PHASE – 1

PROBLEM DEFINITION AND DESIGN THINKING

Submitted by,

G.jayasri

CONTENTS:

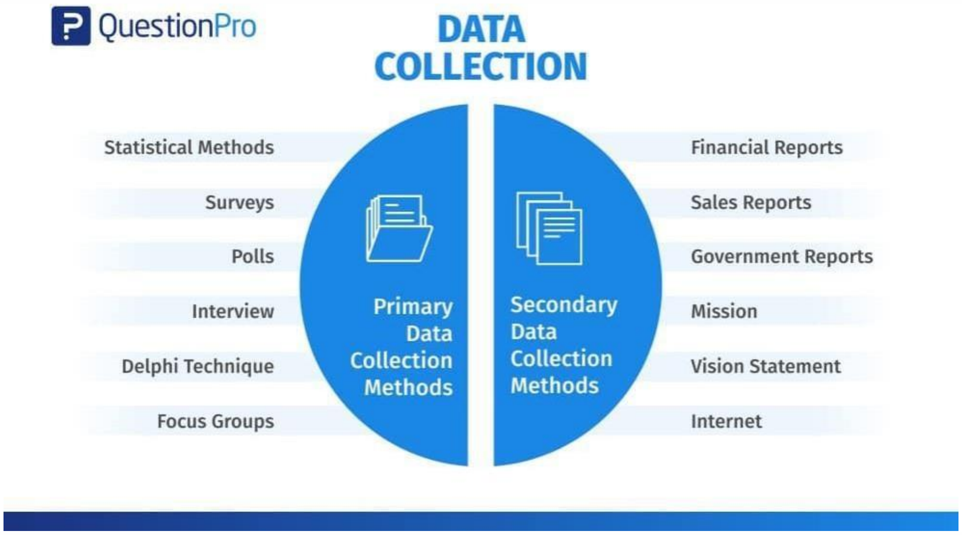
PROBLEM DEFINITION:

- Understanding the project objectives and requirements from a domain perspective and then converting this knowledge into a data science problem definition with a preliminary plan designed to achieve the objectives. Data science projects are often structured around the specific needs of an industry sector (as shown below) or even tailored and built for a single organization. A successful data science project starts from a well-defined question or need.

DESIGN THINKING:

Data collection:

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends, and probabilities, etc., to evaluate possible outcomes is Known as Data Collection.



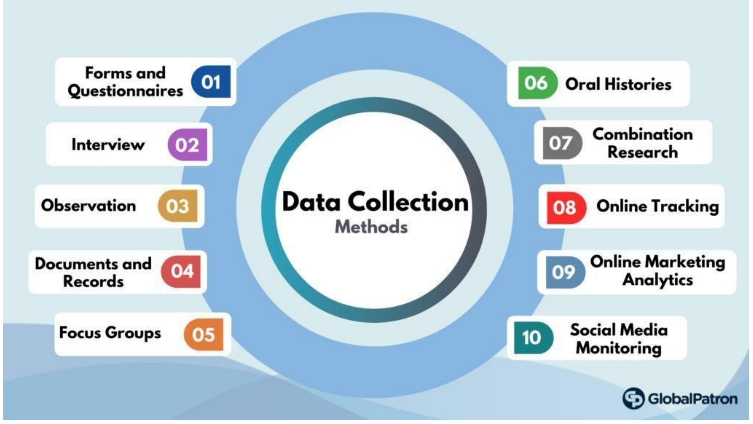
Methods of data collection:

- Surveys, quizzes, and questionnaires.
- Interviews.
- Focus groups.
- Direct observations.
- Documents and records (and other types of secondary data, which won't be our focus here.

Data collection uses:
The data collection uses provides the information that's needed to answer questions, analyze business performance or other outcomes, and predict future trends, actions and scenarios. In businesses, data collection happens on multiple levels.

