Assignment 1 - Regression Models
Submitted By:- Hemang Sharma (24695785)

Part D: A final report detailing this project following the CRISP-DM methodology

Introduction

The aim of this project is to predict cancer mortality based on information related to US counties. The dataset contains 33 different features (demography, medical information). The CRISP-DM methodology was followed to solve this problem.

Business Understanding

Cancer is a major health concern in the US, and predicting cancer mortality can help in developing better policies and strategies to reduce the incidence of cancer. The goal of this project is to accurately predict cancer mortality based on information related to US counties. The project aims to provide insights into the factors that influence cancer death rates and to create a model that predicts the death rates in US counties. The results of the project can be used by healthcare professionals and policymakers to identify areas with high cancer death rates and to develop targeted interventions.

Data Understanding

The dataset used in the project consists of demographic and medical data for 3,047 US counties. The dataset includes variables such as age-adjusted mortality rates, average annual count of cancer diagnoses, income, and population estimates. The dataset contains 33 features and 3,047 instances. The target variable is "TARGET_deathRate", which is the mean per capita (100,000) cancer mortalities. The dataset has a mix of numerical and categorical features, and missing values were imputed with the median.

Data Preparation

In the data preparation step, the data was cleaned. The training set was used to train the models, while the testing set was used to evaluate the performance of the models. The training set was used to train the linear regression model, and the validation set was used to tune the hyperparameters of the model. The testing set was used to evaluate the final model.

Modeling

A linear regression model was chosen to predict the cancer death rates. Three types of models were trained: univariate linear regression, multivariate linear regression, and multivariate linear regression with feature engineering.

In the univariate linear regression models, two features were selected and separate models were trained for each feature. The mean squared error (MSE) was used as the evaluation metric. In the multivariate linear regression model, all numeric features were used. In the multivariate linear regression model with feature engineering, the numeric features were transformed using PolynomialFeatures class with degree 1, 2, 3, 4 and 5.

Evaluation

The performance of the model was evaluated using the mean squared error (MSE) and root mean squared error (RMSE). The MSE and RMSE were calculated for the validation and testing sets. The predicted values were plotted against the actual values for both the validation and testing sets using a scatter plot. The best performing model was the multivariate linear regression model with feature engineering. This model had an MSE of 473.29 on the testing set. The univariate linear regression models had higher MSEs, indicating that a single feature is not enough to accurately predict cancer mortality.

Deployment

The final model can be used to predict cancer mortality based on information related to US counties. The model can be deployed as a web service or integrated into a larger system.

Issues:

One of the issues faced in this project was the high number of features in the dataset. Feature selection techniques could be used to identify the most important features and reduce the dimensionality of the dataset.

Another issue was the use of median imputation to handle missing values. Other techniques such as mean imputation or predictive imputation could be explored to handle missing values.

Recommendations:

To improve the performance of the model, additional features such as the prevalence of risk factors such as smoking and obesity could be included in the model.

In future work, more advanced machine learning models such as decision trees, random forests, and gradient boosting could be explored. Additionally, a more advanced machine learning algorithm such as a neural network could be used to model the complex relationships between the variables. Finally, the model could be extended to include spatial analysis to identify clusters of high cancer death rates and to develop targeted interventions in those

Results:

The linear regression model achieved a mean squared error (MSE) of 417.43 and a root mean squared error (RMSE) of 20.43 on the validation set. On the testing set, the model achieved an MSE of 473.29 and an RMSE of 21.76. This indicates that the model is overfitting when trained on a small amount of data, but the performance improves as more data is used for training.