Will Customers End Credit Card Service?

Xinxian Li 005643690, Xinwei Hu 105641676,

Xiuqi Li 605638474, Zhiting Chen 505433562

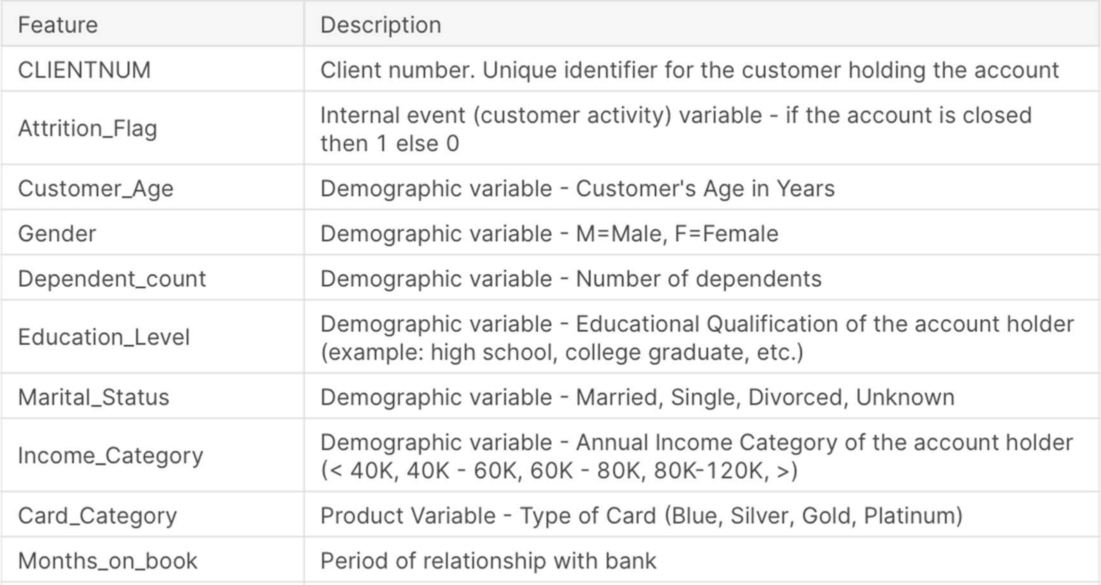
Abstract

In this project, we try to predict whether a customer will end a credit card service based on 8 features, including age, gender, education, income situation and so on. We use Naïve Bayes classification algorithm and give a rather robust result at the end. Firstly, we split the dataset into two parts. Then, apply the algorithm calculated in the training set to the validation set. Finally, evaluate the model by confusion matrix. Overall, the accuracy of the prediction of Naïve Bayes model is about 0.6858.

1. Data

1.1 Data Description

The original data includes 5339 samples with 8 variables. Regarding the variables, we select 8 variables, including customer age, dependent count, educational level, income level and so on. The definition of the variables is listed below in the Table1.



Feature	Description
CLIENTNUM	Client number. Unique identifier for the customer holding the account
Attrition_Flag	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Demographic variable - Customer's Age in Years
Gender	Demographic variable - M=Male, F=Female
Dependent_count	Demographic variable - Number of dependents
Education_Level	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Demographic variable - Married, Single, Divorced, Unknown
Income_Category	Demographic variable - Annual Income Category of the account holder (< 40K, 40K - 60K, 60K - 80K, 80K-120K, >)
Card_Category	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Period of relationship with bank

Table 1: description of variables

The dependent variable is *Attrition_Flag*, which states if the account is closed. We divide our data into two groups: the first group is Existing Customer and the other group is Non-Existing Customer. We could see their distribution of attrition flag in Figure1.

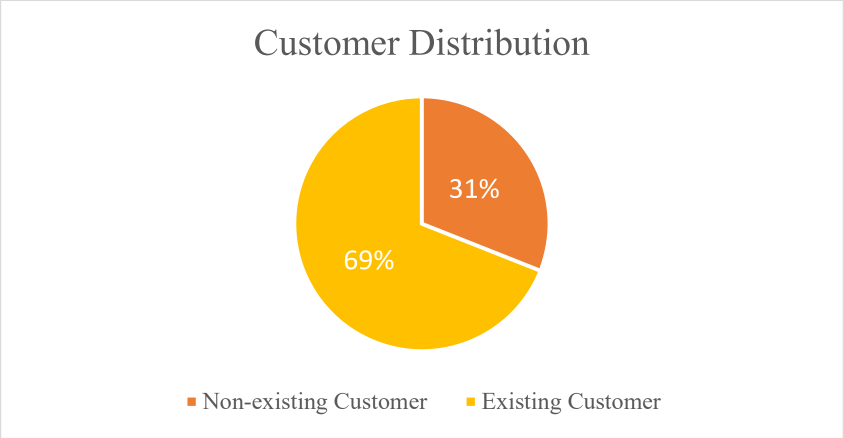


Figure 1: The Distribution of Customer Type (attrition flag)

In the figure 2, we describe the education level of customers in the sample. We found that among the sample, most of the customers have the college and graduate level of education.

