AllLife Bank

Loan Modeling & Predication

PG-DSBA Project 4
Eric Green
March 2020

Background

AllLife Bank has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of `converting its liability customers to personal loan customers` (while retaining them as depositors).

A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with `better target marketing to increase the success ratio` with a minimal budget

Objectives

- To predict whether a liability customer will buy a personal loan or not
- Which variables are most significant
- Which segment of customers should be targeted more

Data Summary

Data Dictionary (raw data)

1. ID: Customer ID
2. Age: Customer's age in completed years
3. Experience: #years of professional experience
4. Income: Annual income of the customer (in thousand dollars)
5. ZIP Code: Home Address ZIP code.
6. Family: the Family size of the customer
7. CCAvg: Avg. spending on credit cards per month (in thousand dollars)
8. Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
9. Mortgage: Value of house mortgage if any. (in thousand dollars)
10. Personal_Loan: Did this customer accept the personal loan offered in the last campaign?
11. Securities_Account: Does the customer have securities account with the bank?
12. CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
13. Online: Do customers use internet banking facilities?
14. CreditCard: Does the customer use a credit card issued by Bank?

Raw shape: 5000 rows x 14 columns

Data Description

	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIPCode	5000.0	93169.257000	1759.455086	90005.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396400	1.147663	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937938	1.747659	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0
Personal_Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0
Securities_Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD_Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0

Observations on raw data

1. Shape of the data is 5000 rows by 14 columns
2. Income, CCAvg and Mortgage need to be converted to thousands
3. Personal_Loan, Securities_Account, CD_Account, Online and CreditCard are binary classes
4. The raw data is free of null/missing values
5. All variables are numeric data types (some are continuous and some are categorical
6. The dependent class variable is Personal_Loan (0 or 1)
7. Experience has a value of -3, which is not a valid value (this needs scrubbing)
8. zip_code ranges 90005 to 966510 (California)
9. Major data preprocessing is not required for raw dataset

Data Preprocessing – Stage 1 (column vectors)

Initial Preprocessed Data

	count	mean	std	min	25%	50%	75%	max
age	4882.0	45.826506	11.155088	25.0	36.0	46.0	55.00	67.0
years_experience	4882.0	20.605899	11.136704	1.0	11.0	21.0	30.00	43.0
annual_income	4882.0	73.870750	46.112752	8.0	39.0	64.0	98.00	224.0
zip_code	4882.0	93167.386317	1760.397727	90005.0	91911.0	93437.0	94608.00	96651.0
family_size	4882.0	2.386112	1.148222	1.0	1.0	2.0	3.00	4.0
avg_monthly_cc_spend	4882.0	1.935412	1.745065	0.0	0.7	1.5	2.60	10.0
education	4882.0	1.874846	0.839329	1.0	1.0	2.0	3.00	3.0
mortgage_value	4882.0	56.844326	102.009136	0.0	0.0	0.0	101.75	635.0
personal_loan_conversion	4882.0	0.096887	0.295833	0.0	0.0	0.0	0.00	1.0
securities_account	4882.0	0.104056	0.305364	0.0	0.0	0.0	0.00	1.0
cd_account	4882.0	0.061450	0.240179	0.0	0.0	0.0	0.00	1.0
online_user	4882.0	0.598730	0.490206	0.0	0.0	1.0	1.00	1.0
credit_card	4882.0	0.294961	0.456072	0.0	0.0	0.0	1.00	1.0

Info

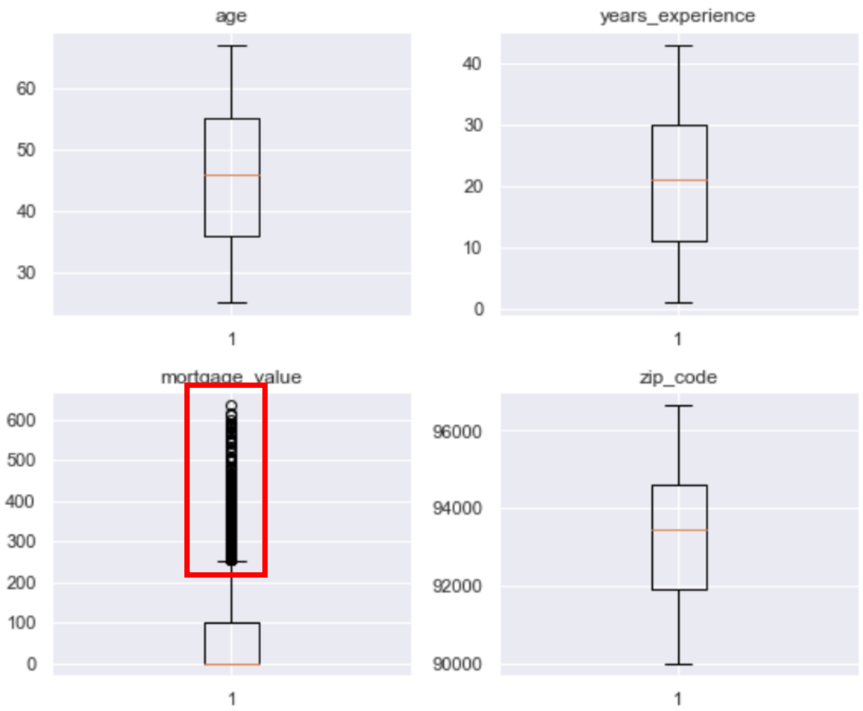
```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       4882 non-null   int64
 1   years_experience          4882 non-null   int64
 2   annual_income             4882 non-null   int64
 3   zip_code                  4882 non-null   int64
 4   family_size               4882 non-null   int64
 5   avg_monthly_cc_spend      4882 non-null   float64
 6   education                 4882 non-null   int64
 7   mortgage_value            4882 non-null   int64
 8   personal_loan_conversion  4882 non-null   int64
 9   securities_account        4882 non-null   int64
 10  cd_account                4882 non-null   int64
 11  online_user               4882 non-null   int64
 12  credit_card               4882 non-null   int64
```

