Team 23: Final Project Report
Rashmi Gehi, Sarah Guo, Tushar Pandey, Ahmet Tas, Alexandra Wang

Introduction

Banks have several different forms of operations, with which different customers engage at different levels. Most customers at many banks are labeled as "liability customers" who primarily interact with the banks through deposits. The number of those who take out loans also called 'borrowers,' is smaller, and the overlap of both is even smaller. To increase profitability, banks must increase the borrowers' customer base through marketing campaigns encouraging more people to take out loans. But how will banks identify these customers, and how can they increase the number of customers who belong to this group? With our model, we seek to understand a bank's customer base through clustering based on their current bank activity. We expect our model to create valuable insights regarding customer segmentation and targeted marketing. By leveraging this information, bank managers can capitalize on their customers who are more likely to become borrowers, and increase marketing efforts towards them. On the other hand, as students entering the workforce, we must understand how banks may implement marketing campaigns toward us.

Data

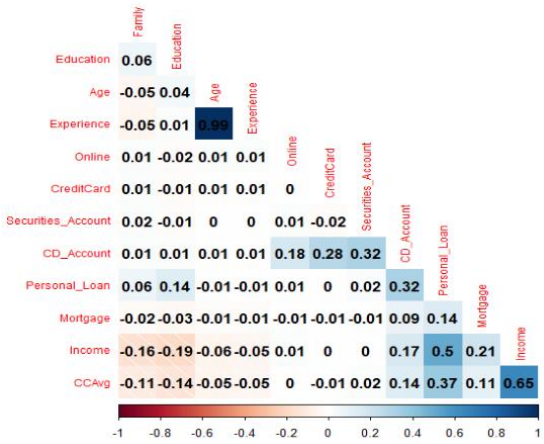
Our data comes from Kaggle and contains 5000 rows and 14 columns. The numerical variables include ID, age, experience, income, zip code, family size, mortgage value, and average monthly credit card spending. The categorical variables are education level, personal loan, securities account, CD account, online, and credit card. There are no missing values in this dataset. We adapted this data in Excel to create a County variable and a Region variable. This is because the number 467 unique zip codes would be difficult to model. Now, we have 5 region variables, all of which are in California, and it will be easier to categorize them by location. We did this by using California census data from 2020.

Exploratory Data Analysis

To begin our EDA, we wanted to get a snapshot of how each of our variables relates to each other and see if they are strongly correlated.

**Definitions and detailed plots of individual variables are present in the appendix*

Figure 1: Correlation between variables



The plot shows us that there are very few variables that share a strong absolute correlation. Age and Experience have a 99% correlation, leading us to drop Experience in our dataset. Other notable correlations* include Income with CCAvg, which is the average spending on credit cards in thousands.

Principal Component Analysis

To better understand the data, we decided to go for dimension reduction with Principal Component Analysis on all the numerical variables and obtained 7 different principal components. When we looked at the proportion of variances of these principal components, we found out that the first three components were enough to explain 70% of our dataset. These are:

- Older people with high experience
- Young people with low income
- Families who have a mortgage

K-means Clustering

Additionally, we performed k-means clustering on the numerical variables to segregate the bank customers into distinct groups. From the elbow plot, we could identify that 3 clusters were the optimal number. We also analyzed the average metrics of the three different clusters to identify their unique characteristics. For example, Cluster 1, which contains high-income members with a mortgage, is the one that has the highest percentage of customers that have accepted loan offers.

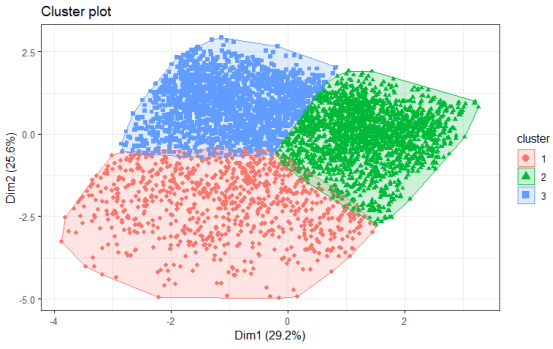


Figure 2: Distribution of clusters based on principal component 1 and principal component 2

Table 1: Average metrics of individual clusters

Group	Group Name	% Customers	%Accepted Loans	Avg Age	Avg Experience	Avg Income	Avg Family	Avg CCAvg	Avg Mortgage	Avg Education
1	Loyalists	17%	**37%**	44	19	146	1.9	5	116	1.5
2	Samplers	43%	4%	56	30	58	2.4	1	44	2.0
3	Potentials	40%	4%	35	10	60	2.6	1	45	2.0

