MATH4432: Final Project

House Prices: Advanced Regression Techniques

Sarah Catherine James
20501098

Tuesday 22nd May, 2018

1 Introduction

In this project, released by Kaggle and viewable at House Prices: Advanced Regression Techniques, the aim is to identify which variables influence house price negotiations, and in turn predict the final price of homes based on their values for such variables. The dataset contains 79 explanatory variables of residential homes in Ames, Iowa and from these, a model will be trained to produces accurate predictions for sale prices.

This project was implemented in R, and involves feature engineering and advanced regression techniques such as random forest.

2 Feature engineering

2.1 Imputation

Before models can be trained to the data, the data must be cleaned. From the raw data, two initial issues are observable. Firstly, there are a high number of NA values. These must be filled or removed in in order to use the data. But, the second problem is that NA in this data set can mean one of two things; either the data is missing, as R would normally interpret an 'NA' value, or the house simply does not have that variable. For example, if the dataset contains the value 'NA' for the variable `GarageType`, then the data is not missing, the house just doesn't have a garage.

For categorical variables, we deal with 'NA' values that imply the absence of the variable altogether, such as with `GarageType` mentioned above, by replacing the 'NA' values with meaningful values such as 'No garage', making our dataset more usable. For the 'NA' values that imply missing data, we used the modal value for the variable. For example, for `MSZoning`, the general zoning classification, the modal value is 'RL', so this is what we replace all 'NA' values with.

For numerical variables, we deal with 'NA' values that imply the absence of the variable altogether, such as with `BsmtUnfS`, unfinished square feet of basement area, by replacing the 'NA' values with '0'. 'NA' values with meaningful values such as 'No garage', making our data set more usable. This was the case for most of the 'NA' values for numerical variables, and the sensible cases for this method were determined after reading the Kaggle-provided descriptions for each variable. We are then left with 3 numerical values that imply missing data- linear feet of street connected to property, year garage was built, and masonry veneer area in square feet. Of these, only the former can be estimated reasonably. This is done by imputing the median. For the remaining two, it is not possible to estimate these values, and so the 'NA' values are instead replaced with illogical values that make it possible for the computer to process the values in an algorithm, while still being able to capture that the data was unavailable.

Now all missing values have been imputed, the training of different models can begin.

3 Model Training Selection

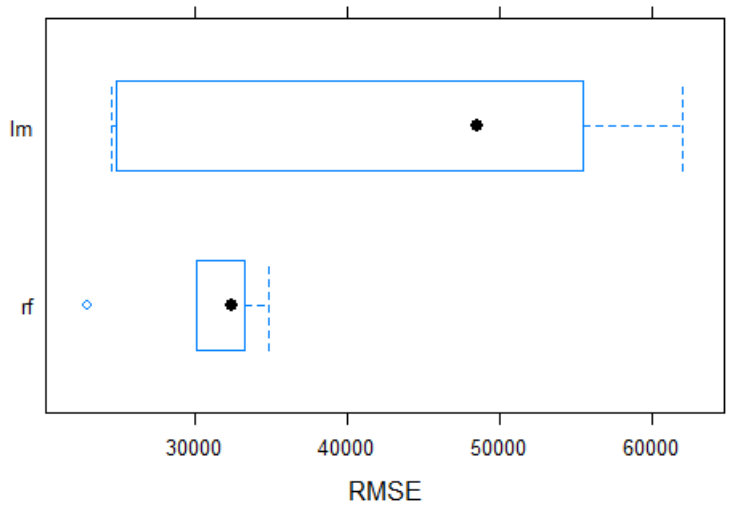
Note that since a continuous variable is being predicted, regression-based algorithms are used, and that the seed is set, arbitrarily to 54 for the sake of reproducibility in cross-validation. The models trained are as follows:

1. Random forest
2. Linear regression
3. Random forest with two mtry values
4. Random forest with 20 most important explanatory variables
5. Random forest with 20 most important explanatory variables and 2 mtry values
6. Linear model with a trimmed training set
7. Regularised linear regression

The first model to be tested is the most basic- a random forest algorithm is used, using all 79 explanatory variables, and ignoring `SalePrice`, as this is being predicted. Note that one mtry value is used. The mtry value is the number of variables available for splitting at each tree node.

