EXPERIMENT REPORT

Name	Monali Patil
Project Name	Cancer-Mortality-Prediction
Deliverables	Model name: Multivariate Linear Regression Notebook name: MachineLearning_Multivariate_LinearRegression_PartB.ipynb Project Repo: https://github.com/monalippatil/MachineLearning-Cancer-Mortality-Prediction.git

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

Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?

This project involves working with a consolidated dataset compiled from census data in the USA to build a regression model. Through data exploration, cleaning, and appropriate feature and model selection, the objective is to create a model that accurately predicts the mortality rate on unseen data. The resulting model can be used to gain insights into the relationships between different variables, and factors that influence cancer mortality rates, and help healthcare professionals (public/private), policymakers, and researchers develop more effective strategies for reducing cancer mortality rates and implement informed decision-making processes improving health results and services.

Accurate results from the model can help identify high-risk areas and populations that require targeted interventions, leading to the development of more effective prevention and treatment strategies and ultimately improving health outcomes for individuals affected by cancer.

Incorrect results from the model could lead to misguided policy decisions and ineffective interventions. Inaccurate predictions could also undermine public trust in the model and the data it relies on, leading to hesitation and resistance to interventions based on the model's findings.

1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: Selecting multiple continuous features with significant correlation with the target variable and analyse its performance. • Upon analyzing the heatmap and identifying the features with the most significant correlation with target 'TARGET_deathRate' attribute, it appears that among all the numerical characteristics, 'incidenceRate' and 'povertyPercent' are correlated with a coefficient of 0.44. Additionally, 'PctPublicCoverageAlone' has a correlation of 0.47, followed by 'PctPublicCoverage' with 0.42, 'PctUnemployed16_Over' with 0.41, and 'PctHS25_Over' with 0.4. • Based on the pairplot chart from the notebook, it seems that all of the selected continuous variables have a relatively linear correlation with the 'TARGET_deathRate' target variable. If a straight line were to be drawn for these variables, it would reasonably fit the data points. • Since the selected characteristics have a more significant correlation with the target variable than the other variables, choosing them as multiple features for our multivariate linear regression.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The expected outcome of this experiment is that the model will fit a linear equation to the data that best describes the relationship between selected features and cancer mortality rate in the dataset. The goal of this experiment is to build the multivariate linear regression model to make predictions about the cancer mortality rate based on the selected features and the model performs well and generalizing well to new data. If the model performs well, it will be able to accurately predict cancer mortality rates based on the new and unseen selected features.

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

Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments.

The process of exploring the data and cleaning it up was done in preparation for using the algorithm.

• Data Understanding

1] Loading Data: Read the data from cvs files into the pandas dataframe for usage and develop a model.

2] Exploring Data:

- df.head(): For checking some datapoints of the dataset.
- df.sample(): For checking sample/any 5 rows of the dataset.
- df.shape: For describing the dataset's dimension/number of rows and columns.
- df.columns: For checking attribute names of the dataset.
- df.info(): For checking attributes summary information(features datatypes) of the dataset.
- df.describe(): For checking the summary statistics of the dataset.
- df.describe(include='all'): For describing summary statistics for all datatype variables of the dataset.
- df.isnull().sum(): To identify if any null values in the dataset.

Since machine learning linear algorithms cannot handle missing values in continuous/any attributes, need to either handle these missing values or remove the columns altogether.

- To drop the null values used drop() function.
  For instance, `df_train.drop(['PctSomeCol18_24'], axis=1, inplace=True)`
- To fill the null values with any data, the mean value is usually substituted as it is an average of the ranges of values present in the attribute.
  df.isna().sum(): To view if any missing values in the dataset.
- To check if any missing values, generate the heaatmap with below syntax.
  `sns.heatmap(df_train.isnull(), yticklabels=False, cbar=True)`

3] Analysing the Relationship between Target and Independent Feature

- To find the relationship between the 'TARGET_deathRate' target variable and other different attributes, plotted Heatmap for Correlation.
  For instance, `sns.heatmap(df_train.corr(), annot=True, cmap='summer_r')`

As higher the value more correlated the two variables are, for attribute selection correlation heatmaps are employed.

- To check the distribution of the chosen numerical features below histograms are used.
  `histogram = df_continuous_features.hist(figsize=(10,10))`
- For examining outliers present in the selected continuous features below boxplots are used.
  `sns.boxplot(x='incidenceRate', data=df_continuous_features, ax=axes[0], color='orange')`
- For checking the data distribution of the independent continuous features with the target 'TARGET_deathRate' variable below pairplots are used.
  `sns.pairplot(df_train, x_vars=['incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone'], y_vars='TARGET_deathRate', height=5)`

- Data Preparation

4] Selecting Target and Independent Feature.

The target 'TARGET_deathRate' feature and selected feature are separated for both training and testing set to build a model.

```
        X = df_train['incidenceRate'].values
        y = df_train['TARGET_deathRate'].values
```

5] Splitting Data into Training, Validation and Testing Sets.

By creating a validation set, it gives freedom to conduct several experiments, as multiple experiments can be run on model using this set. On the other hand, the testing set should only be used a few times. Therefore, splitting the training set (ratio of 90:10) into a validation set, to leverage more flexibility for experimentation. The splitting was performed using train_test_split class from sklearn library.

```
        X_train,  X_validate,  y_train,  y_validate  =  train_test_split(X,  y,
test_size = 0.1, random_state = 19)
```

2.b. Feature Engineering

Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments.

