# **EXPERIMENT REPORT Part A - 2**

| **Student Name** | Shalimar Chalhoub |
| --- | --- |
| **Project Name** | Regression Models |
| **Date** | 31/3/20203 |
| **Deliverables** | <MLAA Assignment 1 Part A>  <Univariate Linear Regression>  <Second Feature> |

| 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 Business objective behind this project is to accurately predict cancer mortality based on information related to US counties.  By interpreting the results, the business can be able to know which factors affect the mortality rate due to cancer and can work towards lowering said rate.  Incorrect results could lead the business to overlooking some counties that need urgent help in lowering the mortality rate as the result was inaccurate | |
| **1.b. Hypothesis** | Can PctPublicCoverageAlone be used to predict DeathRate in a county?  PctPublicCoverageAlone is the percentage of county residents with government-provided health coverage alone which could be directly related to the cancer DeathRate. When I did the correlation between values and deathRate, this feature seemed to have a high correlation value and theoretically it does make sense and thus why I’ve decided to explore it | |
| **1.c. Experiment Objective** | I expected to find a linear relationship between PctPublicCoverageAlone and DeathRate and thus be able to have a correct prediction of the latter.  I estimated it to have a relatively low MSE given that in theory, these two variables seem very correlated and after testing, both have a good correlation index | |

| 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** | 1. Removed all values of PctPublicCoverageAlone smaller than 5 or bigger than 27 in order to avoid outliers 2. Removed all values of TARGET\_deathRate smaller than 100 or bigger than 250 in order to avoid outliers 3. Split my train and test data into x\_train and x\_test containing the incidenceRate variable and y\_train and y\_test containing only the target variable   I didn’t clean any missing values because the variables I’m using don’t have any | |
| **2.b. Feature Engineering** |  | |
| **2.c. Modelling** | The model used for this experiment is univariate Linear Regression which consists of building a Linear Regression model using only one variable to predict the target variable. The reason I’ve used this model is because I have been imposed the constraint. | |

| 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 MSE for this model was 507.82 meaning that there is approximately 22.5 units between the actual values and the prediction, and given that I only have a data range of less than 600, this would be considered very high, however, it is better than my first model.  The reason for this high value is that the data was not linear, and given the restraints I had, I could not transform it, thus, it did not fit a linear model | |
| **3.b. Business Impact** | This experiment does not meet the business objectives as there is a big error margin and hence, will not predict accurate results.  Counties with high DeathRate which have been predicted as low, will be overlooked by the business which might result in higher death rates which are quite a few values in my predictions. | |
| **3.c. Encountered Issues** | The main issue is the non-linearity of the data which I could not solve since it I was restricted from using feature engineering  Another issue was the outliers, which I fixed by setting bounds to my data | |

| 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** | This specific experiment did not produce the results I was looking for and the reason for that is that for such big datasets with multiple features, one feature alone would not be enough to predict accurate results.  The better approach would be to do some feature engineering and make the data more linear as well as use more features. As for this method, it is a dead end with the restrictions imposed | |
| **4.b. Suggestions / Recommendations** | Next steps, from least likely to most likely would be:   1. Try another feature 2. Try multivariate linear regression 3. Try linearizing the data and do another type of regression | |