Predictive Model Assessment for House Prices

Michelle Cleary, Ellen Crombie, Fionnuala Marshall

Executive Summary
Within this report we have created quantitative models which predict the sale price of a property based on information about the property. The models were formulated using data from the sale of houses in an American city over a 5 year period, which includes information about the features of each house, such as house style and lot area. Throughout this analysis we have dedicated a special focus to thorough validation techniques in order to ensure robustness and completeness of reported results.

1. The initial model uses all 29 house features from the dataset, and on average will predict sale price with an error of $20024. To highlight just one result, this model determined that increasing the lot area and number of bedrooms is associated with a higher sale price.
2. One model improvement can be obtained by accounting for dependencies between features. For example, the number of kitchens adjusts the effect of ground living area.
3. Further analysis indicates that using 3 features alone leads to satisfactory model performance, but using 29 produces the best model performance, with an error of $19543. Within the dataset provided, on average the 3 features with the biggest impact on sale price are overall quality, total basement area and ground living area.

Within the second half of the report we developed 3 further models which evaluated several extra features, such as roof material and lot frontage. Under analysis using machine learning techniques, we discovered that the use of different modelling techniques and extra features, such as 1st and 2nd floor square footage, had significant benefit in reducing the error of our previous models. Overall, we have recommended a model which extracts the most relevant features and predicts house prices with an overall error of $17909.

Introduction

We will first formulate a simple linear model to the data using all 29 available predictor features before assessing the impact on performance of different model variations. The initial dataset used to create our models includes 1460 observations based upon 31 features. The data focuses on the sale price of numerous houses in an American city over a 5 year period and included information about features of each house, such as lot area in square feet and neighbourhood. Within the second half of this report, a second dataset is introduced which includes additional features such as roof material and second floor square footage. We will evaluate the performance of each model under leave-one-out cross validation by measuring the mean absolute error of the prediction and the actual sale price.

Preprocessing of the Dataset

When carrying out initial exploration of the data, we looked for outliers and any features which had a significant level of missing values. Lot area appeared to contain some values outside of the usual range. However, these correlated with other features of the properties, and thus, we did not feel it was appropriate to remove them entirely from our dataset at this stage. Within the features describing garage type and electrical system, 5.55%

and 0.07% were missing respectively. We were able to remove any entries corresponding to these values since they only represented a small proportion of entries in the dataset (0.18% in total) and so did not majorly reduce the number of observations.

Next, we coded factor features and made simplifications based on similarity of level descriptions. The house style levels "1.5 story finished" and "1.5 story unfinished" were combined into "1.5 story", and similarly for "2.5 story". We decided that features such as neighbourhood, which had a high granularity, could not be simplified any further because the dataset did not contain any further information or context for us to do so.

Initial Linear Model

We created a linear regression model using all 29 predictor features from dataset of 1378 observations. We defined y_i as the sale price for each house $i = 1, ..., n$, and defined the following model:

$$y_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

where ϵ_i follows a normal distribution with mean zero and variance σ^2. The vector $\boldsymbol{x}_i$ is a row of the model matrix for observation i containing all the features, with dummy variables for factor features such as neighbourhood and street type. The vector $\boldsymbol{\beta}$ contains all the parameters relating to the columns in the model matrix $\mathbf{X}$.

Before analysing the results, we used Figure 1 to check the linearity assumption between features and the sale price. The first plot displays an almost horizontal line without any distinct patterns, indicating the required linear relationship. The Normal Q-Q plot suggests the residuals are normally distributed to some extent, although there is variation within the tails. We noted that this might be improved by model refinement in the future. Furthermore, while the scale-location plot suggests that the variance of residual points is not perfectly constant, the average magnitude of standardized residuals isn't changing significantly as a function of fitted values. To improve this, we would suggest using a log or square root transformation of sale price in future analysis [1]. Finally, by examining the leverage plot, we noted that within the data there are some extreme outliers, that may affect the predictive performance of the linear model.

