2025

Prediction of purchasing behavior

Zhemin Xie(Ayrie)

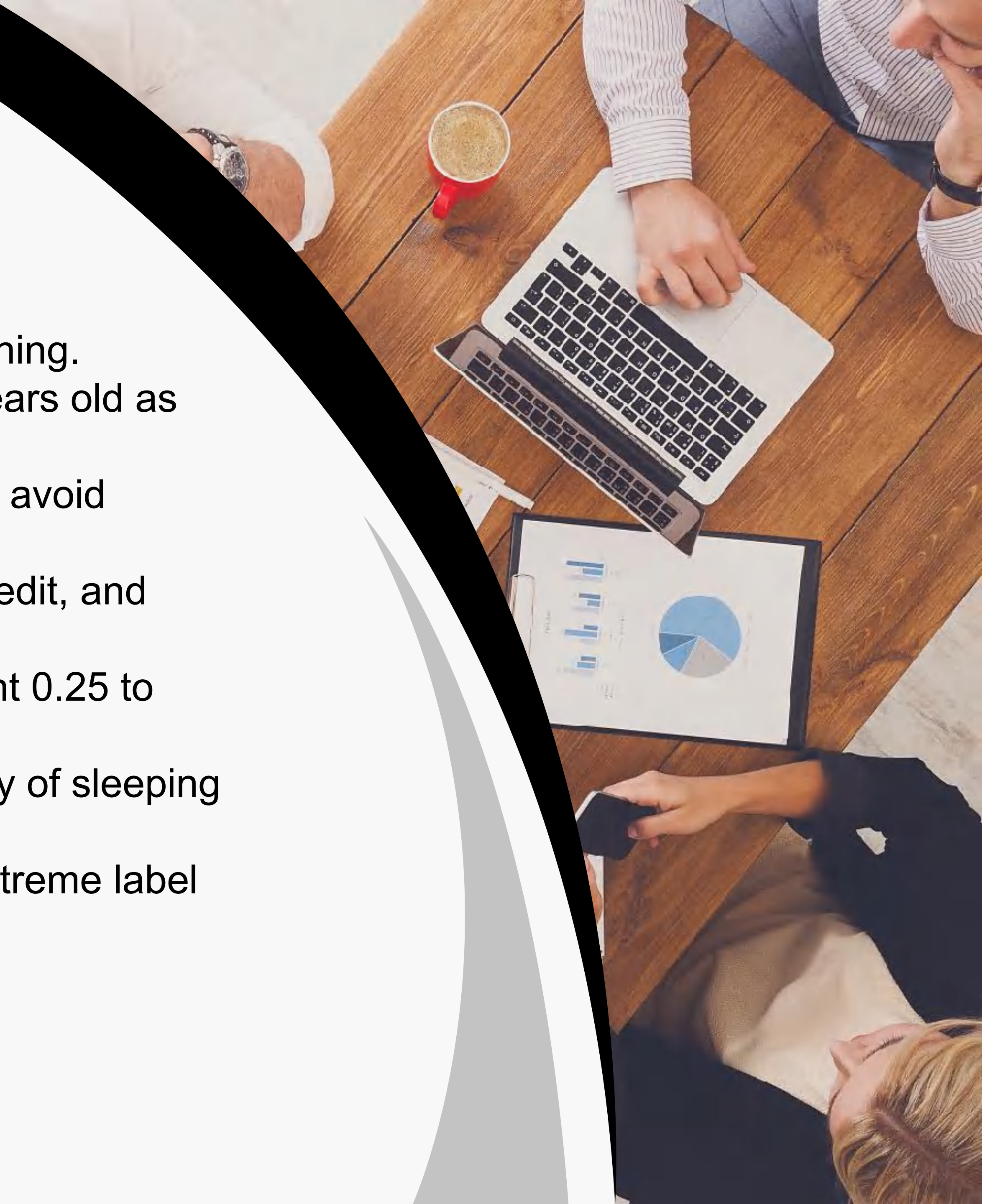
1.1 Data generation

n_samples	5000 samples are generated by default and can be scaled up
random_state	set random seed

