SEEKING SIGNALS FROM ESG DATA

Liwen Ouyang
Bloomberg Enterprise Quants
{louyang11, entquant}@bloomberg.net

Thursday 16th April, 2020

ABSTRACT

More and more attention has been paid in recent years to Environmental, Social, and Governance (ESG) issues at companies and to the data those companies release that seek to quantify those issues. Along with that attention has come a debate about whether ESG data benefits investors in evaluating companies for their portfolios. In this paper, we show that applying Gradient Boosting Trees to Bloomberg's own ESG dataset allows us to create an equity portfolio with higher return and lower volatility than its benchmark Russell 3000 Index. We also investigate the interpretability of our model using SHapley Additive exPlanations and compare the results to a traditional Logistic Regression-based approach.

1 Introduction

In recent years, investors have paid increasing attention to environmental, social, and governance (ESG) issues at their portfolio companies. This rise in ESG awareness has been driven by three factors. First, consumers increasingly favor choices that they see as more sustainable, healthier, and smarter. This is a general shift in consumer preferences and is reflected in consumers' investment decisions by a preference for companies with positive ESG profiles. Second, financial researchers have found evidence that a company's strength in ESG criteria is associated with positive financial impacts in the long term, making those companies attractive investments. Third, the availability of more complete ESG data makes it possible to do more research with it, thus creating more robust and data-driven investment options based on ESG factors.

This paper will focus on the relationship between ESG data and financial impact. Specifically, this paper will discuss the use of advanced machine learning techniques applied to Bloomberg ESG data to create an equity portfolio with higher return and lower volatility than its benchmark. There is existing evidence that such an approach can work. A study (Barnett and Salomon, 2006) showed that there is a curvilinear relationship between social responsibility and financial performance. A more recent study (Eccles et al., 2014) showed that companies with high sustainability had better investment performance than their low sustainability counterparts.

We demonstrate that by applying advanced machine learning techniques to Bloomberg's ESG data, we are able to create a portfolio with superior risk and return characteristics.

2 ESG Data Overview

Bloomberg provides annual ESG data for about 13,000 unique companies globally as far back as 2006 and daily-level governance data for a subset of nearly 4,500 companies back to 2013. For this study, we focus on annual ESG data for US companies as US companies have relatively more complete ESG data. There are about 4,000 unique US companies in Bloomberg's historical annual ESG data, and about 3,100 of these have associated stocks that trade on US exchanges. The study focuses on stocks' performance from 2006 through mid 2018, including the 2008 financial crisis period.

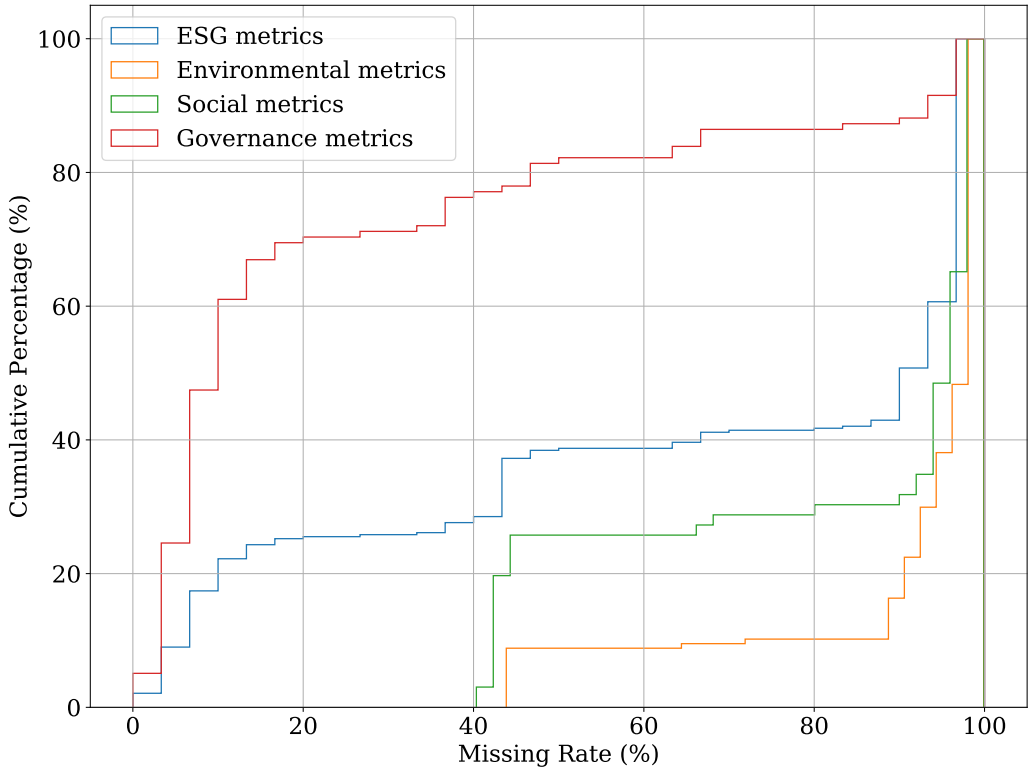


Figure 1: Cumulative step histograms of missing rate for all ESG metrics as well as separate environmental, social, governance metrics.

Of the more than 300 ESG metrics, 147 fall under the Environmental category, 66 fall under Social, and 118 fall under Governance. Environmental metrics include carbon emissions and resource and energy use. Social metrics include human rights and diversity and inclusion. Governance metrics include criteria based on management structure, executive compensation, and employee relations.

Bloomberg's historical ESG data does not provide the disclosure date of annual ESG data, but does provide a date corresponding to the end of the period for the latest available ESG data. We made a relatively conservative assumption that complete annual ESG metrics about a company would be available one year after the current ESG period end date for that company.

