"I’m Just Browsing”
Predicting the Value of Prospective Customers

Estella Dentinger & Brittany Nguyen

Faculty Advisor: Professor Georgia Perakis
PhD Advisor: Leann Thayaparan
Macy's Team: Iris Singhania, Yixin Cai, Donghao Pei
What is a Prospective Customer?

Why are they important? Understanding the potential future value of customers who have engaged with Macy’s, but have not made any purchases is critical to **new customer acquisition**.

Our Project: Understand who the valuable prospective customers are, and how to activate their first purchase and retain them.
Problem Overview

Current State
Macy’s has models to predict the future value of active customers by using their historical purchase data.

Limitation
Prospective customers, by definition, do not have purchase history.

Our Approach
Use prospective customer online activity data to predict their value.
Our Approach: Constructing the Dataset

**Prospective customers** had online activity but no purchases

Prospective customers had online activity but no purchases between Feb 2020 and Jan 2022. New customers have never made a purchase prior to 2022. Inactive customers have purchase history prior to Feb 2020.

Predict prospective customer value in 2022 between Feb 2022 and Jan 2023.

**Features**

- **Click behaviors**
  - Search, browse, add to cart, page view, abandon cart, and others

- **User Profile**
  - Loyalty status, length of loyalty, new/inactive

**Data Limitations**

- **Imbalanced Dataset**
  - Only 8% of prospective customers made a purchase in 2022 - spend is skewed

- **Skewed Distribution for Online Activity**
  - Majority of values indicate little activity

- **Missing Values**
  - Removed demographic and income features
Our Approach: Predictive Modeling

1. **Binary Classification** model to predict whether a customer will purchase in next fiscal year

2. **Regression** model to predict the dollar amount that a customer will spend in next fiscal year

3. **Multi-Classification** model to predict zero/low/high spend in next fiscal year

### 3 Key Questions

1. Which prospective customers will make a purchase?
2. How much will prospective customers spend?
3. Who are the high value prospective customers?
Model Validation through Backtesting

Backtesting: training on recent customer trends and testing on historical data

1. Train model on more recent data

   Jul 2020 – Jun 2022
   Prospective customer online activity

   Jul 2022 – Jun 2023
   Predict prospective customer CLV

2. Test model on older data

   Jul 2019 – Jun 2021
   Prospective customer online activity

   Jul 2021 – Jun 2022
   Predict prospective customer CLV
Our Models Improve upon Existing Methodologies

84% Accuracy of our Best Model +7% Over Baseline

73% Recall of our Best Model +19% Over Baseline

79% AUC of our Best Model +12% Over Baseline

Out-of-Sample Backtesting Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM Binary</td>
<td>0.84</td>
<td>0.73</td>
<td>0.79</td>
</tr>
<tr>
<td>GBM Multiclass</td>
<td>0.79</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>Baseline (Active Customer Churn)</td>
<td>0.77</td>
<td>0.54</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Our Best Model: Binary CatBoost GBM
Baseline: Existing Customer Churn Models
**Top Drivers of Prospective Customer Value**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>New vs. Inactive Customer</strong></td>
<td>Whether a customer is new or inactive</td>
</tr>
<tr>
<td>2</td>
<td><strong>Account Creation</strong></td>
<td>Whether or not a customer signed up for a Macy’s account as a loyalty member or non-loyalty member</td>
</tr>
<tr>
<td>3</td>
<td><strong>Email Opt-In</strong></td>
<td>Whether or not a customer opted-in to email marketing</td>
</tr>
<tr>
<td>4</td>
<td><strong>Count of SMS Sent</strong></td>
<td>Number of SMS messages delivered to customer</td>
</tr>
<tr>
<td>5</td>
<td><strong>Search</strong></td>
<td>Number of days with a search in the past 720 days</td>
</tr>
</tbody>
</table>

*Top Drivers of Binary GBM Model*
# Post-Modeling Analysis: Binary Model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Predicted Purchasers</th>
<th>Predicted Non-Purchasers</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Inactive Customers</td>
<td>91%</td>
<td>0.3%</td>
</tr>
<tr>
<td>% Acct Creation Loyal</td>
<td>6%</td>
<td>1%</td>
</tr>
<tr>
<td>% Email Opt-In</td>
<td>6%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Count of SMS Sent</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>Avg Number of Days w/ Searches Past 720 Days</td>
<td>4</td>
<td>3.8</td>
</tr>
</tbody>
</table>
Post-Modeling Analysis: Binary Model

43% of Predicted Purchasers are Loyalty Members
5% of Predicted Non-Purchasers are Loyalty Members

- Purchasers: 43% Loyalty, 57% Non-Loyalty
- Non-Purchasers: 95% Non-Loyalty, 5% Loyalty
Business Impact

- Targeted email campaigns to valuable prospective customers
- Guide customer personalization, engagement, and retention efforts and act as a data resource for teams across Macy’s