To be able to interpret how effect this model is for predicting `SalePrice` accurately, a linear regression model is made and the root mean square errors (RMSE) are compared. For this model, the variables used and the mtry value are the same as for the simple random forest. For ease of understanding, a box-plot is made to compare the RMSE values. From fig. 1, we can see that the random forest has a much smaller RMSE value and hence performs better than the linear regression model. The linear regression model is therefore deemed unuseful and discarded.

In order to improve the random forest model, it is possible to increase the number of mtry values to two, so now there are two variables available for splitting at each tree node. Again, this model is compared with the current best model, the simple random forest model. From fig. 2, we can see that having two mtry values does reduce the RMSE of the random forest model, and so this becomes the best model for predicting `SalePrice`.



1

Figure 1: random forest vs. linear regression

Another adaptation to the model is to not use all explanatory variables, and instead extract the most important explanatory variables from the data and only use them in the model. One way of doing this it through using the VarImp function from the caret package, that extracts the 20 most important variables from the data i.e. the variables that hold the highest predictive power for `SalePrice`. The 20 most important explanatory variables are found to be:

1. OverallQual	2. GrLivArea	3. TotalBsmtSF	4. GarageArea
5. GarageCars	6. X1stFlrSF	7. YearBuilt	8. ExterQual
9. BsmtFinSF1	10. FullBath	11. KitchenQual	12. LotArea
13. Fireplaces	14. FireplaceQu	15. YearRemodAdd	16. GarageYrBlt
17. X2ndFlrSF	18. TotRmsAbvGrd	19. MasVnrArea	20. LotFrontage

The random forest model using the 20 most important explanatory variables is then compared with the random forest model with 2 mtry values via boxplot. From fig. 3, we see that the model with the 20 most important variables performs better than the model with 2 mtry values. So, the next logical model to try would be the model combining these two features.

Then, after comparing the model with the top 20 variables, the model with 2 mtry values, and the model combining these two features, it is observable that the model with the top 20 variables still has the lowest RMSE, and combining the two features did not improve the model. This was observed from fig. 4.

Now, another option for improving the model is explored. This option is removing outliers from the training set, which obscure the prediction of `SalePrice`. Outliers in the training are found from plotting scatter plots of all 20 most important variables against `SalePrice`. From these variables, outliers are only found for the variable `GrLivArea`- above grade (ground) living area square feet, as shown in fig. 5. Outliers are removed accordingly, i.e. the training set is trimmed. Then, a linear regression model is built using the new, trimmed training set. From fig. 6, this model has a smaller RMSE than the previous best model using the 20 most important variables.

Finally, the last model explored in this project is a regularised linear regression model. From fig. 7, it is observable that the regularised linear regression model performs best, and hence the best model in this project has been reached.

