**Bookclub**

**RETL 603**

**Professor Yu Ma**

**Ellen Jia**

**November 27th, 2023**

**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)**

文本

描述已自动生成

文本

描述已自动生成表格

描述已自动生成

Proportion of Purchases in Training Data: Approximately 28.57% of customers in the training dataset made a purchase of "Skandar and the Unicorn Thief".

**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)**

图形用户界面, 文本, 应用程序

描述已自动生成表格

描述已自动生成

1. **Gender**
   * Coefficient: 1.3417
   * Marginal Effect: 0.2163
   * p-value: < 0.001 (highly significant)
   * Interpretation: The positive coefficient indicates that being male (coded as 1) increases the log odds of purchasing by 1.3417 compared to being female. The probability of purchase for males is 21.63 percentage points higher than for females, holding other factors constant.
2. **Amount Spent**
   * Coefficient: 0.0020
   * Marginal Effect: 0.0003
   * p-value: 0.013 (significant)
   * Interpretation: A small but positive coefficient suggests that each additional dollar spent increases the log odds of purchasing by 0.0020. For every additional dollar spent, the probability of making a purchase increases marginally by 0.03%.
3. **Frequency**
   * Coefficient: -0.0758
   * Marginal Effect: -0.122
   * p-value: 0.003 (significant)
   * Interpretation: The negative coefficient indicates that each additional past purchase decreases the log odds of making a new purchase by 0.0758. The probability of making a new purchase decreases by 12.2% for each additional purchase made previously.
4. **Last Purchase**
   * Coefficient: -0.0979
   * Marginal Effect: -0.0158
   * p-value: 0.110 (not significant)
   * Interpretation: Although the coefficient is negative, suggesting a decrease in the log odds of purchase by 0.0979 with more recent last purchases, it is not statistically significant. Therefore, we cannot reliably interpret its impact on purchase probability.
5. **Tenure**
   * Coefficient: -0.0193
   * Marginal Effect: -0.0031
   * p-value: 0.106 (not significant)
   * Interpretation: The negative coefficient suggests a decrease in the log odds of purchase by 0.0193 with longer tenure, but this is not statistically significant, so its impact on purchase probability is uncertain.
6. **P\_fiction**
   * Coefficient: 0.0090
   * Marginal Effect: 0.0014
   * p-value: 0.729 (not significant)
   * Interpretation: The positive coefficient implies an increase in the log odds of purchase by 0.0090 with a preference for fiction; however, due to the lack of statistical significance, it does not provide a reliable indication of impact on purchase probability.
7. **P\_nonfiction**
   * Coefficient: 0.5401
   * Marginal Effect: 0.0871
   * p-value: < 0.001 (highly significant)
   * Interpretation: The positive coefficient indicates a strong relationship, with each additional nonfiction book purchased increasing the log odds of making a purchase by 0.5401. The probability of purchase increases by 8.71% with each nonfiction book purchased.

**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)** 表格

描述已自动生成

Confusion Matrix: True Negatives (928), False Positives (72), False Negatives (255), True Positives (145).

Accuracy: The accuracy is approximately 76.64%. Which indicates that the model correctly predicted the purchase decision (both yes and no) for about 76.64% of the cases in the training sample.

Precision and Recall: For non-purchase (0): high precision (78%) and recall (93%). For purchase (1): lower precision (67%) and much lower recall (36%).

The model appears to be better at predicting non-purchases than purchases, as indicated by higher precision and recall for the non-purchase category.

**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)** 文本

描述已自动生成手机屏幕截图

描述已自动生成

Confusion Matrix: True Negatives (1769), False Positives (27), False Negatives (197), True Positives (7)

1. **Overall Accuracy:**
   * The model correctly predicted 1776 out of 2000 instances (1769 non-purchases and 7 purchases), resulting in an overall accuracy of 89%
2. **Accuracy for Predicting Non-Purchases (Class 0):**
   * The model correctly predicted non-purchases 1769 times out of 1796 attempts (1769 true negatives and 27 false positives).
   * The accuracy for predicting non-purchases is therefore 1769 / (1769 + 27) \* 100% = 98.50%.
3. **Accuracy for Predicting Purchases (Class 1):**
   * The model correctly predicted purchases 7 times out of 204 attempts (7 true positives and 197 false negatives).
   * The accuracy for predicting purchases is therefore 7 / (197 + 7) \* 100% = 3.43%.

**F1-Scores:**

* The F1-score for non-purchases is 0.94, indicating a high precision and recall balance for predicting non-purchases.
* The F1-score for purchases is 0.06, which is very low, indicating that the model is poor at balancing precision and recall for predicting purchases.

In conclusion, while the overall accuracy might seem high, the class-specific metrics reveal that the model's performance is much better for predicting non-purchases than for predicting purchases. The low F1-score for purchases confirms that the model is not effective in identifying actual purchases, which is a critical insight when the goal is to accurately predict both classes.

