AllLife Bank
Personal Loan Campaign
Business Presentation

Contents

Business Problem Overview and Solution Approach

- AllLife Bank is a US bank that has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and we are interested in expanding this base rapidly to bring in more loan business and earn more through the interest on loans. In particular, explore ways of converting liability customers to personal loan customers (while retaining them as depositors).
- A campaign 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.

Business Problem Overview and Solution Approach

- A model is to be built that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.
- The objective of the model is:
  - 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 Overview

Variable	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)
ZIPCode	Home Address ZIP code
Family	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)?

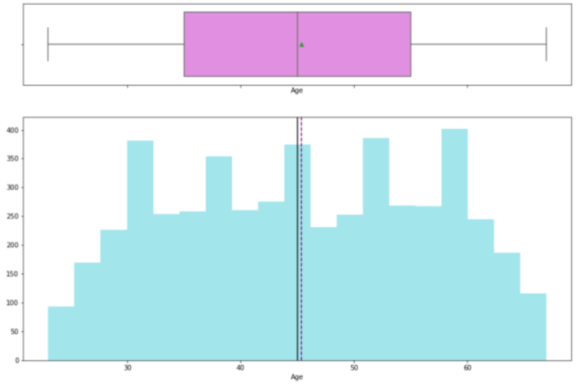
Observations	Variables
5000	14

Note:

- ID column is removed.
- ZIPCode column is removed.
- The Education Column is converted to words instead of numerals.
- Family, Education, Personal_Loan, Securities_Account, CD_Account, Online, CreditCard columns have been converted to category.

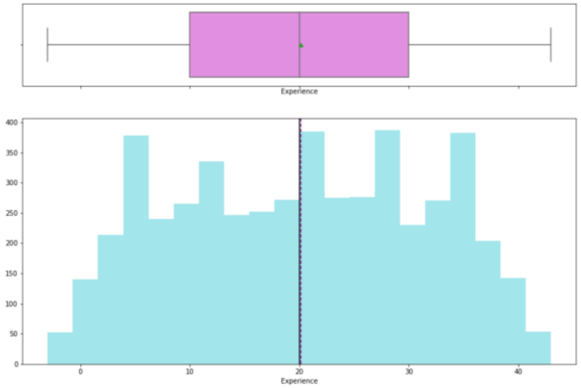
EDA – Age, Experience & Income

Age



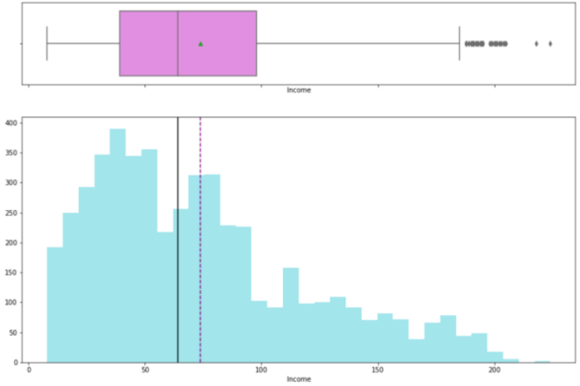
- The distribution of age is normal.
- The boxplot shows that there are no outliers.

Experience



- The distribution of Experience is normal.
- The boxplot shows that there are no outliers.

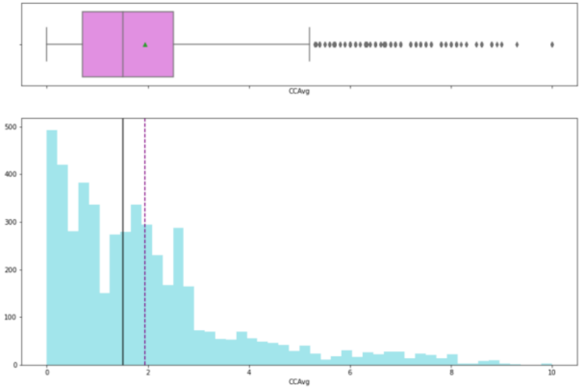
Income



- The distribution of Income is right skewed.
- The boxplot shows outliers to the higher end of the income band.
- We will not treat these outliers as they represent the real market trend.

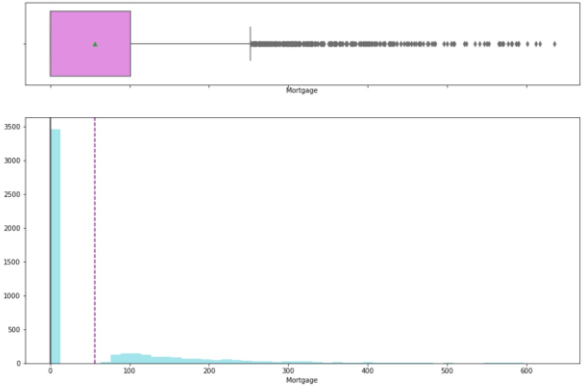
EDA – Credit Card Spending & Mortgage Value

Credit Card Spending



- The distribution of Average spending on credit cards per month is right skewed.
- The boxplot shows outliers to the higher end of the credit card spending.
- We will not treat these outliers as they represent the real market trend.

