"I'm Just Browsing"
Predicting the Value of Prospective Customers

Faculty Advisor: Professor Georgia Perakis **PhD Advisor:** Leann Thayaparan
Macy's Team: Iris Singhania, Yixin Cai, Donghao Pei

Estella Dentinger   Brittany Nguyen

MIT MANAGEMENT BUSINESS ANALYTICS

macy's inc
macy's bloomingdales bluemercury

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

What is a prospective customer?
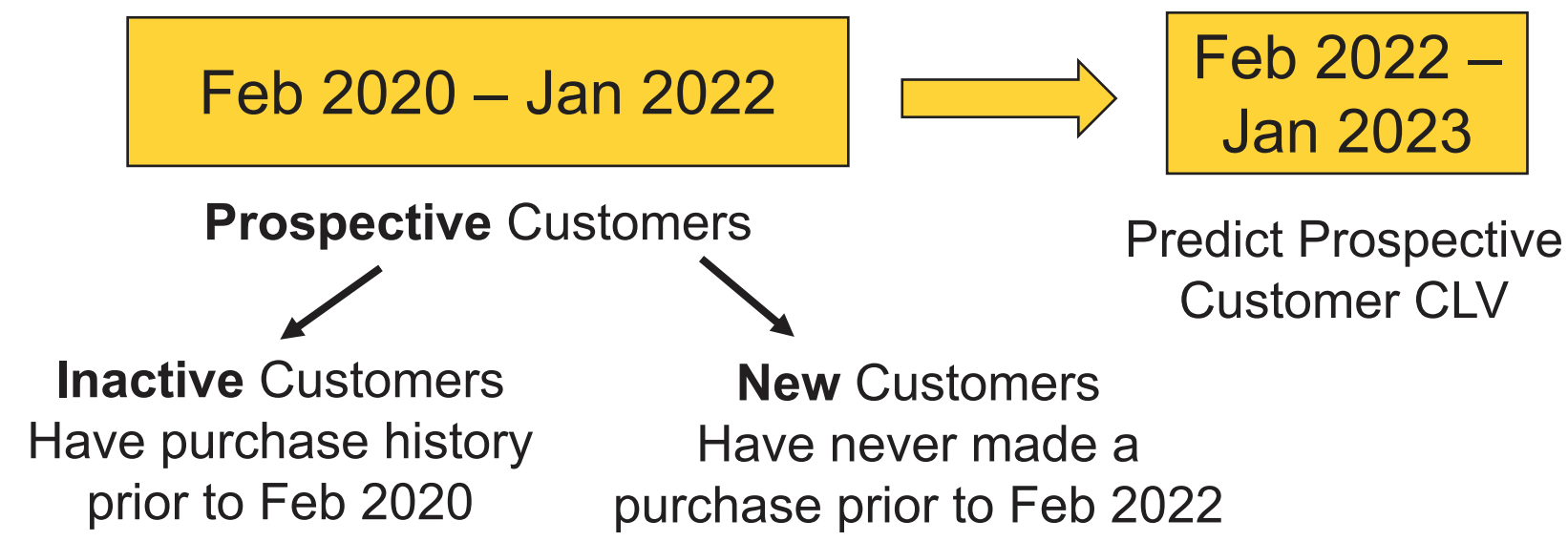Users who had online activity but did not purchase in last 48 months

Online activity − Did not make a purchase = Prospective Customer

Data

Feb 2020 – Jan 2022 → Feb 2022 – Jan 2023
Predict Prospective Customer CLV

Prospective Customers

Inactive Customers
Have purchase history prior to Feb 2020

New Customers
Have never made a purchase prior to Feb 2022

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 customers purchased in 2022

Skewed Distribution for Online Activity
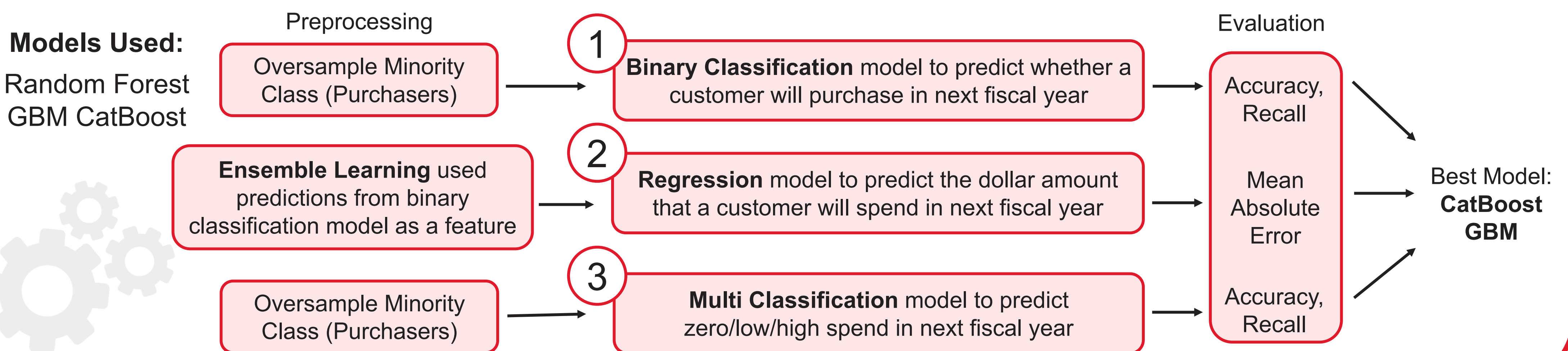Majority of values indicate little activity