feature	generating distribution	range of values	Design Rationale
age	Normal(μ=40, σ=15)	clip(18, 80)	Age is approximately normal, but truncated for minors and extreme advanced age.
time_spent_on_site	Exponential(λ=1/5)	≥0	Online shopping dwell time often has an exponentially long tail, with a few people browsing for a long time.
pages_visited	Poisson(λ=5)	clip(min=1)	The number of page views is a discrete count
is_returning_customer	Binomial(p=0.35)	0/1	35 % are return customers - can be adjusted up or down depending on the specific business。
previous_purchases	Poisson(1)+Poisson(2)*is_returning	≥0	Make return customers ≈2 more historical orders on average, creating relevance
cart_value	Exponential(scale=60)	≥0	There is a large difference between high and low shopping cart amounts, and the long tail is modeled with an exponential distribution.
days_since_last_visit	Exponential(scale=25)	≥0	The longer the interval between recent visits, the more it looks like a sleeping user.

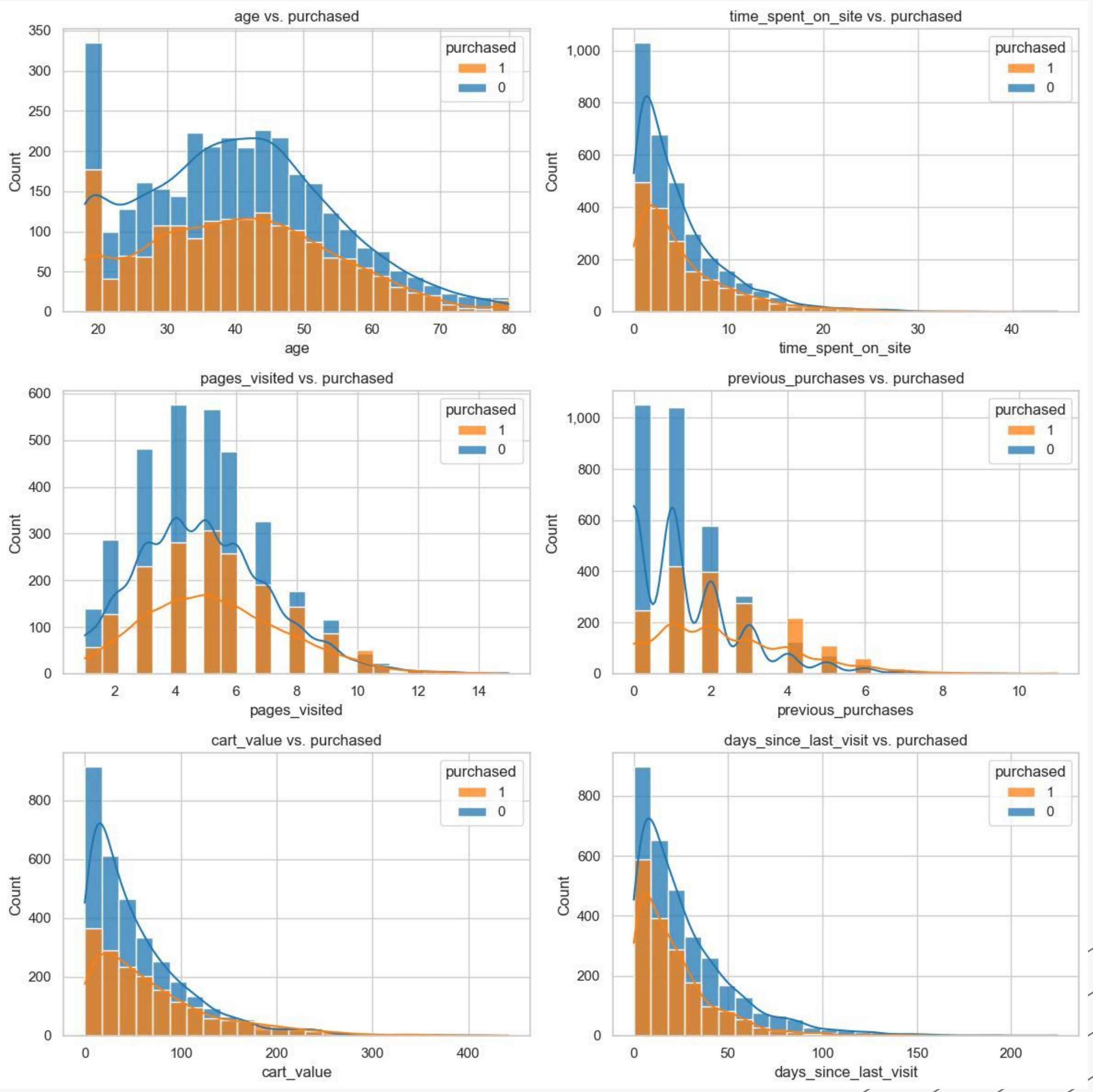
1.2 Calculation of purchase_prob

- Simulates a 3% natural turnover rate even if you do nothing.
- Age fine-tunes the propensity to buy linearly, with 30 years old as the baseline; older and younger are both slightly lower.
- Tanh saturates the gain for >15 minutes of dwell time to avoid extreme values that would blow up the probability.
- Three positive indicators: interaction depth, historical credit, and shopping cart value.