One challenge with ESG data is that there are a lot of missing values. This is due to two reasons. First, most ESG metrics are self-reported and companies are not under any obligation to report them. Second, some specific ESG metrics may apply strongly to certain industries but not at all to others. Overall, about 60% of ESG metrics have a Missing Rate (defined as the proportion of records for which the value is missing) of more than 50% in Bloomberg's ESG dataset on US companies over the past 14 years. Figure 1 shows a visualization of the distribution of Missing Rates for ESG fields.

3 Strategy Empowered by Advanced Machine Learning Models

3.1 Three-Class Classification on Annual Excess Return

Since ESG-focused investing is seen as generally focused on long-term performance, our approach was to tie ESG metrics to long-term returns directly by building a machine learning model that uses Bloomberg's ESG data to categorize the annual excess return of each stock in the dataset. The annual excess returns were measured relative to the Russell 3000 index and were bucketed into three categories: greater than 15%, less than -25%, and in between. These thresholds were chosen empirically to make the resulting positive and negative classes relatively balanced.

3.2 Gradient Boosting Trees

We built a gradient boosting tree (GBT) model to predict the category that a company's excess return would fall into based on its ESG metrics. GBT models can inherently handle missing values quite well, without the need for special treatment.

We used an implementation of the well-known xgboost library, with a learning rate of 0.1, a max depth of 5, a min child weight of 5, an L1 regularization of 80, subsampling and column sampling by tree both set to 0.8, and 200 total boost rounds.

3.3 The Training Process

We tested our model over 4 moving windows. The first window began in January 2007 and contained 7.5 years of training data (through mid-2014). The model calibrated on that data was then tested on a 1-year period from mid-2014 through mid-2015. We then rolled all of those dates forward 1 year, re-calibrated the model, and derived new results from the new training period. A summary of the training and testing periods is given in Table 1 below.

Episode Number	Train Set Start Date	Train Set End Date	Test Set Start Date	Test Set End Date
1	2007-01-03	2014-06-04	2014-06-30	2015-06-02
2	2008-01-02	2015-06-02	2015-06-29	2016-06-07
3	2008-12-31	2016-06-07	2016-06-29	2017-06-07
4	2009-12-30	2017-06-07	2017-07-03	2018-06-07

Table 1: Train and test period for each episode.

3.4 Portfolio Construction and Preliminary Results

In the testing stage, our model evaluated each stock as soon as new ESG data became available for it (at a daily granularity). If our model predicted that a stock would be in the high annual excess return bucket, we considered that stock a suitable long investment. If our model predicted that a stock would be in the low (negative) annual excess return bucket, we considered it a short. However, we did not short stocks whose prices were already low. Furthermore, since our model was predicated on annual excess return, whenever we entered a position, we held it for one year and closed it after that. So even though we could go long or short a stock on any given day, we would hold the position for a year.

All of our long picks were incorporated into an equal-weighted long portfolio, while all of our short picks were incorporated into an equal-weighted short portfolio. We considered the overall portfolio return to be the mean of the long and short portfolio returns.

Assuming a 10bps cost per trade, we obtained an annualized Sharpe Ratio of 1.25 for our long-short ESG portfolio, compared to 0.73 for the Russell 3000 Index over the same testing period (see Table 3 for full the results). For the sake of simplicity, our Sharpe Ratio calculations use a risk-free rate of zero throughout this paper. Since we also report the return and volatility, an alternative Sharpe Ratio could be computed for any desired risk-free rate.

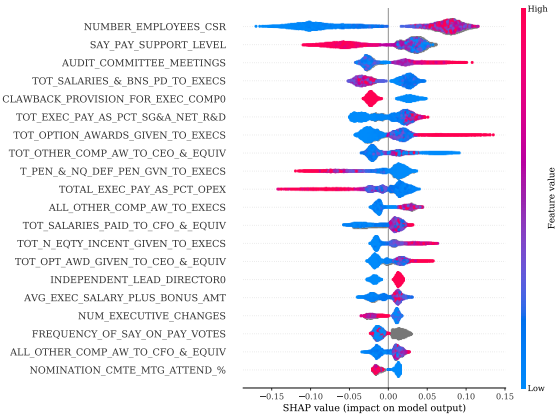
4 From Blackbox Models to Interpretable Models

4.1 Feature Importance with SHAP

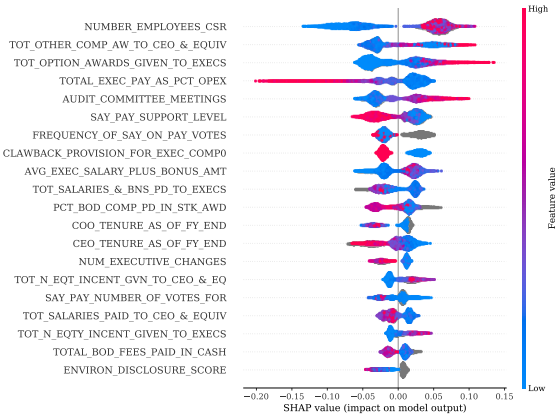
A basic ability of a gradient boosting tree model is assigning an "importance" to every input feature based on how often and how early that feature is chosen to improve the model prediction. Unfortunately, this is not good enough to provide a desirable level of model interpretability. In finance especially, it is desirable to know how a feature affects the final output, and at the very least, the direction (positive or negative) in which it does so. Borrowing the idea from game theory, SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) can help explain the output of blackbox machine learning models. For this project, we obtained SHAP values using a TreeExplainer (Lundberg et al., 2020) designed specifically for tree-based models.

