The data set is customer behavior survey data

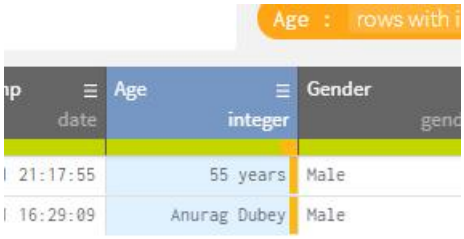
My Data structure：
Timestamp： Date of last purchase
Age
Gender
State
City
MembershipLevel(tier1 is the best)
Marital Status
Occupation
Favorite Product Category
Time for each product selection
Times of purchases per month (1-5)
Cost per purchase
Churn

1.The data is preprocessed with talend first：

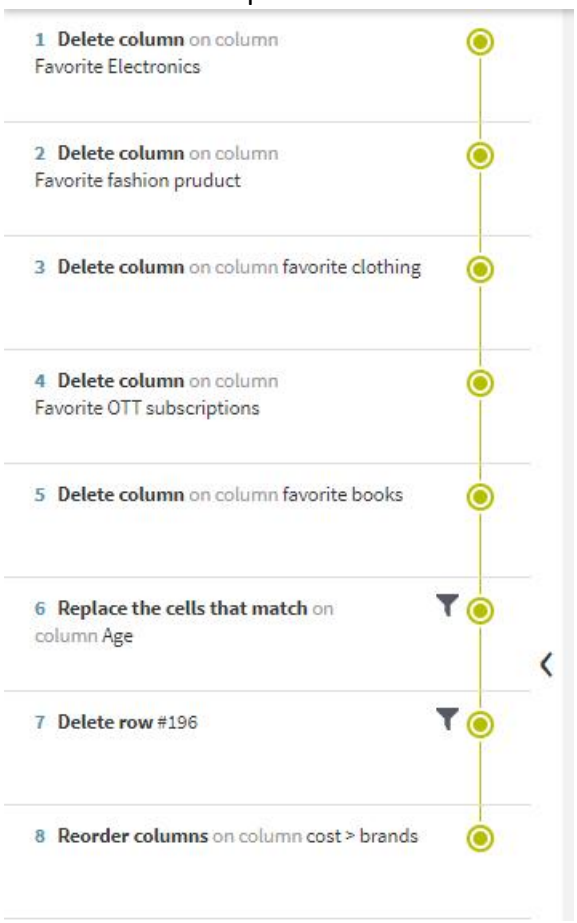
The five columns of customers' favorite electronic products, favorite fashion items, favorite clothes, etc., have too many missing values and are not very helpful to this study, so they are directly deleted.

Favorite Electro... text	Favorite fashion ... text	favorite clothing text	Favorite OTT su... text	favorite books text
Utility Devices (1				
Utility Devices (1				
	Jewellery			
		T-shirts / Shirts		
	Cosmetics			
		T-shirts / Shirts		
			Disney Hotstar	
Mouse / Keyboard				
	Perfumes			
				Fiction
		Shoes		
				Romance
		Lower		
	Purses / Bags			
	Jewellery			

The Age column has two invalid values, and I changed the format of the first value to 55 and deleted the second row.