- Returning customers are a strong bonus, with a constant 0.25 to reflect the role of loyalty.
- “Last visit” is a negative indicator, diluting the probability of sleeping users.
- Business sense: probability cannot be ≥1; also avoid extreme label imbalance caused by 0 or 1.



2.1 EDA

age	Peak orange, in the 20s.	Younger customers are more likely to place orders, and ads/recommendations can focus on this age group.
time_spent_on_site	Unpurchased users mostly stay ≤3 minutes; after longer stays, the proportion of purchases rises significantly, with a longer tail than the unpurchased population	Pop-up offers or customer service floats on page stays of ≥3 minutes are expected to boost conversions.
pages_visited	The more pages viewed, the higher the percentage of orange color, indicating a high conversion rate for deep viewers.	Combined with “page view counts” to trigger personalized recommendations.
previous_purchases	Blue color predominates for 0-1 historical purchases; orange color increases significantly after ≥2 purchases.	Create loyalty programs for regular customers: points, exclusive discounts.
cart_value	The higher the shopping cart amount, the higher the probability of purchase	Send checkout now offers on exit for users with high cart amounts but not checking out.
days_since_last_visit	Purchased users are more concentrated in the last 0-30 days; almost no purchases are made after a visit interval of >60 days.	30 days without access triggers a recall notice with a limited time offer.