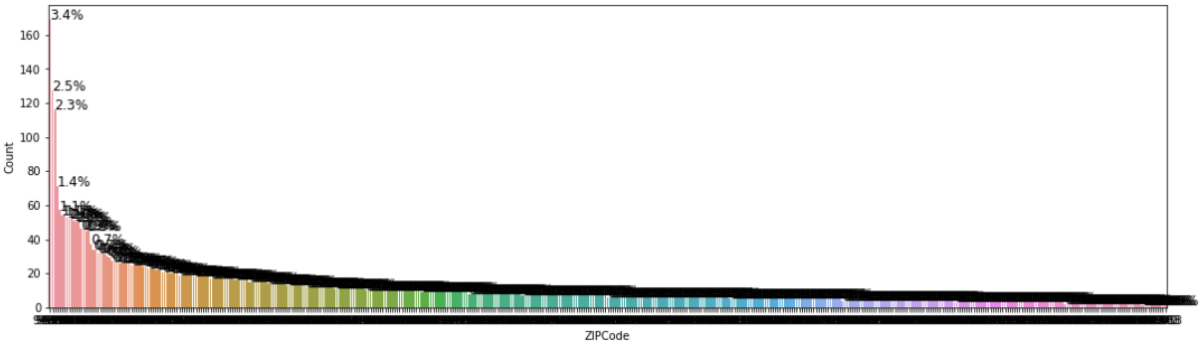
Mortgage Value



- Minority of customers took a mortgage with ranges from 60 to 600K.

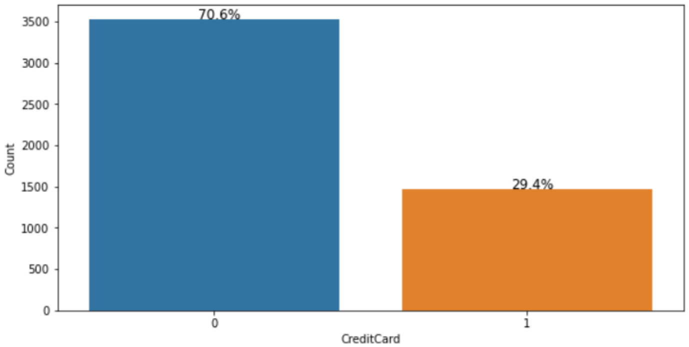
EDA – ZIPCode & CreditCard

ZIPCode



- Customers' locations are dispersed. No discernable trend can be observed. It was eventually removed prior to machine learning modeling.

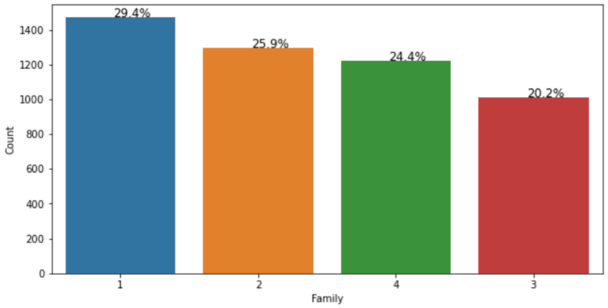
CreditCard



- About 70% of the customers do not use a credit card issued by any other Bank.

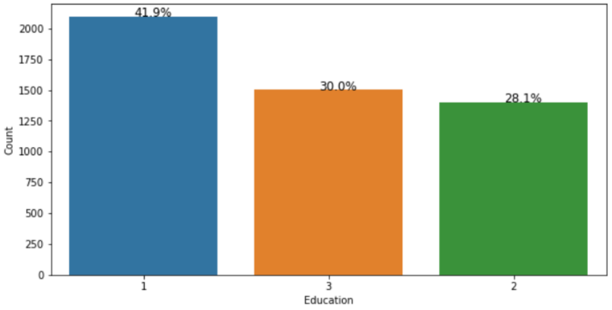
EDA – Family, Education & Personal_Loan

Family



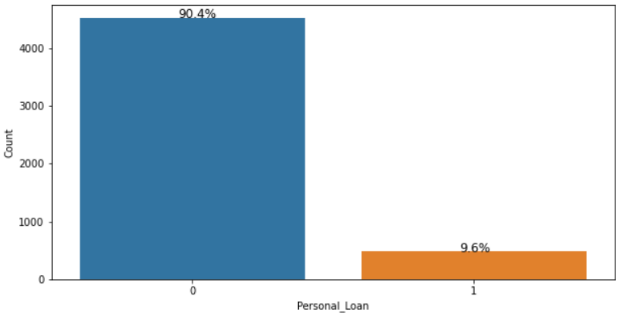
- Family size of customers are also fairly evenly distributed with more trending towards smaller families.

Education



- Almost half of the customers are undergrads while just over half are Grads or have higher education.

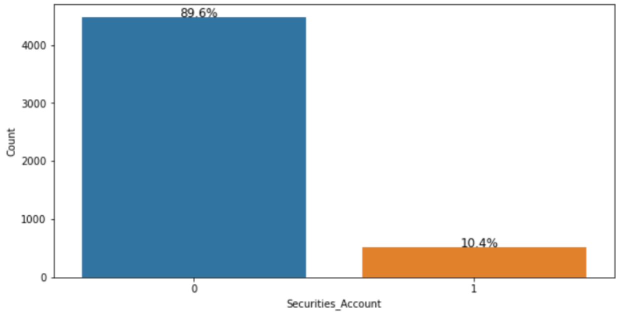
Personal_Loan



- 9.6% of the customers have taken Personal Loans with the bank so far.

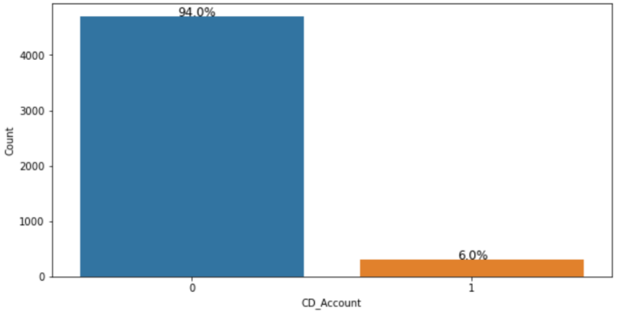
EDA – Securities_Account, CD_Account & Online

Securities_Account



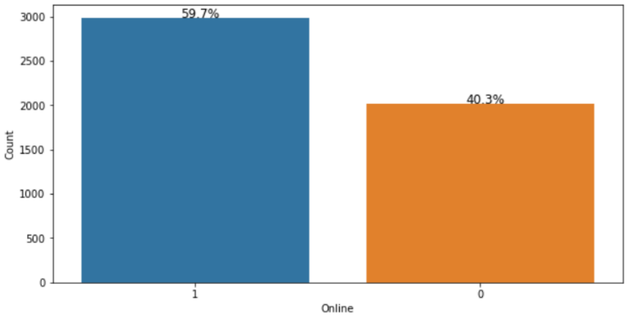
- Almost 90% of the customers do not have securities account with the bank.