This is a series of operations that I do with talend：



2. Data Import and Preprocessing

I set the role of churn to target, the role of ID to ID, and the rest to input

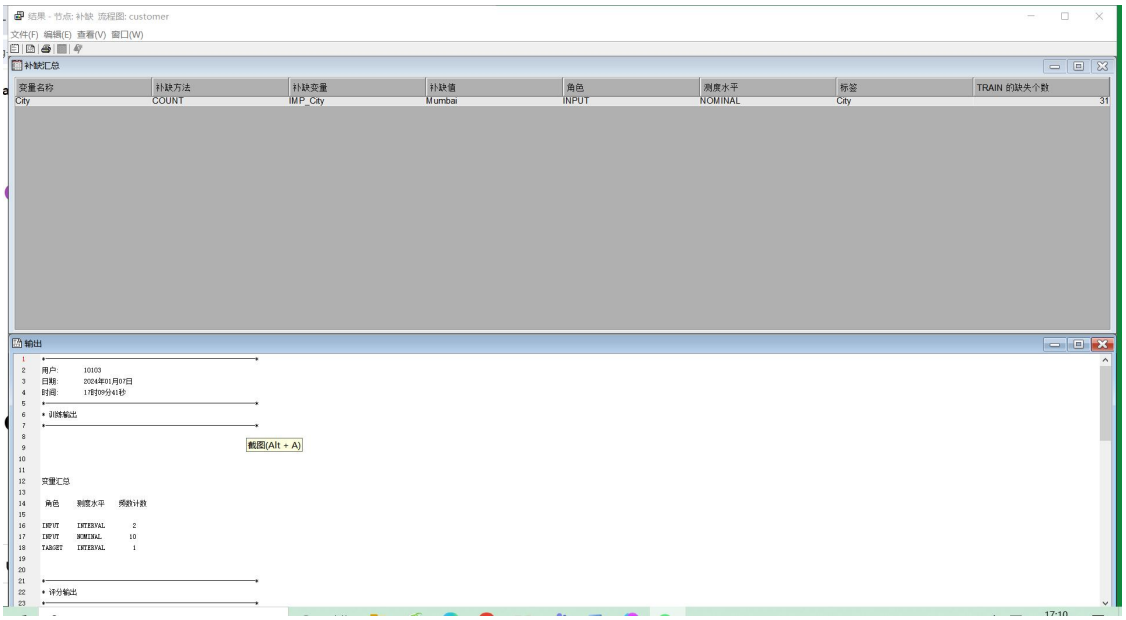
变量 - FIMPORT

(无) 非 等于

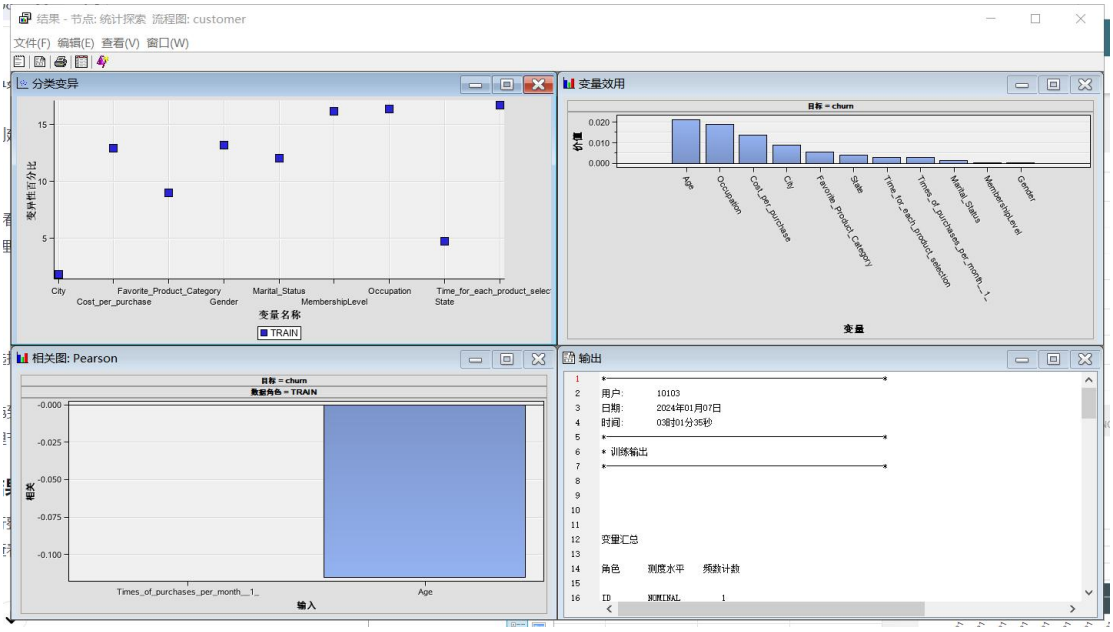
列: 标签(A) 挖掘(M)

名称	角色	水平	报表	顺序	删除	下限	上限
Age	输入	区间型	否		否	.	.
City	输入	列名型	否		否	.	.
Cost_per_p	输入	列名型	否		否	.	.
Favorite_P	输入	列名型	否		否	.	.
Gender	输入	列名型	否		否	.	.
ID	ID	列名型	否		否	.	.
Marital_St	输入	列名型	否		否	.	.
Membership	输入	列名型	否		否	.	.
Occupation	输入	列名型	否		否	.	.
State	输入	列名型	否		否	.	.
Time_for_e	输入	列名型	否		否	.	.
Times_of_p	输入	区间型	否		否	.	.
Timestamp	输入	列名型	否		否	.	.
churn	目标	区间型	否		否	.	.