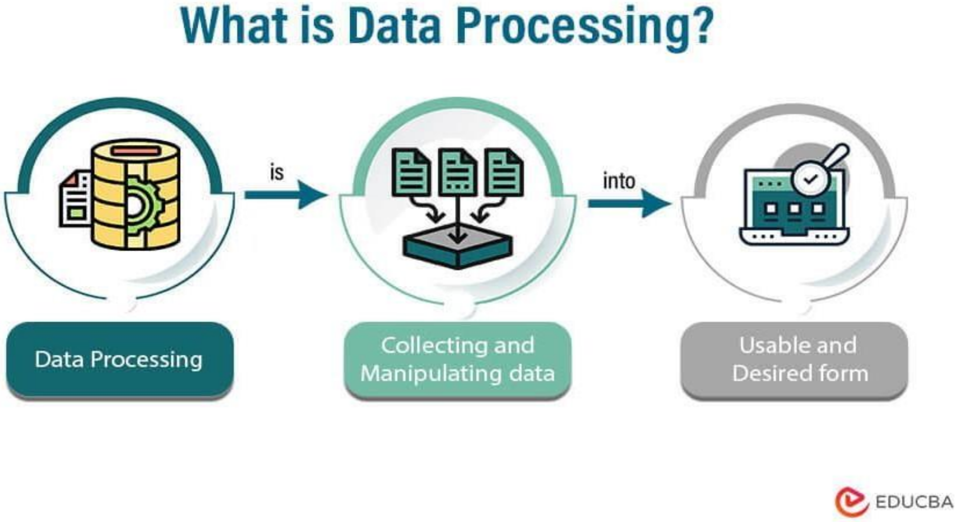
Data collection design:

Last updated on Aug 17, 2023. Data collection instruments are the tools and methods you use to gather and record information for your research or evaluation project. They can include surveys, questionnaires, interviews, focus groups, observations, tests, and more. Last updated on Aug 17, 2023.



Data preprocessing:

Data processing, manipulation of data by a computer. It includes the conversion of raw data to machine-readable form, flow of data through the CPU and memory to output devices, and formatting or transformation of output. Any use of computers to perform defined operations on data can be included under data processing.



The four main stages of data processing cycle are:

•Data collection.

•Data input.
•Data processing.
•Data output.

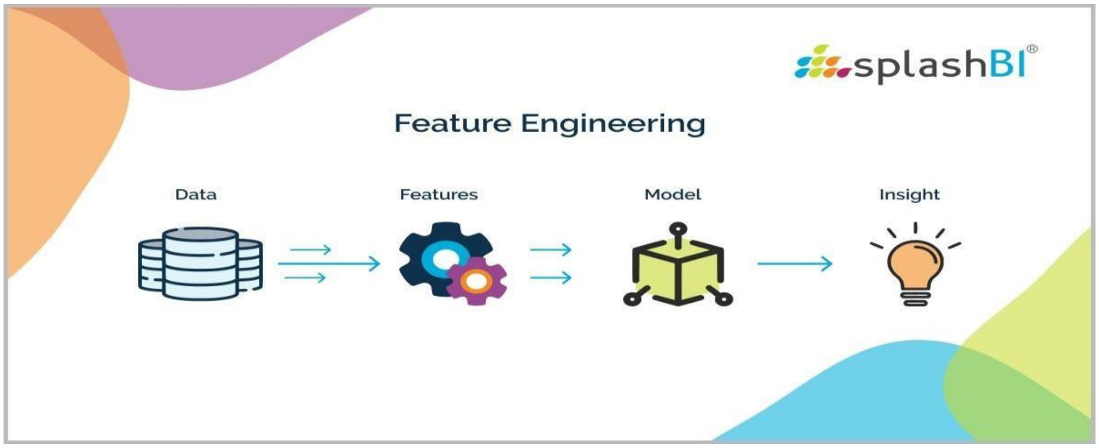
A very simple example of a data processing system is the process of maintaining a check register. Transactions checks and deposits are recorded as they occur and the transactions are summarized to determine a current balance.