Preprocessed shape: 4882 rows x 13 columns

Observations on data preprocessing

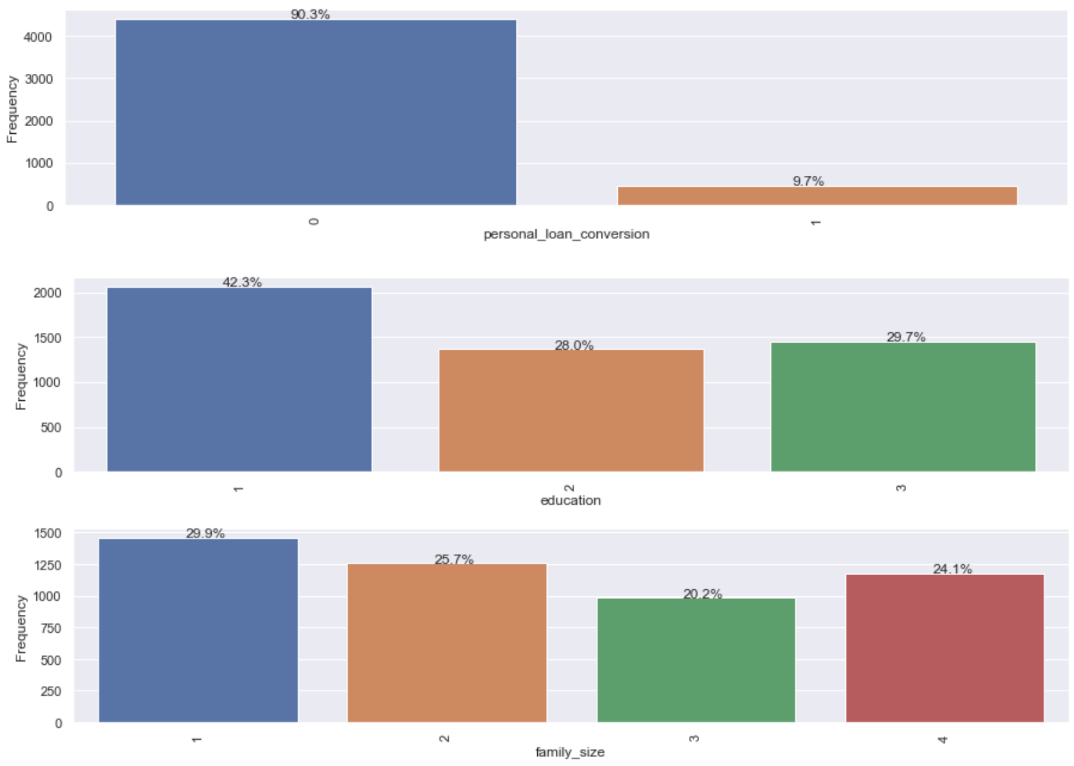
1. Column names were cleaned up to be intuitive to work with
2. Column ID was dropped as it will not be included in modeling
3. Looking at unique values across columns, they appear to be valid values (no odd values)
4. We can note zero values in mortgage_value and avg_monthly_cc_spend
5. Invalid values are observed in experience (e.g., -1, -2, -3)
6. Trade off decision - negative values for experiene were dropped to remove their impact to modeling (removed 2.42% of rows)
7. annual_income, mortgage_value and avg_monthly_cc_spend we scaled by 1000 to their actual values
8. Initial clean dataframe for modeling building and testing

Univariate EDA



Proportions

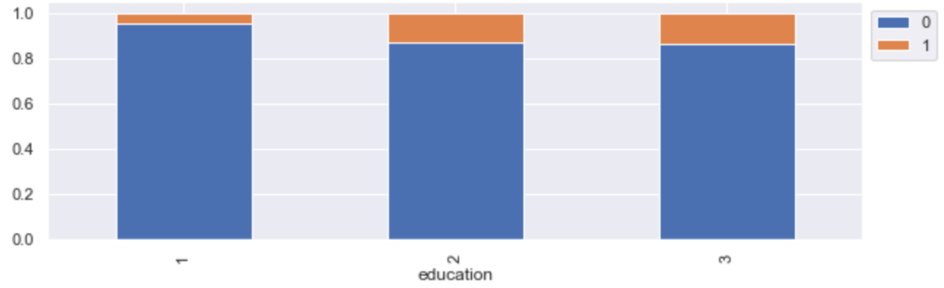
- Personal loans make up 9% of customers
- Largest proportion of customer have undergrad education
- Family sizes are comparable in proportion