**Definitions and detailed plots of individual variables are present in the appendix*

Logistic Regression

It is important for banks to be able to predict if a customer would accept an offer for a loan or not to maximize revenue and minimize loss. We decided to design a logistic regression model which could predict with high accuracy if a customer would accept a loan offer and to understand the customer-level variables that impacted the probability of a customer accepting a loan offer. Below are the two models we designed, where we used *Personal_Loan* as our dependent variable but different sets of independent variables. In both models, we used 80% of the data for training and 20% of the data for testing the models.

Model 1: Logistic Regression with all independent variables.

$$Personal\ Loan \sim \beta_0 + \beta_1\, Age + \beta_2\, \text{Income} + \beta_3\, Family + \beta_4\, CCAvg + \beta_5\, Mortagage + \beta_6\, Securities\ Account + \beta_7\, CD\ Account + \beta_8\, Online + \beta_9\, Credit\ Card + \beta_{10}\, Education\ 1 + \beta_{11}\, Education\ 2 + \beta_{12}\, Region\ Bay + \beta_{13}\, Region\ Central + \beta_{14}\, Region\ LA + \beta_{15}\, Region\ Southern_{\text{l}}$$

This model had an out-of-sample accuracy of 0.96 and an out-of-sample R-squared of 0.53. As the accuracy is very high, we analyzed the confusion matrix of the training and test set to check whether our model was over-fitting. We notice that the percentage distribution of true positives and true negatives are comparable in both the training and test set, suggesting that we are not over-fitting.

Train Data		
	Predicted 0	**Predicted 1**
Observed 0	3129	35
Observed 1	**114**	222

Test Data		
	Predicted 0	**Predicted 1**
Observed 0	1340	16
Observed 1	**44**	100

Table 2: Confusion Matrix of Model 1 on train and test data

A potential risk of Model 1 is that the false negative percentage is high, which means that the model predicts that a person will not accept the loan offer when they actually accept it. This is a huge loss of opportunity for the marketing teams at the bank as they would not target these customers. To minimize the revenue losses caused due to false negatives, we decided to tweak our model by dropping insignificant variables like 'Region' from Model 1 and introducing interaction terms.

Model 2: Logistic Regression with custom interaction terms

For Model 2, we decided to address the fact of multicollinearity in our independent variables and made use of the Lasso Regression model to identify the right set of variables to introduce our model. Lasso Regression penalizes the less important features in data and eliminates them. It

**Definitions and detailed plots of individual variables are present in the appendix*

helps reduce the *variance* in the test dataset. Using Lasso, we introduced custom interaction terms like *Income with Family*, *Income with Education Level 1*, *Online with Credit Card*, and others to predict if a customer would accept a loan. Model 2 has out-of-sample accuracy of 0.97 and an out-of-sample R-Squared of 0.73, a significant increase from Model 1.

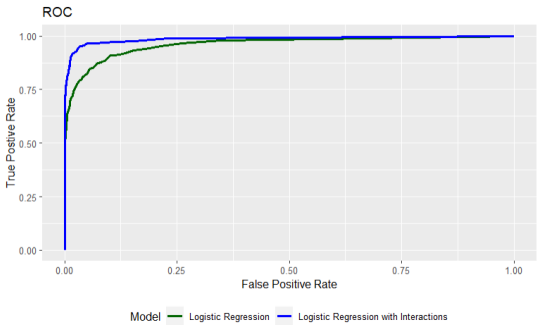


Figure 4: Comparison between the area under a curve of ROC, which suggests that Model 2 with interaction terms is better than Model 1.

The confusion matrix between the train and test set of Model 2 improves upon the issue addressed earlier, and we notice a significant improvement in the false negative rate.

Train Data		
	Predicted 0	**Predicted 1**
Observed 0	3139	25
Observed 1	55	281

Test Data		
	Predicted 0	**Predicted 1**
Observed 0	1342	14
Observed 1	21	123

Table 3: Confusion Matrix of Model 2 on train and test data

Finally, we evaluated and quantified the marginal impact of all the statistically significant variables (p_value < 0.05) that are used to predict whether a customer will accept the loan offer or not.

