PROBLEM STATEMENT

Retail businesses need to proactively identify how much an active customer is likely to spend in the near future in order to optimize marketing strategies and resource allocation.

The goal of this project is to predict the total monetary value a customer will spend in the next 30 days, using historical transaction data and customer behaviour features, by building a supervised machine learning regression model.

WHY THIS MATTERS?

- Not all customers contribute equally to revenue
- Marketing budgets are limited and must be allocated efficiently
- Reactive strategies (after purchase) are often too late
- Predicting short-term spend enables:

A)Targeted promotions

B)Personalized offers

C)Better revenue forecasting

- “Retailers don’t want to treat all customers the same. If we know who is likely to spend more in the next 30 days, we can act proactively instead of reacting after the purchase happens.”

PROPOSED SOLUTION

- We propose a **machine learning–based regression system(Random Forest regressor, GBM regressor,Ridge regressor and Linear Regresssion)** to predict short-term customer value
- The model estimates the **total amount a customer will spend in the next 30 days**
- Predictions are generated using:
- Historical transaction behaviour
- Customer purchase patterns
- Product and channel-level features

“Although the data comes from multiple relational tables, after joining and aggregating them at the customer level, the prediction task becomes a standard regression problem.

Dataset

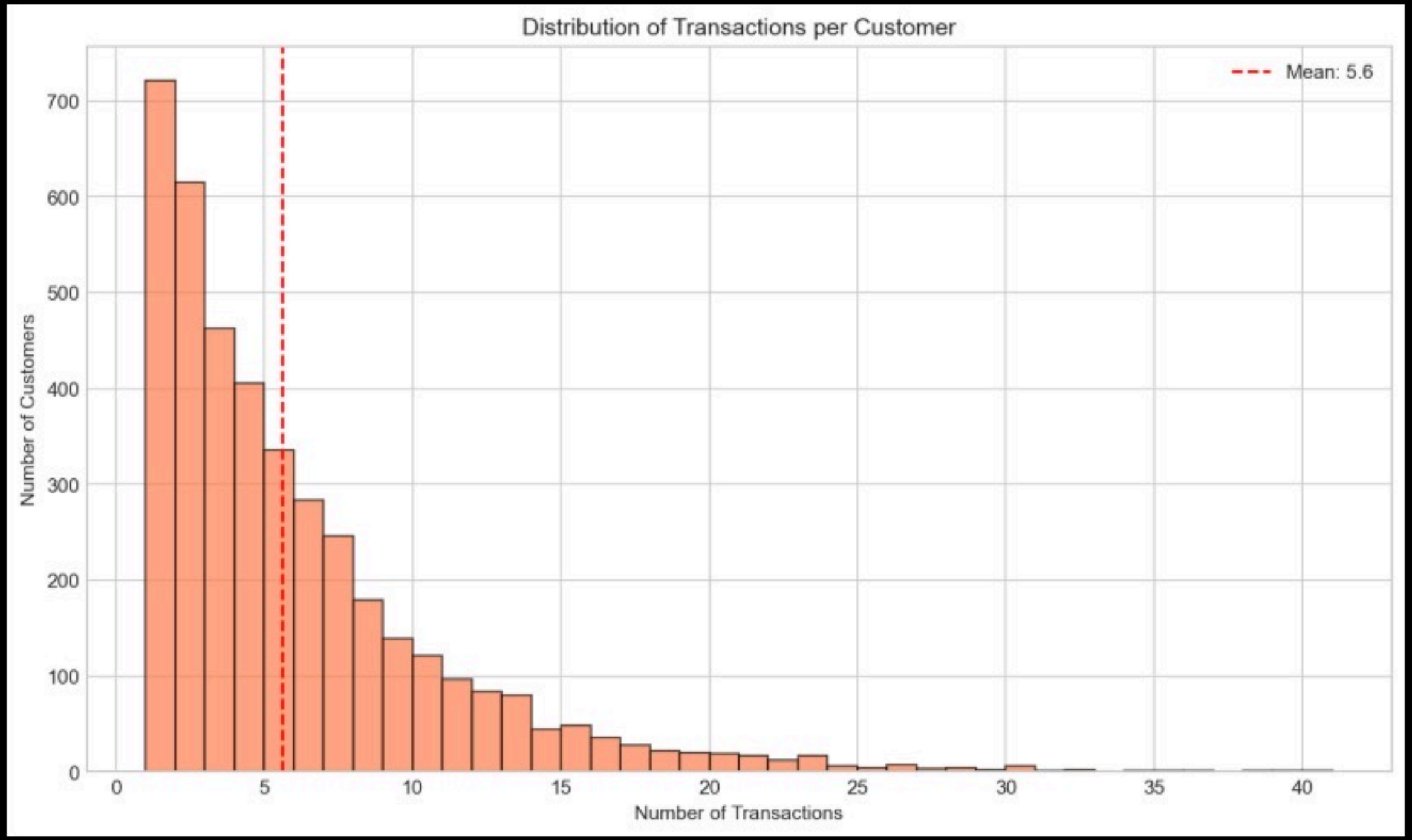
- OVERVIEW
- Load transactions data
- Retail transaction dataset
- Contains customer-level and transaction-level tables
- Each transaction includes:
  - Customer ID
  - Invoice date
  - Product details
  - Quantity and price,etc
- **Dataset Size**
- Rows:5001
- Columns: 32
- *After combining all the 7 tables and storing invalid data to another file.*

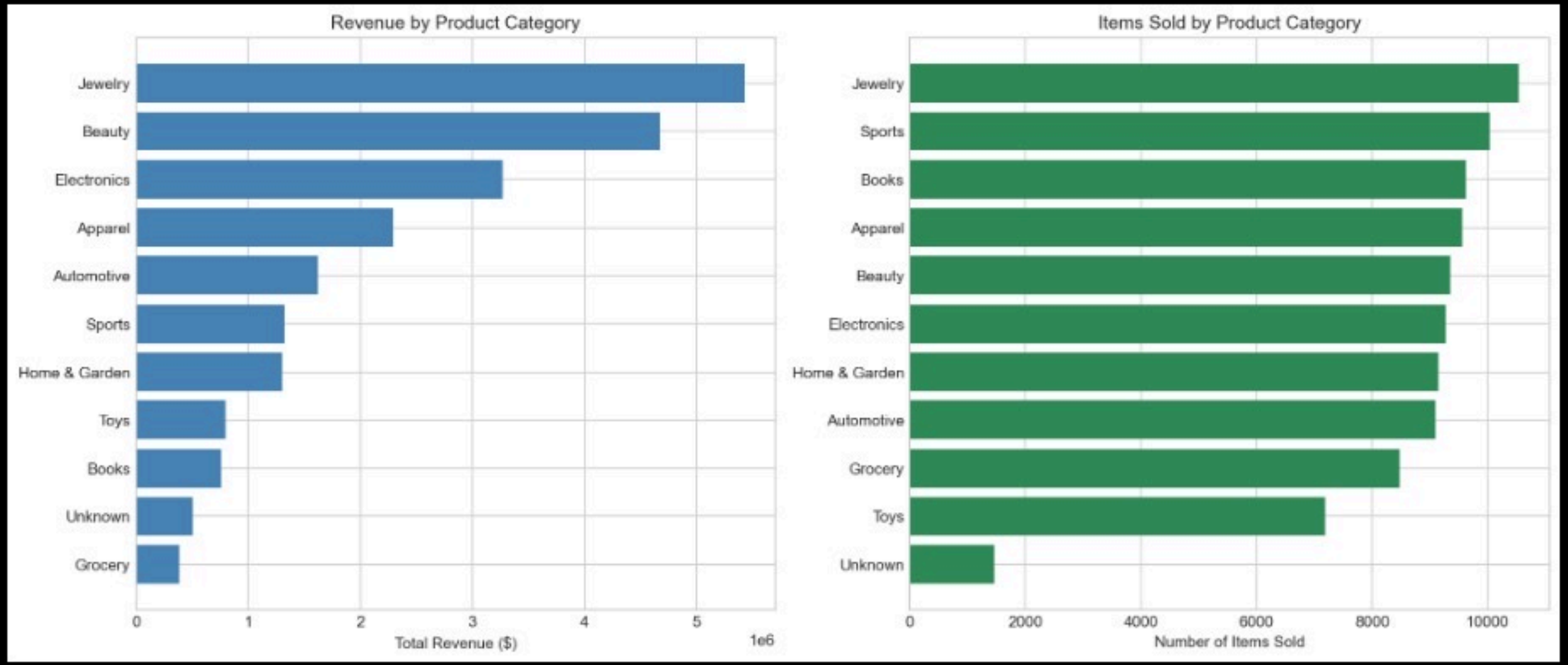
Data Cleaning & Preprocessing:

