**PROPERTY INSPECTION PREDICTION OF LIBERTY MUTUAL GROUP**

**BY PRIYANKA BALINENI**

# **INTRODUCTION**

This report documents the prediction of property hazards before the time of inspection by Liberty mutual group. Liberty mutual insurance provides a wide range of insurance products and services to meet their customer needs. Many newly properties receive a home inspection to ensure that Liberty mutual portfolio aligns their business objectives. Investigators audit the main key attributes of the property along with foundation, roof, windows and siding. These results help Liberty mutual group to decide whether property is one they need to insure.

I collected the data from kaggle website and began the analysis by cleaning the data and splitting into train and test data. Afterwards, I implemented two parametric techniques: Lasso and Ridge regression, and a non-parametric technique: Random forest to predict the hazards. Then, we accessed the models based on the best results given by the above-mentioned metrics.

# **DATA EXPLORATION**

The dataset contains 34 rows which has 32 anonymized variables and a hazard score and ID number. Each row in the dataset corresponds to each property. There are no null values in the dataset.

There are few columns with categorical variables. So, to do analysis I had converted them in to numbers. So, I grouped means of hazard values to convert categorical variables into numbers. After converting all the categorical variables into numbers. I did xgboost feature importance to find out how useful or valuable each feature was in the construction of the model. T1\_V1 has most relative importance followed by T2\_V1 and T1\_V2.

To evaluate our model, we divided the current dataset into two segments: one is used to train model (70%) and the other one is used to test the final model (30%).

# **MODELS**

1. **Parametric Models:**
2. **Multivariate Linear Regression**

Multivariate Linear regression is a basic approach for the prediction problem for modelling between a target variable and explanatory variables. The multiple linear regression is given by:

1. **Ridge Regression**

Ridge regression uses an l2-norm penalty to improve OLS when the covariates are correlated. Like OLS, the ridge estimator has an explicit form

Where λ>0 is the tuning parameter and is chosen to minimize the penalized sum of squares.

1. **Lasso Regression**

Laaso regression uses an l1-norm penalty to improve OLS when covariates are correlated. Lasso estimator explicit form:

Where λ>0 is the tuning parameter and we select λ by minimizing the generalized cross-validation criterion.

1. **Non-Parametric Model:**
2. **Random forest**

A random forest fits the number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The prediction of set of m trees with individual weight functions , it’s predictions are:

# **RESULTS & ANALYSIS**

To predict the hazard score, I first used the multivariate linear regression model with all the variables after data preprocessing. A common way to summarize how well a linear regression model performed is via the coefficient of determination. This can be calculated as the square of the correlation between the observed values and predicted values.

If the predictions are close to the actual values, we would expect to be close to 1. On the other hand, if the predictions are unrelated to the actual values, then =0. In all cases, R2 lies between 0 and 1. The model has a of 8.85% which indicates that 9% of the variance in data is captured.

To improve the model, I used lasso and Ridge models. These two models are closely related and used to prevent overfitting and regularize the coefficients. Like OLS, Lasso and Ridge models try to minimize the residual square errors. The best alpha value for ridge and lasso are determined by choosing the value that minimizes k-fold cross validation error and plots of errors for both models. Lasso has an of 8.87% and ridge has 8.86% of variances. There is very less improvement in the model.

So, further to get a better model I used random forest with 500 number of trees and number of features to split as 11 which are selected randomly. This model results in a better model with an of 10.1% as shown below in figure1.