Variable	Coefficient	Odds Ratio	Probability
Income	5.19	179.58	0.99
Income:Family	1.79	5.99	0.86
CCAvg	1.53	4.6	0.82
Education_1	1.01	2.76	0.73
CD_Account	1	2.73	0.73

Table 4: Interpretation of the top 5 significant variables of Model 2*

Results and Managerial Takeaways

Cluster Analysis and Marketing Strategies:

Loyalists: These are the consumers loyal to our brand and have the highest acceptance rate of our services. This is the best segment to perform cross-selling. Sending them personalized offers on credit cards, Securities, and Certificates of Deposit accounts will result in better conversion rates. Cross-selling via email, push notifications, and calls would help deliver the offers.

**Definitions and detailed plots of individual variables are present in the appendix*

Samplers: This group identifies an older population, which may not be up for signing up for loans. Targeting this group with good returns on securities accounts can lead to better conversion rates. But we do not recommend more than 10% of the marketing budget for this group as the CLV of this segment will not be much.

Potentials: This segment is middle-aged people with a high education level and minimal mortgage value. The income level lies in the lower middle class. These potential customers can be consumers of all products, house loans, credit cards, and securities accounts. We would spend aggressively to market our products to this segment and create the next big new consumers of our services.

Predictive Analytics:

New User Prediction: Once the new user arrives on the platform and we have the personal details of each user and the offer which he opts for. Using the cluster analysis, we can put him in the predicted cluster and implement a cluster marketing strategy. Also, we can predict his chances of accepting the most valuable service, which is the loan. We can predict with 97% accuracy that a consumer will accept the loan offered. This can help us use the marketing budget to reach the highest potential customers and reach higher conversion rates.

Limitations

With our analysis, there are several limitations we recognized along with potential solutions.

- As the current dataset is skewed, with 90% of the people not accepting the loan, our analysis might be biased toward the extreme. One possible way to address this is to test our model on another normally distributed dataset and see if there are any adjustments we need to make.
- While our model could be a general indicator of people's behaviors in accepting loans, we cannot identify exceptional cases when someone with a low income or no education, or low CCavg might accept a loan offer. This is usually when human power comes in to incent loan acceptance.
- We only looked at the data from California. Therefore, it might not be practical for other regions across the country.

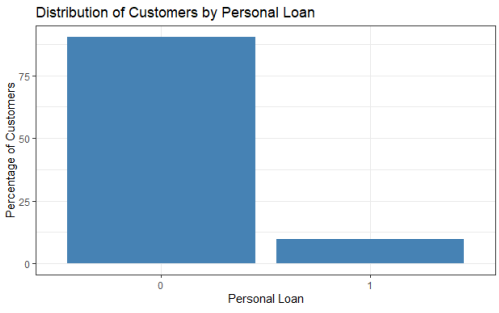
**Definitions and detailed plots of individual variables are present in the appendix*

APPENDIX
Data Description

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: The family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- Mortgage: Value of house mortgage, if any. (In thousand dollars)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities_Account: Does the customer have securities account with the bank?
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Do customers use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)
- County: Name of the California county
- Region: Name of California region

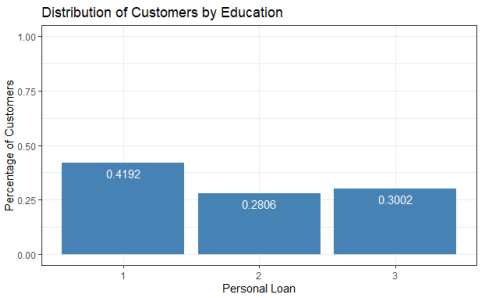
Figure 2: Distribution of Customers by Personal Loan

This bar chart depicts the heavy skew this bank is experiencing in the percentage of customers taking out personal loans versus those not. About 90% of customers do not take out personal loans.



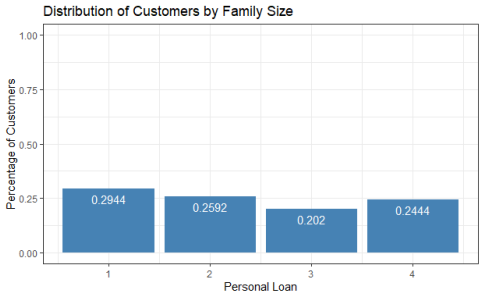
**Definitions and detailed plots of individual variables are present in the appendix*

