# Evaluating Funds Profitability and Sustainability Performance

Qinming Zhang, Tingyu Zhang, Yiyao Guo, Qinruo Hu, and Chengxi Xu (Team USC)

University of Southern California, Los Angeles, United States.

**Abstract**

The rising consciousness among the general public towards sustainability, the evidential irreversible damage from human activity toward the environment, and the recent awakening trend of social equality, viewing investments as mere means of wealth accumulation seems to be inadequate. More and more fund managers take substantial consideration in advocating positive social and environmental impact with their investment decisions. Yet some of them are deeply concerned about the cost of such a good deed.

Overall, this report contains two parts. The first part provides details of data processing, regression and correlation analysis insights, and the second part introduces our new model of scoring funds using Random Forest Regressor.

## 1 Introduction

Since the current Morning Star Rating methodology focuses only on funds' potential risk exposure and its risk management capability, it is hard for investors to know if the fund is socially responsible, if it will generate and quantify a positive societal impact and most essentially if it will result in capital growth. For this project, our group comes up with a different indicator of a fund by examining both its positive or negative effect on the society and its profitability. To achieve that, we first testify the possible factors that contribute to the formation of ESG(environmental, social, governance) values using regressions. After finding out such a relationship, we prove that ESG value is related to the social and environmental factors the public cares about and could be used for measuring the effects of certain funds. Giving different weights to both financial and ESG variables based on reasonable assumptions, we obtain our target variable and apply Random Forest Regressor on other factors to train the model on predicting it.

The final result of the project will be a financial dashboard, allowing investors to input the ticker of their intended investment funds and output the indicator to help them come up with more rational impact-investment strategies.

# Part 1 – Insights

## 2 Data Collection

Our source data is obtained directly from RMDS required Morningstar dataset and Bloomberg USSIF data. The Morningstar fund dataset contains 6003 funds with their size, financial performance and sustainability performance. The Sustainable Investment Mutual Funds and ETFs Chart dataset, which we collect from USSIF website, is provided by Bloomberg. We focus on the Screening  Advocacy part that contains 179 funds with their self-reported information on ESG metrics.

# 3 Data Cleaning

After collecting and organizing the internal and external data, we start to perform data cleaning. Overall, the cleaning steps involve two parts: merging datasets and dealing with null values. By merging both Bloomberg and Morning Star data through Ticker, a comprehensive dataset is formed with information from both datasets. For null values, we either drop empty values for dependent variables or fill in the mode values for categorical variables. Then, we extract three Portfolio ESG scores, Sustainable Investment indices from Morning Star data and Screening Advocacy part from Bloomberg data for further analysis.

# 4 Data Mining

## 4.1 Regression analysis (Portfolio Environmental Score)

Other than the original merge between the two datasets, the complexity of data regarding environmental factors of the funds requires some further data cleaning before regression. First, the column "Climate/Clean Tech" includes information about whether the particular fund has a formal policy to restrict investment in fossil fuels by adding "*" at the end of the category label. The combination leads to up to 7 unique categories in this column, for instance, labels "C*" and "C" indicates the same category on the "Climate/Clean Tech" aspect and differs on the fossil fuels policy only. Too many categories brings unnecessary redundancy in analysis, so our team creates a new binary column called "Fossil Fuels Restrictions" to keep track of the fossil fuels investment policy in funds and replaces all missing values by the mode "C" of the data. Using mode to fill empty cells is used for the remaining columns "Pollution/ Toxics" and "Environment/Other" from the Bloomberg database. Then all 4 columns gathered or generated from the Bloomberg fund database are transferred to dummy variables for regression analysis.

For the Morningstar data, we extract four sustainability-related variables ("Sustainable Investment by Prospectus", "Sustainable Investment - ESG Fund", "Sustainable Investment - Impact Fund", "Sustainable Investment - Environmental Sector Fund") to include in the regression, because these four variables might be highly related to the target we want to analyze. However, interestingly for the variable "Sustainable Investment by Prospectus", all the funds yield the YES value, therefore this variable might not contribute to the regression analysis since we regard it as a constant.

During the coding process, we use the function "get dummies" to convert categorical variables to dummy indicators instead of directly using label encoding and also reduce extra columns. In this way, the regression result can eliminate biases. We then employ the OLS model to generate a comprehensive report.