In Figure 2 and Figure 3, we present the SHAP values for the most important features in the dataset. Since our model is a three-class classification, each feature has three SHAP values (one per class). Figure 2 shows the SHAP values for the Long class (predicted to have greater than 15% annual excess return), while Figure 3 shows the SHAP values for the Short class (predicted to have lower than -25% annual excess return).

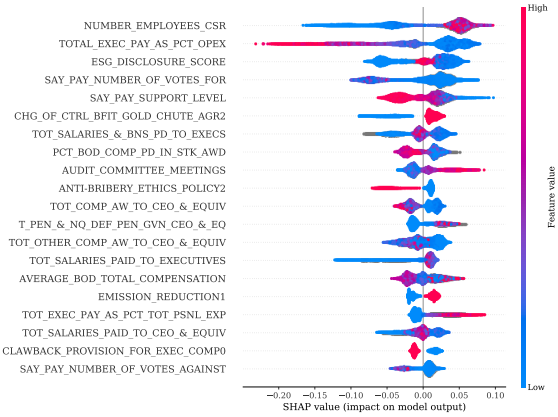
Each figure shows the SHAP values over four different training periods. The thickness of the horizontal line represents the data density: a "bulge" in the line indicates that a lot of that variable's data fell in that range. The color of the line represents the value of the feature, with redder colors indicated higher-value features and bluer colors indicated lower-value features. Finally, the horizontal length of a line indicates the size of the feature's impact. Therefore, an "ideal" feature would be visualized on a SHAP plot by a long horizontal line (meaning a large positive/negative impact) and with relatively clear separation between blue and red regions (meaning that high and low values in this feature will drive different behavior).



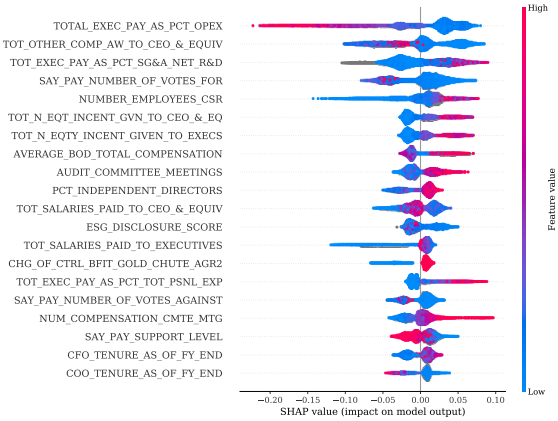
(a) train period 1



(b) train period 2



(c) train period 3

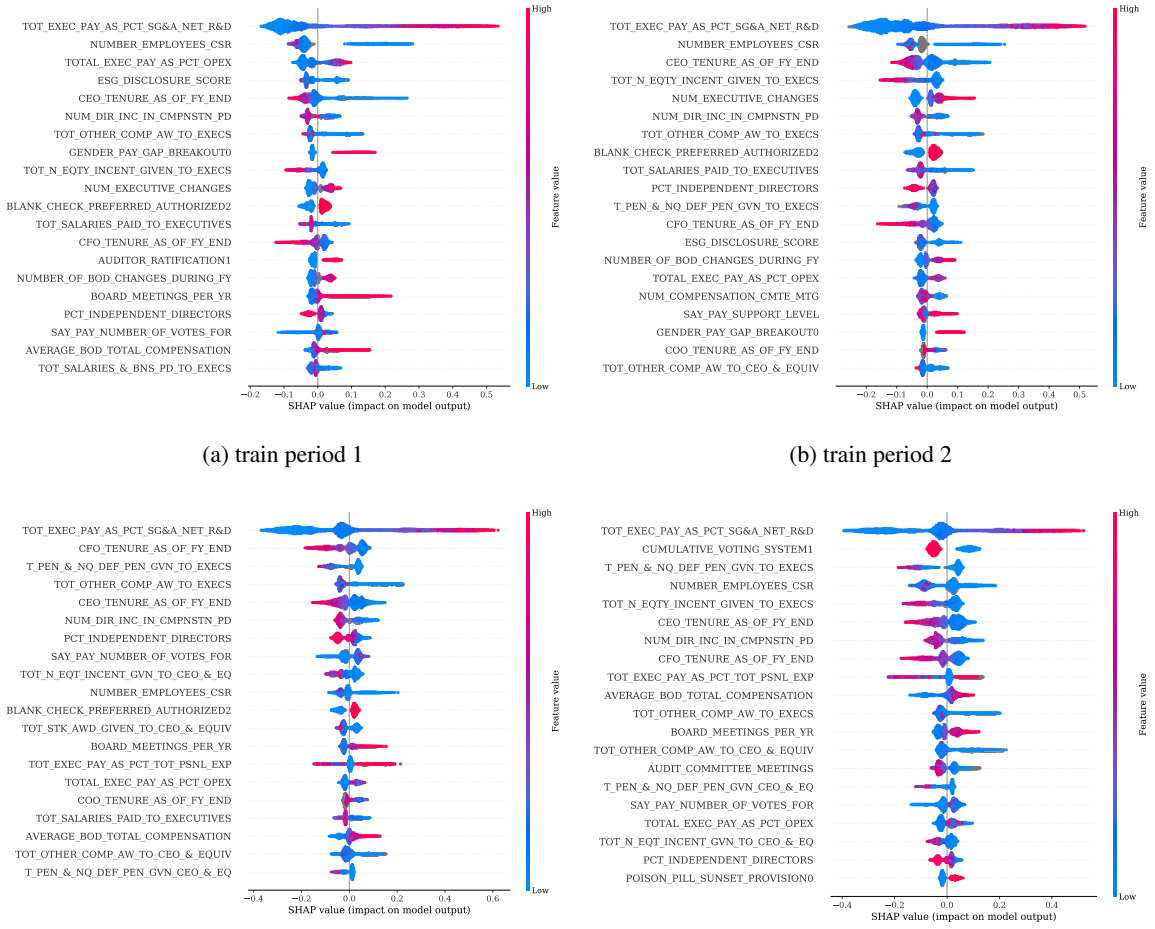


(d) train period 4

Figure 2: Top features' SHAP value for predicting to have more than 15% annual excess return.

