# Integrating Social Sustainability KPIs through APIs in Azure Data Factory: Leveraging Company Data for Performance Benchmarking and Predictive Analytics

Business Analytics MSc Capstone Project Summary

Ian Brandenburg

June 2024

As EU business data transparency regulations tighten, and companies begin implementing social sustainability and other ESG measures, Sustainista plans to be a pioneer in data-driven ESG sustainability. Sustainista strives to assist companies in streamlining their sustainability decision-making to meet achievable and standardized benchmarks. Through the integration of free, open-source APIs and machine learning tools, this project aims to assist Sustainista in developing a simplified and dynamic social sustainability KPI reporting dashboard that can be serve as a prototype for the services Sustainista can offer companies.

Currently, a key challenge for companies is the lack of social sustainability KPI benchmarks, and inconsistent reporting standards, which leads to companies failing to meet ESG related standards. Organizations may not have a standardized method for calculating or integrating KPI benchmarks into their system and could find it challenging to identify problem areas in their company, undermining progress toward meeting social sustainability standards. Finally, companies may be challenged by what KPIs to calculate and benchmark, creating difficulties in performance improvement. Sustainista would like to provide companies with solutions to streamline social sustainability decision-making through using a dynamic dashboard that displays the social sustainability KPI performance measures, including machine learning and regression metrics for predicting the company's performance in comparison to the standardized benchmark.

Four social sustainability KPI benchmarks were developed for this project to be prototyped for clients of Sustainista. The KPI formulas were provided by Sustainista, which include:

- Gender Wage Gap: ((Average Salary of Men – Average Salary of Women) / Average Salary of Men)) * 100
- Female Employment Rate: (Number of Female Full-Time Equivalents / Total Number of Full Time Equivalents) * 100
- Disability Employment Rate (Number of Employees with Disability / Total Number of Employees) * 100
- Occupational Illness Rate: (Total Number of Occupational Diseases / Total Number of Worked Hours) * 200,000 hours

Open-source APIs from Eurostats were utilized and integrated into Azure Data Factory (ADF). The use of ADF was crucial to keeping the entire process in the cloud and easily reproducible. Furthermore, utilizing a cloud service enables collecting newly reported API data used to generate KPI alerts. The necessary Eurostat APIs were identified for the given KPIs, called into ADF using Linked Services, copied into an Azure SQL Database, and fed through a data pipeline for processing and cleaning. Missing values underwent country mean value imputation.

To simulate real-world company experience, mock data, required to calculate company level KPIs, was generated in Python, using parameters to create specific scenarios. Employee data such as department and education levels were included. The target KPIs were calculated, basic feature engineering was conducted on the variables, and the data was analyzed through visual analysis and descriptive statistics. This data was integrated into the Eurostat API data and uploaded to Azure Blob Storage.

From Azure Blob Storage, the data were mounted into Azure Databricks, where further processing, exploratory data analysis, and variable grouping were conducted. A binary performance variable was developed to determine if the company was performing above or below the KPI benchmark, which became the target variable in the Linear Regression and ensemble machine

learning models. The ensemble models deployed included Random Forest, Decision Tree, and Gradient Boosting. Finally, the evaluation metrics for these models were the Accuracy and Area Under the Curve (AUC). Feature importance for the Linear Regression models were also collected for better understanding which features influence the successful performance of the company. Results from the analysis were sent to the Azure Blob Storage and called into Power BI to develop the dynamic KPI and predictive modelling dashboard.

Upon fulfillment of the clients needs, a dynamic KPI performance and ML results dashbioard was delivered to Sustainista. This dashboard shows the capabilities of open-source integration, KPI calculation, and machine learning tools that can measure and monitor the performance of a company's social sustainability KPIs. Through various slicers in the dashboard, a company can visualize underperforming departments, including top features in predicting positive organizational outcomes for heightened decision-making focus. As the delivered dashboard is entirely cloud connected, this prototype can be presented to clients as a representation of API integration KPI benchmarking.

The API collection, integration, analysis, and dynamic dashboard visualization were successfully developed and deployed through Azure cloud services, which will enable simple replicability and use-case demonstrations. The mock data utilized is for simulation purposes, which resulted in very high accuracy and AUC scores in the machine learning evaluation process. These mock datasets may not reflect a realistic company's machine learning or regression metrics but serve as a placeholder for real company datasets. Sustainista can leverage the integrated KPI benchmarks to supply client companies with strategies for data-driven social sustainability decisions. This project has created new learning experiences for the company, through data engineering processes in Azure, open-source API integration into machine learning, and dynamic dashboard visualizations. Through detailed documentation, a careful trail of work accomplished, and methods used were provided for the company's benefit of reproducibility.