.

The first step to identifying outliers is to look at the standardized response data, by subtracting the mean and dividing by the standard deviation. The scaled data points will thus have a mean of 0 and standard deviation of 1. By looking at the extreme low and high outer ranges of the data is it possible to identify potential outliers in the data. In the lowest range, the values are close to one another and are not too far off from 0, suggesting that they are not distinct outliers. The higher range has greater variation away from 0, with standardized values around 5/6, signifying potential outliers.

Analyzing the response data in conjunction with other predictors will help determine if the points identified above are actually outliers. As seen in Figure 5, the top two points, which represent the standardized values of 5 and 6 previously discussed are fairly removed from the other data points. However they still appear to follow the general data trend and thus should not be removed.

Given the complexity of the number of variables involved it is impractical to look at each variable separately and eliminate potential outliers. In particular a data point might appear to be an outlier for one variable, however given the combination of variables fit the data quite well. Thus whilst there is a potential for outliers, this data will not be removed.

3.5 Assumptions of Least Squares MLR

Given the application of the least squares MLR to many predictive models, it is useful to consider the assumptions of the MLR model before starting model selection. This allows for any potential problems to be corrected by transformations, removing predictors or considering alternative models.

As has been previously discussed, the distribution of 'SalePrice' does not follow a normal distribution, with significant positive skew and right tailed. In section 3.4 it is recommended to apply a transformation, such as the log transformation to the response to ensure a normal distribution for the skewed predictors. As shown in Figures 7 and 8 this transformation significant reduces the skew and kurtosis of the response variable. This process was repeated for several other categorical variables with an example shown in Figure 29 of Appendix C.

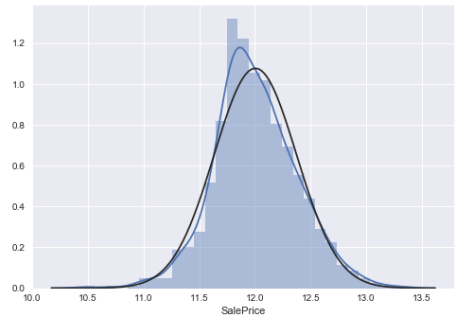

Figure 7: Distribution of Log 'SalePrice'

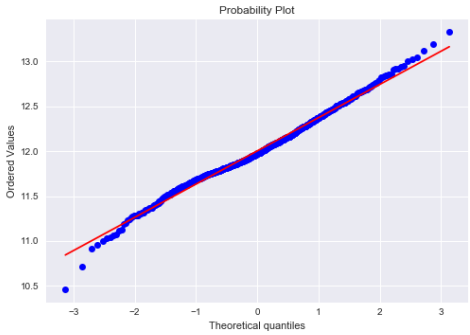

Figure 8: Q-Q Plot of the 'SalePrice'

Another key assumption which is worth mentioning is that of constant error variance. These were checked in accordance with the fitted models and will be discussed in further detail in section 4.

4 Feature Engineering

In order to improve the predictive performance of the two models feature engineering was conducted in which additional predictors were created based on the provided data. Featuring engineering has the potential to produce more simple and flexible models with more accurate predictions through the addition of extra features and transformations as appropriate (Brownlee, 2017).

The additional features created for this report were initially informed via studying the EDA results presented in the earlier section. For example, visualizing the predictor's relationship with the sale price, one could determine if a linear relationship was appropriate, or if non-linear polynomial terms should be constructed. Furthermore, additional features were created via domain knowledge. For instance, a variable representing the total floor area of the house could be constructed via the addition of the basement, and living areas. This in itself is a form of dimension reduction.

Finally, the addition of these extra features was checked to ensure that they yielded improved predictive performance. Variable selection techniques such as LASSO and forward selection were employed and the generated models compared using cross-validation. Generally, the addition of predictors via feature engineering resulted in increased predictive performance of the house prices.

4.1 Dimension Reduction

The purpose of the dimension reduction techniques, by combining existing predictors, was to simplify the complexity of the model. Per the bias-variance trade-off, the simplification of the model would result in decreased variance. For example, a non-exhaustive list of transformed variables can be seen in Table 2.

Table 2: Feature Engineering: Dimension Reduction

New Variable	**Previous Variable**	**Feature Engineering**
SimpleOverallQual	OverallQual	Transforming ten categories into three: 1 - Bad, 2 - Average, 3 - Good
OverallGrade	OverallQual * OverallCond	Creating an overall grade of the house which is the product of the house quality and condition
AllSF	GrLivArea + TotalBsmtSF	Sum of the living and basement area to create a variable representing the total area of the house

4.2 Polynomials

Analyzing the relationship between the predictor and the response, SalePrice, it was possible to determine whether a linear relationship was appropriate. For situations in which this relationship was suspected to be non-linear, higher order polynomial terms were added to the model. A non-exhaustive list of the polynomial features are shown in Table 3.

Table 3: Feature Engineering: Polynomials

New Variable	**Previous Variable**	**Feature Engineering**
GrLivArea-2	GrLivArea	Squaring the living area of the house
GrLivArea-3	GrLivArea	Cubing the living area of the house
GrLivArea-Sq	GrLivArea	Square rooting the living area of the house

4.3 Dummy Variables

Categorical variables were also transformed into dummy variables. This greatly increased the number of predictors in the model as k-1 variables were created from k classes. Therefore, care was taken to not over engineer in the dimension reduction section as to explode the number of potential predictors in the model.

4.4 Skewed Data Transformation

Finally, skewed variables were also log transformed. This resulted in predictors which more closely followed a normal distribution and therefore was more appropriated under

the OLS method assumptions. It has been suggested that variables exhibiting a skewness of greater than 0.5 are appropriate to be transformed (Doane and Seward, 2011). Using this metric, 83 features were transformed via a log transformation in order to produce more normal values for model selection.

5 Methodology

A supervised learning approach is adopted for the purposes of model building in this context, whereby the training data is used to develop a model. Parameters are chosen using a variable selection process. The model with the best predictive performance under cross-validation is selected and used to estimate the responses of the training set, and obtain a validation score.

To predict property prices, various regression models and methods have been considered. Ordinary least squares (OLS) regression was used to form an initial model. Regularization methods of least absolute shrinkage and selection operator (LASSO) regression, ridge regression and elastic net regression are considered along with the dimension reduction methods of principal components regression (PCR) and partial least squares (PLS). Variable selection techniques including best subset and forward selection are used to select predictors. Ultimately, the two best models have been selected based on their cross-validation scores on the training data.

5.1 Model Selection Criterion

The different models considered were compared using a cross-validation root mean squared error (RMSE) score.

Cross validation is based on multiple random data splits of the training data. Given the computational cost increases for larger number of splits and the large number of predictors involved, a 5-fold cross validation was used for these models. For each fold (split of the training data), the model is estimated on all other folds combined, with the fold in question used as the validation set. The cross validation error is taken as the average mean squared error across the 5 validation sets.

5.2 Modelling

Whilst initially several different models were considered, based off the cross-validation scores the models created using LASSO and elastic net were selected.

5.2.1 Justification for Models Considered

Notably a variety of other models could have been considered and analyzed. However considerations of computation intensity, interpretability and domain knowledge results in narrowing down potential models to those considered below.

Only parametric methods have been considered in this analysis mainly due to computational and interpretation considerations. With non-parametric methods, the number of parameters grows with the size of the training data. Given the significant number of initial predictors as well as the large sample size, these methods become computationally infeasible as well as difficult to interpret. Furthermore, given the high-dimensional input, it is likely that non-parametric methods would break down under the curse of dimensionality.

This analysis has further only considered linear models. Based off the original exploratory data analysis, it was seen that the relationship between sales price and other predictors was primarily linear. In the feature engineering section, log transformations were made on data which appeared to have some non-linearities to make these predictors more suitable for application to linear models. Furthermore, from a bias-variance trade-off perspective, considering only linear models reduces the amount of variance without severely increasing the bias.

5.2.2 Baseline Multiple Linear Regression (Baseline MLR)

This model is considered as a benchmark for comparing other models. The MLR is built upon the six basic assumptions of linearity, exogeneity (conditional mean of errors zero), constant error variance, independence of error pairs, arbitrary distribution of input values and no perfect multicollinearity. Often one or more of these assumptions are violated. In particular the model assumes a linear function form, which might not capture the true relationship between a predictor and response. Furthermore, given the large number of predictors, not only is the interpretability of coefficients difficult, but there is potentially high variance from the model complexity. Hence, whilst the MLR provides a strong foundation, it is rarely sufficient to model complex relationships, hence the need for variable selection, regularization and dimension reduction methods.

