EXPERIMENT REPORT

Name	Monali Patil
Project Name	Cancer-Mortality-Prediction
Deliverables	Model name: Multivariate Regression with Lasso, Ridge, Elasticnet, KNN with 50 neighbours Notebook name: MachineLearning_Multivariate_(Various)Lasso_Ridge_Elastinet_KNN_Regression_PartC.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 using multiple features with feature engineering or utilizing any algorithm 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: Based on the multivariate feature selection from Part B, building a multivariate linear regressing model with Lasso, Ridg, Elasticnet, KNN with 50 neighbours and Euclidean Distance algorithms to check if the model can be regularised with feature engineered information from ‘binnedInc’ attribute and evaluating the models performance. • Selecting multiple continuous features that have a considerable correlation with the target variable and evaluating their performance using various algorithms. • The 'binnedInc' feature, which is created by binning the median income per capita by decile from 2013 census predictions, has the potential to take on an infinite number of values. So, it was included in the analysis to determine if it has a significant impact on predicting the mortality rate.
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 with feature engineering and employing different algorithms to review their performance to make predictions about the cancer mortality rate based on the selected features and if 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 predictor 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 the dataset.
- df.describe(include='all'): For describing summary statistics for all datatype variables of the dataset.
- df.isnull().sum(): To identify 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)
```

6] Feature Engineering.

Please refer to below section 2.b.

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.

Feature engineering information from ‘binnedInc’ attribute. That is extracting new datapoints from the existing attribute.

Below steps are carried out to extract new observations.

```
# Removing beginning and tailing brackets '()' and '[]' from the 'binnedInc'
feature from training set.
df_train['binnedInc'] = df_train['binnedInc'].str.strip('[]').astype(str)
df_train['binnedInc'] = df_train['binnedInc'].str.strip('()').astype(str)

# Splitting the two set of information into two new columns 'binnedInc1' and
'binnedInc2' from training set.
df_train[['binnedInc1','binnedInc2']] =
df_train['binnedInc'].str.split(',',expand=True)

# Removing beginning and tailing brackets '()' and '[]' from the 'binnedInc'
feature from testing set.
df_test['binnedInc'] = df_test['binnedInc'].str.strip('[]').astype(str)
df_test['binnedInc'] = df_test['binnedInc'].str.strip('()').astype(str)

