QBUS2820 Assignment2

This version was compiled on November 1, 2020

Task A

Introduction

In a traditional manner, sale prices of houses were predicted by comparing sale prices and costs in the real estate market. There was no general standard to estimate the value of houses. Machine learning techniques therefore play an important role to help establishing models for sale prices of house predictions. As mentioned by Calhoun, the availability of a house price prediction model helps fill up an essential information gap and improve the efficiency of the real estate market (Calhoun, 2003).

This project aims to develop predictive models for sale prices of house with machine learning techniques. With the sale price which is a numerical variable being the response of predictive models, six models are developed and validated.

By comparing the root mean squared errors of predictions, the lasso regression model and random forest model are found to have the best predictive performance for the housing data, compared to elastic net, ridge regression, k-nearest neighbor regression and stepwise regression with forward selection.

Data processing and exploratory data analysis

There are 36 numeric variables and 43 categorical variables in the housing data. By calculating the correlation coefficient, 12 numeric variables are found to be potentially linearly related to sale price, as the absolute values of corresponsing carrelation coefficients are greater than 0.5. The distributions of these variables are visualised in figure 3. ‘TotRms AbvGrd’, ‘Garage Area’, ‘1st Flr SF’ and ‘SalePrice’ are shown to be right-skewed while ‘Garage Yr Blt’ and ‘Overall Qual’ are left-skewed, but these distributions are significantly influenced by outliers in several columns. Moreover, some variables, such as ‘TotRms AbvGrd’, tend to have linear relationships with other variables except sale price, leading to multi-collinearity. This could violate the assumption of some predictive models, such as multiple linear regression, thus robustness to multi-collinearity should be carefully considered when developing predictive models.

Figure 4 shows the distribution of sale price with regard to different categorical features. For most categorical features, sale prices tend to largely different for different groups of the categorical feature, except ‘BsmtFin Type 2’ and ‘Land Slope’. However, although medians sale prices look similar for different groups of ‘BsmtFin Type 2’ and ‘Land Slope’, the distribution of sale prices are not identical. Hence, it can still be worthwhile to includue these two variables as features to predict sale price. In addition, the boxplots also highlight the ouliers of sale price existing in diifferent categorical groups.

Besides affecting the shapes of data distribution, the existing outliers of numeric variables can also post a significant effect on predictive performance when making sale price predicions.Therefore data pre-processing needs to be considered in the stage of feature engineering in order to overcome issues caused by outliers.

Feature engineering

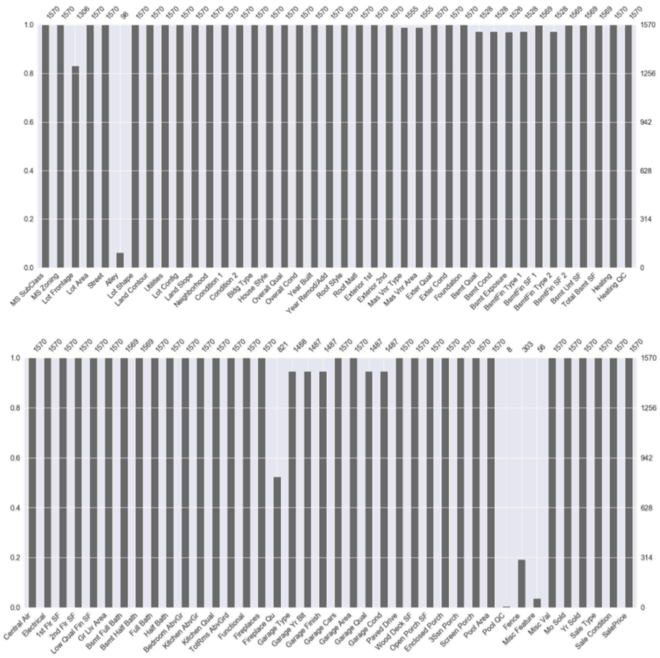

Fig. 1: *Visualizing missingness of housing data in training set.*

As shown in Figure 1, there are huge amounts of missing values in several columns: 'Alley', 'Fireplace Qu', 'Pool QC', 'Fence', 'Misc Feature', with more than 40% missing values within each column. With this issue, such variables are uninformative to be a feature of predictive models as there are too few observations. To deal with this, removing all rows with missing values can lead to significant loss of information, while imputation using small amount of observations can misrepresent the population for largely incomplete columns. Therefore, 'Alley', 'Pool QC', 'Fence', 'Misc Feature' are abandoned due to high missing rates.