CD_Account



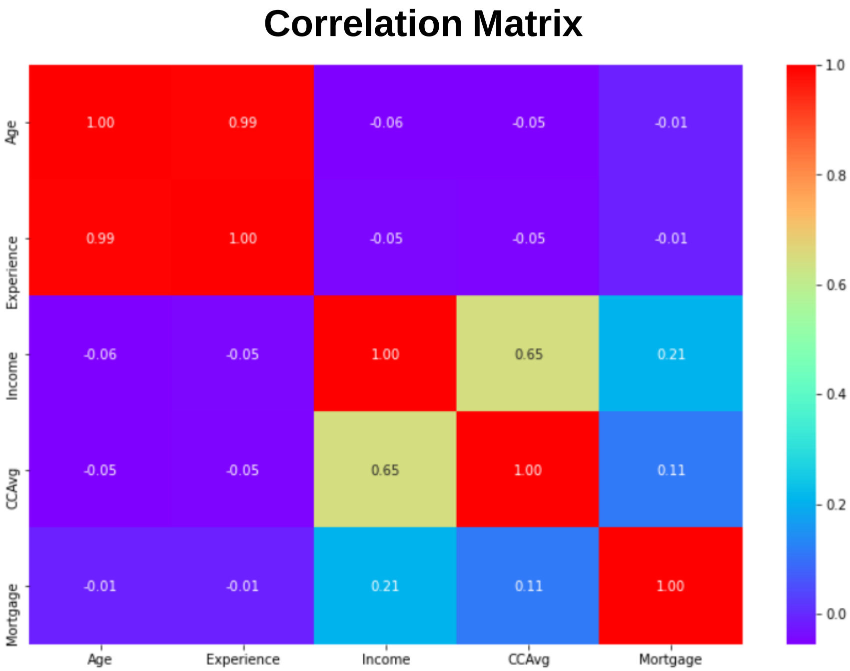
- 94% of the customers do not have CD account with the bank.

Online



- Almost 60% of the customers use internet banking facilities.

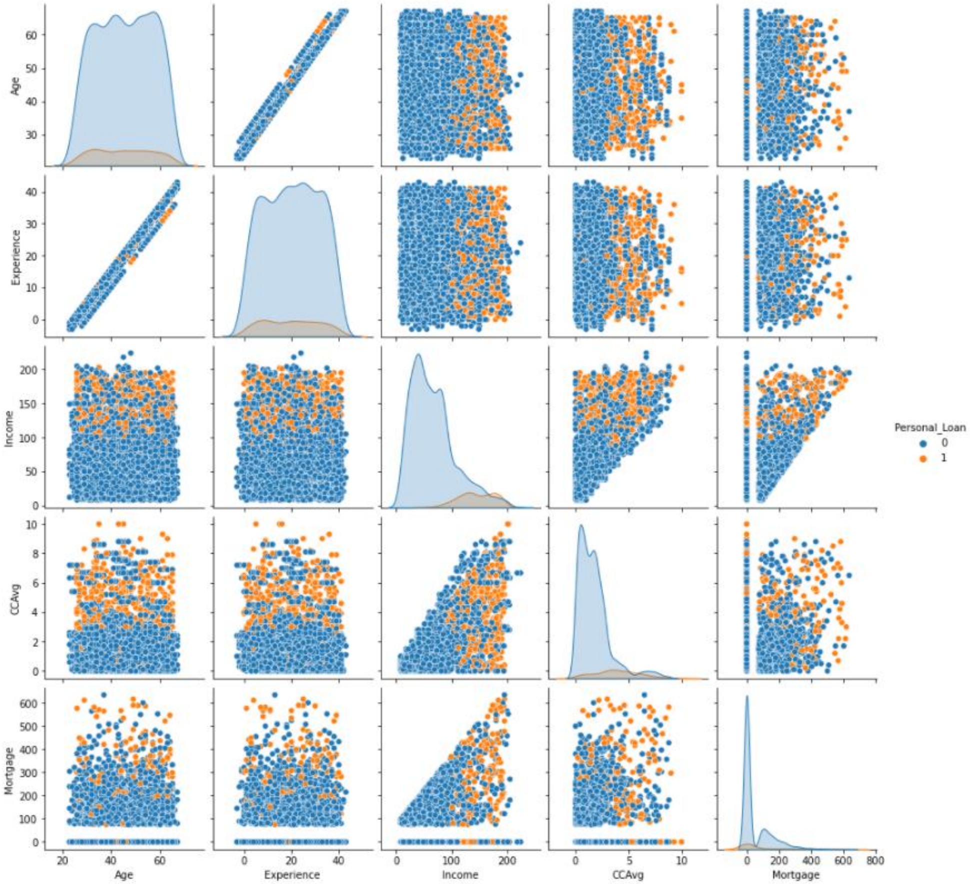
EDA – Correlation Matrix



Observations

- Age and Work Experience of customers is very closely correlated.
- Income and Average credit card spent per month is also correlated.
- Other variables have no significant correlation between them.

