ESG Booster

Data Science Feature – Model Explanation

Synechron

Table of Contents

Revision History

S.No.	Date	Version	Description	Author	Role
1	17/12/2021	1.0	Initial Draft	Sanup Kumar	Lead Business Analyst
2	05/01/2022	1.1	Comments on Sustainability report	Raj Patel	Sr. Associate-Technology
3	05/01/2022	1.2	Comments on Forecasting	Shreya Chandra	Sr. Associate-Technology
4	05/01/2022	1.3	Updated with Feedbacks received	Sanup Kumar	Lead Business Analyst

Objective

The objective of this document is to provide detailed overview and methodology used for each Data science features which were implemented in ESG Booster.

ESG Booster – A Walkthrough

ESG Booster is a web based portfolio management application which enables a Portfolio Manager to analyze his/her Portfolio w.r.t to ESG data and provide insights using various built-in features , which further helps him/her in take timely better Investment Decisions considering ESG factors.

Some of the Key Features are:

- Portfolio ESG rating provides an aggregated overview of portfolio's ESG performance, the trend view, the benchmarks against market indexes and outliner identification for actionable insight
- Incorporated with other data providers (Ravenpack for News, Orenda) and benchmarks (SDGs, SASB, etc.) to provide a comprehensive ESG view across the sectors.
- Enables users to improve the ESG score at the portfolio level, by proposing alternatives for low ESG scoring instruments without compromising financial performance
- Offers cutting edge data science features such as normalization of ESG data for multiple providers, ESG score forecasting, Network Graph Analysis & Sustainable Report Analysis

Below are the brief details of each Data Science feature implemented:

1. Forecasting

Business Objective:

In Order to Forecast the ESG score of a company based on Historical Data available considering the data trend, so that the investor can foresee and plan out his/her investment strategy for the specific company accordingly.

Algorithm/Methodology Implemented:

Fixed Effects Model has been used for prediction without news sentiment scores, additionally last 5 years of scores have been used using mean +/- 1 STD approach to refine the forecasts further if it deviates by a large amount from last year's score. Arimax has been implemented for prediction with news sentiment score.

Specific Reason for using above Algorithm/Methodology:

- Fixed effects model was implemented initially for forecasts with and forecasts without news sentiment score after analyzing the data .
- Later it was observed that the forecasts weren't following latest trend in data. Forecasts from both the models were nearly same (differed only in right side of the decimal point) .
- Fixed effects model was used for forecast without news sentiment score which was further refined by applying mean +/- 1 standard deviation approach for forecasts which deviated largely from last year's score.
- Arimax was implemented for forecasting scores with news sentiment score as fixed effect model has no way to add external variable. auto arima function was used to identify optimal model parameters for Arimax

Synechron' s Value Add:

Implemented multiple different types of Forecasting Models on same set of limited historical data, so result can be compared with including & excluding Ravenpack News Sentiment.

Finalized Content for UI (i) icon for ESG Booster:

i) *Score includes News Sentiment:* In this model, along with historical ESG scores, sentiments determined based on news feeds needs to be considered. Hence, forecasts are generated using ARIMAX model wherein sentiment scores are used as exogenous variable. Model parameters are identified using auto arima function based on last ten years of data after satisfying all model assumptions.

ii) *Score without News Sentiment:* Initially fixed effects model is used in which model parameters are assumed to fixed or non-random quantities. If the predicted values of score using fixed effects model is too far from last year of available data, then it is replaced by computing mean plus/minus one time standard deviation computed using five years of data. This will ensure that predicted values are within the pattern observed most recently.

2. Sustainability Report Analysis

Business Objective:

To provide one level deeper insight to Portfolio Manager, w.r.t a company published sustainability reports, so that Investor can understand on which SASB/SDG Goals company is more providing more focus and accordingly plan his/her investment strategy.

Algorithm/Methodology Implemented:

- *Focus Area:* TF-IDF-based classification model based on SASB taxonomy
- *Sentiment:* VADER-based sentiment model

Specific Reason for using above Algorithm/Methodology:

Our models conduct an analysis of a company's ESG Reports and are able to generate a sentiment and focus score for each relevant SASB category.

- The first step in this process, is to analyze and process the reports text. The text is split into sections and processed via Stemming and Lemmatization.
- After the text is split and preprocessed, we run a TF-IDF-based model to categorize it using the SASB taxonomy. The TF-IDF model is trained with business intelligence via keywords and SASB documentation to develop an understanding of each category and then identifies sections of text with the highest document similarity to each category.
- After the key sections of the report are identified for each SASB category, we apply our sentiment models. We use a VADER-based sentiment model which assesses a body of text for polarity (positive or negative sentiment) and intensity (strength of polarity) to generate a score.
- We adjust this score for any categorical and industry bias to generate a final sentiment score. To generate the focus score, a wider range of text is considered.
- A basic filter is applied to the report text to remove superfluous and non-ESG related text. We measure the document similarity of the remaining sections of text to each SASB category and obtain a raw score using TF-IDF.
- This score is adjusted for categorical and industry bias and then normalized on 0 to 100 scale (50 is average focus within industry).
- These models can be applied to other unstructured sources such as news articles to find mismatches between perceived and reported sentiment.
- Additionally, SASB taxonomy-based sentiment scores derived from news articles can be directly ingested from analytics providers like Ravenpack and compared with sentiment scores gleaned from reports

Synechron' s Value Add:

Custom Classification model based on SASB Taxonomy

Finalized Content for UI (i) icon for ESG Booster:

i) *Sentiment:* This section showcases a heatmap which is generated based on Companies Sustainable Reports published and sentiment scores calculated on report text. The text is categorized into SASB category issues which is further also mapped to Sustainable Development Goals (SDGs)