Data processing is an essential component of modern computing and communication. It involves the manipulation, analysis, storage, and retrieval of data in order to produce. Read More. The future of data processing will be driven by advances in technology, such as artificial intelligence and machine learning.
Without data processing, companies limit their access to the very data that can hone their competitive edge and deliver critical business insights. That's why it's crucial for all companies to understand the necessity of processing all their data, and how to go about it.

Data processing cycle:
The data processing cycle is the set of operations used to transform data into useful information. The intent of this processing is to create actionable information that can be used to enhance a business.

Feature engineering:

Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.



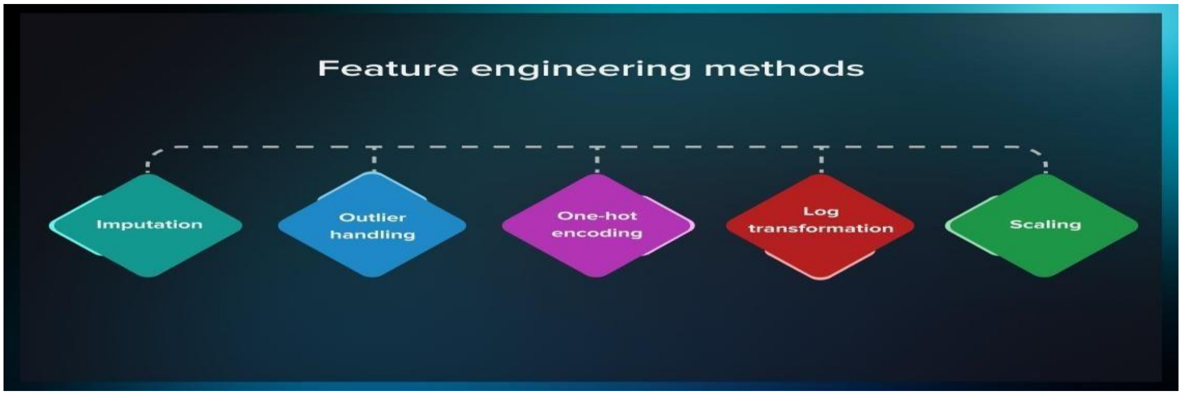
Feature engineering refers to manipulation addition, deletion, combination, mutation of your data set to improve machine learning model training, leading to better performance and greater accuracy. Effective feature engineering is based on sound knowledge of business problems and the available data sources. Feature engineering enables you to build more complex models than you could with only raw data. It also allows you to build interpretable models from any amount of data. Feature selection will help you limit these features to a manageable number.

Types of feature Engineering:

•7 of the Most Used Feature Engineering Techniques. Hands-on Feature Engineering with Scikit-Learn, TensorFlow, Pandas and Spicy.
•Encoding. Feature encoding is a process used to transform categorical data into numerical values that can be understood by ML algorithms.
•Feature Hashing.
•Binning / Bucketizing.
•Transformer.

Feature Engineering methods:

Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models



Model Selection:

Time Series Forecasting:

Time series forecasting is a data science task that is critical to a variety of activities within any business organization. Time series forecasting is a useful tool that can help to understand how historical data influences the future. This is done by looking at past data, defining the patterns, and producing short or long-term predictions.

There are four general components that a time series forecasting model is comprised of:

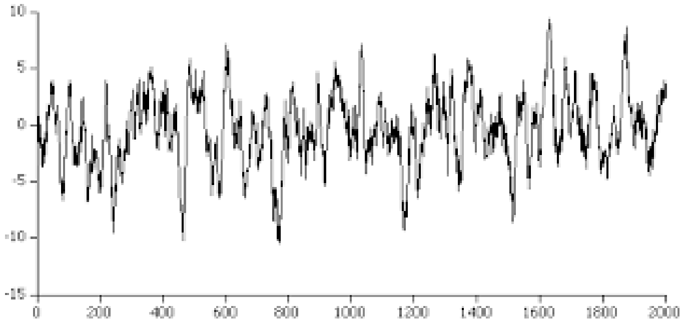
Trend: Increase or decrease in the series of data over longer a period.