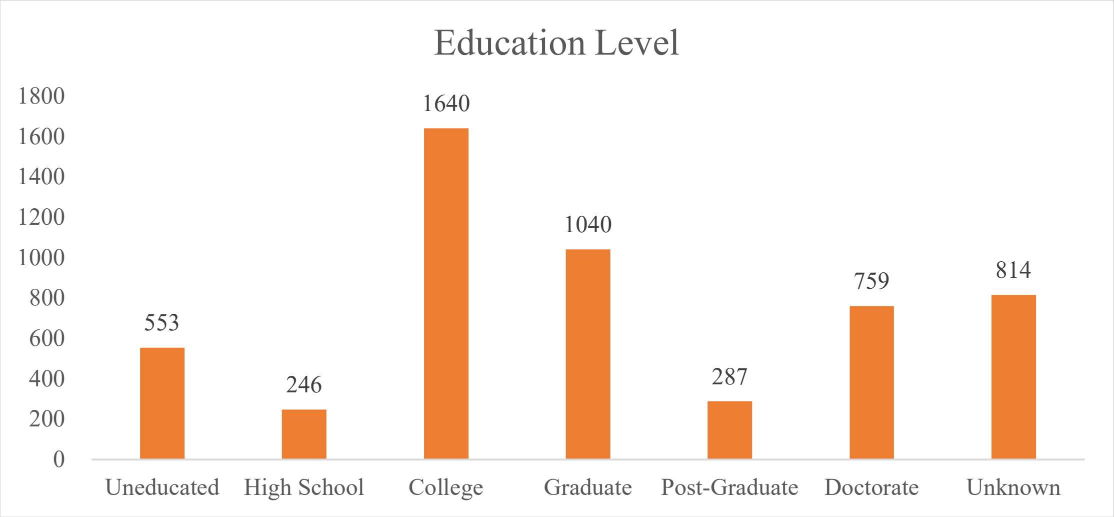


Figure 2: The Distribution of Customer Education Level

For the age variable, according to different age levels, we divide our data samples into three parts, including “25-35” part, “36-50” part and “51-75” part. From Figure 3 we can know that most of the customers are in the “36-50” part, accounting for 64% of the samples.

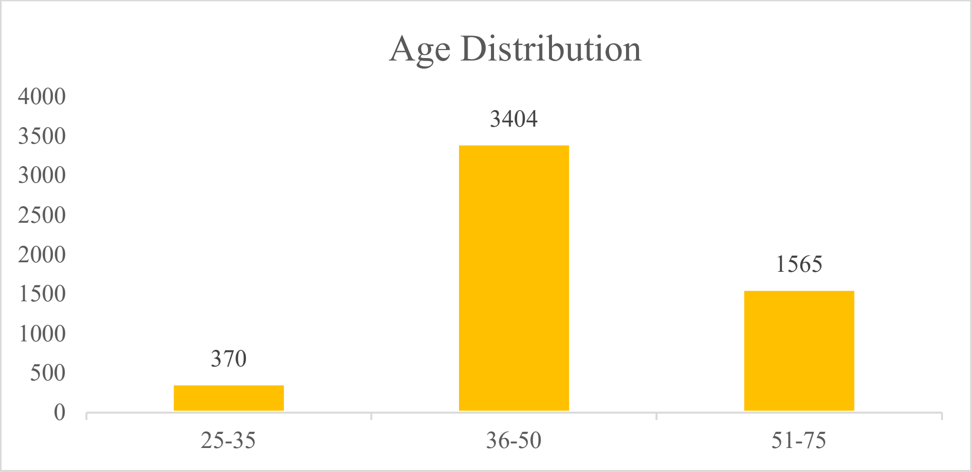


Figure 3: The Distribution of Customer Age Level

In order to better understand the relationship between customers and banks, we divide the period of relationship with bank into three categories. We can see from the figure 4 that most of the customers have a 36-47 months relationship with banks, accounting for 57% of the samples.

