MATH 4432 Project 3: How Much Is Your House Worth?

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Haojie WANG

Department of Civil and Environmental Engineering, HKUST

Kaggle House Price Project 590th of 5407
@Haojie W
Rank 590st of 5407, Top 11%
https://www.kaggle.com/hjwang

1. Introduction

House price might be one of the biggest concerns for young generations all over the word. Now on Kaggle, over five thousand teams are working on predict the house price in Ames, Iowa, based on a detailed Database.

Like many of them, this project is also dedicated to find the secret of how to make an accurate prediction of the price of a certain house.

2. Data Preprocessing

Data cleaning for missing values

- By analyzing the records, there are total 34 features with missing data. To ensure the proper treatment of missing values, each feature was cleaned one by one. (e.g., Fill the NA for **LotFrontage** using random forests, Fill the NA for **MSZoning** using the mode, delete **Utilities** since almost all are "AllPub" (2916/2919), etc.)

Process (ordinal) variables

- Since categorical variables enter into statistical models differently than continuous variables, storing data as factors insures that the modeling functions will treat such data correctly. Hence, character variables are transformed into factors, also some are converted into ordinal integers if there is outstanding ordinality.

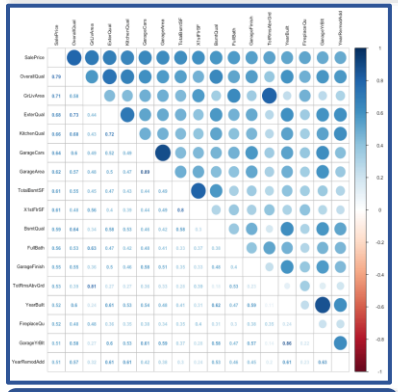

Fig.1 Correlation figure of variables highly correlated to SalePrice(>0.5)

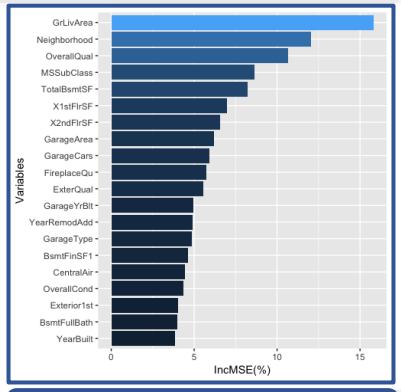

Fig.2 Variable importance figure based on Random Forest

3. Feature Engineering

Guidelines:

It is easy to find out some variables are describing the same aspect of a house(e.g., four variables concerning bathroom, three variables concerning the house age, etc.). By combining these similar variables together to build a relatively stronger predictor, we can have a better prediction performance based on this dataset.

Create variables "SumBath"(total number of bathroom), "Age" (house age), "SumSF"(total living square feet) and so on:

- For **SumBath**: SumBath is a weighting summation of four variables, namely, "FullBath", "HalfBath", "BsmtFullBath" and "BsmtHalfBath".
- For **Age**: "YearRemodAdd" and "YearSold" are used to determine the house age. By default, YearRemodAdd is set to YearBuilt if there has been no Remodeling/Addition.
- For **SumSF**: Generally, the house price is highly affected by the total living space. Hence, SumSF adds up the living space above and below ground.

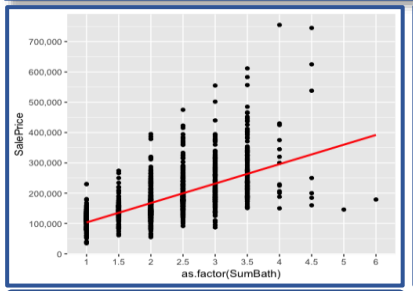

Fig.3 Figure of SumBath vs SalesPrice

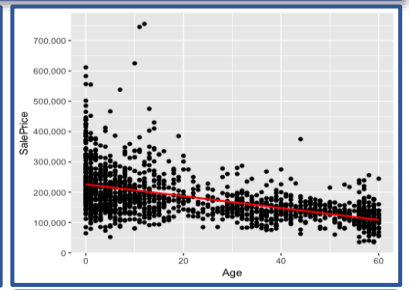

Fig.4 Figure of Age vs SalesPrice

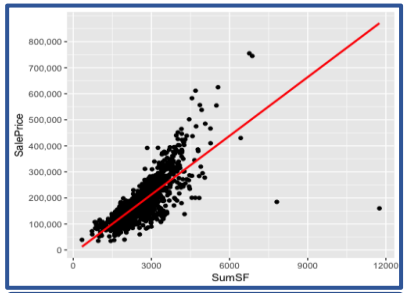

Fig.5 Figure of SumSF versus SalesPrice

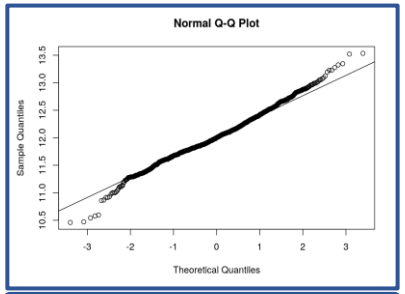

Fig.6 Normal Q-Q plot for log(SalesPrice)

4. Model Design and Predict

Methodology

- **Lasso [Submitted score:0.17125]**
  - Cross validation was used to tune the parameter lambda and when λ equals to 0.002 gives the best prediction.
- **eXtreme Gradient Boosting(XGBoost)[Submitted score:0.16001]**
  - XGBoost was then used with tuned parameter Max_depth=3, eta=0.05, Min_child_weight=4 based on 5 fold cross validation.
- **Support Vector Machine (SVM) [Submitted score:0.12079]**
  - SVM gives the best result comparing to above two models. After psrameter tuning 3, as the cost, gives the best prediction.
- **Weighted ensemble using Caret[Submitted score:0.11747]**
  - Both lasso and XGBoost was ensembled using CaretEnsemble.

5. Analysis & Conclusion

By analyzing the House price data using prescribed machine learning methods, several conclusions can be drawn from this project:

- Houses with larger living space, small value of house age, more bathrooms usually have a higher price.
- The log of Sale price are normally distributed and from the importance analysis of random forest model, "GrLivArea", "Neighborhood", "OverallQual" and "MSSubClass" are the most four important predictors.
- Nerghborhood do have a huge impact on the house price: for houses with neighborhood named"NoRidge", the mean price is over $325,000!

6. References & Acknowledgements

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Space Limited, for acknowledgements, please refer to source codes.

7. Contact Information

Haojie WANG (20371572)
PhD Student
Department of Civil and Environmental Engineering, HKUST
Tel: (852) 5422-9876
Email: hwangbw@ust.hk or h.wang@connect.ust.hk