Seasonality: Fluctuations in the pattern due to seasonal determinants over a period such as a day, week, month, season.

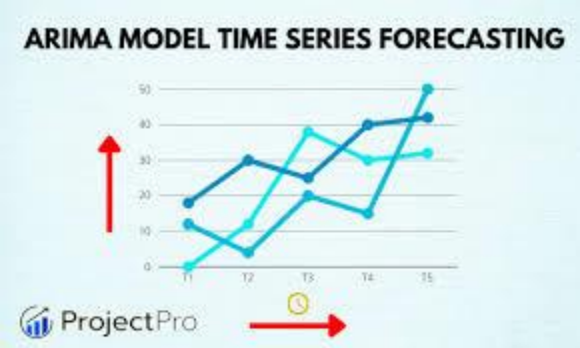
Cyclical variations: Occurs when data exhibit rises and falls at irregular intervals.

Random or irregular variations: Instability due to random factors that do not repeat in the pattern.

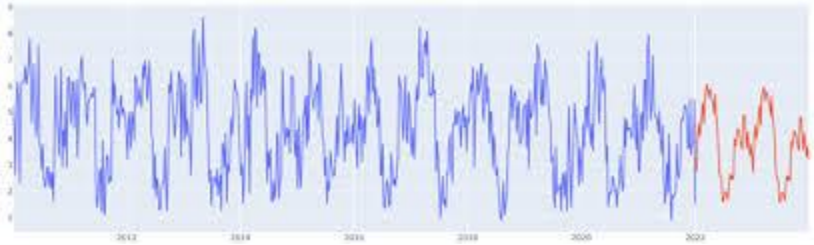
Autoregressive (AR): An autoregressive (AR) model predicts future behavior based on past behavior. It's used for forecasting when there is some correlation between values in a time series and the values that precede and succeed them.



Autoregressive Integrated Moving Average (ARIMA): Auto Regressive Integrated Moving Average, ARIMA, models are among the most widely used approaches for time series forecasting. It is a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

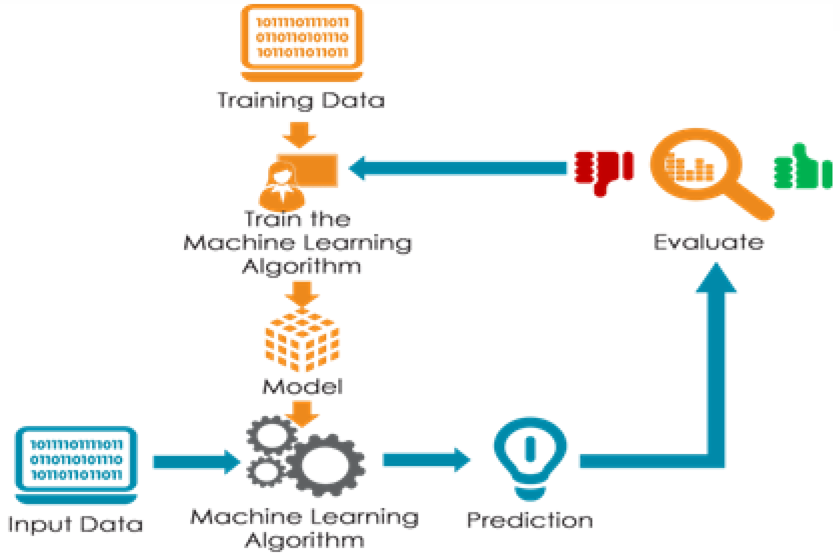


Seasonal Autoregressive Integrated Moving Average (SARIMA): Seasonal autoregressive integrated moving average (SARIMA) models extend basic ARIMA models and allow for the incorporation seasonal patterns.



