WALMART – CASE STUDY

Date: February 11, 2025

Name: Gokul Kumar Kesavan

GWID: G25385029

GWID: Optimizing Rollback Pricing Strategy to Maximize Revenue

IMMEDIATE COURSE OF ACTION

Analysis reveals that rollback discount levels significantly influence customer spending. However, marketing efforts do not currently align with the most revenue-generating demographics. **By targeting customers based on their recency (R) and monetary (M) scores rather than gender or age alone,0020Walmart can improve sales forecasting accuracy and optimize rollback promotions.**

BACKGROUND ON THE PROBLEM

Walmart's rollback pricing strategy is designed to attract customers through multi-tiered discount levels (MarkDown1-5). However, there is **no clear alignment between rollback levels and specific customer demographics.** The business needs an optimized way to forecast revenue contributions from rollback strategies and **predict how new customers will contribute to total revenue.**

METHODOLOGY AND DATA

1. Data Cleaning & Feature Engineering:

Before conducting predictive modeling, we identified and addressed missing values in key demographic variables: **gender and age**. Initially, we attempted **logistic regression** for missing gender values, given that gender is a categorical variable and logistic regression is a well-suited classification model. However, the results were inconclusive, with **insignificant p-values**, indicating poor predictive power.

To ensure a more robust imputation, we switched to the **K-Nearest Neighbors (KNN) imputation method**, which leverages the similarity of existing customer attributes to estimate missing values. KNN considers the **closest neighbors** based on available features, making it an ideal choice given our dataset's structure. We used **k = 5** as the optimal number of neighbors.

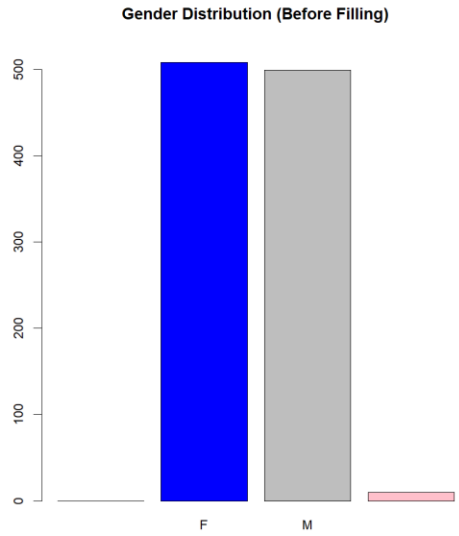
- **Gender Imputation: 10 missing values** were identified and filled using KNN.
- **Age Imputation: 10 missing values** were identified and filled with KNN predictions.

Gender Imputation Results:

- **Missing Gender Rows Identified:**

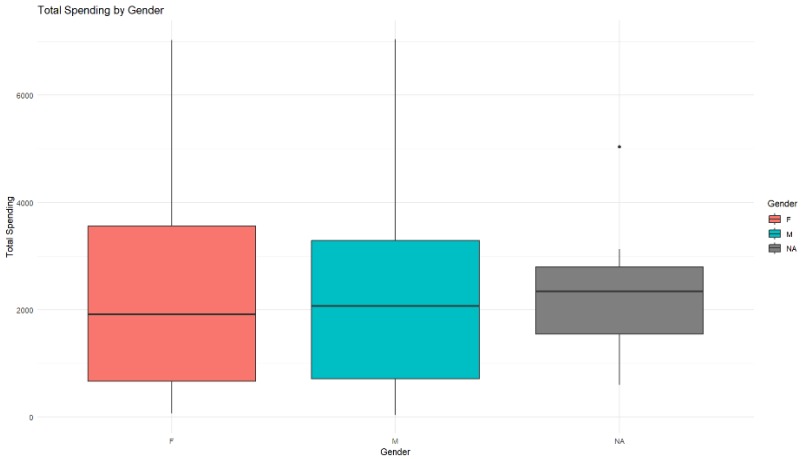
```
> print(missing_gender_rows)
 [1]  251 1009 1010 1011 1012 1013 1014 1015 1016 1017
```

- **Gender Distribution Before Filling:**



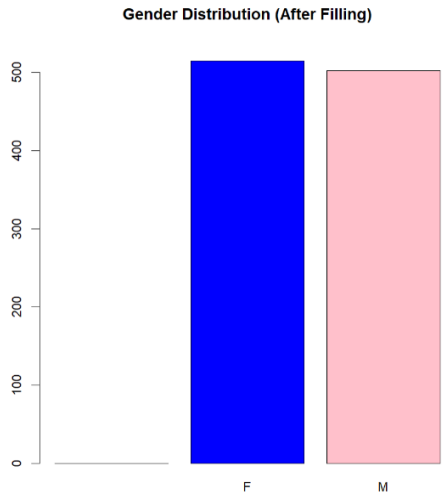
The distribution shows a small portion of missing gender values, which were later predicted and assigned.