5.2.3 Variable Selection

Variable selection involves identifying a subset of predictors (k) from the full data set, which optimizes the bias-variance trade-off.

Forward Selection

Given the large data set, best subset selection (where all possible permutations of models with k predictors are considered) is computationally infeasible. Thus the stepwise selection process of forward selection was used to find the optimal predictors. Starting with the null model (intercept only), further models are created by adding the predictor which has the smallest RSS/highest R^2. The cross-validation score of the various models is thus used to identify the single best model. Whilst this method is not guaranteed to find an optimal solution leading to higher variance, it is able to reduce the number of predictors leading to higher interpretability and accuracy.

5.2.4 Regularization Methods

Regularization methods are used to shrink the coefficients towards zero, improving the prediction accuracy of the model by reducing variance. The three different regularization methods considered follow the no free lunch theorem, that is neither outperform the other, leading to relying on cross-validation to determine the best approach.

Ridge regression

Ridge regression is a common regularization method which aims to minimize the residual sum of squares (least squares method), with the addition of a shrinkage penalty $\lambda\Sigma\beta_j^2$ which shrinks coefficients β_j towards zero. The hyperparameter λ, which controls the degree of shrinkage, is chosen based on cross-validation error for a range of values considered. The ridge regression shrinks, by the same amount, coefficients of similar importance, which reduces the effect of significant correlation between predictors. Notably however, the ridge regression does not perform variable selection, meaning interpretability of the model may still pose a problem.

Least Absolute Shrinkage and Selection Operator (LASSO)

The LASSO is an alternative regularization method which has the penalty term $\lambda\Sigma|\beta_j|$ in addition to the residual sum of squares. This term not only shrinks the coefficients towards zero, but for certain values of λ will force some coefficients to be zero, thus performing both variable selection and regularization. The hyperparameter λ is similarly chosen from a range of potential values by cross-validation. As LASSO has the ability to select variables it is more suited to models where a small subset of predictors are significant to predicting the response. However, removing variables runs the risk of excluding important predictors.

Elastic Net

The elastic net method offers a compromise between the LASSO and ridge regression by retaining the variable selection property of LASSO whilst shrinking together coefficients of correlated predictors, in a similar fashion to ridge regression. However, as per the free lunch theorem elastic net is not guaranteed to outperform ridge and LASSO.

5.2.5 Dimension Reduction Methods

Dimension reduction methods reduce the number of variables in the model by transforming the original predictors, commonly by creating linear combinations of predictors. All dimension reduction methods are conducted on centered and standardized predictors to remove the effects of varying scales.

Principal Component Regression (PCR) Principal component regression involves running a regression on the set of principal components which capture the most variation in the predictor data. Principal components are linear combinations of predictors that maximizes the sample variance. The number of principal components used is selected based on a cross-validation approach. PCR can lead to substantial variance reduction as it assumes the selected components account for a large variation in the response.

Partial Least Squares (PLS)

The partial least squares method involves identifying the best linear combinations of parameters in a supervised manner, but considering the strength of the univariate effect of predictors on the response. Each predictor is assessed based on a simple linear regression (SLR) between itself and the response. Then, each linear combination is computed with the number of regressors in each component equal to or less than the maximum limit. A SLR is then run between each linear combination and the response. The best model is found using a cross-validation approach.

5.3 Model selection

Out of the the seven models developed, the elastic net and the LASSO have been chosen for further analysis. These models performed well on the cross-validation root mean squared error test conducted on the training data (model selection) as discussed in section 6. Furthermore, they have the lowest mean absolute error scores on 50% of the test data as trailed through Kaggle (model evaluation). Finally these models both perform variable selection which reduces the variance of the model, as well as the chance of over-fitting to the training data. The details of these models are summarized below.

5.3.1 Model 1: LASSO

The LASSO model, as previously discussed, performs both regularization and variable selection. For the LASSO model the selection of the hyper-parameter λ, which controls

the shrinkage (and associated variable selection) will thus influence the performance of the model. Using the 5 fold cross-validation root mean squared error for a range of lambda between 0 and 0.30, the optimal λ is equal to 0.0007578. The variation in RMSE over the range of λ considered are shown in Figure 9.

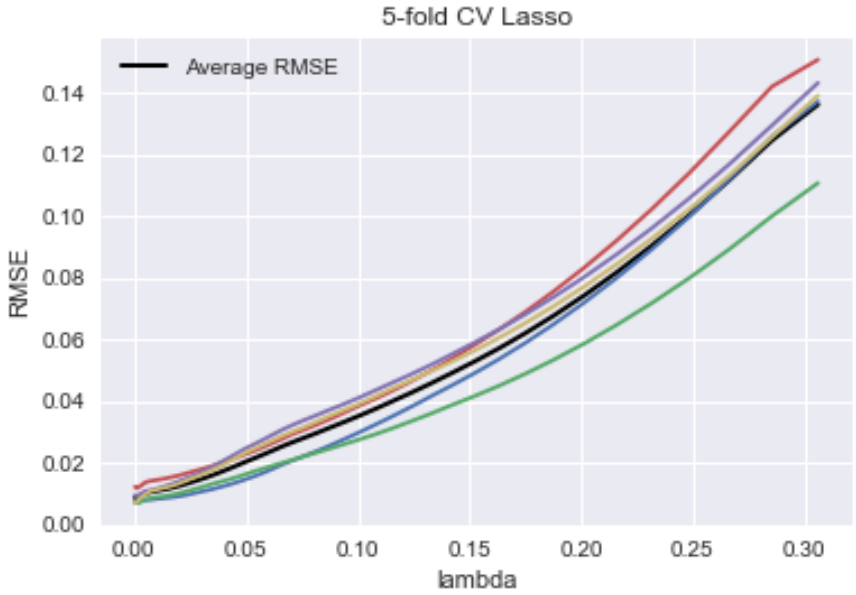

Figure 9: Cross-Validation Results for the LASSO Model

The LASSO model limits the number of coefficients used to predict property prices to 104 features, eliminating the other 304 features, a significant reduction on the initial set. The most significant coefficients included in the model are shown in Figure 10.

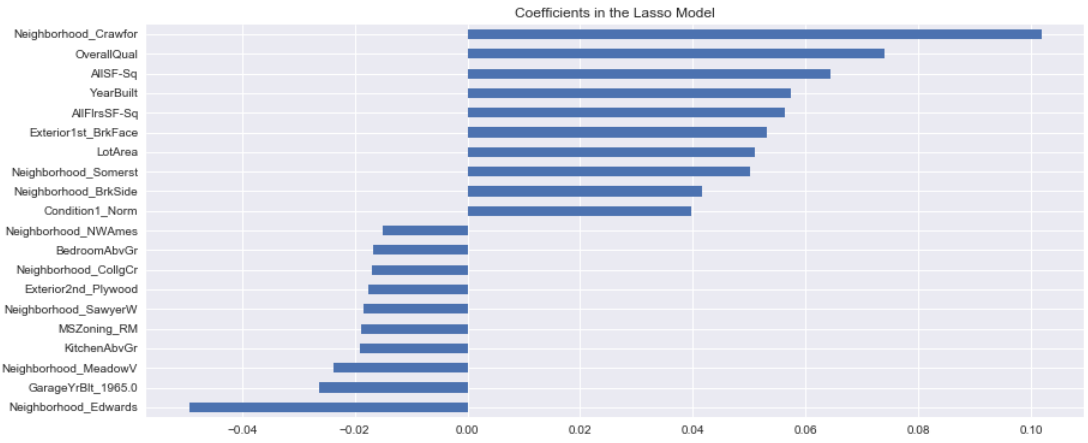

Figure 10: Most Significant Coefficients for the LASSO Model

The significance of parameters is determined by the magnitude of the estimated coefficient, whereby a larger coefficient indicates a unit change in the predictor leads to a larger change in the response. Important to note here however is the presence of dummy variables (like Neighborhood_Crawford), which can only take the value of 0 or 1. In com-

parison variables such as YearBuilt and OverallQual can take on a larger range of values. Thus the coefficients can have a larger impact on the overall property price.

The greatest benefit of the LASSO regularization is derived from its ability to perform both variable selection and regularization. Removing predictors of lower significance leads to lower model complexity, subsequently resulting in lower bias and smaller prediction errors. As LASSO is a continuous method regularization, the variance of the model is reduced compared to variable selection techniques. The use of fewer predictors also results in an increase in interpretability of the model.