- Handling missing values
- Imputing missing values
- Outlier Detection
- Removing invalid transactions and storing it to different file for further analysis
- Fixing data types
- Categorical- One hot encoding/Label Encoding
- AFTER CLEANING
- Rows:3819
- Columns:32

EDA

- **Univariate Analysis**
- Distribution of transaction amount
- Distribution of number of transactions per customer
- Distribution of customer spend
- **Bivariate Analysis**
- Relationship between:
  - Item Sold by Product Category
  - Number of Item Sold
- High-frequency customers show higher future spend


Distribution of Transactions per Customer
Mean: 5.6
Number of Customers
0
100
200
300
400
500
600
700
0
5
10
15
20
25
30
35
40
Number of Transactions


Revenue by Product Category
Jewelry
Beauty
Electronics
Apparel
Automotive
Sports
Home & Garden
Toys
Books
Unknown
Grocery
0
1
2
3
4
5
1e6
Total Revenue ($)
Items Sold by Product Category
Jewelry
Sports
Books
Apparel
Beauty
Electronics
Home & Garden
Automotive
Grocery
Toys
Unknown
0
2000
4000
6000
8000
10000
Number of Items Sold

Feature engineering

Aggregated transaction-level data at **customer level**

Created behavioural features:

- Total historical spend
- Number of past transactions
- Average order value

Created temporal features:

- Recency (days since last purchase)
- Frequency of purchases

Target Variable Definition

Defined target as:

Total amount spent by a customer in the 30 days following a cutoff date

Ensured strict separation between:

- Feature window (before cutoff)
- Prediction window (after cutoff)

Train–Test Split & Model Training

Data split performed at **customer level**

Used a **time-based split** to prevent data leakage

Transactions before the cutoff date used for **training features**

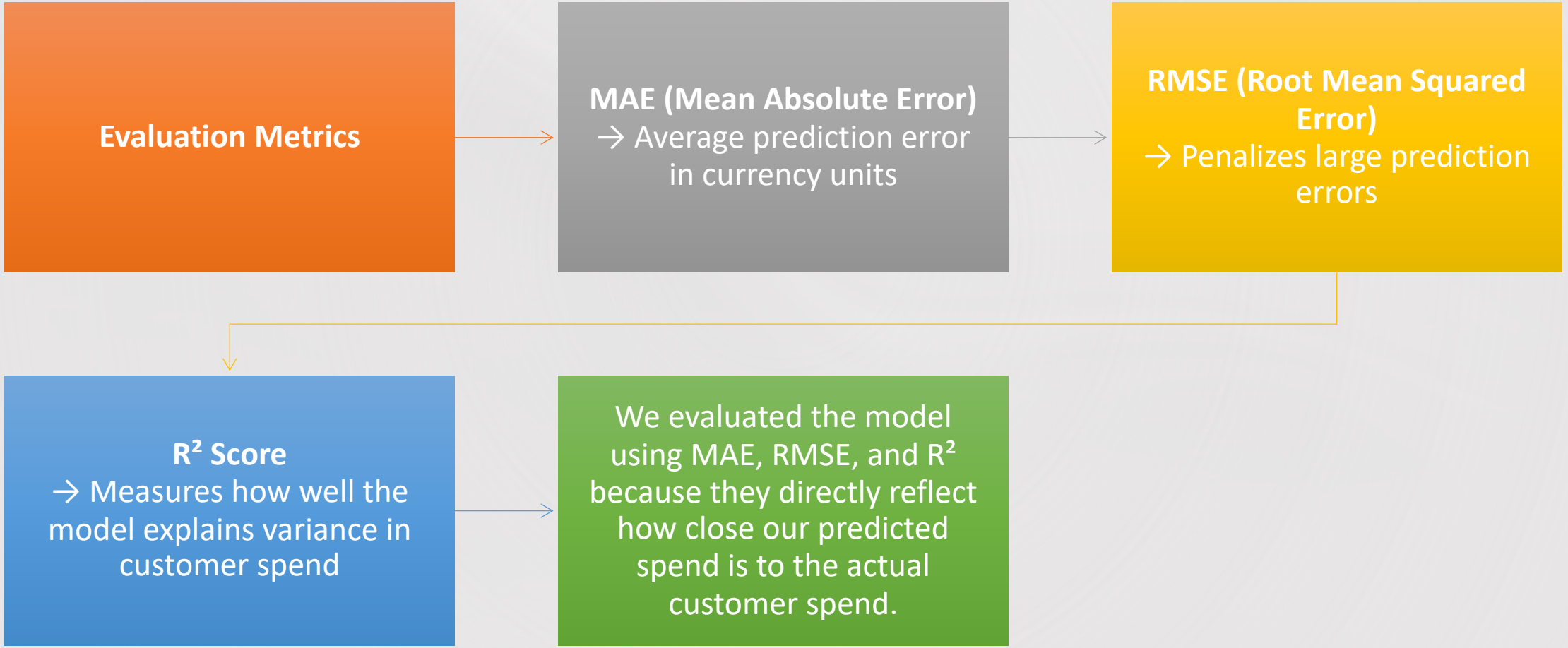
Spend in the next 30 days used as **target variable**