Besides, there are 19 columns containing missing values but the percentages of missing values are less than 20%. This can be deal with by imputation. The missing values are imputed by using the most frequent value of each column.

As mentioned before, outliers exist with most of numeric features, which could be a big concern for predictive performance. Data standardization is therefore performed for numeric variables by subtracting the mean, followed by dividing the standard deviation of the corresponding columns.

After feature engineering, there are 74 informative features, with 36 features being numeric and 38 features being categorical. There are 1570 observations in the training set and 1210 observations in the testing set. For regression models involving categorical features, dummy variables are created for each categorical feature.

Methodology

Five regression models are trained by two sets of data where one training set contains numeric features only and the other set includes both categorical and numeric features. With regards to the issue of multi-collinearity mentioned before, all of the five regression techniques used are capable for addressing multi-collinearity. As such, the features surviving in feature engineering are all fed into predictive models.

To develope the best parameter set for each regression models, hyperparameters of are tunned with 5-fold cross validation. The performance of models with different value of hyperparameters is estimated by negative mean squared error, where a larger score indicates better predictive performance.

Random forest regression. Random Forest uses bootstrap sampling and feature sampling, i.e row sampling and column sampling. Therefore Random Forest is not affected by multicollinearity that much since it is picking different set of features for different models and of course every model sees a different set of data points.

This tells us the most important settings are the number of trees in the forest (n_estimators) and the number of features considered for splitting at each leaf node (max_features).

Lasso Regression. Least absolute shrinkage and selection operator (Lasso), as an extension of liner regression analysis, conducts both feature selection and regularization. This helps to enhance the predictive performance and interpretability of the model developed.

The objective of a lasso regression model is to minimize $\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$, where α is a tuning parameter which represents how strong the L1 regularization penalize coefficients of the lasso regression model. Changing the value of α influences the number of features eliminated. When $\alpha = 0$, coefficients of features are not Palisades such that no feature is removed. As the value of α increases, L1 penalty gets stronger and therefore more features are eliminated, vice versa. Furthermore, the value of α also affects the bias-variance trade-off. An increase of α leads to increase in bias while a decrease of α results in increase in variance.

Ridge Regression. Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

can do certain level of feature selection

Elastic Nets. Several researchers and data scientists have worked hard to explore the value of procedures like ElasticNets to help resolve the L1/L2 debate to multicollinearity correction. Through this technique, we are able to combine the strengths of both Ridge and LASSO regression, while minimizing the negative impact of either of these procedures. Some advantages of Elastic Net is that it is able to (1) enforce sparsity, (2) it has no limitation on the number of selected variables, and (3) it encourages a grouping effect in the presence of highly correlated predictors. A main disadvantage of this technique is that a naïve elastic net can suffer from double shrinkage, therefore, one needs to be careful when employing this option. If a naïve elastic net is found, a correction does exist to help control for this

Stepwise regression with Forward Selection.

Model / Features	Numeric	Numeric and Categorical
Forward selection	34470.77	5.21×10^{15}
Lasso	34543.61	29628.07
Ridge	34475.68	30082.79
Elastic net	35380.58	32844.26
Random forest	28016.33	28805.04

Table 1: *Summary of predictive performance for regression models. The performance measurement metric is the root mean squared error from cross validation.*

Validation set results

Conclusion

References

- Gibson, M., Little, R. and Rubin, D., 1989. Statistical Analysis with Missing Data. The Statistician, 38(1), p.82.
- Scikit-learn.org. 2020. 6.4. Imputation Of Missing Values — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/impute.html.
- Scikit-learn.org. 2020. 6.4. Imputation Of Missing Values — Scikit-Learn 0.23.2 Documentation. [online] Available at: https://scikit-learn.org/stable/modules/impute.html.
- Hintze, J.L., 1992. Chapter 335: Ridge Regression. In Number cruncher statistical system: statistical software. Kaysville, UT: Jerry L. Hintze.

