Life Expectancy Analysis using Regression

Milestone 1: CRISP-DM model

Project plan

Stage 1 - Determining Business Objectives:

Stage 1.1: Assess the current situation

1. Inventory of resources:

- Personnel: the learner, other participants, mentor.
- Data files: education.csv, lifeExpectancy.csv, crime.csv, areas.csv, income.xlsx, and region.txt
- Computing resources:
    - hardware platforms that you have at your disposition (for this project a PC will be enough)
- Software resources:
    - Anaconda with Jupyter Notebook and Spyder
    - Python Libraires: numpy, pandas, matplotlib, seaborn, scipy, and Statsmodels

2. Requirements, assumptions, and constraints:

Following the wikipedia sources, we see that the population, household income, education, area, data comes from the U.S. Census Bureau, which only requires a citation of the sources. The R package page informs us that the data was collected from the Census Bureau. The life expectancy data comes from the CIA World Factbook that is in public domain.

The schedule for completing the project is five weeks.

3: Risks and contingencies

A possible risk is that the variables that we analyze are only mildly related to life expectancy. A solution would be to collect more data.

Regarding the secondary goals, a risk is that the provided files do not contain the necessary information to answer the questions. In this case, we will have to collect more data.

Stage 1.2: Determine business and data mining goals

1. Primary objective

The Jones Family Foundation's efforts are focused on education and research. This year's primary objective is to identify the factors related to the life expectancy of the American population. From a data mining viewpoint this objective traduces into doing a regression analysis to find the variables related to life-expectancy.

2. Measures of success

A measure of success for primary data mining objective can be to find a model with appropriate variables that could explain at least 75% of the variability in life expectancy (R^2 >= 0.75). From a business criterion, a measure of success could be the presentation of clear infographics explaining the findings in terms that will make sense for the client.

3. Secondary Objectives

The secondary objective relates to the data mining questions. From a business perspective the questions are as follow:

1. What percent of the states have a life expectancy greater than 80 years?
2. Which state has the highest life expectancy, and which state the lowest?
3. Is life expectancy equally distributed across the different regions of the U.S.?
4. Is education equally distributed in each of the three levels (high school, bachelor, and advanced degrees)? Does any of the educational levels show a greater spread across the states?
5. Which variables have a direct relation with life expectancy?
6. How does the level of education in the different states relate to life expectancy and income?

From a data mining perspective, the questions are:

7. We can answer this question computing the empirical cumulative distribution function of life expectancy.
8. We will compute standard statistics on the life expectancy data.
9. We will plot the distribution of the life expectancy grouping the observations per region and then analyze the distributions.
10. For the first part of the question, we will do a strip plots and box plots of the education variables. For the second part we will plot the empirical cumulative distribution function.
11. To answer this question, we will do regressions plots between life expectancy and the regressors.
12. To answer this question, we will do a scatterplot for each level of education (high school, bachelor, and advanced) showing life expectancy against

income and we will use the semantic of hue, size, and style to map the third variable (education).

4. Measures of success

Step	Time	Risks
Business understanding	1 week	Since the client is not at our disposition to make questions, we may be at risk of making wrong assumptions.
Data understanding	1 week	We may not have at our disposition a clear definition of the data provided.
Data preparation	1 week	The data provided to run the regression analysis may be incomplete or contain errors.
Modeling	1 week	The chosen modelling method may not be the most appropriate.
Evaluation	1 week	The variables that we have at disposition may only be partially related to life_expectancy. In this case, we may need to collect more data.
Deployment		In this phase, we must be very careful to present the results in terms that the client will understand.

Regarding the secondary objectives, we will consider a success from a data mining point if we produce plots or statistics that answer the data-mining questions. From a business perspective this will constitute a clear presentation clear of the plots and statistics.

Stage 1.3 Producing a project plan

1. Project Plan

2. Initial Assessment of tools and techniques:

We will use the following Python libraries during the project: numpy, pandas, matplotlib, seaborn, scikit-learn, and Statsmodels. We will use linear regression to find factors related to life expectancy.