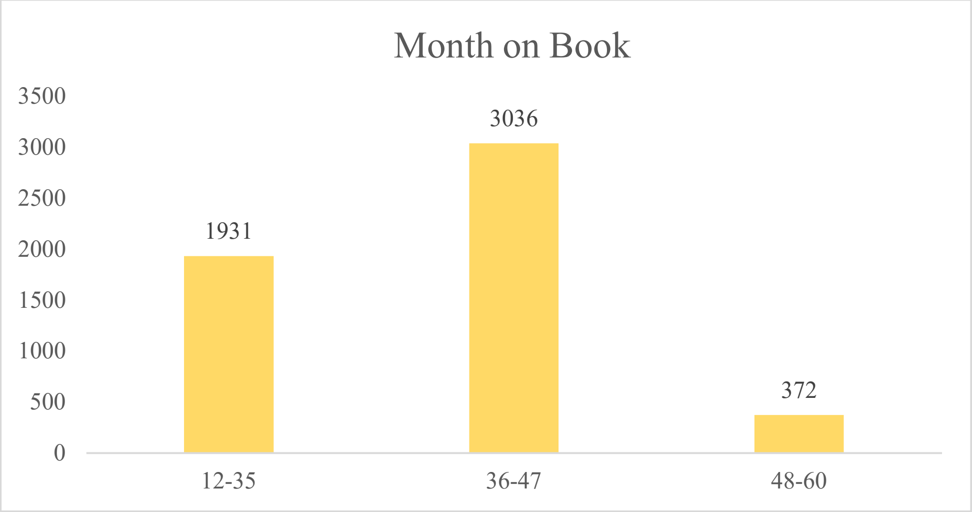


Figure 4: The Distribution of Customers' Relation Period with Bank

And as shown in the Table 1, the income has been divided into four parts which include "<40K", "40K-60K", "60K-80K" and "80K-120K". We can know from Figure 5 that most of our customers' income are under 40K.

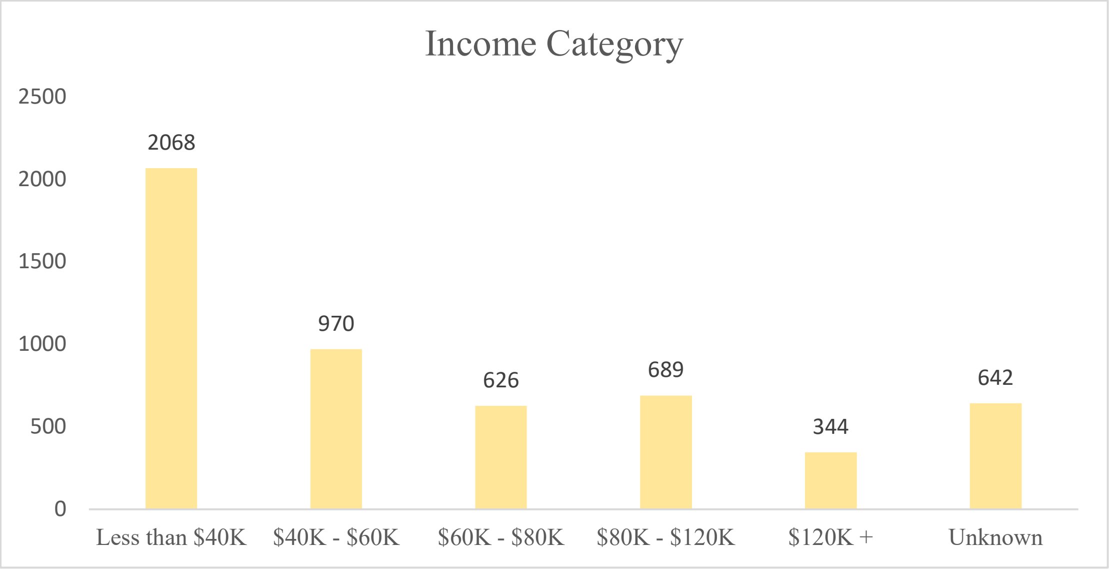


Figure 5: The Distribution of Customers' Income Category

1.2 Data process

Before building up Naïve Bayes model and testing data, we firstly process data with 3 steps.

1) Eliminate the missing data.
2) Classify continuous data by dividing the age variable into 3 types, month on book age variable into 3 types, income variable into 6 types and education period into 7 types.
3) Split the original dataset into training and validation sets.

And other process of data would be shown in the following content.

2. Model: Naïve Bayes

The method used in this project is Naïve Bayes model. Abstractly, Naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector $X = (x_1, x_2, \ldots, x_n)$ where n is 14 specifically representing some n features (independent variables), it assigns to this instance probabilities $p(C_k | x_1, \ldots, x_n)$ for each of K possible outcomes for classes C_k. The main idea of Naïve Bayes is to calculate probability with conditional probabilities, which can be shown as

$$p(C_k|X) = \frac{p(X|C_k)p(C_k)}{p(X)}.$$

As a result, the model process consists of two parts. First one is to calculate the priori probabilities and conditional probabilities with training dataset. Second one is to make the prediction with validation dataset.