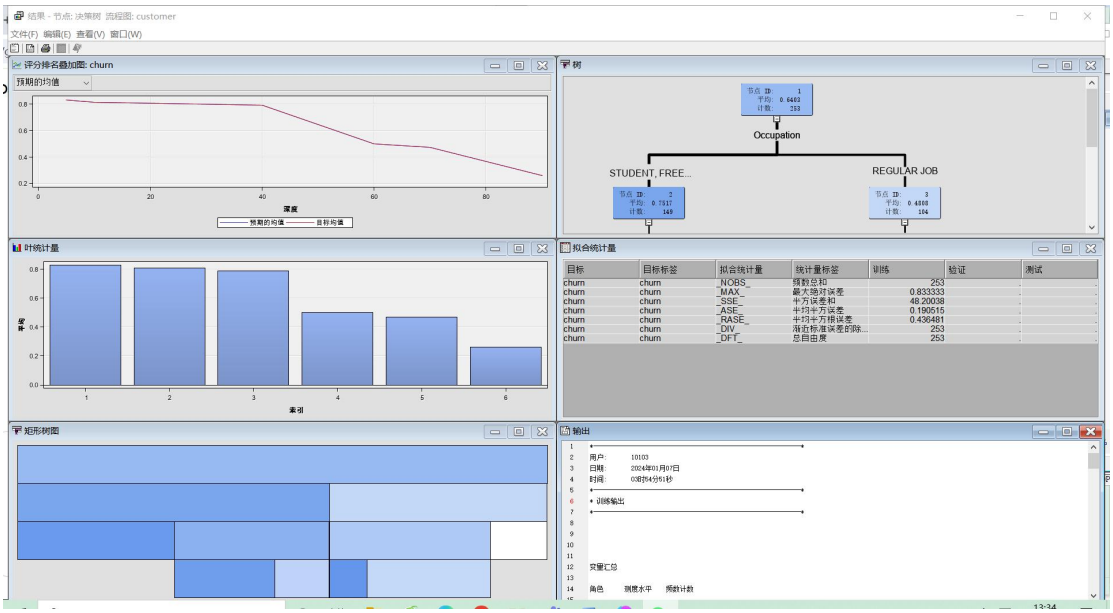
Then the data is cleaned, and the classification variable city is filled with the count method. Only the city column has a few missing values.



The utility of the variables shows the correlation degree of churn between each variable and the target variable. The figure above shows that age, occupation, cost_per_purchase and churn are highly correlated, and it can be inferred that there is a particularly high customer churn rate for a certain age group, occupation, and consumption level.



3. Decision Tree Analysis



Importance of variables: Occupation is considered the most important variable with an importance score of 1.0000, meaning that occupation plays the most critical role in predicting churn. Cost_per_purchase (cost per purchase), City, and several other variables also contribute to the model.

Fitting statistics:

NOBS：253。

MAX ：0.833。

SSE :48.200。

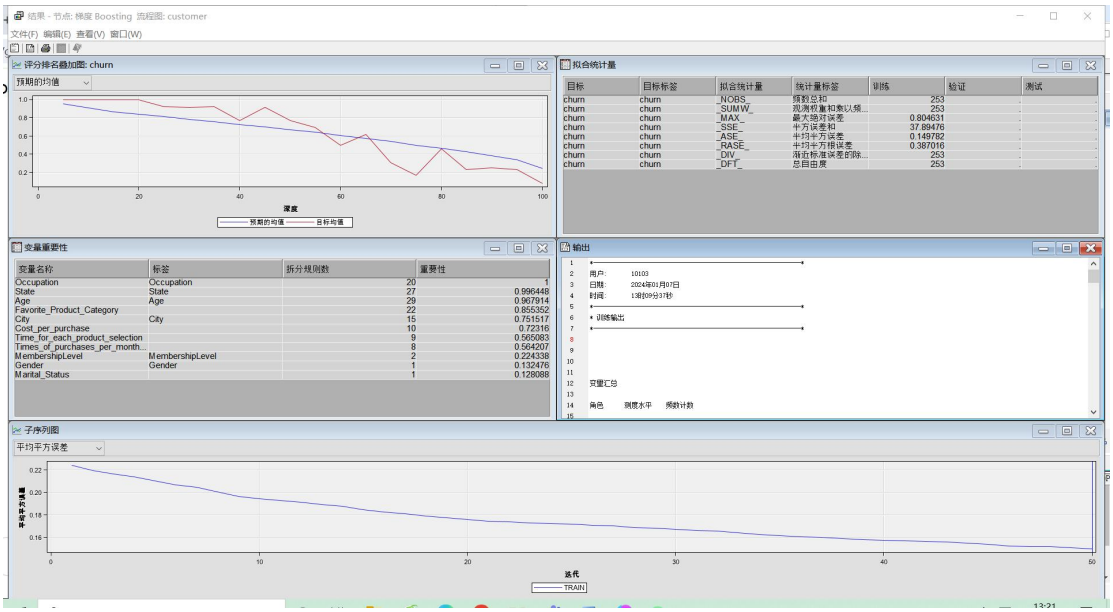
ASE :0.191。

RASE :0.436。

Evaluation score ranking: The target mean is basically consistent with the expected mean,

indicating that the model is accurate in predicting losses at different probability levels.

4.Boosting:



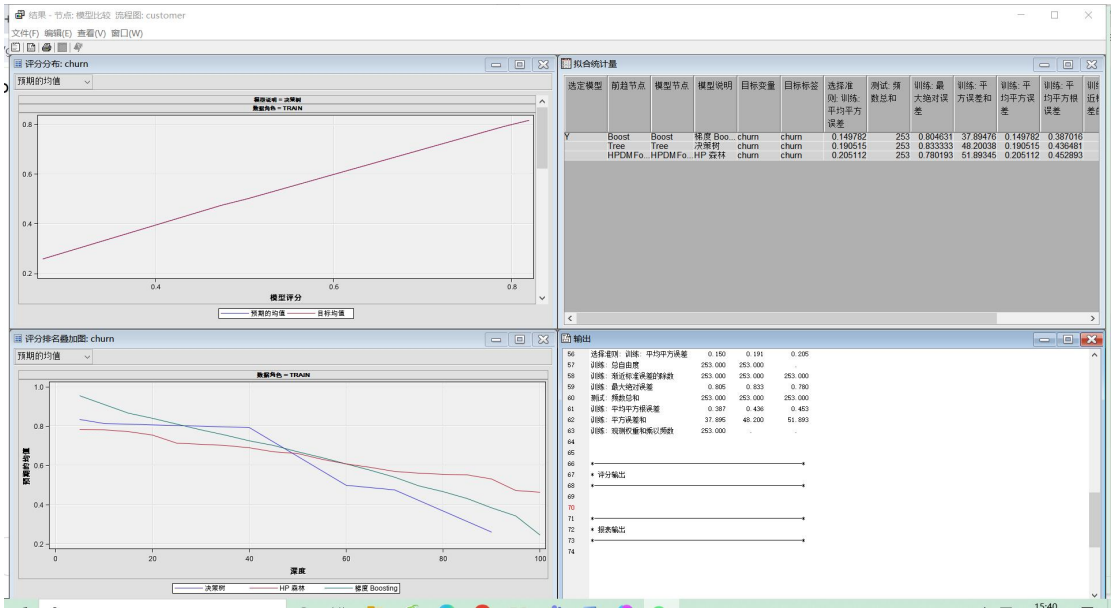
Variable importance: Lists the variables that have a significant impact on the model's predictive churn (customer churn). Occupation (occupation), State (state), and Age (age) are the first three most important predictors. This means that these factors play a key role in predicting whether customers will churn.

The fitting statistics:, ASE　is 0.150, RASE is 0.387.

Evaluation Score Ranking: The top 5% of customers have a 100% churn probability, but the actual churn probability decreases as the ranking decreases. This ranking helps businesses prioritize those customers who are most likely to churn.

