Adhil Payingalil

Predicting Book Purchases & Purchase Patterns

***Abstract*— This analysis aimed to identify factors predicting customer purchases of the specialty "Art History of Florence" book and uncover common book co-purchase patterns for the Charles Book Club. Using historical transaction data, a Random Forest classification model was developed (addressing class imbalance with SMOTE) and market basket analysis was performed using the Apriori algorithm. Key predictors for buying the Florence book were identified as overall customer monetary value (M), time since first purchase (FirstPurch), recency of last purchase (R, Rcode), and Gender. However, the predictive model, while achieving 85% accuracy, struggled to precisely identify the small group of buyers (10.8% precision, 10.5% recall). Market basket analysis revealed strong associations, notably between Youth/DIY books and Cookbooks/Children's books. Recommendations include targeting high-value, long-term, recent customers for specialty books, leveraging co-purchase insights for cross-selling and bundling, and further refining predictive models to improve buyer identification.**

I. Executive Summary

This report details the analysis of Charles Book Club customer purchase data to identify factors influencing the purchase of a specialty travel book on Florence and to uncover common co-purchase patterns among book genres. Using machine learning (Random Forest classification with SMOTE balancing) and market basket analysis (Apriori association rules), we aimed to build a predictive model and derive actionable marketing insights**.**

Key Findings:

Predicting Florence Buyers: Customer engagement and value metrics are the strongest predictors. Customers who have spent more historically (Monetary value M), have been members longer (Time since First Purchase FirstPurch), and have purchased more recently (Recency R & Rcode) are more likely to buy the Florence book. Gender also emerged as a significant differentiating factor. While the predictive model achieved 85.0% accuracy, identifying the small group of actual buyers proved challenging, with low precision (10.8%) and recall (10.5%) for the buyer class.

Book Co-Purchase Patterns: Strong associations exist between practical and family-oriented genres. Notably:

- Customers buying Youth Books are highly likely (68% confidence) to also buy Cookbooks.
- Customers buying Do-It-Yourself Books are also highly likely to buy Children's Books (63% confidence) and Cookbooks (66% confidence).

Key Recommendations:

Targeted Marketing for Specialty Books: Focus promotional efforts for the Florence book (and similar niche titles) on customer segments characterized by high lifetime monetary value, longer tenure, and recent purchase activity. Tailor marketing messages based on gender insights. Use model predictions cautiously to refine targeting lists rather than solely relying on them, given the low recall.

Leverage Co-Purchase Insights: Implement cross-selling and bundling strategies based on identified associations (e.g., "Family Activity" bundles: Youth / Child / Cookbooks ; "Home Skills" bundles: DIY / Cookbooks). Use website recommendations and targeted emails to suggest associated genres.

Refine Predictive Modeling: Due to the model's difficulty in identifying buyers, explore further model tuning, alternative algorithms suited for imbalanced data, or additional feature engineering to improve recall and precision.

This analysis provides a clear profile of likely specialty book buyers and highlights strong cross-genre purchase relationships, offering valuable data for optimizing marketing campaigns and product offerings, while acknowledging the current model's limitations in precisely identifying all potential buyers.

II. Visual Insights

Initial data exploration provided several visual insights into customer behavior and potential predictors:

Target Variable Distribution: The bar chart clearly illustrates a significant class imbalance, with a much larger number of customers who did not purchase the Florence book (label 0) compared to those who did (label 1). This imbalance necessitates careful handling during modeling (like using SMOTE) to avoid biased predictions.