Not in scope of the Part B of the assignment.

2.c. Modelling

Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

The model trained for this experiment is Multivariate linear regression model choosing appropriate independent features to predict cancer mortality rate – 'TARGET_deathRate' target variable.

Although Multivariate Linear Regression is used for assignment objectives, it is an appropriate choice because the target variable, "TARGET_deathRate," and all the chosen independent variables are continuous and can potentially have infinite values. Therefore, Multivariate Linear Regression is a suitable model choice.

Because of time limitations, it was not possible to train regularization algorithms such as Lasso, Ridge, and Elasticnet, and adjust the "fit_intercept" hyperparameter of the model using the StandardScaler method from the sklearn library and would build and test these algorithms in my future experiments.

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

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

Note: The MSE calculates the average of the squared differences between the predicted and actual values. Since the unit of values are doubled, the error is emphasised. Therefore, to mitigate this effect, the RMSE is used for evaluation below, which cancels out the squaring effect and brings the unit of measurement back to its original scale.

- Part B: Experiment 1

Target Variable: 'TARGET_deathRate'
Independent Variables: ('incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage')
Algorithm: Multivariate Linear Regression.
Performance Metrics:

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	397.0357	14.8642	19.9257
Validation	367.6438	14.672	19.174
Testing	453.8926	15.758	21.3047

- Part B: Experiment 2

Target Variable: 'TARGET_deathRate'
Independent Variables: All continuous features except ‘ID’, ‘binnedInc’ and ‘Geography’ as these are object/string/categorical features.
Algorithm: Multivariate Linear Regression.
Performance Metrics:

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	343.6132	13.922	18.5368
Validation	482.1358	14.9528	21.9575
Testing	417.823	15.0174	20.4407

Note: The 'PctSomeCol18_24' feature denotes the percentage of individuals aged between 18 and 24 years who have attained some college education. However, since more than 70% of the records have missing values in this column (only 612 were recorded out of 2438), it is impractical to fill them with mean values. Therefore, removing this feature from both datasets.

* And the difference between the RMSE scores of both experiment 1 and 2 for all three sets are relatively small, and it suggest that the model is slightly overfitting to some extent on the training data in both experiments as there be some datapoints that are specific, and model couldn't enough be generalised to predict unseen observation. Moreover, the model from experiment 1 is performing slightly well than experiment 2, on the testing data compared to the training data.

- Experiment 1: RMSE scores of the training (19.9257), validation (19.1740), and testing (21.3047).
- Experiment 2: RMSE scores of the training (18.5368), validation (21.9575), and testing (20.4407).

* Therefore, it seems that in the experiment 1, selecting multiple features based on correlation was a reasonable approach for the model to perform well and generalize better on unseen data compared to using all the features in experiment 2.

* Additionally, to determine to what extent the model is overfitting and improve its generalization performance, it would be helpful to evaluate the model on other metrics and explore other approaches such as adjusting the model hyperparameters, regularizing the model to reduce its complexity and prevent overfitting.

* The line charts from both experiments informs that some data points are distant from the line, which indicates a need for further analysis and validation of those observations. A different approach, such as manual intervention, may be necessary when predicting those data points.

3.b. Business Impact

Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)

From these two experiments it appears that selecting multiple independent features based on correlation value shows better performance than training a model with all continuous variables and suggest it's reasonably and is a informed selection of predictor attributes.

However, the model is not sufficiently generalized, and there are still features with observations that are specific, which can lead to overfitting. Therefore, it is necessary to address this issue from a business perspective, considering their performance objectives and benchmarks.

3.c. Encountered Issues

List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.

Employed below statement as MSE up until later stages of the learning for this assignment, and then corrected it realising the values of MSE and MAS are relatively close.

```
mse(y_validate, y_validate_predict, squared=False)
```

Also, after going through my lecture notes, learned that a lower MSE score is typically preferred, as it suggests that the model's predictions are more accurate and closer to the actual values. However, the interpretation of this metric depends on the specific problem being addressed and the scale of the target variable, and it places emphasis on the error.

In future experiments, we will consider addressing the issue of overfitting in the model by working on the hyperparameters. However, due to time constraints, we cannot pursue this approach currently.

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

Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.

Selecting multiple features based on correlation was a reasonable approach for the model to perform well and generalize better on unseen data compared to using all the features in the later experiment.

Even though the multiple feature selection was carried out based on correlation value with the target variable, the model is slightly overfitting and needs further examination at other metrics and perform additional analyses, such as examining the residuals, cross-validation, and exploring the model's coefficients.

As stated in the Technical Performance section of 3.a, there is an opportunity for additional evaluation and analysis. I would try to test this model multivariate linear regression by regularising with hyperparameters to ensure that the model could be further generalised.

4.b. Suggestions / Recommendations

Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.

I would mainly carry out these two tasks.

* Employing multivariate linear regression with hyperparameter.
* Develop and evaluate multivariate linear regression with data from different fields recognising their applications.

To deploy a linear regression algorithm successfully in the production environment, it is recommended to scale and transform the model to handle big datasets, choose a suitable deployment environment cloud or on-premises, and transform the model for production settings ensuring compliance to various security, ethical and privacy guidelines, conduct testing and monitoring, update the model periodically with new data, and provide documentation for usage and review the model performance and retrain accordingly.