Evaluation score distribution: The target mean in the highest rating range (0.937-0.983) is 1, and the expected mean is 0.96069, indicating that the model's predictions for high-risk customers are fairly accurate.

5.Random Forest



Model information:

A random forest of 100 decision trees was constructed.

An Inbag score of 0.6 indicates that approximately 60% of the sample for each tree training is used to build the tree (Inbag) and the remaining 40% is used for OOB estimation.

The mean square error is 0.230

The mean squared error varies with the number of trees, from ASE 0.22166 for 1 tree to ASE 0.20511 for 100 trees, and OOB error from 0.22579 to 0.21695, indicating that the stability of the model increases.

Occupation and Cost_per_purchase were identified as the most important predictors and had a significant impact on the model.

The top 5% of customers had a high actual churn mean (1.00000), compared to the expected mean predicted by the model (0.78432), which may indicate higher predictive accuracy of the model in high-risk groups.

From these results, it can be seen that the predictive performance of the model improves with the increase of the tree. Variable importance metrics can help determine which factors have a significant impact on customer churn, potentially guiding business decisions and resource allocation. At the same time, scoring rankings and distribution can be used to identify high-risk customer groups for appropriate customer retention strategies.

According to the mean square error (ASE) on the training set:

Boosting model: ASE is 0.14978, showing the best performance of all models.

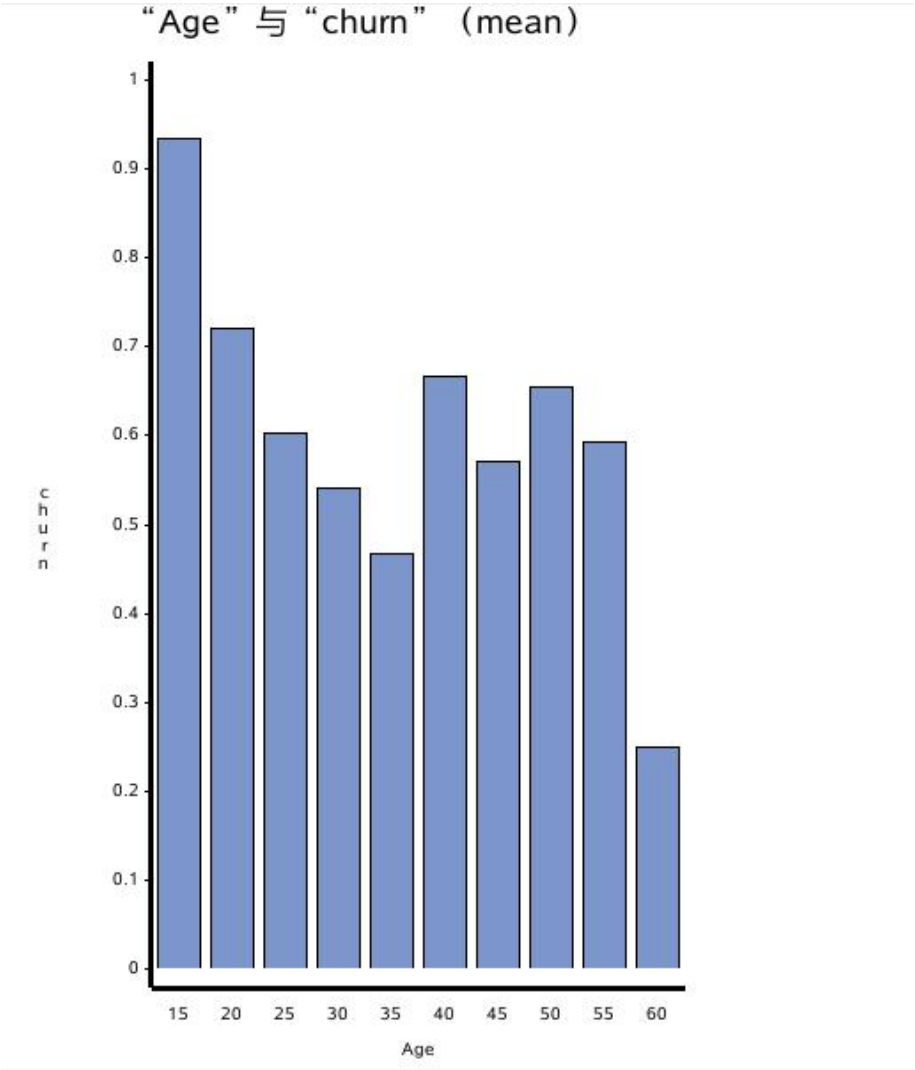
Decision tree model: ASE is 0.19052, with slightly lower performance than Boosting.