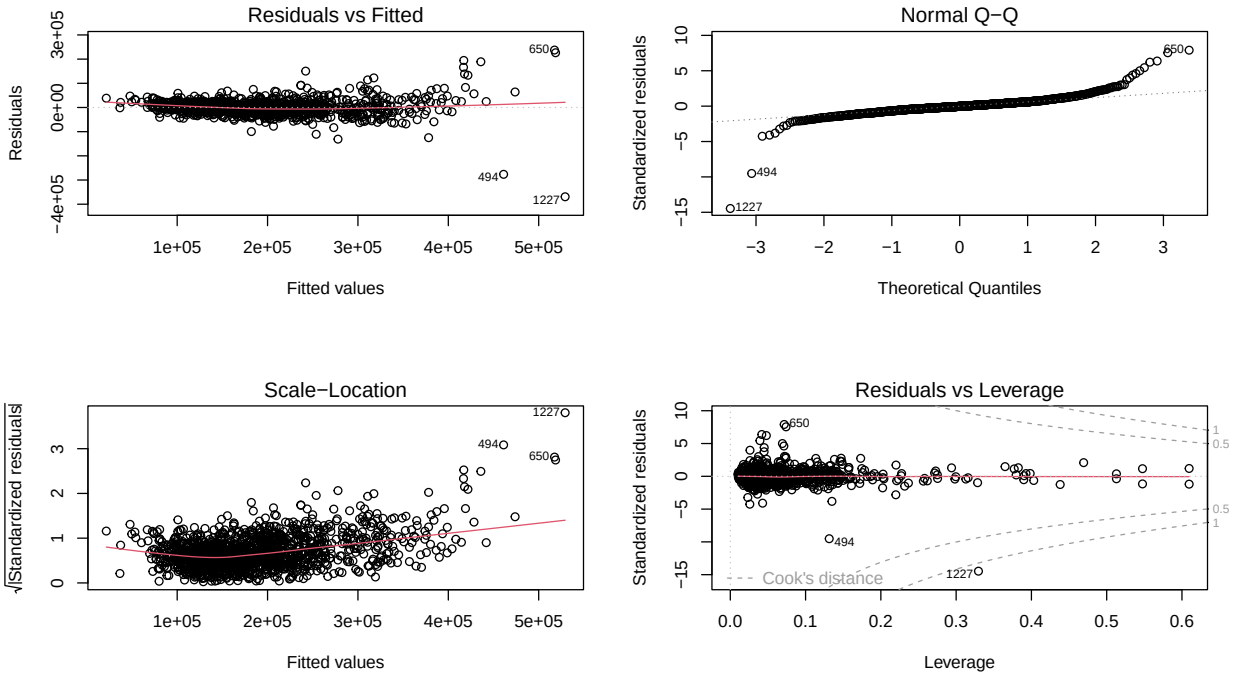

Figure 1: Assessing the linearity assumptions in the initial model.

To highlight a few key findings, an increase in lot area by one square foot is associated with a sale price increase of $0.49 when all other features are held constant. Similarly, an increase in total basement area by one square foot is associated with an increase of $3.95.

Linear Model Assessment

We assessed this model using leave-one-out cross validation (LOOCV), evaluating the mean absolute error as a performance metric. This validation approach belongs to the family of Monte Carlo methods, and is credited with producing an estimate of model performance with low bias across a multitude of statistical research [2]. The method works by creating a model using all the observations in the training dataset except one. From this model, it then predicts the sale price of the observation in the test set and calculates the error of this prediction using the actual price within the dataset. This process is repeated for every observation in the dataset and the overall error for each model is calculated by finding the mean of all the errors. Following this approach, we expect that each time the model predicts a house sale price, the average error is $20024.

One final point to highlight is that features such as the year built and ground living area have high variance inflation factor scores, indicating a degree of collinearity. However, the focus of this analysis is on prediction, and so we will not remove these features at this stage [3].

Improving the Model

Feature Transformation and Interactions

The linear model can be improved by incorporating interactions between features. Figure 2 describes how recent years are associated with higher sales prices, when all other features are controlled. However, the figure also suggests that the majority of recent houses (from 1990 onwards) have overall condition ratings of 4 or 5. Considering this, we included an interaction term to assess whether year built has a different effect on sale price depending on the overall condition rating of the house. In a similar manner, we also included an interaction between the number of kitchens and the ground living area. Intuitively, it is important to investigate the dependence between these two features, since kitchens are most commonly on the ground floor. We also assessed interactions between features such as lot area and the number of bedrooms, however in this particular model, these effects only decreased the model performance.