Furthermore, a number of assumptions are made in the OLS component of the LASSO model which appear to be met verified by diagnostics contained within Appendix D. Firstly, the assumptions of linearity and exogeneity are met as there appears to be no pattern in the residuals in Figure 32. Additionally, the assumption of constant error variance is upheld as there is no systematic pattern in the values of the absolute residuals in Figure 33. Moreover, the distribution of the residuals does appear normal according to Figure 34 with only skewness at extreme positive values. Finally, the highly correlated predictors were removed from the model prior to the model fitting stage and therefore there should be no issues due to multicollinearity. Hence, the LASSO model should be appropriate.

5.3.2 Model 2: Elastic Net

The elastic net, in a similar fashion to the LASSO model performs both regularization (in a similar fashion to the ridge regression) and variable selection (in a similar way to LASSO). The elastic net requires both the selection of the shrinkage parameter λ as well as a complexity parameter α. Using inbuilt cross-validation functions in Python, the optimal α was determined to be 0.095. The optimal λ was found using 5 fold cross-validation based on the RMSE. The variation over a selection of λ values is shown in Figure 11.

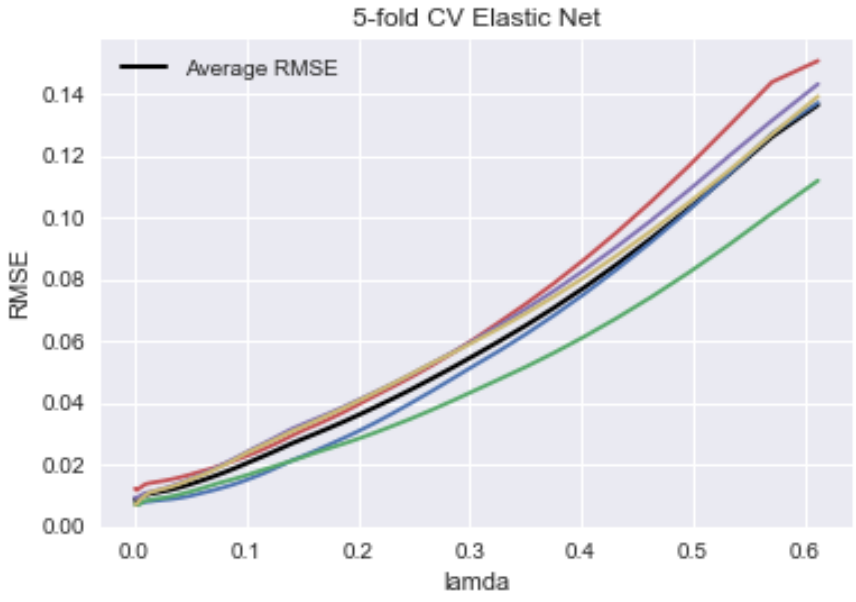

Figure 11: Most Significant Coefficients for the Elastic Net Model

Due to the ability of elastic net to perform variable selection, the resulting model

has only 123 features, eliminating 285 features. The most significant coefficients are summarized in Figure 12.

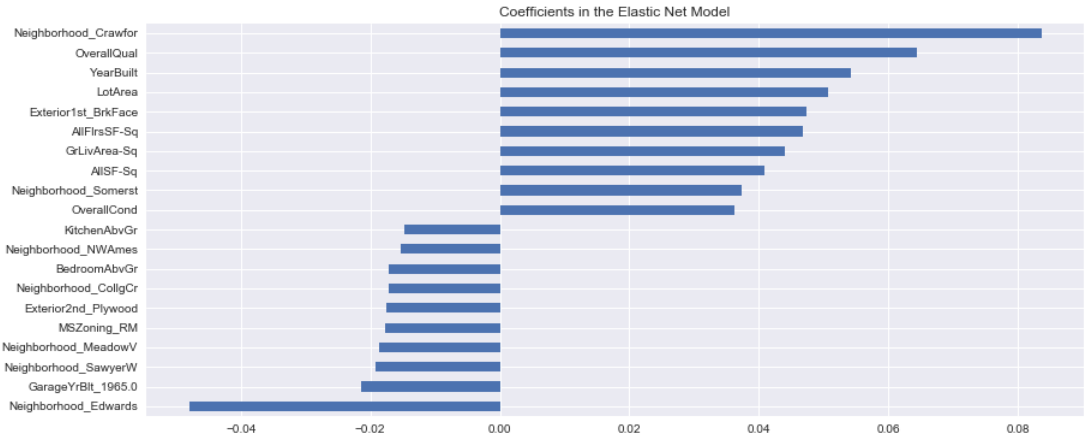

Figure 12: Most Significant Coefficients for the Elastic Net Model

The significance of these parameters likewise determined by the size of the coefficients has a few interesting trends. Compared to the LASSO model, the discrete and continuous variables OverallQual, YearBuilt and LotArea are more significant. Notably the dummy variables still feature heavily, particularly those associated with the neighborhood location.

The elastic net model, in a similar method to the LASSO model results in a model with lower bias and smaller prediction errors by reducing model complexity. Compared to the LASSO model, the elastic net has not eliminated as many features, potentially leading to higher variance. However the elastic net incorporates a sum of squares penalty on the predictors, potentially reducing variance depending on the collinearity of the data. In a similar fashion to LASSO, the exclusion of predictors results in an increased interpretability of the model.

Finally, the assumptions of the elastic net model were verified by diagnostics contained within Appendix D. Firstly, Figure 36 shows no signs of systematic patterns in the residuals so the assumptions of linearity and exogeneity are met. Furthermore, the assumption of constant error variance holds as the values of the absolute residuals in Figure 37 shows no obvious patterns. Furthermore, Figure 38 shows that the residuals are generally normally distribution with only slight positive skewness on the right tail. In addition, the highly correlated predictors were removed from the model prior to the model fitting stage and therefore there should be no issues due to multicollinearity. Hence, the elastic net model should be appropriate.

6 Validation Results

Following the methodology outlined in the previous section a series of models were constructed. In order to select which two models would be submitted in the Kaggle competition the models were first analyzed using the training data as part of the model selection process. Once this had been completed the models were then evaluated on 50% of the test data as part of the model evaluation process. At the conclusion of these stages it was proposed that the LASSO and elastic net models be used in the final submission of

the competition. Furthermore, consideration of the bias-variance trade-off was also made to ensure a model of an appropriate complexity was selected. Finally, attention was also made to verify that the proposed models had not been over-fit on the training data and therefore would perform poorly on the test data. The RMSE CV results for the models are shown in Table 4.

6.1 Model Selection Results

The model selection stage was conducted using the training data provide. This allowed the generalization performance to be estimated, allowing models of varying complexity to be compared in terms of their estimated predictive capability. This was estimated by calculating the cross-validated RMSE using ten folds. This method was selected over the likes of AIC and BIC as it is a more universally applicable approach not requiring strict assumptions of the model, such as constant error variance. The cross validation method also provides a direct estimate of the test error.

Table 4: Model Selection: RMSE CV

Model	**CV RMSE**
OLS	251.157
Forward	0.080
Ridge	0.092
LASSO	0.092
Elastic Net	0.091
PCR	0.244
PLS	0.234

Using Table 4 it is immediately obvious the models which will not be appropriate for the final submission. Firstly, the OLS model performed poorly on the CV RMSE test, likely due to over-fitting of the K-1 training set. Consequentially, the loss function was exaggerated on the Kth validation set for each iteration of the cross validation. However, multicollinearity issues may have also been present, despite highly correlated variables being removed from the data in the preprocessing stage. Furthermore, both PCR and PLS performed poorly, indicating that dimension reduction techniques are inappropriate for this problem. This may be due to the variation in the data not being fully captured in the principle components.

Furthermore, it was also found that the models performing regularization performed well in the cross-validation. Ridge, LASSO and elastic net each had similar RMSE scores. The addition of the complexity penalty had the effect of reducing the variance the predictions and may have alleviated any issues due to multicollinearity.

Finally, Table 4 also indicates that variable selection is an important aspect of the model selection process. Forward selection, elastic net and LASSO had the best CV RMSE on the training data and therefore particular attention was paid to these models during the model evaluation stage.

6.2 Model Evaluation Results

The models trained using the training data were also evaluated on 50% of the test data. The scoring metric used in the Kaggle competition is the mean absolute error (MAE) of the predictions. Due to the limitations that Kaggle imposes for the maximum number

of data uploads only the four best performing models in the CV RMSE in the previous section were evaluated. The results as of 20th September 2017 are contained within Table 5.

Table 5: Model Evaluation - MAE

Model	**MAE**
Forward	12251.394
Ridge	11898.769
LASSO	11769.762
Elastic Net	11795.783

From Table 5 it can be observed that the forward selection model performed the worst on the test data, contrary to the results obtained using the training data. This is likely due to the model, created using variable selection techniques, choosing the optimal predictors which over-fit the training data. As a result, it has a poor generalization performance on the test data.