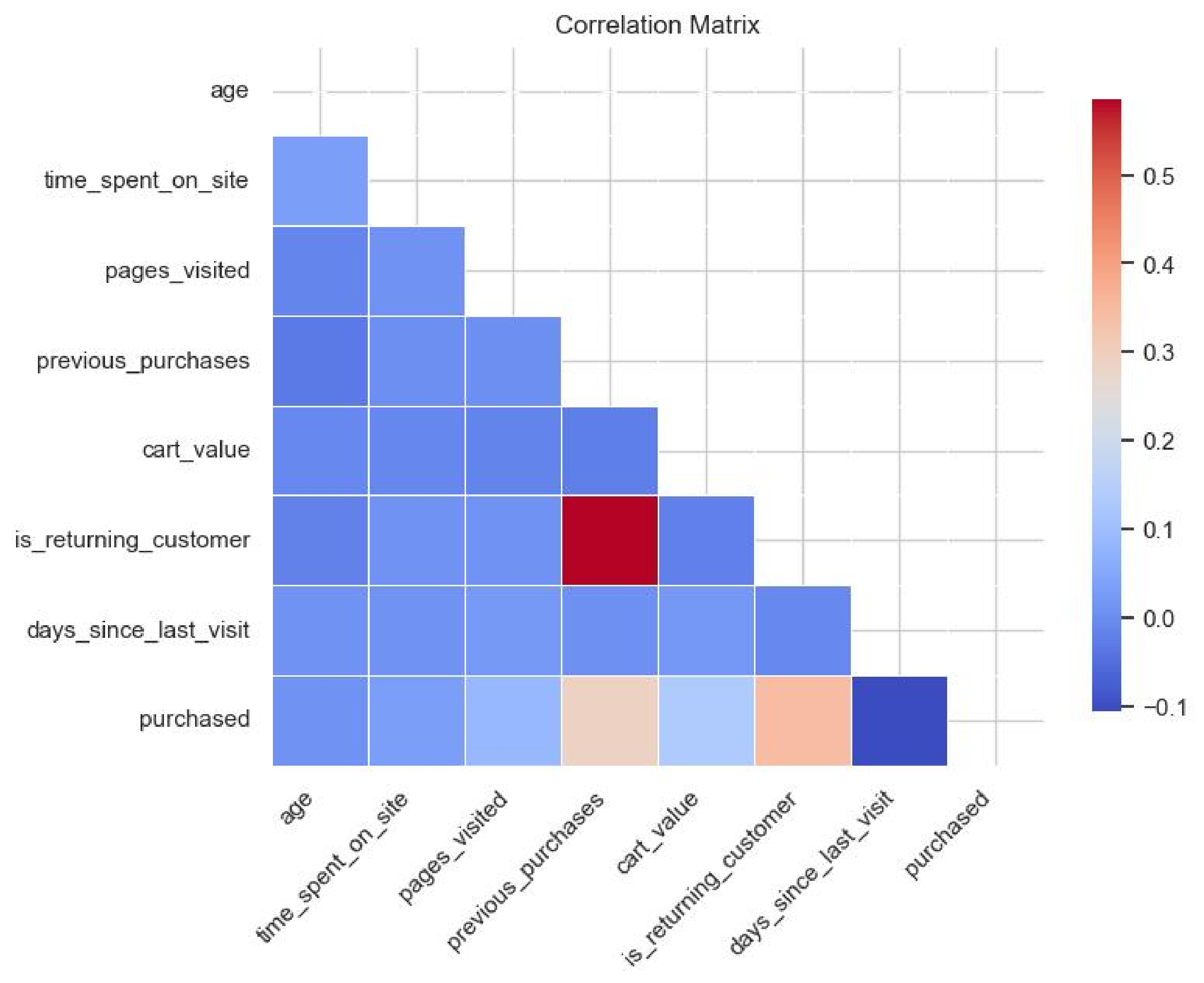


2.2 EDA

- ***is_returning_customer*** ↔ ***previous_purchases*** correlation coefficient ≈ 0.6: returning customers do bring more historical purchases.

- ***purchased*** is positively correlated with **is_returning_customer** and negatively correlated with **days_since_last_visit.**

- The rest of the features are weakly correlated, suggesting that interaction terms or non-linear models can be added to dig deeper relationships.



3.Feature Engineering

```
Feature Importance (sorted by absolute weight):
                   Feature  Coefficient
0  is_returning_customer     0.558644
1  days_since_last_visit    -0.375912
2             cart_value     0.354848
3     previous_purchases     0.350942
4          pages_visited     0.217978
5     time_spent_on_site     0.074769
6       engagement_score     0.059230
7                    age     0.055846
8          recency_score     0.043328
```

- Original 7 characteristics → only describes “pieces of behavior”, hard to capture at once Value, activity, loyalty
- Objective:
    - Extract RFM concept (Recency-Frequency-Monetary)
    - Compress long tail & normalize to mitigate numerical bias

New features:

- recency_score： 1 / (1 + days_since_last_visit)
- engagement_score： time_spent_on_site × pages_visited / 10

Rank	Features	Explain
1	is_returning_customer	Returning customer × 1.75 Chance of purchase
2	days_since_last_visit	Conversion ↓32 % per 1-day delay
3	cart_value	High-value shopping carts lead to high purchase intent
4	previous_purchases	Historical Transaction Count Drive
…	…	…

4. Model Construction

The decision function for logistic regression is：

$$\text{logit}(p) = \ln \frac{p}{1\text{-}p} = \beta_0 + \sum_i \beta_i x_i$$

StandardScaler： Normalization of numerical features

OneHotEncoder： Expand category columns such as age_bin to 0/1

ColumnTransformer: Route to different transformations by column type.