Implementing both of these changes together decreased the model's performance error by $146. To highlight a key finding, we see that when the number of kitchens increases, the positive effect of increasing ground living area is reduced. Alternatively, as the ground living area increases, there is a negative adjustment to the effect of increasing the number of kitchens. Consequently, the effect of increasing the number of kitchens is reduced since it has a positive coefficient.

Figure 2: Modelling sale price by year built, controlling for all other features, over-layed with a bar chart of counts in each year of overall house condition ratings.

In addition to this, Figure 3 suggests that lot area is not linearly associated with sale price, when controlling for other features. Modelling lot area under a log transformation alone did not yield any substantial improvements in the graph, yet implementing both the interactions and this transformation led to an overall decrease in performance error of $215, compared to the initial model.

Applying the log transformation reduced the skewness of the data and led to a slightly improved model fit. However when considering lot area, the most significant benefit to model performance can instead be gained by removing the aforementioned extreme outliers. Incorporating both the interaction terms and the removal of 79 extreme values of lot area contributes to a $2678 decrease in model performance error, compared to the initial model. Despite this, simply removing these values is not feasible because based on the information received regarding the dataset, these data points appear to represent houses within the population. Should further information suggest that these outliers are either data errors or cases with unusual conditions, this analysis indicates that removing them will significantly improve model performance. Overall, we conclude that both log transforming the lot area and incorporating interaction terms for overall condition and year built, and the number of kitchens and the ground living area, leads to a model performance of $215.

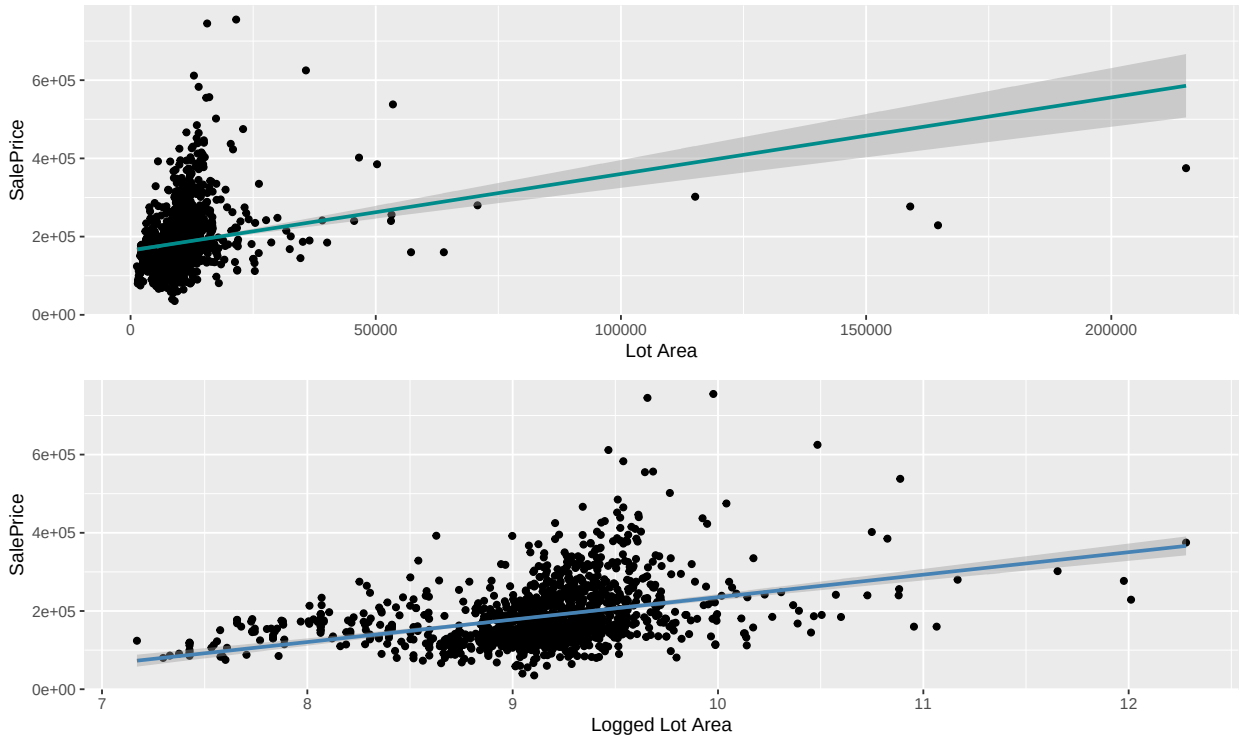

Figure 3: Modelling sale price by lot area, and by lot area with a log transform.