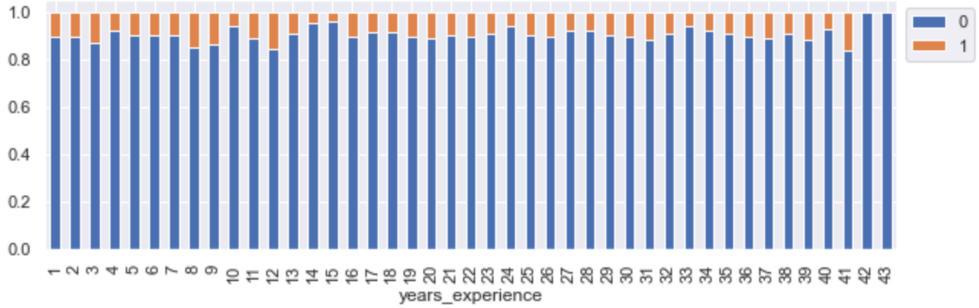


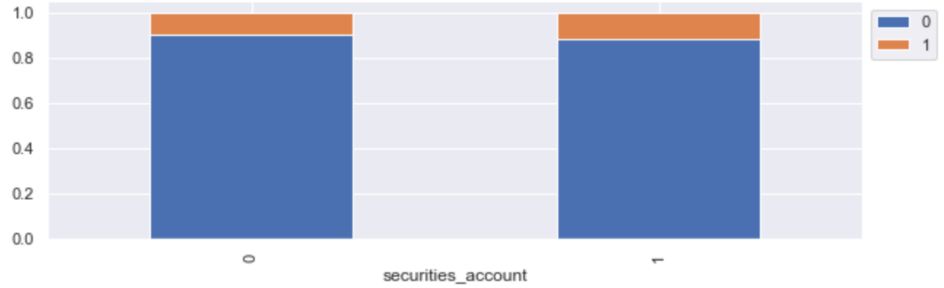
Obserations on univartate analysis

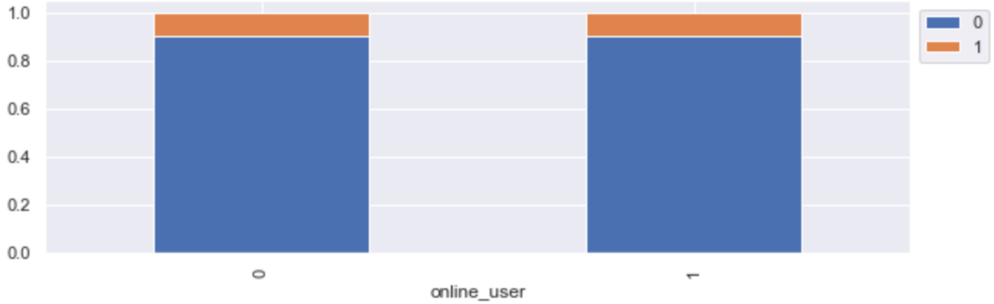
1. Countplot with percentages gives a sense how customers are distributed across variables
2. Majority (60%) of customers are online users
3. Majority (71%) of customers are bank credit card users
4. Majority (90%) of customers do not have a personal loan with the bank (opportunity)
5. Majority (90%) of customers to not have a securities_account with the bank
6. Majority (94%) of customers to not have a cd_account with the bank
7. Family size is roughly evenly distributed across 1, 2, 3 and 4
8. avg_monthly_cc_spend, mortgage_value and annual_income have outliers (which appear within a valid range)

