# **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 C 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 social economic factors. In this part, other factor related to races will be include in the independent variables. |
| **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 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. I perform concatenating the two data frames to one (splitting will be performed in next part). Below is a summary of the dataset.    The second step is to prepare variables for training. I create two subsets 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 independent variables. The reason is that independent variables are skewed. As a result, the ***logarithm transformation*** will bring the variables close to normal distribution. Outlier will be removed to ensure the genetic power of the model. |
| **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 on training set | MAE of model on training set | MSE of model on testing set | MAE of model on testing set | | Model 3: Multivariate linear regression | 527.63 | 18.36 | 300.88 | 13.46 | 300.34 | 13.32 | | Multivariate linear regression  (Using Lasso model) |  |  | 306.14 | 13.62 | 303.12 | 13.43 | | Multivariate linear regression  (Using Ridge model) |  |  | 300.88 | 13.46 | 300.32 | 13.31 | | Multivariate linear regression  (Using Elastic Net model) |  |  | 332.6 | 14.33 | 320.26 | 14.00 |   I perform Lasso and Ridge regularization on the data sets. From this comparison, it appears that the Multivariate Linear Regression model without regularization (baseline) has the highest MSE and MAE on the training set, indicating overfitting. However, on the testing set, the Ridge model and the baseline model have the lowest MSE and MAE, suggesting better generalization performance. The Elastic Net model has the highest MSE and MAE on both training and testing sets, indicating it might not be the most suitable choice for this dataset.  Overall, the Ridge model seems to strike a good balance between bias and variance, offering decent performance on both the training and testing sets.  A graph with blue and orange squares  The above graph demonstrate the coefficients of Ridge Regression model. ***median income*** has the most influence on cancer death rate predictions. This suggests that socio-economic factors, particularly income levels, play a crucial role in determining cancer mortality rates, as higher median incomes are associated with lower cancer death rates.  ***AvgHouseholdSize*** also has a significant influence on cancer death rate predictions.  This suggests that socio-economic factors associated with larger households, such as lower income levels or limited access to healthcare, may contribute to poorer cancer outcomes.  The coefficients for ***PctNoHS18\_24*** and ***PctHS18\_24*** indicate that education levels among young adults correlate with cancer mortality rates. Higher education levels are associated with lower cancer death rates, emphasizing the importance of education in promoting health literacy, awareness of preventive measures, and adherence to healthcare recommendations.  The coefficient for ***studyPerCap*** suggests that areas with higher educational resources or research activities tend to have slightly lower cancer death rates. This implies that access to healthcare resources, including educational institutions and research facilities, can positively impact cancer outcomes by facilitating early detection, innovative treatments, and dissemination of health information.  The coefficient for ***BirthRate*** suggests that areas with higher birth rates tend to have slightly lower cancer death rates. This may reflect the age distribution of the population, with younger populations potentially having lower cancer mortality rates due to different disease prevalence or screening behaviors. |
| **3.b. Business Impact** | The model results have many implications. Higher income levels or larger household sizes may be associated with certain lifestyle choices, access to healthcare, or exposure to environmental risks that can impact cancer risk or mortality. 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 **various socioeconomic factors** and ***the target death rate***. Those factors may affect the living condition, people’s awareness and lifestyles, which may contribute greatly to the cancer risk. |
| **4.b. Suggestions / Recommendations** | My suggestion for these parts is to carry out broader research on the interaction of various economic conditions on the risk of cancer. Broader research allows for a more comprehensive understanding of how different economic conditions interact to influence cancer death rate. |