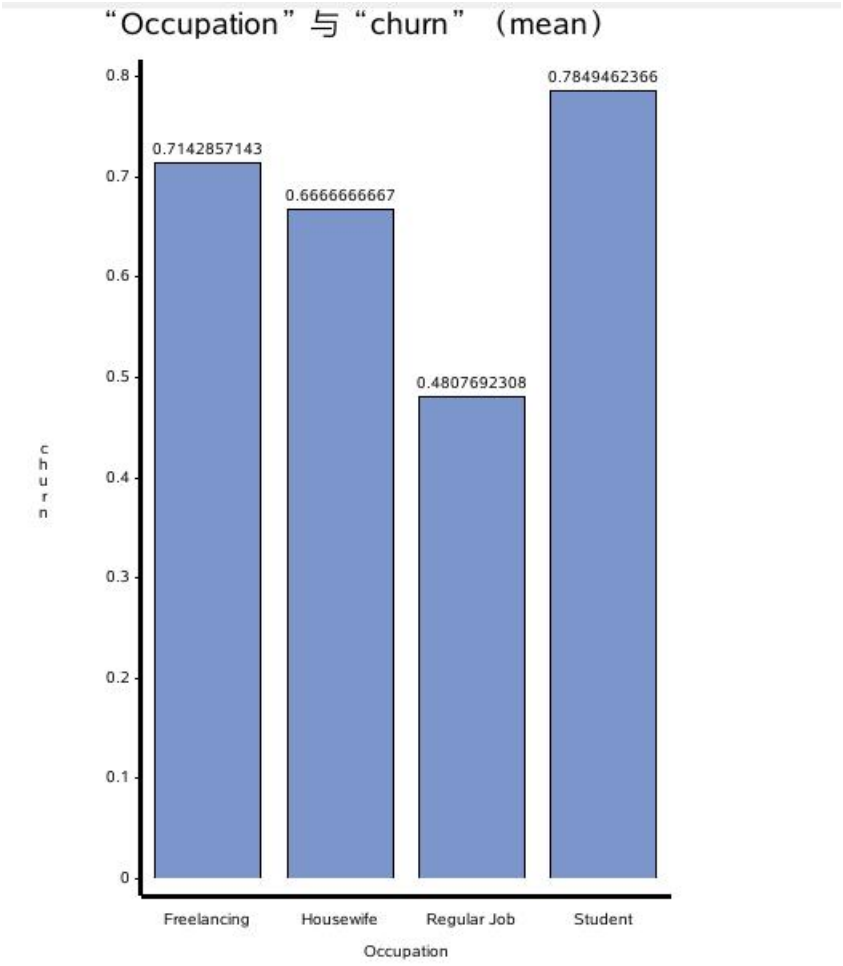
Random Forest model: ASE of 0.20511, the worst performance of the three models.