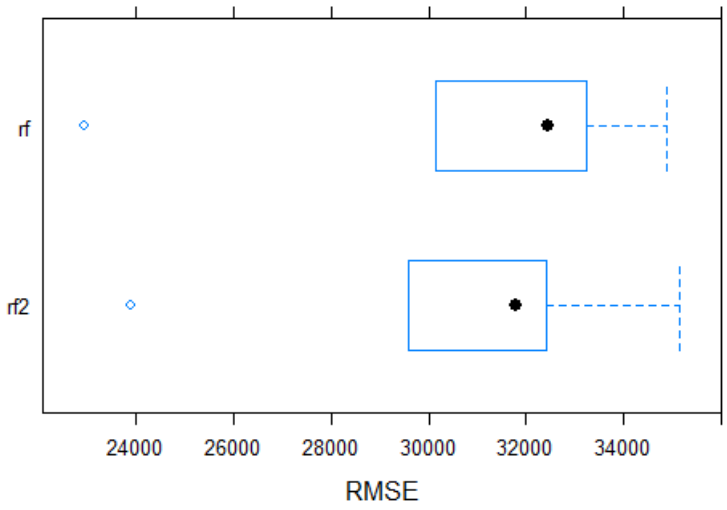


Figure 2: random forest with 1 mtry vs. 2 mtry

4 Conclusion

The best model for predicting `SalePrice` was then found to be regularised linear regression model. However, this model still had a very high RMSE and does not score very well on the Kaggle competition. One further adaption to the model could have been to log-transform the trimmed data set. With a lot more studying of different approaches to model training, for example, gradient boosting, neural networks, and support vector machines, a much better model could be obtained. Given more time, `SalePrice` could be predicted effectively using the 79 explanatory variables given in the data set.

A Individual contributions

All done by Sarah James.