4.2 Interpreting Feature Importance

We used the output of these SHAP plots to refine our feature selection based on interpretability criteria. On the Long position prediction side, taking Figure 2a as an example, we see that **higher** option awards given to executives (TOT_OPTION_AWARDS_GIVEN_TO_EXECS) **increase** the likelihood of a company's annual excess return being greater than 15%, but that **higher** total executive pay as a percent of operating expense (TOTAL_EXEC_PAY_AS_PCT_OPEX) **decrease** the likelihood a company's annual excess return being greater than 15%. Likewise, when a company has **no** compensation clawback provisions (CLAWBACK_PROVISION_FOR_EXEC_COMP0), the company's stock tends to perform **worse**. Similarly, having **more** meetings of the board's audit committee held during the year (AUDIT_COMMITTEE_MEETINGS)



(a) train period 1

(b) train period 2

(c) train period 3

(d) train period 4

Figure 3: Top features' SHAP value for predicting to have less than -25% annual excess return.

has a **positive** impact on a company's annual excess return, while having **higher** support level for the shareholder approval of executive pay (SAY_PAY_SUPPORT_LEVEL) has a **negative** impact on company's annual excess return.

On the short prediction side, taking Figure 3a as an example, **more** total executive pay as a percentage of selling, general, and administrative expenses net of research and development costs (TOT_EXEC_PAY_AS_PCT_SG&A_NET_R&D) has a **negative** impact on a company's annual excess return. Likewise, having **no** quantitative gender pay gap breakout in public filings (GENDER_PAY_GAP_BREAKOUT0) and being **authorized** to issue blank check preferred stock without shareholder approval (BLANK_CHECK_PREFERRED_AUTHORIZED2) both have **negative** impacts on a company's annual excess return. On the other hand, **longer** CFO tenure (CFO_TENURE_AS_OF_FY_END) has **positive** impacts on the company's annual excess return by decreasing the likelihood of that return being in the strongly negative category.

However, there are cases in which a field's impact cannot be clearly interpreted. For instance, in Figure 2d, a **longer** COO tenure (COO_TENURE_AS_OF_FY_END) has a **negative** impact on a company's annual excess return. On the other hand, in Figure 2b, there is no clear separation between high and low values in COO tenure. Since this field had no consistent pattern, it was left out of the final list of important fields.

We performed this kind of analysis on all of the top features across the four training periods, as presented in Figure 2 and Figure 3. We then picked out the common top features that had relatively clear and consistent impacts on the annual excess return prediction. The final top features are listed below in Table 2.

Features	Definiton
NUMBER_EMPLOYEES_CSR	Total number of company employees
AUDIT_COMMITTEE_MEETINGS	Number of meetings of the Board's Audit Committee
SAY_PAY_SUPPORT_LEVEL	Support level for the shareholder approval of executive pay
TOT_OTHER_COMP_AW_TO_CEO_&_EQUIV	Aggregated amount of other compensation paid to the CEO or the equivalent
TOTAL_EXEC_PAY_AS_PCT_OPEX	Total executive compensation percentage of operating expenses
TOT_SALARIES_PAID_TO_CEO_&_EQUIV	Total amount of salary paid to the CEO or the equivalent
TOT_SALARIES_&_BNS_PD_TO_EXECS	Salary and bonus amount paid to the executives
TOT_N_EQTY_INCENT_GIVEN_TO_EXECS	Total amount of non-equity incentives awarded to the executives
SAY_PAY_NUMBER_OF_VOTES_FOR	Number of shareholder votes cast 'For' the approval of executive compensation
TOT_EXEC_PAY_AS_PCT_SG&A_NET_R&D	Total executive compensation percentage of sales, general and administrative expenses net of Research and Development
TOT_OPTION_AWARDS_GIVEN_TO_EXECS	Total amount of options awarded to the executives
TOT_EXEC_PAY_AS_PCT_TOT_PSNL_EXP	Total executive pay percentage of total personnel expense
TOT_N_EQT_INCENT_GVN_TO_CEO_&_EQ	Total amount of non-equity incentives the company awarded to the CEO or the equivalent
PCT_BOD_COMP_PD_IN_STK_AWD	Stock awards given to directors compared to total director compensation as a percentage
NUM_EXECUTIVE_CHANGES	Number of executives changes during fiscal year
AVERAGE_BOD_TOTAL_COMPENSATION	Average total director compensation paid to directors
ESG_DISCLOSURE_SCORE	Proprietary Bloomberg score based on the extent of a company's ESG disclosure
CFO_TENURE_AS_OF_FY_END	Chief Financial Officer and equivalent tenure as of fiscal year end
CHG_OF_CTRL_BFIT_GOLD_CHUTE_AGR	Indicates whether the company has any change of control benefits/severance benefits/golden parachute agreements in place for any of its executives
CLAWBACK_PROVISION_FOR_EXEC_COMP	Indicates whether the company has any compensation clawback provisions in place
GENDER_PAY_GAP_BREAKOUT	Indicates whether the company has quantitative gender pay gap breakout in public filings and/or publicly available company official sources
BLANK_CHECK_PREFERRED_AUTHORIZED	Indicates whether the company is authorized to issue blank check preferred stock without shareholders' approval

Table 2: Top features selected from SHAP.

