Predicting Book Purchase Behavior – Classification & Association Rules

Imadul Islam Chowdhury
Business Insights and Analytics
Longo Faculty of Business
Humber Polytechnic
Toronto, Canada
N01588334@humber.ca

***Abstract*—predicting the book purchase behavior can be challenging sometimes. This paper examines the application of classification and association rules to come up with suggestion for the store. The paper evaluates multiple machine learning models, including Random Forest and Apriori algorithm to extract rules with support, confidence, and lift and highlighting their ability in terms of predictive accuracy and robustness. The paper concludes by discussing the potential uncover book co-purchase trends.**

***Index Terms*—Ensemble classification methods, Random Forest, Apriori.**

I. INTRODUCTION

The Charles Book Club is on a mission to find a comprehensive analysis of their customers historical purchasing behaviors to pinpoint the customers who poses a keen propensity for acquiring specialty books, with a specific focus on travel guides related to Florence. This analysis involves a detailed exploration of customer profiles, aiming to evaluate how past purchasing decisions guidance the likelihood of these individuals purchasing this specific category of literature. To achieve this, I utilized ensemble classification techniques to construct a robust predictive model that can effectively forecast customer preferences. Additionally, I engaged in association rule mining to uncover substantial trends in the co-purchase of books, thereby providing deeper insights into customer behavior and preferences within the same domain book market. This dual approach will not only improve the understanding of customer likings but also expedite the targeted marketing strategies that align with identified purchasing patterns.

To come up with a solution for the above I used Random Forest and Apriori algorithm to potential uncover book co-purchase trends and recommendations for future planning.

II. METHODOLOGY

a) Dataset Description

The CharlesBookClub.csv dataset contains various important columns that provide deep insights into the customer demographics and their purchasing behavior. The dataset consists of 24 distinct columns: 'Seq#', 'ID#', 'Gender', 'M', 'R', 'F', 'FirstPurch', 'ChildBks', 'YouthBks', 'CookBks', 'DoItYBks', 'RefBks', 'ArtBks', 'GeogBks', 'ItalCook', 'ItalAtlas', 'ItalArt', 'Florence', 'Related Purchase', 'Mcode', 'Rcode', 'Fcode', 'Yes_Florence', 'No_Florence'.

b) Data Preparation & Exploration

Python code has been used to exploration and transformation for data. As par my findings, there are 4000 number of rows, and no null values was found.

```
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Seq#              4000 non-null   int64
 1   ID#               4000 non-null   int64
 2   Gender            4000 non-null   int64
 3   M                 4000 non-null   int64
 4   R                 4000 non-null   int64
 5   F                 4000 non-null   int64
 6   FirstPurch        4000 non-null   int64
 7   ChildBks          4000 non-null   int64
 8   YouthBks          4000 non-null   int64
 9   CookBks           4000 non-null   int64
 10  DoItYBks          4000 non-null   int64
 11  RefBks            4000 non-null   int64
 12  ArtBks            4000 non-null   int64
 13  GeogBks           4000 non-null   int64
 14  ItalCook          4000 non-null   int64
 15  ItalAtlas         4000 non-null   int64
 16  ItalArt           4000 non-null   int64
 17  Florence          4000 non-null   int64
 18  Related Purchase  4000 non-null   int64
 19  Mcode             4000 non-null   int64
 20  Rcode             4000 non-null   int64
 21  Fcode             4000 non-null   int64
 22  Yes_Florence      4000 non-null   int64
 23  No_Florence       4000 non-null   int64
```

Fig 1: CharlesBookClub Dataset

Also, we calculated the total purchase summary.

```
Summary of Book Purchases:
ChildBks            2559
YouthBks            1219
CookBks             2925
DoItYBks            1403
RefBks              1025
ArtBks              1156
GeogBks             1550
ItalCook             501
ItalAtlas            150
ItalArt              183
Florence             338
Related Purchase    3540
Mcode              17125
```

Fig 2: Purchase Summary

However, for better performance of the model I have removed the less effective column in this case it's 'Seq#' and 'ID#' column. Also, I have created a single binary target column from Yes_Florence and No_Florence and removed the original Yes_Florence and No_Florence columns.

III. MODELS

a) Random Forest Algorithm

Random Forest is a classifier composed of multiple decision trees that are not related to each other. It is based on the Bagging model in ensemble learning.

Bagging performs subsampling from the training set to create sub-training sets for each base model, building multiple independent classification models. The final classification decision is then made through a voting mechanism, following the principle of majority rule. [2-3].

Random Forest algorithm is more applicable and superior in prediction compared to traditional algorithms, especially when the data has missing values, blanks, or when extracting other data value [4-5].

b) Apriori algorithm

Apriori algorithm [6] is the most famous and most basic association rule mining algorithm. The Apriori algorithm is generally divided into two key steps: one is to find all frequent itemset in the transaction database, and the other is to generate strong association rules. Given a database, initial single scan transaction set D, determine the support of each item, generate frequent 1-itemset L1 by calculating the minimum support, and get candidate 1-itemset C1 by connecting L1. Scanning the database for the second time, find out all itemset with support greater than or equal to the minimum support in C1 to form a frequent 2-itemset L2, and obtain the candidate 2-itemset C2 by connecting L2. In the same way, scanning the database for the kth time, find out all the itemset with support greater than or equal to the minimum support in Ck−1 to form frequent k-itemset Lk, and connect by Lk to obtain candidate k-itemset Ck until there is no new candidate generated. The Apriori algorithm performs the following two operations: connection and pruning. In the connection process, Lk−1 and Lk−1 are connected to generate potential candidate itemset; in the pruning process, the database is scanned to determine the count of each candidate set in Ck and use Lk−1 to remove infrequent itemset in Ck, to determine Lk [7].