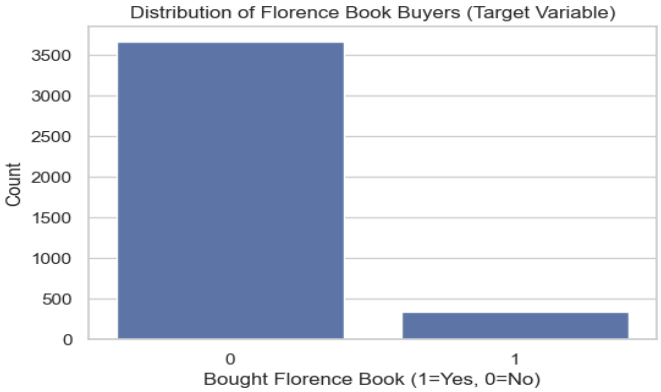

Figure 1. Distribution of Florence Book Buyers plot

Genre Purchase Averages: Comparing average book purchases, Florence buyers tend to purchase more Art, Geography, Cookbooks, and Italian-themed books (ItalCook, ItalArt, ItalAtlas) on average than non-buyers. Conversely, non-buyers show slightly higher average purchases for ChildBks and YouthBks.

Figure 2. Average Book Purchases by Genre plot

RFM Code Averages: The mean values for the coded RFM variables show subtle differences. Florence buyers (1) have slightly higher average Mcode (Monetary) and Fcode (Frequency) values and a slightly lower average Rcode (Recency - lower code indicates more recent purchase) compared to non-buyers (0). This hints that buyers might be slightly more valuable and recently active customers.

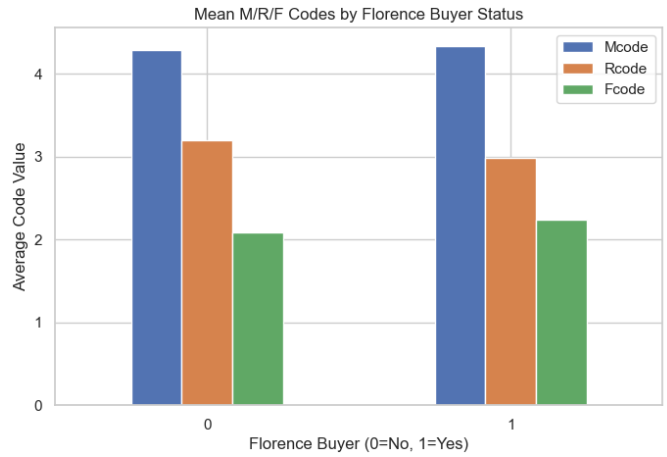

Figure 3. Mean M/R/F Codes plot

Florence & Related Purchase Averages: There's a stark difference in the mean counts for these specific features. The average Florence count is near 1.0 for buyers and 0 for non-buyers, confirming its direct link to the target. More importantly, the average Related Purchase count is significantly higher for buyers (around 1.4) than for non-buyers (around 0.85), indicating it's a strong potential predictor.

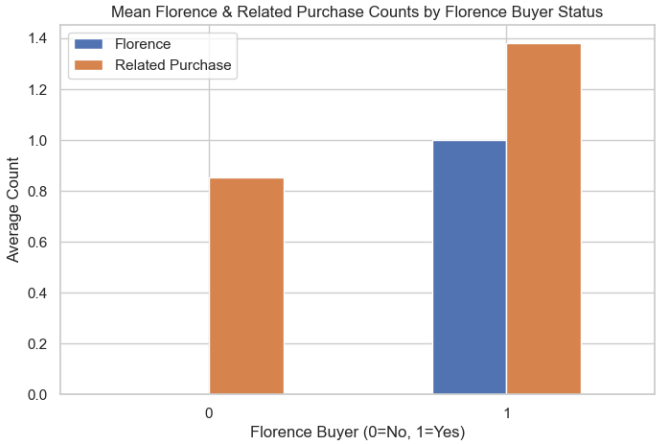

Figure 4. Mean Florence & Related Purchase Counts plot

Genre Correlations: The heatmap reveals positive correlations between related genres. For example, CookBks show moderate correlations with ChildBks (0.40), DoItYBks (0.33), and GeogBks (0.23). Italian genres (ItalCook, ItalArt, ItalAtlas) show correlations among themselves and with ArtBks and CookBks. This suggests natural groupings of interests.

