# **EXPERIMENT REPORT**

| **Student Name** | Bui The Hai |
| --- | --- |
| **Project Name** | Part C: Experiment on multivariate linear regression with feature engineering |
| **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 using all other factors to test the relation | |
| **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. | |
| **1.c. Experiment Objective** | The objective of this part is to transform the variables by logarithm function and run the regression test. The reason for transforming the data is that the data is not normally distributed. As a results, the log transformation will make it close to normal distribution. | |

| 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 and perform logarithmic transformation | |
| **2.b. Feature Engineering** | In this part I will perform the log transformation of the variables .The reason is that independent variables are skewed. As a result, the logarithm transformation will bring the variables close to normal distribution | |
| **2.c. Modelling** | In this part, I will transform all the independent variables by using logarithm transformation and perform linear regression model | |

| 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** | On this part, after transforming the variables, the model is overfitted. As a result, I will perform regularizations using Lasso, Ridge and Elastic models. The table below is the MSE of the three models.  Table 3. Regression result of model 3   |  | MSE of baseline | MAE of baseline | MSE of model | MAE of model | | --- | --- | --- | --- | --- | | Model 3: Multivariate linear regression | 758.055 | 21.31 |  |  | | Multivariate linear regression  (Using Ridge model) |  |  | 3.0086761285757553e-10 | 1.2567825116756712e-05 | | Multivariate linear regression  (Using Lasso model) |  |  | 1.4225342818338516e-05 | 0.002828996939191898 |   I perform Lasso and Ridge regularization on the data sets. The Lasso model brings all the coefficients of independent variables to 0. However, in Ridge model, by adding some penalty, **medIncome, studyPerCap,avgAnnCount,PctWhite** show negative relationship with TargetDeathRate, while the other show a positive relationship with TargetDeathRate. It is interesting because it shows an inverse relationship between medIncome and TargetDeathRate, which is totally different with the second part of the experiment. | |
| **3.b. Business Impact** | The model results have many implications.it may mean that people with higher levels of income or white people are more prone to cancer. It may has impacts on treatment of cancer when the patients are white people, especially this study is carried out in the US, the country has many white people. | |
| **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 **medIncome, studyPerCap,avgAnnCount,PctWhite** and the target death rate. It is interesting while it shows that white people are more prone to cancers than other races | |
| **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 on different countries to obtain the impact of different areas | |