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

Classification Task Report

Abstract

In this analysis, a classification method is applied to predict whether or not a customer is likely to open a fixed term deposit account. The dataset comes from the UCI Machine Learning Repository with customer demographic and financial information. Some steps involved data preprocessing, exploratory data analysis, model building employing logistic regression and decision trees, hyperparameter tuning, and feature selection. Logistic regression gave more generalizable performance while decision trees started to show overfitting signs. Important features influencing customer decision included job-type, previous campaign contacts, account balance, call duration, etc. Missed values and class imbalance issues were solved through preprocessing techniques. Hyperparameter tuning and feature selection were done to optimize parameters, reduce complexity, and improve model performance. The final model selected showed an accuracy of 91.62% with precision at 67.48%, recall at 52.33%, and F1 score at 58.95%. Future improvements could involve other ensemble techniques like Random Forest and XGBoost, SMOTE balancing on the dataset, and deep learning techniques that would enhance prediction. accuracy.

1. Introduction

1.1 Problem

This project's goal is to use a customer's financial and demographic data to decide if they should open a fixed term deposit account. In the banking sector, this classification problem might be an effective strategy for focused marketing.

1.2 Dataset

Bank marketing data from the UCI machine learning repository served as the dataset's foundation. Customer characteristics including age, occupation, marital status, education level, bank balance, and specifics of previous marketing efforts are all included. Because it seeks to enhance financial inclusion and responsible consumption, this dataset is in line with the UNSDGs.

1.3 Objectives

The objective of this analysis is to create a predictive classification model that predicts whether a customer will open a fixed deposit account based on certain characteristics.

2. Methodology

2.1 Data Preprocessing

We handled missing values and coded categorical variables to tidy up the data before creating the model. The variables were then subjected to statistical transformations, such as altering the scale of numerical features, in order to get them ready for analysis.

2.2 Exploratory Data Analysis (EDA)

This either suggests that in the process of analysis, one would use histograms, bar charts, and correlation matrices for an interpretatively shedding of exploratory analysis. EDA disclosed that the nature of activity and previous interactions with customers affected the target variable substantially.

Consequently, a class imbalance exists in the dataset: more No responses than Yes.

2.3 Model Building

The Chapter Model Execution Process went as follows:

- -The two classification models, logistic regression and decision tree, were considered for this task.
- -Data are partitioned into two parts, whereby one is the training set and the other is the test set.
- -Train the model based on training data.
- -Test performance on the test set.

2.4 Model Evaluation

Evaluation metrics:

- Accuracy: Percentage representation of total predictions illustrating the ratio of correctly predicted positive and negative labels.
- Precision: Proportion of predicted positive cases that are actually correct.
- Recall: The proportion of positive instances that the model correctly identifies.
- F1 Score: The harmonic mean between precision and recall giving equal weight to both.

The best model acquired an accuracy of 91.62%, precision of 67.48%, recall of 52.33%, and an F1 score of 58.95%.

2.5 Hyperparameter Optimization

Hyperparameter tuning on the models was carried out succinctly using GridSearchCV.

- The models' optimal parameters were matched with:
- Logistic Regression-L2 regularization with C equal to 1.
- Decision Tree-Maximum depth 6; criterion equals Gini.
- Best model: A random forest model with n_estimators equals 150; min_samples_split equals to 5; min_samples_leaf equal to 4; max_features equal to log(2); and max_depth equal to None.

2.6 Feature Selection

Recursive feature elimination was carried out properly so that the really significant features could be found. The selected factors included type of work, past contact with this campaign, balance, and length of last call.

3.Conclusion

The performance of the classification models in this regard was fairly good, with logistic regression outdoing the others as the most effective for giving an optimal trade-off between precision and recall. They provided insight into the most important features, such as employment type, account balance, and previous contact during the campaign, for predicting the customer response. The final selected model achieves 91.62% accuracy, 67.48% precision, 52.33% recall, and 58.95% F1 score.

Considerations like handling missing values, countering class imbalance, and preventing overfitting in the decision tree model came up as challenges during implementation. The advancements into the future may consider putting ensemble methods into play, including Random Forest and XGBoost, as well as SMOTE for balancing the data, and taking neural networks into account for further enhancement of expected results.

4.Discussion

Logistic regression emerged as a generalization good, whereas decision trees a better fit for training data. This however leads to very little generalization ability to predict outcomes on unseen test data very well; this is known as overfitting.Auto-tuning other hyperparameters brings out required regularization and tree depth that drive accuracy for better prediction. Applying feature selection also relieved us of some complexity without sacrificing performance, which helped overcome the hurdle. Factors like type of activity and contact during previous campaigns influenced a customer's decision reasonably well in line with business expectation. However, the following issues identified in this study still remain: B. The data set is imbalanced and provides further opportunities for improvement through feature engineering. In addition to those, future research can explore ensemble methods, neural networks, customer behavior features, etc., as means of enhancing the accuracy of prediction models.