4.3 Retraining with Top Features

Once we had a list of top features based on the insights from the SHAP plots, we retrained our GBT model using only the top features. Interestingly, by reducing the available information to just the top features, we obtained a model that yielded a superior portfolio. The new portfolio had a Sharpe Ratio of 1.56, compared with 1.25 for the all-feature portfolio and 0.73 for the benchmark. We believe that by excluding the noise in the unimportant features, our model was better able to "focus" on the meaningful variations in the important features.

Note that we always used the same set of top features across all training windows. This served as a way to prevent the model from overfitting the features to any particular timeframe. One might also pick different feature sets in different training periods to further tailor the model to be more specific to the economic and market properties of the timeframe. In theory, this could produce even better performance. However, in this case one needs to pay extra attention to make sure that the top features picked are not overfitted to the training data or to any one specific set of validation data.

4.4 A Fully Interpretable Model

For the interest of clients who prefer to have a fully interpretable model, we also crafted a more traditional strategy using logistic regression (LR) on the top features in Table 2. In order to allow the logistic regression model to capture non-linear relationships in the data, we added second-degree polynomials of the underlying features.

Note that logistic regression models cannot handle missing values, so we had to fill in any missing values beforehand. We tested two variants: filling in missing values with the average value of the feature and imputing the missing values by using an additional model based on a variational auto-encoder (VAE). Missing-value imputation using a VAE model is a rich area of research independently of its application here. Discussing the VAE imputation model in detail is beyond the scope here and is the topic of a separate paper from our team (Gopal, 2020).

The portfolio constructed by the LR model with missing values filled in by averages performed poorly, with both return and Sharpe Ratio worse than those of the index. By using the VAE model to impute missing features, however, the LR model was able to beat the Sharpe Ratio of the index while still lagging in return. More details can be found in section 5.

5 Discussion of Results

5.1 Results Summary

Table 3 contains summary statistics of the portfolios generated by the different models we investigated. All four portfolios constructed from ESG data had lower volatilities compared to the Russell 3000 Index. The index performed quite well in the test period already, but the two portfolios constructed from the GBT models beat this benchmark in annualized return and Sharpe Ratio as well as volatility. It is also clear from the table that our VAE imputation led to superior results for the LR-based models as compared to mean-value imputation.

Model behind Portfolio	Return	Volatility	Sharpe Ratio
GBT (All Features)	9.9%	7.9%	1.25
GBT (Top Features)	11.8%	7.5%	1.56
LR (Impute Mean)	4.4%	8.5%	0.51
LR (Impute VAE)	5.4%	6.5%	0.82
Russell 3000 Index	9.4%	12.9%	0.73

Table 3: Annualized return, volatility and sharpe ratio during the test period (2014/07 to 2018/07).

5.2 Improving the Efficient Frontier

Furthermore, we checked how the efficient frontier for a stock and bond investor can be improved by the addition of our ESG portfolios. For a benchmark, we used a combination of the Russell 3000 Index to represent equity returns and the Bloomberg Barclays US Aggregate Bond Index (USAGG USD) to represent bonds. The results are shown in Figure 4 below. Note that when either ESG portfolio is available, the efficient frontier is pushed to the upper-left corner to various degrees. The best efficient frontier is given by adding the portfolio from the GBT model trained on the top features. It is interesting that even the worst-performing portfolio improves the efficient frontier: this is evidence for the orthogonality of ESG factors to the broad market. We decided to investigate this further.

5.3 Backtesting with a Fama/French Five-Factor Adjustment

In order to verify that our ESG portfolios cannot be explained by common factors such as market return (Mkt-RF), size (SMB), value (HML), profitability (RMW) and investment (CMA), we attempted to attribute the returns of our ESG portfolios to the Fama/French five-factor model (Fama and French, 2015). Table 4, given below, shows the regression summary for the four portfolios.

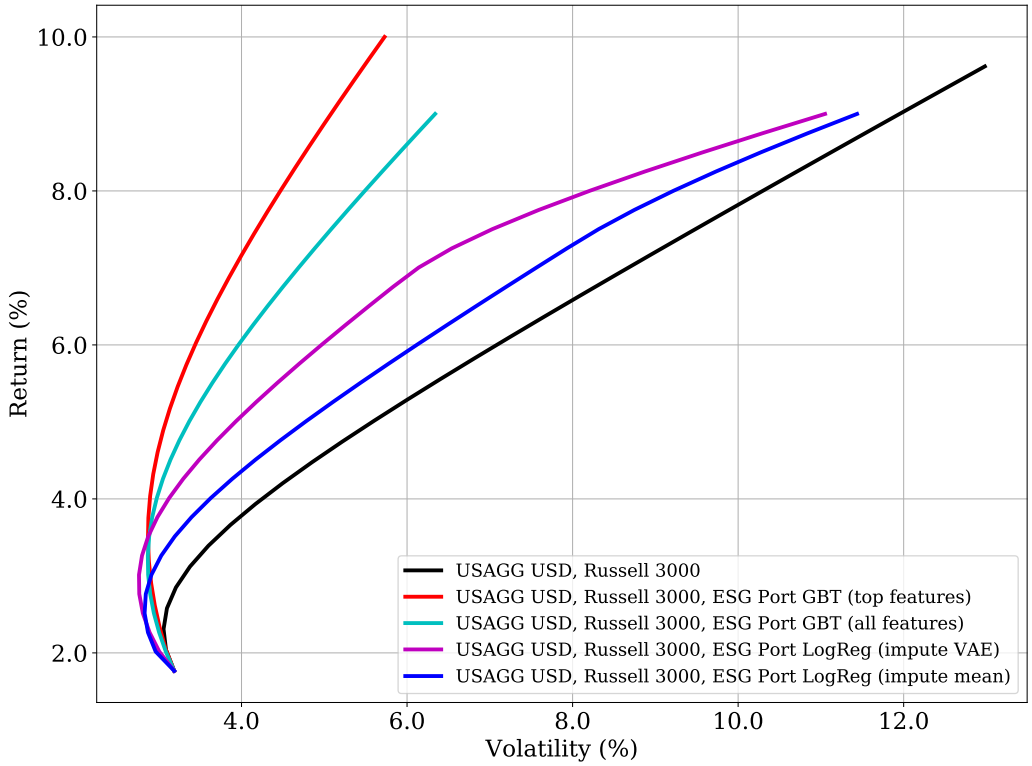


Figure 4: Efficient frontiers formed by different assets.