Having the variable "**Portfolio Environmental Score**" from the Morningstar database as the target variable, the first regression is conducted with other environment-related variables from both datasets. As it is possible for all the variables to be 0 in this regression, in another word, the fund does not directly involve in any sort of climate related investment, the environment score of a particular fund can be 0. To accommodate such speciality in the data, the model applied does not have an intercept. Subsequently, this regression performs well with 0.953 R-square and 209.4 F-statistic, which means the model highly fits the selected data and proves the predictive capability of the model, preliminarily supporting the assumption that the variables chosen for the regression are strongly related to the target variable, Portfolio Environment Score. The model yields the below results:

With the detailed information of each variable, variables like "% Fossil Fuels", "Environment/Other-P", "Sustainable Investment - ESG Fund_Yes" and "Sustainable Investment - Environmental Sector Fund-Yes" have statistically significant coefficients with p-value close to 0. Among them, "Sustainable Investment - ESG Fund_Yes" has

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| % Alcohol | 0.3940 | 0.317 | 1.242 | 0.216 |
| % Fossil Fuels | 0.0532 | 0.023 | 2.349 | 0.020 |
| % Small Arms | −0.1093 | 0.373 | −0.293 | 0.770 |
| % Thermal Coal | 0.0035 | 0.067 | 0.052 | 0.958 |
| % Tobacco | 0.1961 | 0.665 | 0.295 | 0.768 |
| Fossil Fuels Restrictions | 0.0720 | 0.204 | 0.352 | 0.725 |
| Climate/Clean Tech_C | 0.2205 | 0.204 | 1.080 | 0.282 |
| Climate/Clean Tech_P | −0.1949 | 0.484 | −0.403 | 0.688 |
| Pollution/Toxics_P | 0.9153 | 0.524 | 1.745 | 0.083 |
| Pollution/Toxics_R | −0.3628 | 0.297 | −1.221 | 0.224 |
| Pollution/Toxics_X | 0.3397 | 0.624 | 0.544 | 0.587 |
| Environment/Other_P | −0.6986 | 0.292 | −2.393 | 0.018 |
| Sustainable Investment - ESG Fund_Yes | 2.7682 | 0.208 | 13.289 | 0.000 |
| Sustainable Investment - Impact Fund_Yes | 0.2134 | 0.157 | 1.357 | 0.177 |
| Sustainable Investment - Environmental Sector Fund_Yes | 1.7756 | 0.286 | 6.199 | 0.000 |

Figure 1: Portfolio Environmental Score Regression No.1

the coefficient 2.7682, highest among other significant variables. Interestingly, variable "Environment/Other_P" which indicates the fund seeks investment with positive impact in the environment has a negative impact on the Portfolio Environmental Score because the coefficient is -0.6986. Moreover, some of the variables supposed to positively affect the environment score have insignificant coefficients with p-value larger than 0.6. With some counter-intuitive results in mind, we decide to further examine the collinearity among variables and make the regression results more precise and convincing. The correlation matrix is as follows.


Figure 2: Portfolio Environmental Score Correlation Matrix No.1

After repetitive comparison of variable pairs that have correlation magnitudes greater than 0.75 like "% Small Arms" and "% Alcohol" and regraphing of the correlation table, we eventually drop 5 variables with high correlations and keep the other. The finalized regression results still performs well with 0.950 R-squared and 299.7 F-statistic.

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| % Thermal Coal | 0.1196 | 0.047 | 2.537 | 0.012 |
| % Small Arms | 0.3800 | 0.186 | 2.043 | 0.043 |
| % Tobacco | 0.3962 | 0.669 | 0.592 | 0.555 |
| Fossil Fuels Restrictions | 0.1060 | 0.192 | 0.551 | 0.582 |
| Pollution/Toxics_R | −0.4162 | 0.276 | −1.506 | 0.134 |
| Pollution/Toxics_X | 0.2483 | 0.583 | 0.426 | 0.671 |
| Climate/Clean Tech_C | 0.3077 | 0.148 | 2.081 | 0.039 |
| Sustainable Investment - ESG Fund_Yes | 2.8406 | 0.163 | 17.427 | 0.000 |
| Sustainable Investment - Impact Fund_Yes | 0.0467 | 0.150 | 0.311 | 0.756 |
| Sustainable Investment - Environmental Sector Fund_Yes | 1.9219 | 0.267 | 7.197 | 0.000 |

Figure 3: Portfolio Environmental Score Regression No.2

In this case, "Sustainable Investment - ESG Fund_Yes" and "Sustainable Investment - Environmental Sector Fund_Yes" are highly significant with p-value 0, and both of the variables indicate a positive impact on the Portfolio Environmental Score, reasonably to conclude that the fund is claimed as ESG fund or also claimed as environmental sector fund is associated with increase in Portfolio Environmental Score. Furthermore, variables like "% Thermal Coal" and "Climate/Clean Tech_C" have p-values smaller than 0.05. These variables with significant coefficients will improve the accuracy of our random forest model later as valuable inputs. The improved correlation matrix is below with no highly correlated variables.