# Splitting the two set of information into two new columns 'binnedInc1' and
'binnedInc2' from testing set.
df_test[['binnedInc1','binnedInc2']] =
df_test['binnedInc'].str.split(',',expand=True)
```

Note: Since we are training the model with the feature-engineered attribute, it is necessary to include it in the testing set as well. Therefore, performing the above steps on both the training and testing sets.

- The aim of including the 'binnedInc' feature, which is derived by categorizing the median income per capita by decile from the 2013 census predictions, in the analysis is to evaluate its effect on predicting the mortality rate.
- Also, this attribute is continuous and has the capacity to assume infinite number of values so it is suitable for supervised multivariate linear regression algorithm.

Geography attribute can also be featured engineered but given the time constraints will perform it in future experiments.

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

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.

Based on the multivariate feature selection, building a multivariate linear regressing model with Lasso, Ridge Elasticnet, KNN with 50 neighbours and Euclidean Distance algorithms to check if the model can be regularised with feature engineered information from 'binnedInc' attribute and evaluating the models performance with different algorithm. Utilising inbuilt algorithms parameters and not specifying hyperparameters with given time constraints.

Because of time limitations, it was not possible to include analysing the model's precision, recall, F1 score, accuracy, and/or ROC curves. Also, would like use different approaches to examine the model's predictions visually and compare them to the ground truth.

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.

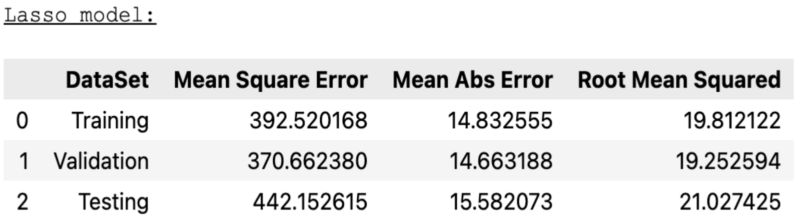
3.a. Technical Performance

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 C: Experiment 1

Target Variable: 'TARGET_deathRate'
Independent Variables: ('incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2')
Algorithm: Lasso Linear Regression.
Performance Metrics:

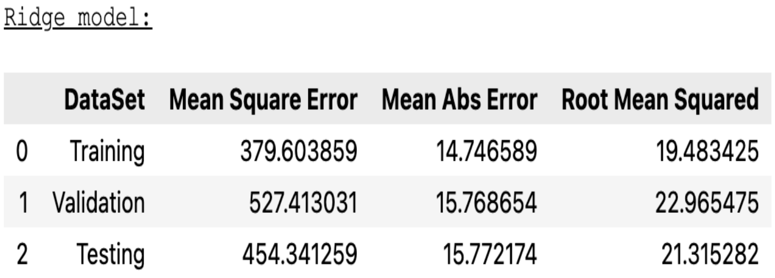


Lasso model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	392.520168	14.832555	19.812122
1	Validation	370.662380	14.663188	19.252594
2	Testing	442.152615	15.582073	21.027425

- Part C: Experiment 2

Target Variable: 'TARGET_deathRate'
Independent Variables: ('incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2')
Algorithm: Ridge Linear Regression.
Performance Metrics:



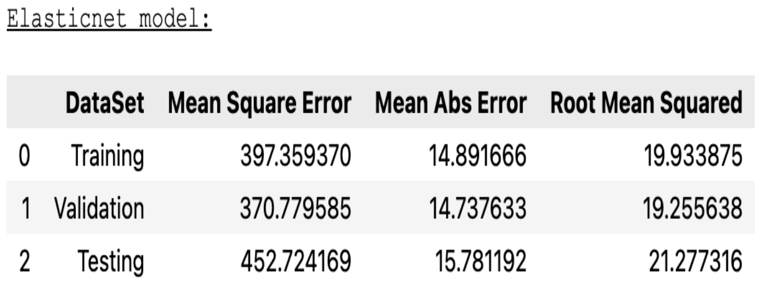
Ridge model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	379.603859	14.746589	19.483425
1	Validation	527.413031	15.768654	22.965475
2	Testing	454.341259	15.772174	21.315282

- Part C: Experiment 3

Target Variable: 'TARGET_deathRate'
Independent Variables: ('incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2')

Algorithm: Elasticnet Linear Regression.
Performance Metrics:

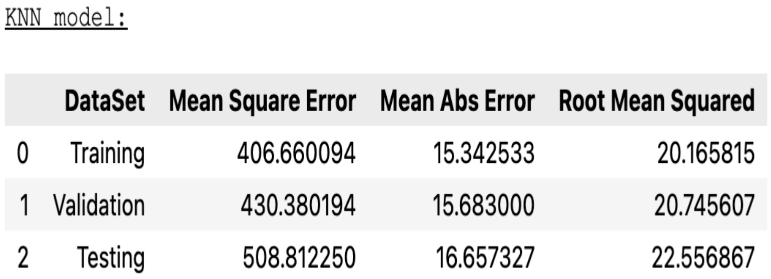


Elasticnet model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	397.359370	14.891666	19.933875
1	Validation	370.779585	14.737633	19.255638
2	Testing	452.724169	15.781192	21.277316

- Part C: Experiment 4

Target Variable: 'TARGET_deathRate'
Independent Variables: ('incidenceRate', 'povertyPercent', 'PctPublicCoverageAlone', 'PctHS25_Over', 'PctUnemployed16_Over', 'PctPublicCoverage', 'binnedInc1', 'binnedInc2')
Algorithm: KNN with 50 neighbours and Euclidean Distance.
Performance Metrics:



KNN model:

	DataSet	Mean Square Error	Mean Abs Error	Root Mean Squared
0	Training	406.660094	15.342533	20.165815
1	Validation	430.380194	15.683000	20.745607
2	Testing	508.812250	16.657327	22.556867

* The difference between the RMSE scores in comparison with multivariate linear regression from part B informs that experiment 1 (from this Part C of assignment) which used Lasso algorithm and experiment 3 that used ElasticNet has improved slightly performance score (RMSE), while the other two experiments using Ridge and KNN with 50 neighbors and Euclidean Distance have only shown a negligible impact.

- Note: Part B Experiment 1: MSE scores of the training (19.9257), validation (19.1740), and testing (21.3047).

* Moreover, to gauge the level of overfitting in the model and enhance its ability to generalize, it is significant to further consult with business and understand the outliers and observations that are far from line from the graphs above.

* To get a more comprehensive understanding of the model's performance, additional evaluation metrics and techniques could be employed.

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 experiments it appears that selecting multiple independent features with featuring engineering is slightly contributing to the prediction and better performance for some of the algorithms namely Lasso() and ElasticNet().

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.

Upon further investigation, it was discovered that the entire dataset was split into 10 bins/sections, but it was unclear if it was based on ‘medianIncome’ or any other feature. Realizing this late in the process of assignment, couldn’t take any different approach with the time constraint.

Also, since the data is aggregated and not actual values, it is challenging to obtain precise insights on the overall numbers and would also consider gaining a deeper understanding of the information contained in the data points.

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 employing multivariate linear regression with different algorithms and including feature-engineered additional information the model is slightly overfitting and still has some specific data points that needed to be analysed.

And as mentioned in the Technical Performance section of 3.a, to gauge the level of overfitting in the model and enhance its ability to generalize, it is significant to further consult with the business and understand the outliers and observations that are far from line with the graphs from python notebook.

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.

For this part, would consider working on the below.

* Analysing the model's precision, recall, F1 score, accuracy, and/or ROC curves.
* Use different approaches to examine the model's predictions visually and compare them to the ground truth.

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.