

**IBM**

**Naan Mudhalvan**

Phase 1: Problem Definition and Design Thinking

Problem Definition:

The problem is to build a machine learning/data science model that predicts stock prices based on the previous year's / Historical stock market data. Our goal is to create a model or tool that helps us to predict the future Stock prices and to help us make correct decisions and to correctly apply our investment strategies. This project consists of six various steps which are: data collection, data preprocessing, feature engineering, model selection, training, and evaluation.

Design Thinking:

Data Collection:

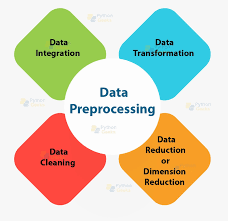
This is the first stage of this project. In this phase, we can acquire data through various methods such as collecting new data,converting/transforming legacy data, sharing/exchanging data, and purchasing data.In this project as we are already given a data set, we will be using that in this project. It is a Microsoft historical stock market data set which includes features like valuation measures, financial highlights, trading information, EPS Revisions, Growth estimates and various other features.

The Data set link is given below:

<https://www.kaggle.com/datasets/prasoonkottarathil/microsoft-lifetime-stocks-dataset>

Data Preprocessing:

This is the second stage of this project. This phase refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. We clean and preprocess the data, handle missing values, convert categorical features into numerical representations etc. The goal is to improve the quality of the data and make it ready for the next stage. For this, we will be using various tools such as Pandas library, Scikit learn library, Orange etc.

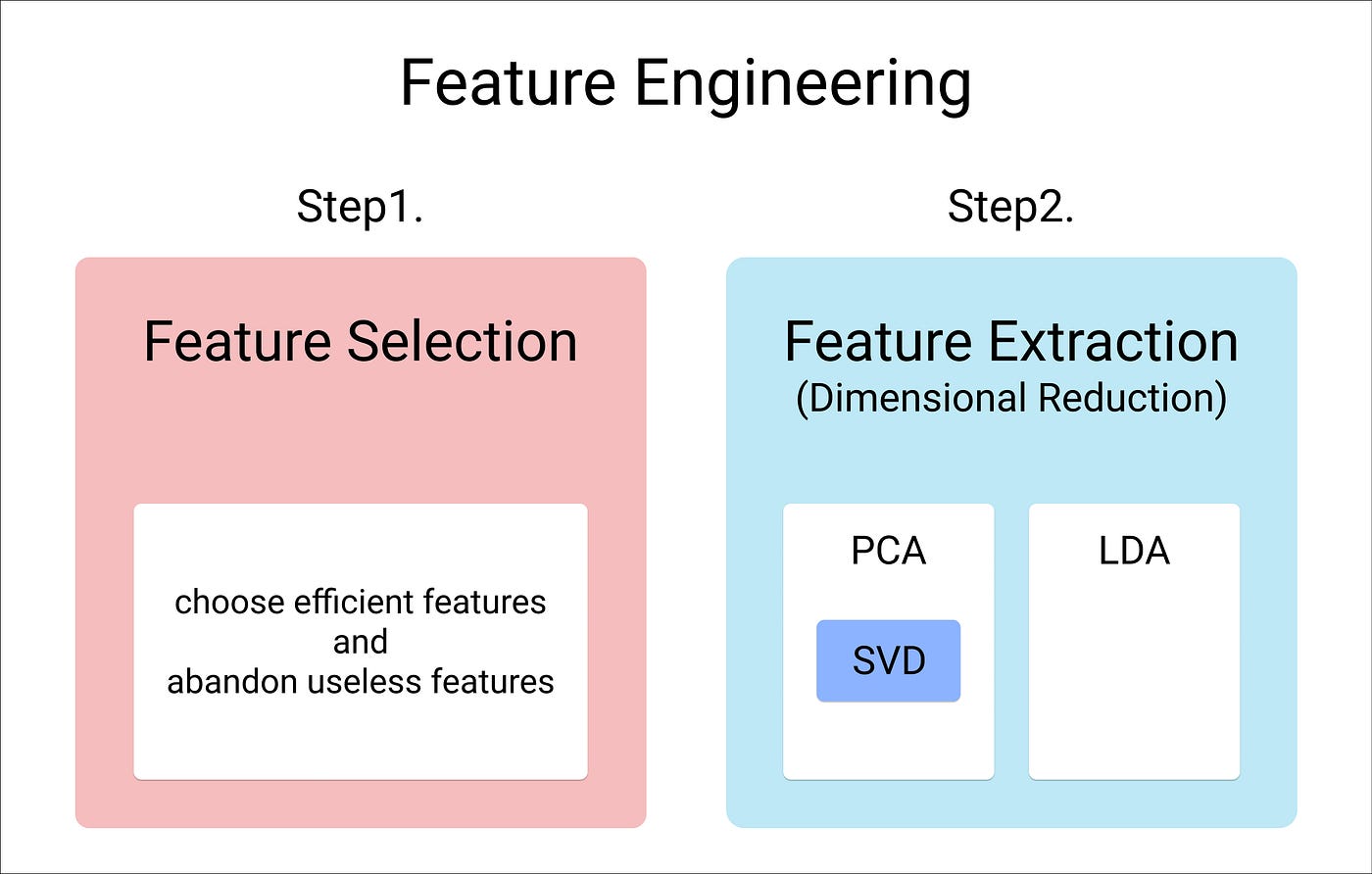


Feature Engineering:

This is the Third stage of this project. In this process we will be creating new features or transforming existing features to improve the performance of a machine-learning model. Creation of additional features will enhance the predictive power of the model. The data is transformed into a format that can be easily understood by the model.

We engineer features to improve the performance of machine learning models by providing them with relevant and informative input data. Raw data may contain noise, irrelevant information, or missing values, which can lead to inaccurate or biased model predictions. By engineering features, we can extract meaningful information from the raw data, create new variables that capture important patterns and relationships, and transform the data into a more suitable format for machine learning algorithms.

Some of the tools used in this process are: Featuretools,PyFeat, etc.



Model Selection:

This is the fourth stage of this project. This stage consists of choosing one among many candidate models for a predictive modeling problem(future stock prediction). We will be choosing suitable algorithms to predict future Stock prices. Careful consideration should be taken when evaluating machine learning algorithms for stock predictions. The Techniques/ Algorithms we will be using are:

* Long short-term memory (LSTM)
* Regression
* Classifier
* support vector machine (SVM)

Model Training:

This is the fifth stage of this project. In this phase the model is fed with sufficient training (preprocessed) data to learn from. The model is trained till it becomes efficient and the results are accurate.

Evaluation:

This is the sixth and last stage of this project. In this phase we will be evaluating our model's performance using appropriate values and measures such as:

* Concordant – Discordant ratio
* Root mean square error
* Root mean squared logarithmic error
* Cross validation, etc.

Conclusion:

Hence, the above steps will be followed by us to build a machine learning/data science model that predicts stock prices based on the previous year's / Historical stock market data.