Figure 3: Distribution of Customers by Education



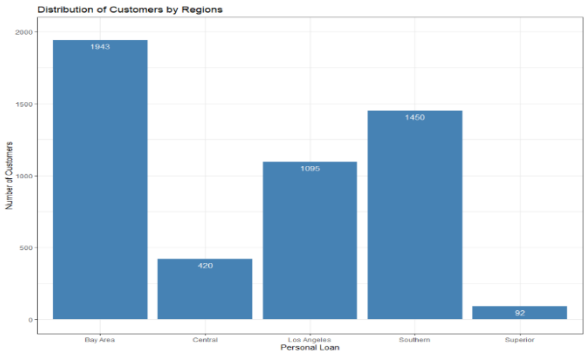
This next graph shows the percentage of customers with different levels of education. 42% of the bank's customers have an undergraduate education, 28% have a graduate education, and 36% have an advanced or professional degree.

Figure 4: Distribution of Customers by Family Size



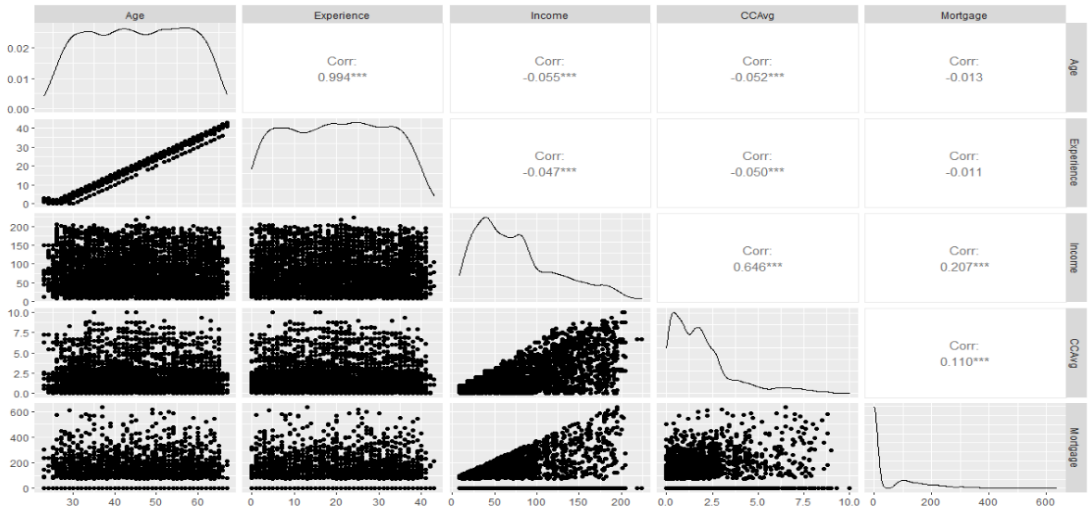
The distribution of customers by family size is more uniformly distributed. The most common size is 1, with almost 30%, while the least common is 3, with 20%. Family sizes 2 and 4 comprise 26% and 24%, respectively.

Figure 5: Distribution of Customers by Regions



Finally, we analyzed the number of customers in each California region. As seen in the figure, the most populous region was the Bay Area with 1943, followed by the Southern region with 1450, Los Angeles with 1095, Central with 420, and Superior with 92.

