EXPERIMENT REPORT

Student Name	Hemang Sharma
Project Name	**Assignment 1 - Regression Models Part B**
Date	30th March 2023
Deliverables	< HemangSharma_24695785_PartB.ipynb> < reg > < 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 project for the business is to build a model that accurately predicts the target variable, i.e., the cancer death rate, based on various demographic and socioeconomic factors. The results of this project can be used to identify areas where cancer death rates are higher than expected and to target interventions and resources to those areas to improve health outcomes. Accurate results can lead to improved public health outcomes by identifying at-risk populations and providing targeted interventions to reduce cancer death rates. On the other hand, incorrect results can lead to misallocation of resources and missed opportunities to improve public health outcomes, which can have negative impacts on the population's health and well-being. Therefore, it is important to ensure that the model is accurate and reliable before using it to make decisions that affect public health.
1.b. Hypothesis	Hypothesis: The inclusion of all features in the multivariate linear regression model may not necessarily lead to the best performance, and a subset of the features may provide a better model. Question: Can a model with a subset of features outperform the model with all the features in terms of prediction accuracy for cancer incidence rates? Insight: By comparing the performance of the two models, we can determine whether including all features in a model is necessary or if a model with fewer features can provide similar or better performance. This information can be valuable for researchers and policymakers who are interested in identifying key predictors of cancer incidence rates and developing targeted interventions. Additionally, if a model with fewer features

	performs better, it can reduce the complexity and cost of data collection and analysis.
1.c. Experiment Objective	Expected outcome: The expected outcome of this experiment is to compare the performance of two multivariate linear regression models, one with all the features and the other with a subset of features. The goal of this experiment is to determine if using a subset of features can improve the performance of the model without sacrificing too much accuracy. Possible scenarios: 1. The first model with all features outperforms the second model with fewer features in terms of MSE and MAE for both training and test data. This would suggest that using all features is necessary for predicting the target variable accurately, and using a subset of features would result in a less accurate model. 2. The second model with fewer features outperforms the first model with all features in terms of MSE and MAE for both training and test data. This would suggest that using a subset of features is sufficient for predicting the target variable accurately, and using all features is unnecessary and may even introduce noise into the model. 3. The second model with fewer features performs similarly to the first model with all features in terms of MSE and MAE for both training and test data. This would suggest that using a subset of features can achieve similar performance as using all features, and using a subset of features may be preferable due to its simplicity and reduced complexity. 4. Both models perform poorly in terms of MSE and MAE for both training and test data. This would suggest that linear regression may not be a suitable model for this dataset, and more sophisticated models or feature engineering may be required to improve the performance.

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.
2.b. Feature Engineering	In this experiment, no feature engineering was performed as we used all available numeric features for our model. However, we did remove some features that were not relevant to our prediction task, such as 'Geography', 'binnedInc' and 'popEst2015'. These features did not add any significant predictive value to our model and removing them helped simplify our model and reduce the risk of overfitting. variables.
2.c. Modelling	For this experiment, we have trained two multivariate linear regression models. The first model includes all the numeric features available in the dataset, while the second model includes a subset of the features. We have chosen multivariate linear regression because it is a simple yet powerful algorithm for predicting numerical values based on a set of input features. For the first model, we have not tuned any hyperparameters as we are using the default values provided by scikit-learn's LinearRegression function. For the second model, we have selected a subset of the features to reduce the number of input variables and potentially improve the model's performance. The hyperparameters were not tuned for this model as well. In terms of hyperparameters that may be important for future experiments, the regularization parameter (alpha) in linear regression could be tuned to potentially improve the model's performance. Additionally, feature selection techniques such as Lasso or Ridge regression could be applied to select the most important features for the model.

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

For the first multivariate linear regression model, the baseline mean squared error is 601.8066308092469, and the model's mean squared error is 450.9372592193991. The mean absolute error of the model is 16.357179639018987. The coefficients and intercept of the model suggest that the most significant factors in predicting the target variable are 'medIncome,' 'PctSomeCol18_24,' 'PctEmployed16_Over,' 'PctPrivateCoverage,' and 'PctMarriedHouseholds.' The scatter plot shows that the predicted values of the test data align closely with the actual values, indicating that the model performs well in predicting the target variable.

For the second multivariate linear regression model, with a subset of features, the model's MSE is 477.07279521201883, with a mean absolute error of 16.819920852603918 for the test data. The model's performance on the test data is slightly worse than the training data, where the model's MSE is 521.5510489687584, with a mean absolute error of 16.32459550322298. Compared to the first model, this model performs slightly worse, likely due to the fewer features considered. The scatter plot shows that the predicted values of the test data align relatively closely with the actual values, indicating that the model performs reasonably well in predicting the target variable.