Pipeline: serial encapsulation, followed by LogisticRegression

LogisticRegression: last step, solving for minimizing logarithmic loss

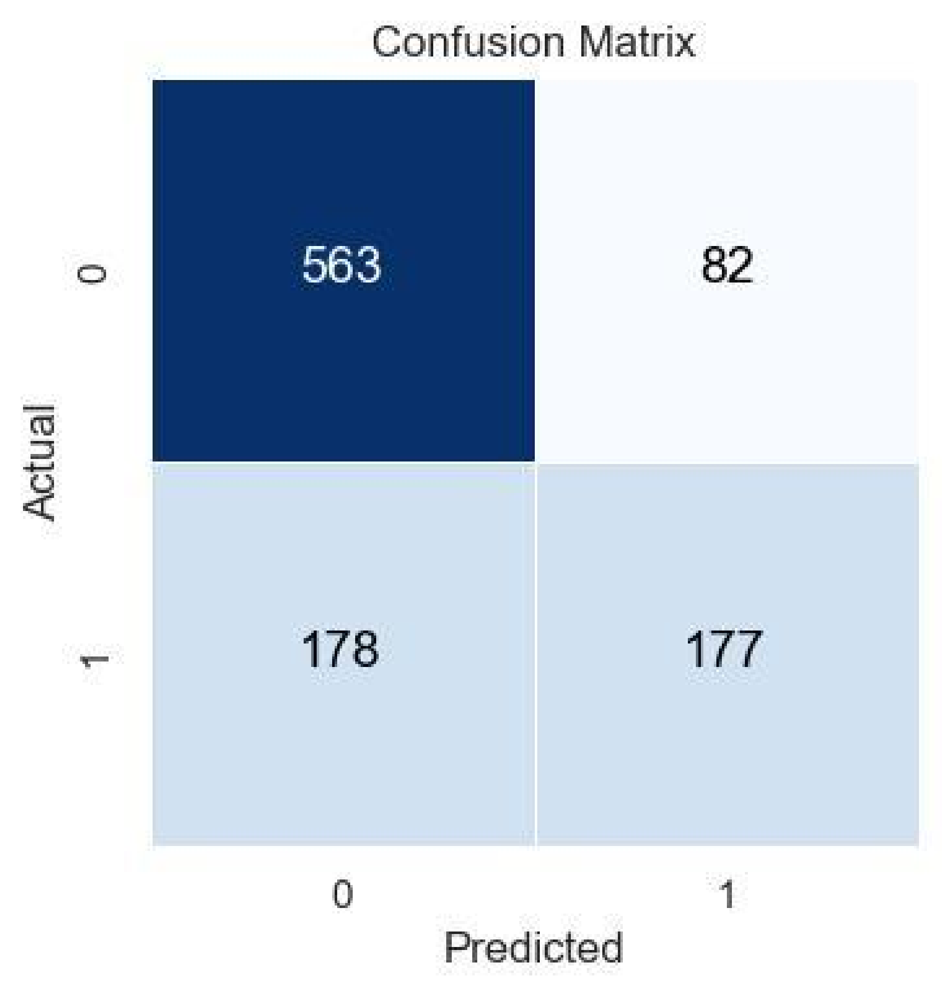
5.1 Model Evaluation

	precision	recall	f1-score	support
0	0.76	0.87	0.81	645
1	0.68	0.50	0.58	355
accuracy			0.74	1000
macro avg	0.72	0.69	0.69	1000
weighted avg	0.73	0.74	0.73	1000

