Improving E-commerce Marketing Effectiveness:

Predicting Changes in Consumer Behavior and Influencing Factors

Group 9

Yi-Ting Lin, Guan- yu Chen, Liang-ru Wu

91APP

Agenda

1. Research Motivation
2. Data Preprocessing
3. Statistical Analysis
4. Model Training
5. Business Validation and Discussion

Customer Relationship Management → Customer Segmentation(RFM) → Personalized Marketing

Proposing More Dynamic Metrics to Capture Changes in Consumer Behavior

	Past Marketing	Current Marketing
Right person	✓	✓
Right content	?	✓
Right time	?	✓

Data Preprocessing Workflow

Data Filtering and Merging	● Read **Member Data** and **Behavior Data from Aug 2021 to Feb 2023** and perform data merging. ● Delete columns with non-quantifiable data (e.g., columns with too many categories that cannot be converted into dummy variables, or with too few "1"s after conversion, resulting in low discriminative power).
Handling Missing Values	● Missing data includes fields such as **gender, registration date** (e.g., 1900/1/1), **ad campaign push frequency**, etc. ● Mark missing gender as **-1**. ● Replace missing age values with the **median age (41 years)**. ● Replace other missing values (e.g., ad push frequency) with the **mean** for that feature. ● Delete rows with more than **6 missing values**; for other rows, fill the remaining missing values with **the mean**.
Categorical Variable Encoding	**Variable Conversion using One-Hot Encoding:** ● Convert variables like ad push acceptance, app ownership, referral status, gender, etc. **Variable Conversion using Frequency Encoding:** ● Convert variables such as eight types of behavior data, device usage frequency, ad platform push frequency, etc. **Date Conversion:** ● Convert dates into years or days from the current date (e.g., registration date, date of birth, engagement date).

Data Filtering Results: Behavioral Data

Eight Types of Behavioral Data

session count: the total number of times of various behaviors of this customer
Includes: member registration, page browsing, product page browsing, search, add to shopping cart, start checkout, purchase, click items
Segmented according to page type, and finally a total of 13 features were taken

Device Usage Frequency

Behavior Frequency (Count):

Devices include:
- iOS App
- Android App
- Desktop
- Mobile Web

4 device-related features in total.

Average Product Price

Objective: To segment customers based on different price categories.
Method: Calculate the average price of products viewed, purchased, and added to cart by each customer during the month.

Data Filtering Results: Behavioral Data

Notifications for Ad Campaigns	Notifications per Advertising Platform	Average Web Page Viewing Time
• Filter the **Top 20 Campaigns** by Push Notification Count for Each Month • **Includes:** googleAD, HBDgift, eCoupon, outlet, ecsite, Valentine’s Day KV, appdownload, Flexible Packages, Multi-Item Discounts	• Take out the **total number of pushes** from each advertising platform. • **Includes:** Yahoopush, IGpush, FBPO... etc. • Total **14 features**	• **Purpose:** To retrieve the **average duration** of user session. • **Method:** Since the project is assigned to the right side of the strict sum of the browsing time, **take the log of the browsing time.**

Data Filtering Results: Customer Data

Dummy

Mark Yes as 1 and No as 0:

- APPRefereeId,
- IsAppInstalled,
- IsEnableEmail, IsEnablePushNotification

Year/Day

- **RegisterAge:** The number of days since the date of data marking (total time of registered members).
- **Age:** Age of the user

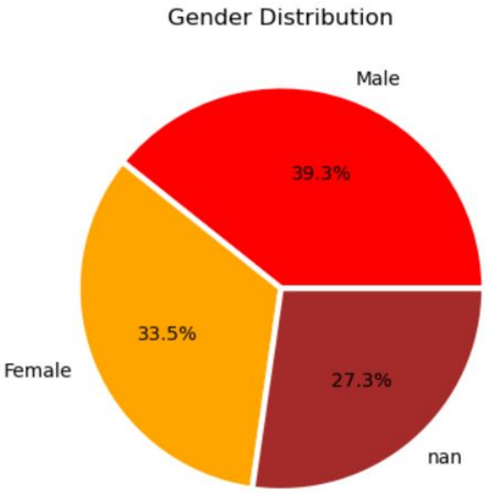
Membership Level

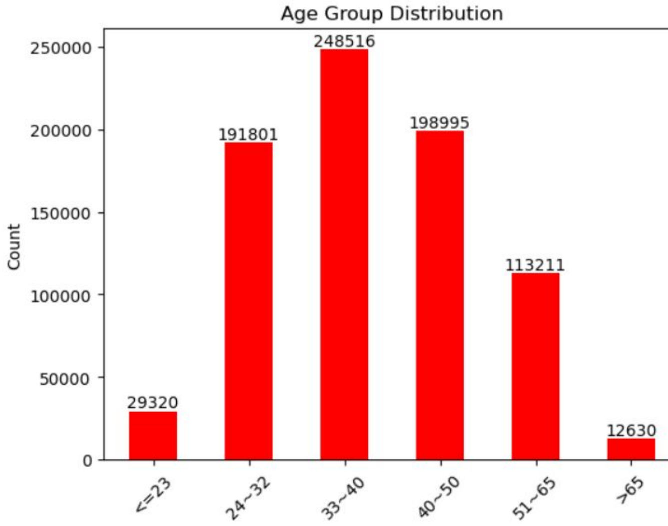
- **MemberCardLevel:** Level 10/20/30

Data Labeling Process

Customer Activity Indicator (Calculated using CAI)	1. The consumption behavior of customers in six months (the source of information is the master list data), and the number of days between consumption is calculated. 2. Calculate the CAI of different customers from **February 2022 to January 2023.** 3. The customers with **higher CAI are labeled as 1**, while the lower ones are **labeled as 0** (about the top 30%).
Data Labeling Inspection	● **Average number of tags:** 1 (active): 9664, 0 (inactive): 11978 ● **Big difference in number of tags between months**: Average: 1803.5, Standard deviation: 482 ○ (Highest: 2651, lowest: 1244 => pay attention to sample size)

Exploratory Data Analysis: Member Data Analysis





- Membership card level is mainly 10

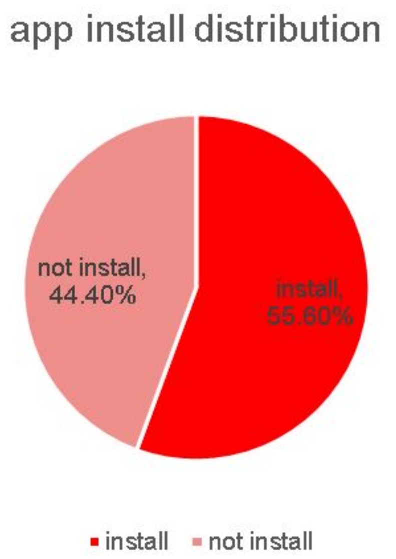
NaN values are labeled as -1.

