Analysis of **Unemployment Index**

Gopalakrishnan Kumar, IIT –B Alumnus
Consultant Data Scientist
LinkedIn URL –
https://www.linkedin.com/in/gopalakrishnan-kumar-a73301110/
Git Repository **https://github.com/Gopalakrishnan-Kumar/Python-for-Data-Science**
Website/blog URL https://www.kaggle.com/gopalkk1

Project Goals

Objectives

- To classify the unemployment index of survival and unemployment group with respect to timeline estimate using Supervised Machine Learning method.
- To apply suitable machine learning model for classifying the unemployment index with respect to spell and event
- To create awareness of unemployment rate so as to suggest ways and means to minimize unemployment rate

Constraints

Inadequate information regarding column names
Missing values
Maximize-Maximize the customer satisfaction by good healthcare services

Business pipeline:

CRISP-ML(Q) Methodology

Phase 1: Business and Data Understanding

This phase helps us to ensure the feasibility of the project. As the data are available through from team lead the data quality has been assured.

Phase 2: Data Engineering

This includes data selection, data cleaning, feature engineering, and data standardising tasks.

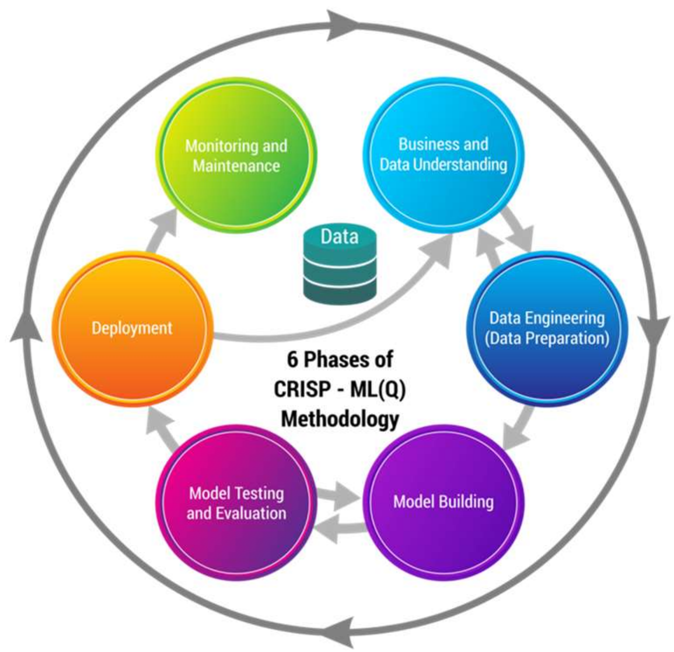
Data Selection- In this phase, filter method is used for data selection.

Data Cleaning- Error detection has been performed in the form of outliers, maximum, minimum, mean, average and mean deviation to all of these.

Feature engineering- In this project clustering and discretization of continuous attributes has been performed.

Data Standardization- In this process, the ML tools' input data are unified.

Data preprocessing has been done to ready the model fitting procedure by replacing with NA, NaN and so on. This enhances data reusability.



Business pipeline:

CRISP-ML(Q) Methodology

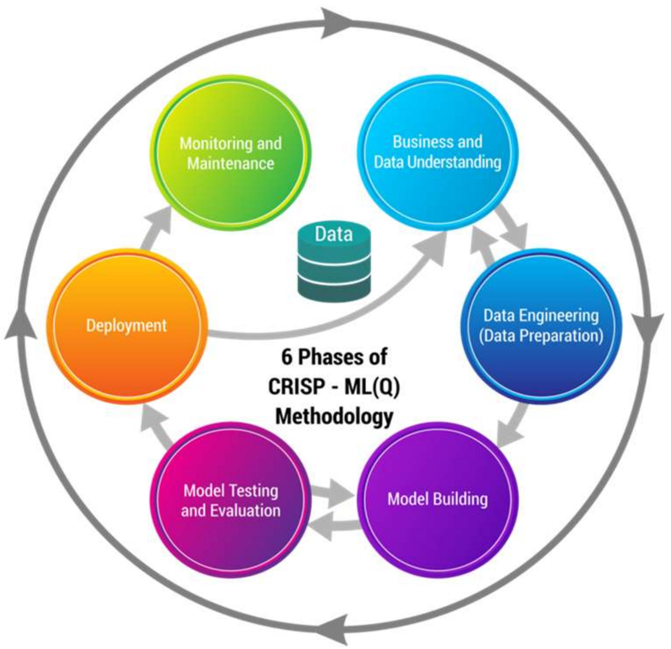
Phase 3: Machine Learning Model Engineering

