# **EXPERIMENT REPORT Part C**

| **Student Name** | Shalimar Chalhoub |
| --- | --- |
| **Project Name** | Regression Models |
| **Date** | 31/3/2023 |
| **Deliverables** | <MLAA Assignment 1 Part C>  <Lasso Linear Regression>  <Feature Engineering> |

| 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 of this project for the business is to help the business predict deathRates due to cancer in different US counties by the use of a regression model in order to be able to see in which counties they should invest more funds into research and healthcare to lower these rates.  Inaccurate results may cause some counties with high death rates to be overlooked. | |
| **1.b. Hypothesis** | By using the existing features of the data set and adding an extra feature which consists of reconstructing the target variable using the predictor features, we will be able to get an accurate model.  In theory, this model would perform greatly thus why I decided to try it out. | |
| **1.c. Experiment Objective** | I anticipate that the results of my model will be very favorable, with an error rate of less than 10%. This would allow us to utilize the model effectively.  The two outcomes of my model could be   1. Error will be more than 10% and thus unsuitable for this purpose, 2. Error will be less than 10% and we will be able to employ it | |

| 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** | Steps taken to prepare the data   1. Removing all non numeric data 2. Putting the target rate in one dataframe and all other features in another 3. Dropping PctSomeCol18\_24 for having too many missing values and the column Id because it is irrelevant 4. Creating a new feature and adding it to the model which is the average death per year per county because of cancer multiplied by 100k and divided by the population. This was done in hopes of getting a close replica of the target value. 5. Dividing the training set into Train and Dev Datasets.   Steps 1-4 were done on both training and testing datasets  Some of the steps I did not execute are the following   1. I did not set bounds to the target variable because I wanted my model to be able to predict for all the variables 2. I did not drop irrelevant features from my model because I did not see any reason to do so 3. I did not do any feature scaling because if my data is nonlinear, scaling it will not make it linear. And I also don’t have a big difference in my data   More feature engineering can be done for further studies | |
| **2.b. Feature Engineering** | The feature engineering I’ve done is creating a new feature called “new” that takes the avgDeathsPerYear, multiplies it by 100,000 and then divides by popEst2015. By doing this, we will get the average death per year per 100 000 per capita which is a replication of the target variable.  Features I removed consisted of all non numeric features, as well as the Id column because of irrelevancy and then the column PctSomeCol18\_24 because it had too many missing values.  Finding a way to work with incident rate will be a good start to try and engineer a new variable that will assist in making predictions better. The reason why I think that incident rate will be good | |
| **2.c. Modelling** | The model trained for this experiment is the Lasso which was chosen because it is very efficient when it comes to feature selection, regularization and producing sparse solutions. The main reason I decided on Lasso is because it can handle multicollinearity and after testing other models, Lasso performed best.  I decided not to train Ridge or ElasticNet but did not go with them as my data has outliers and Ridge doesn’t deal well with outliers and since my data is correlated, ElasticNet wasn’t producing the best results either.  I also decided not to test KNN because first, it doesn’t do any feature selection, and KNN is based on feature similarity, however, my data is quite independent and not a lot of similarity can be found.  Training a decision tree model might be good for the future | |

| 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** | This model performed the best with an MSE of 166.055 and an RMSE of 12.886 which is less than 10%/ That being said, this model produced good results in my opinion,  The cases that did not perform very well are those where the target is below 120 or above 250 the reason being that these particular variables are scarce and can be counted as outliers. | |
| **3.b. Business Impact** | The results fit the business objective and can help predict accurately cancer mortality rates in counties. With values less than 140, they were either classified less or more, however, those results don’t really matter that much as the county does not have an alarming rate of deaths, however, what causes a significant problem, is high values that are predicted as low, which, my model, unfortunately, has some, but now a lot | |
| **3.c. Encountered Issues** | The issues faced and their solutions are listed below   1. Outliers: I did not solve that issue because I wanted my data to be able to predict on all values, however, I could put some bounds for future experiments 2. MIssing Values: I either removed or fixed them 3. Linearity of data: I created a new variable that had a better linear relationship with my data 4. Scalability of data: I did not scale my data but I could have | |

| 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 experiment presented a good outcome that can be used by the business. Reflecting back on this assignment, I found that feature engineering is a powerful tool that should not be overlooked when performing or building a machine learning model.  This experiment should be explored further on a bigger scale, like statewide or even countrywide. It can be also further explored by providing bounds to the data to remove outliers and make predictions more accurate | |
| **4.b. Suggestions / Recommendations** | .  Potential next steps are to try to do this experiment statewide or worldwide or even try to take this model and apply it to other countries and see which parts of the country need more investments in healthcare and cancer research.  Other additional steps that can be added into this model are below ranked from least important to most important   1. Set bounds to the data (might produce better results but not by much) 2. Try Decision tree models (Might work better) 3. Try more feature engineering( Might work the best, however, I was unable to explore it further than I did due to time constraint) | |