Predicting Book Purchases Using Classification and Association Rule Mining

Haleh Arjmand
line 2: dept of Business
Humber College
Toronto, Canada

***Abstract*— This report focuses on predicting customer interest in a specialty travel book using past purchase behavior. Using a dataset from a book club, various classification and association rule mining techniques were applied to understand what influences customers to buy the Florence travel book. Three key steps were followed: data cleaning and exploration, building an ensemble classification model, and applying the Apriori algorithm for market basket analysis. The Random Forest model was used to classify likely buyers, and its performance was evaluated using accuracy, precision, and recall. Additionally, association rules uncovered strong genre relationships that can be useful for cross-selling. The insights from this analysis can support targeted marketing and inventory decisions.**

Keywords—forecasting, Ensemble Learning, Random Forest demand, Apriori Algorithm

I. Introduction

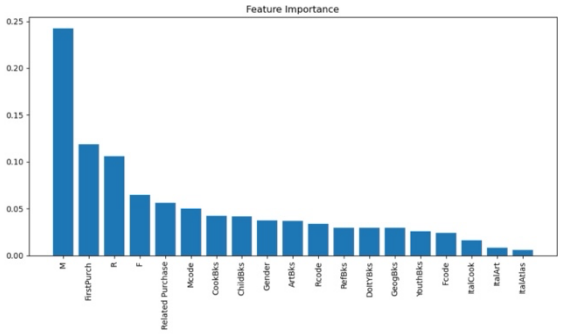
This report looks at how to predict whether a customer will buy a specific travel book based on their past book purchases. The data comes from a book club and includes different genres people have bought. I used classification and association rule mining to find patterns in their behavior and figure out what influences them to buy. The idea is to use these insights to help with more targeted marketing and better decisions around what to promote or keep in stock.

II. Executive Summary

This case study focuses on predicting if a customer will purchase a specialty travel book using past purchase data. I used a classification model to identify which features have the most impact, and applied association rule mining to find patterns between book genres. The Random Forest model performed well overall and showed that customers who spent more and bought certain genres were more likely to buy the Florence book. The association rules revealed strong links between genres like Child, Youth, and Cookbooks. These findings can help with targeting the right customers and making smarter marketing and inventory decisions.

III. Visual Insights

To start, I explored the data to spot any patterns in customer behavior. Customers who bought the Florence book also tended to spend more overall and purchased more frequently. They were also more active in genres like Art, Geography, and Youth books. The feature importance plot from the Random Forest model confirmed that monetary value (M), first purchase timing, and CookBooks were among the top drivers. Association rule mining also showed strong relationships between Child and Youth books, and between CookBooks and Child books. These patterns gave a better idea of what kinds of buyers are more likely to go for specialty books like Florence.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	cert
0	(ChildBks)	(YouthBks)	0.39400	0.23825	0.1475	0.374365	1.571314	1.0	0.053629	1.217564	0.599983	0.304281	0.17
1	(YouthBks)	(ChildBks)	0.23825	0.39400	0.1475	0.619098	1.571314	1.0	0.053629	1.590959	0.477309	0.304281	0.37
2	(CookBks)	(ChildBks)	0.41550	0.39400	0.2420	0.582431	1.478251	1.0	0.078293	1.451256	0.553507	0.426432	0.31

III. Methodology

I started by cleaning the data and creating a binary target column to show whether a customer bought the Florence book or not. Then I explored the features to understand purchase behavior across genres. For the classification part, I used a Random Forest model because it works well with mixed data and can handle feature importance. I split the data into training and testing sets and evaluated the model

using accuracy, precision, and recall. After that, I applied the Apriori algorithm on the genre columns (converted to binary) to find association rules. These rules helped reveal which genres are often bought together and gave extra insight into customer behavior.

IV. Recommendations

Based on the results, customers who spend more and are active in certain genres like Art, Youth, and CookBooks are more likely to buy specialty books like Florence. These customers can be good targets for future promotions. The model can also help identify potential buyers early on, so marketing efforts can be more focused. Also, since Child and Youth books were strongly linked, recommending them together could boost sales. It's a good idea to keep updating the model with new customer data and continue tracking purchase trends to improve targeting over time.

APPENDIX

- Created a binary target column (FlorenceBuyer) from Yes_Florence.
- Used Random Forest for classification; evaluated with accuracy, precision, and recall.
- Feature importance showed M, FirstPurch, CookBks, and Mcode as top drivers.
- Applied Apriori algorithm on binary genre data; strong associations found between ChildBks, YouthBks, and CookBks.
- Key insights supported by both model results and association rules.

References

[1] "IEEE - The world's largest technical professional organization dedicated to advancing technology for the benefit of humanity." [Online]. Available: https://www.ieee.org