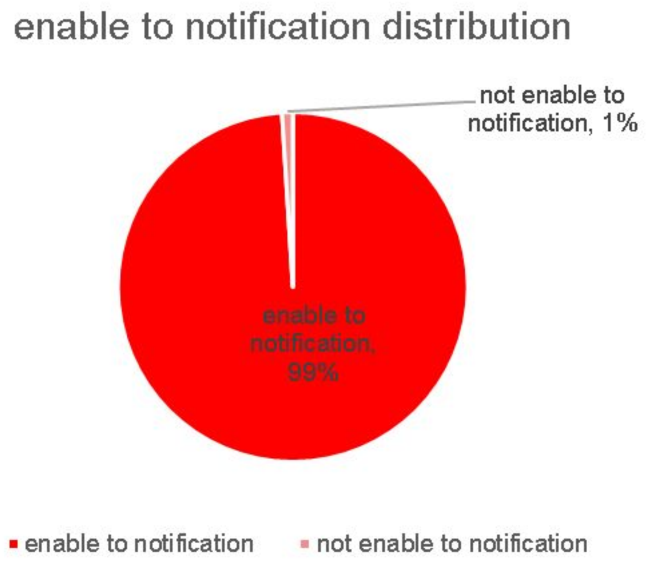
The median age is 41 years.

Exploratory Data Analysis: App Usage and Push Notification Analysis

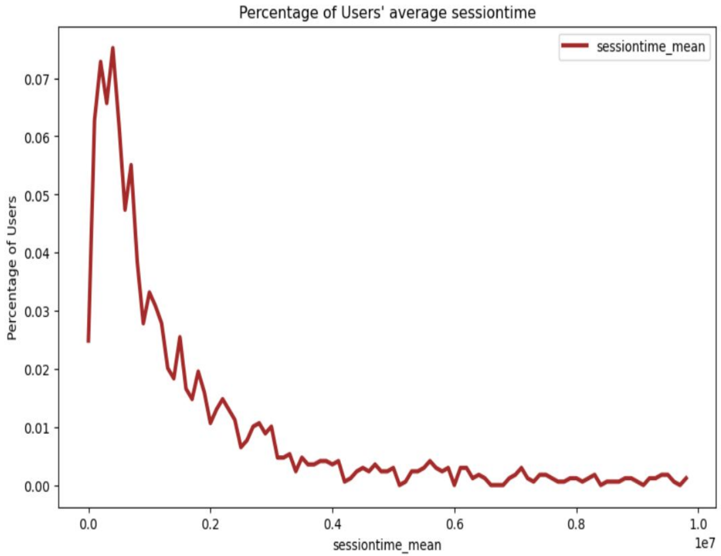
- App Install Rate:
  The app install rate exceeds 50%.

- Push Notification Acceptance:
  99% of users have enabled marketing notifications.

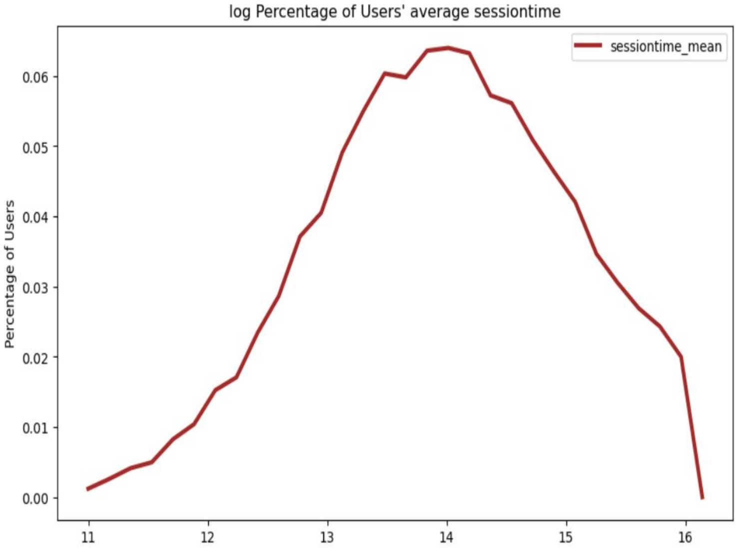




Exploratory Data Analysis: The distribution of both average spending time and total spending time is right-skewed.