The model created using Ridge regression could have potentially been selected based on the results in Tables 4 and 5. However, not performing variable selection is a drawback of this method. As observed in the EDA, not all variables are as important as one another in predicting the house price and therefore the inclusion of all predictors in the model is likely to generate additional noise, rather than predictive signal. It is expected that this model would not perform as well as the models created using LASSO and elastic net on the remaining 50% of the test data.

Hence, it is proposed that the models created using LASSO and elastic net are used in the competition as both these methods include variable selection and regularization properties. Only selecting a subset of the predictors will reduce the model complexity and under the bias-variance trade-off will reduce the amount of variance. Furthermore, the regularization properties of each ensure that the most important predictors are more heavily weighted in the model. Combining these features is likely to result in good generalization on the remaining test data.

7 Conclusion

This report details the process for developing two predictive models which would allow city governments to make more informed decisions regarding property prices in their area. The models with the best predictive performance, based on their cross-validation scores using the training data (root mean squared error) and their predictive accuracy on the test data (mean absolute error), were developed using the elastic net and the LASSO approaches.

Both approaches yielded models with a cross-validation Root Mean Square Error (CV RMSE) of 0.091 and 0.092 during the model selection phase using the training data, which was significantly lower than their rivals (with the exception of the forward selection regression model). The MAEs for both the EN and LASSO models were $11,795.78 and $11,769.76 representing the average expected deviation of the predicted values from the actual values (measured on half of the test set).

The relative success of both model approaches can be attributed to their variable selection properties, which enabled both of them to produce sparser models without compromising on variance. Given the bias-variance trade-off inherent to predictive modeling, lower model complexity tends reduce the risk of over-fitting to the data, increasing bias but reducing the variance. However, both the LASSO and EN address this issue by optimizing predictive accuracy with respect to both bias and variance. These models are also more interpretable as the number of variables in the model are reduced, allowing councilors more easily understand the models they are using. Therefore, these two proposed models should provide an appropriate method for determining property prices for local city governments.

8 References

Brownlee, J. (2017). *Discover Feature Engineering, How to Engineer Features and How to Get Good at It - Machine Learning Mastery.* [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/ [Accessed 15 Sep. 2017].

Doane, D. and Seward, L. (2011). Measuring Skewness: A Forgotten Statistic?. *Journal of Statistics Education*, 19(2), pp.1-18.

A Further Information on Dataset

Predictor	Description	Type
MS SubClass	Type of dwelling	Nominal
MS Zoning	General zoning classification	Nominal
Lot Frontage	Linear feet of street connected	Continuous
Lot Area	Lot size in square feet	Continuous
Street	Type of road access	Nominal
Alley	Type of alley access	Nominal
Lot Shape	General shape of property	Ordinal
Land Contour	Flatness of property	Nominal
Utilities	Type of utilities available	Ordinal
Lot Configuration	Lot configuration	Nominal
Land Slope	Slope of property	Ordinal
Neighborhood	Physical locations within city	Nominal
Condition 1	Proximity to conditions	Nominal
Condition 2	Proximity to conditions (additional)	Nominal
Building Type	Type of dwelling	Nominal
House Style	Style of dwelling	Nominal
Overall Quality	Rates material and finish	Ordinal
Overall Condition	Rates overall condition	Ordinal
Year Built	Original construction date	Discrete
Year Renovations	Remodel date	Discrete
Roof style	Type of roof	Nominal
Roof material	Roof material	Nominal
Exterior 1	Exterior covering	Nominal
Exterior 2	Exterior covering (additional)	Nominal
Masonry Veneer Type	Type of veneer	Nominal
Masonry Veneer Area	Veneer area in square feet	Continuous
Exterior Quality	Evaluates the quality of the material on the exterior	Ordinal
Exterior Condition	Evaluates the present condition of the material on the exterior	Ordinal
Foundation	Type of foundation	Ordinal
Basement Quality	Evaluates the height of the basement	Ordinal
Basement Condition	Evaluates the general condition of the basement	Ordinal
Basement Exposure	Refers to walkout or garden level walls	Ordinal
Basement Finish 1	Rating of basement finished area	Ordinal
Basement Finish 1 SF	Type 1 finished square feet	Continuous
Basement Finish Type 2	Rating of basement finished area (if multiple types)	Ordinal
Basement Finish SF 2	Type 2 finished square feet	Continuous
Basement Unfinished SF	Unfinished square feet of basement area	Continuous
Total Basement SF	Total square feet of basement area	Continuous

Table 6: Variables in the Data 1

Predictor	Description	Type
Heating	Type of heating	Nominal
Heating QC	Heating quality and condition	Ordinal
Central Air	Central air conditioning	Nominal
Electrical	Electrical system	Ordinal
1st Flr SF	First Floor square feet	Continuous
2nd Flr SF	Second Floor square feet	Continuous
Low Qual Fin SF	Low quality finished square feet (all floors)	Continuous
Gr Liv Area	Above grade (ground) living area square feet	Continuous
Bsmt Full Bath	Basement full bathrooms	Discrete
Bsmt Half Bath	Basement half bathrooms	Discrete
Full Bath	Full bathrooms above grade	Discrete
Bedroom	Bedrooms above grade (does NOT include basement bedrooms)	Discrete
Kitchen	Kitchens above grade	Discrete
KitchenQual	Kitchen quality	Ordinal
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	Discrete
Functional	Home functionality (Assume typical unless deductions are warranted)	Ordinal
Fireplaces	Number of fireplaces	Discrete
FireplaceQu	Fireplace quality	Ordinal
Garage Type	Garage location	Nominal
Garage Yr Blt	Year garage was built	Discrete
Garage Finish	Interior finish of the garage	Ordinal
Garage Cars	Size of garage in car capacity	Discrete
Garage Area	Size of garage in square feet	Continuous
Garage Qual	Garage quality	Ordinal
Garage Cond	Garage condition	Ordinal
Paved Drive	Paved driveway	Ordinal
Wood Deck SF	Wood deck area in square feet	Continuous
Open Porch SF	Open porch area in square feet	Continuous
Enclosed Porch	Enclosed porch area in square feet	Continuous
3-Ssn Porch	Three season porch area in square feet	Continuous
Screen Porch	Screen porch area in square feet	Continuous
Pool Area	Pool area in square feet	Continuous
Pool QC	Pool quality	Ordinal
Fence	Fence quality	Ordinal
Misc Feature	Miscellaneous feature not covered in other categories	Nominal
Misc Val	Value of miscellaneous feature	Continuous
Mo Sold	Month Sold (MM)	Discrete
Yr Sold	Year Sold (YY)	Discrete
Sale Type	Type of sale	Nominal
Sale Price	Sale Price $	Continuous

Table 7: Variables in the Data 2

B Additional EDA

Covariance of Predictors

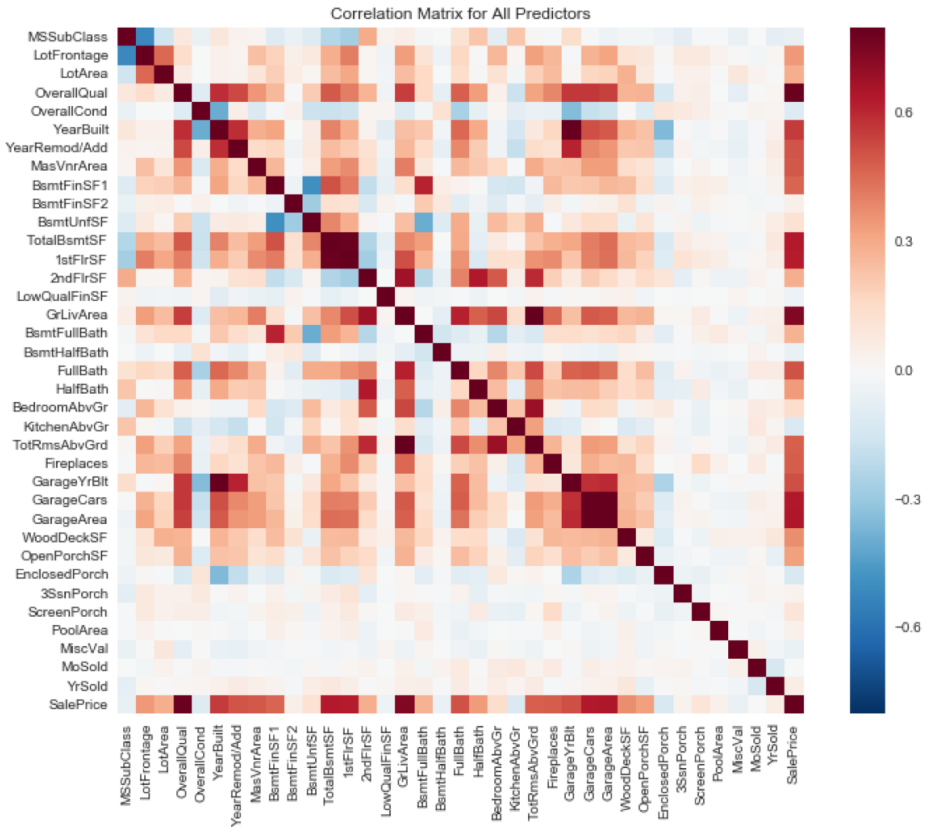

Figure 13: Correlation Matrix for All Predictors

Figure 14: Correlation Matrix with Scatter Plots for Strongly Correlated Predictors

