Predicting Mutual Fund Returns

Mary Donovan Martello

DSC 550 Final Case Study

Introduction

The analysis problem addressed in the case study is: Can mutual fund year-to-date returns be predicted from data describing the mutual fund and from historical performance data of the mutual fund? Are any certain features of the mutual fund more correlated to year-to-date returns?

Publicly traded mutual funds are a type of investment that may be purchased and sold on public stock exchanges. A mutual fund is a collection of stocks, bonds, and/or other securities wrapped together as a single investment. Investors need information on the mutual funds in order to make purchase and sale decisions. Thus, information on publicly traded mutual funds are available on websites like Yahoo! Finance and API's like FMP Finance, and investors would benefit from a model that prediction year-to-date returns of the mutual funds.

In attempting to predict mutual fund year-to-date returns, the data features examined include: net assets of the mutual funds, size and investment approach of the funds, portfolio indicators (as cash, stocks, bonds, and sectors), year-to-date and historical returns, and financial ratios (as price/earnings, Treynor and Sharpe ratios, alpha, and beta).

Analysis

The key components of the analysis that I would like to highlight are: (1) dimensionality reduction, (2) feature selection, (3) model evaluation and selection, and (4) assessment of models.

Dimensionality Reduction

Given the very large number of features in the dataset, dimensionality reduction was important. After cleaning the data for missing values and messy and nonunique features, features with low variance and high multicollinearity were addressed by performing a Principal Component Analysis (PCA) on the quantitative features. The PCA analysis reduced the number of features by 31 features. The categorical features were then added to the PCA features for an adjusted data frame to use in modeling.

Feature Selection

One of the goals of the analysis is find which of the many features are significant predictors of the target year-to-date return. Another purpose of feature selection is to see if a smaller subset of the predictor features is better at predicting the target than the entire set of predictive features.

The exploratory data analysis (EDA) was helpful to show high correlations between the target variable (ytd_return) and: portfolio stocks and bonds (i.e., portfolio makeup), technology sector investments (a type of holding of some mutual funds), Category ytd_return (year-to-date returns of other mutual funds in the same category), and Fund and Category 3-year returns (3-year returns of the Fund and of other mutual funds in the same category).

I also ran a function to see which of the PCA components are the best features. I ran various basic regression models using those variables to see how well they did compare to the

baseline linear regression model with all variables. The top PCA components did not result in good baseline regression models. Next, I created subsets of the features based on the EDA results and ran baseline linear regression models with these subsets to try to identify a smaller set of variables to use in model evaluation instead of using all the features. The best subsets of predictive features, based on the coefficient of determination results, include subsets with categories of mutual fund factors identified in the EDA:

- Portfolio Makeup (i.e., what types of investments does fund hold): .71
- Category and Fund Returns (for prior periods): .96
- Mean Annual Returns (year-to-date returns for prior periods): .81

Model Evaluation and Selection

Model evaluation includes assessing the general performance of a predictive algorithm and evaluating how well the model will work in the real world. I decided to evaluate two predictive algorithms. Linear Regression is used for predictive analysis for continuous data, and thus seems appropriate for the ytd_return target. I also tried a LASSO (least absolute shrinkage and selection operator) linear regression model in order to try to reduce features because my dataset still has many features. LASSO regression reduces the magnitude of coefficients in the model, and accordingly reduces the inputs as the coefficients of some features are reduced to zero. Also, a goal of LASSO regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero (Jain, 2017).

The baseline models performed on the linear regression and LASSO linear regression models showed high coefficient of determination (R-Squared) results and thus I determined that they were viable models for model selection.

Feature selection involves selecting which predictive features and hyperparameters will be used as inputs in the models. For model selection, I tested each of the linear regression and LASSO linear regression models with the full set of predictive features and with only the best subset of key features identified in the feature selection process discussed above (i.e., Category and Fund Returns for prior periods). A grid search was performed to tune the LASSO model's hyperparameters (i.e., using different values for the hyperparameters).

Assessment of Models

The linear regression and LASSO regression models utilizing the best subset of key features as inputs (i.e., Category and Fund Returns for prior periods) produced the better coefficient of determination. I made predictions on the regular linear regression and LASSO regression models with the subset of key features as inputs, and then calculated performance metrics. The results for both models were very good, with the following results:

	Linear Regression / Cross-validation (Categorical and Fund Return features	**LASSO Regression / Cross-validation (Categorical and Fund Return features**
R-Squared	0.963044	0.962770
Root Mean Squared Error	0.997036	1.000720