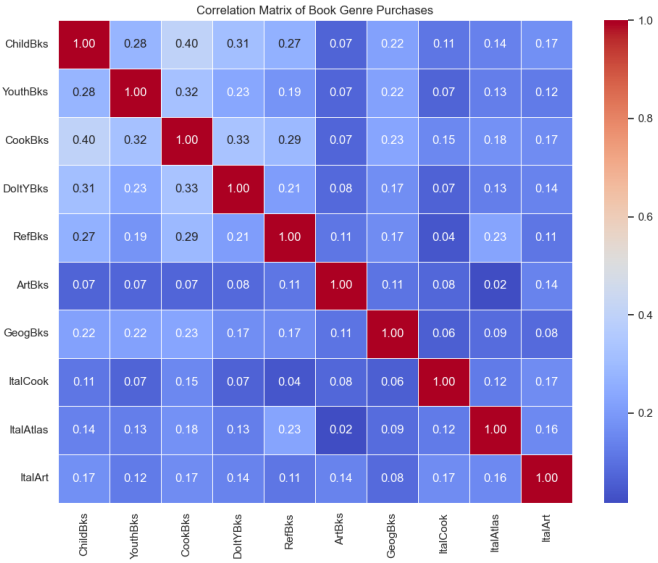

Figure 5. Correlation Matrix of Book Genre Purchases plot

Model Performance (Confusion Matrix): The confusion matrix for the trained Random Forest model (after SMOTE and excluding the 'Florence' feature) visually confirms the model's performance. It correctly identified many non-buyers (663 True Negatives) but struggled with buyers, correctly identifying only 7 (True Positives) while misclassifying 60 actual buyers as non-buyers (False Negatives) and 58 non-buyers as buyers (False Positives). This highlights the challenge posed by the class imbalance and the difficulty in predicting the minority 'buyer' class.

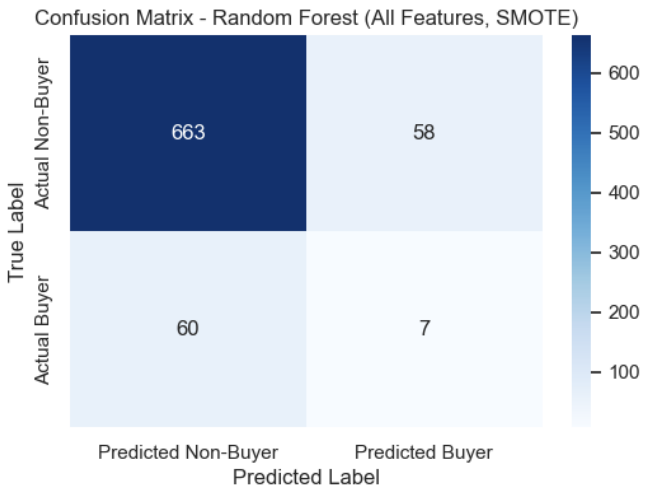

Figure 6. Confusion Matrix

Feature Importance: The bar chart ranking feature importances visually confirms the dominance of Monetary (M), First Purchase (FirstPurch), and Recency (R, Rcode) as the top predictors identified by the Random Forest model (when the direct 'Florence' feature is excluded). Gender also ranks relatively high, while specific book genres like CookBks and ChildBks follow, indicating their contribution, albeit less than the primary behavioral metrics.

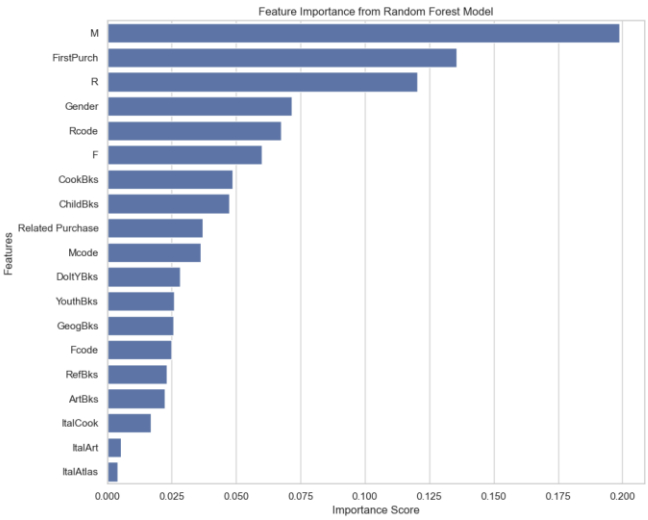

Figure 7. Feature Importance Plot

III. METHODOLOGY

The analysis involved the following steps:

A. Data Cleaning & Preparation

The raw dataset was loaded, inspected for missing values (none found) and duplicates (none found). Irrelevant identifier columns (Seq#, ID#) were removed. The target variable, Florence_Buyer, was created as a binary indicator (1 for purchase, 0 for no purchase) from the original Yes_Florence column.