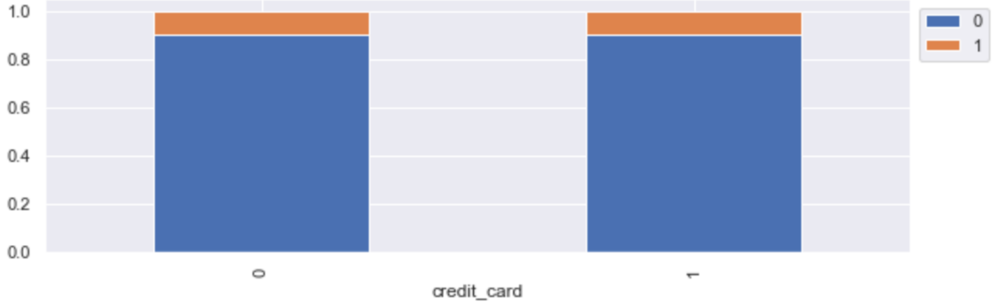
Multivariate EDA







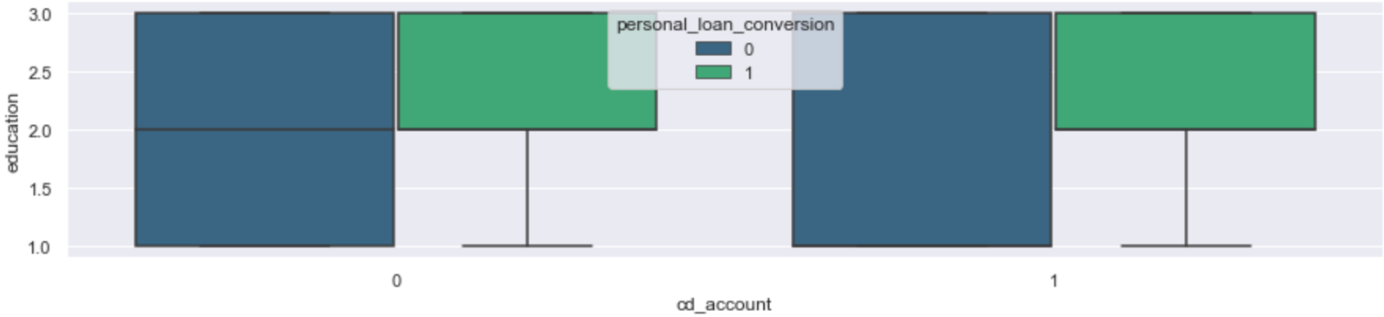




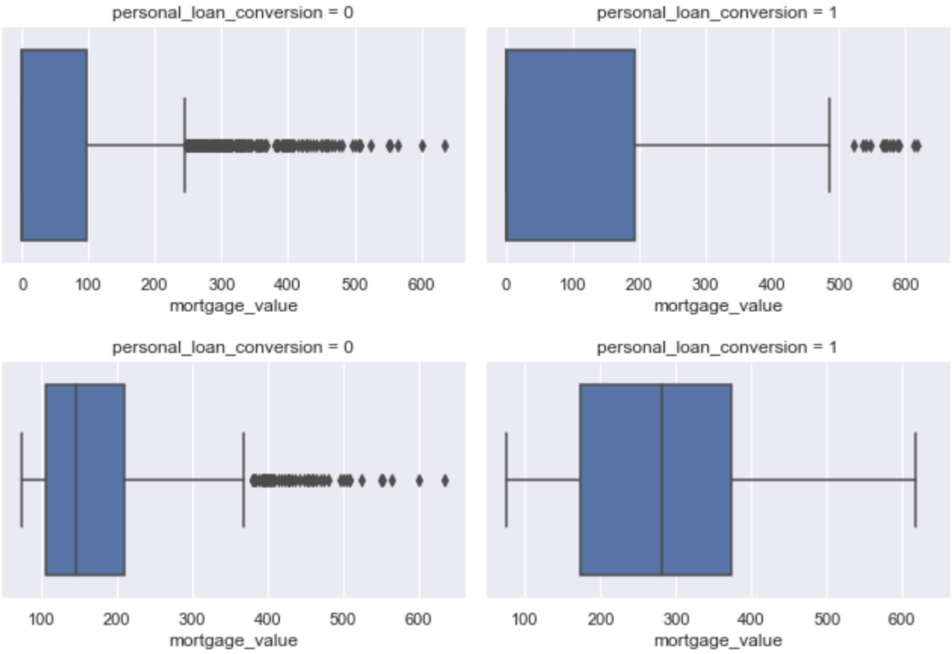
Observations on crosstab

Performing a crosstab between the target variable personal_loan_conversion and other independent variables shows how the positive class value (1) is distributed. We can see the customers with a cd_account have close to a 50% probability of taking out a personal loan.

Multivariate EDA



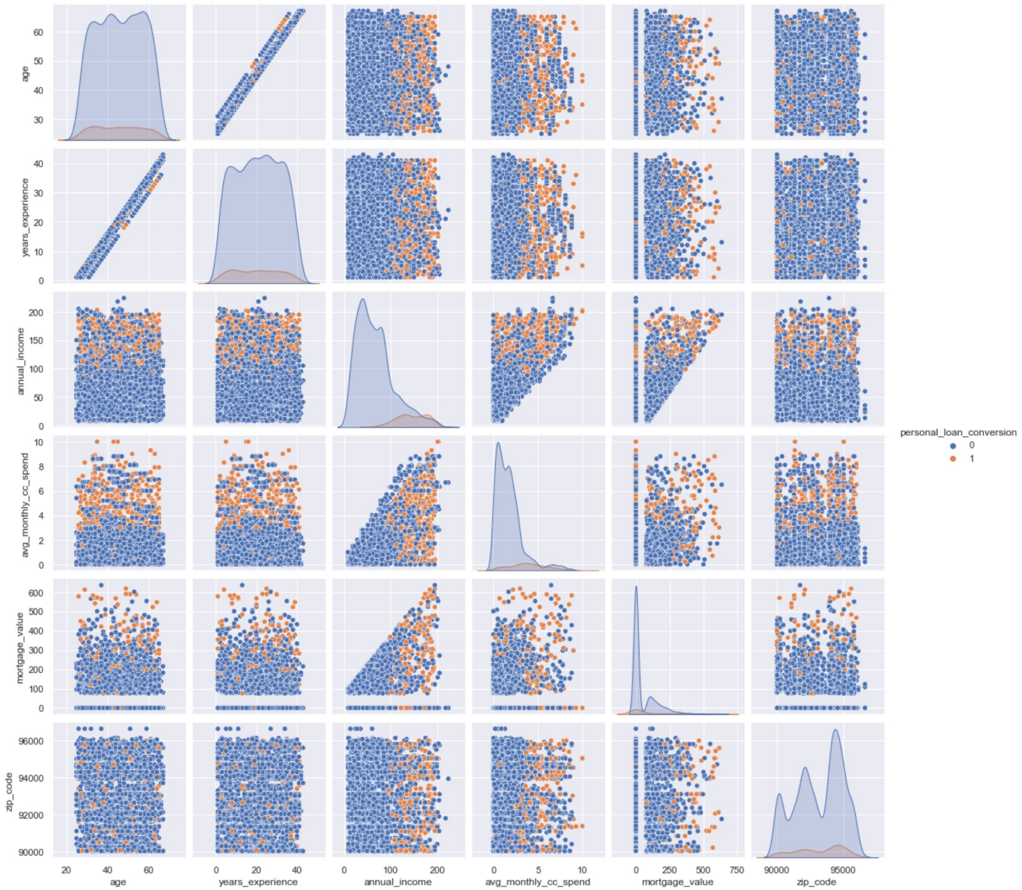
Boxplot shows that regardless of having a cd_account with the bank, customers who take out a personal loan have higher education levels than those who do not take out a personal loan



