EXPERIMENT REPORT

Name	Monali Patil
Project Name	Cancer-Mortality-Prediction
Deliverables	Model name: Univariate Linear Regression Notebook name: MachineLearning_Univariate_LinearRegression_PartA-I.ipynb MachineLearning_Univariate_LinearRegression_PartA-II.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: Generally speaking, the greater number of cancer cases that are recorded and diagnosed will result in a more positive interpretation of the mortality rate.

- After examining the heatmap and identifying the attributes with the strongest connection, it appeared that of all numerical features, 'incidenceRate' and 'povertyPercent' have a correlation of 0.44, while 'PctPublicCoverageAlone' has a correlation of 0.47.

- Based on the pairplot graph the data distribution of the 'incidenceRate' feature in relation with target 'TARGET_deathRate' variable showed relatively better linear relationship than other independent candidate features. Also, if we were to draw a straight line for 'incidenceRate' attribute against the target variable, it would fit more closely to its data points compared to the other two features.

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 cancer incidence rate and cancer mortality rate in the dataset.

The goal of this experiment is to build the univariate linear regression model to make predictions about the cancer mortality rate based on the incidence rate 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 incidence rate.

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 attributes 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 any missing values in the dataset.
- To check if any missing values, generate the heatmap 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 A 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 Univariate linear regression model choosing any single appropriate independent feature to predict cancer mortality rate – ‘TARGET_deathRate’ target variable.

Although Univariate Linear Regression is being used for learning purposes, it is appropriate because both the target variable, 'TARGET_deathRate,' and independent variable, 'incidenceRate,' are continuous and have the potential to take upto infinite values. Therefore, Univariate Linear Regression is a suitable choice of model.

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.

	Further regularising the mode with fit_intercept hyperparameter would be appropriate to generalised the model to predict unseen data and as it was explained in the lectures so would like to experiment on it.

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.

- Experiment: Part A - I

Target Variable: 'TARGET deathrate'
Independent Variable: 'incidence Rate'
Algorithm: Univariate Linear Regression.
Performance Metrics:

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	612.89	19.3498	24.7566
Validation	614.1857	19.2027	24.78277
Testing	622.488	19.7412	24.9497

* As the difference between the RMSE scores of the training (24.7566), validation (24.7827), and testing (24.9497) sets are relatively small, it suggests that the model is performing consistently across the three sets and is not overfitting to the training data and this univariate linear regression model with single 'incidenceRate' independent feature is generalizing well to new/unseen data.

* However, further it is important to evaluate other metrics and perform additional analyses such as examining the cross-validation and comparing the model's performance to other models and with expected business benchmarks before considering it to deploy in the operational environment.

- Experiment: Part A - II

Target Variable: 'TARGET_deathRate'
Independent Variable: ‘PctPublicCoverageAlone’
Algorithm: Univariate Linear Regression.
Performance Metrics:

Dataset	MSE	MAE	RMSE
Baseline	744.585	21.1836	27.287
Training	579.9413	18.3111	24.0819
Validation	719.2265	19.9816	26.8184
Testing	698.2437	19.2253	26.4243

* The difference between the RMSE scores of the training (24.0819), validation (26.8184), and testing (26.4243) sets is small and it appears that the model's performance is decent but not ideal. Therefore, it appears relying solely on the 'PctPublicCoverageAlone' independent feature is inadequate to make precise predictions about cancer mortality rates.

* Additionally, the difference between the training and validation/testing MSE scores suggests that the model is marginally overfitting to the training data as some of the datapoints might be too specific and may be penalised and that the model's performance could be improved by reducing overfitting, using regularization techniques or adjusting the model's hyperparameters. Therefore, it is not advisable to implement this model in the operational setting if it’s not fulling the business performance or benchmark.

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)

Among these experiments, it was found that using the "incidenceRate" feature in univariate linear regression was better at predicting cancer mortality rates compared to the other features. This model is relatively more generalized and can effectively predict new, unseen data.

Therefore, the higher the incidence of cancer is recorded, the better the estimation of the mortality rate due to cancer. However, before taking further steps, it is important to analyze the business objective and benchmark accordingly.

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.

While building univariate linear regression it is important to understand difference between below. As I have used `y_base = np.full((2194,1), y_train_mean),` I encountered error and later could resolve through `y_base = np.full((2194,), y_train_mean)`

```
y_base = np.full((2194,), y_train_mean)
y_base = np.full((2194,1), y_train_mean)
```

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.

Despite the fact that the "PctPublicCoverageAlone" feature had a higher correlation than the "incidenceRate" feature, its performance score was lower than that of the latter.

	Therefore, when conducting machine learning experiments to assess model performance and determine whether it is worthwhile to continue analysing it, the business context and perspective are of utmost importance.to evaluate the model performance and to consider worth for further analysis. As stated in the Technical Performance section of 3.a, there is an opportunity for additional evaluation and analysis. I would like to perform univariate linear regression on each feature that has a significantly higher correlation with the target variable and then with rest to assess all features and confirm its correlation value.
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.** The model from first experiment performed consistently for all the 3 sets and is generalised fairly to predict cancer mortality rate, whereas the model from second experiment is sightly overfitting the training set. Primarily, would like to perform these two tasks. * Develop Univariate Linear Regression on all features to ensure its relatability with the target variable. * Regularise this univariate linear regression algorithm and review if its performance is increasing. To deploy a linear regression algorithm successfully in the operational settings, it is recommended to scale and transform the model to handle big datasets, choose a suitable deployment environment cloud or on-premises, 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.