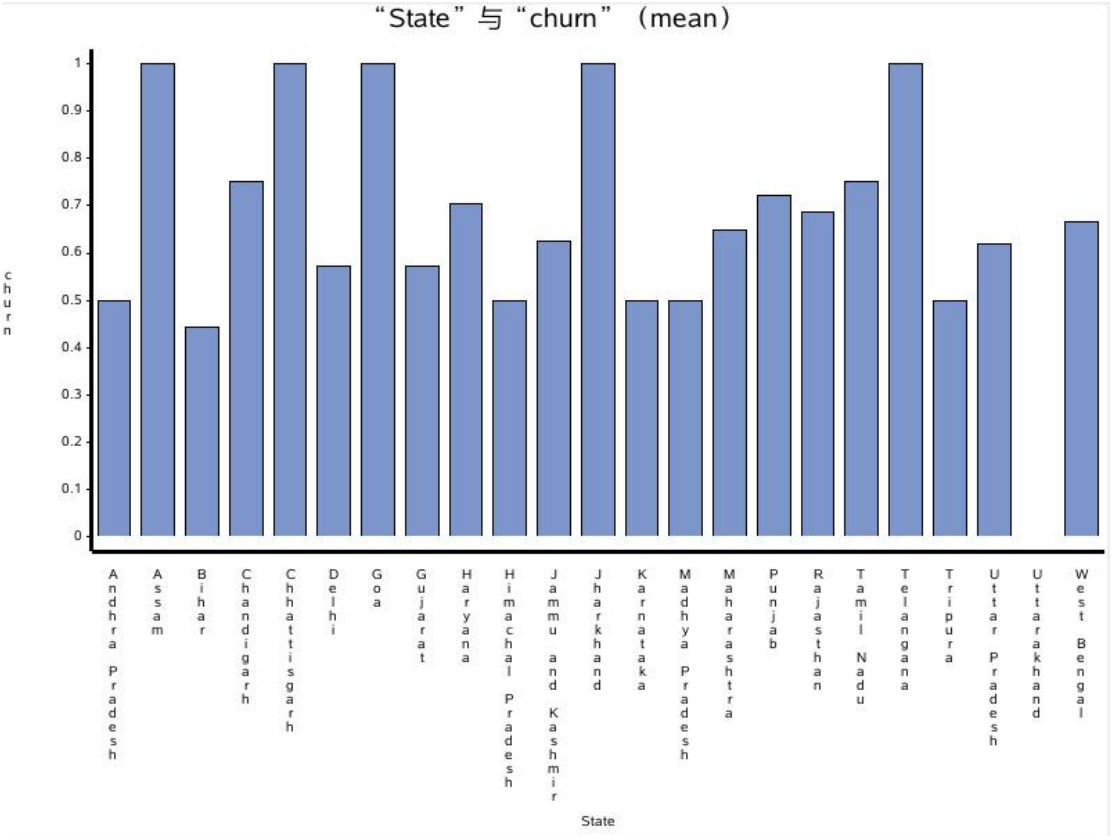
MAX shows that the maximum deviation between the predicted results and the actual values of Boosting model and Random forest model is small, indicating that these two models are more accurate in predicting extreme values.

6. Analysis and strategy

Age,occupation and state are the three factors with the highest correlation to churn.




“Occupation”与“churn”（mean）
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
churn
0.7142857143
0.6666666667
0.4807692308
0.7849462366
Freelancing
Housewife
Regular Job
Student
Occupation


“State”与“churn”（mean）
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
churn
Andhra Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu and Kashmir
Jharkhand
Karnataka
Madhya Pradesh
Maharashtra
Punjab
Rajasthan
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
State

As can be seen from the polygraph：

Young people aged around 15 years old have the highest turnover, and students have the highest turnover, both of which are younger people with low economic income.
Assam,Chhattisgarh,Goa,Jharkhand and Telangana are the top five states with the highest loss

The highest churn among teenagers around the age of 15 May mean that users in this age group are less loyal to the platform service or product. This may have something to do with teenagers' rapidly changing interests and preferences, or their spending power.
If students have the highest churn, it may be related to the fact that the specific needs of the student population are not being met. Students may be more price-sensitive, or they may be more inclined to pursue the latest trends and products.
The reasons for the high turnover in the first five states may be related to local characteristics, and local characteristics can be deeply investigated to analyze the reasons for customer loss

Strategy:
1.Add youth-related products and services, such as the latest electronics, fashion accessories, pop culture merchandise, etc.
2.Analyze the purchasing characteristics of the top five states with the highest turnover, and develop marketing strategies. It is recommended to conduct market research among these states to investigate the reasons for customer turnover
3.Customized promotions: Create special discounts and promotions for the student population and regions in the first five states, such as seasonal back-to-school offers, holiday sales, etc.
4.Social media marketing: Teens and students are often active on social media, and leveraging social media marketing can increase their engagement and brand loyalty.
Offer educational benefits: Give students specific educational benefits, for example, offer discounts on textbooks, educational software, online courses.
4. Parental involvement: Since teenagers around the age of 15 May not be fully independent, parental involvement has a greater impact on their consumption behavior. Therefore, platforms can consider introducing parental controls or family account plans to increase parental engagement.

5.