Figure: Detailed Correlation between numerical variables



**Definitions and detailed plots of individual variables are present in the appendix*

Figure: Age and Family Distribution

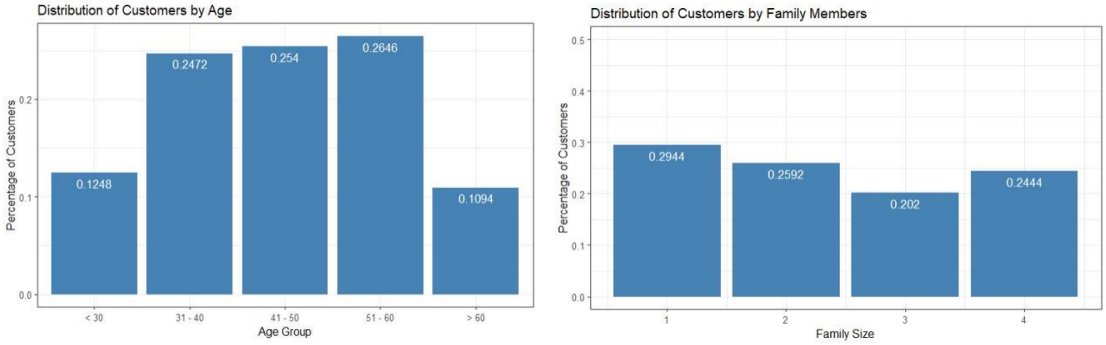
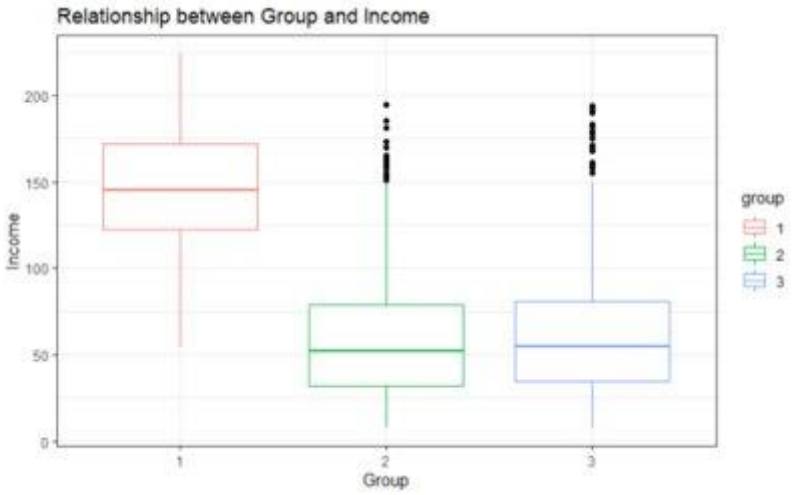
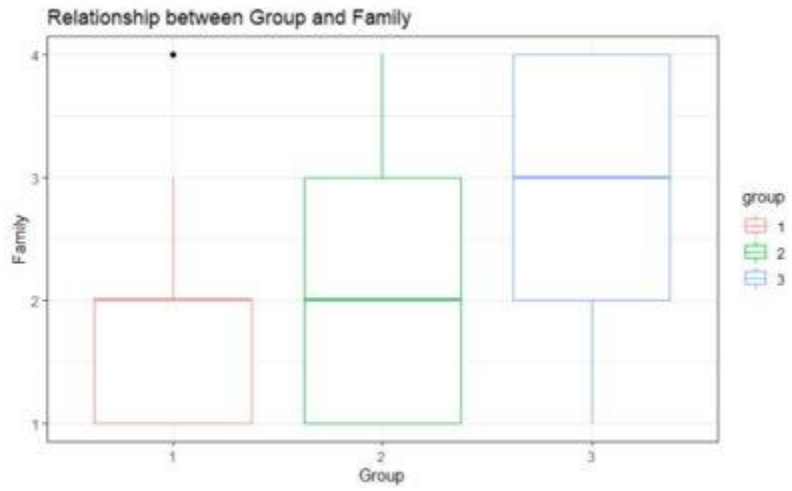
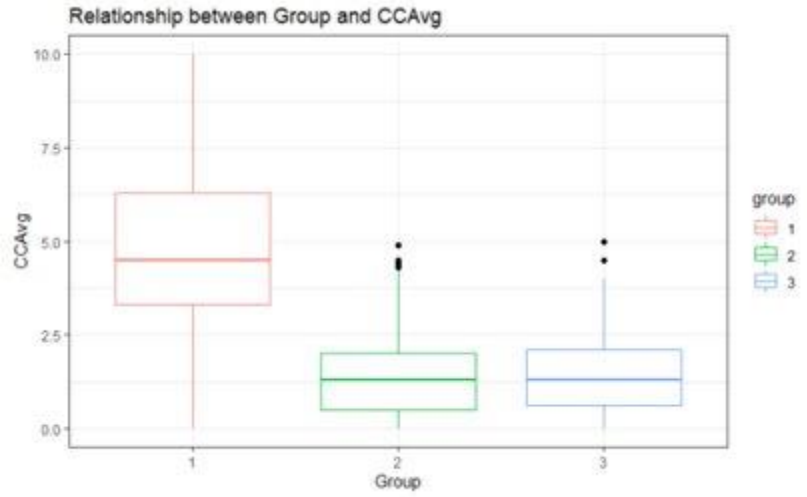