Boxplot of mortgage_value shows that customers who take out a personal loan with the bank have higher mortgage values

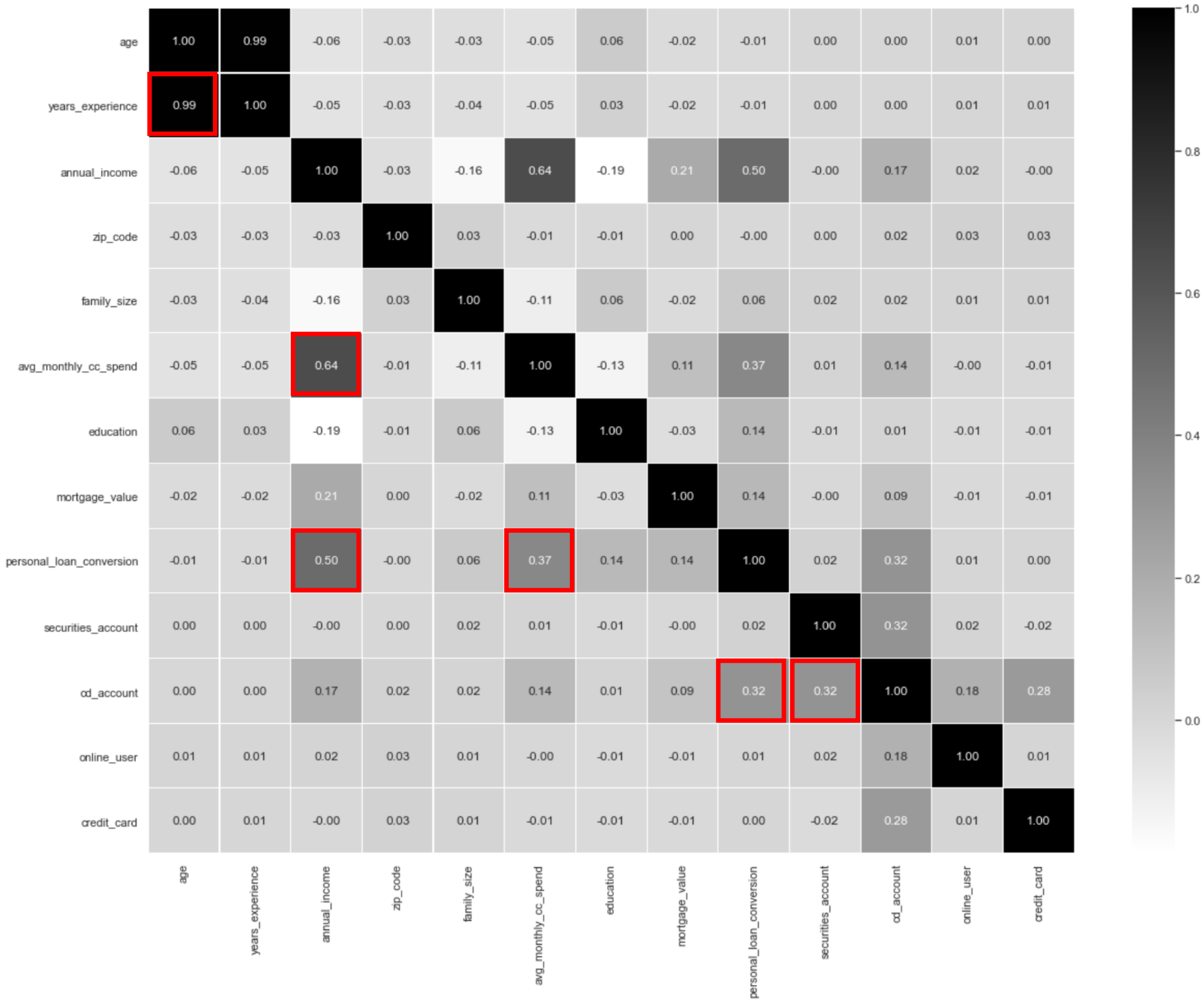
Boxplot shows that regardless of having a credit card with the bank, customers who take out a personal loan have higher annual incomes than those who do not take out a personal loan

Multivariate EDA



Pair plot shows a very strong collinearity relationship between age and years_experience. Other variables do not show visible linear relationships.

The correlation heatmap show associations between variables indicated by the red boxes.