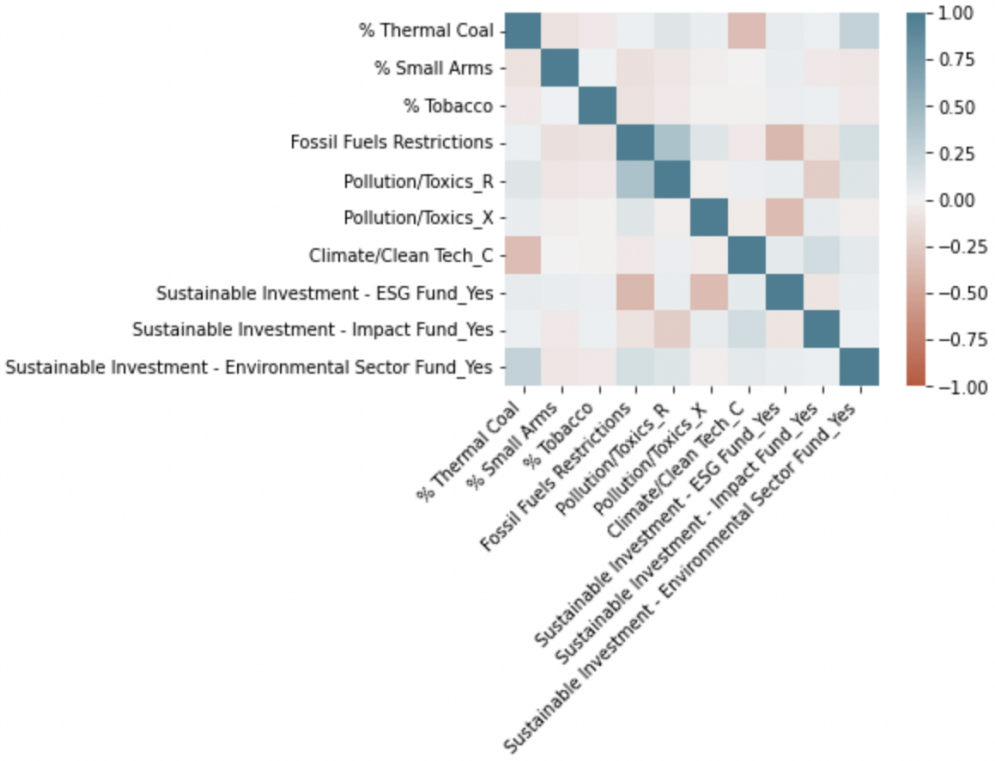


Figure 4: Portfolio Environmental Score Correlation Matrix No.2

## 4.2 Regression analysis (Portfolio Social Score)

Similarly, another regression on the target variable **Portfolio Social Score** is performed to investigate how investment strategies of the fund (whether the fund is involved in certain industry/field) contribute to the portfolio social score. Since we are focusing only on the social score, only socially related factors are included in the preliminary regression, which are "Community Development", "Human Right", "Diversity", "Labor Relation" and "Conflict Risks". The other two features are "ESG" and "Impact", showing if the fund is categorized as an ESG fund or Impact fund by Morningstar.

Joining the data from Bloomberg and the Moningstar Fund yields 180 overlapping funds. After dropping rows of funds without a social score, we acquire a total of 169 observations. To handle missing feature data, we use the most frequent value to fill in null values. Since all the variables are qualitative, we generate dummy variables for each category and created a total number of 11 variables. The model performs well with 0.932 R-squared value and 195.7 F-statistics, which shows the validity of the regression in fitting with the observed data.