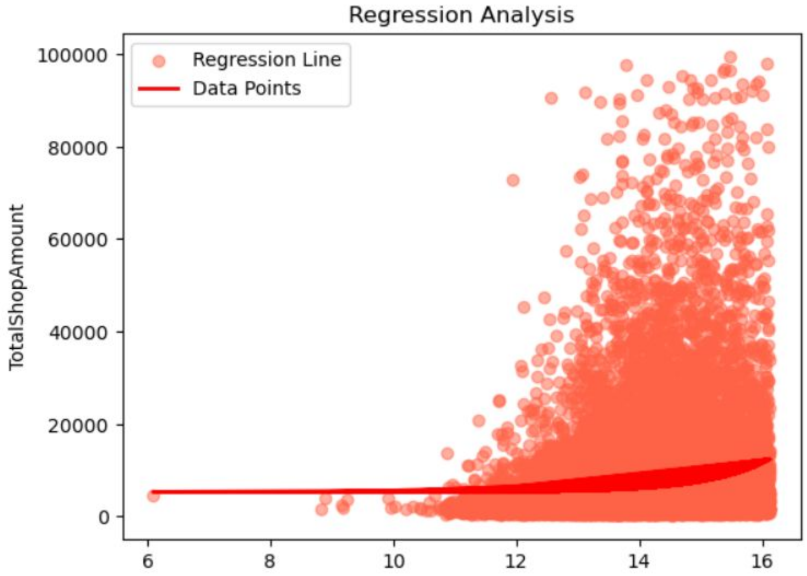


Take log



Data range: February 2022 - October 2022

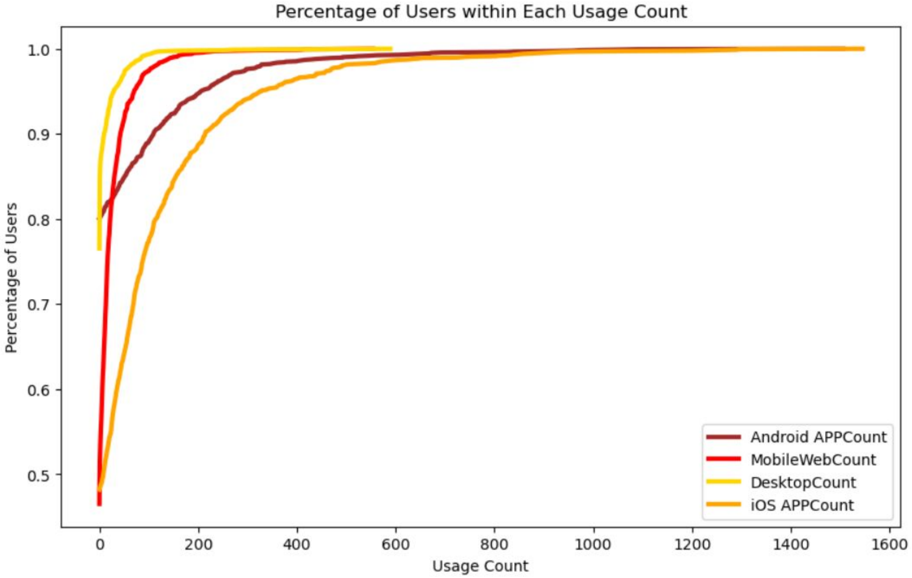
Exploratory Data Analysis: Average spending time and shop amount show a significant positive correlation.