Logistic Regression – Modeling & Performance

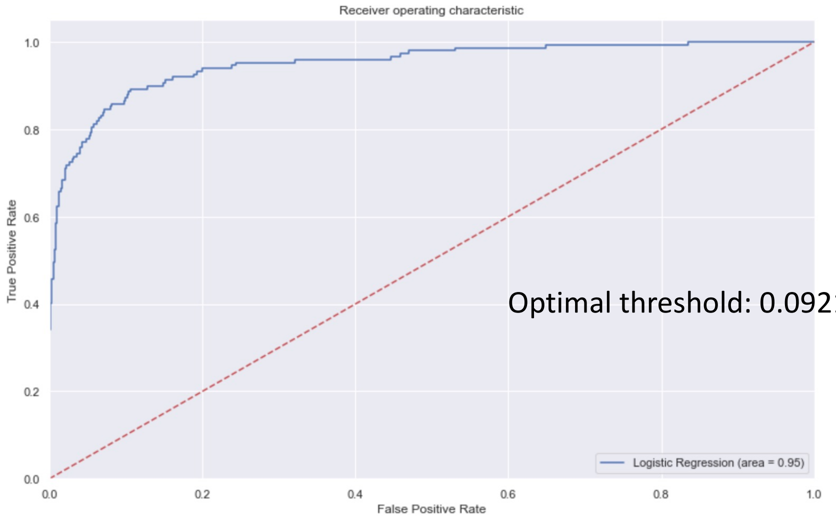
LR Method 1: statsmodels

```
Optimization terminated successfully.
         Current function value: 0.125000
         Iterations 9
                              Results: Logit
==========================================================================
Model:              Logit                    Pseudo R-squared: 0.601
Dependent Variable: personal_loan_conversion AIC:              878.2524
Date:               2021-03-20 21:09         BIC:              951.8907
No. Observations:   3417                     Log-Likelihood:   -427.13
Df Model:           11                       LL-Null:          -1071.4
Df Residuals:       3405                     LLR p-value:      1.3193e-269
Converged:          1.0000                   Scale:            1.0000
No. Iterations:     9.0000
--------------------------------------------------------------------------
                          Coef.  Std.Err.    z    P>|z|   [0.025   0.975]
--------------------------------------------------------------------------
const                  -15.3311   4.8202 -3.1806 0.0015 -24.7785 -5.8837
years_experience         0.0145   0.0081  1.7897 0.0735  -0.0014  0.0304
annual_income            0.0552   0.0032 17.4694 0.0000   0.0490  0.0614
zip_code                 0.0000   0.0001  0.3174 0.7509  -0.0001  0.0001
family_size              0.6777   0.0898  7.5430 0.0000   0.5016  0.8538
avg_monthly_cc_spend     0.1368   0.0490  2.7934 0.0052   0.0408  0.2328
education                1.7700   0.1425 12.4178 0.0000   1.4906  2.0493
mortgage_value          -0.0001   0.0007 -0.1303 0.8964  -0.0014  0.0013
securities_account      -0.9446   0.3503 -2.6964 0.0070  -1.6313 -0.2580
cd_account               3.9284   0.4052  9.6961 0.0000   3.1343  4.7225
online_user             -0.5227   0.1937 -2.6983 0.0070  -0.9024 -0.1430
credit_card             -1.2689   0.2556 -4.9645 0.0000  -1.7699 -0.7679
==========================================================================

Recall on train data: 0.6358024691358025
Recall on test data: 0.6577181208053692

Accuracy on train data: 0.9534679543459175
Accuracy on test data: 0.9535836177474403
```

Regression Modeling
Method 2: sklearn

Recall on train data: 0.6358024691358025
Recall on test data: 0.6577181208053692

Accuracy on train data: 0.9531752999707346
Accuracy on test data: 0.9535836177474403



Optimal threshold: 0.09218191027302287

Observations on Logistic Regression modeling

1. Scaling (log and 1000) variables appear to have little effect on model scoring
2. We are able to achieve Accuracy on Test data of .95 (quite high, % of accurate predictions over all data)
3. We are able to achieve a Recall of Test data of .66 (this is the % of actual 1s captured by prediction)
4. We have utilized logistic regression algorithms from both statsmodels and sklearn with compareable results

Decision Tree - Modeling

Decision Tree binary classifier – no optimization

model = DecisionTreeClassifier(criterion='gini', class_weight={0:0.10,1:0.90}, random_state=1)

Gini classifier results:

Train

0 (no loan conversion) 0.90518

1 (loan conversion) 0.09482

Name: personal_loan_conversion, dtype: float64

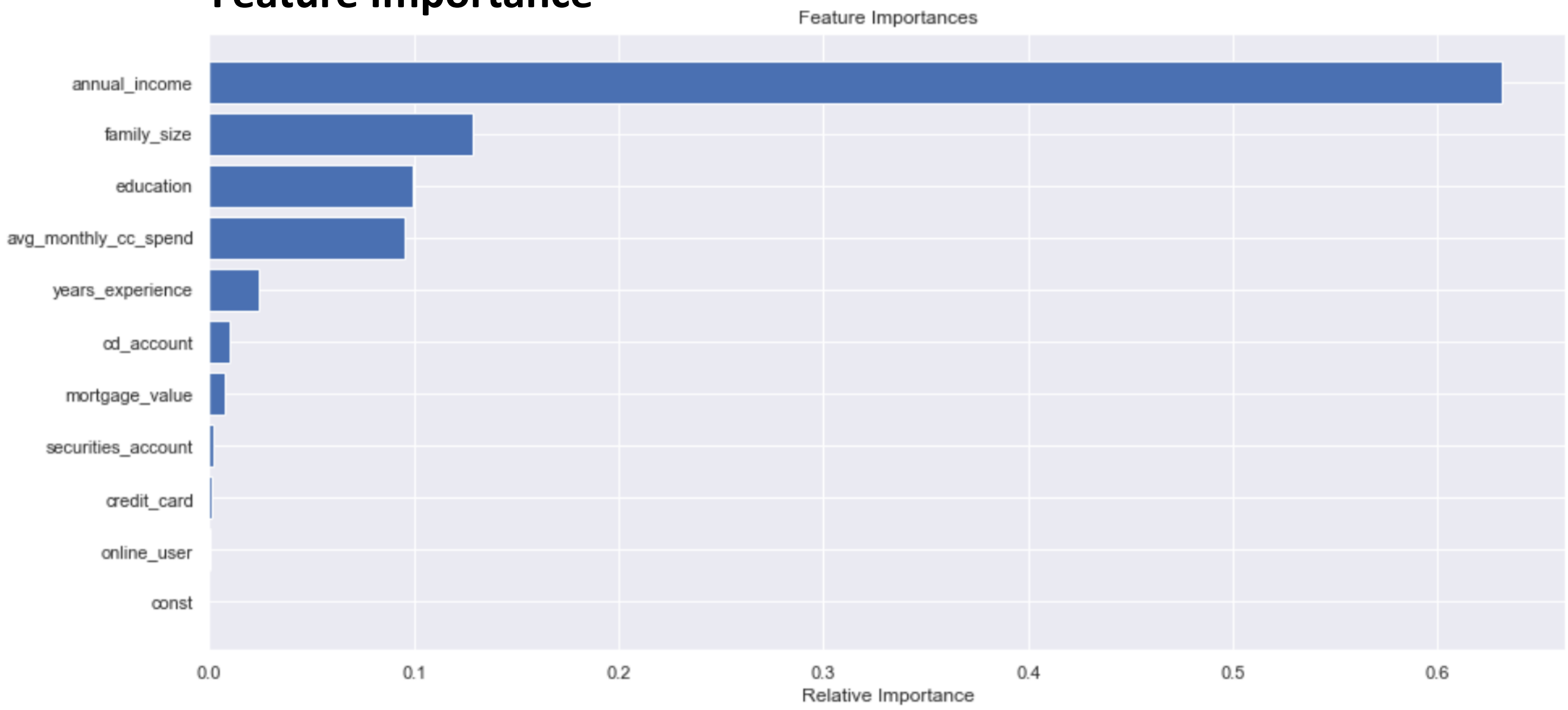
Test

0 (no loan conversion) 0.898294

1 (loan conversion) 0.101706

Name: personal_loan_conversion, dtype: float64

Feature Importance



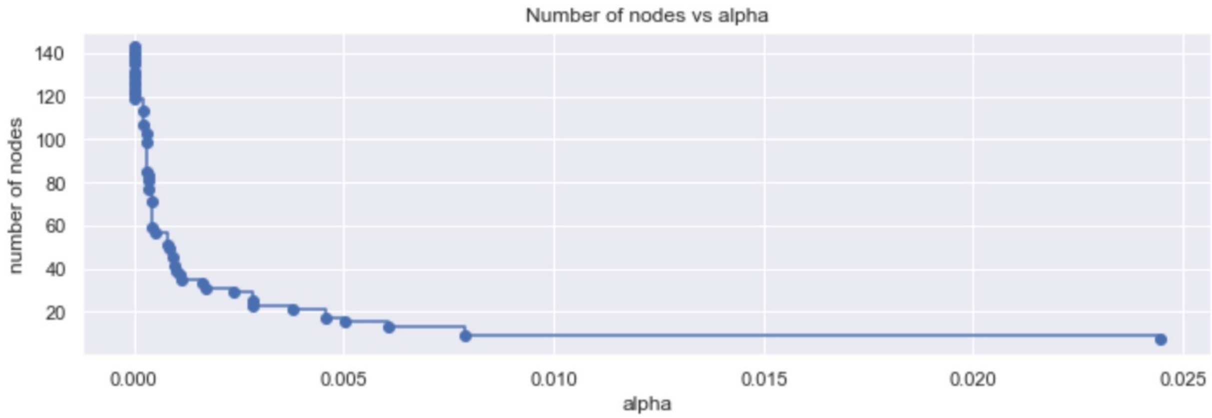
We can use feature importance to target marketing efforts based on top impacting variables

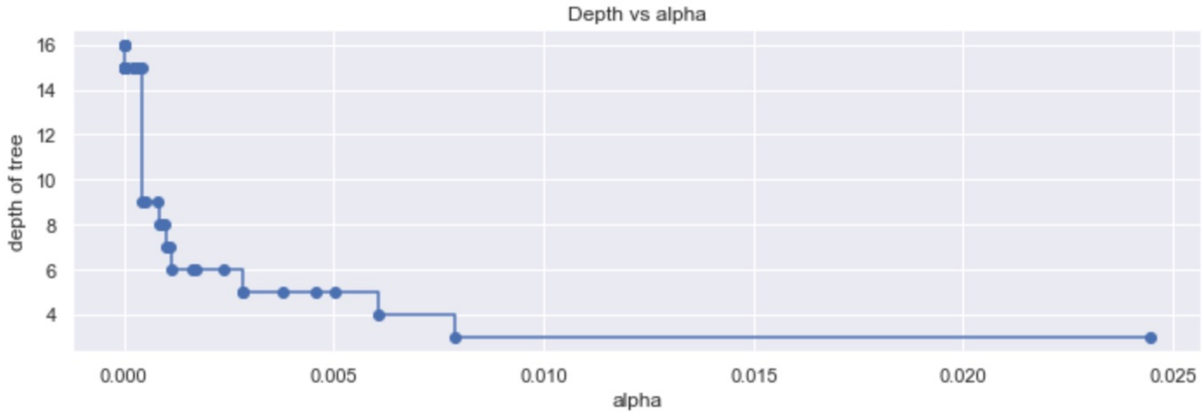
Decision Tree – Performance & Optimization

