# **EXPERIMENT REPORT**

| **Student Name** | Tarun Krishnan |
| --- | --- |
| **Project Name** | Assignment 1, Part C |
| **Date** | 22/03/2023 |
| **Deliverables** | MLAA Assignment 1, Part C.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 linear 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 value in reducing our error rate by rationally reducing the dimensionality of the dataset by applying feature normalisation.  Additionally, I believe that we can develop a new feature for our use case, based on educated populace percentages.  In the previous experiment, specific groups showed the highest correlation against the target variable, and intuitively, reducing our error rate should help fortify our conclusions. | |
| **1.c. Experiment Objective** | It is expected that applying feature scaling will reduce our overall error to the point where our confidence in these metrics will increase.  If proved correct, we could suggest changes in county policies that would in turn drastically reduce the rate of cancer related deaths, and further extrapolation could have massive ramifications. | |

| 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** | Looking at the trends within the dataset, we can see that the education percentage seems to make an impact on the overall death rate. In light of that, feature engineering was applied to generate a measure of overall education percentages across a county, to provide a better metric for training.  It was also believed that the large error rate in the previous experiments was due to the overall spread of the dataset, and by decreasing the dimensionality of measurement, our predictions would become more accurate. | |
| **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 scores obtained were as follows, after scaling:  LinearRegression : 0.7463373849051153  LassoRegression : 0.9999999999999998  RidgeRegression : 0.7463669070649771  ElasticNetRegression : 0.9999999999999998 | |
| **3.b. Business Impact** | Our scaled results prove that our actual error is far lower than the initial experiments showed us. However, given the scale of impact this experiment’s outcome could have, further research and investigation would be required before implementing changes. | |
| **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. | |

| **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** | Reducing the datasets dimensionality has in turn made our predictions more reliable. In light of this, we can more strongly inform the business of our findings and suggest more impactful changes.  Further experimenting with linear regression will not shed more light, and we should switch to more advanced models to predict better results. | |
| **4.b. Suggestions / Recommendations** | Given the outcome seen here, perhaps training more complex and advanced models could shed a better light on the overall situation, and help track even deeper trends that we connect as of yet see. | |