The result of the preliminary linear regression is shown below:

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| social_comm_develop_P | 4.4090 | 1.063 | 4.147 | 0.000 |
| social_diversity_P | 4.2038 | 1.368 | 3.073 | 0.002 |
| social_human_rights_P | -4.6239 | 1.753 | -2.638 | 0.009 |
| social_human_rights_R | -0.5519 | 1.119 | -0.493 | 0.623 |
| social_labor_P | -3.8759 | 1.506 | -2.574 | 0.011 |
| social_labor_R | -3.0153 | 2.126 | -1.418 | 0.158 |
| social_conflict_P | 1.8358 | 2.398 | 0.765 | 0.445 |
| social_conflict_R | 0.3704 | 0.445 | 0.832 | 0.407 |
| social_conflict_X | 1.0462 | 0.795 | 1.316 | 0.190 |
| ESG_Yes | 7.3011 | 0.479 | 15.251 | 0.000 |
| Impact_Yes | 0.3805 | 0.503 | 0.756 | 0.451 |

Figure 5: Portfolio Social Score Regression No.1

As for the significance of variables, variables including "social_comm_develop_P", "social_diversity_P" and "social_human_rights_P", "social_labor_P" and "ESG_Yes" are have statistically significant coefficients with p-value close to 0. Since the notion P indicates a positive engagement in certain social field of the fund, for the categories "social_comm_develop" and "social_diversity" and "ESG_yes", the positive coefficients indicate a positive effect on the social score.

A correlation matrix is generated to show the relationship between variables. As seen from the results, several factors such as "social_com_develop_P", "social_diversity_P", "social_human_rights_P" are highly correlated with a correlation coefficient greater than 0.75. Therefore for each pair one of the variables are removed and the regression is performed again.

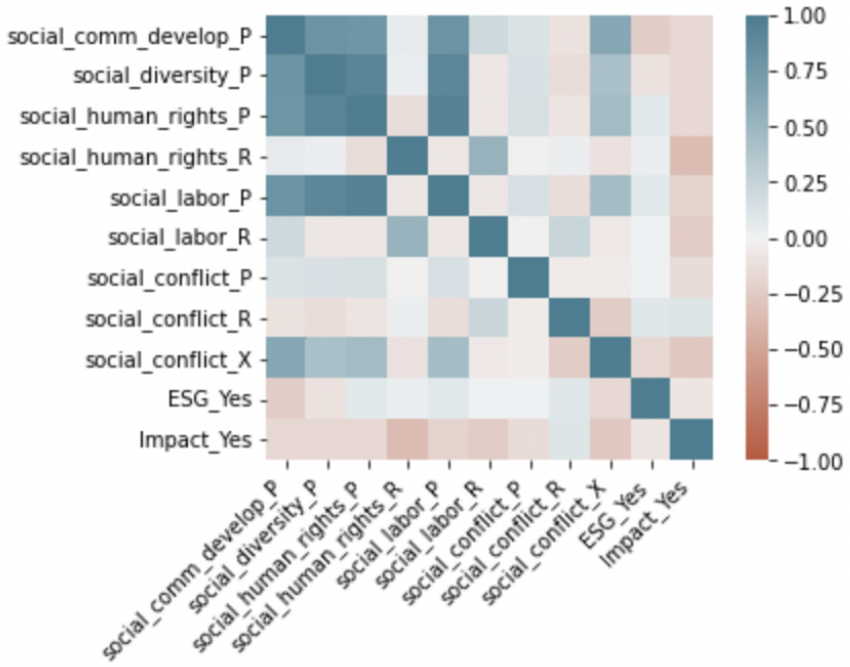


Figure 6: Portfolio Social Score Correlation Matrix No.1

The result of the second regression is shown below. The R-squared has slightly decreased, which is reasonable with fewer number of variables. In terms of p-value, "Sustainable Investment - ESG Fund_Yes" and "Sustainable Investment - Environmental Sector Fund_Yes" are statistically significant with a p-value of almost 0. Other statistically significant factors are "social_human_rights_R" and "social_conflict_X". Since all coefficients in the updated regression are positive, we can conclude that most positively influence the Portfolio Social Score is the fund being in the ESG category, with coefficient almost reaching to 6. On the other hand, the fund seeks investment of the negative impact on labor has the least effect in Portfolio Social Score, with coefficient only 0.1.

| | coef | std err | t | P>\|t\| |
|---|---|---|---|---|
| social_diversity_P | 0.2611 | 0.520 | 0.502 | 0.616 |
| social_human_rights_R | 2.1337 | 1.047 | 2.039 | 0.043 |
| social_labor_R | 0.1000 | 1.857 | 0.054 | 0.957 |
| social_conflict_P | 3.0636 | 2.643 | 1.159 | 0.248 |
| social_conflict_R | 0.3544 | 0.494 | 0.718 | 0.474 |
| social_conflict_X | 2.8801 | 0.652 | 4.416 | 0.000 |
| ESG_Yes | 5.9252 | 0.478 | 12.390 | 0.000 |
| Impact_Yes | 1.9771 | 0.486 | 4.064 | 0.000 |