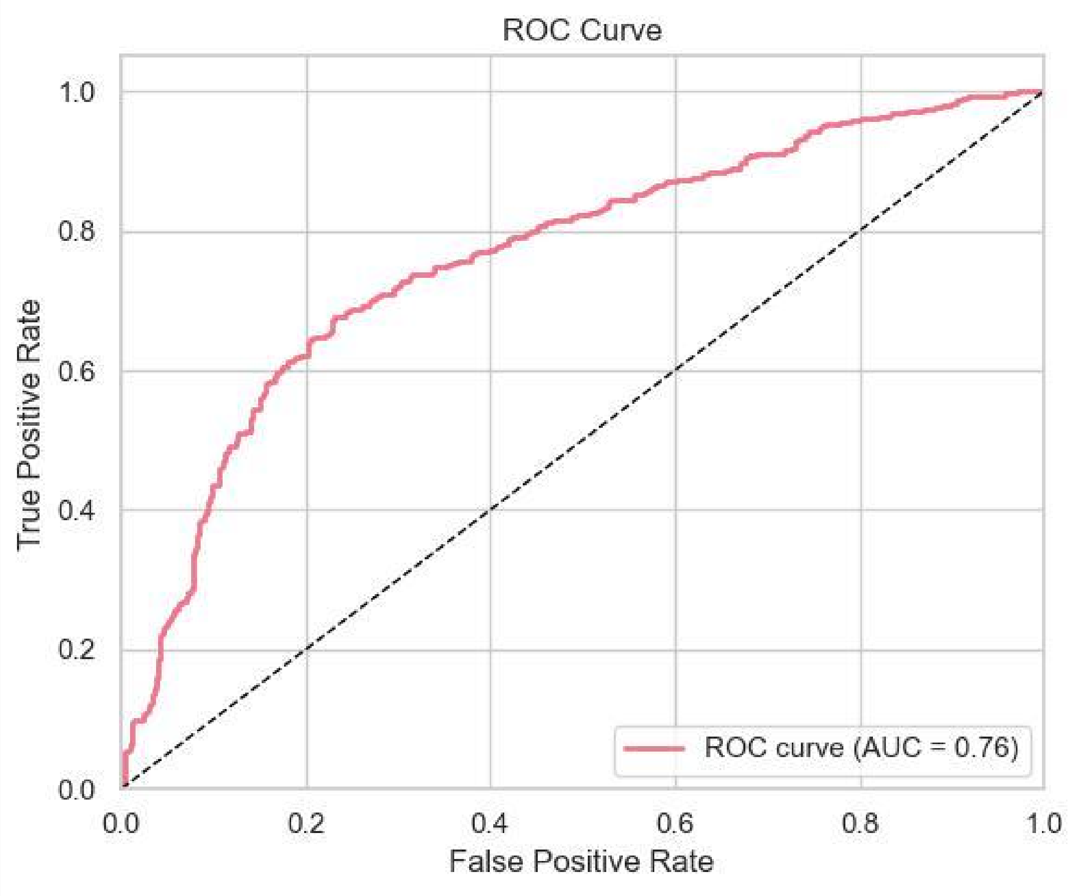
Overall 74 % of the sample was correctly categorized. Accuracy alone is not sufficient due to category imbalance. Unpurchased recognition is very stable, with few misses.

- **TP 178** Correctly found 178 potential buyers
- **FN 177** Missed 177 potential buyers → Insufficient recalls

Purchasing users were "fished" half the time; a further 32% were incorrectly pushed to marketing.