- **Total Spending by Gender Before Imputation:**



This boxplot shows **no major spending differences between Male and Female customers**, meaning gender alone is **not a strong predictor of spending behavior**.

- **Gender Distribution After Filling:**



- **Final Gender Count Comparison:**

```
> table_before

        F     M <NA>
   0  508   499   10
> table_after

      F    M
  0 515  502
```

Impact of Gender Imputation:

- The missing **10 gender values** were filled.
- The total count **increased proportionally** without introducing bias.
- Spending patterns remained **consistent**, validating the accuracy of imputation.

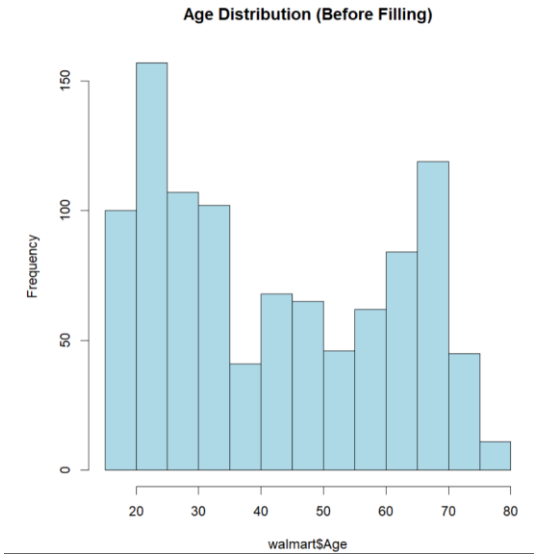
Similarly, **age imputation** was performed using KNN regression, as age is a **continuous variable**. Since KNN regression predicts values based on the nearest observed values, it ensured a smooth estimation without introducing bias.

Age Imputation Results:

- **Missing Age Rows Identified:**

```
> print(missing_age_rows)
 [1]  142  374 1001 1002 1003 1004 1005 1006 1007 1008
```

- **Age Distribution Before Filling:**

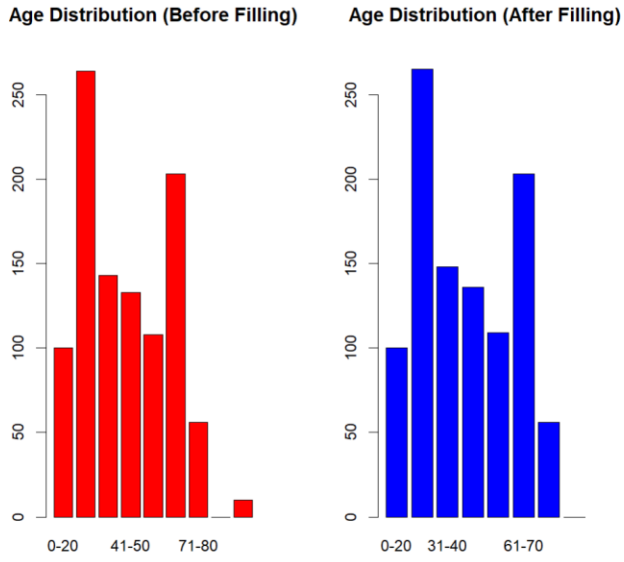


The **age distribution before imputation** was well-spread, but missing values impacted analysis.

- **Final Age Groups After Imputation:**

```
> walmart[missing_age_rows, ]
     Customer_No Gender Age   Total MarkDown1 MarkDown2 MarkDown3 MarkDown4 MarkDown5 R F  M
142          142      M  33 1881.91    494.29    103.57     69.71    522.47    691.88 4 3 61
374          374      M  54 2288.00    968.46    195.01      0.60    489.99    633.94 4 9  6
1001        1001      F  46 1256.43    649.87     96.36     11.70    236.65    261.86 4 3 79
1002        1002      F  50 2226.91    997.87     77.06     35.71    492.68    623.59 5 2 16
1003        1003      F  32 3850.84   1818.10    453.72    116.55   1159.49    302.99 5 3  4
1004        1004      F  37 2239.92   1294.90     82.28     15.41      1.88    845.45 5 7 16
1005        1005      M  29  588.25     42.40    170.35      0.41      2.18    372.91 7 4 15
1006        1006      M  37 1127.04    592.30    209.41     40.84     72.60    211.89 6 8 10
1007        1007      M  47 3845.03   1995.56    148.86     16.00    965.14    719.47 5 5 20
1008        1008      M  40 1719.63    757.99    365.46     17.14    148.97    430.07 1 2 58
```