Figure 7: Portfolio Social Score Regression No.2

# Part 2 – Practicability

## 5 New Model Background

With the above two regressions which have high $R^2$ and significant coefficients, we can conclude with confidence that evaluating a fund's performance on making positive social and environmental impact is possible with the information known to the public and funds' managers. In another word, rather than underestimating the complex rating system developed by Morningstar and collecting numerous data, it is possible to build a model that helps investors estimate the sustainability-related consequences with existing accessible data. Furthermore, although it should be preferred to take consideration of social and environmental issues when investing, profitability is still the ultimate goal of investing and thus should not be neglected. There, paying attention on both aspects is an important task for investors and demands careful analysis.

Knowing that the Morningstar Portfolio Sustainability Rating is not a merit or performance based evaluation, our team decide to construct a new comprehensive index, called Profitability and Sustainability Score, to guide investors' decision making process. The Profitability and Sustainability Score, taking into account both the financial performance and the social and environmental impact of a fund, would allow investors to systematically compare funds with both wealth growth and sustainability in mind. Additionally, although the score is constructed based on the Morningstar ratings during the training stage, once high accuracy of the model is achieved with selected inputs variables, new investment opportunities would obtain their Profitability and Sustainability Scores by providing values of these input variables rather than waiting for their Morningstar rating. Thus, investors receive a comprehensive rating of their prospecting investments with accessible data.

## 6 Algorithm

The new score includes four components, 30% Portfolio Environmental Score, 30% Portfolio Social Score, 25% 5 Years Annualized Return, and 15% Credit Quality. To construct the score, we first drop the rows with null values in Portfolio Environmental Score and Portfolio Social Score which leaves us with a total of 5993 rows in the dataset. Then, we normalize the numerical values in Portfolio Environmental Score, Portfolio Social Score and 5 Years Annualized Return respectively to a 0 to 1 range by the Min-Max normalization formula z = (x - min(x)) / (max(x) - min(x)). For Average Credit Quality that has 6 categories: B(lowest), BB, BBB, A, AA, AAA(highest), we encode them from 0 to 5 accordingly. We times 30 with the new Portfolio Environmental Score, 30 with the new Portfolio Social Score, 25 with the new 5 Years Annualized Return, and 3 with the new Credit Quality classes conforming to the weights we define at first. Adding all of these numbers together and normalizing using the previous formula and time each score with 100, we get the final scores so that they are ranged from a fixed range of 0-100, which is also the Y variable in our machine learning model for prediction.

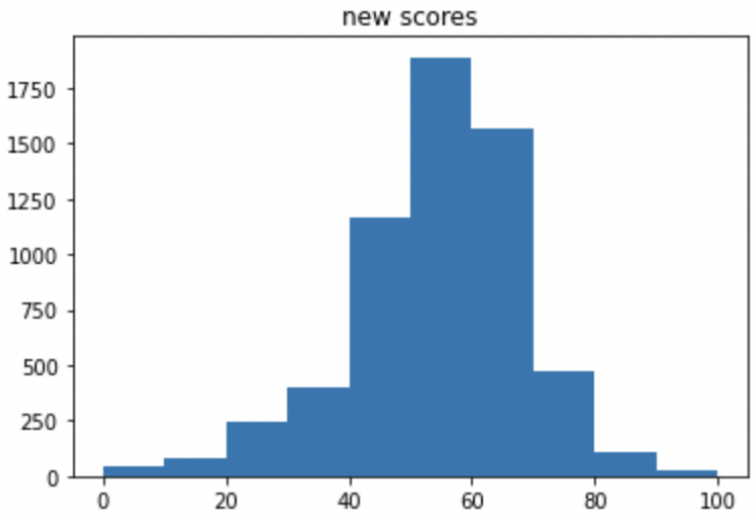


Figure 8: New Score Distribution

We decide to use columns with information that is easily gained by the users to train the model. As is shown in the image below, there are missing values in Yield (%), 10 Years Annualized (%) and Sustainable Investment - Environmental Sector Fund. For Yield we simply fill with the median, and for the Environmental Sector Fund we fill with the mode.