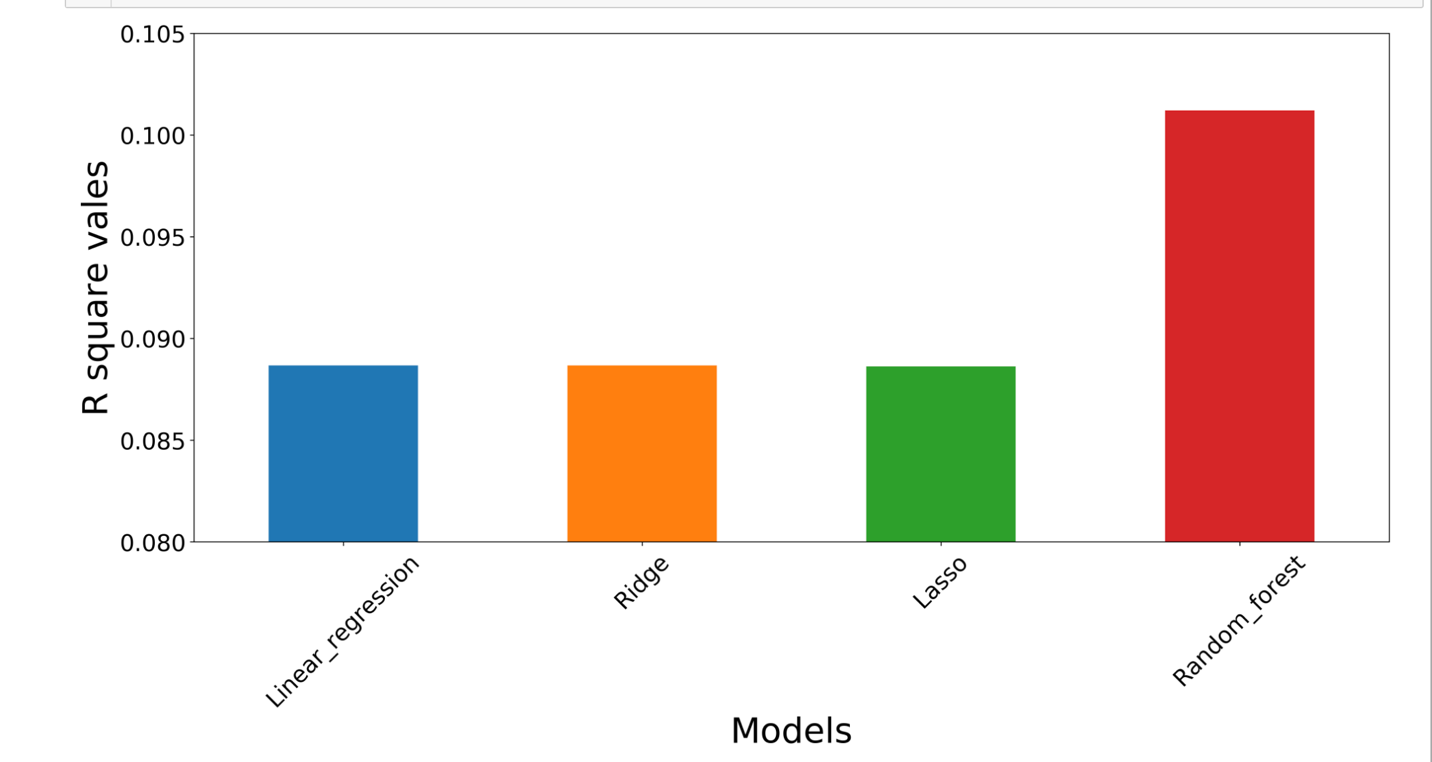


Figure 1

# **CONCLUSION**

I have predicted models using parametric and non-parametric methods and the best result is obtained using Random forest, a non-parametric method with an of 10.1%. We could choose parametric models if the interpretability is important or else non-parametric model for better prediction accuracy.

# **REFERENCES**

1. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.
2. Jain, S., Srivastava, T., Dar, P., & Shaikh, F. (2017, June 22). A comprehensive beginners guide for Linear, Ridge and Lasso Regression. Retrieved April 16, 2018, from <https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/>
3. Retrieved April 16, 2018, from <https://www.kaggle.com/c/liberty-mutual-group-property-inspection-prediction/data>
4. Ridge Regression and the Lasso. (2017, May 23). Retrieved April 16, 2018, from <https://www.r-bloggers.com/ridge-regression-and-the-lasso/>