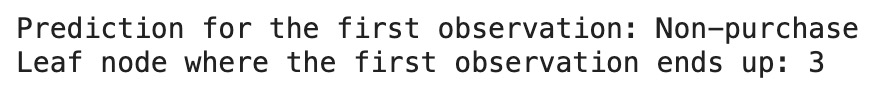
**Q5. Build a classification tree with 2 levels (max\_depth=2, random\_state=603) and visualize it. (5 points)**

图形用户界面, 文本, 应用程序

描述已自动生成图示, 日程表

描述已自动生成

**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)** 图形用户界面, 文本

描述已自动生成

Validation dataset:

Customer: 1

Choice: 1 (indicating a purchase)

Gender: 0

Amount Spent: 287

Frequency: 12

Last Purchase: 4

Tenure: 24

Purchases of Fiction: 6

Purchases of Non-fiction: 1

The root node checks if p\_nonfiction <= 2.5. Since our observation has a p\_nonfiction value of 1, we move to the left child node. The left child node checks if frequency <= 11.0. Our observation has a frequency of 12, so we move to the right child node. The right child node is a leaf node, with the class = No Purchase.

**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)** 图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成文本

描述已自动生成

The OOB accuracy of the random forest model is approximately 77%. This metric is a measure of the prediction error of the random forest model when predicting unseen data.

The following variables are key indicators in predicting whether a customer will purchase the book:

* P\_nonfiction: The number of nonfiction books purchased in the prior year is the most important variable, indicating a strong influence on the likelihood of purchasing "Skandar and the Unicorn Thief".
* P\_fiction: The number of fiction books purchased in the prior year is the second most important variable.
* Frequency: The total number of purchases in the prior three years is the third most important variable.

**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)** 图形用户界面, 文本, 应用程序

描述已自动生成图形用户界面, 文本, 应用程序

描述已自动生成手机屏幕截图

描述已自动生成

Confusion Matrix for Random Forest Model: True Negatives (1750), False Positives (46), False Negatives (197), True Positives (7)

Accuracy on Validation Data: The random forest model has an accuracy of approximately 87.85% on the validation data.

Precision and Recall for Random Forest Model: For non-purchase (0): High precision (90%) and recall (97%). For purchase (1): Low precision (13%) and very low recall (3%).

Comparison of Logistic Regression and Random Forest Models:

Logistic Regression Model:

* Better at predicting non-purchases (high precision and recall for non-purchase).
* Struggles with predicting purchases (low precision and recall for purchase).
* Overall accuracy: 88.8%.

Random Forest Model:

* Similar performance in predicting non-purchases.
* Slightly worse performance in predicting purchases.
* Overall accuracy: 87.85%.

In evaluating the preferred model for the dataset, the logistic regression model seems slightly better suited for this dataset, especially given its marginally higher overall accuracy and similar performance characteristics. Additionally, the logistic regression model also offers the advantage of interpretability, allowing for a clearer understanding of how different variables influence the likelihood of a purchase.

**Q9. 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)** 文本

描述已自动生成

Key factors to consider:

* Cost per Brochure: $2
* Cost of Procurement and Mailing per Purchase: $20
* Selling Price per Book: $42
* Customers' Actual Choices: Recorded in the “choice” column of the validation dataset.

Profitability Analysis:

* Number of Customers Who Made a Purchase: In the validation sample, 204 customers made a purchase.
* Total Revenue: The total revenue generated from these purchases is $8,568.
* Total Costs: The total cost of the promotion, including the cost of sending brochures to all 2,000 customers and the additional costs per purchase, is $8,080.
* Net Profit: The net profit from this blanket promotion strategy is $488.

This analysis suggests that while the blanket promotion strategy is profitable, the margin is relatively small given the scale of the campaign.

**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. 图形用户界面, 文本, 应用程序

描述已自动生成文本

描述已自动生成**

Profitability Analysis:

* Predict Purchase Probabilities:
  + Using the logistic regression model from Question 1, we predicted the purchase probabilities for each customer in the validation sample.
* Rank Customers by Purchase Probability:
  + We ranked the customers based on their predicted probabilities of making a purchase.
* Select Top 50% of Customers:
  + We selected the top 50% of customers who are most likely to make a purchase according to the model's predictions.
* Calculate the Number of Actual Purchases:
  + We counted how many of these targeted customers actually made a purchase (as indicated in the validation data).
* Calculate Costs and Revenue:
  + Costs: Included the cost of sending brochures to the top 50% of customers and the additional cost per actual purchase made.
  + Revenue: Calculated based on the number of actual purchases and the selling price per book.
* Determine Net Profit:
  + Net profit was calculated as total revenue minus total costs.

Based on this analysis, the results were:

* Number of Actual Purchases by Targeted Customers: 118
* Total Revenue: $4,956 (calculated as 118 purchases \* $42 per book)
* Total Costs: $4,360 (calculated as $2 \* 1000 for brochure distribution + $20 \* 118 for purchase costs)
* Net Profit: $596 (calculated as $4,956 in revenue - $4,360 in costs)

Overall, the targeted promotional strategy is more profitable than the blanket strategy, yielding a higher net profit while reducing the scale and costs of the campaign. This analysis demonstrates the effectiveness of using predictive modeling to optimize marketing strategies and improve profitability.