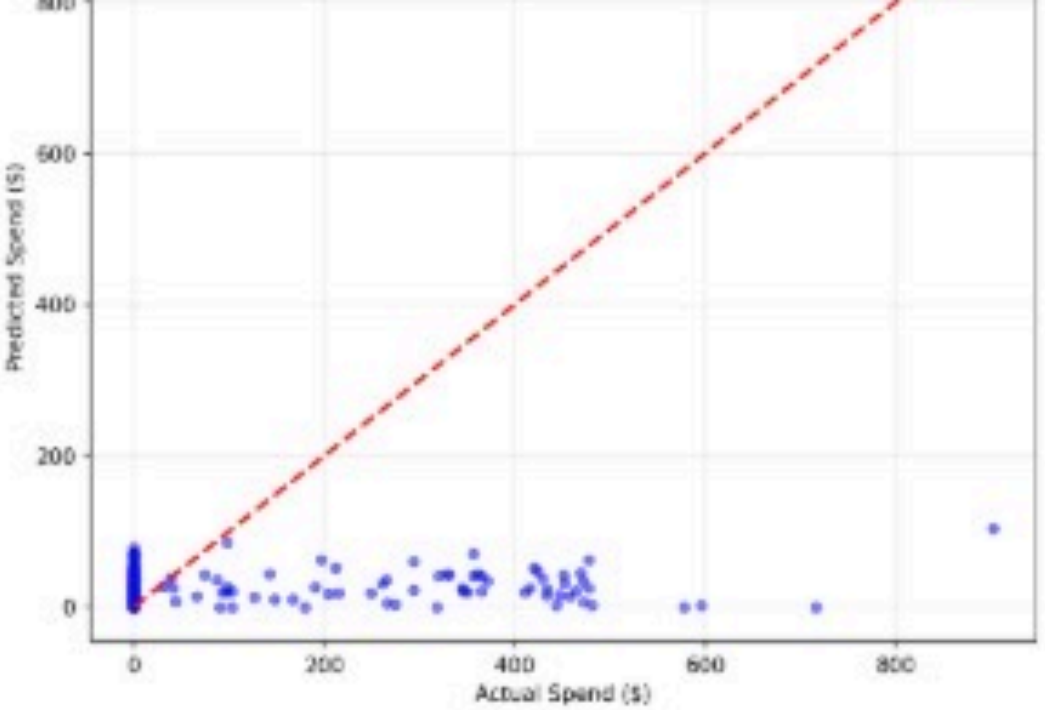
Split Ratio

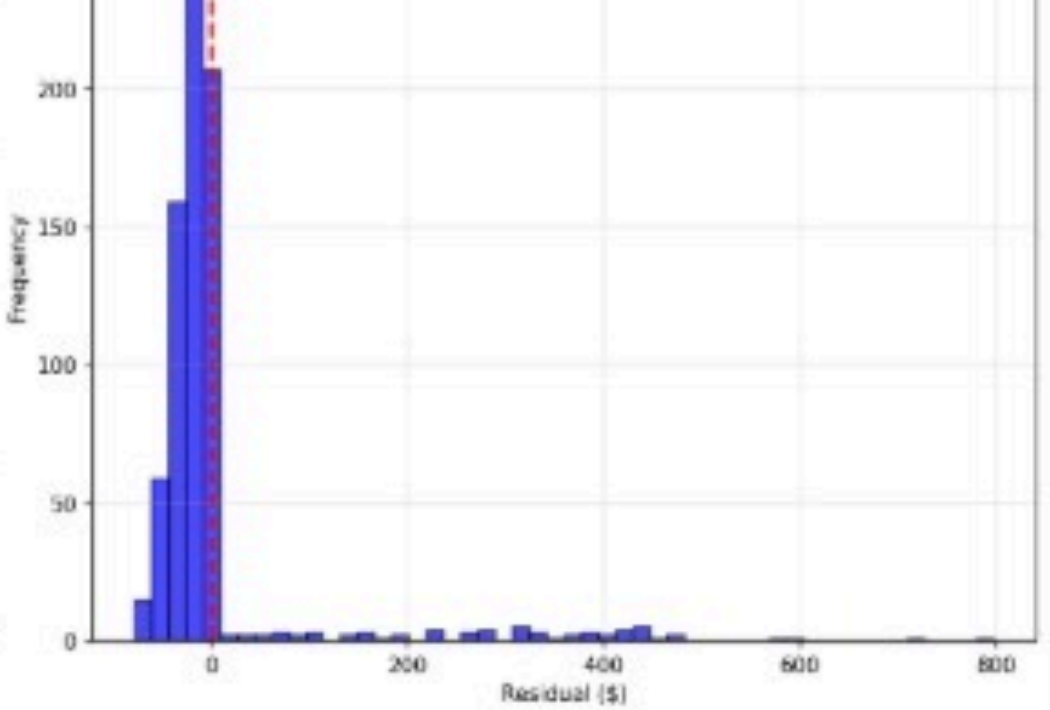
Training set: **80% of customers**

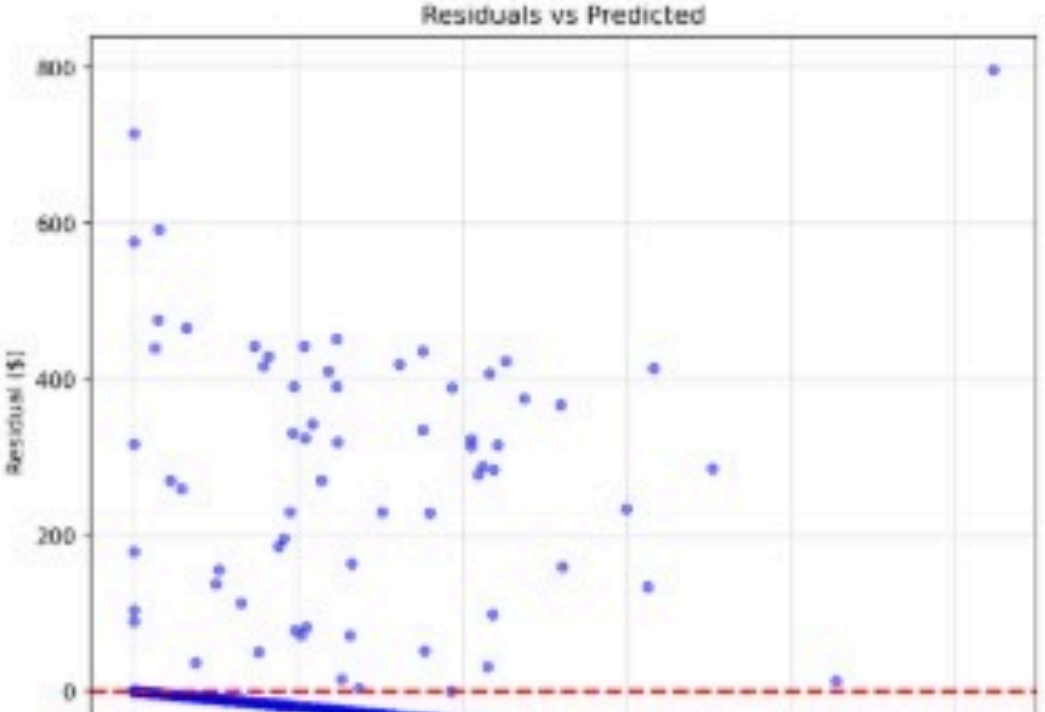
Test set: **20% of customers**

Model Evaluation (Random Forest Regressor)