Business and data understanding phase will shape this phase. Model assessment metrics might include performance metrics, robustness, fairness, scalability, interpretability, model complexity degree and model resource demand.

Model Selection / Specialization-

Kaplan-Meier Estimator

Being a non-parametric estimator, Kaplan-
Meier doesn’t require making initial assumptions about the distribution of data. It also takes care of right-censored observations by computing the survival probabilities from observed survival times. It uses the product rule from probability and in fact, it is also called a product-limit estimator.



$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

where:

d_i: number of events happened at time t_i

n_i: number of subjects that have survived up to time t_i

The survival probability at time *t_i* is equal to the product of the probability of surviving at prior time *t_i-1* and the percentage chance of surviving at time *t_i*.

Model training tasks-

Pretrained models and ensemble learning methods-

Business pipeline:

CRISP-ML(Q) Methodology

Phase 4: Model Testing and Evaluation

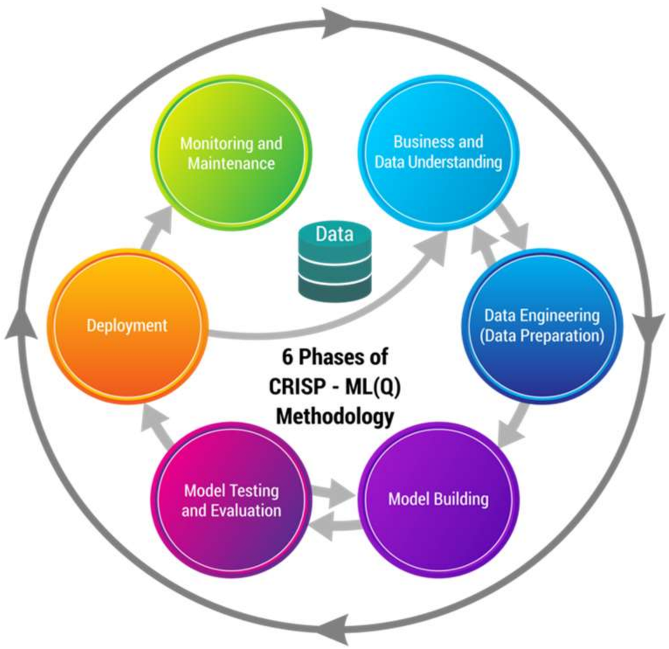
Consequently, model training is followed by a model evaluation phase also known as offline testing. Here, the performance of trained model needs to be validated on a test set. Then, the model deployment decision to be taken.

Phase 5: Deployment

The ML model deployment denotes a process of ML model integration into the existing software system.

Phase 6: Monitoring and Maintenance

Once the ML model has been put into production, it is essential to monitor its performance and maintain.



Technical Stacks

Languages: Python, html, cloud, sql, heroku, R, visual studio, git bash
AI/ML: Pytorch, skikitlearn,
Libraries: lifelines, pandas, numpy, skikitlearn
Database: Jupyter Online
Warehouse: Jupyter classic notebook
ETL: Python, Jupyter
Visualizations: Plotly()
Tracking & SC: Github

Project Architecture / Data Pipeline

- Data Processing and Preprocessing
- Explaratory Data Analysis

- Model Building
  - Model Selection/ Specialization
  - Kaplan Mier Estimator
  - Model Training Tasks

