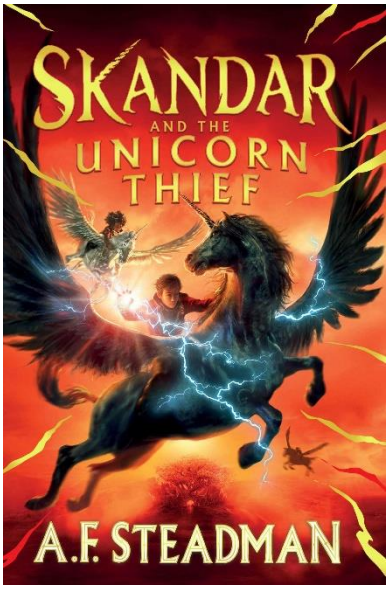
Case 3 – Skandar and the Unicorn Thief

In an era where pixels populate most of our reading experiences, Canada's children's book clubs are fostering a literary renaissance with a tangible twist. Despite the world's move to the digital sphere, there's a resurgence of something more traditional—physical mailers. These are not just simple pieces of paper but are gateways to fantastical worlds, meticulously crafted to entice young readers and their guardians into the magical realm of books.



The joy of receiving a physical mailer from a book club is akin to finding a treasure in one's mailbox. It's a sensory experience that a digital screen cannot replicate. The texture of high-quality paper between one's fingertips, the vibrant colors springing to life in the illustrations, and the distinct scent of ink all contribute to a child's excitement and anticipation. This excitement often translates into a child's eagerness to engage with the books themselves, a marketing marvel that online images alone cannot achieve.

These mailers are more than a list of recommended titles; they often contain engaging content such as author interviews, fun facts, and reading challenges that can be shared among friends and family. They might include stickers, bookmarks, or other book-related trinkets that celebrate the joy of reading in a way that digital content simply can't. Some might consider this an environmentally unfriendly approach in a world that is increasingly aware of its carbon footprint. However, many book clubs are countering this by using recycled materials and eco-friendly inks, thus embracing a more sustainable practice that aligns with the values of their readers and their families.

Magic Book Nook

Founded in 1990, Magic Book Nook (MBN thereafter) has become a distinguished entity in the world of children's book distribution through the channel of direct marketing. MBN dispatches physical mailers to its clients frequently as a core marketing strategy.

MBN is currently exploring the potential benefits of using predictive modeling approaches to enhance the effectiveness of its direct mail program. In a recent mailing campaign, the company selected 10,000 customers in Ontario from its database and included with their regular mailing a specially produced brochure for the book: "Skandar and the Unicorn Thief". MBN then developed a predictive model to identify the factors influencing these customers' purchase decisions.

For the purpose of this case analysis, we will utilize two subsets of the database available to MBN. The first dataset is a training sample of 1,400 customers. All of these customers received the brochure. The dependent variable for the analysis is choice, indicating whether customers placed an order for "Skandar and the Unicorn Thief" after receiving the brochure. MBN also selected several independent variables believed to explain the observed choice behavior. This dataset will be used to calibrate the predictive model. Below is a description of the variables used for the analysis:

- **Choice**: Whether the customer purchased "Skandar and the Unicorn Thief". Yes =1 and No= 0
- **Gender**: 0 = Female and 1 = Male. This categorial variable has already been encoded as a number. Therefore, you don't need to do one-hot encoding in python.
- **Amount spent**: Total money spent on MBN books.
- **Frequency**: Total number of purchases in the prior three year.
- **Last purchase** (recency of purchase): Number of months since the last purchase.
- **Tenure**: Number of months since the first purchase.
- **P_fiction**: Number of fiction books purchased in the prior year.
- **P_nonfiction**: Number of nonfiction books purchased in the prior year.

To assess the performance of the model, MBN has provided an additional dataset of 2,000 customers as a validation sample. All of these customers in the validation sample also received the brochure, and the company observed whether they purchased the book or not. They are included to validate the model's performance beyond the training data.