Model behind Portfolio	**Mkt-RF**	**SMB**	**HML**	**RMW**	**CMA**	α	R^2
GBT (All Features)	0.0184 (0.363)	0.0381 (0.239)	-0.0035 (0.931)	0.1533 (0.002)	-0.0611 (0.339)	8.6940 (0.028)	0.011
GBT (Top Features)	0.0076 (0.693)	-0.0306 (0.317)	0.0819 (0.031)	0.2113 (0.000)	-0.0699 (0.246)	10.9116 (0.004)	0.028
LR (Impute Mean)	0.0135 (0.529)	-0.1226 (0.000)	0.0761 (0.074)	0.2062 (0.000)	0.0006 (0.993)	3.7800 (0.368)	0.041
LR (Impute VAE)	-0.0006 (0.971)	-0.1375 (0.000)	0.1136 (0.000)	0.2708 (0.000)	-0.0081 (0.871)	4.9392 (0.110)	0.114

Table 4: Regression summary for ESG portfolios with Fama/French Five-Factor Model. The second to seventh columns show the estimated coefficients and their corresponding p-value.

Consistent with the pre-adjustment results in Table 3, portfolios based on the GBT models had higher and more significant alpha after attributing to the Fama/French five factors. Moreover, as Table 4 shows, the ESG-based portfolios are not strongly influenced by the Fama/French factors. For the best model (GBT with top features), only 2 factors were statistically significant (HML and RMW) and that at a low magnitude. Based on these results, we plotted the cumulative return for these four portfolios before and after the Fama/French five-factor adjustments in Figure 5, shown below.

5.4 Change of Benchmark

We also explored the results when the S&P 500 is used as a benchmark instead of the Russell 3000. The top features and model performances changed slightly, but the overall conclusions remained the same. We provide detailed results for the S&P 500 as a benchmark in Appendix A.

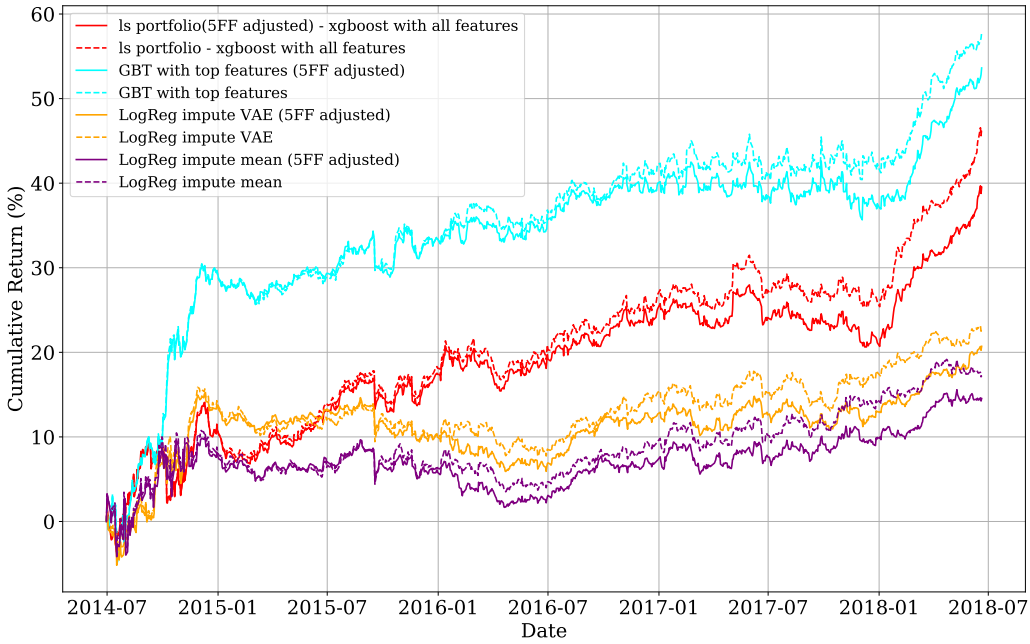


Figure 5: Cumulative returns for four portfolios before and after Fama/French five-factor adjustments during the test period.

5.5 Long Only vs Long-Short

We also compared the performance of a long-only strategy instead of a long-short strategy. As expected, the long-only portfolios tended to be more volatile and more correlated with common factors like market performance, size, etc. Please refer to Appendix B for a more detailed discussion.

6 Conclusion

We were able to beat the index on return, volatility, and Sharpe Ratio by using Bloomberg's ESG data combined with machine learning techniques. Although ESG data suffers from missing values, GBT and VAE techniques mitigated this problem. The GBT models specifically were able to achieve both higher return and lower volatility than the index. Guided by an interpretability analysis using SHAP plots, a more traditional LR-based approach that used VAE techniques to fill in missing values resulted in a completely explainable model. This model sacrificed some of the return and volatility advantages of the GBT-based portfolios but nevertheless maintained positive alpha and improved the efficient frontier.