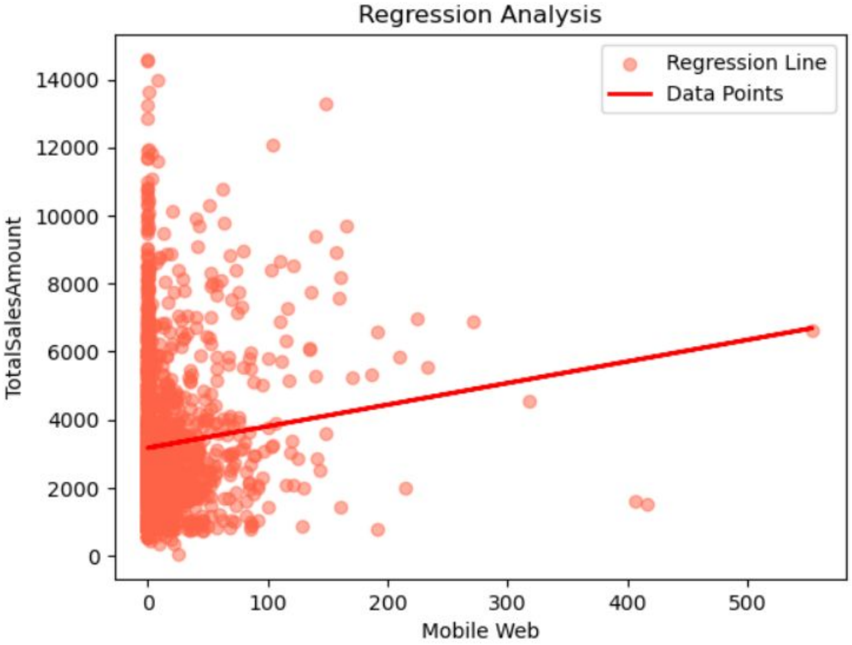



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        TotalShopAmount   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     564.2
Date:                Wed, 31 May 2023   Prob (F-statistic):          2.36e-123
Time:                        15:29:31   Log-Likelihood:            -2.6551e+05
No. Observations:               25022   AIC:                         5.310e+05
Df Residuals:                   25020   BIC:                         5.310e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            5322.4661     87.097     61.109      0.000    5151.750    5493.182
sessiontime_mean    0.0007   2.98e-05     23.753      0.000       0.001       0.001
==============================================================================
Omnibus:                    20507.910   Durbin-Watson:                   1.860
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           506319.180
Skew:                           3.911   Prob(JB):                         0.00
Kurtosis:                      23.602   Cond. No.                     4.10e+06
==============================================================================
```

Data range: February 2022 - October 2022

Exploratory Data Analysis: Frequency of mobile browser usage has the most Positive effect on monthly shop amounts





- Use of mobile apps is the most common
- Slope = 6.36

Feature Selection Process

54features → Elastic Net → 30features → Random Forest → 20features → Correlation Analysis → 16features

Elastic Net

- Select the top 20 most important features from a set of 30 variables.

Random Forest

- From a set of 54 variables, select the top 30 features with the highest absolute coefficient values.

Correlation Analysis

- From the 20 features, remove 4 highly correlated variables with lower importance based on Random Forest.

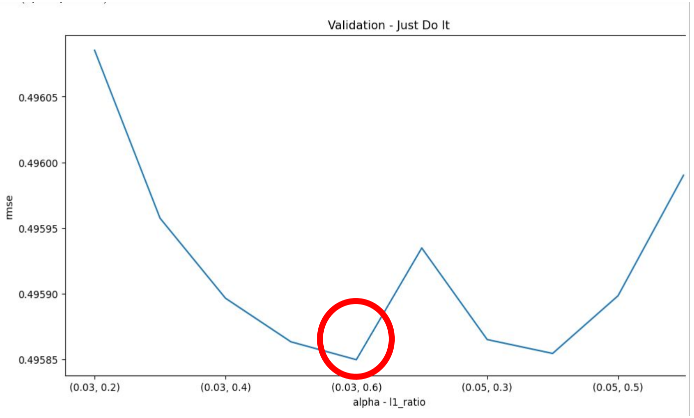
Feature Selection: Elastic Net

Optimal Model Parameters:

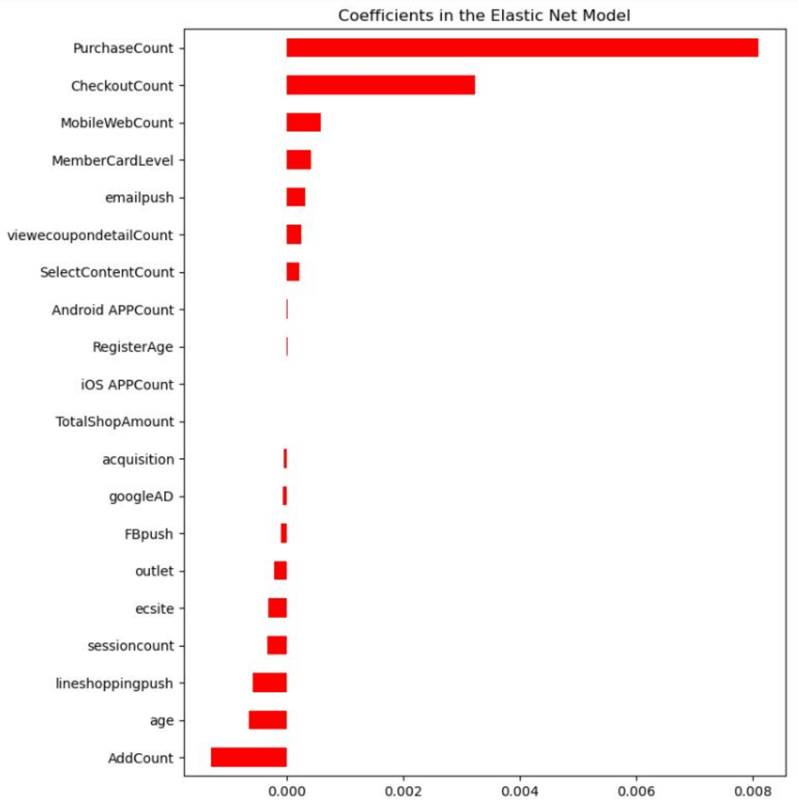
Use **cv_elastic** to select the optimal **alpha** and **l1_ratio** that minimize RMSE. (Minimum RMSE: 0.49)

Select 30 variables

Variables are ranked by the absolute value of their coefficientsbased on Elastic Net., with the top 30 chosen.



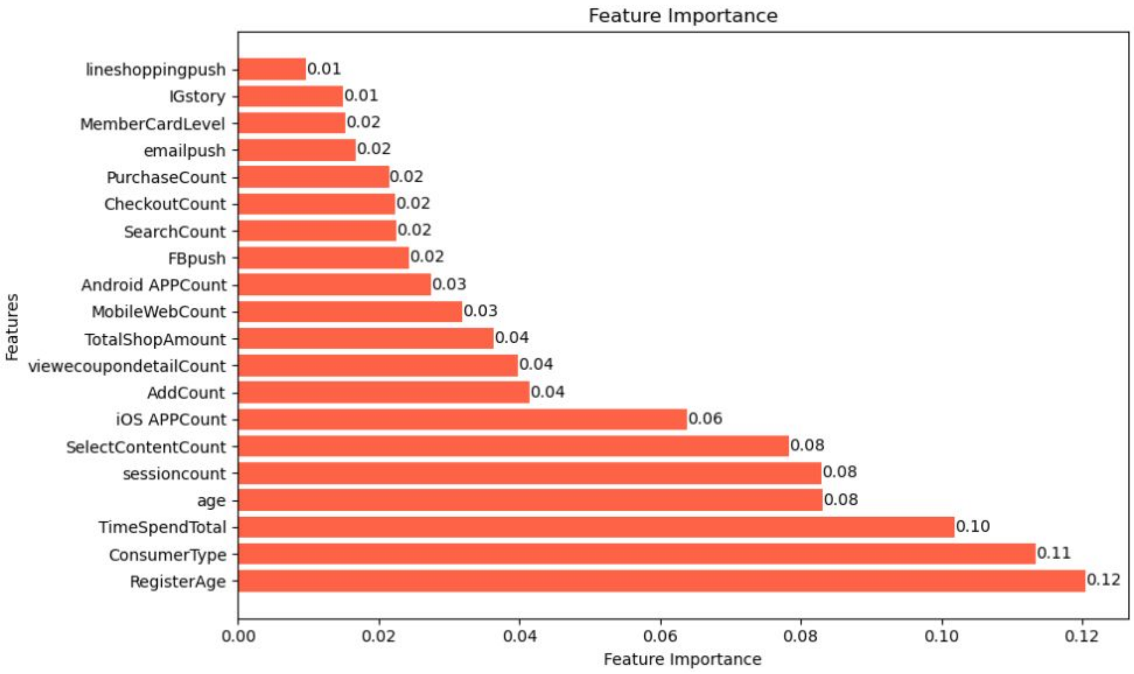
- (alpha,l1_ratio) = (0.03, 0.6) achieves the minimum RMSE.



Data range: February 2022 - October 2022