2.1 Train Model with Training Data

With the training dataset, we could run the Naïve Bayes model to get the priori probability and to directly calculate the conditional probabilities from the data.

2.2 Prediction and Model Evaluation

In this part, we can make the prediction according to classification probability calculated with the previous conclusion and test the performance of the model.

Priori Probability	
Non-Existing Customer	0.30003
Existing Customer	0.69997

Table 2

Table 2 shows that the priori probability of closing the account is approximately 0.30 and the probability of continuing the account is approximately 0.69.

The conditional probability of the different features is also concluded in our result. We use the different conditions as the classification basis for forward steps.

Conditional Probability on Gender		
Gender	Female	Male
Non-Existing Customer	0.55879	0.44121
Existing Customer	0.60883	0.39117

Table 3

Table 3 shows the conditional probability on gender. It means for a person whose decides not to continue the credit card service, there is 55.8% probability that she is a female, and 44.1% probability that he is a male. Similarly, for a person who is an existing customer for the credit card service, there is 60.9% probability that she is a female, and 39% probability that he is a male.

Conditional Probability on Income						
Income Category	Less than $40K	$40-$60K	$60K-$80K	$80K-$120K	$120K+	Unknown
Non-Existing Customer	0.36733	0.16337	0.12383	0.15297	0.08117	0.11134
Existing Customer	0.38492	0.19447	0.11775	0.12444	0.05665	0.12177

Table 4

Table 4 shows the conditional probability on Income. It means for a person who is a non-existing customer, the probability for his income less than $40K is 36%, 16% for $40K - $60K, 12% for $60K - $80K, 15% for $80K - $120K, 8% for $120K+ respectively. Similarly, for a person who is an existing customer, the probability for his income less than $40K is 38%, 19% for $40K - $60K, 12% for $60K - $80K, 12% for $80K - $120K, 6% for $120K+ respectively.

Except for the two tables above, there are also conditional probabilities based on other 6 features which can be found in R results.

2.2.1 Robustness with training data

Prediction	Non-Existing Customer	Existing Customer
Non-Existing Customer	15	9
Existing Customer	946	2233

Table 5: Confusion Matrix on training sample

According to the performance result, Confusion Matrix, we can calculate the several ratios to test the robustness of the model with training data. The results are shown below.

$$Accuracy = \frac{15+22}{15+2233+946+} = 0.70184$$

$$Precision\_\text{non-existing} = \frac{15}{15+9} = 0.625$$

$$Precision\text{\_existing} = \frac{2233}{2233+946} = 0.70242$$

$$Recall\ rate\text{\_non-existing} = \frac{15}{15+} = 0.01561$$

$$\text{Recall rate\_existing} = \frac{2233}{2233+} = 0.99599$$

It is safe to say that the model performs quite good with 0.70184 accuracy, and high precision. However, the recall rate of non-existing customers has a low number of 0.01561. This may imply the exist of type 2 error. The high false-negative number also imply the low risk control level. There are some customers intend to cancel the credit card can't be detected from the model. Comparatively, the model fits better for customers who have intention to continue the card service due to the larger sample size.

2.2.2 Robustness with Testing Data

Prediction	Non-Existing Customer	Existing Customer
Non-Existing Customer	7	12
Existing Customer	659	1458

Table 6: Confusion Matrix on testing sample

Based on Confusion Matrix, we can calculate the several ratios to test the robustness of the model with testing data. The results are shown below.

$$Accuracy = \frac{7+145}{7+1458+659+12} = 0.68586$$

$$Precision\text{\_non-existing} = \frac{7}{7+12} = 0.36842$$

$$Precision\text{\_existing} = \frac{1458}{1458+659} = 0.68871$$

$$Recall\ rate\text{\_non-existing} = \frac{7}{7+659} = 0.01051$$

$$Recall\ rate\text{\_existing} = \frac{1458}{1458+12} = 0.99184$$

Similar to the training sample, the model to predict the classification for salary is relatively solid with 0.68586 accuracy.

3 Conclusion and future work

Using the Naïve Bayes method, we calculated the algorithm of conditional probability and successfully predict whether based on 8 features. This method has significant important results with the high accuracy, and precision. Furthermore, the little difference of the robustness measurements between validation and training samples, showing that there's no overfitting problem in this case.

We could improve our work by trying Gaussian Naïve Bayes, which including both priori and posterior probabilities. Another method is finding a database with more customers who cancel the card service. In this way, the precision rate can be Also, we could try the principal components analysis to capture more variables so as to improve the prediction performance.

References

[1] https://www.kaggle.com/sureshmecad/credit-card-churn-evalml-autoeda

[2] https://www.kaggle.com/prashant111/naive-bayes-classifier-tutorial

[3] https://www.kaggle.com/sakshigoyal7/credit-card-customers