AllLife Bank

Personal Loan Campaign Modelling

June 2021

Brian Hall – DSBA – Project 4

Problem

AllLife Bank would like to rapidly increase the amount of interest on personal loans by converting existing liability customers (depositors) to personal load customers while maintaining them as depositors. In a previous campaign the bank showed a 9% success rate for converting depositors to personal loan customers.

Objective

Utilizing data obtained from the previous successful campaign we will build machine learning models with the potential to predict what types of customers Marketing could be targeted in a new campaign.

Recommendations (Next Campaign)

Generally, customers who use more of the banks services and have higher averages for those services were more likely to have accepted a personal loan. This makes some sense as most of these customers have higher debt with AllLife Bank and may be more willing to extend that debt further. These same customers also tend to have higher incomes.

- For the next marketing campaign identify customers with higher education, income and larger family sizes.
- External marketing campaigns to attract those with higher education, incomes and family sizes could bring in more customers who would be liability customers as well as personal loan customers.
- Target incomes above 100k
- Target CCAvg > 3.5k

- Target customers with CD_Accounts
- Target Mortgages above 200K
- Target those who do not use Online banking facilities
- Target those who do not have Securities Account
- Examine customers in Santa Clara County more closely and possibly market heavier to them.
- Marketing other services, such as CD_Accoutns, could increase willingness to accept a personal loan
- Extend the types of data used to predict personal loan acceptance - such as spending patterns, savings & balances for other services

Data Overview

The data is assumed to be a random subset of data collected from previous year marketing campaign on conversion rates of liability customers to personal loan. Data has been provided via CSV (load_modeling.csv – 204k)

Variable	Description
ID	**Customer ID**
Age	Customer's age in completed years
Experience	Number of years of professional experience
Income	Annual Income of the customer (in thousand dollars)
ZIPCode	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)?

The Data Contains:

- 5000 rows and 13 columns
- Column Keys match data description provided
- 12 variables of int64
- 1 variable of type float64
- No missing data
- No Duplicate rows

Data Treatment (pre-model):

- Experience contained 52 negative values as was imputed with the absolute of the value
- Family, Education, Personal_Loan, Securities_Account, CD_Account, Online, CredicCards were converted to type – category
- ZIPCode was binned by 'County' with counts <= 50 going into 'Other_County' and converted to type - category
- Other variables were binned for specific models, details located in notebook.

EDA Overview

5000 observations

Variable	Mean	Range
Age	45.3	23 – 67
Experience	20.1	0 – 43
Income	73.3k	8k – 224k
CCAvg	1.9k	0 – 10k
Mortgage	56.5k	0 – 635k

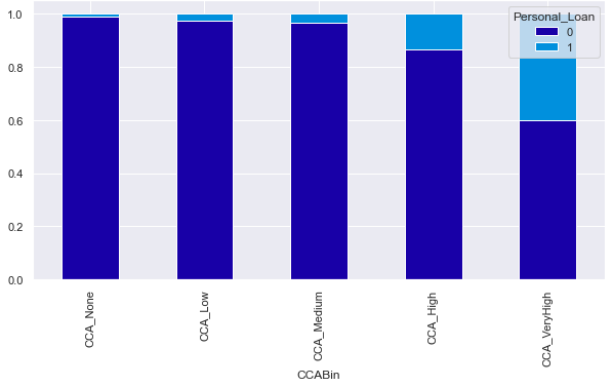
Variable	Unique	Top
County	19	LA County
Family	4	1
Education	3	1
Personal_Loan	2	0 – False
Securities_Account	2	0 – False
CD_Account	2	0 – False
Online	2	1 – True
CreditCard	2	0 - False