Variable Selection

In many contexts, it will be important to understand which features have the biggest impact on the sale price. For example, if a homeowner is looking to make renovations before selling their house, they might want to consider whether increasing the number of bedrooms or improving the overall quality of the house might have more of an impact on the sale price. We used the full linear model created above which includes all features within the data set, and then used subset selection in order to limit the model to a certain number of predictor features.

Initially, we investigated a model which used only the 3 features with the biggest impact on sale price. To obtain this model, we implemented forwards stepwise selection, which works by gradually increasing the complexity of the model. This approach evaluates separate best models of all sizes for $k = 1 \ldots, K$ features, where in this analysis, K was set to 3. For each value of k, the best model is chosen to be the one with the largest R^2 value. Finally, this approach was assessed by calculating the mean absolute error under leave-one-out cross validation. Across all but one set of the training data, the three variables which had the largest effect on performance were the overall quality, total basement square footage and ground floor living area. However, using forward stepwise selection on one set of the training data determined that overall quality, garage area and ground floor living area had the largest effect on performance. Assessing this approach overall led to a performance error of \$26099. This error is 30.3% higher than when using all of the variables.

Figure 4 shows how the mean absolute error changes as more variables are added to the model. It is important to note here that this approach considers each level within a factored feature as a feature on its own, hence we have described the effect on "variables" instead of features. It is evident that when starting with the most basic model, simply adding one variable had a large impact on performance. This rate of improved performance decreased as more variables were added to the model, with visibly lower rates of improved performance after the 3 variable threshold. Therefore, if aiming to predict house prices using a simple model with only a few variables, we suggest that using forward stepwise selection for 3 variables is an effective method to do so.

As the model became more and more complex, the error typically decreased, except for some slight deviations.

The model which used 29 variables appears to have the optimal performance, with a mean absolute error of \$19543. This suggests that implementing forward stepwise selection with $K > 29$ is not necessary as it simply increases complexity of the model without improving performance, most likely due to overfitting.