EDA – Pairplot



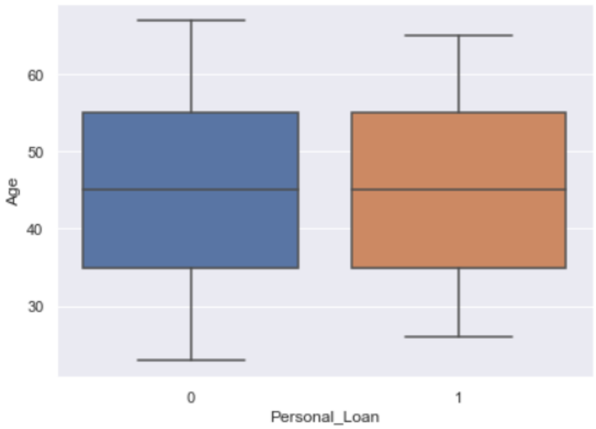
Observations

- Customers higher income and average credit card spending are more likely to accept personal loans over lower income and average credit card spending.

- Customers with higher mortgage value have a slight tendency to accept personal loans.

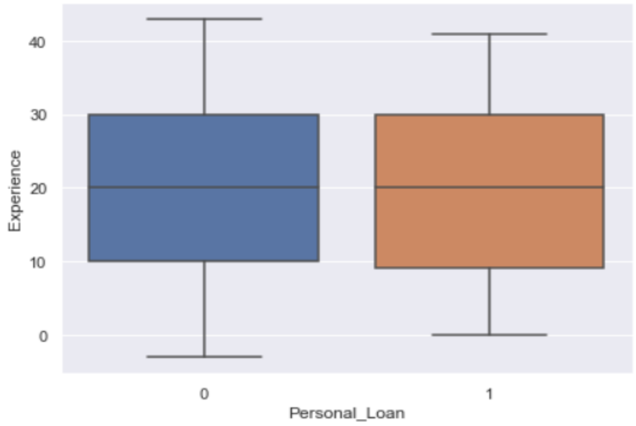
EDA – Personal Loan with Age & Experience

Personal Loan Vs Age



- We can see that median ages and age distribution between 25th percentile to 75th percentile of personal loaners and non personal loaners are similar.
- The age ranges of customers taking up personal loan is between ~ 27 to 66.
- There are no outliers in boxplots of both class distributions.

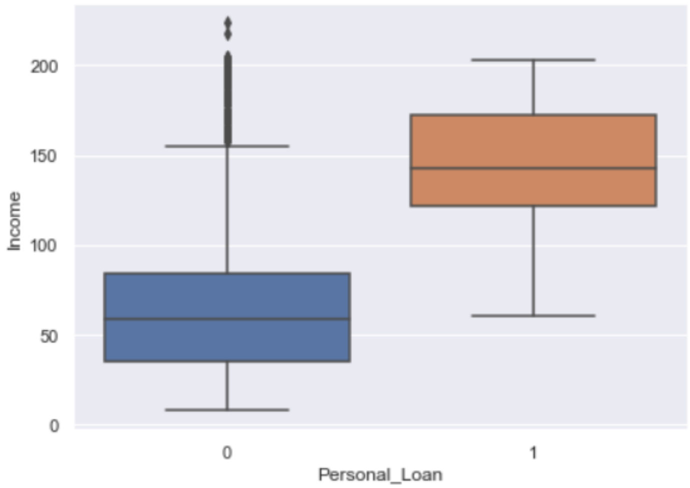
Personal Loan Vs Experience



- We can see that median work experience and work experience distribution between 25th percentile to 75th percentile of personal loaners and non personal loaners is about similar with the IQR range a little wider.
- The work experience ranges of customers taking up personal loan is between ~ 0 to 41.
- The negative work experience values may need fixing.
- There are no outliers in boxplots of both class distributions.

EDA – Personal Loan with Income

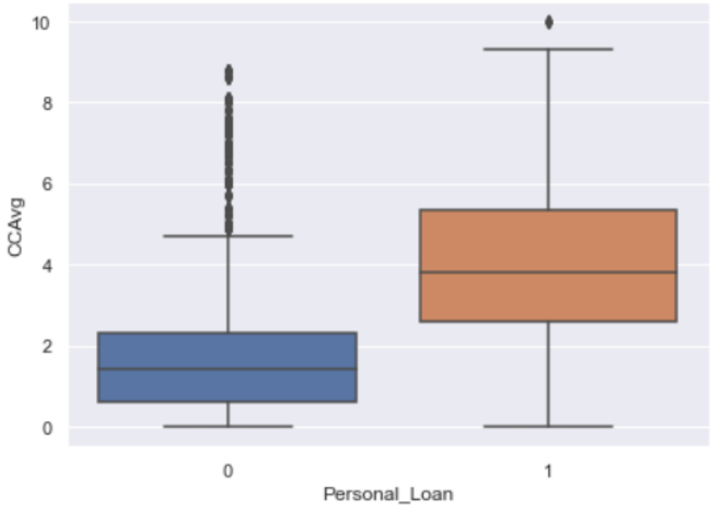
Personal Loan Vs Income



- We can see that median and distribution of personal loaners' income levels are higher at just under 150K and IQR between about 125K to 175K. There are no outliers.
- This is compared to non personal loaners at median income of ~ 60K and IQR between close to 0 to just above 150K.
- The personal loaners income ranges from just above 50K to ~ 200K compared to non loaners range for 1.5 IQR of just above 0 income to just above 150K.
- There are outliers in boxplots of class distributions of non personal loaners with higher income ranges.