Missing Values
Removed demographic and income features

7M Prospective customers in dataset

77% Prospective customers are new

21% Prospective customers are loyalty members

8% Prospective customers made a purchase in 2022

Methodology

Build models to predict prospective customer CLV

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)

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

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

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

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

Out-of-Sample Performance

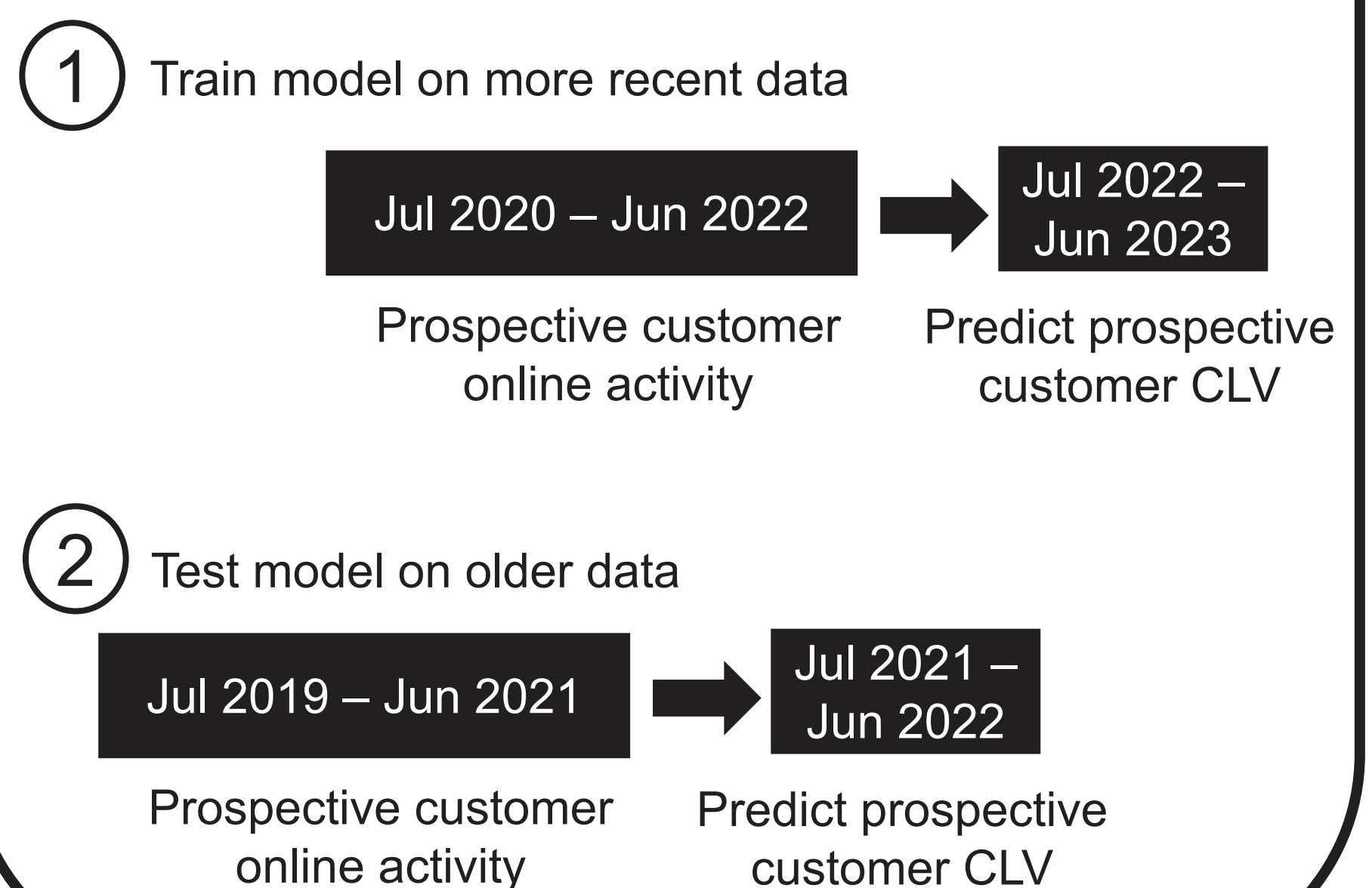| Model	Accuracy	Recall	AUC
Binary Classification	0.84	0.73	0.79
Multi Classification	0.79	0.79	0.75
Baseline (Churn)	0.77	0.54	0.67

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
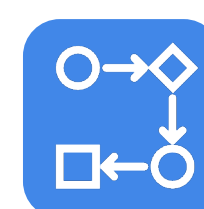
Business Impact

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

Deployment of prospective customer CLV models