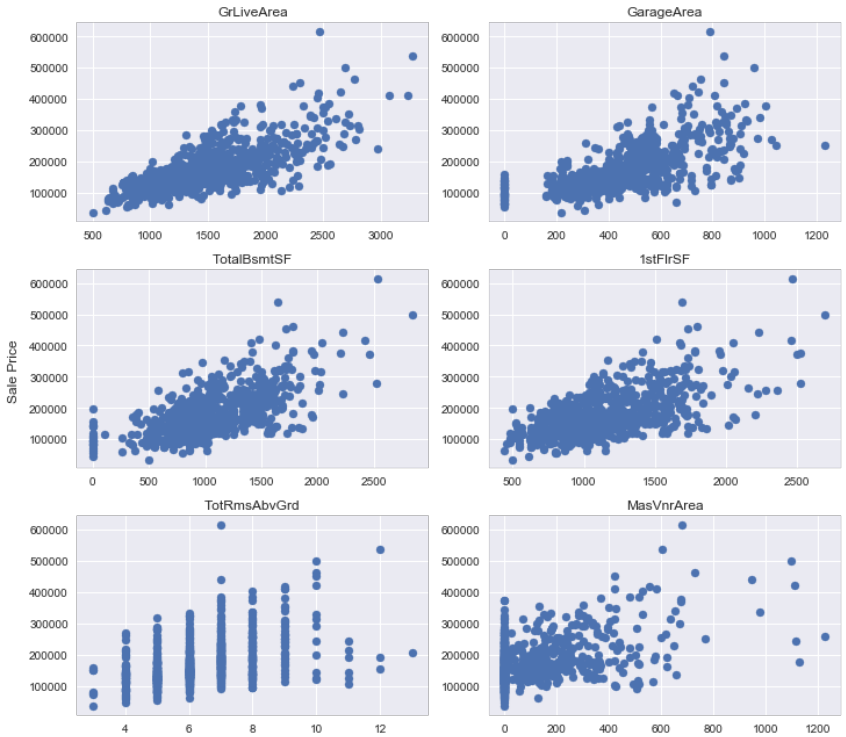

Figure 15: Scatter Plots of Highly Correlated Variables

Further Distribution of Variables

Figure 16: Frequency Histogram of Sales Price

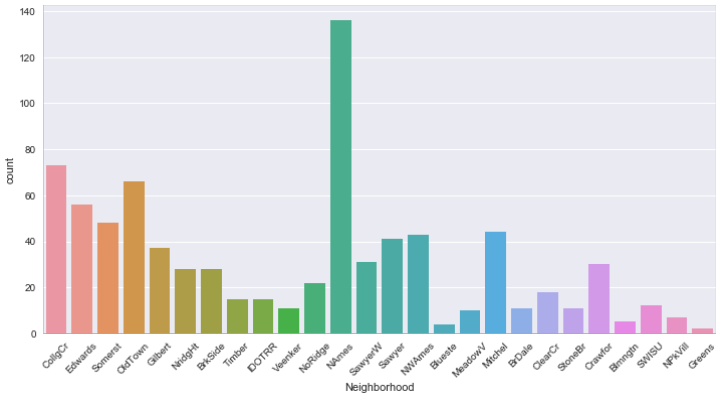

Figure 17: Frequency Histogram of Different Neighborhoods

Further Bivariate Analysis

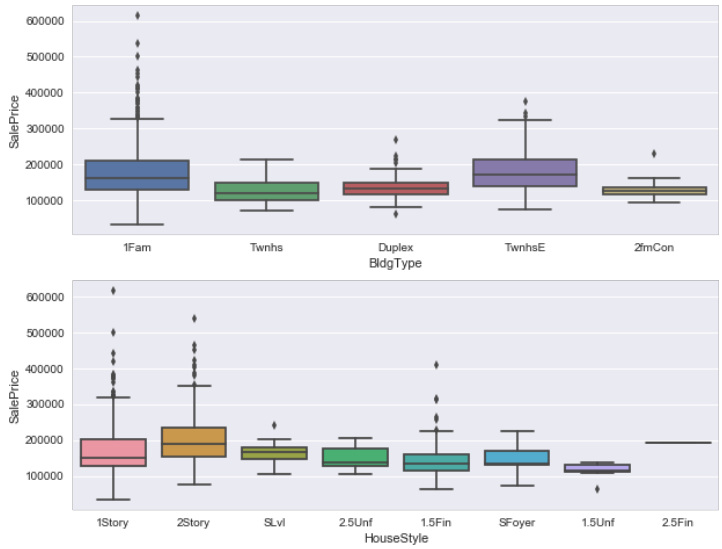

Figure 18: Box Plots of Sale Price vs Building Type and House Style

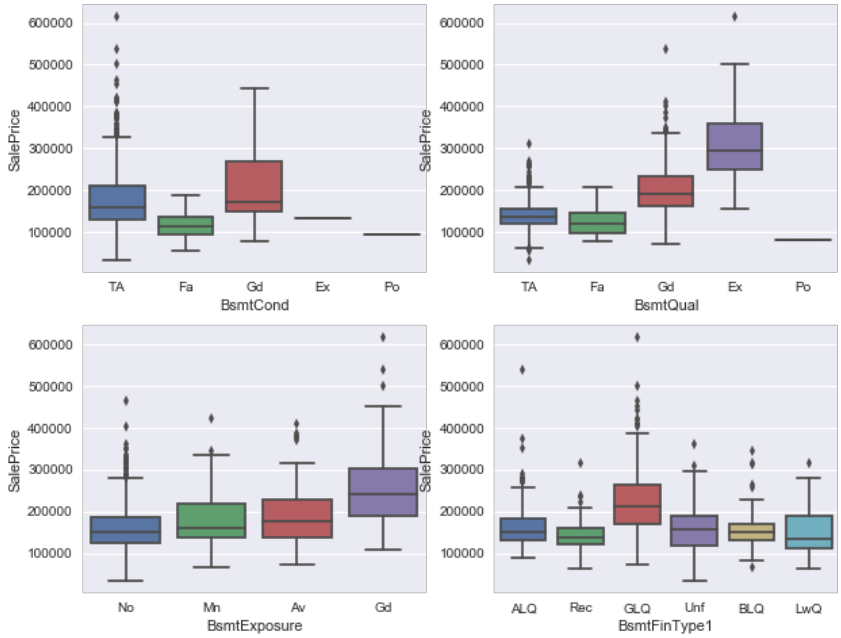

Figure 19: Box Plots of Sale Price vs Basement Variables

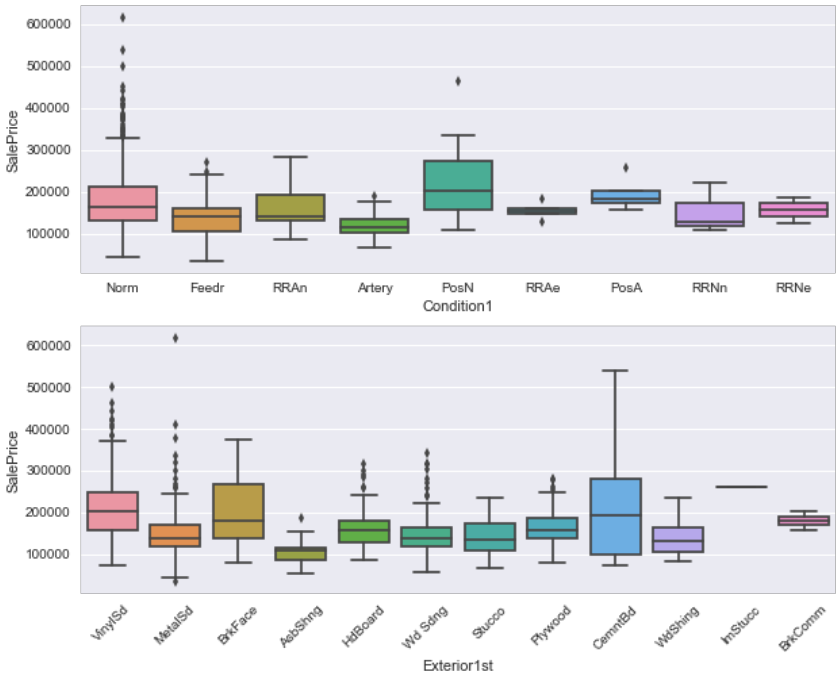

Figure 20: Box Plots of Sale Price vs Condition of the Property and Exterior Material

