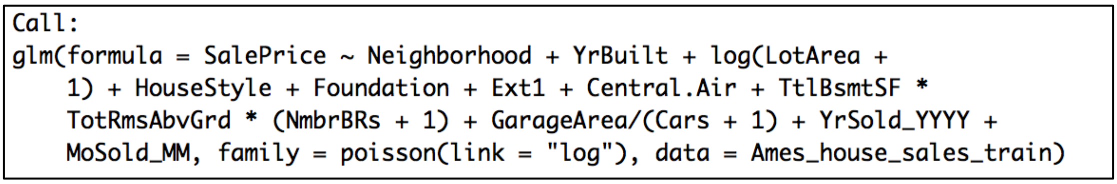
ETC5242 - Project

13 October 2016

‘Australian Actuaries’ have spent the last few months attempting to create a model to predict house sale prices for houses in Ames, Iowa, USA.

The team was presented with data on over 1700 houses, with information including house sale prices, the number of bedrooms, size of the house in terms of land size and garage area, and the location of the house, among other factors. Through analysis of this data, we created the most accurate model possible, to be used to predict house prices.

This aforementioned model was created using the code shown below:

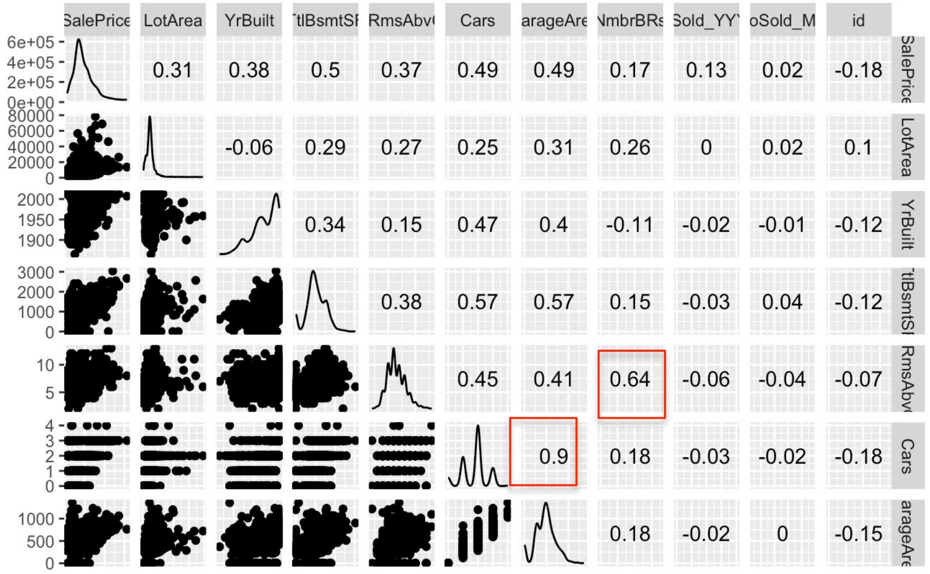


```
Call:
glm(formula = SalePrice ~ Neighborhood + YrBuilt + log(LotArea +
    1) + HouseStyle + Foundation + Ext1 + Central.Air + TtlBsmtSF *
    TotRmsAbvGrd * (NmbrBRs + 1) + GarageArea/(Cars + 1) + YrSold_YYYY +
    MoSold_MM, family = poisson(link = "log"), data = Ames_house_sales_train)
```

Ideas Behind the Model

When creating the model, data analytics was performed in order to improve the model. Several approaches that were taken to improve the model were as follows:

- A generalized linear model was used, with a link function of ln(μ);
- By analyzing the correlation between the explanatory variables, we could determine which variables to add interaction terms to. As shown below, Number of Cars and Garage Area are highly correlated, and so too are the Total Rooms Above Ground and the Number of Bedrooms;



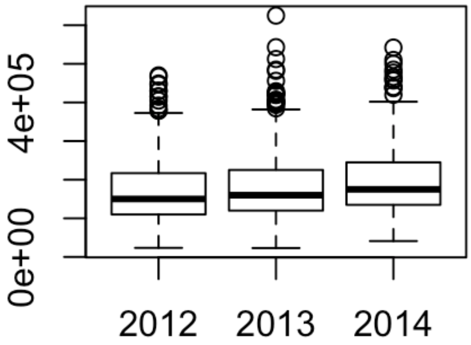
- In order to create these interaction terms, several variables had to be altered. If the value of one variable, for example Number of Bedrooms, were zero, the effect of the term being interacted with would be removed, creating a large bias.

Therefore, all values for the Number of Cars and Number of Bedrooms has been increased by one;

- Furthermore, the values of Lot Area have been increased by one to remove and null figures. By doing so, this allows us to compute the logarithm of every term;
- Finally, several extreme outliers were removed from the data. By doing so, the large effects that these outliers have on the data are removed, and the accuracy of our model increased significantly.

Point of Interest

By analyzing the data, it was found that the data was filled with several outliers, as shown below. This fact was alarming to us, as the presence of outliers would make it very difficult to fit an accurate model. By removing several outliers we improved our model's accuracy, but removing all of the outliers would prove difficult, and would likely reduce the accuracy of our model.



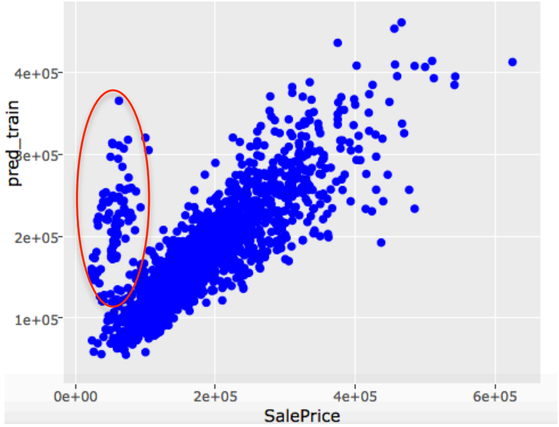
Accuracy of Predictions

In order to assess the accuracy of our model, our group created an error test, according to the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i\,|$$

This equation was used, and the predicted values from our model were assessed against the actual values for the training data. Throughout the project, we managed to improve the accuracy of our model from an initial mean absolute error of just under $40,000, to our final absolute error of just over $36,000.

A plot of the actual sale prices and our predicted sale prices for the training data is shown below. Emphasis was placed on reducing the amount of values as shown in the red circle below. However, this proved difficult, and the reduction in the number of values was unfortunately minimal.


4e+05
3e+05
2e+05
1e+05
pred_train
0e+00
2e+05
4e+05
6e+05
SalePrice