Model training:

Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range. The p urp ose of model training is to build the best mathematical representation of the relationship between data features and a target label (in supervised learning) or among the features themselves (unsupervised learning). Loss functions are a critical aspect of model training since they define how to optimize the machine learning algorithms. Depending on the objective, type of data and algorithm, data science practitioner use different type of loss functions. One of the popular examples of loss functions is Mean Square Error (MSE).


Training Data
Train the Machine Learning Algorithm
Model
Evaluate
Input Data
Machine Learning Algorithm
Prediction

Why is it Important?

Model training is the key step in machine learning that results in a model ready to be validated, tested, and deployed. The performance of the model determines the quality of the applications that are built using it. Quality of training data and the training algorithm are both important assets during the model training phase. Typically, training data is split for training, validation, and testing. The training algorithm is chosen based on the end use case. There are several tradeoff points in deciding the best algorithm–model complexity, interpretability, performance, compute requirements, etc. All these aspects of model training make it both an involved and important process in the overall machine learning development cycle.

Evaluation:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Maybe the most popular and simple error metric is MAE:

MAE**:** The Mean Absolute Error is defined as:

While the MAE is easily interpretable (each residual contributes proportionally to the total amount of error), one could argue that using the sum of the residuals is not the best choice, as we could want to highlight especially whether the model incur in some large errors.

MSE & RMSE: For those cases, maybe MSE (Mean Squared Error) or RMSE (Root Mean Squared Error) are a better choice. Here the error grows quadratically and therefore extreme values penalize the metric to a greater extent.

$$\mathbf{MSE = \Sigma\ (y_i - p_i)^2 n,}$$

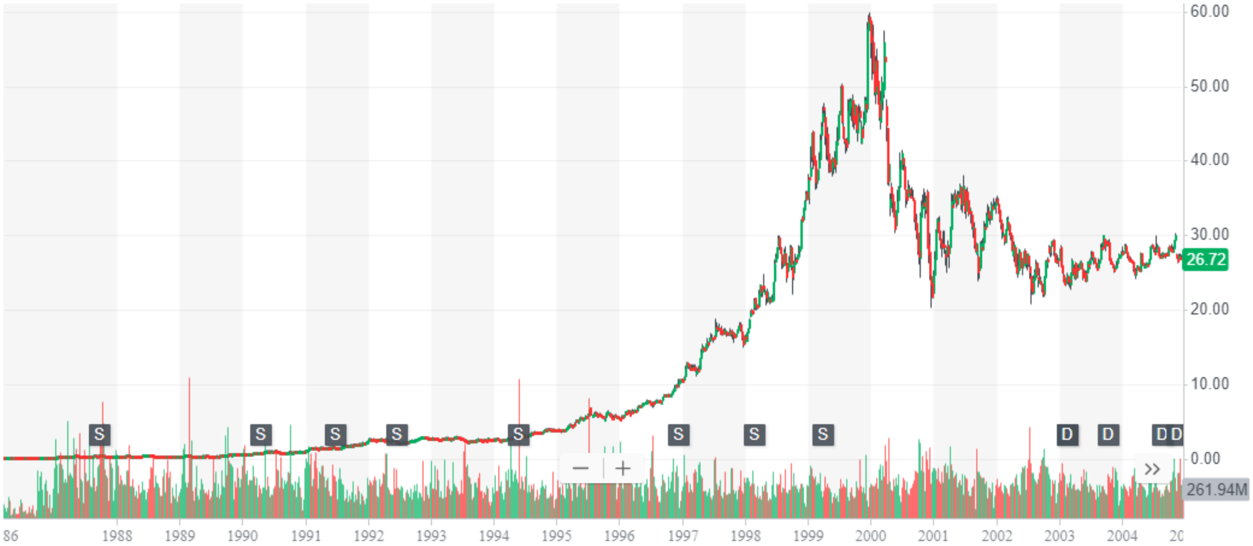
RMSE = Square root of MSE.