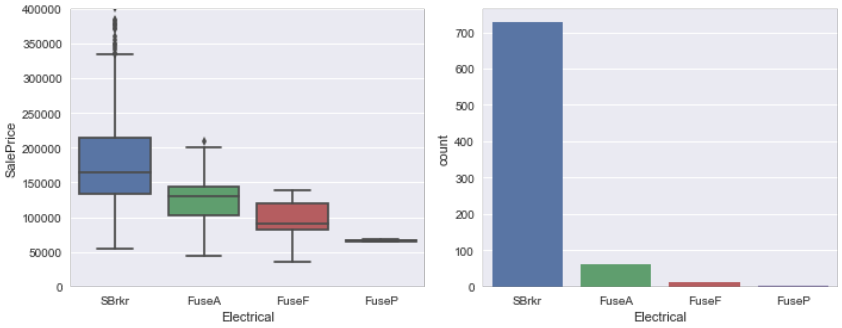

Figure 21: Box Plots of Sale Price vs Electrical Variables

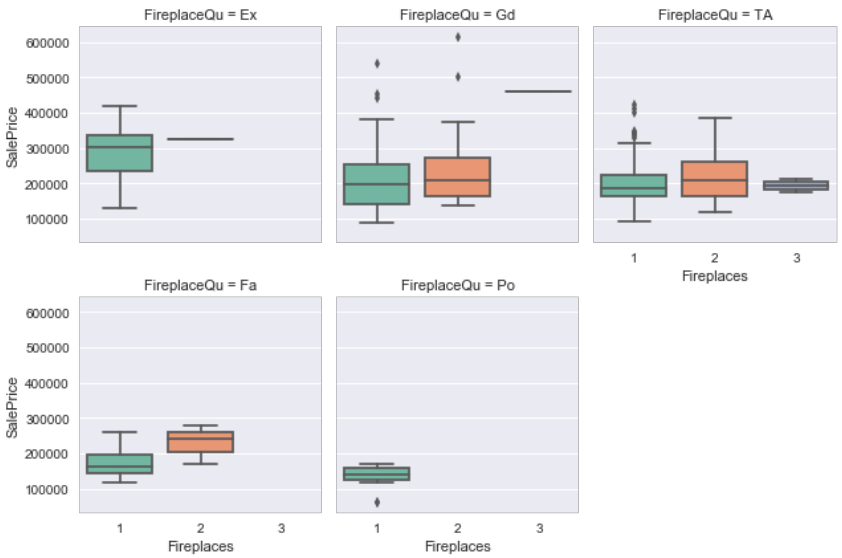

Figure 22: Box Plots of Sale Price vs Fireplace Quality

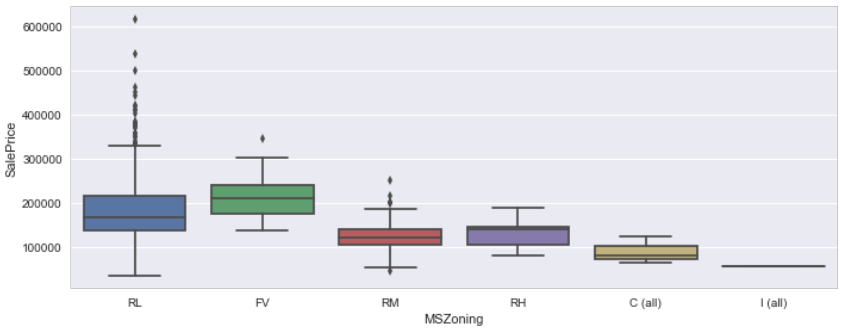

Figure 23: Box plots of Sale Price vs Zoning Types

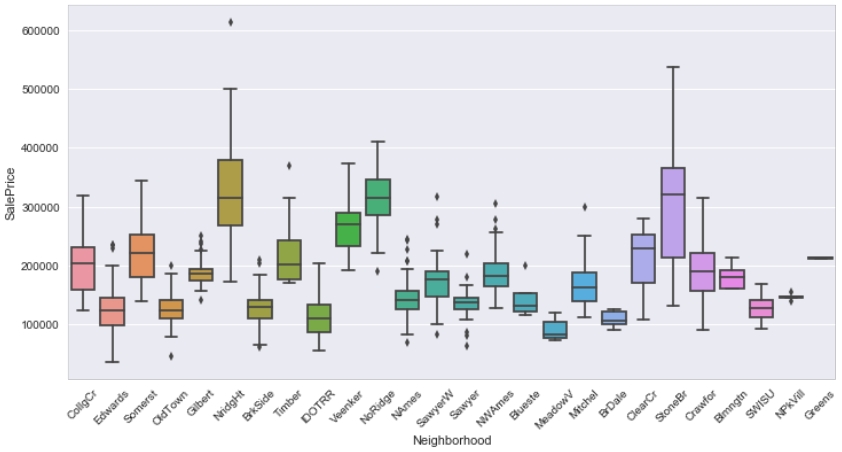

Figure 24: Box Plots of Sale Price vs Neighborhood

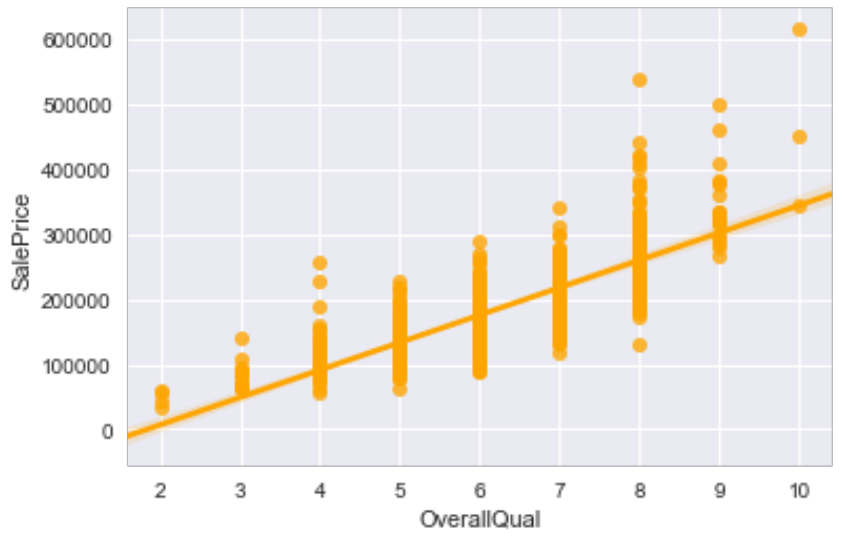

Figure 25: Regression plot of Sale Price vs Overall Quality

