REGRESSION PROJECT #1

Predicting Housing Prices: Regression Task (Boston Dataset)

Prepared By: Kisha Taylor
Date completed: 29-Sept.-2019

Report Outline

1. Objective
2. Executive Summary
3. Methodology of experiments
4. Data Exploration & Feature Selection
5. Model Performance Evaluation
6. Conclusion

1. Objective

We built various predictive models to predict the median price based on features such as Crime rate and Number of rooms etc. and identified the highest performing model based on specific evaluation metrics: Adjusted R-squared (Adj-R2) and Mean Squared Error (MSE).

2. Executive Summary

We used the Boston dataset to predict the median price based on features such as Crime rate and Number of rooms etc. and identified the highest performing model based on specific evaluation metrics: Adjusted R-squared (Adj-R2) and Mean Squared Error (MSE). The Adj-R2 measures the goodness of fit of the model adjusted for the number of predictors while the MSE is an error measure where the squared error for each data point is calculated and summed across all the datapoints and a mean value is calculated. Error is the difference between the actual and predicted value.

We followed the standard Machine Learning Model building pipeline (as noted below). The six distinct models (however over eight in total given sub-setting of features) varied in terms of complexity with the simplest model being the Linear Regression Model, then Penalized Linear Regression, Polynomial Linear Regression, Polynomial Penalized Linear Regression, Decision Trees, Random Forest. Finally, the model complexity peaked at an ensemble model called Gradient Boosted Machine. Note that a subset of features were used for some models and or the entire set of features.

Top Performer – Gradient Boosted Machine (MSE: 11.144 and Adj-R2: 0.867) where all features were used.
2nd Top – Polynomial Penalized (Lasso) Linear Regression (MSE: 12.818 & Adj-R2: 0.83).

Note that if interpretability was a high priority then I recommended choosing the 2nd Top performer (Lasso Regression -Polynomial using all features) since Gradient Boosted Machines (similar to other ensemble models) are much harder to explain the basis for the predictions.

The six distinct machine learning models which applied regression techniques and were evaluated were:

(i) Linear Regression
A statistical model which measures the linear relationship of the continuous dependent variable and multiple predictor (independent) variables. Note the linear relationship is defined with respect to the model parameters (feature coefficients). The prediction is a linear additive of the parameters of the predictive variables.

(ii) Polynomial Linear Regression
A variant of linear regression where the features have been transformed to polynomial features of n degrees (eg. X^2 where n=2 or X^3 where n=3 etc). This is effected in order to fit a non-linear function in cases where the underlying pattern of the data appears to be non-linear. It is still considered Linear Regression since the fitted line/function is still linear in its parameters.

(iii) Penalized (Lasso) Linear Regression

Another variant of a linear regression model where the model parameters are shrunk towards zero. In this case of Lasso regularization, the parameters can actually be shrunken all the way to zero (for Ridge Regression, the shrinkage is towards zero but never zero). In light of this, this model innately conducts feature sub-setting.

(iv) Decision Tree

A hierarchical tree-based statistical method that uses a top-down greedy search through the input space and iteratively partitions the input space based on the feature which is the best at reducing the mean squared error at that level in the tree (before versus after a split of the current data on that feature) in classification the best predictor is assessed based on an impurity measure (eg. Entropy).

(v) Random Forest

This is an ensemble method that uses a decision tree as the base learner. The method randomizes by row and column thereby randomly selecting a subset of the dataset to grow a tree (a learner) and randomly selects the candidate features from which a best feature will be chosen. It uses bootstrap aggregation (Bagging – random sampling with replacement) to derive the final prediction using multiple learners. Bagging reduces the variance of the model given that we average over the individual learners. Note in Random forest we generate the base learners in parallel, one independent of the other. This is in contrast to gradient boosting which we will delve into in the next section below.

(vi) Gradient Boosted Trees

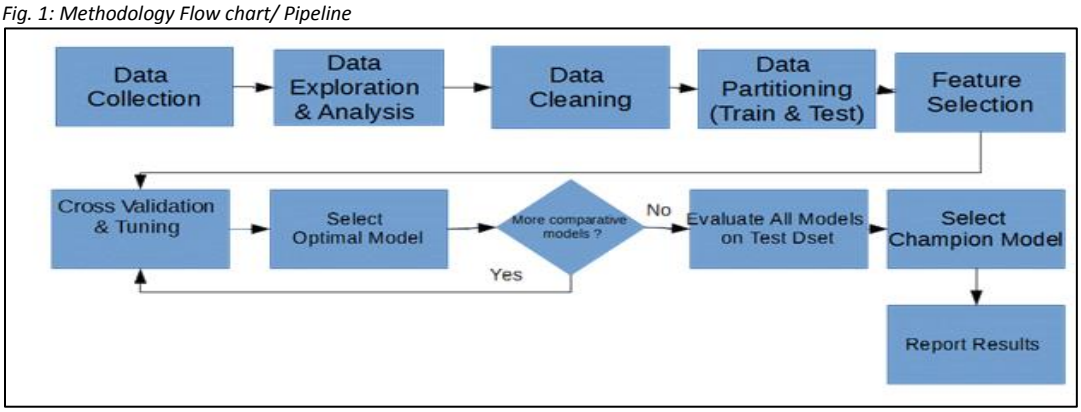
An ensemble method where more than one decision models work together to produce best results with each model that is gradually added (sequentially in contrast to random forest were multiple leaners are generated in parallel), trying to correct the mistakes of the previous model.