Key Observations

- **More than Half of those with Income above 175k accepted an offer for the personal loan**
- **The majority of personal loans are to incomes above 100k**
- **Nearly half of those with income between 126 and 175 accepted an offer for the personal loan**
- Customers with mortgages higher then 200k show a 30% acceptance rate
- Mortgages between 76k and 200k are around 5% acceptance
- **Only ~6% of customers have a CD_Account, of those who do nearly half (46%) have accepted a personal loan.**
- **The higher the monthly average for credit cards the more likely someone was to accept a personal loan with the majority of loan acceptance having an average above 3.5**
- **40% of customers with Very High CCAvg accepted personal loans**
- No one with an income below 75k accepted a personal loan
- All CCAvg >= 8k have incomes higher the 143k
- Most Customers 36.2% have a low monthly CCAvg
- Santa Clara county has the highest acceptance rate, 13%, but has about half the customers of LA County
- 69% of customers do not have a mortgage

Income, Credit Card Average & Personal Loan Acceptance

Credit Card Average & Personal Loan Acceptance



Model Performance

By developing Machine Learning models that are based on historical data we can predict within a range of certainty who is more likely to accept a personal loan and what factors play a key role.

22 models trained with F1 Score as the metric

During EDA and feature engineering the decision was made to make 2 versions of the sample set for training.

1. Maintain continuous variables
2. Binning continuous variables

- Both sets were trained using SKLearn - Logistic Regression (solvers - newton-cg & liblinear) - **4 models**
- Both sets were trained with StatsModel - Logistic Regression (reducing complexity with VIF & P Value) - **5 models**
- Best results from StatsModel (logit5) was tuned with various thresholds (ROC-AUC calculated & recall / precision plot) - **8 Models**
- Decision Tree training used Binned set with Hyper Parameter Tuning (GridSearchCV) & Cost Complexity Pruning - **5 models**

NOTE: Initial sample set could be considered imbalanced with only 9.6% of data representing target variable, no oversampling was performed

Conclusion

The best performing model was the **Decision Tree Model with Post Pruning** and could predict liability customers that would accept a personal loan **89.04%** of the time. The model is well balanced and should generalize well with new data.

- F1 on training set : 0.9132569558101473
- **F1 on test set : 0.8904109589041096**
- Precision on training set : 0.9789473684210527
- Precision on test set : 0.9420289855072463
- Recall on training set : 0.8558282208588958
- Recall on test set : 0.8441558441558441

	Model	Train F1	Test F1
0	LR_SKLearn Continuous Variables, Solver=newton-cg	0.690300	0.723800
1	LR_SKLearn Continuous Variables, Solver=liblinear	0.691600	0.726500
2	LR_SKLearn Binned Variables, Solver=liblinear	0.693500	0.669200
3	LR_StatsModel logit5 - Threshold=0.50	0.704000	0.730600
4	StatsModel logit5 - Threshold=0.63	0.542600	0.603400
5	LR_StatsModel logit5 - Threshold=0.45	0.720600	0.743800
6	Decision Tree Initial	0.993800	0.842800
7	Decision Tree - Max Depth=3	0.823100	0.824400
8	Decision Tree Hyperparameter Tuned	0.865700	0.854000
9	Decision Tree Post Pruned	0.913200	0.890400

Model Performance

Significant Variables from Logistic Regression

Income, Family, CCAvg, Education, Securities_Account, CD_Account, Online, CreditCard

Observations on coefficient change odds

1 unit change in feature if all other features remain constant:

- Income will increase the odds of someone accepting a personal loan by 5.66%
- CCAvg will increase the odds of someone accepting a personal loan by 11.33%
- Securities_Account will decrease the odds of someone accepting a personal loan by 67.7%
- Online will decrease the odds of someone accepting a personal loan by 64.5%

Decision Tree Gini Importance for features

- Education is the most important feature, with Income & Family being next.
- Most features other than County & Mortgage are showing some level of importance in this model. This would likely be a better generalized model when new data is presented.

EDA of missed predictions observations

No patterns stand out. Some features do initially stand out with higher or lower percentages for the misses, however when compared to the EDA from the original sample we find that the misses patterns closely match those for the original sample.