Feature Selection: Random Forest

- Random forest models are used for training and the top 20 most important coefficients are selected based on feature importance.



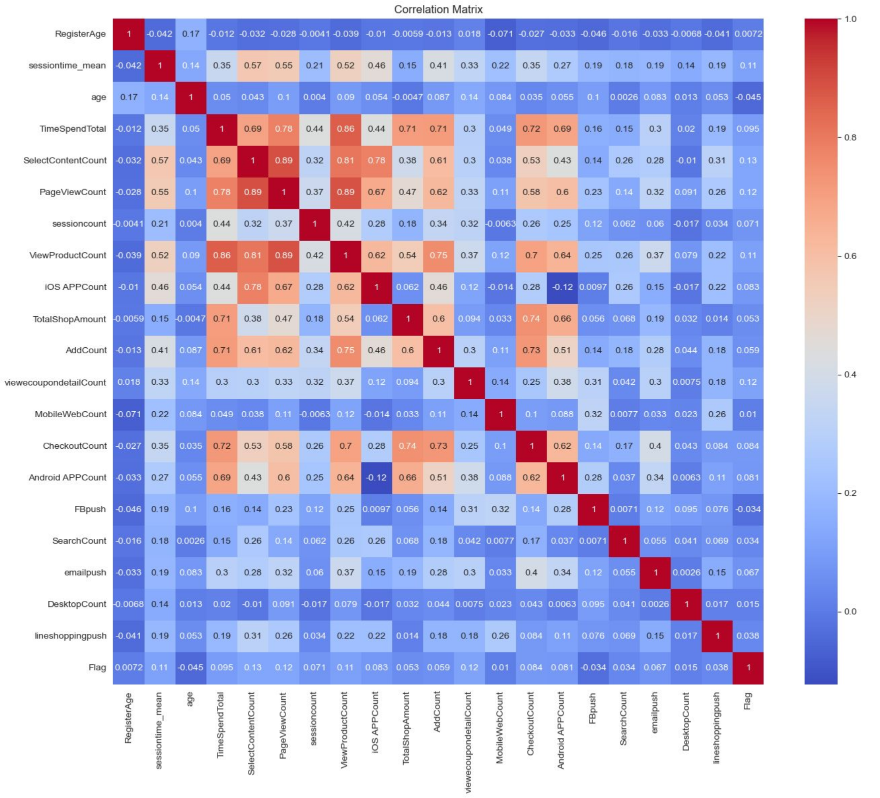
Data range: February 2022 - February 2023

Feature Selection: Correlation Analysis

- Check the correlation coefficients between the selected variables => Delete the coefficients with high correlation coefficients and significance of features that occur together in different months.

- **Delete the variable:**
  1. ViewProductCount
  2. SelectContentCount
  3. AddCount
  4. CheckoutCount

Data range: February 2022 - February 2023



Statistical analysis results: for different age groups

	Top 3 variables name	p-value	Correlation
Age group 1: **age<=25**	1.sessioncount 2.MobileWebCount 3.emailpush	0.018~0.107 0.040~0.382 0.048~0.190	negative for all
Age group 2: **25<age<=45**	1.FBpush 2.DesktopCount 3.emailpush	0.067~0.592 0.134~0.416 0.170~0.634	negative positive negative
Age group 3 **45<age<=65**	1.session_timemean 2. Android APPCount 3. MobileWebCount	0.003~0.139 0.104~0.352 0.008~0.723	negative positive positive

Statistical analysis results: for different membership levels

	Top 3 variables name	p-value	Correlation
membership levels 1: **MemberCardLevel=10**	1. viewecoupondetailCount 2. emailpush 3. Android APPCount	0.093~0.559 0.217~0.262 0.104~0.646	negative negative positive
membership levels 2: **MemberCardLevel=20**	1. Android APPCount 2. PageViewCount 3. iOS APPCount	0.026~0.725 0.042~0.766 0.129~0.473	postive for all
membership levels 3: **MemberCardLevel=30**	1. iOS APPCount 2. Android APPCount 3. DesktopCount	0.080~0.580 0.030~0.786 0.034~0.710	postive for all

Statistical analysis results: for different activities promoted each month

March (202203)

Activity Name	coef	p-value
SS22	165.5591	**0.000**
onsale	-44.1313	0.356
ss2210off	73.0761	0.259

June (202206)

Activity name	coef	p-value
220618	53.1933	**0.122**
618encore	1.9034	0.912
pre618	7.1724	0.758

Statistical analysis results:

Conclusion

1. Emailpush indicators show negative correlation in different subgroups, it is inferred that the content of emailpush is too general and **lacks personalized suggestions or design highlights**.
2. androidappcount and iosappcount indicators have significant impacts on members of different levels, and the correlation is positive, which suggests that modern consumers are used to browsing with cell phones, and the **app interface is more convenient than the mobile webpage**.