- Chakon, O. (2017). Practical Machine Learning: Ridge Regression Vs Lasso. Coding Startups: Coders With Entrepreneurial Mindset. Published August 3rd, 2017.
- Allison, P. (2012). When Can You Safely Ignore Multicollinearity? Statistical Horizons.
- Cross Validated. (2015). What is elastic net regularization, and how does it solve the drawbacks of Ridge (L2) and Lasso (L1)? [online] Available at: https://stats.stackexchange.com/questions/184029/what-is-elastic-net-regularization-and-how-doesitsolve-the-drawbacks-of-ridge/184031#184031.

Task B

Exploratory data analysis

The time series plot figure 2 shows an upward trend from 1991 to 2016, with a seasonal pattern as symtematic changes occur in short periods which are fixed. Furthermore, the variation of number of visitors within the fixed period becomes greater as time moves. As such, a multiplicative forecasting model may be more suitable for this data compared to an additive model.

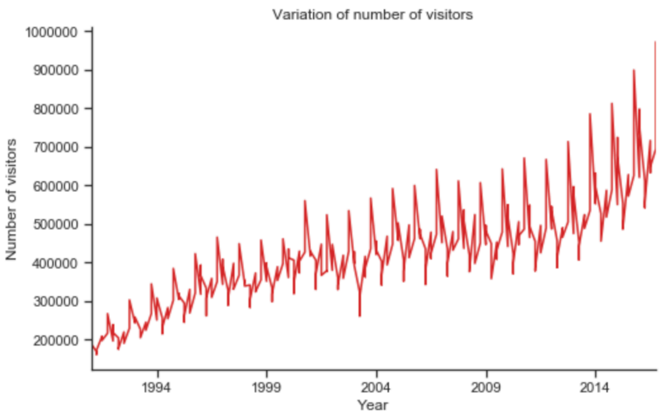

Fig. 2: *Time series of number of visitors from 1991 to 2016.*

Forecasting models

Seasonal random walk.

Drift model.

Exponential smoothing.