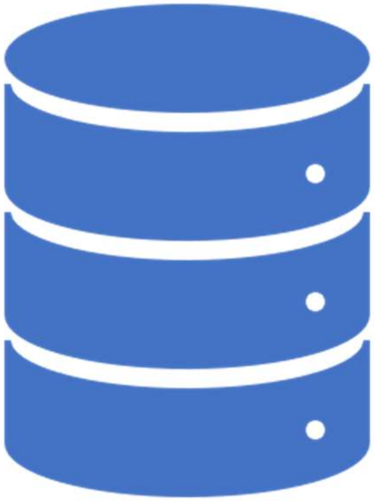
- Model Evaluation
  - Validation of performance of trained model

Model Deployment- ML model integration into the existing software system/ User acceptance and usability.

Data Preparation

Data used for training the model has been collected from Github and Wikipedia.

- Data cleansing- This includes describing mean, standard deviation, average, count, maximum, minimum, and removing NaNs.
- Data has been cleaned by removing NaNs.
- Outliers have been noted in value_counts()



Model Building

Model Specialization/ Standardization.
Kaplan Meir Fitter Model is the model to be used.
Tasks to be performed

Model Description

Algorithms used:

- **Kaplan-Meier Estimator**
- Being a non-parametric estimator, Kaplan-
- Meier doesn't require making initial assumptions about the distribution of data. It also takes care of right-censored observations by computi the survival probabilities from observed survival times. It uses the product rule from probability and in fact, it is also called a product-limit estimator.

$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

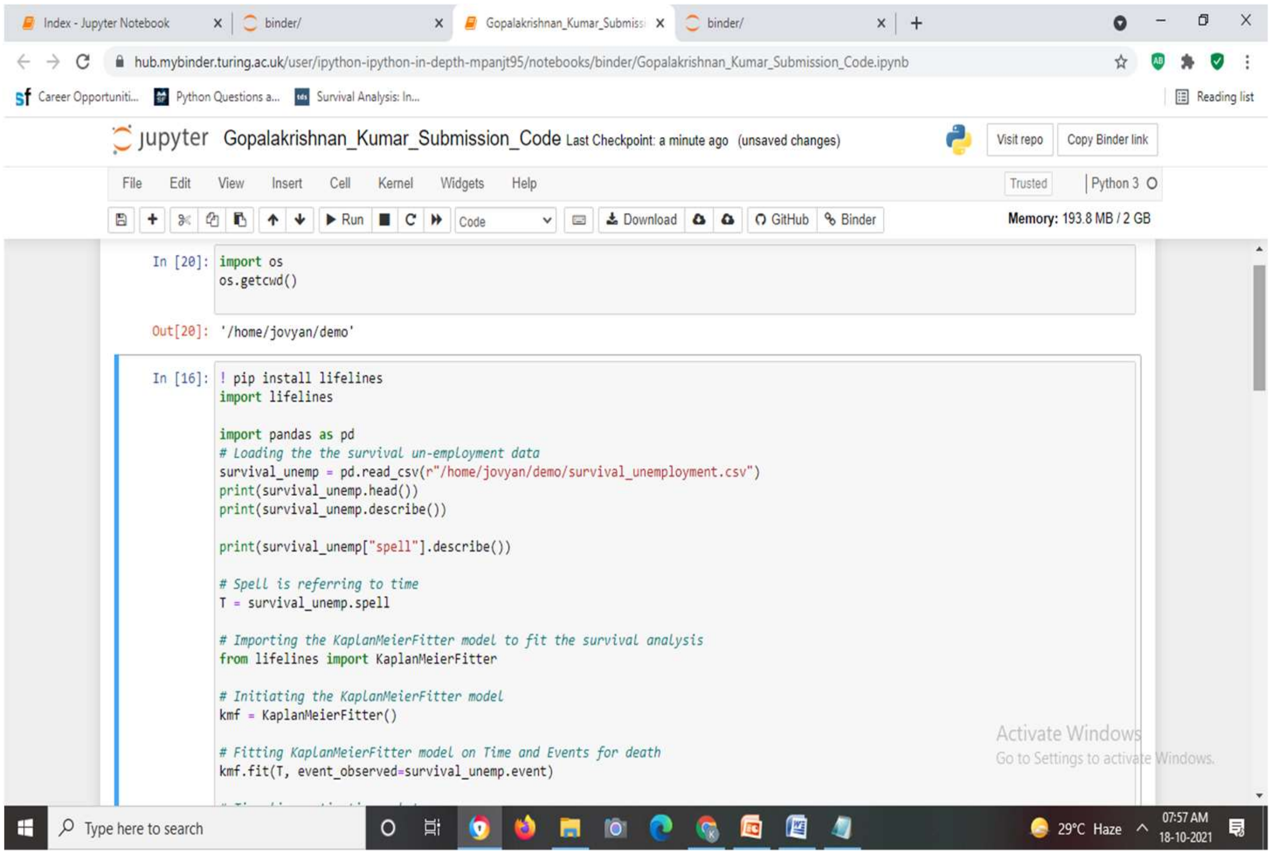
- where:
- **d_i**: number of events happened at time t_i
- **n_i**: number of subjects that have survived up to time t_i
- **Descriptions:**