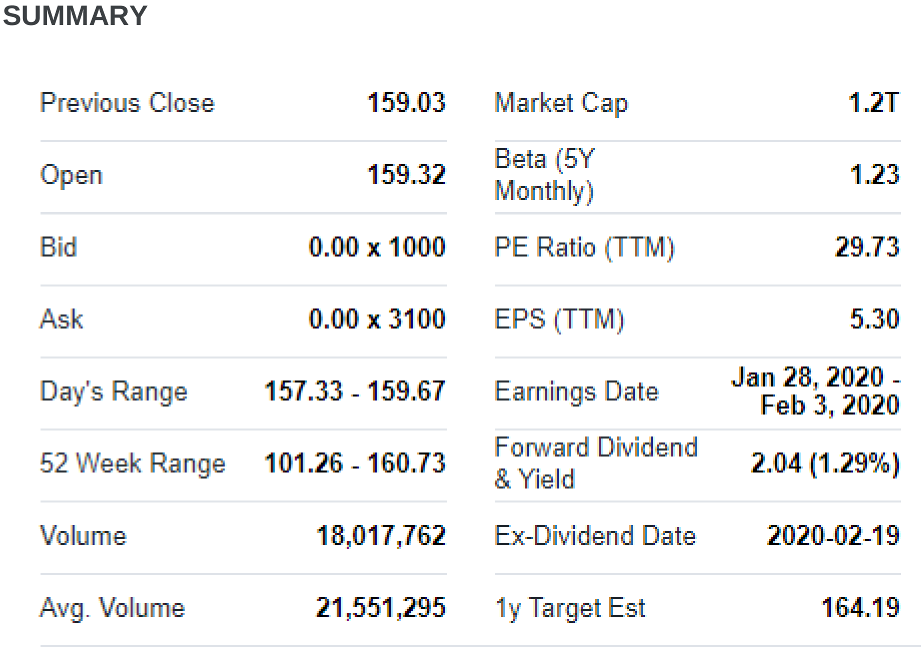
The main problem with scale dependent metrics is that they are not suitable to compare errors from different sources.

In our case, the capacity of the power plants would determine the magnitude of the errors and therefore comparing them between facilities would not make much sense. This is something we should try to avoid when choosing the metric.

ABOUT DATASET:

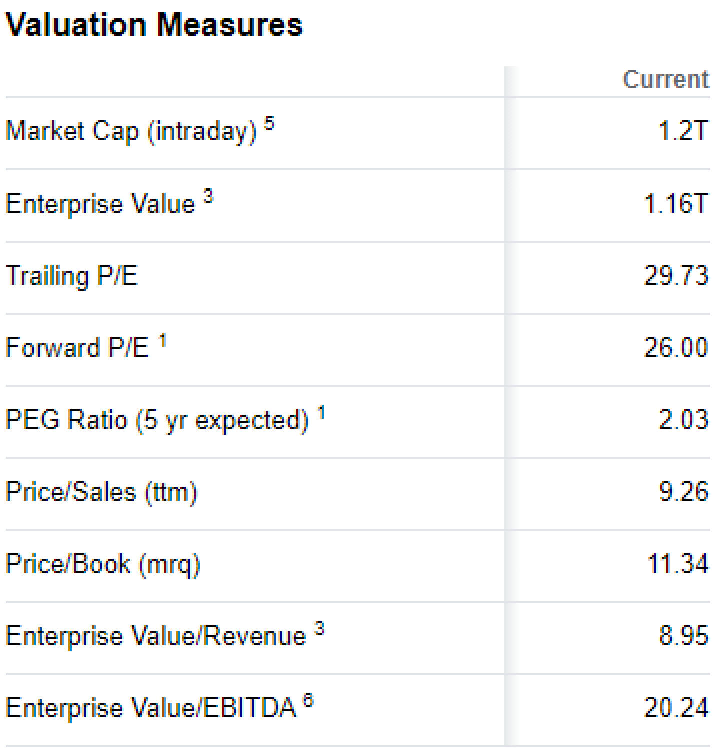
If you reach this DATASET, please UPVOTE this dataset to show your appreciation.