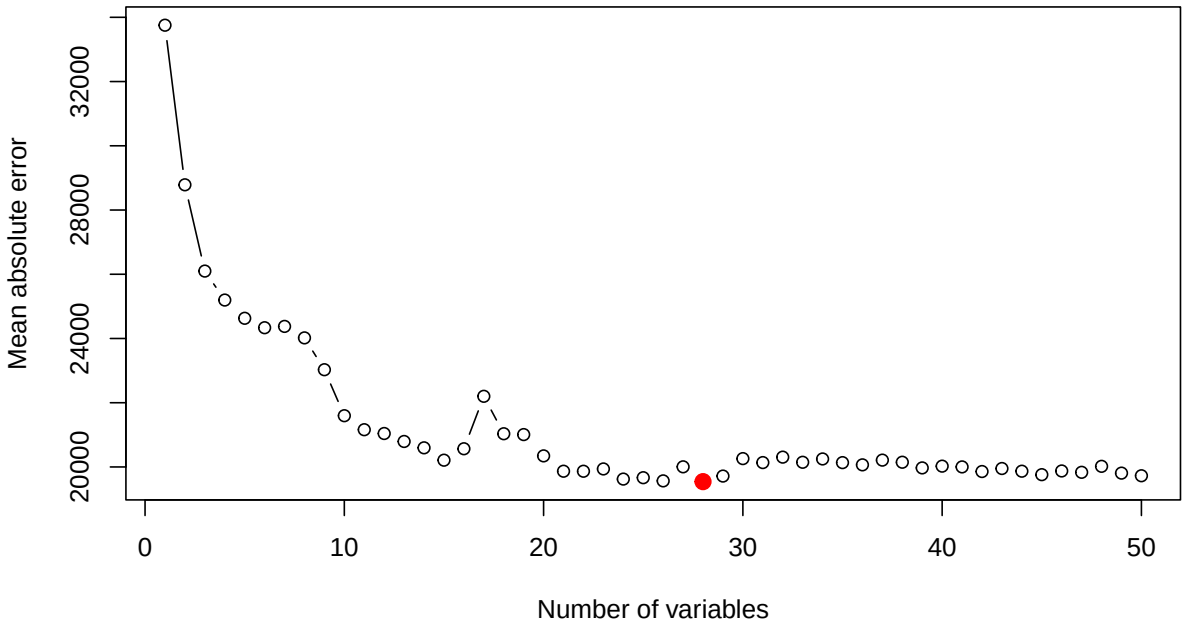

Figure 4: Mean absolute error for best forward stepwise selection models with the minimum error highlighted in red.

Investigation of Extra Features

The next step in this analysis was evaluating a new dataset containing information on 81 features of 1460 observations. This dataset has several extra property features compared to the one previously used, including garage finish, quality, and condition; exterior quality and condition; and roof material. We fit three models to this data - a simple linear regression, a lasso, and a random forest. We again assessed each model using leave-one-out cross validation, evaluating the mean absolute error as a performance metric. We also investigated whether the extra features in this dataset have much benefit in predicting sale price compared to using only those in the original dataset.

Preprocessing

When carrying out initial exploration of the data, we looked for any features which had a significant level of missing values. We found that five features - alley type, fireplace quality, pool condition, fence type, and miscellaneous features - had over 600 missing values, as seen in Figure 5, and so we removed these features from the dataset. We then removed any observations with missing values from the data, leaving 1094 remaining.

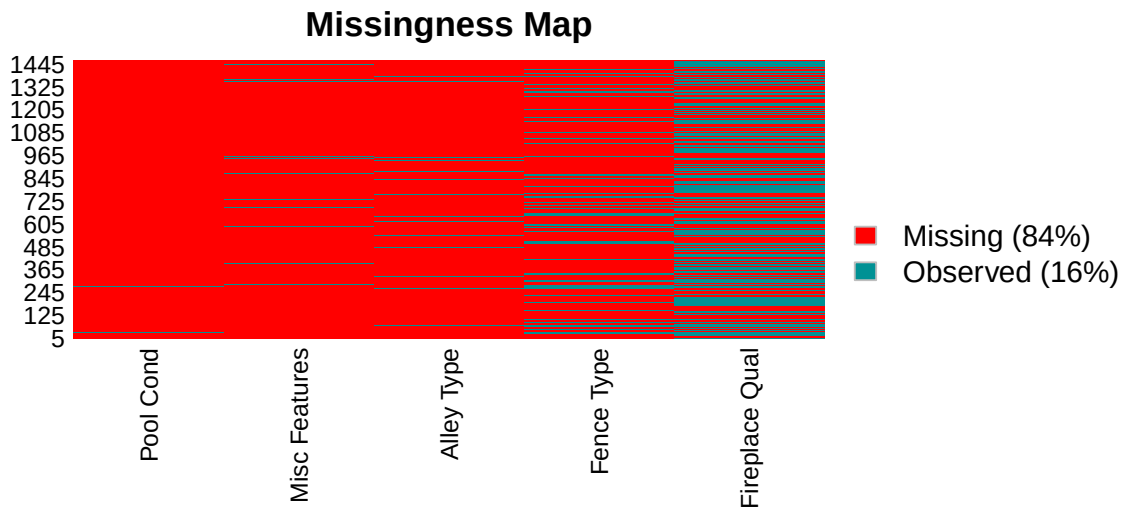

Figure 5: Missingness map showing missing values for features alley type, fireplace quality, pool condition, fence type, and miscellaneous feature.

We then noted three possible issues that may cause problems when fitting our models. Firstly, the features 1st floor square footage and total basement square footage had a correlation of 0.91. Secondly, the values for total basement square footage were a linear combination of those for basement finished and unfinished square footage. Finally, ground living area was a linear combination of the values for total basement square footage and second floor square footage. As total basement square footage is a common factor in all three of these issues, we excluded this feature.