B. Exploratory Analysis

Initial analysis revealed a significant class imbalance (approx. 8.5% buyers vs. 91.5% non-buyers). Visualizations explored average genre purchases, mean RFM code values (Mcode, Rcode, Fcode), and mean Related Purchase counts between buyers and non-buyers. A correlation matrix focused specifically on book genre co-purchases was generated.

C. Predictive Modeling (Classification)

- **Feature Selection:** All potentially relevant features were used, excluding the Florence feature itself to prevent data leakage.
- **Data Splitting**: Data was split into 80% training and 20% testing sets, stratified by the target variable.
- **Imbalance Handling**: The SMOTE technique was applied to the training data to create a balanced set for model training.
- **Model Building**: A Random Forest Classifier (an ensemble learning method) was trained on the SMOTE-balanced data.
- **Evaluation**: The model's performance was assessed on the unseen test set using Accuracy, Precision (for buyers), Recall (for buyers), and a Confusion Matrix.
- **Feature Importance**: The relative importance of each predictor variable was extracted from the trained model.

D. Market Basket Analysis (Association Rules)

- **Data Transformation**: Purchase counts for book genres were converted into a binary (boolean) format.
- **Apriori Algorithm**: Used to identify frequent itemsets with a minimum support threshold of 15%.
- **Rule Generation**: Association rules were generated, filtered by minimum confidence (50%) and lift (>=1.0), and ranked to identify the strongest co-purchase relationships.

```
Frequent Itemsets (Support >= 0.15):
    support              itemsets
2   0.41550             (CookBks)
0   0.39400            (ChildBks)
6   0.26675             (GeogBks)
3   0.25475            (DoItYBks)
7   0.24200   (ChildBks, CookBks)
1   0.23825            (YouthBks)
5   0.22300              (ArtBks)
4   0.20475              (RefBks)
10  0.16875   (DoItYBks, CookBks)
8   0.16150  (DoItYBks, ChildBks)
9   0.16100   (YouthBks, CookBks)
11  0.15625    (GeogBks, CookBks)
```

Figure 8. Frequent Itemsets

Association Rules (Lift >= 1.0 and Confidence >= 0.5):

	antecedents	consequents	support	confidence	lift
6	(YouthBks)	(CookBks)	0.16100	0.675761	1.626380
4	(DoItYBks)	(ChildBks)	0.16150	0.633955	1.609022
2	(DoItYBks)	(CookBks)	0.16875	0.662414	1.594258
0	(ChildBks)	(CookBks)	0.24200	0.614213	1.478251
1	(CookBks)	(ChildBks)	0.24200	0.582431	1.478251
8	(GeogBks)	(CookBks)	0.15625	0.585754	1.409758

Figure 8. Association Rules

IV. Key Findings

A. *Target Variable & Data Characteristics*

- **Class Imbalance**: Only 8.5% of customers purchased the Florence book. SMOTE was used to address this during model training.
- **Data Quality**: The dataset was clean and complete for the used columns.

B. *Predictive Model Insights*

Model Performance:

- **Accuracy**: 85.0% - While seemingly high, this is below the baseline of simply predicting 'non-buyer' (91.5%).
- **Precision (Buyers):** 10.8% - Indicates a high rate of false positives; when the model predicts a buyer, it's often incorrect.
- **Recall (Buyers):** 10.5% - Indicates a very high rate of false negatives; the model identifies only a small fraction of the actual buyers.
- **Confusion Matrix**: Confirmed the model predicts few buyers (7+58=65) and correctly identifies only 7 of the 67 actual buyers in the test set.
- **Implication**: The model struggles significantly to reliably identify the target group (Florence buyers).