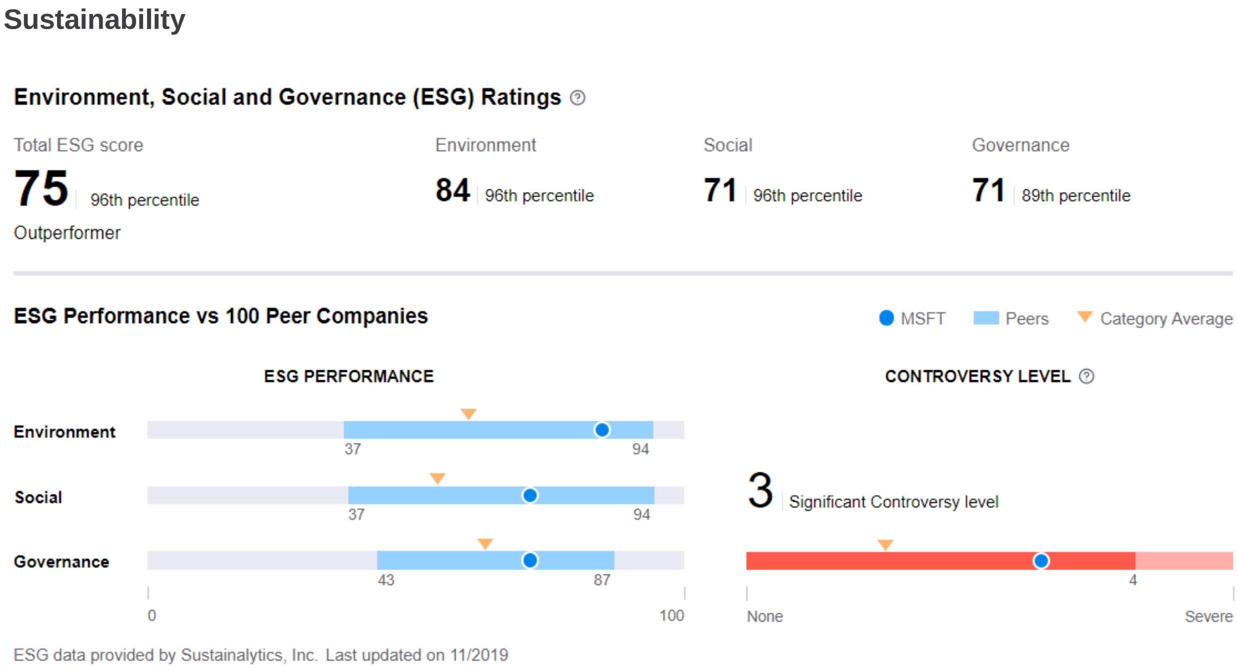
SUMMARY

Previous Close	159.03	Market Cap	1.2T
Open	159.32	Beta (5Y Monthly)	1.23
Bid	0.00 x 1000	PE Ratio (TTM)	29.73
Ask	0.00 x 3100	EPS (TTM)	5.30
Day's Range	157.33 - 159.67	Earnings Date	Jan 28, 2020 - Feb 3, 2020
52 Week Range	101.26 - 160.73	Forward Dividend & Yield	2.04 (1.29%)
Volume	18,017,762	Ex-Dividend Date	2020-02-19
Avg. Volume	21,551,295	1y Target Est	164.19