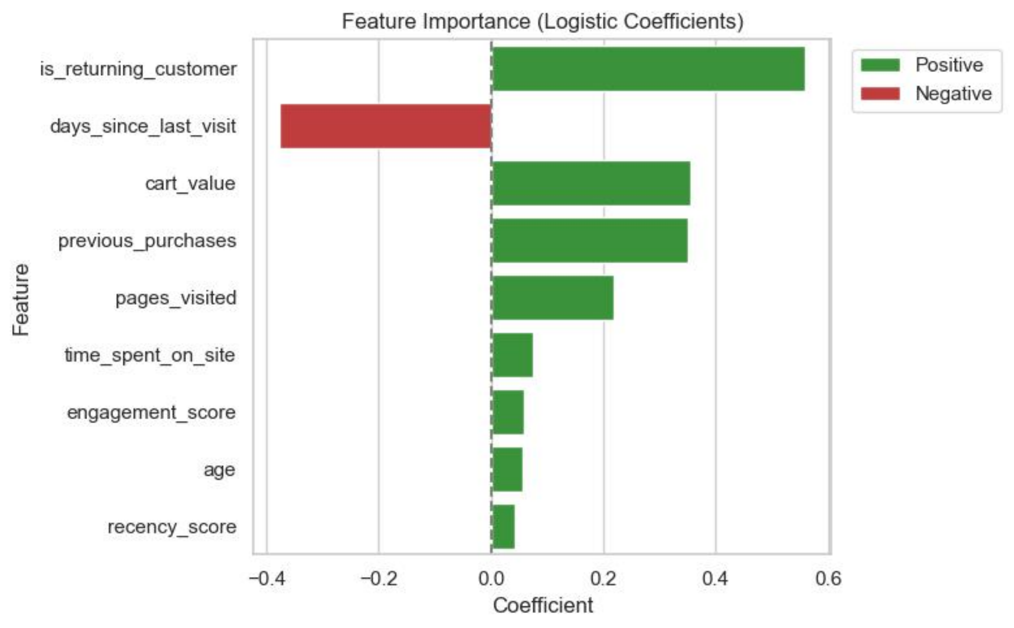
5.2 Model Evaluation



According to the **ROC-AUC chart,** the model has a **76%** probability of ranking positive samples first at all possible thresholds for a random pair of "positive/negative" samples - a medium to high performance.
Cross-validation Acc 0.72 ± 0.02 indicates that the 5-fold results are stable with low variance and the model has some generalization ability to different sampling splits.

The current thresholds focus on **accuracy**, if the business is more concerned about not missing potential buyers, try adjusting the thresholds downwards or use **F2 / Recall** as an optimization target.
Also can use business KPI quantification (e.g., marketing costs) to determine the model's focus.

6. Business Insights



Odds Ratios (impact on purchase odds):

Feature	Coefficient	Odds_Ratio
is_returning_customer	0.559	1.748
cart_value	0.355	1.426
previous_purchases	0.351	1.420
pages_visited	0.218	1.244
time_spent_on_site	0.075	1.078
engagement_score	0.059	1.061
age	0.056	1.057
recency_score	0.043	1.044
days_since_last_visit	-0.376	0.687

According to the bar chart it is clear that different consumer behaviors drive willingness to buy positively and negatively. Based on the factors that drive the most, the company should primarily consider:

- Returning customers are 1.75 times more likely to place an order than new customers → *Maintaining members and secondary marketing is the highest priority.*
- For every +50 of shopping cart amount, the chance of purchasing increases by 43% → *Pushing "Discount Coupon / Free Shipping" for high shopping cart users can maximize the revenue.*
- For each additional day of no visit, the chance of purchase decreases by 31% → *Recall should be emphasized for 30 days of no visit.*
- The more historical orders, the more likely to repurchase → *Design loyalty programs such as points and VIP.*

7. Model Application

01

Pipeline One-Click Prediction

○ predict_purchase_probability(df_raw, model)
○ Auto-completion of project features → Align training columns → Output

02

Threshold adjustable

○ Default 0.5; can be adjusted to 0.3 by ROI to increase recall, or 0.7 to lock in precision

