Assignment 1

Luis Dale Gascon
Computer Science
Towson University
lgascon1@students.towson.edu

***Abstract*—Using a dataset**

***Index Terms*—shopping, classification, random forest tree**

I. Introduction

This experiment attempts to create a classifying model based on customer transactions to determine if a customer has a subscription based on the multiple variables. Predicting whether a customer is subscribed provides insights as to what factors will make a customer want to apply for a subscription

II. Related Works

[1] compares classification algorithms that are commonly found in ecological studies, namely linear discriminant analysis (LDA), logistic regression, additive logistic regression and classification trees across 3 different experiments. Benchmark values are: percentage correctly classified (PCC), sensitivity, specificity, κ

Concludes that random forest, in principle, outperforms LDA and logistic regression when there are strong interactions among variables. The author notes 2 key features that sets the random forest algorithm apart. One is its ability to evaluate the importance of features and the possibility of imputing data using proximities, a metric on how often 2 data points end up in the same leaf node across all trees in the forest.

[2] is trying to find out the reason for factors and motivations that cause undergraduate students to drop out. Their research looks at using the random forest algorithm to predict the level of student performance. Students were surveyed on 24 questions on the freshmen population. Significant factors needs identification, so the researcher performed an analysis of variance to check the correlation between the different factors. The researcher ended up with 11 significant factors. The accuracy of predictions is at 96.88% and random forest was also able to determine 2 key factors, which [1] noted that random forest was able to do: effort and creating good notes. The author suggests comparing random forests against support vector machines and various deep learning algorithms. The author also suggests that more data on other universities would be beneficial.

III. Dataset

The dataset contains 3,900 rows, each representing a customer's transaction. For each entry, information such as the customer's age, gender, and unique ID is provided. Details about the purchased item: including category, size, color, price, and payment method are also included. Additional context captures the season and store location where the transaction took place. Customer behavior is recorded through review ratings, subscription status, and whether a discount or promo code was used. The dataset also tracks each customer's purchase history, including the total number and frequency of their transactions. This dataset was artificially generated using ChatGPT for educational purposes.

IV. Methodology

A. Toolset

We begin our experiment with Python and its vast libraries. Scikit learn provides a random forest classifier class that provides functionality to build forest trees from our dataset. Polars will represent data in dataframes. Marimo is a Jupyter notebook alternative that comes with

B. Approach

a) *Preprocessing:* Since this dataset is artificially created, there wasn't any missing data that I would need to impute or perform data cleaning with.

To avoid biasing in one location, we'll have to perform stratified random sampling on train-test data on location.

The random forest function from scikit learn expects columns to have numerical value and there are plenty of categorical columns. We would have to encode those columns. Shirt sizes and frequency will be encoded using label encoding.

The Subscription Status, Discount Applied, and Promo Code used columns consists of binary values, so I'll just map Yes = 1 and No = 0.

Due to the artifical nature of this dataset, the values of the location column range from 96 to 63, with 50 unique values. There are multiple locations with the same frequency after exploring the data, so frequency encoding would be a poor choice for this feature. I went with target encoding with Subscription Status as the target feature.

One-Hot encoding is used for the rest of the categorical data as those columns have low cardinality. Although my dataset has the gender column consists of values as male or female, I plan to use One-Hot as its encoding technique.

After encoding categorical data and setting a standard scale for numerical data, we perform feature selection using a Chi-Square test to determine the most relevant features to the target variable.

b) *Training:*

V. Results

As this is a binary classification problem, an ROC curve is used to visually showcase the performance of our model

A. Findings

References

[1] D. R. Cutler *et al.*, “RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY,” *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, doi: https://doi.org/10.1890/07-0539.1.

[2] S. Kumar and F. Janan, “Prediction of Student's Performance Using Random Forest Classifier,” 2021, p. .