Case questions

Q1. What proportion of customers made a purchase (choice = 1) in the training data? Build a binary logistic regression model using all the variables in the training data. Use a linear format, e.g., y ~ x1 + x2 + x3 … and do not utilize polynomial, spline functions, or interactions). Provide a screenshot of the model summary (use the statsmodels library for this task). (5 points)

Q2. Interpret the results of the models. Discuss how these variables influenced customers' decision to buy or not to buy the book. Include information about the significance, the direction of impact, and the magnitude of impact on purchase probability (i.e., the impact on purchase probability if X increases by 1 unit – this is called the marginal effect). (15 points)

Q3. Report the confusion matrix and accuracy using the training sample. Does the model perform equally well in predicting purchases and non-purchases? Note: the logistic regression lecture code has been updated on myCourse to fix a bug that caused an error when running the classification_report for the second time. (5 points)

Q4. Assess the performance of the model in the validation sample, using a confusion matrix and accuracy. Does the model perform equally well in predicting purchases and non-purchases? (5 points)

Q5. Build a classification tree with 2 levels (max_depth=2, random_state=603) and visualize it. (5 points)

Q6. What would the classification tree predict (purchase or non-purchase) for the first observation in the validation sample (customer = 1), and which node will it ultimately end up in? (5 points)

Q7. Build a random forest model for the training data using all the variables (30 trees, max depth = 4, random_state=603). What is the OOB accuracy? What are the top three most important variables in predicting the outcome? (10 points)

Q8. Assess the performance of the random forest model in the validation sample by reporting the confusion matrix and accuracy. Which model would you prefer, the random forest model or the logistic regression model (Q1)? And why? (10 points)

Q9. MBN is considering a similar mail campaign in British Columbia. They would like to use the data obtained from Ontario to project what would happen in British Columbia, assuming consumer behavior is similar.

The allocated cost of the promotional brochure and mailing is $2 per addressee including postage. For every purchase (conversion from the promotional brochure), the actual book costs MBN $20 to procure and mail and to account for overhead. The selling price of the book is $42. The management would like to evaluate the profitability of different promotion strategies based on the observed behavior in the validation sample.

Evaluate the profitability of a blanket promotion strategy without using a predictive model -- sending the promotional brochure for the book "Skandar and the Unicorn Thief" to everyone. Use the validation sample of 2,000 customers for this evaluation. How many customers made a purchase in the validation sample? Calculate total revenue, total costs, and net profits of this promotion strategy on the 2,000 customers. Explain how you arrive at these final numbers. Hint: customers in the validation sample had all received the brochure, and their actual choices (to buy or not to buy) are recorded in the choice column of the validation dataset. (20 points)

Q10. A potentially better way to execute the direct mail campaign is to leverage the predictive model. A straightforward strategy is to send the promotional brochure exclusively to the top 50% of customers (based on their predicted purchase probability). Use the logistic regression model from Q1 and the validation sample to assess the profitability of this targeting strategy. How many of the targeted customers (top 50%) would actually make a purchase? Calculate the total revenue, total costs, and net profits of this promotional strategy for the 2,000 customers. Explain how you arrive at these final numbers.

Hint: This question requires a what-if analysis. For example, if the company did not mail the brochure to some customers in the validation set, those customers would NOT have made a purchase since they wouldn't be aware of MBN's offer for this specific book. In other words, not sending the brochure to certain customers would result in missed opportunities to sell the book. However, not all customers have an equal likelihood of buying the book. Sending promotional brochures to customers unlikely to make a purchase would be a wasteful use of resources. Thus, a more effective strategy is to target customers with a high purchase probability (e.g., ranking them by purchase probability and sending offers to the top 50% or top 1,000 customers). However, the success of this strategy hinges on your ability to construct a predictive model for customers' purchase probabilities. (20 points)

Appendix: Python code to load data

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, classification_report
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.ensemble import RandomForestClassifier

# load training and validation datasets
bookclub_training =
pd.read_csv('https://raw.githubusercontent.com/yumayuma/retl603/main/bookclub_training.csv')
bookclub_val = pd.read_csv('https://raw.githubusercontent.com/yumayuma/retl603/main/bookclub_validation.csv')
```