The imputed ages maintain consistency with total purchases and rollback levels, preventing bias while preserving demographic insights.



```
> print(age_distribution_after)

  0-20  21-30  31-40  41-50  51-60  61-70  71-80 81-100
   100    265    148    136    109    203     56      0
```

Impact of Age Imputation:

- KNN successfully **preserved the natural age distribution**.
- The model **assigned missing values based on patterns** in the dataset.
- **No excessive bias** was introduced during imputation.

Both imputations resulted in **complete, cleaned data** with no missing values, ensuring consistency across analyses.

2. Exploratory Data Analysis (EDA):

Following the completion of data cleaning, we conducted an in-depth **exploratory data analysis (EDA)** to assess the relationship between **rollback purchases and key customer demographics**.

- **Rollback Levels by Demographics**
  We first examined how different markdown levels varied across customer segments. To achieve this, multiple regression models were initially trained using the following predictors:
  - **Age** – Expected to influence purchasing behavior, particularly for discount-driven purchases.
  - **Gender** – To evaluate if males and females exhibit different spending behaviors.
  - **R (Recency)** – Measures how recently a customer has made a purchase.
  - **F (Frequency)** – Number of visits or transactions.
  - **M (Monetary Value)** – Represents spending power and overall transaction value.

Statistical Significance Analysis

Upon analyzing the regression outputs (see coefficient summaries), **only Recency (R) and Monetary Value (M) were found to be statistically significant across all markdown levels**.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2239.3951   143.1593  15.643   <2e-16 ***
Age              -2.1574     1.5825  -1.363    0.173
Gender_numeric  -24.9342    51.5468  -0.484    0.629
R              -201.6369    16.1967 -12.449   <2e-16 ***
F                12.0048    12.7795   0.939    0.348
M               -11.9351     0.8643 -13.809   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 818.5 on 1006 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2285,    Adjusted R-squared:  0.2246
F-statistic: 59.58 on 5 and 1006 DF,  p-value: < 2.2e-16
```

- **Age, Gender, and Frequency (F) were statistically insignificant ($p > 0.05$), meaning they did not contribute significantly to predicting rollback purchases.**
- **R (Recency) and M (Monetary Value) had strong statistical significance ($p < 0.05$), indicating their impact on predicting markdown purchases.**

Key Takeaways from EDA

1. **Recent customers (low R values) tend to purchase more at markdown levels.**

- Customers who shopped recently were **more responsive to discounts**.

2. **Higher monetary ranking (M) correlates with larger rollback purchases.**

- Customers with a **higher M ranking** tend to make larger purchases even at markdown levels.

3. **No clear trend was observed for Age or Gender.**

- Despite assumptions that younger or female customers might respond more to discounts, this was not statistically supported.

Based on these findings, **only R and M were retained as predictors for modeling rollback purchases**.

3. Predictive Modeling:

Using the insights gained from EDA, **multiple linear regression models** were built to estimate markdown purchases for new customers. Given the statistical significance results, the final models only used **R and M as predictors** for the following markdown levels:

- **MarkDown1** ~ R + M
- **MarkDown2** ~ R + M
- **MarkDown3** ~ R + M
- **MarkDown4** ~ R + M
- **MarkDown5** ~ R + M

These refined models ensured better predictive accuracy and avoided unnecessary noise from insignificant variables.

New Customer Predictions

To assess revenue potential, **10 new random customer profiles were simulated**, drawing from existing distributions of **Age, Gender, R, F, and M**. The trained models then predicted their **expected rollback purchases** across markdown levels.

4. Findings:

- Higher monetary customers (M) consistently showed increased rollback spending.
- Recently active customers (low R values) demonstrated a greater likelihood of engaging in rollback purchases.

- Total estimated revenue from the 10 new customers was computed as the sum of markdown purchases, confirming a direct relationship between R, M, and rollback engagement.