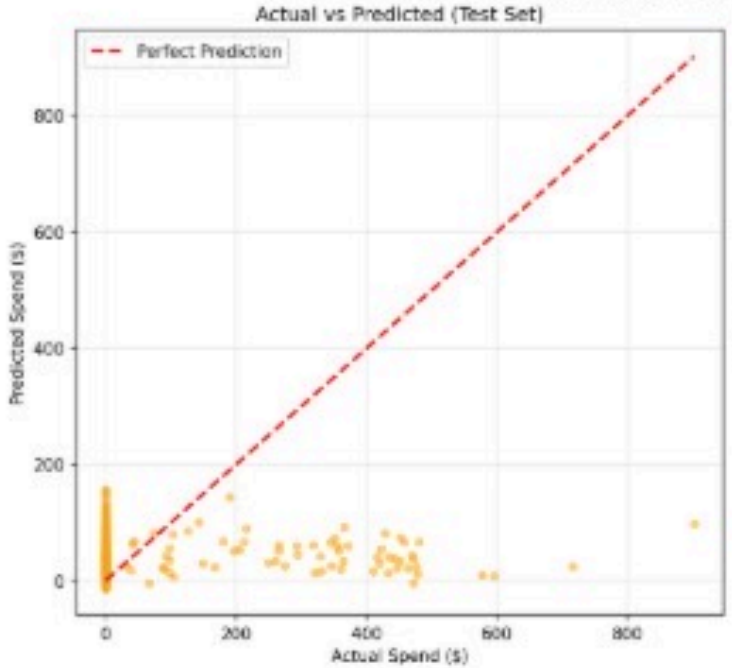


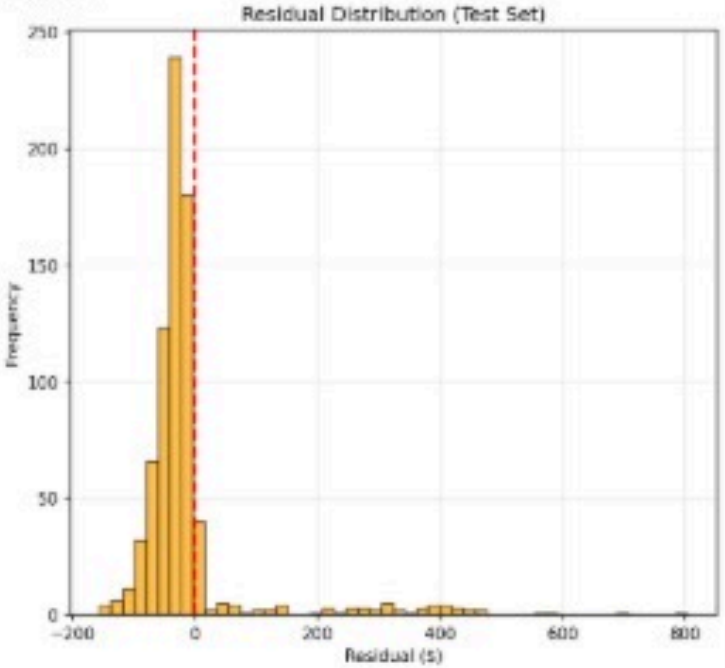


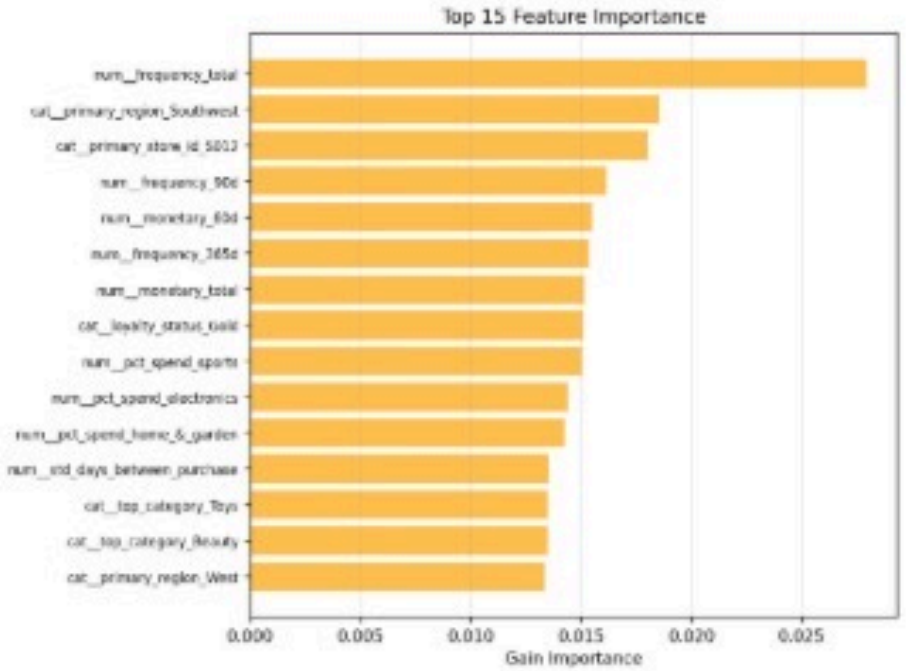



```
LINEAR REGRESSION METRICS
========================================

TRAINING SET:
  RMSE:  $124.09
  MAE:   $75.23
  R²:    0.0020

TEST SET:
  RMSE:  $100.34
  MAE:   $42.93
  R²:    0.0137

========================================
Train/Test RMSE Ratio: 1.237
```

XGBoost Model Evaluation

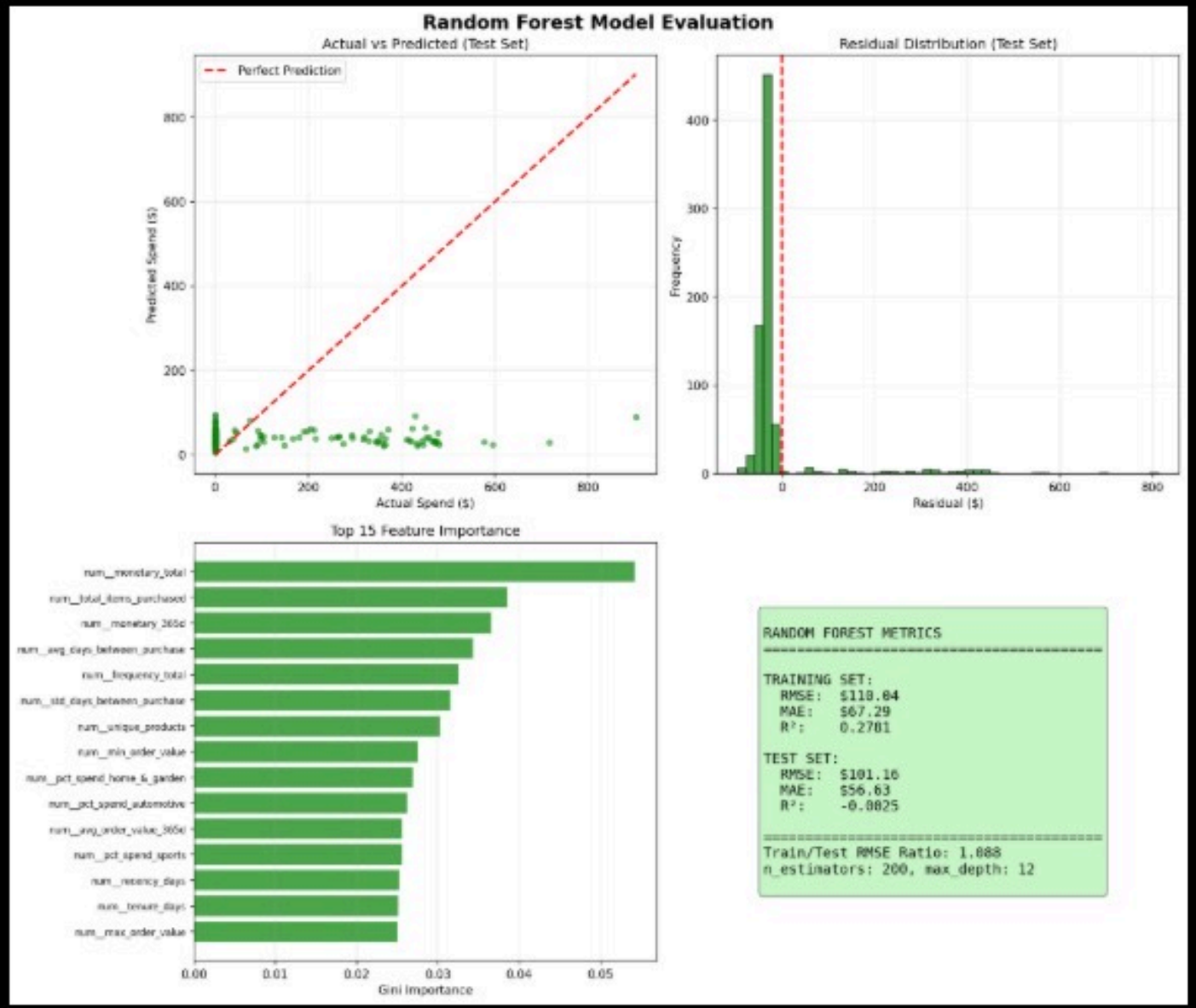







```
XGBOOST METRICS
========================================

TRAINING SET:
  RMSE:  $67.22
  MAE:   $40.33
  R²:    0.7306

TEST SET:
  RMSE:  $103.53
  MAE:   $57.00
  R²:    -0.0501

========================================
Train/Test RMSE Ratio: 0.649
n_estimators: 200, max_depth: 6, lr: 0.05
```

random_forest_evaluation.png



Results:

Business Insight:
1. Predicts which customers will generate revenue in the next 30 days, enabling proactive decision-making.
2. Helps focus marketing spend on high-value customers, improving ROI and reducing wasted discounts.
3. Enables early intervention for declining spenders, supporting customer retention strategies.
4. Transforms raw transaction data into customer-level revenue insights usable by business teams.
5. Uses short-term, forward-looking value (30-day spend) instead of static historical reporting.
6. Deployable via UI/API, allowing real-time use by non-technical stakeholders.

