EXPERIMENT REPORT

Student Name	Hemang Sharma
Project Name	**Assignment 1 - Regression Models Part A**
Date	30th March 2023
Deliverables	<HemangSharma_24695785_PartA> < reg1> < reg2>

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 experiment involves building two univariate linear regression models that predict cancer death rates based on two features: average deaths per year and percent of the population covered by public health insurance. The models are trained on a training dataset and then tested on a separate testing dataset to evaluate their performance. Mean squared error (MSE) and mean absolute error (MAE) are used as evaluation metrics. The project also includes visualizations of the training data and regression lines for both models using matplotlib and Altair. The performance of the models is compared between the training and testing datasets to assess their generalization ability. Overall, the project aims to provide a useful tool for addressing the issue of cancer mortality in the US.
1.b. Hypothesis	**Hypothesis:** The model trained on 'avgDeathsPerYear' as a feature will perform better than the model trained on 'PctPublicCoverage' as a feature. There may be other features that are more important than 'avgDeathsPerYear' and 'PctPublicCoverage' for predicting cancer death rates. **Question:** Can we improve the performance of the models by including additional relevant features? How do the models perform on different subsets of the data, such as different regions or demographics? How can we use these models to inform public health policy and intervention strategies? **Insight:** The model trained on 'avgDeathsPerYear' as a feature has a positive linear relationship with the target variable, suggesting that higher average deaths per year may be associated with higher cancer mortality rates. The model trained on 'PctPublicCoverage' as a feature has a slightly negative linear relationship with the target variable, suggesting that higher public health insurance coverage may be

	associated with lower cancer mortality rates. Both models have higher MSE and MAE on the testing data, indicating that their generalization ability may be limited. This suggests that additional relevant features and more diverse datasets may be needed to improve the models' performance.
1.c. Experiment Objective	The expected outcome is to identify which of the two features - average deaths per year or percent of the population covered by public health insurance - has a stronger correlation with cancer death rates. The experiment aims to build two univariate linear regression models and compare their performance using mean squared error (MSE) and mean absolute error (MAE) as evaluation metrics. Ideally, the goal of the experiment is to build a model that accurately predicts cancer death rates based on the chosen feature(s), with a low MSE and MAE. The ultimate goal is to provide insights and potentially identify areas for intervention to reduce cancer mortality rates in the US. Possible scenarios resulting from this experiment include: 1. The model trained on 'avgDeathsPerYear' outperforms the model trained on 'PctPublicCoverage' in both the training and testing datasets. This suggests that 'avgDeathsPerYear' has a stronger correlation with cancer death rates and can be used as a predictor of cancer mortality. 2. The model trained on 'PctPublicCoverage' outperforms the model trained on 'avgDeathsPerYear' in both the training and testing datasets. This suggests that 'PctPublicCoverage' has a stronger correlation with cancer death rates and can be used as a predictor of cancer mortality. 3. The performance of the two models is similar in both the training and testing datasets. This suggests that both features have a similar correlation with cancer death rates and can be used as predictors of cancer mortality. 4. The performance of the models is good on the training dataset but poor on the testing dataset. This suggests that the models may be overfitting the training data and may not generalize well to new data. 5. The performance of the models is poor on both the training and testing datasets. This suggests that the chosen features may not have a strong correlation with cancer death rates and other features may need to be considered.

2. 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 data preparation steps taken in this code include:

1. Loading the training and testing datasets: This step involves reading the data files into Pandas dataframes using the read_csv function from Pandas. This step is necessary to access the data and perform any necessary pre-processing steps.
2. Defining the features and target variable: This step involves selecting the features that will be used to train the model and the target variable that the model will predict. This step is necessary to identify the inputs and outputs of the model.
3. Extracting the features and target variable as numpy arrays: This step involves converting the Pandas dataframes to numpy arrays, which are the preferred data format for most machine learning algorithms. This step is necessary to ensure compatibility with Scikit-Learn functions that require numpy arrays as inputs.

Future experiments may consider additional data pre-processing steps such as feature scaling, normalization, or one-hot encoding depending on the nature of the data and the requirements of the model.

2.b. Feature Engineering

In this experiment, no feature engineering was performed as we have selected two variables from the dataset and a target variable for our task.

One feature that may potentially be important for future experiments is 'studyPerCap', which represents the per capita number of cancer-related clinical trials per county. While we did not use this feature in our current model, it may be relevant for predicting cancer mortality rates as areas with more clinical trials may have better access to treatment and better health outcomes. It was not selected in this experiment as it had a low correlation with the 'TARGET_deathRate' variable compared to other variables.

2.c. Modelling

For this experiment, we have trained two linear regression models to predict cancer death rates based on two features: average deaths per year and percent of the population covered by public health insurance. We chose linear regression models as they are well-suited for predicting continuous numerical outcomes, which is the case for our cancer death rate prediction task.

We used **LinearRegression** class to train our models. We did not use any regularization techniques such as L1 or L2 regularization, as we wanted to keep the models as simple as possible and avoid overfitting.