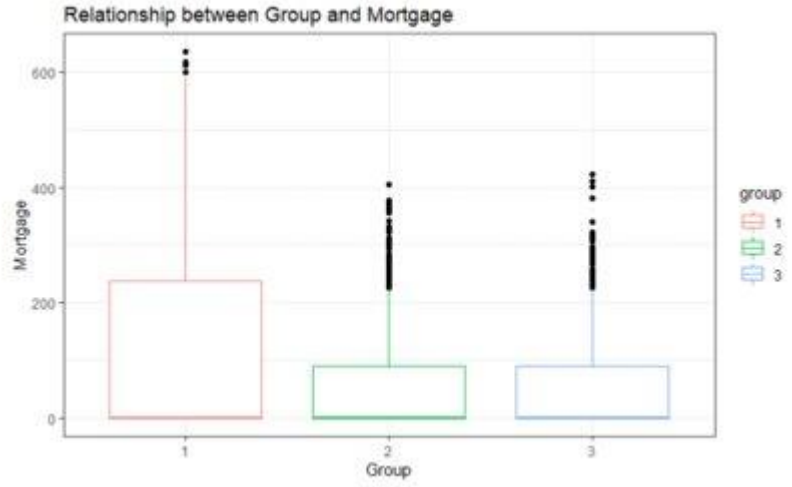
Figure: Box Plots for Cluster Analysis

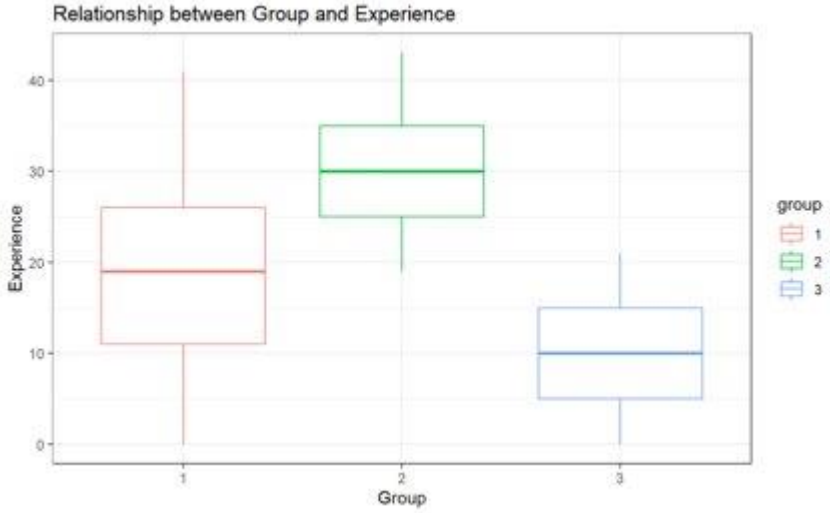




**Definitions and detailed plots of individual variables are present in the appendix*







**Definitions and detailed plots of individual variables are present in the appendix*

Figure: Distribution of Loan Acceptance by Cluster

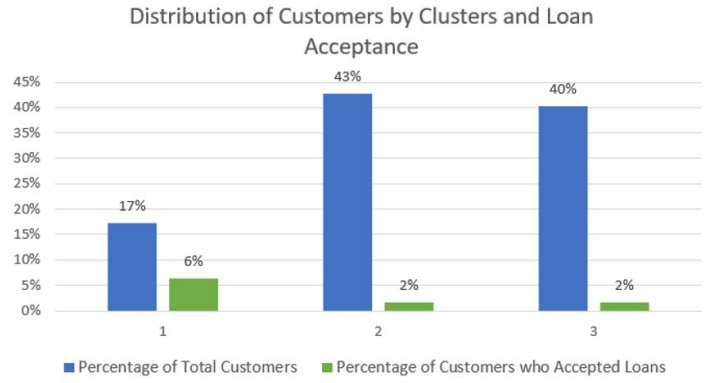


Table : Interpretation of all significant variables of Model 2

Variable	Coefficient	Odds Ratio	Probability
Income	5.19	179.58	0.99
CD_Account	1.00	2.73	0.73
Family	-0.50	0.61	0.38
Education_1	1.01	2.76	0.73
CCAvg	1.53	4.60	0.82
Online	-0.57	0.57	0.36
CreditCard	-0.88	0.41	0.29
Income:Family	1.79	5.99	0.86
Income:Education_1	-2.92	0.05	0.05
Family:CCAvg	0.52	1.69	0.63
Education_1:CCAvg	-0.88	0.42	0.29
CCAvg:Education_2	0.51	1.67	0.63
Online:CreditCard	-0.67	0.51	0.34

**Definitions and detailed plots of individual variables are present in the appendix*