Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037/HJ1: Concepts and Technologies of AI (Herald College, Kathmandu, Nepal)	⭐ Final portfolio Project	An End- to- End Machine Learning Project on Regression and Classification Task

Student Name: Saif Siddiqui

Student ID: 2407733

Module Leader: Mr. Siman Giri

Tutor: Ms. Durga Pokharel

Abstract

The purpose of this report is to predict a continuous variable through regression techniques. The dataset chosen for this study is "Health.csv," which contains financial data related to customer transactions and loans. The investigation works from exploratory data analysis (EDA) and model-development perspective based on regression algorithms. Important methodologies utilized in the analysis are data preprocessing for missing values and outliers, followed by feature selection for better model performance and hyperparameter tuning for better optimization. The evaluation of the models was through R-square and mean square errors to find their accuracy. The findings suggest that the regression model is reasonably accurate in predicting the target variable. Results from the analysis may provide insights that are beneficial for financial decision-making and risk assessment.

1. Introduction

1.1 Problem Statement

Through this project, it is aimed to predict a continuous target variable based on financial data.

1.2 Dataset

The data was cleaned as per the considerations made before modeling. The processes done include handling missing values, udaq-ing outliers, and normalizing features.

1.3 Objective

Develop a regression model that predicts the target continuous variable with the aid of given financial features.

2. Methodology

2.1 Data Preprocessing

Only visualizations like scatter plots, histograms, and correlation matrices are used in such analyses in order to give a better understanding of the data. Some of the salient points derived from the same EDAs have been provided.

2.2 Exploratory Data Analysis (EDA)

Besides understanding the dataset and identifying major distributions, EDA was performed. In that context, variable distributions were visualized with histograms and pound plots, and relationships between features were shown by a correlation heatmap. It was further found that high cholesterol and hypertension were highly correlated with heart diseases, thus these characteristics can be used as good predictors.

2.3 Model Building

Two regression models that will be considered for this task are: [Model 1, e.g., Linear Regression] and [Model 2, e.g., Decision Trees]. These were fitted by taking the data split into training sets and testing sets so that the model could be properly trained.

2.4 Model Evaluation

To characterize the predictive performance of the model, a number of statistics were calculated: R-squared: This statistic indicates the ratio by which the variance of the dependent variable is explained by the independent variables.Mean square error: This contains the mean square of the difference between the actual values and the fitted values.

2.5 Hyperparameter Optimization

Hyperparameter tuning was done using either [GridSearchCV or RandomizedSearchCV] . At last, the best parameters were found to be [List of Optimal Parameters].

2.6 Feature Selection

[RFE] was used for feature selection. This recursion selects the features that contribute best to prediction of the target variable. Th chosen features, in this case, were [`Symbols of Selected Features']

3. Conclusion

In this study, the regression analysis effectively provided a model that predicts a target variable as a continuous variable through financial data. The final model performed well with the help of evaluation metrics like R-squared and MSE. This shows the value of data preprocessing and feature selection in enhancing model accuracy. However, the process was met with hurdles regarding data quality aspects and model generalization. While these challenges will be passed on, it is most fortunate that the model gives an insight into, and refinements could be made for even better predictive results. Future work may investigate the inclusion of other features, the use of more advanced regression techniques, and a larger dataset for more robustness of the model.

4. Discussion

Model results indicate that the model is capable of learning in such a way as to reasonably capture relationships in the data, unhappy with the fact that limitations were imposed. Hyperparameters optimization and feature selection worked well to avoid overfitting while improving the accuracy of the model by employing significant predictors. The results are as anticipated, underscoring the significance of suitable data preprocessing and optimization of the model. However, limitations such as the size of the dataset and the underlying assumptions of the regression model may also have contributed to the findings. Future research could consider other regression alternatives such as ensemble learning or deep learning models to improve the prediction accuracy even further. In addition, collecting more diverse and representative data can increase the generalizability of the model on real-world finance problems.