Key Drivers (Feature Importance):

- **Monetary Value (M):** ~19.9% importance. Past spending is the strongest predictor.
- **Time Since First Purchase (FirstPurch)**: ~13.6% importance. Customer tenure is highly relevant.
- **Recency (R)**: ~12.0% importance. How recently a customer purchased matters significantly.
- **Gender:** ~7.2% importance. A notable demographic differentiator.
- **Recency Code (Rcode)**: ~6.8% importance. Reinforces the importance of recency.

RFM Code Means: While Mcode and Fcode means were similar between groups, the mean Rcode was lower for buyers (2.99 vs 3.19), supporting the importance of Recency (lower code = more recent)

Related Purchase Mean: Buyers had a higher mean count of related purchases (1.38 vs 0.85 for non-buyers), confirming its relevance, though it wasn't a top 5 feature in the RF model.

C. *Market Basket Analysis Insights*

Strong Associations Found: Several actionable co-purchase patterns were identified.

Top 3 Rules (by Lift):
{YouthBks} -> {CookBks} (Lift: 1.63, Confidence: 67.6%)
{DoItYBks} -> {ChildBks} (Lift: 1.61, Confidence: 63.4%)
{DoItYBks} -> {CookBks} (Lift: 1.59, Confidence: 66.2%)

Interpretation: These rules highlight segments likely composed of families or individuals interested in practical skills and home activities. The high confidence and lift values indicate strong, non-random associations useful for cross-marketing.

V. Recommendations

A. *Targeting Strategy for Specialty Books*

Focus on High-Value/Engaged: Prioritize customers with high M, long FirstPurch, and low R/Rcode. These are the clearest signals, despite the overall model's limitations.

Utilize Gender Insights: Analyze the specific gender preference indicated by the model and tailor campaigns if the difference is substantial.

Use Model as a Secondary Filter: Given the low precision and recall, do not rely solely on the model to find buyers. Use it to potentially score or rank leads generated through other methods (e.g., RFM segmentation, past purchase of highly correlated genres like ItalArt/ItalCook from EDA) rather than as a primary selection tool. Be prepared to accept a lower conversion rate if using the model's predictions directly for outreach.

B. *Leverage Association Rules for Cross-Selling & Bundling*

The market basket analysis provides strong, reliable insights. Actively implement cross-promotions:

- Recommend Cookbooks when Youth or DIY books are viewed/added to cart.
- Recommend Children's books when DIY books are viewed/added to cart.

Create thematic bundles ("Family Activities," "Home Skills") with discounts.

Use these rules to inform email marketing segments and website personalization.

C. Model Improvement Strategies

Address Imbalance Differently: Experiment with other techniques beyond SMOTE, such as undersampling the majority class or using algorithms inherently better suited to imbalance (like balanced Random Forest implementations or specific boosting algorithms with appropriate parameter tuning).

Feature Engineering: Create new features that might capture interactions or patterns better (e.g., ratio of Art books purchased, flags for buying specific combinations identified in EDA).

Alternative Models: Test simpler models like Logistic Regression (with class weights and potentially interaction terms) or more complex ones like Gradient Boosting (carefully tuned) to see if recall/precision can be improved without sacrificing too much accuracy.

-

VI. CONCLUSION

The analysis confirms that customer value, tenure, and recency are key indicators for identifying potential buyers of the specialty Florence book, along with gender. However, the predictive model built on these features struggles to accurately isolate this small buyer group, suggesting caution in its direct application for lead generation. In contrast, the market basket analysis yielded strong, actionable rules about co-purchase behavior, particularly linking DIY/Youth books with Cookbooks/ChildBks. The primary recommendation is to leverage these robust association rules for cross-selling and bundling, while using the predictive model's insights (especially RFM and gender) as supplementary factors in refining broader marketing campaigns for specialty titles, acknowledging the need for further model improvement to reliably predict buyers.