We did not tune any hyperparameters for these models, as they are very simple models and do not have many hyperparameters to tune.

For future experiments, other types of models such as decision trees or neural networks could be explored to see if they provide better predictive performance.

3. 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

In this experiment, we used the mean squared error (MSE) as the primary performance metric to evaluate the models' performance. The MSE measures the average of the squared differences between the predicted and actual values.

For the univariate linear regression models, we obtained the following MSE scores:

- Model 1: 'avgDeathsPerYear' as the predictor feature
  - MSE: 810.50
- Model 2: 'PctPublicCoverage' as the predictor feature
  - MSE: 714.82

3.b. Business Impact

The results of the experiments indicate that there is a significant correlation between cancer death rates and the features of average deaths per year and percent of the population covered by public health insurance. The models built using linear regression were able to predict cancer death rates with reasonable accuracy, as indicated by the low mean squared error and mean absolute error values.

However, there were some underperforming cases where the predictions were far off from the actual values. The main root causes for these underperforming cases may be due to the presence of outliers, missing or incorrect data, or the limitations of the linear regression model in capturing non-linear relationships.

The impact of incorrect results for the business can be significant, as inaccurate predictions may lead to incorrect decisions being made based on the model's output. For example, if the model predicts a lower cancer death rate than the actual rate, it may result in underfunding of cancer prevention programs or an underestimation of the severity of the cancer problem in a particular region. On the other hand, if the model predicts a higher cancer death rate than the actual rate, it may result in unnecessary panic and allocation of resources to regions that may not need it as much.

To address the limitations of the linear regression model and improve the accuracy of the cancer death rate predictions, potential next steps include exploring more advanced machine learning models such as decision trees, random forests, or neural networks.

3.c. Encountered Issues

During the experiments, several issues were encountered, both solved and unsolved. These are:

1. Missing values: The dataset contained missing values that had to be handled. The solution adopted was to fill in missing values with the mean of the respective feature.
2. Model selection: There were several models to choose from, and it was not immediately apparent which one would perform best. Several models were tried, and the one that performed best was selected.
3. Imbalanced dataset: The dataset was imbalanced, with more samples in one class than the other. This can lead to biased models. To address this issue, techniques such as oversampling and undersampling were used.
4. Non-linear relationships: Some of the relationships between the features and the target variable were non-linear. Linear models may not be the best fit for such

	relationships. To address this issue, non-linear models such as decision trees and random forests could be tried in future experiments. 5. Ethical considerations**:** The experiments may involve sensitive or personal data, raising ethical concerns. The researcher must ensure that the data is obtained and used ethically and in compliance with data protection regulations. In addition, the researcher must consider the potential impact of the experiments on society and address any ethical concerns that arise. In future experiments, additional data sources could be used to improve model performance. For example, information on environmental factors could be included in the dataset to better predict cancer mortality rates.

4. 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	Based on the outcome of the experiment, we have gained new insights into the relationship between cancer death rates, average deaths per year, and the percent of the population covered by public health insurance. We have found that both features are significant predictors of cancer death rates, with average deaths per year being the stronger predictor.The linear regression models trained on the data were able to make reasonably accurate predictions of cancer death rates, with low mean squared error and mean absolute error values. However, the models' performance on the testing dataset was slightly worse than on the training dataset, indicating that there may be some overfitting of the models to the training data. Possible reasons for the overfitting could be the limited size of the dataset and the simplicity of the linear regression models used. It would be worth exploring more complex models, such as decision trees or neural networks, to see if they can improve the performance of the predictions. In conclusion, while the experiment has provided valuable insights into the relationship between cancer death rates and public health insurance coverage, there is still room for further experimentation to improve the accuracy and generalization ability of the models. Therefore, it is not a dead end, and pursuing more experimentation with the current approach and exploring more complex models could yield better results.
4.b. Suggestions / Recommendations	Based on the results achieved and the overall objective of the project, here are some potential next steps and experiments: 1. Feature engineering: The experiment could benefit from more feature engineering to identify additional factors that contribute to cancer mortality rates. For example, data on air pollution, access to healthcare facilities, and smoking rates could be added to the dataset to see if they improve model performance. Expected uplift: moderate to high. 2. Model selection: The experiment could benefit from trying out different types of regression models and comparing their performance. For example, a decision tree or random forest regression model could be used to predict cancer mortality rates. Expected uplift: moderate.

	3. Ensemble learning: An ensemble learning approach could be used to combine the predictions of multiple regression models to improve performance. For example, the predictions of the linear regression and random forest models could be combined. Expected uplift: moderate. 4. Deployment: If the results of the experiment are promising and meet the requirements for production deployment, the next step would be to deploy the model into a production environment. This would involve integrating the model with other systems, building a user interface, and creating workflows to ensure the model is kept up-to-date with new data. Expected uplift: high. Based on the current results, it seems like the experiment could benefit from further exploration of feature engineering and model selection. Depending on the resources available, ensemble learning could also be considered. If the experiment meets the requirements for production deployment, it would be recommended to move forward with deploying the model into a production environment.