3. Most of the significant variables are the channels for receiving messages and browsing, so we can guess **whether consumers are aware of the activities or not has a greater impact on the customer activity level.**

Model Training Process

Objective

- Predict the change in Purchase Count for the next month based on the customer's behavior in the current and previous months.

Data Processing and Feature Selection

- **Selected Features:** 20, **Data Points:** 2698
- **Label:** Change in next month's purchase count,
- **Data Range:** February 2022 to December 2022

Data Splitting

Training | Validation | Testing

Model Building

- **Models used:** Linear, Logistic, Random Forest, XGBoost, LSTM, Stacking
- **Performance metrics:** RMSE, MAE

Model Data Preprocessing Methods

Feature Selection → **Customer Data Filtering** → **Prediction Baseline Setup**

Feature Selection

Select key features using **Elastic Net and Random Forest**.
Use the top **20** features as model inputs.

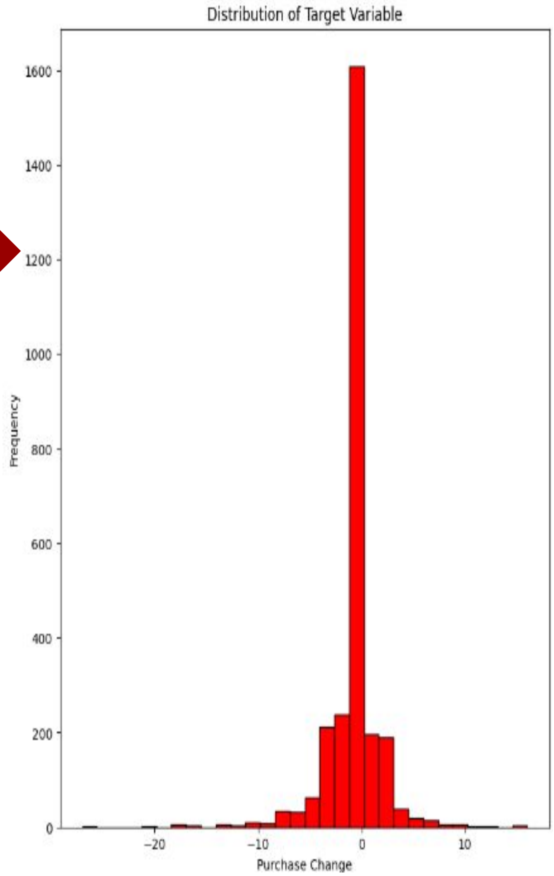
Customer Data Filtering

Filter out data with an active or inactive flag for the next month and subtract the current month's behavioral data from the previous month's data to obtain the change volume.

Prediction Baseline Setup

Use the median change in purchase count (0) as the baseline for predictions.

- RMSE = 2.73
- MAE = 1.47

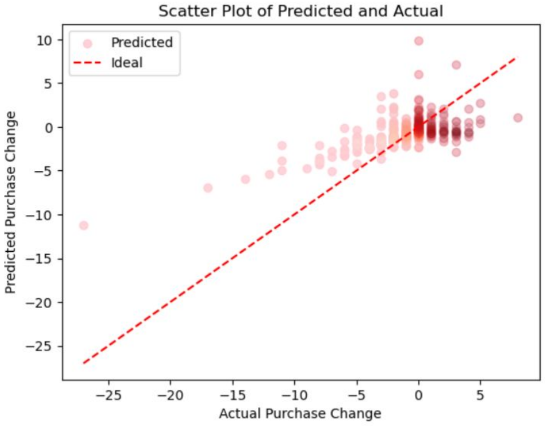


Model Building: Linear and Logistic Regression

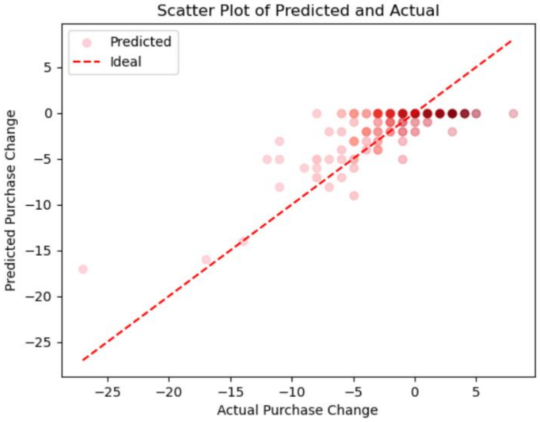
Description of Model Results

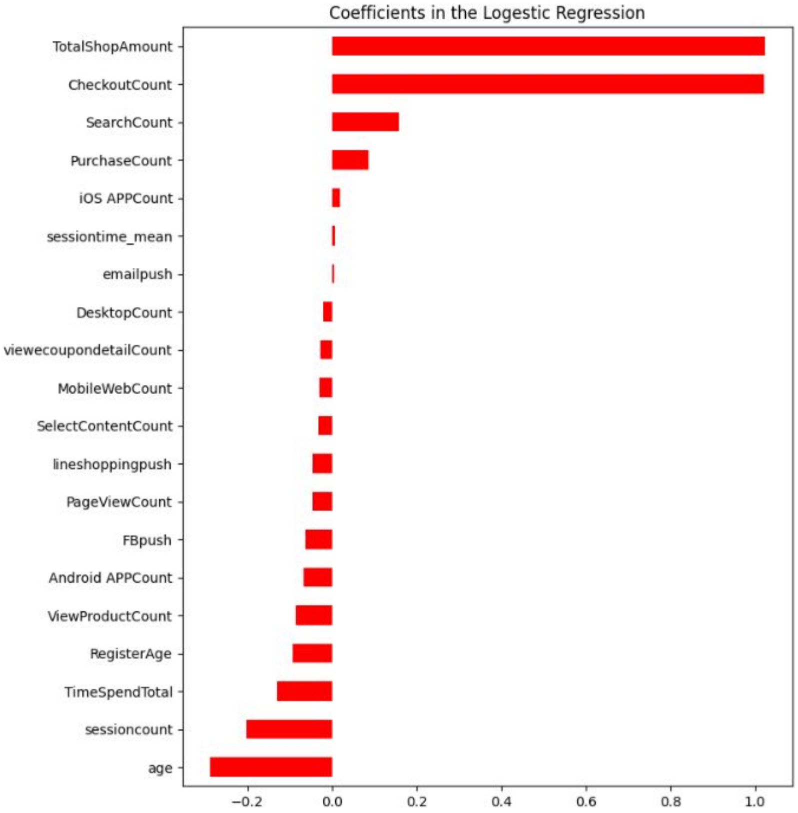
Linear	Better predictor of customers with increase purchase **RMSE = 2.09、MAE = 1.32**
★ **Logistic**	The standardized Features training model is more accurate than the Linear regression model **RMSE = 1.78、MAE = 1.26**

- linear



- logistic





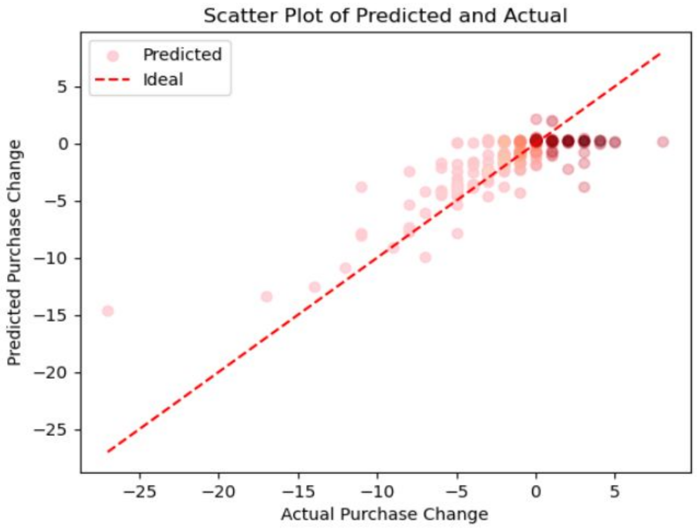
Model Building: Random Forest

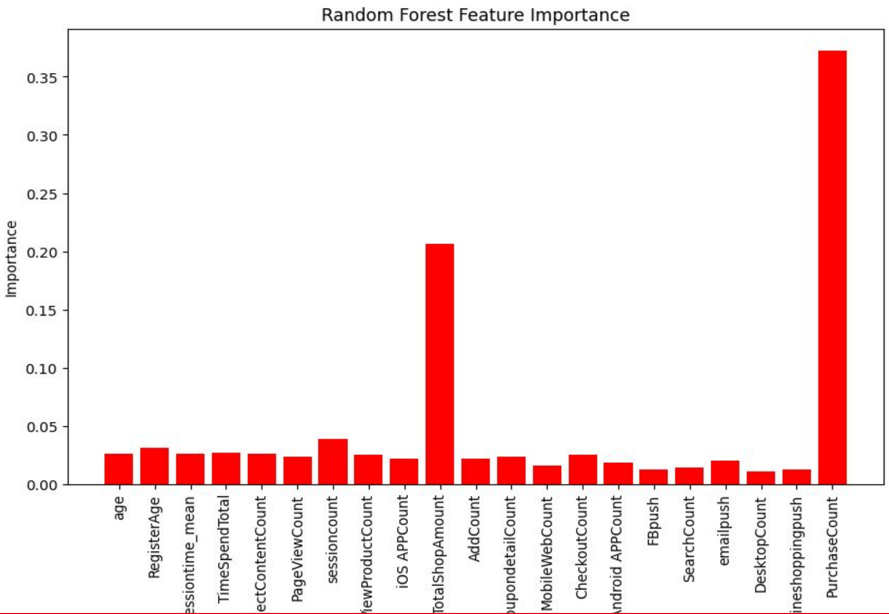
Description of Model Results

Randomized forest prediction results are relatively **good for outliers**: RMSE = 1.55, MAE = 2.42

Parameter Settings

Use **GridsearchCV** to search for the best combination of parameters: max_depth: 5, n_estimators: 150, n_jobs: **-1**





Model Building: XGboost

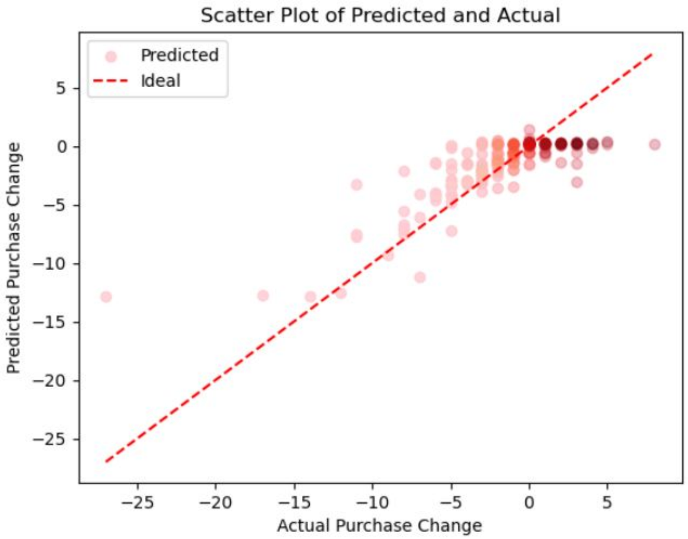