EDA – Personal Loan with CCAvg

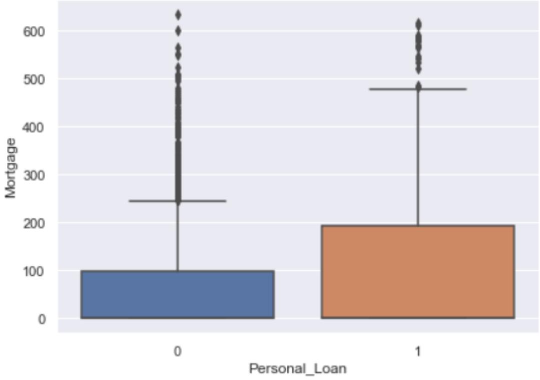
Personal Loan Vs CCAvg



- We can see that median and distribution of personal loaners' credit card average spending are higher at just under 4K and IQR between about 2.2K to 5.5K. There are outliers at high credit card spending at 10K.
- This is compared to non personal loaners at median credit card spending of ~ 1.5K and IQR between less than 1K to just above 2K.
- The personal loaners credit card average spending of 1.5 IQR ranges up to above 9K compared to non loaners up to just under 5K.
- There are outliers in boxplots of class distributions of personal and non personal loaners with higher credit card spending.

EDA – Personal Loan with Mortgage & Family

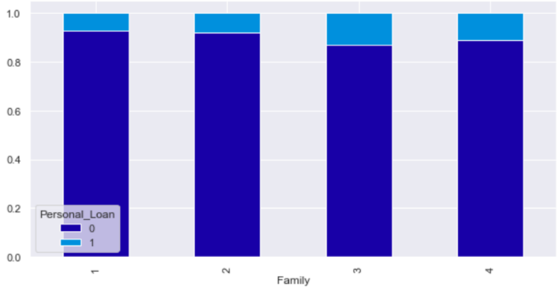
Personal Loan Vs Mortgage



- We can see that distribution of personal loaners' mortgage are higher at IQR up to just under 200K.
- This is compared to non personal loaners at IQR up to just 100K.
- The personal loaners mortgage ranges of 1.5 IQR is also higher at just under 500K compared to non loaners at 250K.
- There are outliers in boxplots of class distributions of personal and non personal loaners on the higher end of mortgage values.

Personal Loan Vs Family

Personal_Loan	0	1	All
Family			
1	1365	107	1472
2	1190	106	1296
3	877	133	1010
4	1088	134	1222
All	4520	480	5000

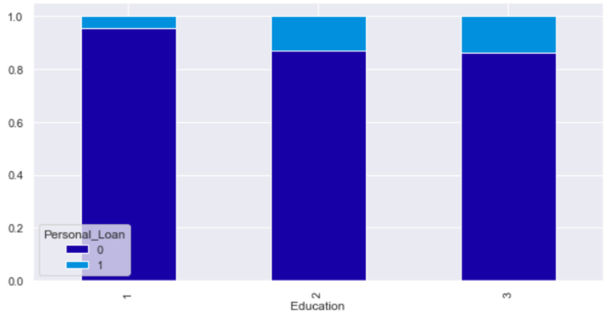


- Customers with larger families are more likely to take personal loans than smaller families.

EDA – Personal Loan with Education & Securities_Account

Personal Loan Vs Education

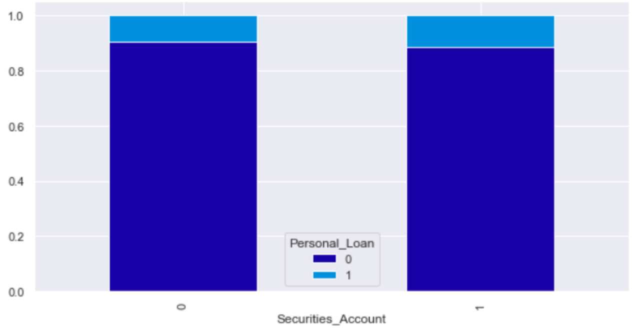
Personal_Loan	0	1	All
Education			
1	2003	93	2096
2	1221	182	1403
3	1296	205	1501
All	4520	480	5000



- Customers who have graduated or with advanced degrees are more likely to take personal loans.

Personal Loan Vs Securities_Account

Personal_Loan	0	1	All
Securities_Account			
0	4058	420	4478
1	462	60	522
All	4520	480	5000

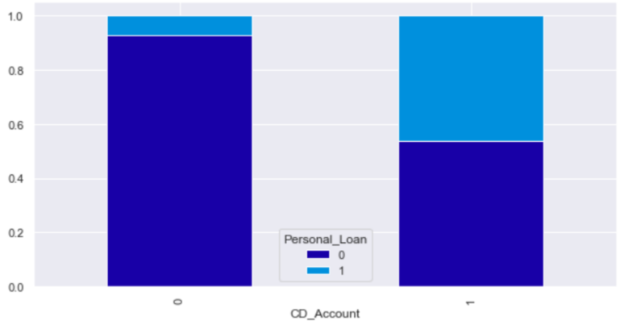


- There is no discernable differences between proportion of personal loan takers among those with or without securities accounts.

EDA – Personal Loan with CD_Account & Online

Personal Loan Vs CD_Account

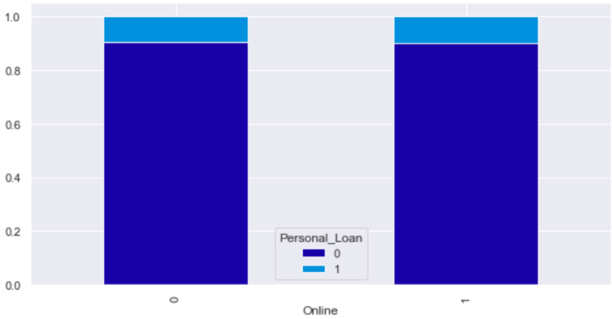
Personal_Loan	0	1	All
CD_Account			
0	4358	340	4698
1	162	140	302
All	4520	480	5000



- Customers who have CD accounts (fixed term deposits) are very likely to take personal loans close to half of them.