After the key sections of the report are identified for each SASB category, we apply our sentiment models. We use a VADER-based sentiment model which assesses a body of text for polarity (positive or negative sentiment) and intensity (strength of polarity) to generate a score. We adjust this score for any categorical and industry bias to generate a final sentiment score.

ii) *Focus Area:* This section showcases a heatmap which is generated based on Companies Sustainable Reports published and then generating a focus score categorized into SASB category issues which is further also mapped to Sustainable Development Goals (SDGs)

To generate the focus score, a wider range of text is considered. A basic filter is applied to the report text to remove superfluous and non-ESG related text. We measure the document similarity of the remaining sections of text to each SASB category and obtain a raw score using TF-IDF. This score is adjusted for categorical and industry bias and then normalized on 0 to 100 scale (50 is average focus within industry).

3. Network Analysis

Business Objective:

To help Portfolio Manager evaluate the company supply chain network companies and their impact on parent company based on -ve ESG specific News in market.

Algorithm/Methodology Implemented:

- *Weight Determination Model:* RavenPack Co-Mention Indicators
- *Sentiment Score:* RavenPack Sentiment Models
- *Score Generation Methodology:* Custom Centrality Graph Algorithms

Specific Reason for using above Algorithm/Methodology:

- The objective is to map relations between a company and other companies in its supply chain and gauge the influence that the sentiment scores of the suppliers and their associated subsidiaries have on the sentiment score of the company under observation.
- These relationships and their associated degrees of influence (High, Medium, Low) are stored in a Neo4J graph database with sentiment scores calculated from RavenPack data.

- A general observation is that suppliers often tend to greenwash by engaging in ESG-negative actions via their subsidiaries and thereby avoiding the negative sentiment that arises from the aforementioned actions.
- The relationships in the network graph allow us to transmit these negative sentiments upstream from subsidiaries to their respective suppliers and in turn to the companies consuming the supplies.
- This would result in an adjusted sentiment score which might shift from positive to negative based on the additional information gleaned from the network graph.
- The UI is also user customizable i.e. the user can add/remove suppliers and associated subsidiaries and the ISINs of the added companies would be leveraged to retrieve sentiment scores from the Datalake.
- The Datalake service can also be expanded to substitute RavenPack IDs for LLCs and privately held companies since these entities would not have associated ISINs.

Synechron' s Value Add:

Usable code implementation of RavenPack's Co-Mention Indicators

Finalized Content for UI (i) icon for ESG Booster:

The objective is to map relations between a company and other companies in its supply chain and gauge the influence that the sentiment scores of the suppliers and their associated subsidiaries have on the sentiment score of the company under observation. These relationships and their associated degrees of influence (High, Medium, Low) are stored in a Neo4J graph database with sentiment scores calculated from RavenPack data.

The relationships in the network graph allow us to transmit these negative sentiments upstream from subsidiaries to their respective suppliers and in turn to the companies consuming the supplies. This would result in an adjusted sentiment score which might shift from positive to negative based on the additional information gleaned from the network graph.

4. Portfolio Outliers

Business Objective:

To help Portfolio manager evaluate portfolio companies ESG, E, S and G scores w.r.t to multiple data providers and identify the ones who have maximum score difference/outliers.

Algorithm/Methodology Implemented:

Outlier/difference score's calculation is dependent on the industry's average and standard deviation. Here, z-score and Chebyshev rules normalized the means and the scores.

Below is the sample for outlier calculation:

Specific Reason for using above Algorithm/Methodology:

While developing this outlier feature, statistical methods such as rank, spearman's rank correlation etc., were tried. Due to the interpretation of ESG and pillar scores provided by different rating providers, Chebyshev's rule with average normalization fitted well.

Below is the methodology which we use to calculate the outlier:

- First calculate Average and Std deviation based on ESG score data for each industry.
- After this Z score can be calculated for company by formula – (ESG score - Avg score) / Std Deviation
  (ESG score - Avg score) / Std Deviation
- Then we calculate the Four-point range i.e. Divide the z score series into four range points that are from 1.25 to -1.25
- Post calculating Four-point range we Normalize the scores via Dividing the values (four range points) by 5 to make it 1-point range and add 0.5 (so that mean of both the series can be the same and comparable)
- Finally, we calculate the Outlier score, when one score is above average and other score is below average, we calculate the outlier score on absolute difference via taking the difference between the normalized score of both the methodologies.
- The same logic is applied also to calculate E,S and G level outliers.

Synechron' s Value Add:

Dynamic percentile value, the normalized score between two or more ESG rating providers.

Finalized Content for UI (i) icon for ESG Booster:

Outlier/difference score's calculation is dependent on the industry's average and standard deviation. Here, z-score and Chebyshev rules normalized the mean and the scores.

Below is the methodology which we use to calculate the outlier:

- First calculate Average and Std deviation based on ESG score data for each industry.
- After this Z score can be calculated for company by formula – (ESG score - Avg score) / Std Deviation
  (ESG score - Avg score) / Std Deviation
- Then we calculate the Four-point range i.e. Divide the z score series into four range points that are from 1.25 to -1.25

- Post calculating Four-point range we Normalize the scores via Dividing the values (four range points) by 5 to make it 1-point range and add 0.5 (so that mean of both the series can be the same and comparable)
- Finally, we calculate the Outlier score, when one score is above average and other score is below average, we calculate the outlier score on absolute difference via taking the difference between the normalized score of both the methodologies.
- The same logic is applied also to calculate E,S and G level outliers.