Valuation Measures

	Current
Market Cap (intraday) [5]	1.2T
Enterprise Value [3]	1.16T
Trailing P/E	29.73
Forward P/E [1]	26.00
PEG Ratio (5 yr expected) [1]	2.03
Price/Sales (ttm)	9.26
Price/Book (mrq)	11.34
Enterprise Value/Revenue [3]	8.95
Enterprise Value/EBITDA [6]	20.24



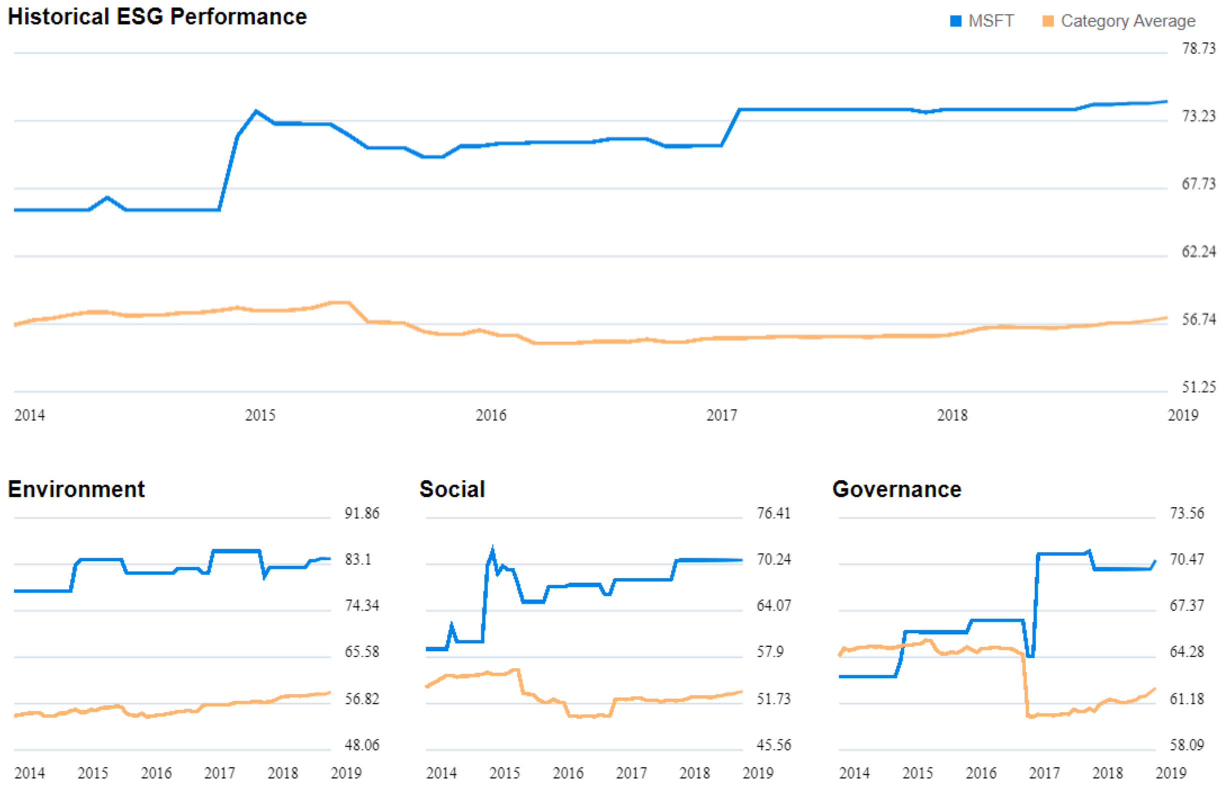
Sustainability

Environment, Social and Governance (ESG) Ratings

Total ESG score	Environment	Social	Governance
75 96th percentile Outperformer	**84** 96th percentile	**71** 96th percentile	**71** 89th percentile

ESG Performance vs 100 Peer Companies

ESG data provided by Sustainalytics, Inc. Last updated on 11/2019



Historical ESG Performance

Environment

Social

Governance