Results

Technical Insight:

1. Time-based target definition prevents data leakage by strictly separating historical features from future spend.

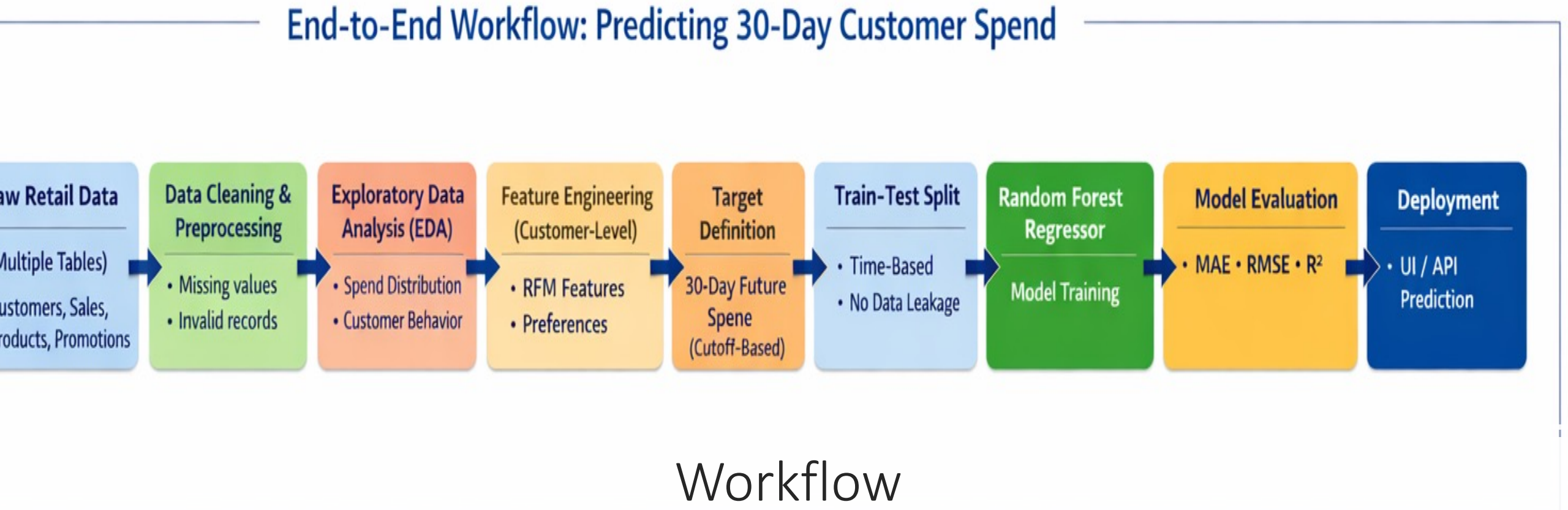
2. Customer-level feature engineering (RFM) significantly improves predictive power over transaction-level data.

3. Non-linear models (Random Forest, XGBoost) outperform linear models by capturing complex purchase behavior.

4. XGBoost achieves the lowest MAE and RMSE, making it the most reliable model for skewed spend distributions.

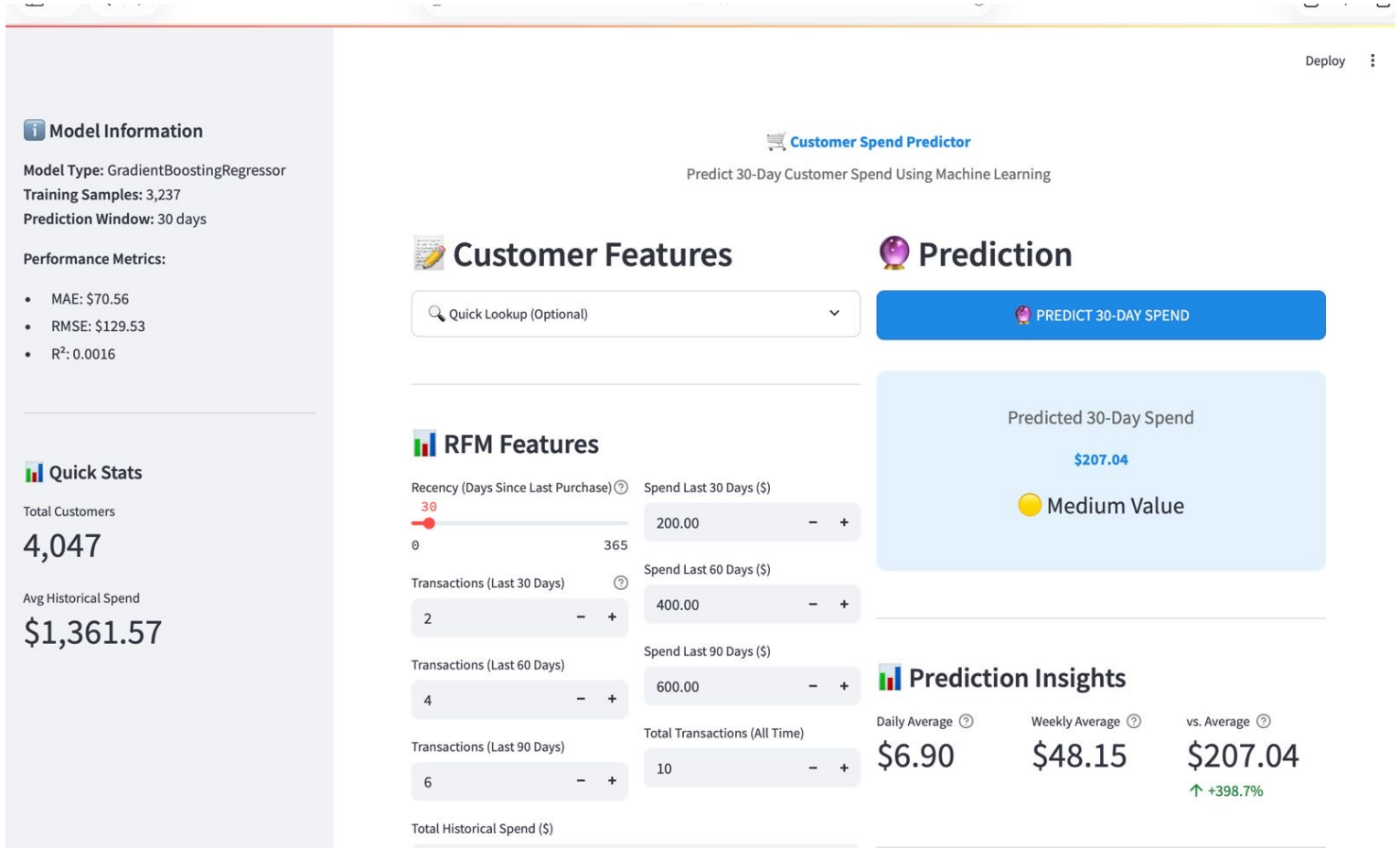
5. Tree-based feature importance highlights recency and frequency as dominant predictors of short-term spend.

