# **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 socioeconomic factors.  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 live in areas with less educational resources will suffer a higher risk of death due to cancer. |
| **1.c. Experiment Objective** | The objective of the experiment is to examine the effect of various socioeconomical factors on target death rate. It is expected that people who obtain a higher level of education will face less risk of cancer. |

|  |  |
| --- | --- |
| 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. I perform concatenating the two data frames to one (splitting will be performed in next part). Below is a summary of the dataset.    In my regression model, I will perform label encoding on these two variables. This step gives us an overall picture of our datasets. However, the two columns 'PctSomeCol18\_24' and 'PctEmployed16\_Over' miss 152 and 609 rows respectively, which will be imputed by its mean values.  On the other hand, columns PctSomeCol18\_24 will be eliminated because it is missing a substantial amount of data (2285 out of 2438 rows)  ***TARGET\_deathRate*** has outliers, which should be removed to make it symmetric by the following code:  ***df=df[(df['TARGET\_deathRate']<=240) & (df['TARGET\_deathRate']>=120.2)]***  A screenshot of a computer  Description automatically generated  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 ‘**incidenceRate', 'medIncome', 'studyPercap', 'MedianAge', 'Geography', 'AvgHouseholdSize', 'PctNoHS18\_24', 'PctHS18\_24', 'BirthRate'** (independent variables).  The above variables cover a broad range of factors like health outcomes, demography, education and geography, which could potentially explain variations in death rates across different populations or geographic regions.  Below is correlation matrix of independent variables:    The data will be splatted for the next modelling phase by the following code:  ***from sklearn.model\_selection import train\_test\_split***  ***# Split the data into training and testing sets (60% training, 40% testing)***  ***X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.4, random\_state=42)*** |
| **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 various factors on cancer death rate, two independent variables named ‘**incidenceRate', 'medIncome', 'studyPercap', 'MedianAge', 'Geography', 'AvgHouseholdSize', 'PctNoHS18\_24', 'PctHS18\_24', 'BirthRate'** 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 (baseline) | MAE (baseline) | MSE (training set) | MAE  (training set) | MAE (validation set) | MSE (validtion set) | | Model 3: Multivariate linear regression | 585.27 | 19.42 | 377.37 | 15.05 | 347.7 | 14.5 |   Table 3. Coefficient table   |  |  | | --- | --- | | ***Intercept*** | 177.70 | | ***incidenceRate*** | 9.16 | | ***medIncome*** | -9.58 | | ***studyPerCap*** | -0.94 | | ***MedianAge*** | -0.45 | | ***Geography*** | 0.33 | | ***AvgHouseholdSize*** | 1.10 | | ***PctNoHS18\_24*** | 1.03 | | ***PctHS18\_24*** | 4.20 | | ***BirthRate*** | -1.73 |   The regression model performs better than baseline model. The coefficient table shows that areas with higher median incomes tend to have lower cancer death rates. |
| **3.b. Business Impact** | The negative coefficient for median income indicates that there is indeed a negative relationship between median income and cancer death rate, which is aligned with finding on the first experiment. So, areas with higher median incomes tend to have lower cancer death rates, which aligns with what we would typically expect in terms of access to healthcare, better lifestyle choices, and other socioeconomic factors associated with higher income levels.  Higher study per capita is associated with lower cancer death rates. This suggests that areas with more educational resources or research tend to have lower cancer death rates.  There appears to be a positive relationship between the incidence rate and cancer death rate. This suggests that areas with higher incidence rates tend to have higher cancer death rates.  Overall, the regression model suggests that socioeconomic factors such as income and education, as well as demographic factors like age and birth rate, may play significant roles in determining cancer death rates within a given area. Additionally, incidence rate and study per capita also seem to be important factors.  The above results may provide various business impacts. Healthcare providers and policymakers can use this information to better allocate resources for cancer prevention, screening, and treatment. For example, areas with higher incidence rates and lower median incomes may require additional funding for cancer screening programs or access to affordable treatment options.  Moreover, health insurance companies, pharmaceutical companies, and healthcare providers can use this information to tailor their marketing and outreach efforts. For instance, they can focus on promoting cancer prevention and early detection services to communities with higher incidence rates and lower education levels. |
| **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 complex interplay between socio-economic and demographic factors and cancer death rates, thereby emphasizing the impact of developing strategies to reduce disparities and improve health outcomes for all communities. |
| **4.b. Suggestions / Recommendations** | My suggestion for these parts is to carry out broader research on socio-economic and demographic factors influence cancer death rates. This could involve qualitative research methods such as interviews or focus groups to gather insights from affected communities. We may divide the education areas into smaller subsets, such as Economic education IT education, and to obtain the impact of different areas |