In terms of underperforming cases or observations, it would be helpful to investigate the data points that deviate significantly from the diagonal line in the scatter plot. These data points may indicate outliers or cases where the model is performing poorly. Additionally, further feature engineering or data preprocessing may be necessary to improve the model's performance.

3.b. Business Impact

Based on the results of the multivariate linear regression models, we can see that the second model with a subset of features (['avgAnnCount', 'incidenceRate', 'medIncome', 'popEst2015', 'povertyPercent']) performed better than the first model with all features.

The MSE and MAE for the second model are lower for both the training and test data compared to the first model. This indicates that the second model is better at predicting the target variable and has better generalization performance.

The business objective set earlier was to predict the target variable (cancer mortality rate) based on the given features. The multivariate linear regression models help achieve this objective by providing a model that can predict the target variable based on the features.

Incorrect results from the model can have various impacts on the business depending on the severity and nature of the incorrect results. For example, if the model overestimates the cancer mortality rate for a particular region, then it may lead to unnecessary panic and lower real estate values in that region. On the other hand, if the model underestimates the cancer mortality rate for a particular region, then it may lead to a lack of attention and resources towards that region, which can have negative health outcomes for the residents. Therefore, it is important to carefully evaluate the model's performance and consider the potential impacts of incorrect results on the business objectives and stakeholders.

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. Additional data sources: 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 experiment, we gained several insights about the performance of multivariate linear regression models on this dataset. The first model, which used all numeric features, had a lower MSE and MAE compared to the baseline model, indicating that it is performing better than simply predicting the mean of the training target variable for all test data points. The coefficients and intercept of the model can also provide insights into which features are more strongly correlated with the target variable. The second model, which used a subset of the features, had slightly higher MSE and MAE for the test data compared to the first model. However, the MSE for the test data was still lower than the baseline MSE, indicating that this model is still performing better than the baseline model. The coefficients of this model can also provide insights into which subset of features are more important in predicting the target variable. Overall, these results suggest that multivariate linear regression models can be effective in predicting the target variable for this dataset, and further experimentation could be pursued to optimize the model's performance. However, it is important to note that linear regression models may not be able to capture non-linear relationships between the features and target variable, and other machine learning models may need to be explored if linear regression proves to be a dead end for this dataset.

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: We could create new features by combining existing features or extracting new features from the existing data. For example, we could create a new feature by combining the average income and poverty percentage, or extract the age distribution of the population. This could potentially improve the performance of the model. Expected uplift or gains: The expected uplift or gains from this experiment are moderate. While creating new features may improve the model's performance, it may also lead to overfitting or introduce noise into the data. Ranking: Medium 2. Data preprocessing: We could experiment with different data preprocessing techniques such as scaling, normalization, or outlier removal. This could potentially improve the performance of the model by reducing the impact of outliers or bringing the data into a more manageable range. Expected uplift or gains: The expected uplift or gains from this experiment are moderate. While data preprocessing techniques may improve the model's performance, they may also introduce bias into the data or remove important information. Ranking: Medium 3. Model selection: We could experiment with different types of models such as decision trees, random forests, or neural networks. This could potentially improve the performance of the model by leveraging different algorithms or combining multiple models. Expected uplift or gains: The expected uplift or gains from this experiment are high. Trying different models can potentially lead to significant improvements in performance and can also help us understand which models are best suited for this problem. Ranking: High 4. Hyperparameter tuning: We could experiment with different hyperparameters such as the learning rate, regularization strength, or number of hidden layers. This could potentially improve the performance of the model by optimizing the hyperparameters for the specific problem. Expected uplift or gains: The expected uplift or gains from this experiment are high. Hyperparameter tuning can significantly improve the performance of the model and is often necessary to achieve state-of-the-art results. Ranking: High 5. Ensemble methods: We could experiment with ensemble methods such as bagging, boosting, or stacking. This could potentially improve the performance of the model by combining multiple models or reducing the impact of noisy data. Expected uplift or gains: The expected uplift or gains from this experiment are high. Ensemble methods have been shown to improve the performance of models in many applications and can often lead to state-of-the-art results. Ranking: High If any of the experiments achieve the required outcome for the business, the next step would be to deploy the solution into production. This would involve integrating the model into the existing software infrastructure and creating an interface for users to interact with the model. We would also need to set up monitoring and maintenance procedures to ensure that the model continues to perform well over time.