❖ Defining the model as 'Kaplan Meier Fitter '

❖ Fitting the model with parameters like Time, Spell, Event

❖ Plot

Outputs of the Model in Codes

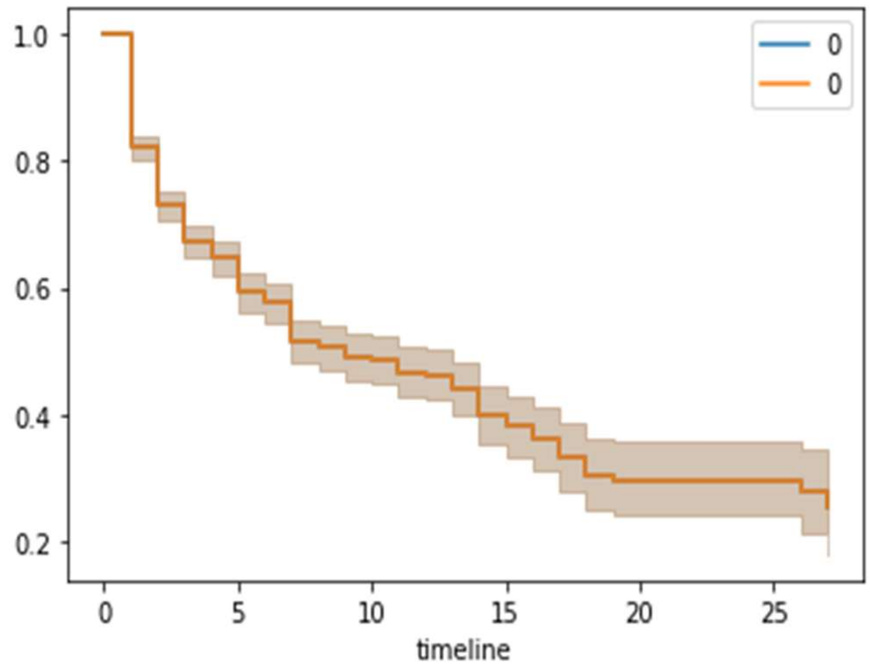


Deployment Strategy

- Web Based- HTML applications
- Enterprise Based- AI algorithm/ Kaplan Meier Fitter Model

