BM 655: Essential AI for Managers

Case Study 1: Building regression model to predict life expectancy

Instructor:	Manu K. Gupta	**Due date:**	3^{ed} September
Email:	manu.gupta@ms.iitr.ac.in		11:59 PM

Overview: In this case study, we build multi linear regression model using sklean to predict the life expectancy against the relevant response variables.

Problem Statement: The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this case study we have considered data from year 2000-2015 for 193 countries for further analysis. Dataset consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors and Social factors.

Requirement: You are required to build multi linear regression model with the purpose of predict the life expectancy against the relevant response variables.

Dataset: Dataset link: https://www.kaggle.com/kumarajarshi/life-expectancy-who
Dataset name: Life Expectancy Data.csv

Guidelines: Your activities towards the problem statement would involve:

1. Missing Value Treatment and outliers
2. Explore the relationship between the Life expectancy variable and the response variables.
3. Feature selection
4. Split the data in to train & test using sklearn
5. Building model & performance (RMSE & R-squared (R2)) evaluation
   (a) Multi Linear Regression