6. Modular pipeline design enables easy retraining, model comparison, and seamless deployment via UI/API.

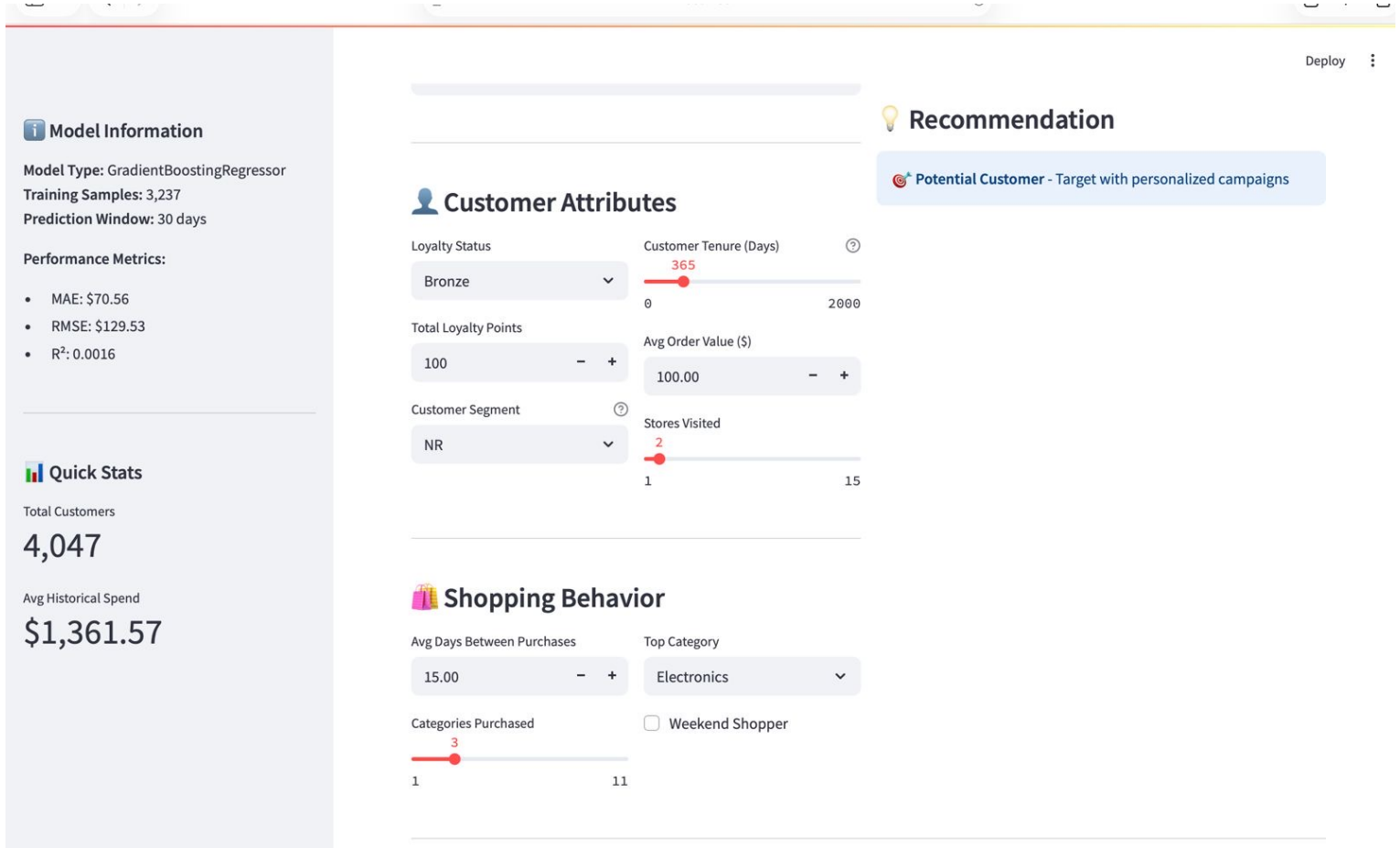

End-to-End Workflow: Predicting 30-Day Customer Spend
aw Retail Data
Multiple Tables)
ustomers, Sales,
roducts, Promotions
Data Cleaning & Preprocessing
• Missing values
• Invalid records
Exploratory Data Analysis (EDA)
• Spend Distribution
• Customer Behavior
Feature Engineering (Customer-Level)
• RFM Features
• Preferences
Target Definition
30-Day Future Spene
(Cutoff-Based)
Train-Test Split
• Time-Based
• No Data Leakage
Random Forest Regressor
Model Training
Model Evaluation
• MAE • RMSE • R²
Deployment
• UI / API Prediction