Personal Loan Vs Online

Personal_Loan	0	1	All
Online			
0	1827	189	2016
1	2693	291	2984
All	4520	480	5000

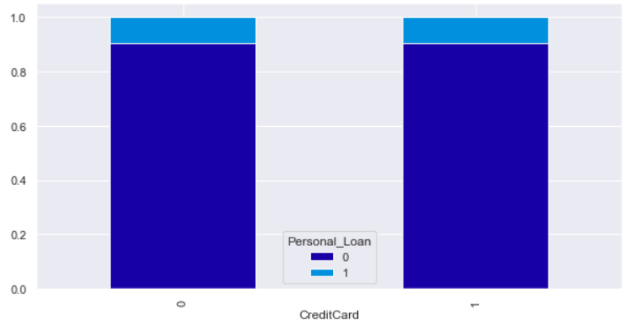


- There is no discernable differences between proportion of personal loan takers among using online banking facilities or not.

EDA – Personal Loan with CreditCard & ZIPCode

Personal Loan Vs CreditCard

```
Personal_Loan     0    1   All
CreditCard
0              3193  337  3530
1              1327  143  1470
All            4520  480  5000
--------------------------------------------------------------------------------
```

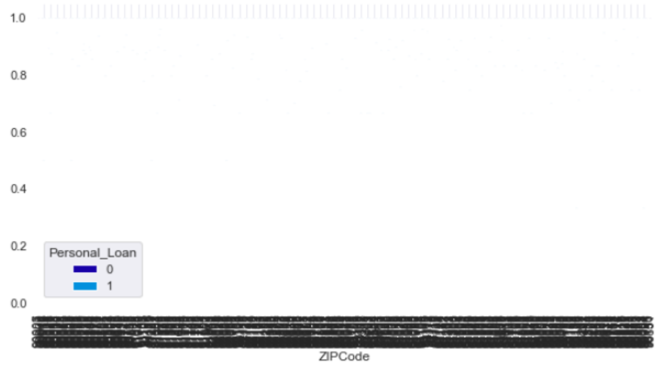


- There is no discernable differences between proportion of personal loan takers among those who use other banks issued credit cards or not.

Personal Loan Vs ZIPCode

```
Personal_Loan     0    1   All
ZIPCode
90005             5    0     5
90007             6    0     6
90009             8    0     8
90011             3    0     3
90016             1    1     2
...             ...  ...   ...
96094             2    0     2
96145             1    0     1
96150             4    0     4
96651             6    0     6
All            4520  480  5000

[468 rows x 3 columns]
--------------------------------------------------------------------------------
```



- There is no discernable trend of more or less personal loan takers from any Zip code.

EDA – Insights & Data Pre-Processing

- Insights
  - Personal loaners tended towards high income customers with higher credit card spending.
  - Mortgage takers who accepted personal loans also tended towards those with higher mortgage values.
  - Customers with CD accounts are most likely to accept a personal loan from the bank.
  - Customers with family sizes more than 2 or who have graduated or with advanced/professional degrees are more likely to accept a personal loan.

- Data Pre-Processing
  - There are no missing values and duplicate entries.
  - We will not treat these outliers as they represent the real market trend.
  - Education variable values will be converted to words.
  - ZIPCode variable will be removed as it has over 400 unique values with the highest frequency at 3.4%. only so it is too widely dispersed to serve any trending and it has no bearing on personal loan takers.

Data Preparation

- The data set is split into 70% for training and 30% for testing
- Dummy variables were prepared for categorical variables
- The list of variables/features used for all the models are as below:
  - 'Age', 'Experience', 'Income', 'CCAvg', 'Mortgage', 'Family_2', 'Family_3', 'Family_4', 'Education_Grad', 'Education_Undergrad', 'Securities_Account_1', 'CD_Account_1', 'Online_1', 'CreditCard_1'

Model Performance Summary – Logistic Regression

- Model Evaluation Criterion
  - Model can make wrong predictions as:
    - False Positive: Predicting a customer is a personal loan convertable but actually not convertable.
    - False Negative: Predicting a customer is a personal loan non-convertable but actually convertable.

- Which case is more important?
  - Both the cases are important as:
  - If we predict a customer is a personal loan convertable but actually not convertable then a wrong person will be getting the targeted marketing effort wasting resources.
  - If we predict a customer is a personal loan non-convertable but actually convertable, that person will not be able to receive targeted marketing effort and hence may not be aware of the personal loan service and thus a loss of business.

Model Performance Summary – Logistic Regression

- How to reduce losses?
  - We can use accuracy but since the data is imbalanced it would not be the right metric to check the model performance.
  - Therefore, f1_score should be maximized, the greater the f1_score higher the chances of identifying both the classes correctly.

Model Performance Summary – Logistic Regression Model 1 (lg)

```
Optimization terminated successfully.
         Current function value: 0.117686
         Iterations 9
                           Results: Logit
====================================================================
Model:              Logit            Pseudo R-squared: 0.618
Dependent Variable: Personal_Loan    AIC:              853.8055
Date:               2021-06-04 22:55 BIC:              946.2133
No. Observations:   3500             Log-Likelihood:   -411.90
Df Model:           14               LL-Null:          -1077.3
Df Residuals:       3485             LLR p-value:      1.3033e-275
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
--------------------------------------------------------------------
                     Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------
const               -7.5362   2.1241  -3.5480 0.0004 -11.6993 -3.3731
Age                 -0.0359   0.0790  -0.4549 0.6492  -0.1908  0.1189
Experience           0.0480   0.0786   0.6104 0.5416  -0.1061  0.2021
Income               0.0584   0.0035  16.7634 0.0000   0.0516  0.0653
CCAvg                0.1899   0.0558   3.4014 0.0007   0.0805  0.2993
Mortgage             0.0013   0.0007   1.8080 0.0706  -0.0001  0.0026
Family_2            -0.0327   0.2718  -0.1201 0.9044  -0.5654  0.5001
Family_3             1.7003   0.3024   5.6231 0.0000   1.1076  2.2929
Family_4             1.6825   0.2776   6.0601 0.0000   1.1383  2.2267
Education_Grad      -0.1574   0.2268  -0.6939 0.4878  -0.6018  0.2871
Education_Undergrad -3.8520   0.3200 -12.0384 0.0000  -4.4791 -3.2249
Securities_Account_1 -0.8738  0.3695  -2.3650 0.0180  -1.5980 -0.1496
CD_Account_1         3.6328   0.4133   8.7889 0.0000   2.8227  4.4429
Online_1            -0.6448   0.2004  -3.2171 0.0013  -1.0377 -0.2520
CreditCard_1        -1.0800   0.2620  -4.1221 0.0000  -1.5935 -0.5665
====================================================================

Accuracy on train data: 0.9597142857142857
Accuracy on test data: 0.9633333333333334
Recall on train data: 0.6656346749226006
Recall on test data: 0.732484076433121
Precision on train data: 0.8669354838709677
Precision on test data: 0.8984375
f1 score on train data: 0.7530647985989493
f1 score on test data: 0.8070175438596491
```