Simple Linear Regression Model

We created a linear regression model using all 72 predictor features from the dataset of 1094 observations. We defined y_i as the sale price for each house $i = 1, ..., n$, and defined the following model:

$$\mathrm{y}_i = \omega + \boldsymbol{x}_i^T \boldsymbol{\gamma} + \delta_i$$

where δ_i follows a normal distribution with mean zero and variance σ^2. The vector $\boldsymbol{x}_i$ is a row of the model matrix for observation i containing containing all the features, with dummy variables for factor features such as neighbourhood and street type. The vector $\boldsymbol{\gamma}$ contains all the parameters relating to the columns in the model matrix $\mathbf{X}$.

Before analysing the results, we used Figure 6 to assess the linear regression assumptions for the data. The first plot displays an almost horizontal line without any distinct patterns, but with some significant outliers. The Normal Q-Q plot suggests the residuals are normally distributed to an extent. However, there is deviation from the fitted line in the tails. The upwards trend in the scale-location plot suggests that the assumption of constant variance may not be valid. Finally, by examining the leverage plot, we noted that there are some extreme outliers within the data that may affect the predictive performance of the linear model. As a result, we would expect that a linear regression model may not perform particularly well for this data.

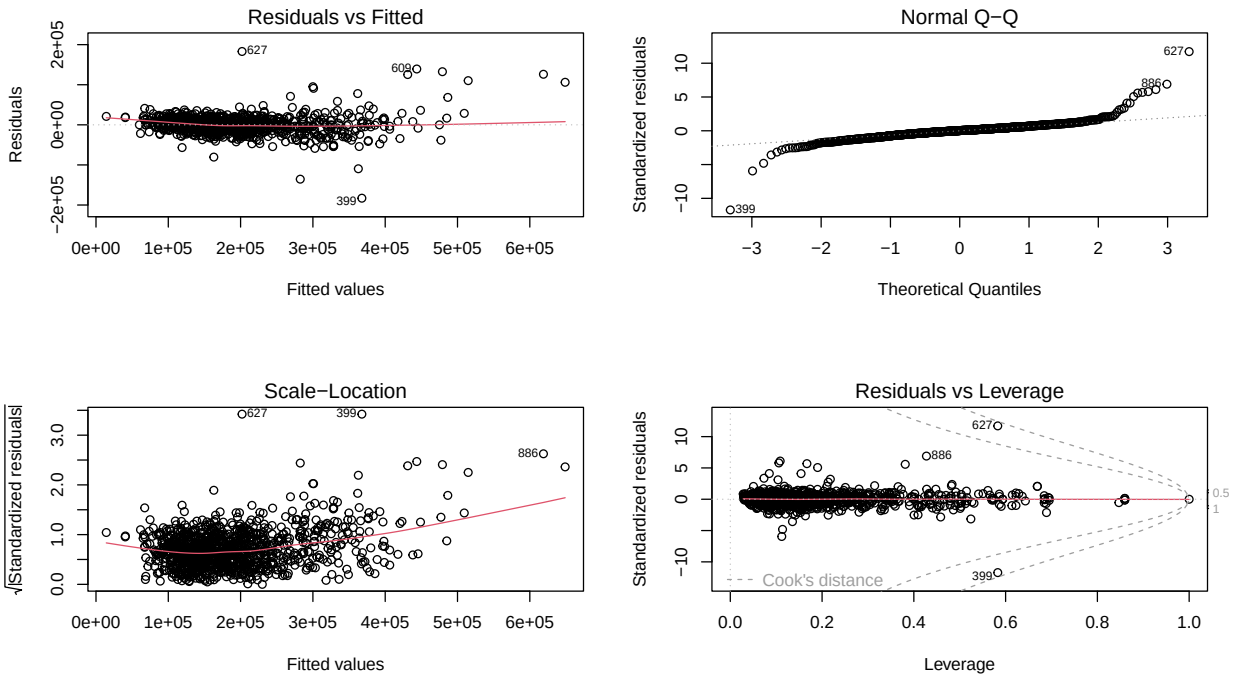

Figure 6: Assessing the linear regression assumptions for the data.