Acknowledgements

We would like to thank Aaron Key, Ruslan Tepelyan and Achintya Gopal for their fruitful insights and useful comments.

References

Barnett, M. L. and Salomon, R. M. (2006). Beyond dichotomy: the curvilinear relationship between social responsibility and financial performance. Strategic Management Journal *27*, 1101–1122.

Eccles, R. G., Ioannou, I. and Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science *60*, 2835–2857.

Gopal, A. (2020). ESG Imputation Using DLVMs. `https://www.bloomberg.com/professional/blog/imputation-of-missing-esg-data-using-deep-latent-variable-models/`.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence *2*, 2522–5839.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30, (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds), pp. 4765–4774. Curran Associates, Inc.

A Using S&P 500 as a Benchmark

We tried S&P 500 as a benchmark instead of Russell 3000, and used exactly same approach otherwise. The top features picked based on SHAP value are listed below in Table 5.

Features	Definitions
NUMBER_EMPLOYEES_CSR	Total number of company employees
TOT_OPTION_AWARDS_GIVEN_TO_EXECS	Total amount of options awarded to the executives
TOTAL_EXEC_PAY_AS_PCT_OPEX	Total executive compensation percentage of operating expenses
TOT_OTHER_COMP_AW_TO_CEO_&_EQUIV	Aggregated amount of other compensation paid to the CEO or the equivalent
AUDIT_COMMITTEE_MEETINGS	Number of meetings of the Board's Audit Committee
SAY_PAY_SUPPORT_LEVEL	Support level for the shareholder approval of executive pay
TOT_EXEC_PAY_AS_PCT_SG&A_NET_R&D	Total executive compensation percentage of sales, general and administrative expenses net of Research and Development
TOT_OPT_AWD_GIVEN_TO_CEO_&_EQUIV	Total amount of options the company awarded to the CEO or the equivalent
ESG_DISCLOSURE_SCORE	Proprietary Bloomberg score based on the extent of a company's ESG disclosure
TOT_EXEC_PAY_AS_PCT_TOT_PSNL_EXP	Total executive pay percentage of total personnel expense
NUM_EXECUTIVE_CHANGES	Number of executives changes during fiscal year
CLAWBACK_PROVISION_FOR_EXEC_COMP	Indicates whether the company has any compensation clawback provisions in place
PRESIDING_DIRECTOR	Indicates whether the company has a presiding director in its board of directors

Table 5: Top features selected from SHAP when using SP 500 as benchmark.

As we can see, actually most of the top features are the same as before, with only a few differences. The performances for four portfolios are summarized in Table 6. Again, all four portfolios have lower volatilities compared to the benchmark index, and both GBT-based portfolios can beat the benchmark's returns, and the VAE-empowered LR model can beat the benchmark's Sharpe Ratio. One interesting wrinkle here is that we found that using imputed values from the VAE model in the top-features GBT model actually yielded improved performance compared to relying on the GBT structure to handle the missing values. So in this case, the results reported here for the GBT with top features actually use imputed values from our VAE model.

Model behind Portfolio	Return	Volatility	Sharpe Ratio
GBT (All Features)	13.1%	9.6%	1.37
GBT (Top Features)	13.8%	9.1%	1.51
LR (Impute Mean)	5.7%	8.3%	0.68
LR (Impute VAE)	7.6%	8.7%	0.87
SP 500 Index	9.4%	13.0%	0.74

Table 6: Annualized return, volatility and sharpe ratio during the test period (2014/07 to 2018/07) when using SP 500 as benchmark.

For the Fama/French five-factor attribution, we again found consistent results. Table 7 below contains the details. We also plotted cumulative returns for the ESG portfolios both before and after the Fama/French five-factor adjustments in Figure 6. Although the magnitudes of cumulative returns are slightly different from before, the rankings for the four portfolios remain similar.

Model behind Portfolio	Mkt-RF	SMB	HML	RMW	CMA	α	R^2
GBT (All Features)	0.0143 (0.561)	0.0242 (0.536)	0.0352 (0.468)	0.1938 (0.001)	-0.0875 (0.257)	11.8692 (0.013)	0.011
GBT (Top Features)	-0.0355 (0.120)	-0.0964 (0.008)	0.1011 (0.025)	0.2733 (0.000)	0.0506 (0.481)	13.9356 (0.002)	0.059
LR (Impute Mean)	0.0166 (0.429)	-0.1806 (0.000)	0.0933 (0.024)	0.1521 (0.003)	0.0563 (0.392)	5.5188 (0.176)	0.059
LR (Impute VAE)	0.0342 (0.121)	-0.1748 (0.000)	0.0237 (0.587)	0.0940 (0.085)	0.0837 (0.227)	7.1316 (0.097)	0.037

Table 7: Regression summary for ESG portfolios with Fama/French Five-Factor Model when using SP 500 as benchmark, the second to seventh columns show the estimated coefficients and their corresponding p-value.