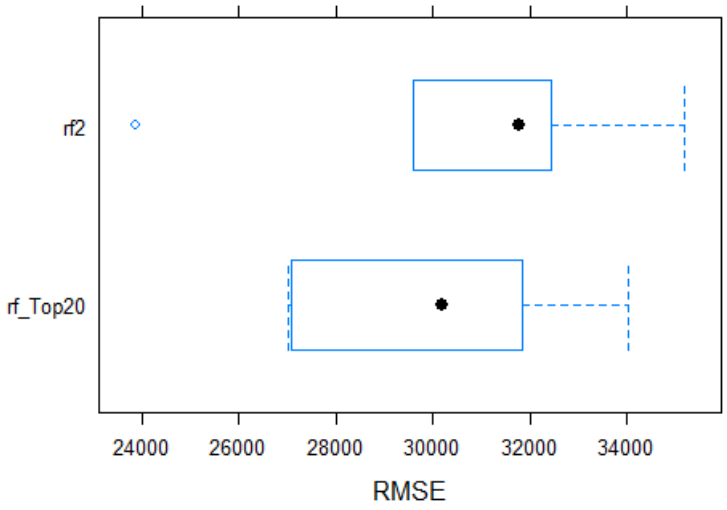


Figure 3: RF with 20 top variables vs. with 2 mtry

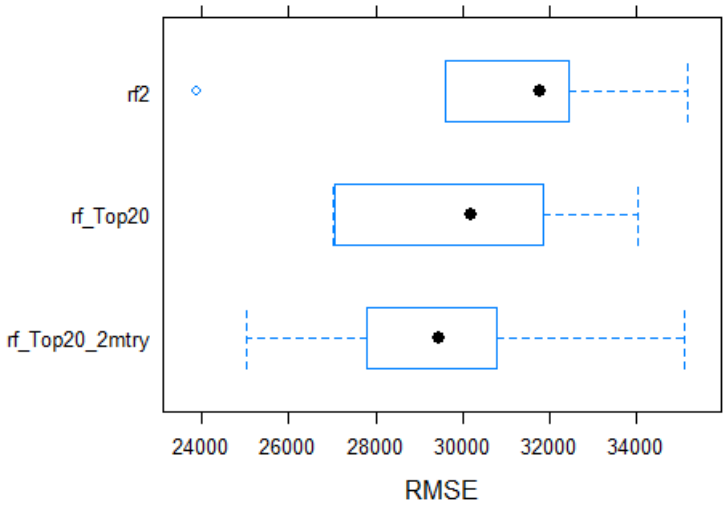


Figure 4: RF with 20 top variables vs. with 2 mtry vs. with both features

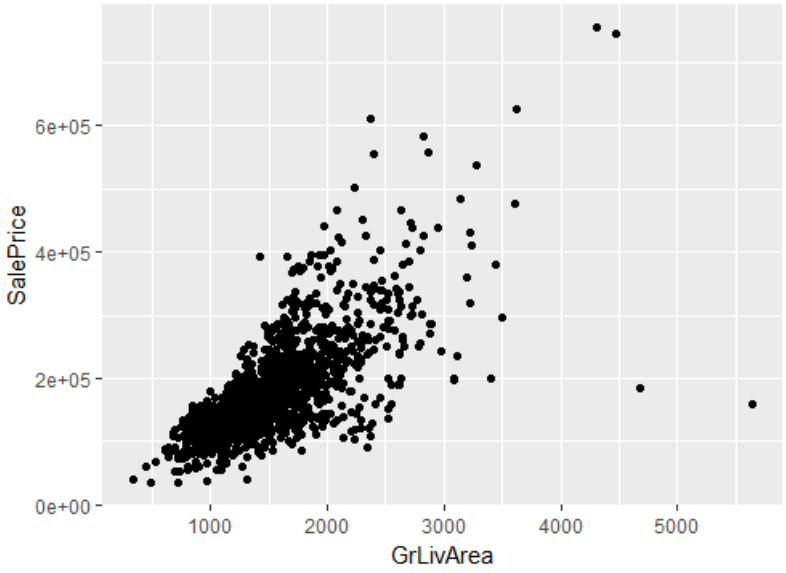


Figure 5: GrLivArea vs. SalePrice

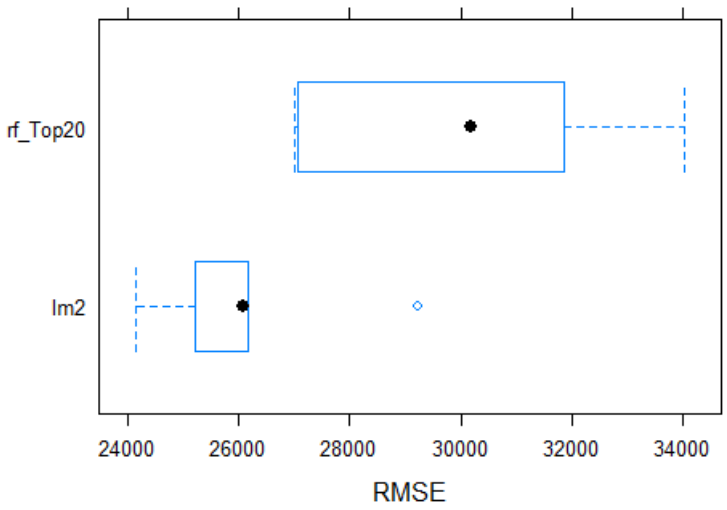


Figure 6: Trimmed linear regression model vs. random forest with top 20 variables

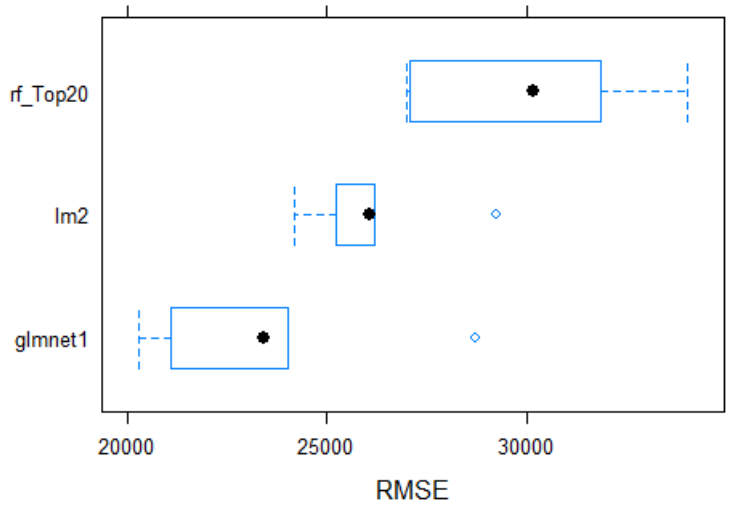


Figure 7: Trimmed linear regression model vs. random forest with top 20 variables vs. regularised linear regression