ADVANCED R

PREDICTING HOUSE PRICES

Andrés Llerena Icochea

# INTRODUCTION

This report provides a summary of regression insights gleaned from observations of 17,277 houses sold in King County, Washington – USA between 2014 and 2015. A regression analysis was performed on the data provided with the **objective to predict future house prices in the area** based on the information available.

# REGRESSION RESULTS

To conduct the regression analysis, several models which include a linear regression, a random forest, and gradient tree boosting were constructed and executed on the provided dataset. The **final model produced has an R-square of 89.48% and Mean Absolute Percentage Error (MAPE) of 12.48% in predicting future house prices based on the house features that were available** in the dataset**.**

# SUMMARY OF TECHNICAL APPROACH & METHODOLOGY

The dataset *“house\_price\_train.csv”* was the foundation for the analysis and the basis for the proposed predictive model. The dataset contains information of 17,277 houses sold in Washington, USA between 2014 and 2015. Among the 21 total variables available, 20 represent variables regarding house features, location, condition, etc., and one represents a target variable, indicating the final price the house was sold at. An ID variable identifies each observation, as it is assumed that each observation represents a different employee and there are no duplicates.

To following Data Preparation, Feature Engineering and Modelling steps were undertaken using **R Studio**:

* Data preparation (dataset loading, information integrity check and initial descriptive analysis)
* Feature Engineering (variable construction and treatment, univariate analysis, multivariate analysis)
* Model building (dataset split, model training and validation)

A detailed analysis of data fields, definitions and additional content is provided in **Appendix A.**

1. Baseline

A first attempt of a logistic regression model was tried on the variables as is (without any treatment or feature engineering). The resultant baseline model had an **R-square of 61.82% and MAPE of 26.39%.**

1. Feature Engineering and Modelling

Refer to R-markdown file attached.

1. Model Selection

Different models were trained using both the target value as it is (Price) and the transformed target value (Log(Price)). This transformation was conducted to correct skewness on its original distribution.

The following table shows a summary of the metrics obtained for each of the models trained.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Price** | | **Log(Price)** | |
| **R-square** | **MAPE** | **R-square** | **MAPE** |
| Linear Regression | 61.82% | 26.39% |  |  |
| Random Forest | 85.72% | 13.12% | 88.13% | 12.59% |
| GBM | 82.90% | 15.58% | 86.51% | 14.28% |
| XGB | 86.57% | 13.40% | 89.48% | 12.48% |
| XGB + PCA | 84.70% | 13.82% | 88.33% | 13.27% |

It is observed that metrics improve when using Log(Price) as the target variable. Particularly Random Forest and XGB models perform better than the other ones. XGB model is selected to its slightly superior improvement.

# CONCLUSION & ADDITIONAL RECOMMENDATIONS

This report delineated a **logistic regression model** which goal is to predict which employees are more likely to leave the company based on past employees’ information. The model is considered **Good** in terms of accuracy rate and performance metrics such as AUC-ROC.

It is, nevertheless, recommended to monitor the model performance on a monthly or quarterly basis in order to identify any anomaly or loss of predictor power in order to take immediate actions such as adjusting the coefficients or, eventually, the construction of a new version of the model.