Upon fitting the model using all 72 available features, we found that the coefficient values for ground living area, the mix level of the electrical feature, and the cinderblock level of the 2nd exterior feature were not defined. This indicates that the remaining variables explain all of the variation in the data and these features do not add any information to the model, so we dropped ground living area. For the latter two factor level features, as each of these levels appear only once in the data, we removed these two observations. We fit the model again, using the remaining 71 predictor variables for 1092 observations.

We assessed this simple linear regression model using leave-one-out cross validation and found it to have a mean absolute error of \$18850. This is comparatively better performance than that of the simple linear model for the original dataset, which had a mean absolute error of \$20024.

Lasso Model

Next, we applied the lasso regularisation technique to attempt to improve upon the performance of the simple linear model. In order to fit a lasso model, we first encoded all of the factor features in the data using dummy variables. With each individual factor level now being considered as its own variable, this encoded dataset has 252 predictor variables.

In a lasso regression model with n observations and K variables, we choose θ to minimise the loss function:

$$C(\theta) = \sum_{i=1}^{n}(y_i - \theta^T x_i)^2 + \lambda \sum_{k=1}^{K} |\theta_k|.$$

The lasso technique performs variable selection, with the absolute value penalty forcing some coefficients to zero. Using trial and error, we found the optimal value of λ to be 1596, and used this to fit a lasso model.

The lasso model uses 60 variables as predictors, forcing the coefficients of the other 192 variables to zero. These 60 variables correspond to 45 of the 71 features used - 19 continuous features and particular levels of 26 factor

features. Notably, of these 45 selected features, only 16 were present in the original, smaller dataset, such as lot area, year built and overall quality. The remaining 29 consist of the extra features available in the new dataset, including roof material, sale type, and garage finish and quality. This suggests that the extra variables have some benefit when modelling the data, compared to using only those from the original dataset.

We assessed the lasso model using leave-one-out cross-validation and find it to have a mean absolute error of \$17909. The lasso model performs better than the simple linear model, which had a mean absolute error of \$18850, which is equivalent to a 5% decrease. This suggests that the simple linear model may have been overfitting the data, and that some of the features were irrelevant.

Random forest

Finally, we fit a random forest model to the data by training 40 decision trees under a bootstrapping approach. As the dataset contained $K = 252$ variables, we chose $P = \frac{K}{3} = 84$ variables at random. We then sampled $M = 1092$ observations with replacement from the total 1092 observations. Each decision tree was grown on a bootstrap sample, using only these 1092 sampled observations for the randomly chosen 84 variables. This process was repeated until 40 decision trees were formed. The final prediction for a new observation is the average prediction of all trees.

Figure 7 illustrates the importance of the 15 most important variables as measured by the random forest model based on mean absolute error. Of these 15 variables, only 5 were present in the original dataset, indicating that the extra variables improve model performance, compared to using only those from the original dataset. In addition, 2 of the top 3 most important variables were not present in the original, smaller dataset, providing further evidence for this claim. We note that the 12 most important variables in the random forest model were also used in the lasso model, also suggesting that they explain variation within the data.