Vif Score

```
Series before feature selection:

const                   485.134684
Age                      93.542430
Experience               93.412885
Income                    1.886842
CCAvg                     1.725779
Mortgage                  1.061917
Family_2                  1.386231
Family_3                  1.385494
Family_4                  1.418310
Education_Grad            1.445589
Education_Undergrad       1.554669
Securities_Account_1      1.144488
CD_Account_1              1.342380
Online_1                  1.042382
CreditCard_1              1.113117
dtype: float64
```

Observations

- The outputs are pretty reliable for our targeted marketing prediction purposes but data might contain multicollinearity so variables can be removed based on insignificance where pvalue > 0.05.
- f1 score can still be improved at 75.3% for train data and 80.7% for test data.
- Age has pvalue=0.6492 so it could be dropped due to insignificance.
- Experience has pvalue=0.5416 so it could be dropped due to insignificance.
- Mortgage has pvalue=0.0706 so it could be dropped due to insignificance.
- Some variables of Family and Education are significant, so we won't drop any of these.
- Age and Experience seemed to be correlated so one or both has to be removed.

Model Performance Summary – Logistic Regression Model 2 (lg1)

```
Optimization terminated successfully.
         Current function value: 0.117716
         Iterations 9
                          Results: Logit
====================================================================
Model:              Logit            Pseudo R-squared: 0.618
Dependent Variable: Personal_Loan    AIC:              852.0153
Date:               2021-06-04 22:55 BIC:              938.2625
No. Observations:   3500             Log-Likelihood:   -412.01
Df Model:           13               LL-Null:          -1077.3
Df Residuals:       3486             LLR p-value:      1.4011e-276
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
--------------------------------------------------------------------
                     Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------
const               -8.4791   0.4919 -17.2357 0.0000 -9.4433 -7.5149
Experience           0.0124   0.0081   1.5412 0.1233 -0.0034  0.0282
Income               0.0586   0.0035  16.8576 0.0000  0.0518  0.0654
CCAvg                0.1894   0.0557   3.3976 0.0007  0.0801  0.2987
Mortgage             0.0012   0.0007   1.7889 0.0736 -0.0001  0.0026
Family_2            -0.0305   0.2718  -0.1123 0.9106 -0.5633  0.5023
Family_3             1.7004   0.3024   5.6238 0.0000  1.1078  2.2931
Family_4             1.6827   0.2778   6.0570 0.0000  1.1382  2.2272
Education_Grad      -0.1450   0.2253  -0.6437 0.5198 -0.5866  0.2966
Education_Undergrad -3.8322   0.3170 -12.0899 0.0000 -4.4535 -3.2110
Securities_Account_1 -0.8668  0.3685  -2.3520 0.0187 -1.5891 -0.1445
CD_Account_1         3.6340   0.4128   8.8035 0.0000  2.8250  4.4431
Online_1            -0.6431   0.2003  -3.2107 0.0013 -1.0357 -0.2505
CreditCard_1        -1.0745   0.2616  -4.1078 0.0000 -1.5872 -0.5618
====================================================================

Accuracy on train data: 0.9594285714285714
Accuracy on test data: 0.9633333333333334
Recall on train data: 0.6625386996904025
Recall on test data: 0.732484076433121
Precision on train data: 0.8663967611336032
Precision on test data: 0.8984375
f1 score on train data: 0.750877192982456
f1 score on test data: 0.8070175438596491
```

Vif Score

```
Series before feature selection:

const                   13.943249
Experience               1.009660
Income                   1.881267
CCAvg                    1.719837
Mortgage                 1.061910
Family_2                 1.385833
Family_3                 1.375617
Family_4                 1.417950
Education_Grad           1.418695
Education_Undergrad      1.455419
Securities_Account_1     1.144219
CD_Account_1             1.340403
Online_1                 1.042342
CreditCard_1             1.113011
dtype: float64
```

Observations

- ‘Age’ variable is dropped.
- The output scores did not change.
- Experience has pvalue=0.1233 so it could still be dropped due to insignificance.
- Mortgage has pvalue=0.0736 so it could still be dropped due to insignificance.
- Some variables of Family and Education are significant, so we won't drop any of these.
- None of the variables seems to be correlated, so the values in summary are reliable.