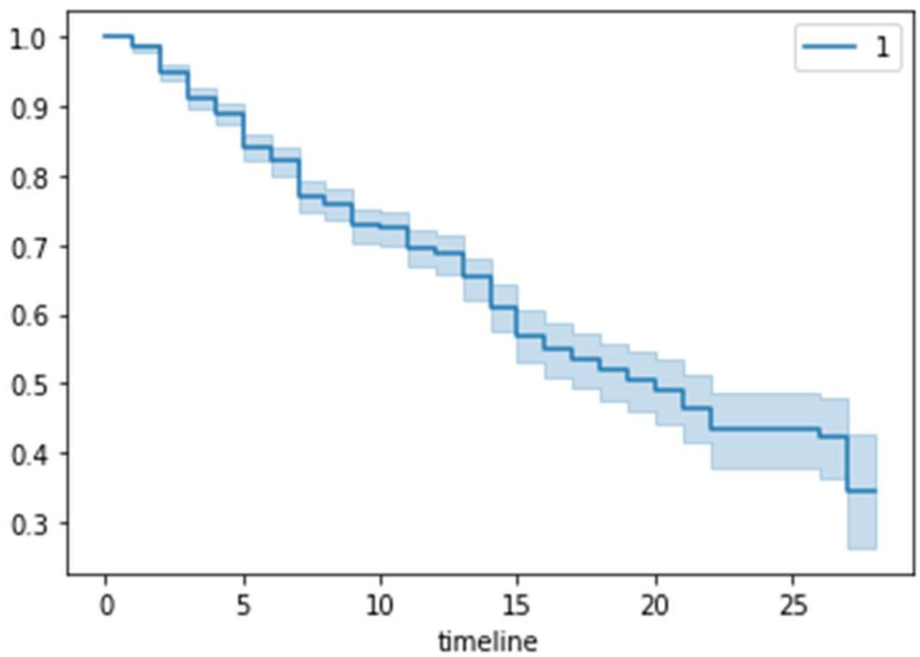
Outputs of the Deployed Model

<u>Graph 1- Time and events from Group “0”</u>



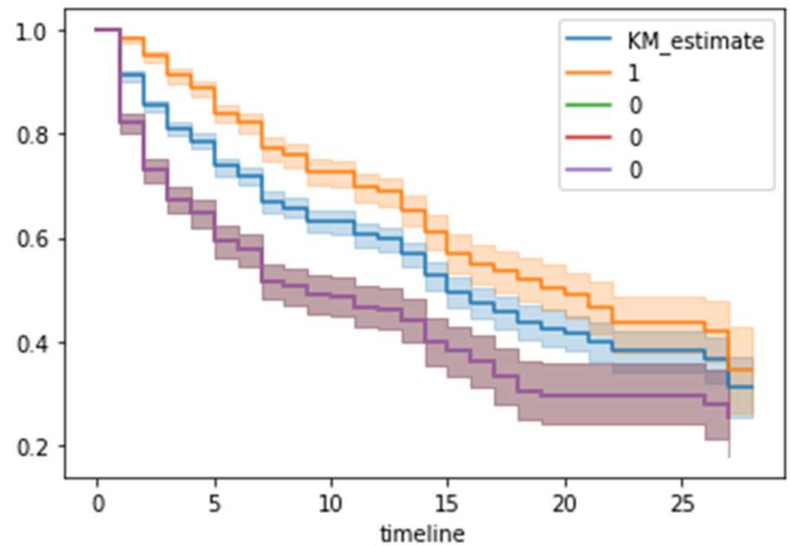
Outputs of the Deployed Model

<u>Graph 2- Time and events from Group “1”</u>



Outputs of the Deployed Model

<u>Graph 3- Time-line estimations plot</u>



The graph is plotted KM estimate vs. timeline

Outputs of the Deployed Model

1. The graph is plotted KM estimate vs. timeline
2. The Kaplan-Meier plot can be interpreted as follows:
The horizontal axis (x-axis) represents time in days and the vertical axis shows the probability of surviving or the proportion of people surviving.
3. The line represents survival a curve of the lines represents survival or proportion of people surviving. A vertical drop in the curves indicates an event. Survival function can be interpreted as the probability that a certain object of interest will survive beyond a certain time'‘t’.
4. The value of the function lies between 0 and 1 (both inclusive and it is a non-increasing function). The value of the function is above the KM curve for occurrence of unemployment and the value of function is below the KM curve for survival group.
5. Comparing the KM curves in Figure 3, the survival duration of the survival group is longer than unemployment.


Thanks