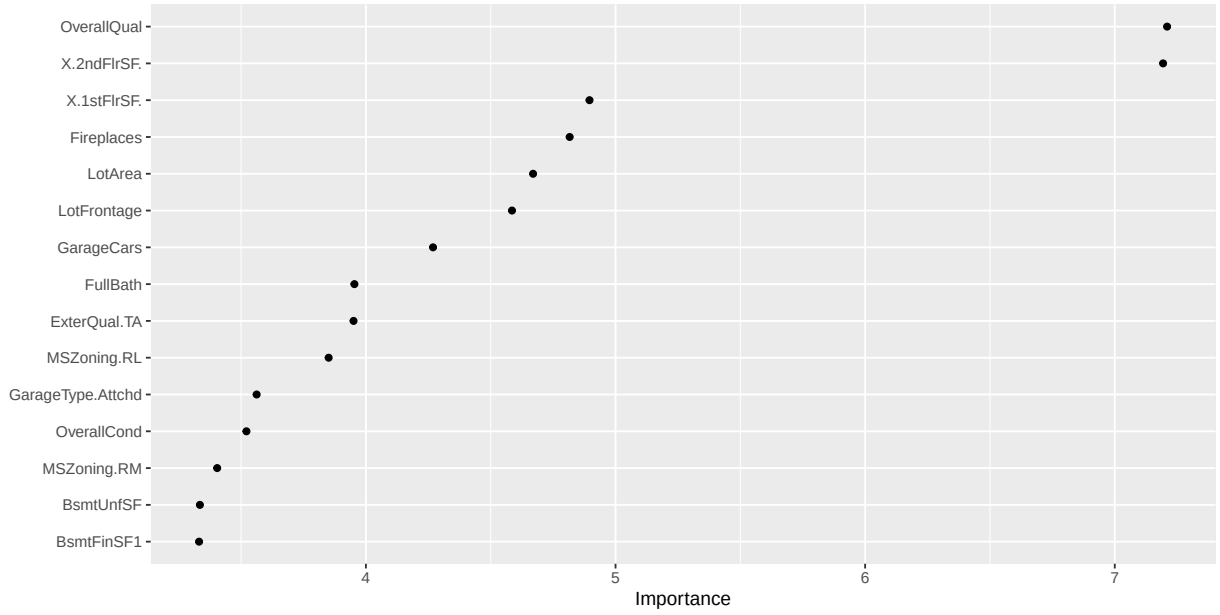

Figure 7: Variable importance of 15 most important variables based on mean absolute error as measured by the random forest model.

We assessed the random forest model using 50-fold cross-validation and found it to have a mean absolute error of \$18309, which is equivalent to a 3% decrease. It performs better than the simple linear model, but not as well as the lasso model.

Conclusion

In this report, we investigated the performance of various models in predicting sale price of a house, given particular features. Throughout the process, we used leave-one-out cross validation in order to estimate the mean absolute error of each model, as summarised in Table 1.

Table 1: Summary of mean absolute prediction errors for each model.

Model	Mean Absolute Error under LOOCV ($)
Simple	20024
Interactions and Transformation	19808
3 Features	26099
29 Variables	19543
Simple with Extra Features	18850
Lasso	17909
Random Forest	18309

Based on these results, when there is limited number of house features available, such as those provided in the original, smaller dataset, the implementation of a three variable model under forward stepwise selection is not recommended. However, this selection method has the potential to perform well when the number of variables selected is increased. In fact, this approach produced the best model for the original, smaller dataset, when choosing 29 variables. Moreover, if a model is required with a specific complexity, we would advise using this approach. The results also show that implementing a linear regression model with interactions and transformations performs very similarly, with only a $266 increase in error compared to the former approach. Considering that stepwise selection can be time-consuming and computationally expensive, we propose this method as an alternative if a simpler method is desired.

When extra house features are available, such as those included in the second dataset, we have created 3 quantitative models which more accurately predict the sale price of a house, compared to any of the models which use only the features in the smaller dataset. This indicates that there is benefit in including these extra features when formulating predictive models. Considering the results in Table 1, we conclude that, out of all the models developed, the lasso regression model most accurately predicts the sale price of a house. As a note of comparison, the mean absolute error is $1634 less than that of the 29 variable stepwise selection model.

Overall, the evaluation of the mean absolute error under leave-one-out cross validation has reduced bias in our results, and we are confident in our recommendation that the lasso model provides the best prediction of house sale prices.

References

[1] STHDA, *Linear regression assumptions and diagnostics in R: Essentials* http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/year=2018, month=Mar

[2] Hawkins, Douglas M. (2004) *The Problem of Overfitting*. In Journal of Chemical Information and Computer Sciences, 44,1, pp1-12 https://doi.org/10.1021/ci0342472

[3] Dormann, C. F. and Elith, J. and Bacher, S. and Buchmann, C. and Carl, G. and Carré, G. and Marquéz, J. R. and Gruber, B. and Lafourcade, B. and Leitão, P. J. and Münkemüller, T. and McClean, C. and Osborne, P. E. and Reineking, B. and Schröder, B. and Skidmore, A. K. and Zurell, D. and Lautenbach, S. (2012) *Collinearity: A review of methods to deal with it and a simulation study evaluating their performance*. Ecography, 36(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x