Surprisingly, the null of 10 Years Annualized rows have consecutive indexes and similar names so we go back to the original dataframe to find some insights. We check the 64 rows with names containing Fidelity Advisor Freedom and their 1 Year Annualized, 3 Years Annualized, 5 Years Annualized and 10 Years Annualized return. As these values are quite close among the rows, we fill the null values with the mean of other non-null 10 Years Annualized returns.

Overall, the variables we include in the Random Forest model are: Yield (%), Sustainable Investment - ESG Fund, Sustainable Investment - Impact Fund, Sustainable Investment - Environmental Sector Fund, Animal Testing, % Alcohol, % Fossil Fuels, % Small Arms, % Thermal Coal, % Tobacco, YTD Return (%), 10 Years Annualized (%), Fund Size (Mil), Average Market Cap (Mil).

With all the variables handled, we split our data into a training set (80%) and a test set (20%) by the "train_test_split" method and fit the training data in Random Forest Regressor since its robustness towards outliers and towards non-linear relationships among variables. For faster runtime, we use Randomized Search Cross Validation instead of GridSearchCV for hyperparameters tuning.

# 7 Conclusion

When applying the Randomized Search CV function, we acquire the hyperparameters with 300 numbers of estimators, 3 maximum depth, and using bootstrap samples as the best solution for building the model. After applying those parameters, the resulting Random Forest Regressor model generates the test accuracy of 0.91, which proves that our model is quite optimal. We then evaluate the importance of each feature in the dataset that we used to perform the model. According to the graph below, finance-related variables such as "Average Market Cap (Mil)", "10 Years Annualized (%)" and "YTD Return" are important features. While interestingly, sustainability or environment-related features do not show much significance in determining the score of funds. To test this model, we create a virtual fund

example by entering the above features to predict the fund's score, and it turns out our model performs well providing a reasonable score. For example, we make up a fund's data based on the distribution of sample data, with 3.88% Yield, being an ESG, Impact and Environmental Sector Fund, with Animal Testing, 0% Alcohol, 0% Fossil Fuels, 0% Small Arms, 0% Thermal Coal, 1% Tobacco, 11.98% YTD Return, 12.57% 10 Years Annualized, 200.59 Million Fund Size, and 35000.37 Million Average Market Cap. The score predicted by Random Forest Regressor is 50.32, which within expectation.

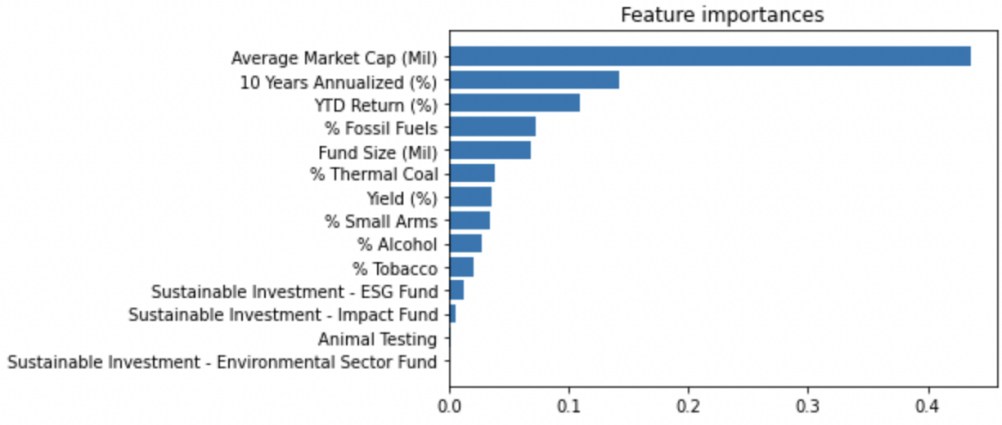


Figure 9: Feature Importance for Random Forest Regressor

In order to improve our current model, we decide to drop variables that do not contribute much, in this case, with feature importance lower than 0.03. Therefore, variables below "Yield (%)" were removed. After rerunning the model using variables considered with higher importance values, the model generates 0.9 test accuracy, which has similar test accuracy as before. Also, the example fund removing related variables receives the score of 49, also similar to the previous model.

To sum up, from the model we created above, we conclude that "Average Market Cap (Mil)", "10 Years Annualized (%)", "YTD Return (%)", "Fund Size (Mil)", and "Yield (%)" are important financial factors determining the fund score. Interestingly, for the environmental factors, "% Fossil Fuels" and "% Thermal Coal" contribute more to the fund score. We believe our model would provide directions for choosing optimal investments.

## References

https://charts.ussif.org/mfpc/