# **EXPERIMENT REPORT**

| **Student Name** | Tarun Krishnan |
| --- | --- |
| **Project Name** | Assignment 1, Part B |
| **Date** | 20/03/2023 |
| **Deliverables** | MLAA Assignment 1, Part B.ipynb  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 goal for training this experiment is to try and train a machine learning model to determine the multivariate 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** | Using the previous experiment as a stepping stone, I believe that there exists a deeper relationship between the target variable and the individual counties’ income groups, healthcare coverage groups and degree of education.  In the previous experiment, these specific groups showed the highest correlation against the target variable, and intuitively, these do make sense as sensitive factors in our predictions. | |
| **1.c. Experiment Objective** | The expected outcome here is to successfully build a multivariate linear regression model that can effectively predict the trends we hope to see among the variables that we have chosen.  Additionally, we can also use the findings to further improve on our model by implementing feature scaling in the next stage of this assignment. | |

| 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** | The dataset itself is extremely clean. The only stages of cleaning that needed to be performed are those of dropping entries where we encountered null values.  Of note is the age field, which would need to be cleaned if this field provided data of use, however, in light of the lack of trends it offers, we have safely ignored it. | |
| **2.b. Feature Engineering** | Since this stage of the assignment too focuses on primarily experimenting with the data without implementing feature engineering, I have decided to leave it to Part C of the assignment to tackle.  Of note, the following features in the dataset will be used to build the multivariate linear regression model:  [ "medIncome", "povertyPercent", "PctUnemployed16\_Over", "PctEmployed16\_Over", "PctHS25\_Over", "PctBachDeg25\_Over", "PctPrivateCoverage", "PctPrivateCoverageAlone", "PctEmpPrivCoverage", "PctPublicCoverage", "PctPublicCoverageAlone", "PctMarriedHouseholds", "PercentMarried" ] | |
| **2.c. Modelling** | The models trained for this regression include the standard LinearRegression model, as well as the Lasso, Ridge and ElasticNet Regression models. While the LinearRegression model remains as before, the Lasso and Ridge models provide us with a degree of regularisation over the base LinearRegression. ElasticNet is unique in that it offers both the regularisations that Lasso and Ridge offer, in one compact package. | |

| 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** | The primary metric for performance was Mean Squared Error, and the scores obtained by each model were as follows;  LinearRegression : 547.7813962647862  LassoRegression : 546.9354304580315  RidgeRegression : 547.7792095883966  ElasticNetRegression : 546.3250861443997 | |
| **3.b. Business Impact** | Plotting a graph between each regression and the corresponding test case shows that the model is fairly accurate towards the middle, however, at either end of edge cases, the accuracy does fall off.  While there does appear to be a decent level of accuracy, these results should be used as a measure for potential changes after further investigation and research. However, if correct these results could dictate a decrease in the rate of cancer related deaths. | |
| **3.c. Encountered Issues** | At present, no new functional issues were faced in working with the dataset or experiments. Potential issues may arise in feature scaling and feature engineering, however, those will have to be dealt with in future experiments. | |

| 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** | The primary takeaway from this experiment is that by increasing the number of variables, we have in turn reduced the overall error and have built better estimations. However, given the spread of the data, reducing its dimensionality may in turn reduce the overall error we perceive and make our results more or less accurate.  Feature scaling may shed more light, however, there may not be a further tangible outcome using linear regression alone. | |
| **4.b. Suggestions / Recommendations** | Going forward, we should implement a set of feature engineering, namely feature scaling to reduce the overall error rate. Given the outcome of that, fitting a different model may make sense to account for the extreme ends of the graphs that linear regression cannot compensate for. | |