03

Batch & Real-time compatible

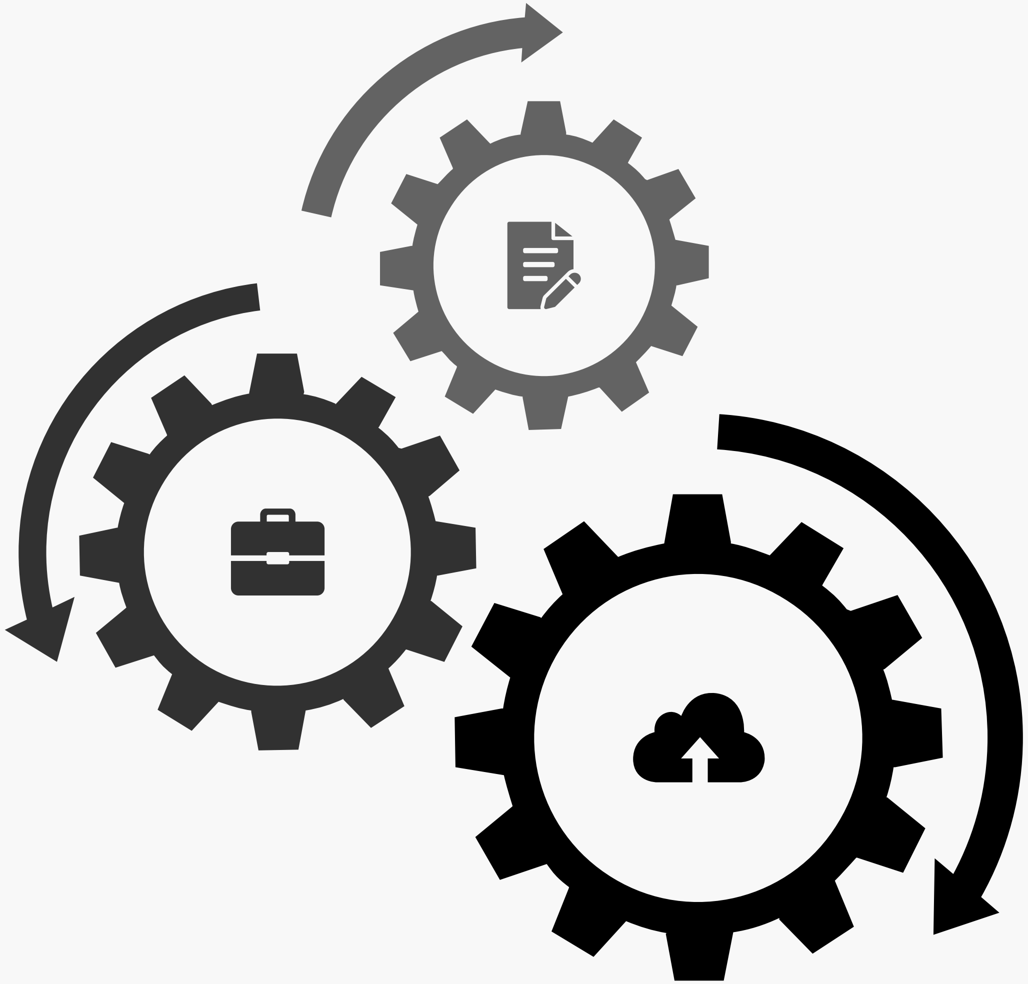
○ Notebook batch：model.predict_proba(df)
○ API realtime：Flask/FastAPI + joblib.load('purchase_prediction_model.pkl')

04

Robustness guarantees

○ Missing columns are automatically filled with 0
○ Empty table throws explicit error
○ Version Number + Training Time Writemodel.metadata_

8. Business Scenarios



01

Targeted marketing

Increase ROAS by focusing your budget on the 25% most likely ($p \geq 0.7$) to buy.

02

Dynamic pricing

Issuance of small discounts / coupons to reduce churn and protect margins ($0.4 \leq p < 0.7$)

03

Personalized Customer Journey

In-site AB test
Increase Stay & Conversion, Reduce Bounce

04

Inventory management

Aggregate by day $\Sigma\, p_i$: Predict category demand
Reduce out-of-stocks / over-stocks and optimize cash flow

Thank you for watching

Zhemin Xie