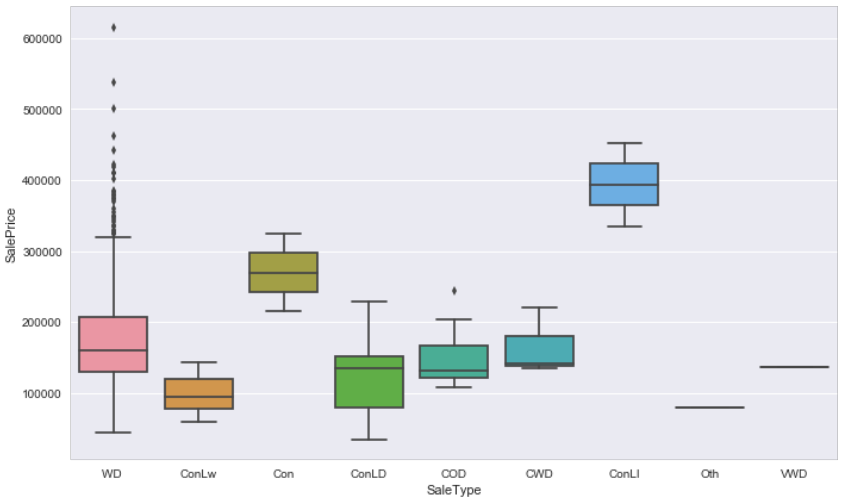

Figure 26: Box Plots of Sale Price vs Sale Type

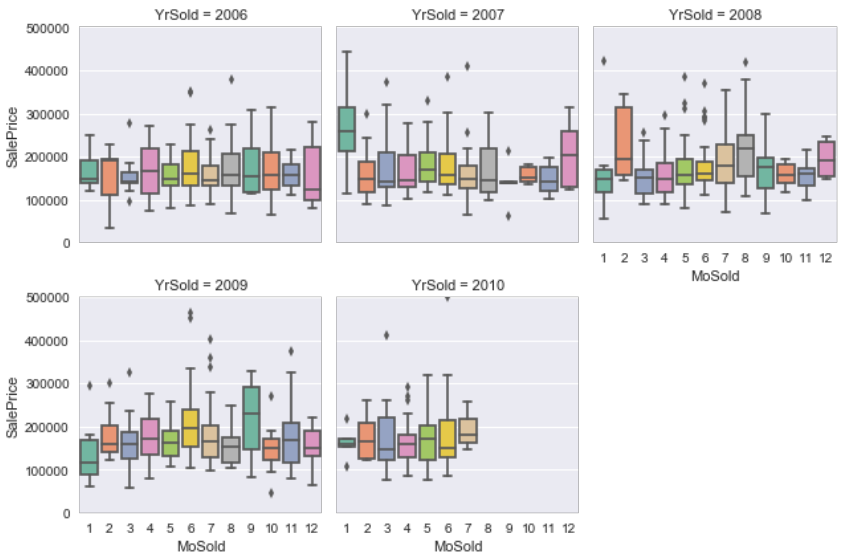

Figure 27: Box Plots of Sale Price vs Year Sold (2006-2010)

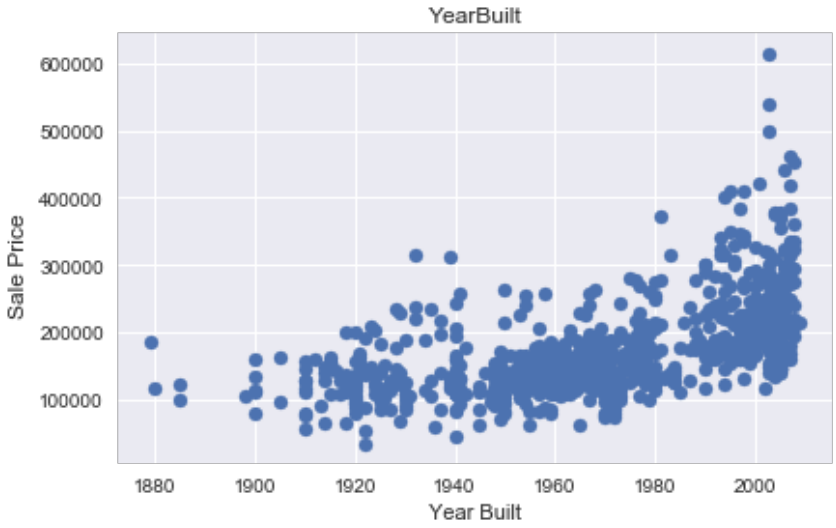

Figure 28: Box Plots of Sale Price vs Year Built

C Log Transformation Examples

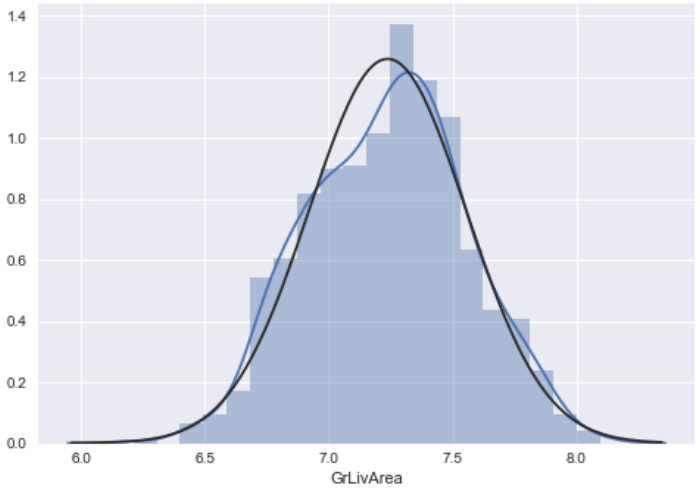

Figure 29: Distribution of the 'GrLivArea'

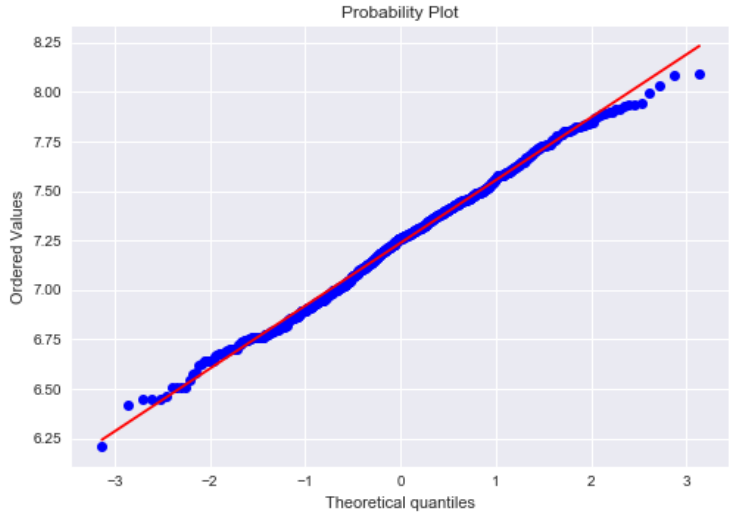

Figure 30: Q-Q Plot of the 'GrLivArea'