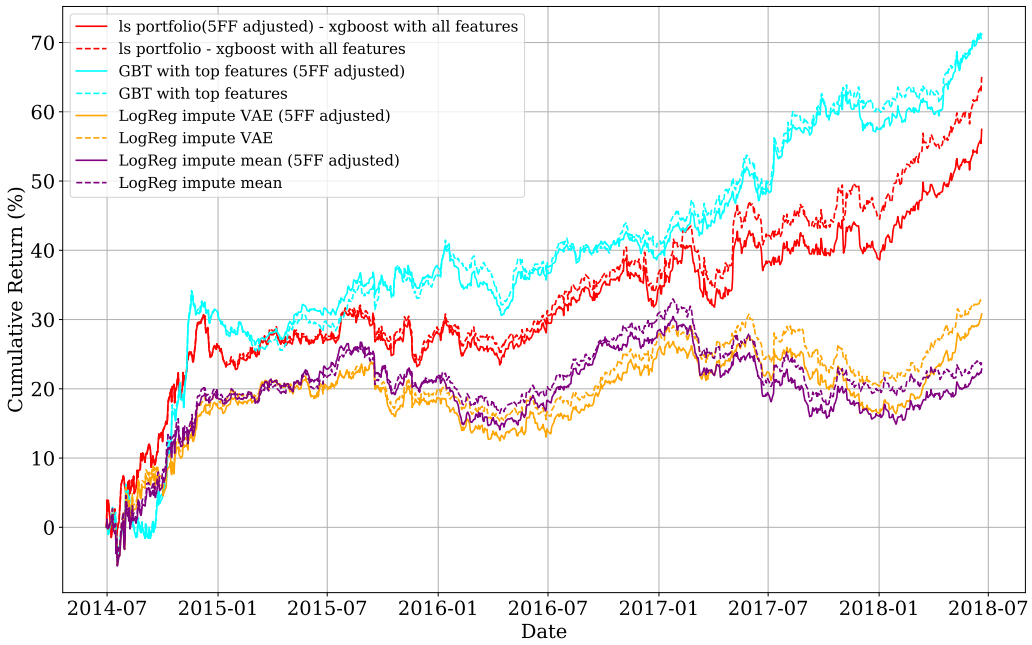


Figure 6: Cumulative returns for four portfolios before and after Fama/French five-factor adjustments during the test period when using SP 500 as benchmark.

B Long Only Strategy

Considering that short positions might be restricted for some clients, we also tested a long-only portfolio. We still used our three-class model as described above, but now we were only allowed to take the long side. In Table 8 below, we present the long-only results for the GBT model with top features and with the Russell 3000 as the benchmark. The annualized return for the long-only portfolio is higher than the benchmark's but its volatility is also higher. Nevertheless, its Sharpe Ratio is still higher than the benchmark's.

Portfolio Position	Return	Volatility	Sharpe Ratio
Long Only	25.1%	23.2%	1.08
Long Short	11.8%	7.5%	1.52

Table 8: Annualized return, volatility and sharpe ratio for ESG portfolios constructed from GBT with top features during the test period (2014/07 to 2018/07).

When attributing the returns of the GBT top feature long-only portfolio to the Fama/French five-factor model, we found that the portfolio return is more significantly explained by the five factors while the significance of the alpha coefficient is lower, although the magnitude of alpha is still larger than the magnitude of any of the other coefficients. The details are presented below in Table 9. Figure 7 shows the cumulative return for the long-only portfolio versus the long-short portfolio before and after the Fama/French five-factor adjustments. Although the total cumulative return for the long-only portfolio seems higher than that of the long-short portfolio, after adjusting for the five factors, the long-only portfolio has a slightly lower cumulative return in the test period.

Model behind Portfolio	**Mkt-RF**	**SMB**	**HML**	**RMW**	**CMA**	α	R^2
GBT long only (Top Features)	0.9976 (0.000)	1.1470 (0.000)	-0.5500 (0.000)	-0.9128 (0.000)	-0.4566 (0.000)	8.694 (0.013)	0.759

Table 9: Regression summary for ESG portfolios with Fama/French Five-Factor Model, the second to seventh columns show the estimated coefficients and their corresponding p-value.

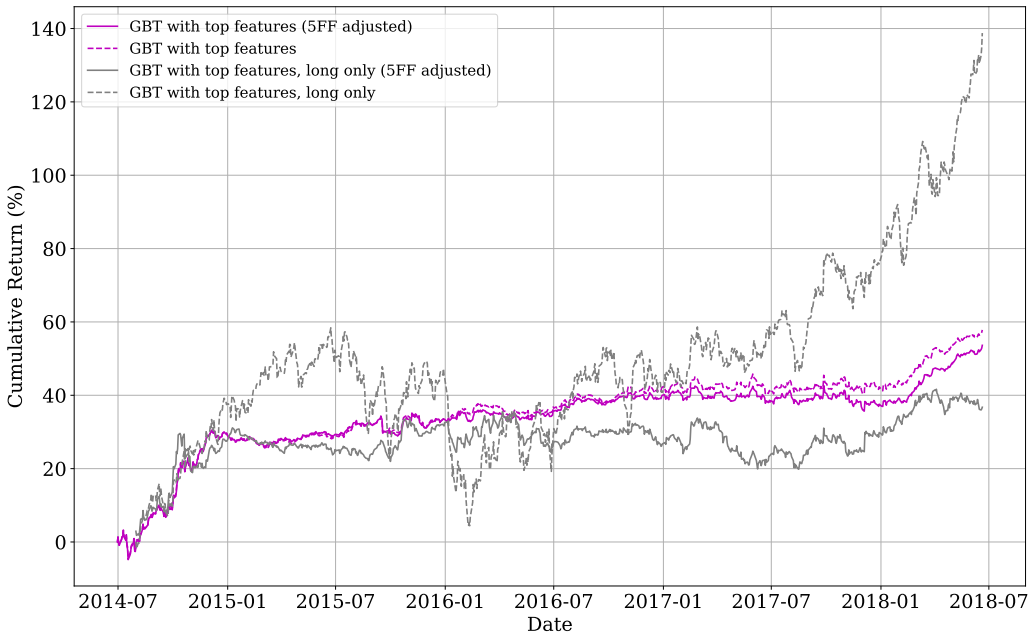


Figure 7: Cumulative returns for long only v.s. long short portfolios from GBT with top features before and after Fama/French five-factor adjustments during the test period.