Next Steps

- Integration into active customer CLV workflow
- Predict CLV for future time frame 2023-2024
- Deployment of prospective customer CLV models
Acknowledgements

Thank you to:

- Iris Singhania, Yixin Cai, and Donghao Pei
- Professor Georgia Perakis and Leann Thayaparan
- MIT MBAn Program Team
Predictive Features Selected

- Online activity metrics (search, browse product, page view, add to cart, abandon) *
- Loyalty tier & age of loyalty
- Email opt-in & SMS opt-in flag
- App download flag
- Prospective customer flag (1=never made a purchase)
- Count & Duration (seconds) of visits
- Device medium for visits (mobile phone, mobile app, tablet, desktop)
- Source sites (Google, Facebook, Bing, etc.)
- SMS data (sent, clicked, ordered, click rate, days since sms sent, days since sms clicked, days since sms ordered) *

*note: time frame: across 30, 60, 90, 180, 360, 720 days, 2 years
Online Activity Average Counts are Similar Across Purchasers and Non-Purchasers

Add to Cart

Search

Page View

FY22 Purchase Flag

- No Purchase
- Purchase

*Outliers filtered out for all
## App Download, Email Marketability, Session Length

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
</table>
| App Download Flag             | • Binary  
  • Whether user downloaded Macy’s app                                                                                                   |
| Email Marketability Flag      | • Binary  
  • Whether user was email marketable at time of downloading Macy’s app                                                                       |
| Session Length                | • Total duration spent in seconds over last 7 days                                                                                         |
## SMS Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td># SMS Sent</td>
<td>• # of SMS messages sent to user</td>
</tr>
<tr>
<td># SMS Clicked (Total)</td>
<td>• # of SMS messages clicked by user</td>
</tr>
<tr>
<td># SMS Clicked (Unique)</td>
<td>• # of unique SMS messages clicked by user</td>
</tr>
<tr>
<td>Click Rate</td>
<td>• # SMS messages clicked / # sent</td>
</tr>
<tr>
<td>Days Since SMS Sent</td>
<td>• # of days since SMS was sent</td>
</tr>
<tr>
<td>Days Since SMS Clicked</td>
<td>• # of days since SMS was clicked</td>
</tr>
</tbody>
</table>

*Measured for all features over 30, 60, 90, 360, 720 days*
# Post-Modeling Analysis: Multi-Class Model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Predicted Zero Tier (spend = $0)</th>
<th>Predicted Low Tier* (spend &lt;= $119)</th>
<th>Predicted High Tier (spend &gt; $119)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Acct Creation Loyal</td>
<td>0%</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>% New Customers</td>
<td>99.9%</td>
<td>15%</td>
<td>19%</td>
</tr>
<tr>
<td>Avg Number of Days w/ Searches Past 360 Days</td>
<td>0.3</td>
<td>0.4</td>
<td>10</td>
</tr>
<tr>
<td>Avg Number of Days w/ Abandons Past 720 Days</td>
<td>1</td>
<td>0.5</td>
<td>3.5</td>
</tr>
<tr>
<td>Avg Number of Days w/ Page Views in Past 360 Days</td>
<td>0.4</td>
<td>0.5</td>
<td>12</td>
</tr>
<tr>
<td>% Email Opt-In</td>
<td>0%</td>
<td>5%</td>
<td>18%</td>
</tr>
<tr>
<td>Count of SMS Sent</td>
<td>0.4</td>
<td>6.6</td>
<td>6.8</td>
</tr>
</tbody>
</table>

*$119 cutoff determined by median spend value
## Post-Modeling Analysis: Multi-Class Model

<table>
<thead>
<tr>
<th>Feature</th>
<th>Predicted Zero Tier (spend = $0)</th>
<th>Predicted Low Tier* (spend &lt;= $119)</th>
<th>Predicted High Tier (spend &gt; $119)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Acct Creation Loyal</td>
<td>0%</td>
<td>10%</td>
<td>11%</td>
</tr>
<tr>
<td>% Inactive Customers</td>
<td>0.1%</td>
<td>85%</td>
<td>81%</td>
</tr>
<tr>
<td>Avg Number of Days w/ Searches Past 360 Days</td>
<td>0.3</td>
<td>0.4</td>
<td>10</td>
</tr>
<tr>
<td>% Email Opt-In</td>
<td>0%</td>
<td>5%</td>
<td>18%</td>
</tr>
<tr>
<td>Count of SMS Sent</td>
<td>0.4</td>
<td>6.6</td>
<td>6.8</td>
</tr>
</tbody>
</table>

*$119 cutoff determined by median spend value
Post-Modeling Analysis: Multi-Class Model

3% of Zero Tier are Loyalty Members
43% of Low Tier are Loyalty Members
53% of High Tier are Loyalty Members

- Zero: 97% Non-Loyalty, 3% Loyalty
- Low: 57% Non-Loyalty, 43% Loyalty
- High: 47% Non-Loyalty, 53% Loyalty