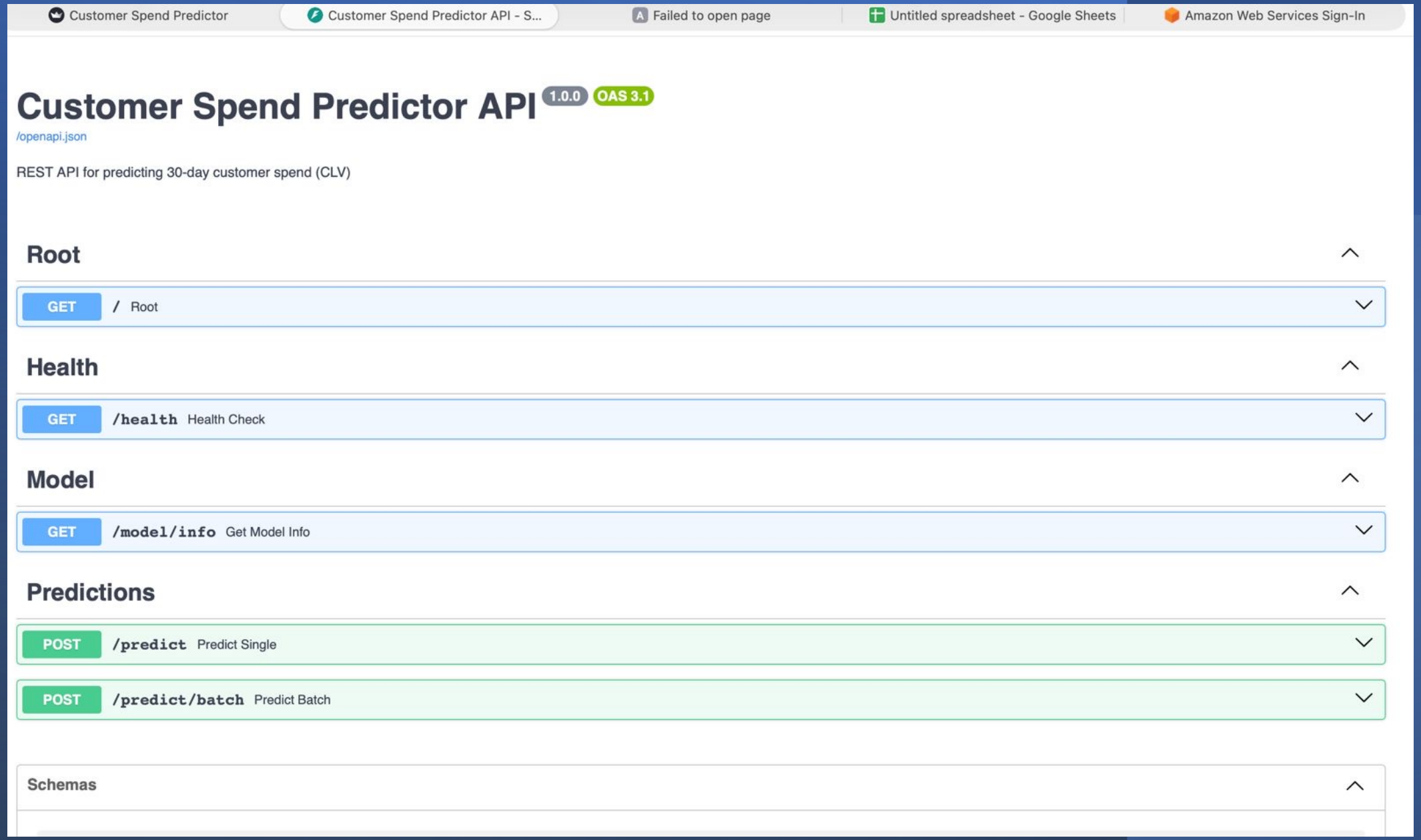
Workflow

Front-end UI

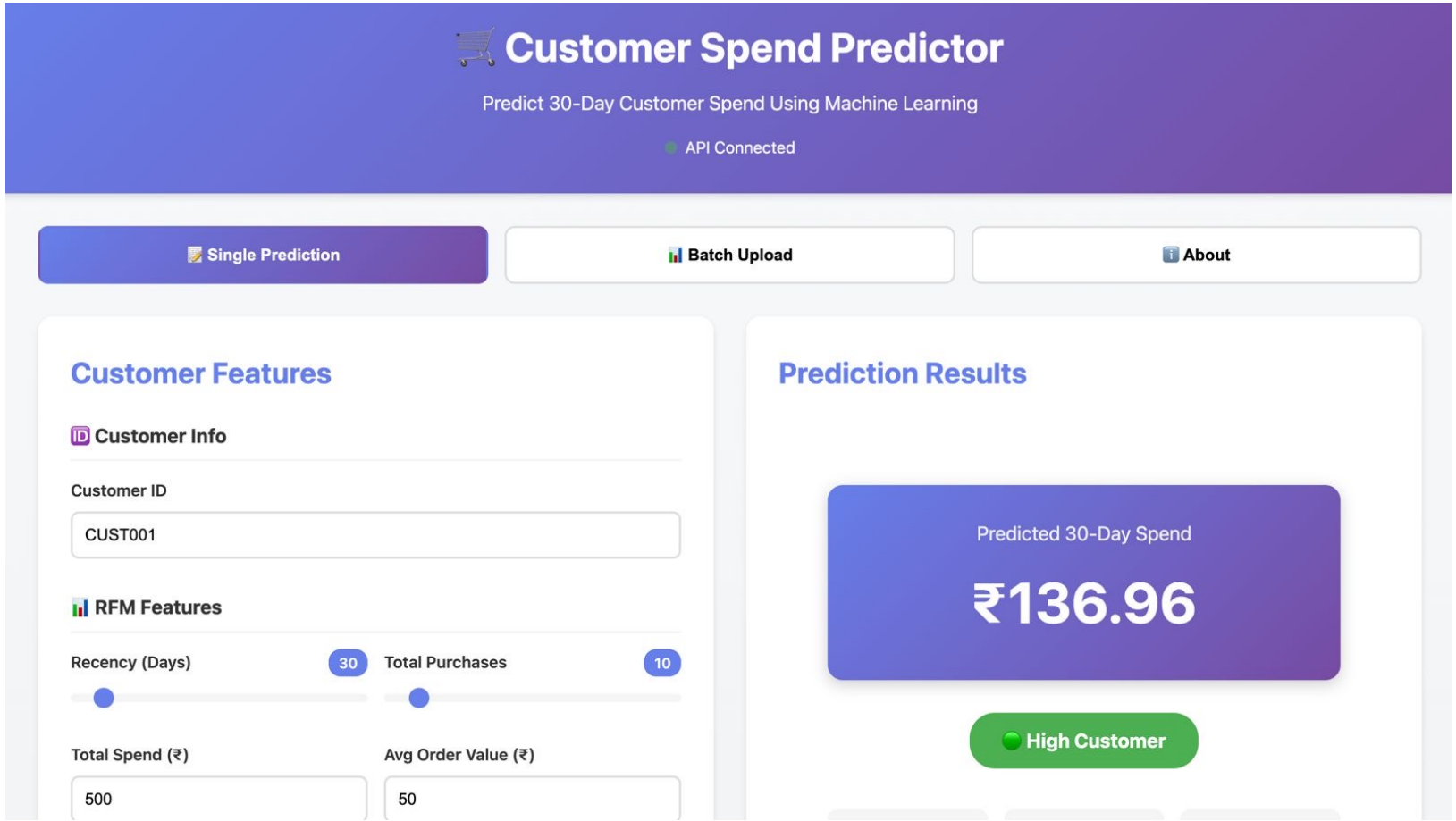

Deploy
Model Information
Model Type: GradientBoostingRegressor
Training Samples: 3,237
Prediction Window: 30 days
Performance Metrics:
MAE: $70.56
RMSE: $129.53
R²: 0.0016
Quick Stats
Total Customers
4,047
Avg Historical Spend
$1,361.57
Customer Spend Predictor
Predict 30-Day Customer Spend Using Machine Learning
Customer Features
Quick Lookup (Optional)
RFM Features
Recency (Days Since Last Purchase)
30
0
365
Transactions (Last 30 Days)
2
Transactions (Last 60 Days)
4
Transactions (Last 90 Days)
6
Total Historical Spend ($)
Spend Last 30 Days ($)
200.00
Spend Last 60 Days ($)
400.00
Spend Last 90 Days ($)
600.00
Total Transactions (All Time)
10
Prediction
PREDICT 30-DAY SPEND
Predicted 30-Day Spend
$207.04
Medium Value
Prediction Insights
Daily Average
$6.90
Weekly Average
$48.15
vs. Average
$207.04
+398.7%

Front-end UI

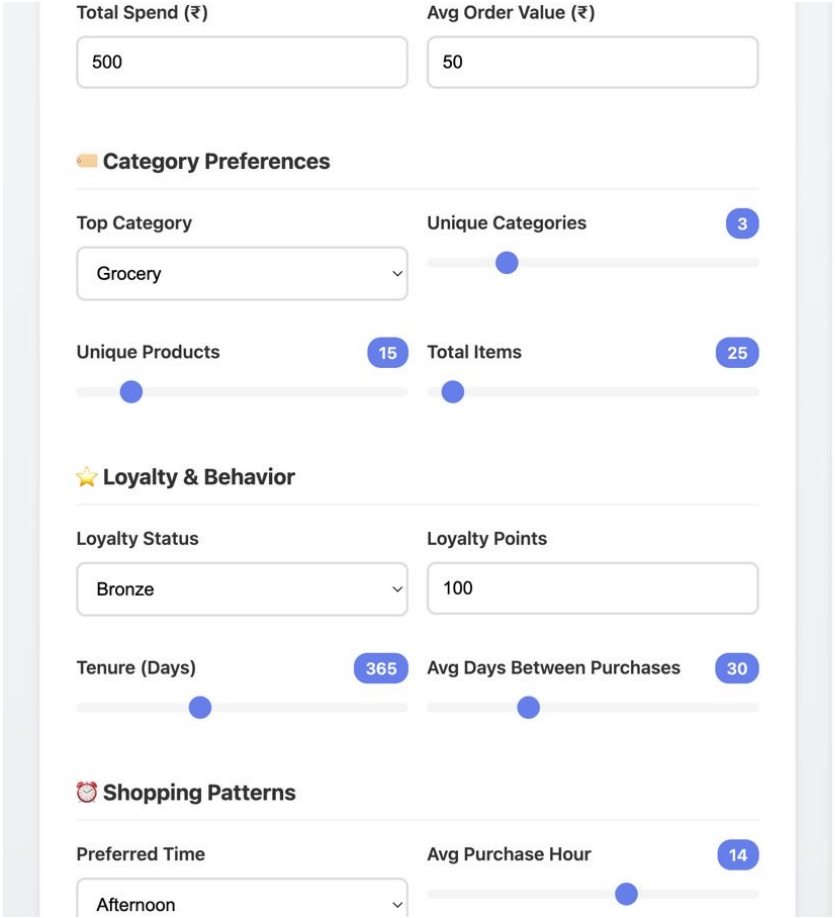

Deploy
Model Information
Model Type: GradientBoostingRegressor
Training Samples: 3,237
Prediction Window: 30 days
Performance Metrics:
MAE: $70.56
RMSE: $129.53
R²: 0.0016
Quick Stats
Total Customers
4,047
Avg Historical Spend
$1,361.57
Customer Attributes
Loyalty Status
Bronze
Customer Tenure (Days)
365
0
2000
Total Loyalty Points
100
Avg Order Value ($)
100.00
Customer Segment
NR
Stores Visited
2
1
15
Shopping Behavior
Avg Days Between Purchases
15.00
Top Category
Electronics
Categories Purchased
3
1
11
Weekend Shopper
Recommendation
Potential Customer - Target with personalized campaigns

