Modeling House Prices

Random Forest model with Feature Transformation, Hyperparameter Optimization, and Cross-validation

By Ahmad Qadri (arqchicago@gmail.com)

House Price dataset

This dataset[1] is a variation of the House Sales dataset in King County, Washington. The data is from Kaggle website and comes with Public Domain license. A variable called Weight was added to this dataset to specify the number of similar homes that were sold in the area. The target variable represents sale price of houses. The features measured in the dataset are age of the house, square feet area of the living area, square feet area of the lot, number of bedrooms, number of bathrooms, age of appliances, crime rate in the area, number of years since last major renovation and grade condition of the home.

Random Forest Regression

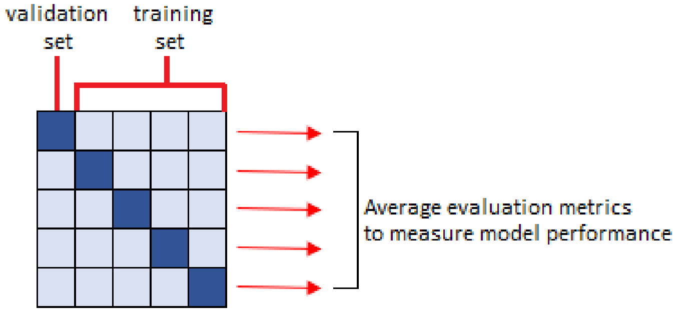
Random Forest is an ensemble method in which many decision trees are trained by bootstrapping the data and aggregating the results at the end. For regression problems, the average of predictions from all decision tree is taken as the overall regression prediction. In each decision tree, data is continuously divided based on values of randomly selected features and purity of resulting nodes is computed to evaluate the effectiveness of the split. In regression Random Forests, the function that measures the quality of split in each decision tree is based on reducing the variance in a node. The idea is to divide the node into sub-nodes to create relatively pure nodes such that the overall variance is reduced. The formula to calculate the sample-based variance is shown below.

$$Variance = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

In each split, variance of each child node is calculated. Overall variance for the split is calculated as the weighted average variances of child nodes and the split with the lowest variance is picked as the best split. This process continues until most homogeneous nodes are achieved. It is important to note that expanding the tree until completely homogenous nodes are achieved is not the most appropriate approach since it leads to the model overfitting the training data and performing poorly on unseen test or validation data. Additional criteria is specified for the terminal nodes to stop the splits after a minimum number of data points remain in those nodes. In other words, a compromise is reached in the terminal nodes such that perfect homogeneity is avoided in favor of the model generalizing well over the underlying distribution of data.

Cross Validation

This Random Forest regressor models sale price of houses which is a quantitative variable. Standard regression metrics are used to optimize the model. This includes Root Mean Squared Error and Root Mean Absolute Error. These evaluation metrics are collected using 5-fold cross validation to avoid overfitting on the training set. This data is collected for each iteration in the hyperparameter optimization process.



5-fold Cross Validation

Hyperparameter Optimization

Machine learning models use various parameter settings that have an impact on the cost function. Random Forest models involve a set of parameters that developers can optimize to evaluate the cost function. For example, one parameter setting is maximum depth of decision trees that are built. This setting allows developers to cap the depth of decision trees to avoid overfitting model on the training set. If this setting is not optimized, the tree is expanded until all leaves contain data points from the same class. This can lead to severe overfitting. The parameters tuned for modeling heart disease data optimized parameters including the number of decision trees, the number of features to consider for the best split, maximum depth of each tree, minimum number of data points required for splitting a node, minimum number of data points required to be a leaf node and number of data points required to be a leaf node and whether a bootstrap should be used to build a tree instead of the full data set.

Pipeline

This model utilizes feature transformation techniques before model training step. A Scikit-Learn class called Pipeline is used to combine the steps of feature transformation and model building. This makes the process of applying parameter optimization through grid search and cross-validation data collection easy to implement. Pipeline allows us to build several critical steps in which data is pre-processed, analyzed and modeled using simple function setup and calls. Pipelines class can be used to efficiently apply many different pre-processing steps to test model optimization using different combinations of pre-processing and modeling steps. In this model for housing prices, Standard Scaler feature transformation was applied to the feature set followed by modeling and evaluating cross-validation data. In the standard scaler pre-processing step, features are standardized by removing the mean and scaling to their unit variances.

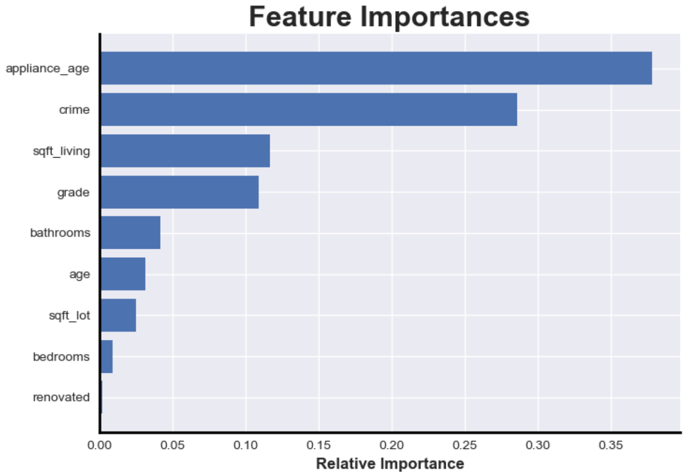
Model Results

Random Forest regressor model was run on the data set with 5-fold cross validation and hyperparameter tuning. The best parameter settings picked were based on minimizing the root mean squared error. Model evaluation metrics including Root Mean Squared Error, Root Mean Absolute Error and Coefficient of Determination were collected. The metrics for the train and test sets based on the best set of hyperparameters are shown below.

Metric	Training Set	Testing Set
Root Mean Squared Error	47,105	69,290
Root Mean Absolute Error	190	230
Coefficient of Determination	0.9130	0.8147

Feature Importance

Once the best model is selected, the next step is to evaluate the significance of each feature in predicting the target variable using that model. The importance score can be calculated that is useful in understanding the model, patterns in the data and to reduce the number of features to simplify the model. Feature importance scores were calculated for the model trained on heart disease data set.



The 5 most important features adding value in the model included age of appliances, crime, living area, grade of the home and the number of bathrooms. In general all variables except for years since renovation are important and should be included in the model.

Model Persistence

Once the best model is obtained and the evaluation metrics are satisfactory, the next step is to persist the model so that it can be used in the future without the need to retrain or reproduce it. This can be accomplished by object serialization in which a model is converted into byte stream and saved on the server. When the model is needed for predictions in the future, the byte stream is converted back into the model. Shell commands can be used to encrypt the model file and to restrict access to select users.

Modeling House Prices

Random Forest model with Feature Transformation, Hyperparameter Optimization, and Cross-validation

By Ahmad Qadri (arqchicago@gmail.com)

References

[1] This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.
Kaggle.com
Retrieved from
https://www.kaggle.com/harlfoxem/housesalesprediction

Modeling House Prices

Random Forest model with Feature Transformation, Hyperparameter Optimization, and Cross-validation

By Ahmad Qadri (arqchicago@gmail.com)