Philip Demeri

Capstone Project 2: Milestone Report 2

Machine Learning – Overview

The machine learning phase largely involves, at each month, training classification models on 75% of the observations and testing on the remaining 25%, and then looping over all months. The 75/25 partition is accomplished by train\_test\_split.

The three base estimators selected for this project are the logistic regression, K-nearest neighbors, and random forest classifiers. A random state is initialized for purposes of reproducibility.

Ensemble models that this project integrates includes the voting classifier, bagging, and boosting. The AdaBoost classifier is selected for boosting. All three classifiers are legs of the voting classifier, whereas only the random forest classifier is the base estimator for bagging and boosting.

From a high level, the machine learning phase is subdivided into (1) running the base and ensemble models using the base models’ default hyperparameters, and (2) tuning the base models’ hyperparameters on a subset of the data (the first 25 months), and then re-running the base and ensemble models on the remaining N – 25 months.

In all cases, the objective is to predict, at a given month, which stocks will outperform (target = 1) and which will underperform (target = 0) in the next month. Consideration is given to the accuracy on the test set, but as a sanity check, the accuracy on the training and validation sets will briefly be examined.

Machine Learning – Default Hyperparameters

The project begins by implementing the logistic regression, KNN, and random forest models, using default hyperparameters, at each month, looping over all months, and then reporting the mean test set accuracy over all months. The results for logistic regression, KNN, and random forest, respectively, are: 58.6%, 56.8%, and 56.1%. Plots of these accuracies when indexed by time indicate a slight incline. Correlations between each of the accuracies and the number of stocks at each month reveal that the accuracies are weakly positively related to stock count (0.26 to 0.37).

The models are then re-run over all months, this time implementing 10-fold cross validation on the training set. The purpose is to gauge the reliability of the results of the test set. For all models, the validation set accuracies closely mirror those of the test set. Further, the training set accuracies are all higher than those of the validation and test sets: 63.3% for logistic regression, 72.3% for KNN, and 97.0% for random forest. These results are intuitive in that the nonlinear KNN and random forest models outperform the logistic regression model when applied to the same set of data. No reason, then, exists to scrutinize the integrity of the data, and hence the project moves forward to ensemble learning.

The voting classifier rests on the hard outcomes of each of the three classifiers without regard for the probability of each outcome. The mean accuracy of the voting classifier over time is 58.1%, which outperforms the accuracy of each of the KNN and random forest models but underperforms that of the logistic regression model.

Bootstrap aggregation uses the random forest as its base estimator. Accuracy is reported both on the test set and on out-of-bag realizations. The accuracies are close to each other, and close to those of other models: 58.1% for test set accuracy and 58.4% for out-of-bag accuracy.

The AdaBoost classifier is selected for boosting, using the random forest as in bagging. The reported accuracy is 56.1%, which lies on the low end of those witnessed thus far.

Machine Learning – Tuned Hyperparameters

The original dataset ranges from December 1994 to July 2015, with monthly periodicity. To tune the hyperparameters for the logistic regression, KNN, and random forest classifiers, the dataset is partitioned into (1) the first 25 months (December 1994 to December 1996) for tuning, and (2) the remaining N – 25 months in order to run the newly tuned models using the same workflow as described earlier (in this phase, however, the 10-fold cross validation phase is skipped, as its function was to measure the integrity of the data within a machine learning workflow). GridSearchCV is implemented over the entire 25-month period using ten-fold cross validation. The tuned hyperparameters are: logistic regression – C; KNN – n\_neighbors; and, random forest – n\_estimators, max\_depth, and max\_features.

With the now-tuned models, re-running each from January 1997 to July 2015 leads to a decline in performance for the logistic regressor but an improvement for KNN and the random forest: 58.1% for logistic regression (previously 58.6%), 58.8% for KNN (previously 56.8%), and 58.8% for random forest (previously 56.1%).

The ensemble models also witness performance improvement, even though no hyperparameters were tuned at the ensemble level: voting classifier accuracy is now 59.4% (previously 58.1%), bagging now 59.1% for both test set and out-of-bag accuracy (previously 58.1% and 58.4% for test set and OOB accuracy, respectively), and boosting now 56.4% (previously 56.1%).

Regarding the variance of the accuracies between the tuned and untuned implementations, no appreciable difference of tendency exists, as is the case for the mean of the accuracies. Hence, the tuned models, broadly speaking, more accurately predict the next month’s outperformers from underperformers without witnessing greater variation among predictions.

Finally, as in the case of the untuned models where accuracy is weakly positively correlated with the number of stocks at each month, such correlation manifests itself in the case of the tuned models, where the correlation ranges from 0.30 to 0.39.