IV. RESULTS AND RECOMMENDATION

The primary factor which is influencing the customer behavior in the purchase of the Florence travel book include previous purchases of books relating to Florence (with .8635), the total expenditure by the customer, and the timing of their initial purchase.

```
Key Drivers of Customer Behavior:
Florence            0.863519
M                   0.023918
FirstPurch          0.013962
Related Purchase    0.012549
ArtBks              0.011293
R                   0.011193
F                   0.009860
GeogBks             0.005950
CookBks             0.005654
Rcode               0.005582
```

Fig 3: Key Drivers of Customer Purchases

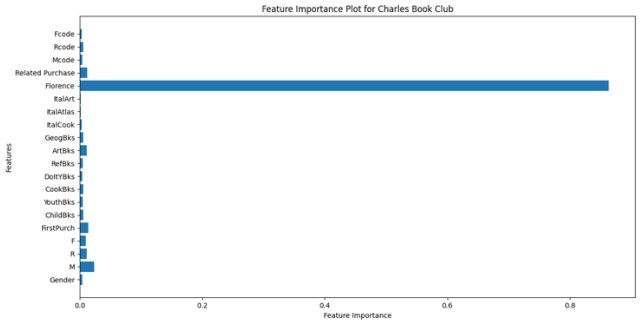

Fig 4: Key Features

Furthermore, connected purchases like the quantity of art books bought, as well as the recentness and frequency of purchases, are also decisive elements. Analyzing these pieces can enhance the ability to predict and effectively target prospective buyers.

```
Model Accuracy: 1.0
Model Precision: 1.0
Model Recall: 1.0
```

Fig 6: Confusion Matrix

As per the confusion matrix, the Random Forest model demonstrated flawless performance, attaining an accuracy, precision, and recall rate of 1.0.

The Apriori model specifies that customers purchasing genres such as Related Purchase, ItalArt, and CookBks are also tending to purchase ItalCook and ArtBks, showing significant confidence and lift.

```
Top 3 Association Rules:
                                  antecedents  ... kulczynski
1757                      (ItalCook, ArtBks)  ...    0.348986
902                       (ItalArt, CookBks)  ...    0.348986
1744  (Related Purchase, ItalArt, CookBks)  ...    0.348986
```

Fig 7: Top 3 Association Rules

```
Rule 1751:
Antecedents: Related Purchase, ItalArt, CookBks
Consequents: ItalCook, ArtBks
Support: 0.011
Confidence: 0.362
Lift: 10.575
Actionable Insights:
Customers who buy Related Purchase, ItalArt, CookBks are likely to also buy ItalCook, ArtBks.
Consider bundling these genres or offering combo discounts on Related Purchase, ItalArt, CookBks + ItalCook, ArtBks in marketing campaigns.

Rule 1752:
Antecedents: ItalCook, ArtBks
Consequents: Related Purchase, ItalArt, CookBks
Support: 0.011
Confidence: 0.336
Lift: 10.575
Actionable Insights:
Customers who buy ItalCook, ArtBks are likely to also buy Related Purchase, ItalArt, CookBks.
Consider bundling these genres or offering combo discounts on ItalCook, ArtBks + Related Purchase, ItalArt, CookBks in marketing campaigns.

Rule 1759:
Antecedents: ItalArt, CookBks
Consequents: ItalCook, ArtBks, Related Purchase
Support: 0.011
Confidence: 0.362
Lift: 10.575
Actionable Insights:
Customers who buy ItalArt, CookBks are likely to also buy ItalCook, ArtBks, Related Purchase.
Consider bundling these genres or offering combo discounts on ItalArt, CookBks + ItalCook, ArtBks, Related Purchase in marketing campaigns.
```

Fig 8: Top 3 Support, Lift and Confidence

To effectively address these buying behaviors, marketers could consider bundling these genres or providing combo discounts in their promotional strategies.

REFERENCES

[1] "Research on User Behavior Prediction Based on Random Forest Algorithm with RFM Model," by R. Li, J. Zhang and S. He, 4th International Conference on Computer Science and Blockchain (CCSB), Shenzhen, China, 2024, pp. 133-137. http://ieeexplore.ieee.org/document/10735665

[2] MUCHLINSKI, D., SIROKY, D., HE, J.R., et al. (2016) Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. J. Political Analysis, 24(1):87-103.

[3] Barandela, R., Valdovinos, R.M., Sánchez, J.S., et al. (2004) The imbalanced training sample problem: Under or over sampling. In: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Berlin. 806-814.

[4] Xiang, J.Y., Wang, Z.H., Deng, Y.Y. (2024) A Review of Machine Learning Classification Based on the Random Forest Algorithm. J. Journal of Artificial Intelligence and Robotics Research, 13(1): 143-152.

[5] Xiong, Z.M., Guo, H.Y., Wu, Y.X. (2021) A Review of Methods for Handling Missing Data. J. Journal of Computer Engineering & Applications, 57(14): 27.

[6] "Mining association rules between sets of items in large databases" by R. Agrawal, O. Imieliński and A. Swami, vol. 22, no. 2, pp. 207-216, 1993.

[7] "Research on Community Consumer Behavior Based on Association Rules Analysis," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 2021, pp. 1213-1216, https://ieeexplore.ieee.org/document/9408917