3. Methodology/ ML Pipeline

A standard methodology was applied which involved cross validation (10-fold used), tuning of models and evaluation on test dataset. The top model was selected based on the MSE and Adjusted R2-squared value. The models used a subset of features and/or the entire set of features. The Adj-R2 measures the goodness of fit of the model adjusted for the number of predictors while the MSE is an error measure where the squared error for each data point is calculated and summed across all the datapoints and a mean value is calculated. Error is the difference between the actual and predicted value.

As shown in the Methodology flow-chart below, we first started by collecting the data. The dataset was imported from one of the packages in Python, sklearn dataset. ***Data exploration*** was conducted to get a better understanding of the dataset and features, the distribution of the data, the range of values etc. (descriptive statistics). This was followed by ***cleaning of the data***, a check for missing values (in this case there were none). A check was also done for outliers and categorical variables (none in this case). ***Data Analysis*** was performed to assess the existence of a linear relationship between the dependent & independent variables and the strength thereof via a correlation matrix. Subsequently, ***Data Partitioning*** was performed to separate the dataset into training and test. Then feature selection was conducted using Variable Inflation Factor (VIF) technique to identify redundant features. VIF assesses a feature for multicollinearity by a computation mainly involving the Rsq (or R2) value where the feature is regressed against all the other predictor features. The formula is: $VIF_j = 1/(1-Rsq_j)$. Note that the input values for each model as mentioned earlier varied from a subset of features to the entire set of independent features***. Cross-validation (CV) & tuning*** was performed (K-fold training and testing/validation & tuning). The ***optimal model was selected*** and the CV was repeated for all models explored. Finally, the optimal models were each ***evaluated on the test data,*** a ***champion model selected*** and the ***results reported and conclusions*** drawn.

Fig. 1: Methodology Flow chart/ Pipeline



4. Data Exploration & Feature Selection

The main results of the data exploration and feature selection are shown below. Note that there was moderate to high correlation of a few features versus the median price (dependent variable) as shown in the heat map Fig. 2A and Fig 2B (table of correlation values). After applying the VIF technique and dropping the redundant features (features with VIF over 5 dropped) the subset of features selected were CHAS, RAD, ZN, CRIM, DIS and LSTAT with correlation values ranging from 0.175 to a high of 0.738 (magnitude), see Fig. 2C below. Note that this dropping of redundant features was only necessary because our first model being explored was Linear Regression and high multicollinearity could impact the interpretation of the results given the effect of increased variance in the coefficient estimates and thus increased sensitivity of the model to small changes (eg. using other predictive features).

Fig. 2A: Correlation matrix - Heatmap

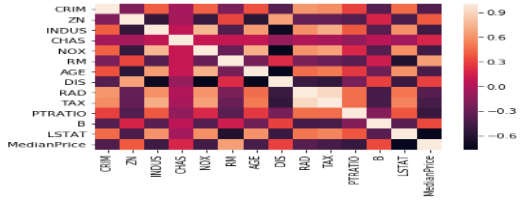


Fig. 2C: Correlation values for selected features

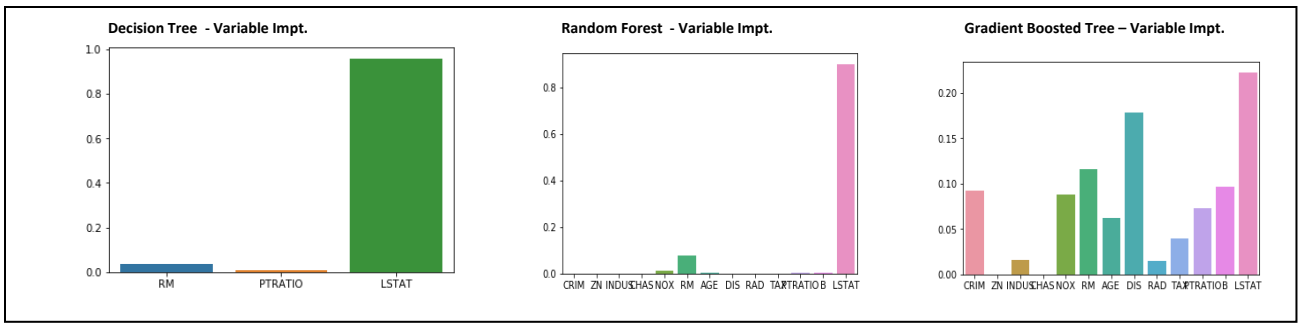
Correlation values vs. median price- after redundant features dropped						
	CHAS	RAD	ZN	CRIM	DIS	LSTAT
MEDIAN PRICE	0.175	-0.382	0.36	-0.386	0.2499	-0.738

