**Problem Statement**

Who are Macy’s valuable prospective customers? What is their Customer Lifetime Value (CLV)?

**Current State:** Macy’s predicts CLV for 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 CLV

---

**Data**

- **Prospective Customers:**
  - Feb 2020 – Jan 2022
  - Feb 2022 – Jan 2023
  - Predict Prospective Customer CLV

- **Inactive Customers:**
  - Have purchase history prior to Feb 2020

- **New Customers:**
  - Have never made a purchase prior to Feb 2022

**Data Limitations**

- **Imbalanced Dataset:** Only 8% of customers purchased in 2022
- **Skewed Distribution for Online Activity:** Majority of values indicate little activity
- **Missing Values:** Removed demographic and income features

**Features**

- **Click behaviors:** Search, browse, add to cart, page view, abandon cart, and others
- **User Profile:** Loyalty status, length of loyalty, new/inactive

**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**

---

**Methodology**

**Models Used:**

- Random Forest
- GBM
- CatBoost

**Preprocessing**

- Oversample Minority Class (Purchasers)
- Ensemble Learning used predictions from binary classification model as a feature
- Oversample Minority Class (Purchasers)

**Build models to predict prospective customer CLV**

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

**Evaluation**

- Accuracy, Recall
- Mean Absolute Error
- Accuracy, Recall

**Best Model:** CatBoost GBM

**Results**

- **Accuracy of our Best Model:** 84%
  - +7% Over Baseline

- **Recall of our Best Model:** 73%
  - +19% Over Baseline

- **AUC of our Best Model:** 79%
  - +12% Over Baseline

**Out-of-Sample Performance**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Recall</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary Classification</td>
<td>0.84</td>
<td>0.73</td>
<td>0.79</td>
</tr>
<tr>
<td>Multi Classification</td>
<td>0.79</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>Baseline (Churn)</td>
<td>0.77</td>
<td>0.77</td>
<td>0.75</td>
</tr>
</tbody>
</table>

**Why Do We Want High Recall?**

- We wish to limit instances where Macy’s misses out on sending promotional emails to customers who would have been valuable shoppers

**Top Drivers of Prospective Customer Value**

1. New vs. Inactive Customer
2. Account Creation
3. Email Opt-In
4. Count of SMS Sent
5. Search

**Solution Validation**

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

1. Train model on more recent data
   - Jul 2020 – Jun 2022
   - Jul 2022 – Jun 2023
   - Prospective customer online activity
   - Predict prospective customer CLV

2. Test model on older data
   - Jul 2019 – Jun 2021
   - Jul 2021 – Jun 2022
   - Prospective customer online activity
   - Predict prospective customer CLV

---

**Next Steps**

- Integration into active customer CLV workflow
- Deployment of prospective customer CLV models

---

**Acknowledgements:** We would like to thank the Macy’s Inc. EDA team (especially Donghao Pei, Yixin Cai, and Iris Singhania) for their time and support, and our faculty advisor Professor Georgia Perakis and PhD advisor Leann Thayaparan for their insights and guidance. Together they have helped us develop the work presented in this poster.