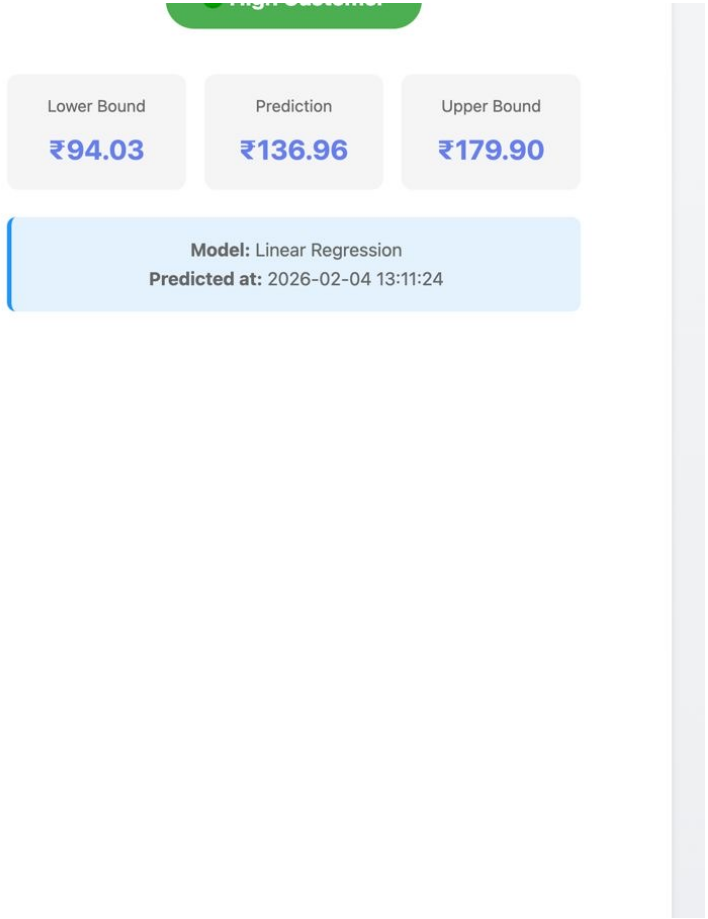

Customer Spend Predictor
Customer Spend Predictor API - S...
Failed to open page
Untitled spreadsheet - Google Sheets
Amazon Web Services Sign-In
Customer Spend Predictor API
1.0.0
OAS 3.1
/openapi.json
REST API for predicting 30-day customer spend (CLV)
Root
GET / Root
Health
GET /health Health Check
Model
GET /model/info Get Model Info
Predictions
POST /predict Predict Single
POST /predict/batch Predict Batch
Schemas

Deployed on AWS

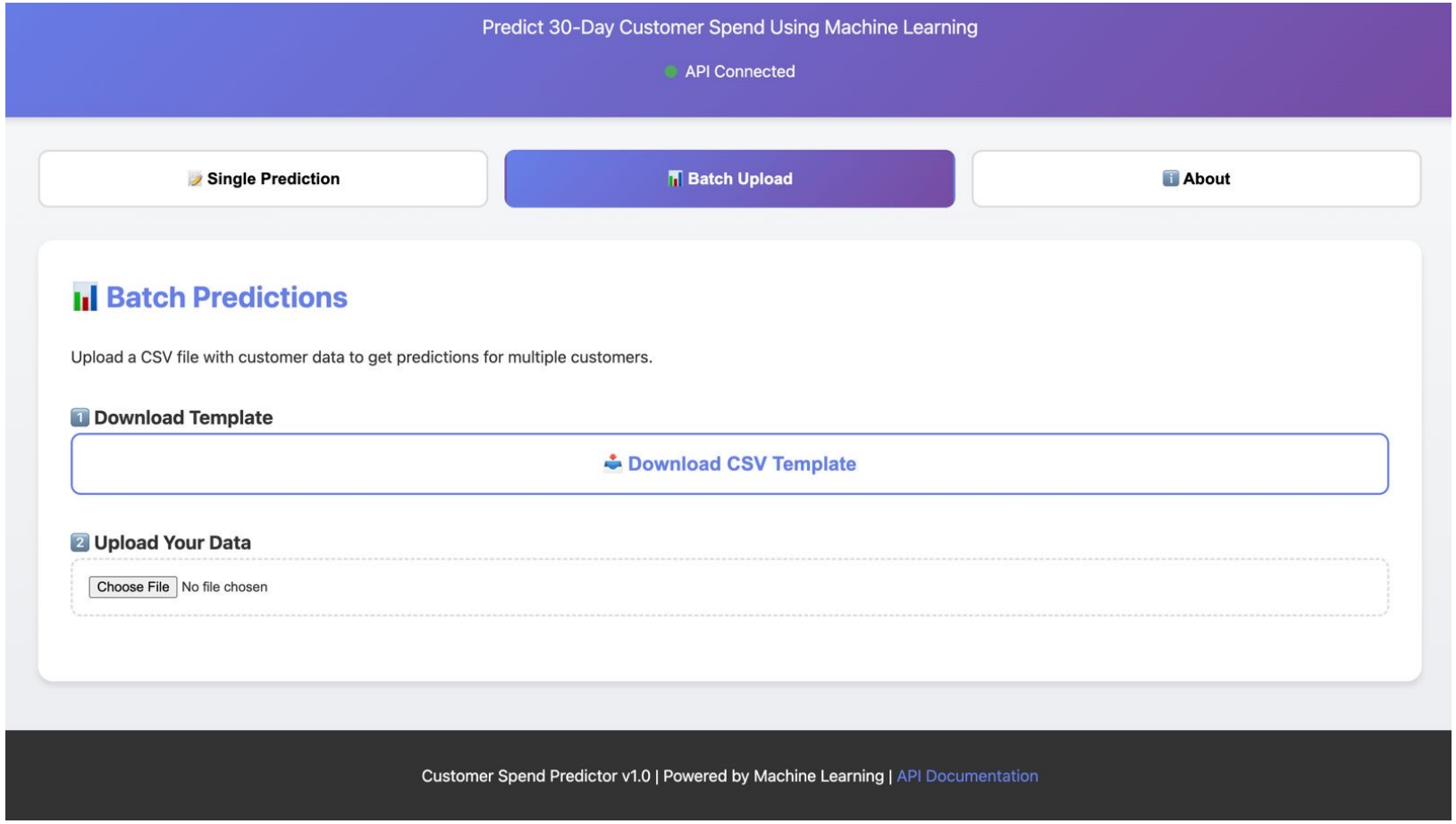


Deployed on AWS





Deployed on AWS



🛒 Customer Spend Predictor

Predict 30-Day Customer Spend Using Machine Learning

🟢 API Connected

📝 Single Prediction | 📊 Batch Upload | ℹ️ About

ℹ️ About This Application

🎯 Purpose

This application predicts customer 30-day spend (CLV - Customer Lifetime Value) using machine learning. It helps businesses identify high-value customers and optimize marketing strategies.

🔧 How It Works

1. **Input Features:** Customer behavioral data (recency, frequency, monetary value, etc.)
2. **ML Model:** Trained on historical customer data
3. **Prediction:** Estimated 30-day spend with confidence range
4. **Classification:** Customers are segmented into value tiers (VIP, High, Medium, Low, Zero)

📊 Model Information

Model	Features	Training Date

THANK YOU