Fig. 2B : Correlation values of features versus Median Price (target/dependent variable)

LSTAT	PTRATIO	INDUS	TAX	NOX	CRIM	RAD	AGE	CHAS	DIS	B	ZN	RM	MedianPrice
-0.738	-0.508	-0.484	-0.469	-0.427	-0.386	-0.382	-0.377	0.175	0.250	0.333	0.360	0.695	1.000

The variable importance results below in Fig3 across the Decision Tree, Random Forest and the Gradient Boosted Tree have some similarities. All the models reflect LSTAT (% lower status of the population) as the most important feature. This is reflected in the correlation values above. The RM feature (average number of rooms per dwelling) is also common across models in the top 3 most important features (2nd place for Decision Tree and Random Forest but 3rd place for Gradient Boosted Tree).

Fig 3: Variable Importance across selected Tree-based models



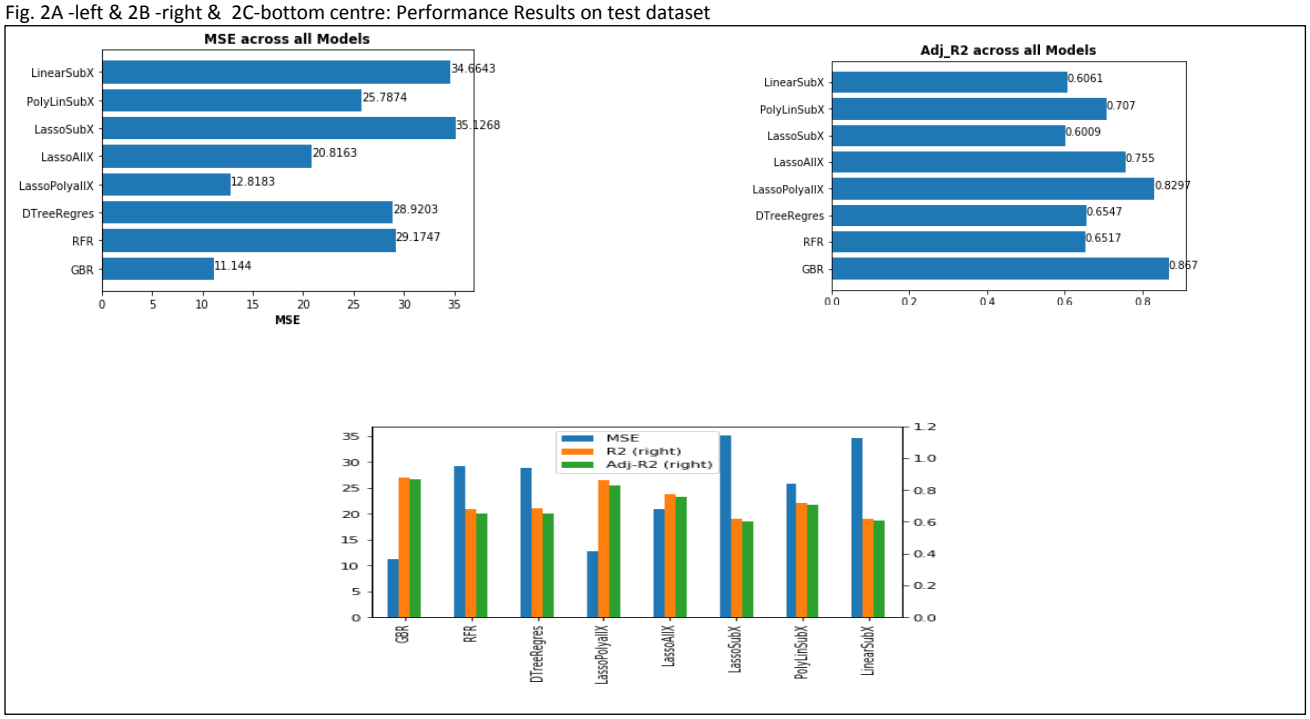
Attribute Information:

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per $10,000
- PTRATIO pupil-teacher ratio by town
- B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in $1000's

5. Performance of Models

The results below (in Table 2) reflect the top performer as Gradient Boosted Tree based on two main evaluation metrics: MSE and Adj-R2 with values of 11.144 and 0.867 respectively. This top model used all features or training and derived the top three (3) most important features as LSTAT, DIS and RM. The Polynomial Lasso Penalized Linear Regression model followed in 2nd place with respective MSE and Adj-R2 values of 12.818 and 0.83, this model also used all the features in the dataset during training.

Fig. 2A -left & 2B -right & 2C-bottom centre: Performance Results on test dataset



6. Conclusion

Based on the results the top performing model, Gradient Boosted Tree, should be deployed as it yielded the best results on the test dataset based on lowest MSE and highest Aj-R2. However, note that the Polynomial Linear (Lasso) Regression model which held 2nd place should be used if interpretability is a high priority.