Assignment 1

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

***Abstract*—Using a dataset generated by ChatGPT to mimic real world consumer behavior, we want to determine the factors influencing customer subscription decisions by developing a predictive classification model using a classification algorithm.**

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

I. Introduction

This experiment attempts to create a classification model based on customer transactions to determine if a customer has a subscription based on multiple variables. Predicting whether a customer is subscribed provides insights as to which factors influence a customer to subscribe.

II. Related Works

[1] compares several 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 datasets. All 3 datasets were used to create a predicting model

This paper 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 two data points end up in the same leaf node across all trees in the forest. The author did point out that random forest shouldn't be a tool for traditional statistical inference such as ANOVA or hypothesis testing.

Similarly, [2] investigates the factors and motivations that contribute to undergraduate students dropping 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 required identification, so the researcher performed an analysis of variance to check the correlation among the different factors. The researcher identified 11 significant factors with model accuracy reaching 96.88%. The study highlights 'effort' and 'note taking' as key determinants, and recommends comparing random forest performance with support vector machines and deep learning methods, as well as expanding the dataset to include multiple universities.

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. Approach

a) *Preprocessing:* Since this dataset is artificially created, there were no missing data to impute or clean.

To avoid any location-based biases, we randomly stratified the samples and split the dataset into training and testing subsets according to the store 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. We want to standardize numerical values so that no numerical feature has greater significance than another simply due to the magnitude of their values.

The Subscription Status, Discount Applied, and Promo Code used columns consist of binary values, which were mapped as Yes = 1 and No = 0.

Due to the artificial 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. We went with target encoding for the location feature, with Subscription Status as the target feature.

One-hot encoding is applied to other categorical variables with low cardinality to accommodate potential future values, such as additional genders appearing in unseen data.

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

Discount Applied and Promo Code Used emerged as the most significant features (score: 771), followed by Gender (score: 315).

Looking back at the related works, specifically [1], they mentioned how random forest is able to determine the important features. I was able to extract those important features via the feature_importances_ variable from the classifier object. After sorting, I was presented with the same features from the Chi-Square test.

V. Results

Using the default setting of 100 trees in Scikit-learn's RandomForestClassifier, the model achieved an accuracy of 100%, which raises concerns regarding overfitting.

We wanted to experiment by setting the hyper-parameter of the number of trees to a low number and testing without feature selection.

In the worst case where we set the number of trees to 1 and without feature selection, the model achieved an accuracy of 89% but once we trained with feature selection and with the same number of trees, the model achieved 100% accuracy after 20 predictions.

This suggests the synthetic dataset contains patterns that are influenced from the large language model that generated it, potentially limiting the generalizability of these findings when deployed in the real world.

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

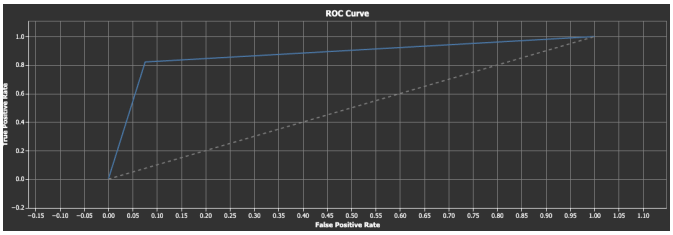

Fig. 1. ROC with accuracy score of 89% without feature selection

This experiment's limitation is its AI-generated dataset. Incorporating real-world data in the future would improve the model's ability to capture authentic consumer behavior and lead to a model that is more representative of practical business scenarios.

A. Connection to relate works literature

As highlighted in the related works, the random forest classifier perfroms effectively with high-dimensional data. Similar to the works of [1] and [2], our experiment leverages feature importance analysis.

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. .