# **EXPERIMENT REPORT**

| **Student Name** | Bui The Hai |
| --- | --- |
| **Project Name** | Part B: Experiment on multivariate linear regression |
| **Date** | Mar 30, 2023 |
| **Deliverables** | Assignment 1 Part B Experiment on multivariate linear regression |

| 1. **EXPERIMENT BACKGROUND** | | |
| --- | --- | --- |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | | |
| **1.a. Business Objective** | The objective of this experiment is to illustrate the relationship between the cancer death rate and all other factors listed in  The results of this study may point to a potential problem with people's perception of cancer. People with more education are more knowledgeable about cancer risk and may not be as susceptible to cancer. | |
| **1.b. Hypothesis** | The hypothesis for this part is people who have a lower education level or are unemployed will suffer a higher risk of death due to cancer and poor people face higher risk of cancers | |
| **1.c. Experiment Objective** | It is expected that people who obtain a higher level of education will face less risk of cancers. Therefore, the regression model of ***TARGET\_deathRate***and ***dependent variables listed below*** will show a negative coefficient | |

| 1. **EXPERIMENT DETAILS** | | |
| --- | --- | --- |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | | |
| **2.a. Data Preparation** | For data preparation, the first step is to import data and examine the summary of the data.The train dataset includes 2438 rows and 35 attributes while the test dataset has 609 rows and 35 attributes. In 35 attributes, the dataset has two object data type variables (Geography and binned Inc). The remaining variables are numerical.  In my regression model, I will eliminate these two variables from the regression model and only keep float and int data types. This step gives us an overall picture of our datasets.  The second step is to prepare variables for training. I create two subset of the two dataframes, each containing ***TARGET\_deathRate*** (dependent variable) and independent variables listed as **PctNoHS18\_24','PctHS18\_24','PctSomeCol18\_24','PctBachDeg18\_24','PctHS25\_Over','PctBachDeg25\_Over','PctEmployed16\_Over','PctUnemployed16\_Over'** (independent variables). | |
| **2.b. Feature Engineering** | The feature engineering steps will be performed on the part C of this experiment. | |
| **2.c. Modelling** | In order to illustrate the effect of education level and cancer death rate, two independent variables named **PctNoHS18\_24', 'PctHS18\_24', 'PctSomeCol18\_24', 'PctBachDeg18\_24', 'PctHS25\_Over', 'PctBachDeg25\_Over', 'PctEmployed16\_Over', and 'PctUnemployed16\_Over'** will be chosen to train a univariate regression model for each of them, with the dependent variables being ***TARGET\_deathRate.*** | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | | |
| **3.a. Technical Performance** | Table 2. Regression result of model 2   |  | MSE of baseline | MAE of baseline | MSE of model | MAE of model | | --- | --- | --- | --- | --- | | Model 3: Multivariate linear regression | 758.055 | 21.31 | 641.2 | 18.2 |   Table 3. Coefficient table   | Intercept | 169.0139999 | | --- | --- | | PctUnemployed16\_Over | 2.279320084 | | PctHS25\_Over | 0.644776806 | | PctHS18\_24 | 0.151037299 | | PctBachDeg18\_24 | 0.100588005 | | PctSomeCol18\_24 | -0.016590731 | | PctEmployed16\_Over | -0.133932735 | | PctNoHS18\_24 | -0.483402997 | | PctBachDeg25\_Over | -1.497963072 |   Firstly, comparing the results with the baseline shows that the model performs better. Secondly, the coefficient table shows that unemployed people face a higher risk of cancer death. People aged 25 and over who achieve bachelor’s degrees face less risk of cancer death. The level of education will have a positive effect on the cancer death rates. | |
| **3.b. Business Impact** | The model results have many implications because it highlights the tremendous impact of education on reducing cancer death rates. People’s lifestyles is one of the factors contributing to the risk of cancers. Therefore, the government should be in charge of educating people about the effect of lifestyles on their health. | |
| **3.c. Encountered Issues** | The issue of these experiments is that they are tested in US society. We need a bigger dataset to test the influence of education level on cancer death rate. | |

| 1. **FUTURE EXPERIMENT** | | |
| --- | --- | --- |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | | |
| **4.a. Key Learning** | My key learning from the experiment is the connection between the education level and the target death rate, thereby emphasizing the impact of education on people's risk of cancer. | |
| **4.b. Suggestions / Recommendations** | My suggestion for these parts is to carry out broader research on the impacts of education level on people. We may divide the education areas into smaller subsets, such as Economic education IT education, and to obtain the impact of different areas | |