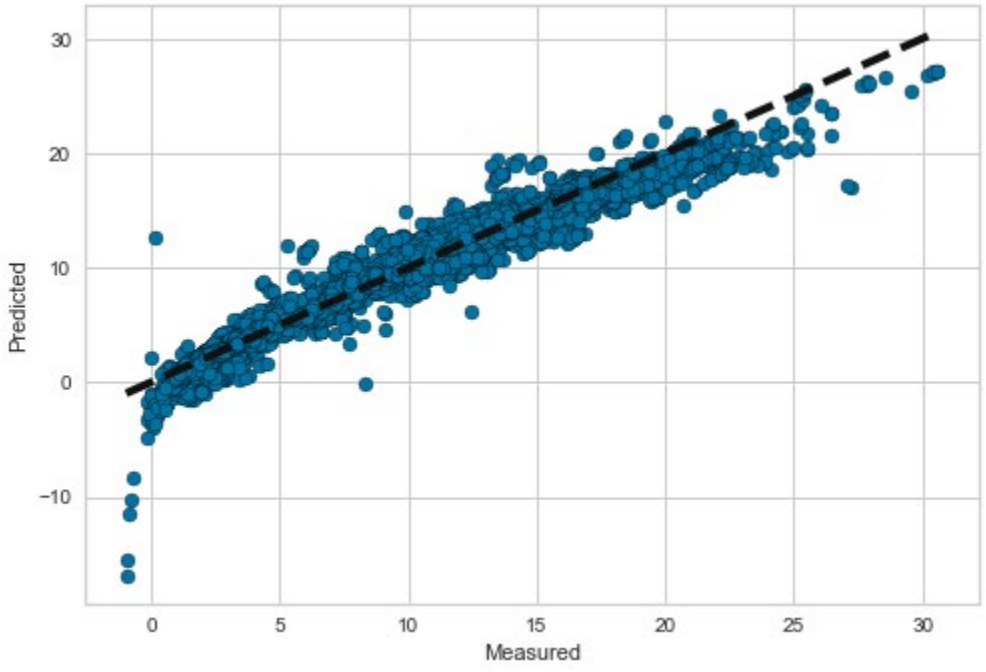
The coefficient of determination (R-Squared) tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variables. R-Squared is the square of the correlation between the observed values of the outcome variable and the values of the outcome variable predicted by the multiple regression model. Its values range from 0 to 1; 1 means that the predictive feature inputs perfectly explained the target (year-to-date return) through the model; 0 means the predictive feature inputs explained nothing of the target

feature. Both the linear regression and LASSO regression models' R-Squared results are almost 1 and thus indicate the models are highly predictive of the target.

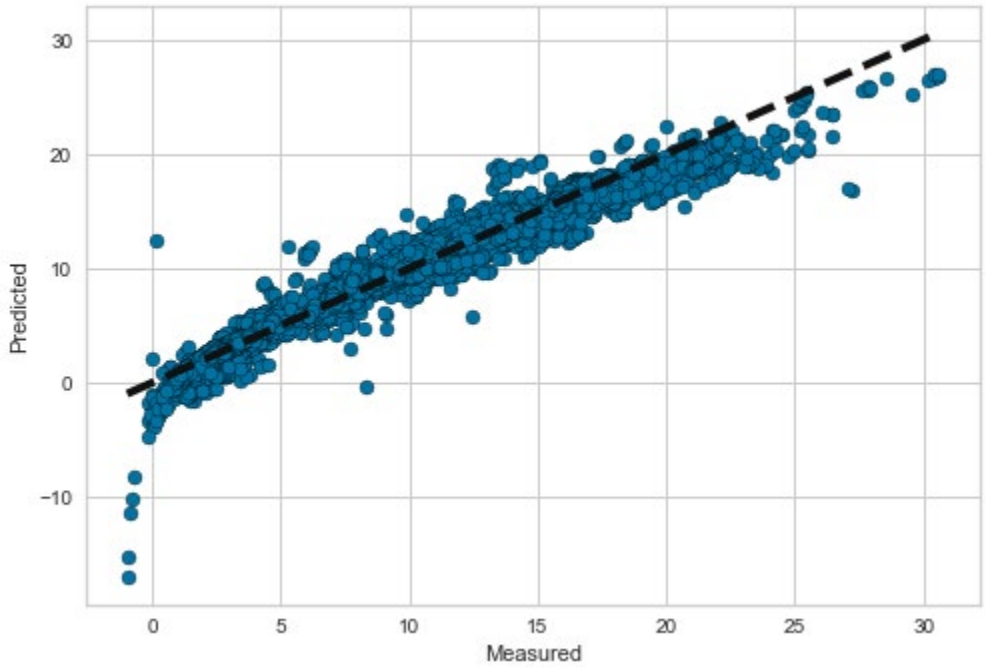
Root mean squared error (RMSE) is the square root of mean squared error. Mean squared error measures the average squared difference between the estimated values and the actual value in the prediction. Thus, it measures the error of the prediction. Accordingly, the smaller the score, the more accurate the model. The square root of the mean squared error is taken to convert the result to base units.

The linear regression and LASSO regression models' RMSE are just under 1 and just over 1, respectively, and thus indicate that the predictions are quite accurate. The following graphs, which chart the predictions versus the actual target values, shows how accurate the models performed:

Linear Regression Predictions vs Actual



LASSO Linear Regression Predictions vs Actual



Conclusion

The results of my linear regression and LASSO linear regression models indicate that mutual fund year-to-date returns can be predicted to a high degree of accuracy from the mutual fund's historical return performance. Specifically, the following features are significant in making the year-to-date return predictions as of 5/3/19:

- fund_return_1month
- category_return_1month
- fund_return_3months
- fund_return_1year
- category_return_1year
- fund_return_3years
- category_return_3years
- fund_return_5years

- category_return_5years
- fund_return_10years
- category_return_10years
- fund_return_2018
- category_return_2018
- fund_return_2017
- category_return_2017
- fund_return_2016
- category_return_2016
- fund_return_2015
- category_return_2015

These linear regression and LASSO linear regression models show that these predictive models almost entirely explain the predicted year-to-date returns and that the errors in the predictions are minimal (RMSE of .9970 and 1.0007, respectively). Thus, investors would benefit from these models predicting year-to-date returns of the mutual funds.

Finally, although these two completed linear regression and LASSO linear regression models performed well, further work may show improvement. A larger dataset for training the models would also help to improve the performance of each model because there would be a larger amount of data points for each model to utilize. Also, ensemble models could be tested to see if they could improve performance.

References

Jain, S. (2017, June 22). A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/