Appendix

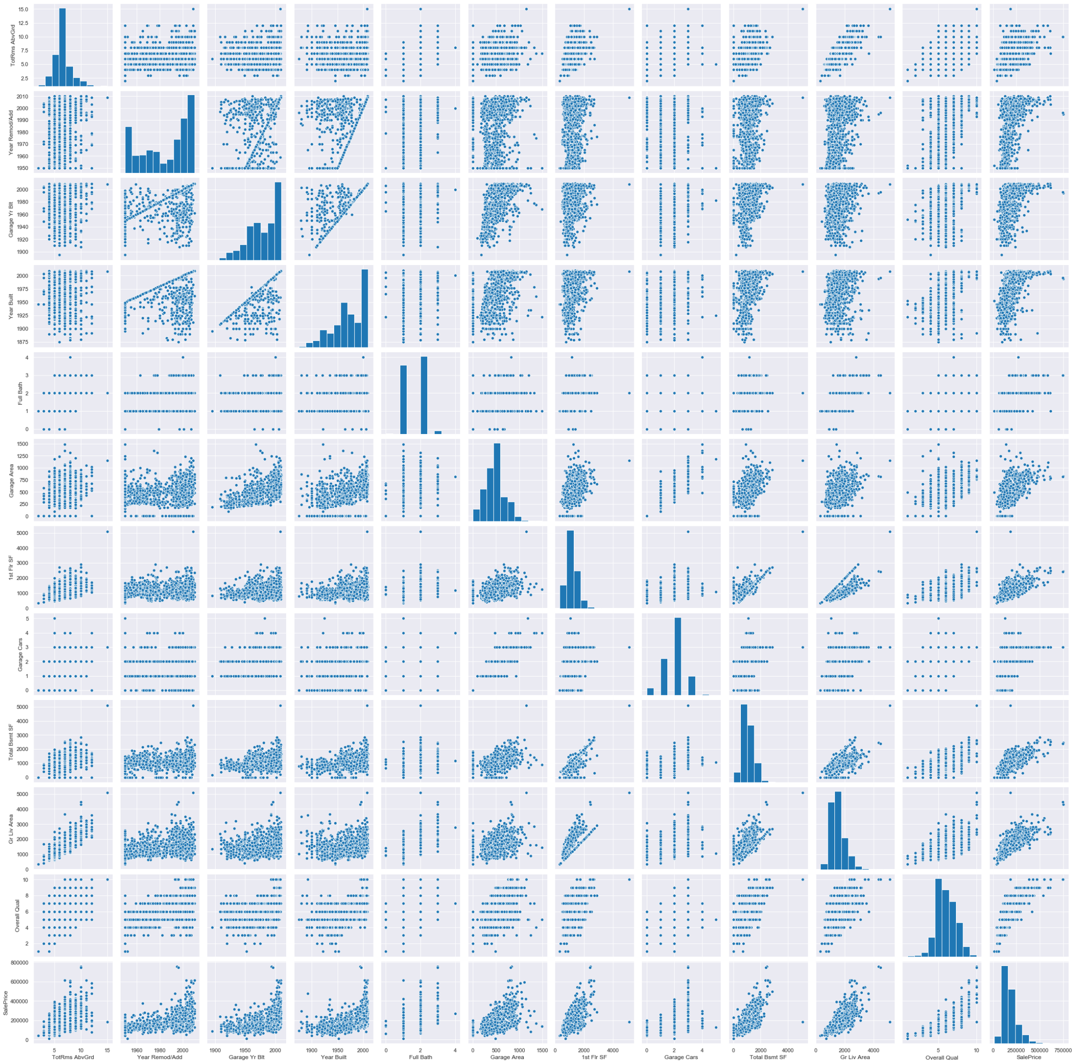

Fig. 3: *Distribution of numeric variables in housing data.*

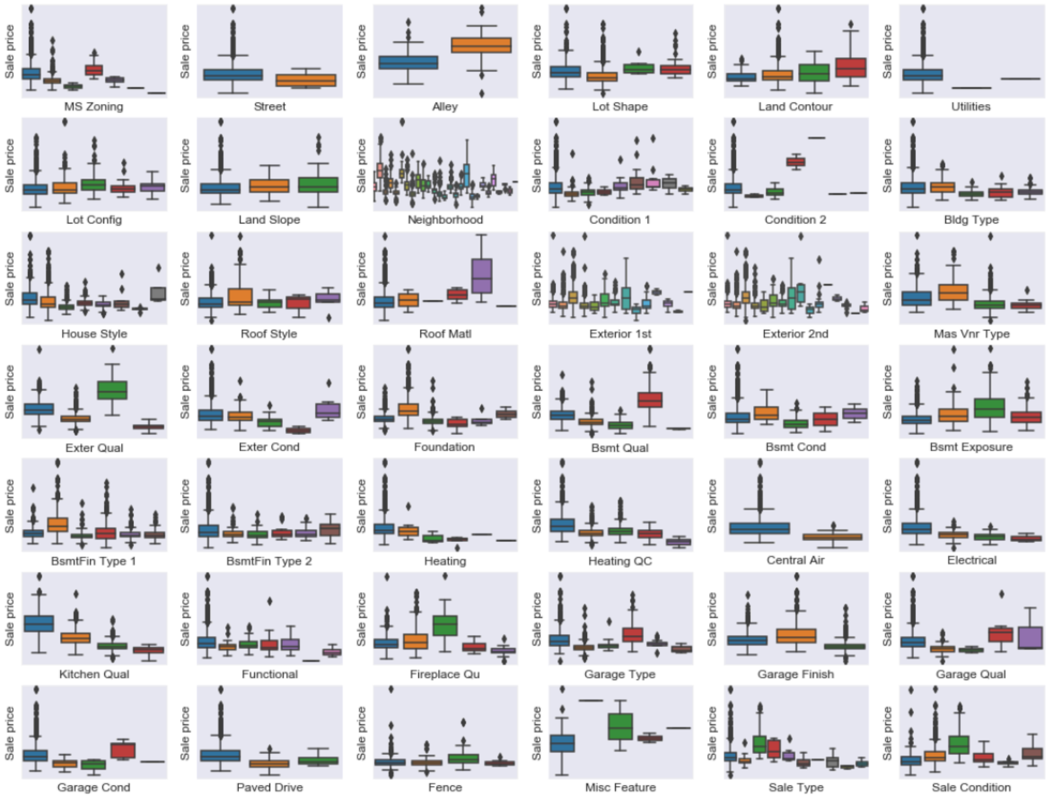

Fig. 4: *Boxplots demonstrating distribution of sale price for houses with different categorical features in housing data.*