Description of Model Results

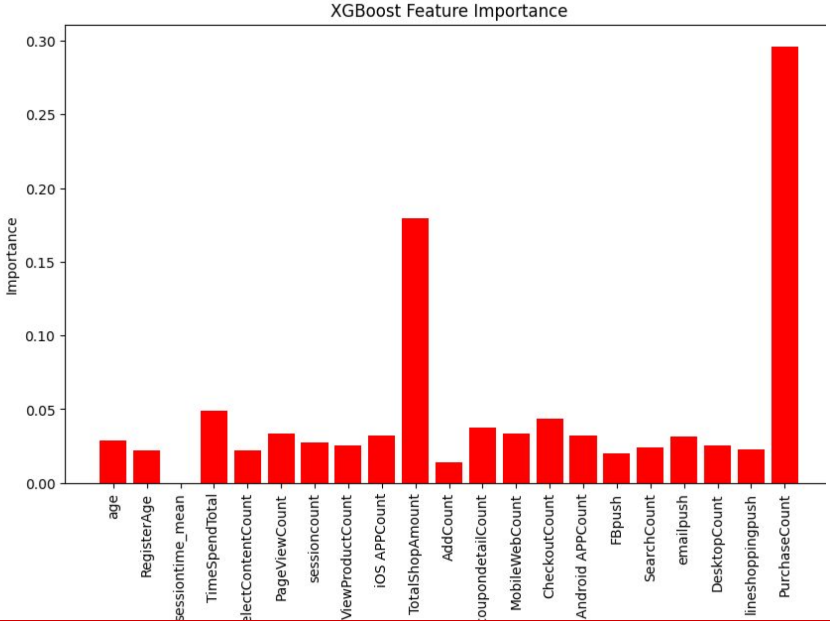
Accuracy is relatively high, but the **ability to predict customers with increased purchases is relatively poor**. RMSE = 1.57, MAE = 0.95

Parameter Settings

Use **GridsearchCV** to search for the best parameters.

learning_rate: 0.01, max_depth: 3, n_estimators: 300





Model Building: Stacking

Model Selection

Based on **model relevance**. Since Logistic Regression is only suitable for categorization, Linear Regressor, which is relatively less accurate, is used as a substitute, and the parameter settings are **based on the results of the former grid search method**.

Base Learners

- Linear Regression
- XGBoost
- Random Forest

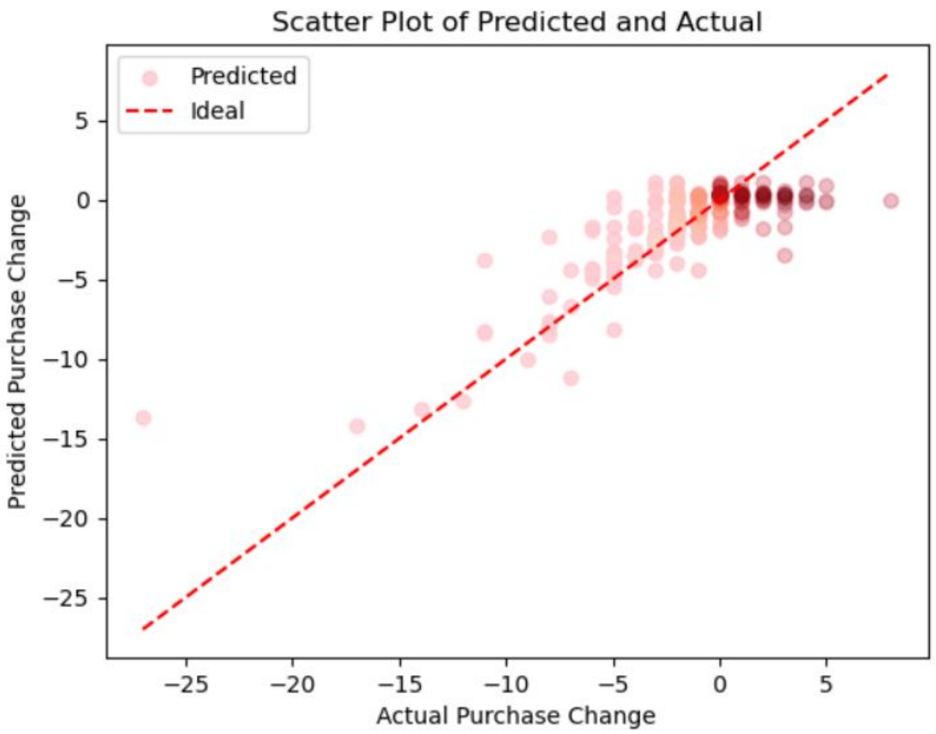
Final Estimator

MLP Regressor

hidden_layer_sizes=(8, 8)

Description of Model Results

Combining the three models gives t**he most accurate output**, especially for predicting **customers with declining purchases**. RMSE = 1.53、MAE = 0.92



Model Prediction Results

All models are generally effective in predicting the change in the number of purchases in the next month, while **Stacking** has the best performance in the test data, followed by **XGBoost**, and **LSTM** has the worst performance. Therefore, the important features selected by **XGBoost** will be analyzed in the following section of the result discussion.

	RMSE	MAE
Logistic Regression	1.78	1.05
Linear Regression	2.09	1.32
Random Forest	1.55	2.42
XGBoost	1.57	0.95
LSTM	2.75	1.47
Stacking	1.53 ☆	0.92 ☆
Base Line	2.73	1.47

Discussion

Data Processing Issues

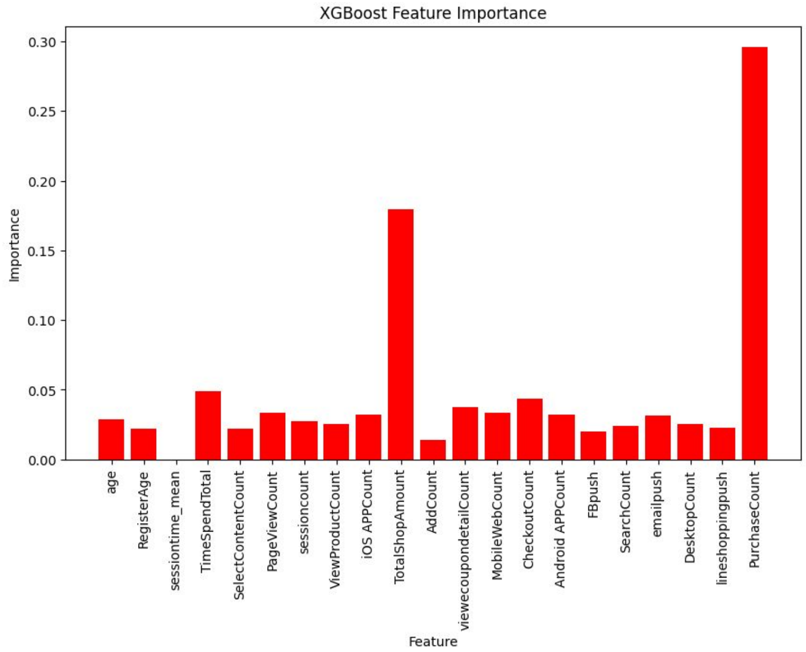
- CAI is calculated based on the interval between customer purchases. However, we identified **discrepancies between order records and behavior data**, likely due to differences between online and offline records or data inconsistencies.
- As a result, the **monthly change in purchase count** was used as the label, focusing on customers with significant changes in CAI.

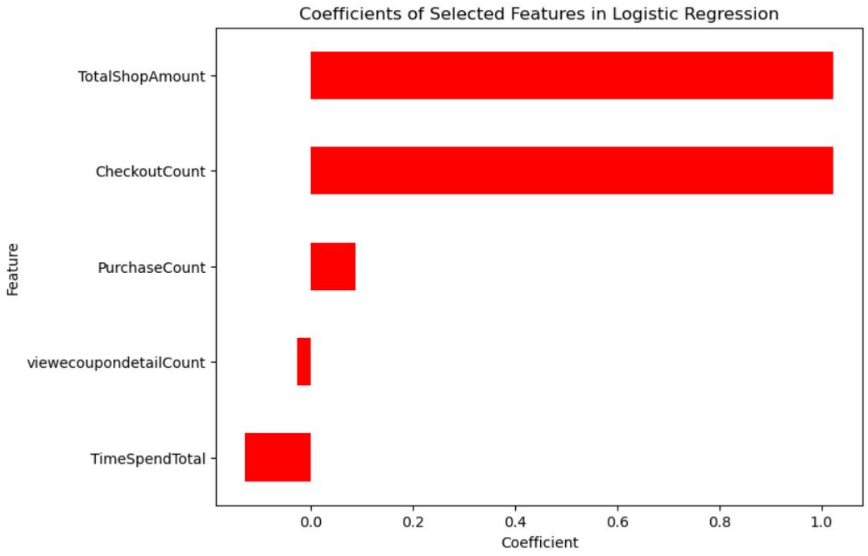
Model Improvement Directions

- Use **longer time periods** as training data to forecast purchase changes over specific months.
- **Balance customers** with significant increases and decreases in purchase volume to enhance prediction accuracy for purchase growth.
- Consider the **rate of behavioral data change**, not just the absolute change, for improved insights.