Cost Complexity Pruning

Start with a full tree and remove sub-trees

Relative error decrease per node (complexity param αlpha)



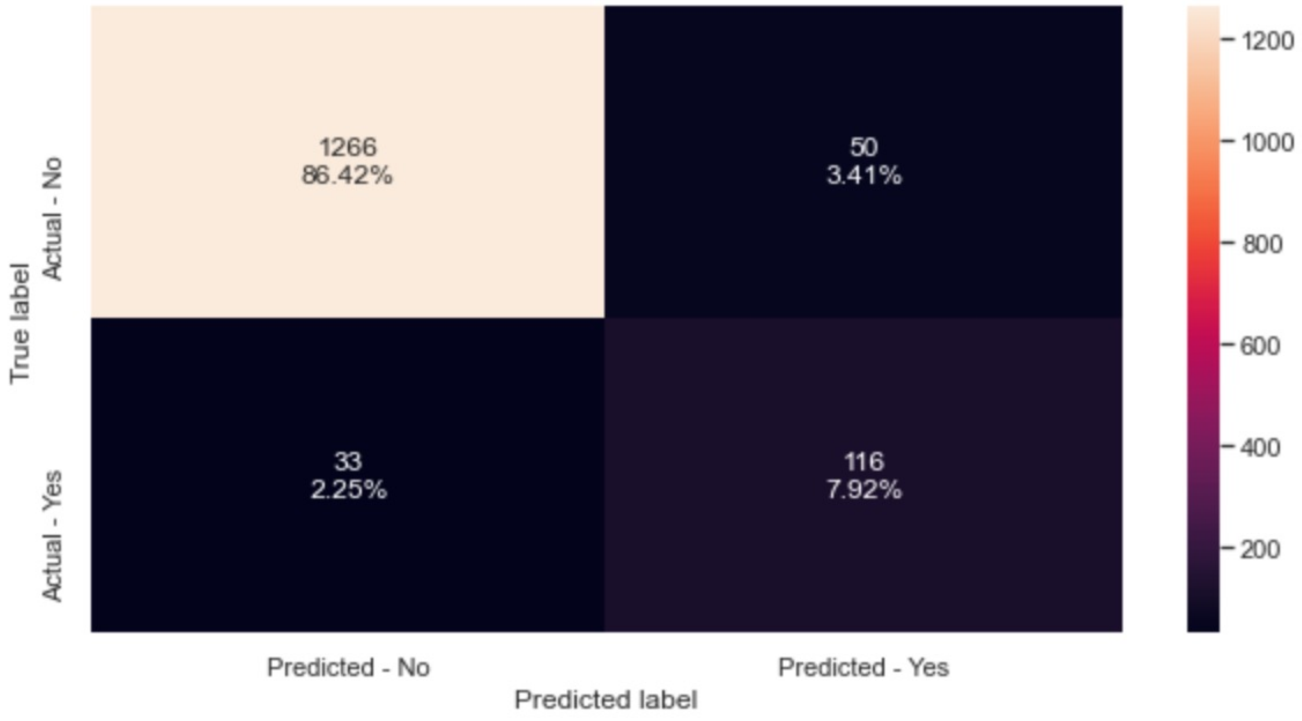


Decision Tree – Performance & Optimization

GridSearchCV
Try different combinations of hyper-parameters to determine the best modeling configuration

Best hyper-parameters:
DecisionTreeClassifier(
class_weight={0: 0.15, 1: 0.85},
max_depth=4,
max_features='log2',
min_impurity_decrease=1e-06,
random_state=1)

Recall Scoring
Train data: 0.7993827160493827
Test data: 0.7785234899328859



Insights & Recommendations to Business

Insights - Logistic Regression

- Data preprocessing required to build a CART model-based prediction is minimal
- Model performance of logistic regression algorithm is less easy to tune compared with decision tree algorithm

Insights - Decision Trees

- By using GridSearchCV, we were able to improve Recall to .7785
- Features importance:
  - annual_income
  - family_size
  - education
  - avg_monthly_cc_spend

Recommendations to Business

- Build a marketing campaign around importance features
- Enhance bank website to present message to customers which have a annual_income higher 100k with special offer interest rate for personal loans
- Enhance bank website to present message to customers which have a family size of greater than 2 with special offer interest rate for personal loans
- Enhance bank website to present message to customers which have an avg_monthly_cc_spend greater than 3k with special offer interest rate on personal loans
- Enhance bank wbesite to present message to customers which have education greater than 1 (undergraduate degree) with special offer interest rate on personal loans