D Training Predictions and Residuals

LASSO Model

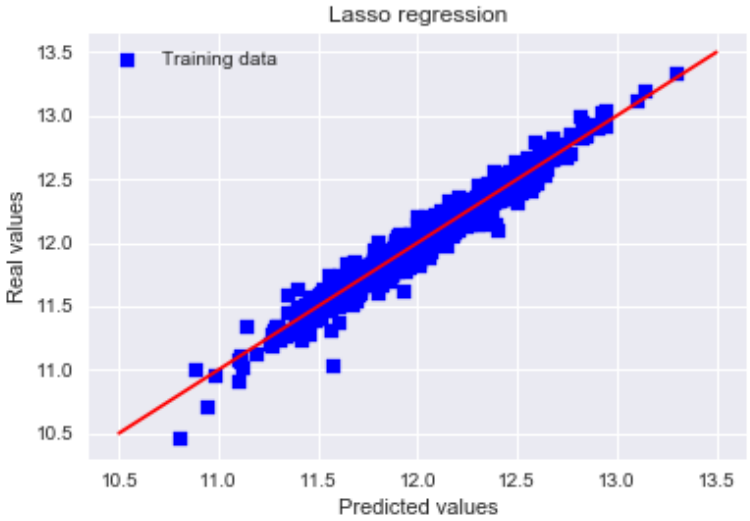

Figure 31: LASSO Predictions on Training Data

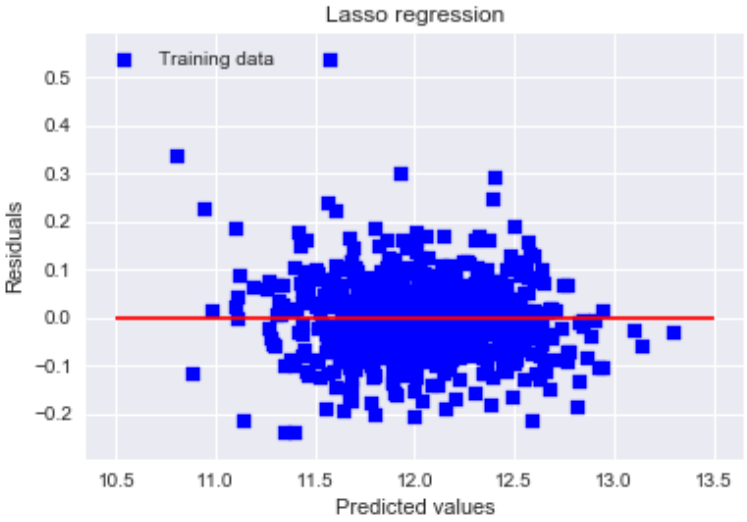

Figure 32: LASSO Residuals on Training Data

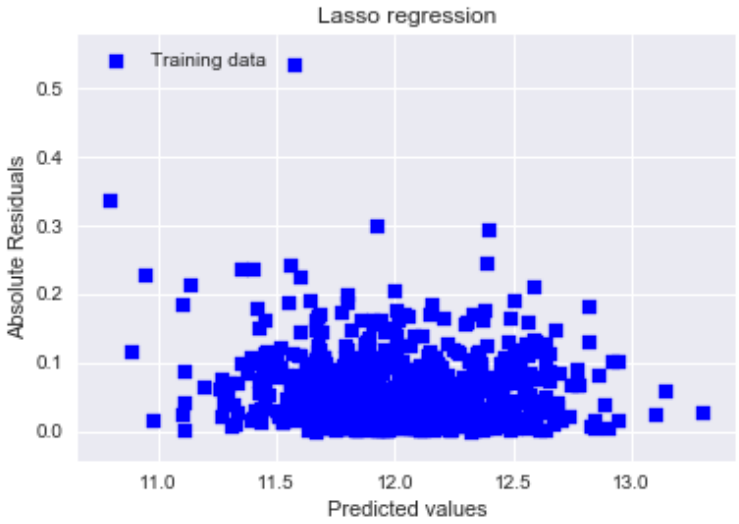

Figure 33: LASSO Absolute Residuals on Training Data

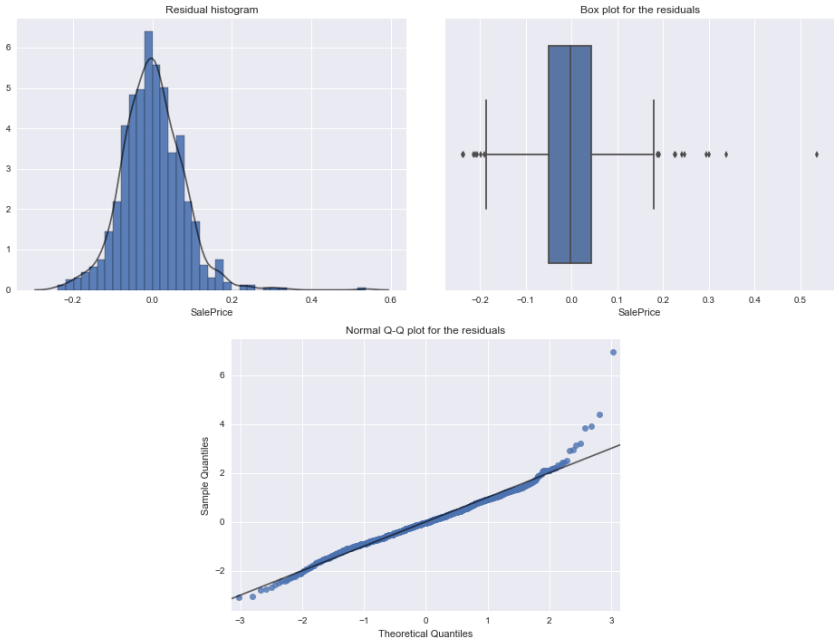

Figure 34: LASSO Residual Distribution on Training Data

Elastic Net Model

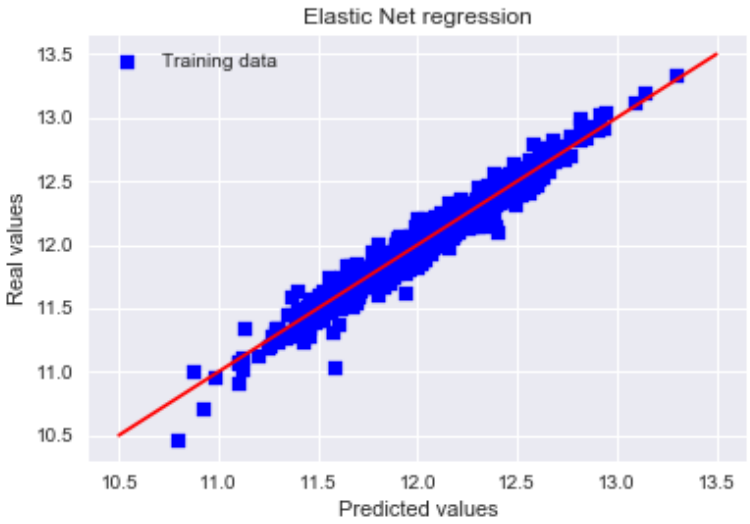

Figure 35: Elastic Net Predictions on Training Data

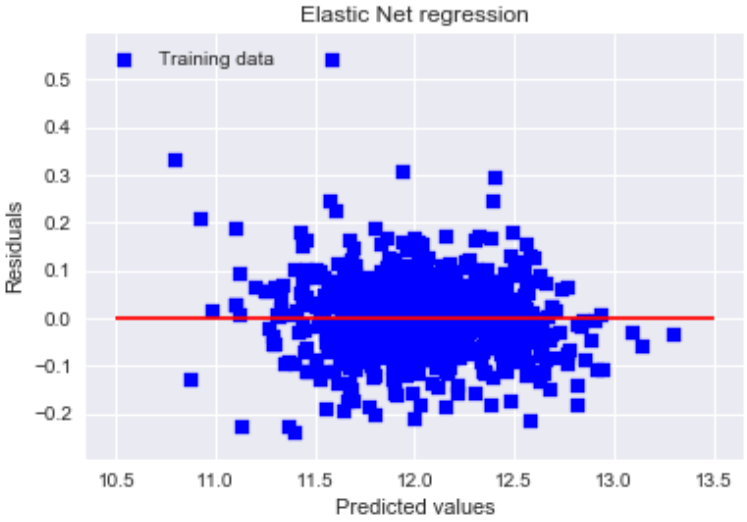

Figure 36: Elastic Net Residuals on Training Data

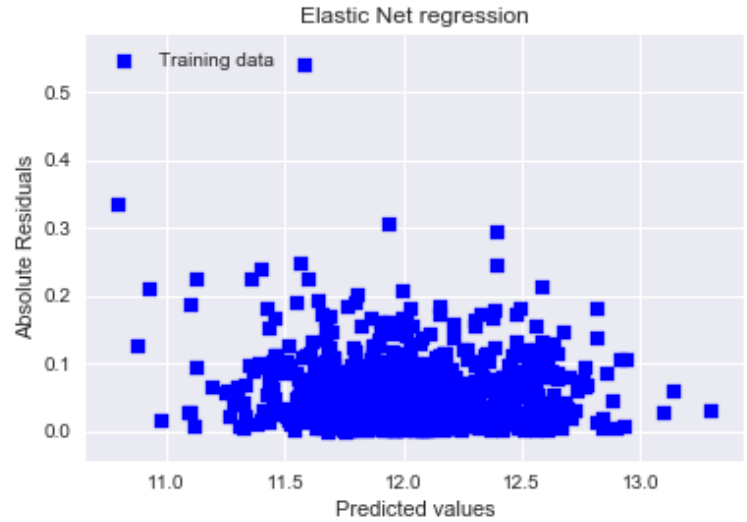

Figure 37: Elastic Net Absolute Residuals on Training Data

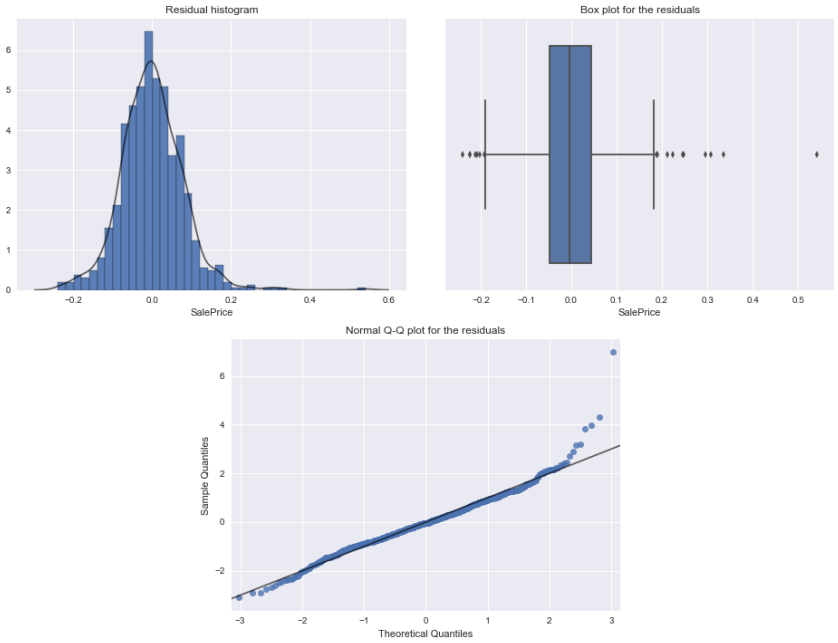

Figure 38: Elastic Net Residual Distribution on Training Data