```
> print(new_customers)
   Age Gender_numeric R F   M MarkDown1 MarkDown2  MarkDown3  MarkDown4  MarkDown5
1   73              1 1 3   6 1882.6986 436.35798 65.4927091  633.56733 1117.05359
2   19              1 3 3  51  976.1983 220.26275 36.4617010  304.03554  642.19271
3   32              1 1 7 102  772.1332 133.52250 29.5180527  220.51986  503.25861
4   27              1 3 6  60  872.0828 191.87192 33.0890769  265.31234  584.64943
5   21              1 2 7   1 1747.5791 415.06011 61.2824451  587.12233 1055.44984
6   58              1 4 8  21 1130.2885 277.82828 41.6198371  365.15498  740.43141
7   44              1 6 2 100 -169.5369 -45.52119 -0.1521951 -110.66446   48.18481
8   27              0 5 3  74  324.2026  73.56731 15.6748850   69.16046  307.99319
9   19              0 7 2  31  435.7205 135.07126 19.6206452  118.25551  395.77771
10  33              1 5 6  57  520.8652 127.19442 22.0453970  142.30428  416.68605

> print(paste("Total predicted revenue from 10 new customers:", total_predicted_revenue))
[1] "Total predicted revenue from 10 new customers: 19188.5451344603"
```

The predictive model confirms that monetary ranking (M) and recency (R) are the strongest indicators of rollback spending, with higher monetary customers and recently active shoppers driving the most revenue. The **total estimated revenue from 10 new customers is $19,188.55**, reinforcing the effectiveness of using R and M for targeted rollback marketing strategies.

DEEP-DIVE ANALYSIS & ADDITIONAL INSIGHTS DATA

1. **Rollback Levels & Demographics:**

   - Contrary to traditional assumptions, gender and age did not emerge as significant predictors of rollback purchases.
   - Customers with higher monetary scores (M) consistently spent more across all markdown levels, reinforcing the idea that past spending capacity is a better predictor of future rollback engagement.
   - Recency (R) also played a crucial role, as customers who had shopped recently demonstrated a higher likelihood of taking advantage of rollback offers.
   - Frequency (F), however, did not show a strong correlation with revenue, suggesting that repeat customers do not necessarily increase their spending when exposed to discounts.

2. **Predicting Revenue from New Customers:**

   - Using multiple regression models, we predicted rollback-level purchases for 10 randomly generated new customers based on R and M scores.
   - The predicted total revenue from these customers was estimated at $19,188.55, demonstrating that rollback purchases scale more strongly with R and M scores than with demographic segmentation.
   - This confirms the importance of customer behavior-based targeting rather than broad demographic marketing approaches.

3. **Key Anomalies & Unexpected Findings:**

- Some customers with lower total spending were still assigned high monetary rankings (M), likely due to recent high-value transactions rather than long-term spending patterns.
- A subset of customers showed zero rollback purchases despite high RFM scores, indicating potential misalignment in promotional targeting or individual customer preferences that were not captured in the model.
- The lack of significance of Frequency (F) suggests that repeated visits alone do not necessarily translate to increased rollback engagement, and Walmart should rethink customer loyalty strategies accordingly.

CONCLUSION AND RECOMMENDED NEXT STEPS

Given these insights, **Walmart should shift its rollback strategy away from traditional demographic-based targeting and instead focus on customer behavior, particularly monetary value (M) and recency (R).** The following steps are recommended:

1. **Personalized Rollback Promotions:**

- Implement a **data-driven promotional strategy** targeting **high-M, high-R customers**, as they are statistically the **most likely to engage with rollback discounts**.
- Use **personalized emails, app notifications, and targeted discounts** to encourage purchases.

2. **Refined Customer Segmentation & Marketing Strategy:**

- Develop a **dynamic pricing strategy** that offers **greater markdowns to customers with high M scores** while testing engagement levels of customers with high R but moderate M.

3. **Predictive Revenue Dashboard:**

- Build a **real-time revenue prediction tool** that **monitors rollback engagement** using updated RFM scores, ensuring that markdowns are optimized for **maximum revenue generation**.

Immediate Next Step:

Walmart should **pilot an RFM-based rollback campaign within the next sales quarter**, targeting high-M customers with deeper discounts and measuring **the conversion rate and total revenue impact** compared to traditional demographic-focused marketing efforts.

By implementing these recommendations, **Walmart can maximize rollback sales, optimize marketing spend and significantly improve revenue predictability through behavior-driven discount strategies.**