# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Bui The Hai |
| **Project Name** | Part A: Experiment on univariate linear regression |
| **Date** | Mar 30, 2023 |
| **Deliverables** | Assignment 1 Part A Univariate 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 goal of this experiment is to examine the relationship between wealth and cancer death rate (***TARGET\_deathRate***). Therefore, I will choose two variables, namely ***medianIncome*** and ***povertyPercent*** as independent variables and train two univariate linear regression models.  ***medianIncome*** variable represents the median income per US county while ***povertyPercent*** calculates percent of the populace in poverty. As a result, these two independent variables should demonstrate the relationship between wealth and cancer rate.  The results of this study may indicate a potential inequality in healthcare treatment between the rich and the poor. The costs of treatment or standard of living may be the reasons for this imparity. Despite the high fee of cancer treatment, the fee for insurance is more affordable. A reasonable price insurance package with an effective mechanism for people with low incomes can be the solution to shorten the gap in cancer diagnosis and treatment. As a result, micro insurance products, which offer coverage for poor people with little savings, should be promoted. |
| **1.b. Hypothesis** | My hypothesis for this section is that: There is a positive correlation between ***TARGET\_deathRate***and ***medianIncome*** and vice versa, and a negative correlation between ***TARGET\_deathRate*** and ***povertyPercent***.  There are many reasons for the above hypothesis. First, people who are in poverty will not be easily accessible to the public health service. Early diagnosis and aggressive treatment of cancer are important factors in overcoming this disease. Second, poor people often have a poor diet, which is also a potential cause of high cancer rates.  As a result, people living in poverty may face a higher risk of cancer than those with higher incomes. |
| **1.c. Experiment Objective** | It is expected that the coefficient of the regression model of ***TARGET\_deathRate***and ***medianIncome*** will be negative, and vice versa; the regression model of ***TARGET\_deathRate***and ***povertyPercent*** will show a positive coefficient. However, there are circumstances that the two models above do not show clear patterns. |

|  |  |
| --- | --- |
| 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 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)  The second step is to prepare variables for training. I create two subsets of the two data frames, each containing ***TARGET\_deathRate*** (dependent variable) and ***medIncome*** (independent variables). This dataset is used to train my first model. The same method is applied for the training of the second independent variable, namely ***povertyPercent.*** |
| **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 wealth level and cancer death rate, two independent variables named ***medIncome and povertyPercent*** 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 1. Regression result of model 1   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | MSE of baseline | MSE of training | MSE of testing | Coefficients | | Model 1: Regression models for ***TargetDeathRate*** and ***povertyPercent*** | 768.13 | 624.95 | 640.07 | 1.86 | | Model 2: Regression models for ***TargetDeathRate*** and  ***medIncome*** | 768.13 | 631.07 | 618.54 | -0.00096 |   This experience will use Mean Square Error for assessing the performances of linear regression models. The mean of variables serves as a baseline for evaluating. First, as can be seen from the table, MSE of the two modes are smaller than that of baseline. It means that two models achieve improvement compared to the baseline. Second, Model 2 (using ***medIncome***) has a lower testing MSE compared to Model 1 (using ***povertyPercent***). This suggests that Model 2 might be performing slightly better in terms of prediction accuracy on unseen data.  Based on the Coefficient, the model shows the exact result as we predicted. **medIncome** has a negative impact on ***TargetDeathRate***while ***povertyPercent*** and ***TargetDeathRate*** show a positive correlation.  Therefore, it is evident that the wealth of households and cancer death rate are negatively correlated. |
| **3.b. Business Impact** | The regression results agree with the hypothesis stated at the beginning of this experiment. It indicates that impoverished people are more likely to die from cancer than rich people. As mentioned earlier, it is a business opportunity for insurance companies. People with low levels of income will not afford high fee treatment for cancers. As a result, an affordable insurance package will be their priority. |
| **3.c. Encountered Issues** | The experiments only compare two univariate regression models to illustrate the potential impacts of wealth level on cancer death rate. However, income reflects the temporary status of individuals and it may change rapidly over time, which may not immediately reflect on the cancer 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** | Experimental results prove the relationship between wealth and risk of dying from cancer. As mentioned above, this relationship may have many profound and different meanings in many areas of our society. For example, from the government view point, this experiment emphasizes the bad influence of income parity in US society and they need to find a way to tackle these problems. However, dealing with cancer needs the joint efforts of everyone in our society. The participation of entrepreneurs is a good solution when they can balance social responsibility and profit maximization. For example, they can get an edge on their rivalry in business areas like insurance. |
| **4.b. Suggestions / Recommendations** | In my view, this experiment should be carried out globally to obtain a broader picture. When the investigation is conducted in American society, we cannot give an overall conclusion and recommendation from the global perspective.  However, it is good evidence of a link between wealth and cancer mortality. Many experts suggest that early diagnosis is essential in cancer treatment. As suggested above, this is an excellent opportunity for insurance companies. The cost of routine screening and insurance is much cheaper than cancer treatment. Their next step is to conduct more thorough surveys to obtain a broader picture of demography. Insurance companies should evaluate each specific class of customers and offer appropriate product packages. |