Result Analysis

When enterprises predict the change of purchasing frequency in the next month, they can refer to the **change of customers' purchasing quantity, the change of shop amount, the change of total spending time on the platform, the change of checkout count, and the change of the times of viewing coupons** as the basis of analysis, so as to rescue the possible loss of customers as early as possible.





Application of Machine Learning Model Results

Business Application

1. Email broadcasting is relatively ineffective, so we can **appropriately reduce the proportion of email broadcasting or increase the personalized content of email ads**.
2. Regardless of android or ios system, apps are the main browsing path for members, so we can try to **optimize the app experience and increase the willingness to download apps by giving away shopping bonuses for downloading apps.**
3. Under different segmentation methods, the more significant variables mainly focus on the difference between the receiving channel and the browsing channel. Therefore, it is more important to **determine whether the channel of the marketing campaign is appropriate and whether the campaign message can be successfully delivered to customers** than the content of the campaign and promotion.
4. To predict changes in customer purchases, we should **take into account changes in purchase amount, number of purchases, time spent on the platform, and viewing of coupons. Utilizing these indicators to track real-time consumption trends** makes it easier to maintain good customer relationships and increase customer loyalty.

Review and Improvement Direction

Problem Review

1. For different product categories, **different time intervals may be suitable for the calculation of CAI indicators**, for example, e-commerce data such as FMCG industry can be shortened to 6 months, and vice versa, longer time intervals are suitable, so we can make different attempts in the data exploration stage.
2. In the part of statistical model, we can try more different kinds of regression models, maybe we can get a more suitable model for this data type.
3. Try more grouping methods for regression and find out more important variables that are unique to each group.
4. Replace the binary categorization (very active, moderately active, ...) with a more hierarchical categorization of CAIs.

Questions?