# **EXPERIMENT REPORT**

| **Student Name** | Tarun Krishnan |
| --- | --- |
| **Project Name** | Assignment 1, Part A |
| **Date** | 13/03/2023 |
| **Deliverables** | Assignment 1, Part A.ipynb  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 for training this experiment is to try and train a machine learning model to determine the univariate correlation between different measurements for the demographics in various counties in the United States and the death rate in those counties due to cancer.  The results obtained could potentially be used to reduce the overall rate of cancer related deaths across various counties, and extrapolated to have a wider effect.  Incorrect results, however, could result in no change, or even more deaths and so must be verified going forward. | |
| **1.b. Hypothesis** | Given the dataset, I would like to examine the correlation between the level of education, employment/income and healthcare in order to determine if an accurate model can be trained to predict the likelihood of cancer deaths in a county.  I believe that there should exist a correlation between the levels of education and income that affect the target variable, and that by visualising the regression models we can better understand how to reduce the death rate. | |
| **1.c. Experiment Objective** | The expected outcome of this experiment is to detail the existence of trends, or lack thereof, between the aforementioned parameters of the dataset as well as ascertain the strength and reliability of the predictors on the rate of cancer deaths.  Additionally, we shall try to build some ideas about variables we can use to train the upcoming multivariate linear regressions. | |

| 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** | A high level observation of the data shows us that for the most part, the dataset is clean.  The only fields with obvious outlier data are those for age and household size.  However, since we can also observe that there exists no clear correlation between the fields and the target variable, we can ignore cleaning them for the sake of including more data points. The code for cleaning them is included if required.  The majority of the dataset may be cleaned by dropping the NA values present in the columns we can visually identify as important and having some trend/relationship with the target variable. | |
| **2.b. Feature Engineering** | The assignment guidelines state to use univariate machine learning, and that feature engineering will be required for Part C. In light of that, I have omitted the stages of feature engineering for this report, however, the code for use will be kept in the notebook for reference. | |
| **2.c. Modelling** | As mentioned above, the scope of this experiment only details univariate linear regression, and so no other models were trained or considered.  The model used in the simple univariate linear regression model from the scikit\_learn library. | |

| 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** | Of all the fields tested against, the score obtained using MSE when training the model on:  TARGET\_deathRate v/s PctBachDeg25\_Over was found to be 659, and that of TARGET\_deathRate v/s PctPublicCoverageAlone was found to be 695.  The performance values are more accurately documented in the jupyter notebook. | |
| **3.b. Business Impact** | The data and analysis performed shows a decreasing rate of death trend for counties who have a population with more bachelor degrees, and an increasing rate of death trend for those with a population who only have public healthcare coverage.  This indicates that increasing the number of people who have bachelor's degrees or private health coverage may in turn reduce the number of cancer related deaths.  Due to the randomness associated with this type of dataset, using univariate linear regression may not be the ideal approach but can still provide a decent indicator. | |
| **3.c. Encountered Issues** | No significant issues were raised during this experiment. A few fields were labelled incorrectly in the Canvas assignments page and had led in turn to a few correctional steps during the research and discovery phases. Additionally, a few fields did need to be cleaned, however, since they were not relevant in terms of correlation, they were safely ignored. | |

| 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** | While the resulting correlation discovered is fairly straightforward, there are other variables that need to be further assessed in combination with the ones studied in this experiment. It would make sense to perform further experiments to discover any underlying connections and correlations.  For the time being, by increasing the percentage of residents in a county who have a bachelor’s degree or by increasing the access to private health care one could potentially decrease the rate of cancer deaths in that county. | |
| **4.b. Suggestions / Recommendations** | The next steps include performing a series of multivariate linear regressions on the dataset to see if any further information or correlation can be deduced. | |