Model Performance Summary – Logistic Regression Model 3 (lg2)

```
Optimization terminated successfully.
         Current function value: 0.118057
         Iterations 9
                           Results: Logit
======================================================================
Model:               Logit              Pseudo R-squared:  0.616
Dependent Variable:  Personal_Loan      AIC:               852.3995
Date:                2021-06-04 22:55   BIC:               932.4862
No. Observations:    3500               Log-Likelihood:    -413.20
Df Model:            12                 LL-Null:           -1077.3
Df Residuals:        3487               LLR p-value:       4.2512e-277
Converged:           1.0000             Scale:             1.0000
No. Iterations:      9.0000
----------------------------------------------------------------------
                      Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------------
const               -8.2188   0.4578 -17.9526 0.0000 -9.1160 -7.3215
Income               0.0585   0.0035  16.8402 0.0000  0.0517  0.0653
CCAvg                0.1813   0.0556   3.2621 0.0011  0.0724  0.2902
Mortgage             0.0013   0.0007   1.8008 0.0717 -0.0001  0.0026
Family_2            -0.0333   0.2714  -0.1226 0.9025 -0.5652  0.4987
Family_3             1.7098   0.3023   5.6552 0.0000  1.1172  2.3023
Family_4             1.6748   0.2780   6.0247 0.0000  1.1299  2.2196
Education_Grad      -0.1373   0.2246  -0.6115 0.5409 -0.5774  0.3028
Education_Undergrad -3.8149   0.3160 -12.0719 0.0000 -4.4343 -3.1956
Securities_Account_1 -0.8739  0.3664  -2.3855 0.0171 -1.5920 -0.1559
CD_Account_1         3.6515   0.4118   8.8673 0.0000  2.8444  4.4585
Online_1            -0.6313   0.1996  -3.1624 0.0016 -1.0226 -0.2401
CreditCard_1        -1.0732   0.2615  -4.1045 0.0000 -1.5857 -0.5608
======================================================================

Accuracy on train data: 0.9585714285714285
Accuracy on test data: 0.9626666666666667
Recall on train data: 0.6470588235294118
Recall on test data: 0.7197452229299363
Precision on train data: 0.8708333333333333
Precision on test data: 0.904
f1 score on train data: 0.7424511545293072
f1 score on test data: 0.8014184397163121
```

Observations

- ‘Age’ and ‘Experience’ variables have been dropped.
- The precision score has improved by 1% point to 87.1% for train data and 90.4% for test data.
- The recall and f1 score dipped slightly and accuracy score remained the same.
- Mortgage has pvalue=0.0717 so it could still be dropped due to insignificance.
- Some variables of Family and Education are significant, so we won't drop any of these.

Model Performance Summary – Logistic Regression Model 4 (lg3)

```
Optimization terminated successfully.
         Current function value: 0.118517
         Iterations 9
                          Results: Logit
====================================================================
Model:              Logit            Pseudo R-squared: 0.615
Dependent Variable: Personal_Loan    AIC:              853.6211
Date:               2021-06-04 22:55 BIC:              927.5473
No. Observations:   3500             Log-Likelihood:   -414.81
Df Model:           11               LL-Null:          -1077.3
Df Residuals:       3488             LLR p-value:      1.8717e-277
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
--------------------------------------------------------------------
                      Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------
const                -8.1757   0.4570 -17.8915 0.0000 -9.0713 -7.2800
Income                0.0590   0.0035  17.0159 0.0000  0.0522  0.0658
CCAvg                 0.1697   0.0550   3.0823 0.0021  0.0618  0.2775
Family_2              0.0058   0.2699   0.0216 0.9828 -0.5232  0.5348
Family_3              1.7206   0.3042   5.6572 0.0000  1.1245  2.3168
Family_4              1.7040   0.2784   6.1200 0.0000  1.1583  2.2497
Education_Grad       -0.1399   0.2243  -0.6237 0.5328 -0.5794  0.2997
Education_Undergrad  -3.7567   0.3114 -12.0630 0.0000 -4.3670 -3.1463
Securities_Account_1 -0.8803   0.3677  -2.3943 0.0167 -1.6009 -0.1597
CD_Account_1          3.6836   0.4124   8.9332 0.0000  2.8754  4.4918
Online_1             -0.6335   0.1993  -3.1793 0.0015 -1.0240 -0.2430
CreditCard_1         -1.0868   0.2605  -4.1724 0.0000 -1.5974 -0.5763
====================================================================

Accuracy on train data: 0.9582857142857143
Accuracy on test data: 0.964
Recall on train data: 0.653250773993808
Recall on test data: 0.7197452229299363
Precision on train data: 0.8612244897959184
Precision on test data: 0.9186991869918699
f1 score on train data: 0.7429577464788731
f1 score on test data: 0.8071428571428572
```

Observations

- ‘Age’, ‘Experience’ and ‘Mortgage’ variables have been dropped.
- The accuracy score for test data improved to reach 96.4%.
- The precision score for test data improved to reach 91.87%.
- The recall and f1 score stayed the same.
- All variables are significant.
- Some variables of Family and Education are significant, so we won't drop any of these.

Model Performance Summary – Logistic Regression Model 4 (lg3)

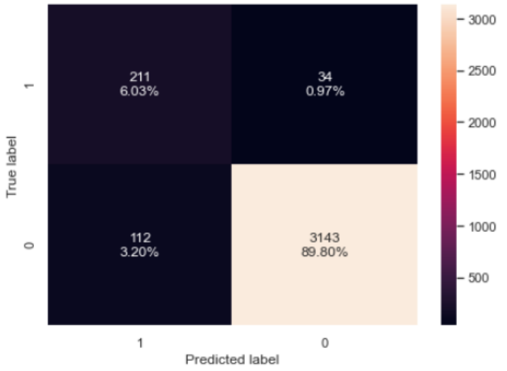
Odds from Coefficients

	odds
const	0.000281
Income	1.060735
CCAvg	1.184907
Family_2	1.005842
Family_3	5.588106
Family_4	5.495865
Education_Grad	0.869468
Education_Undergrad	0.023361
Securities_Account_1	0.414659
CD_Account_1	39.790847
Online_1	0.530732
CreditCard_1	0.337281

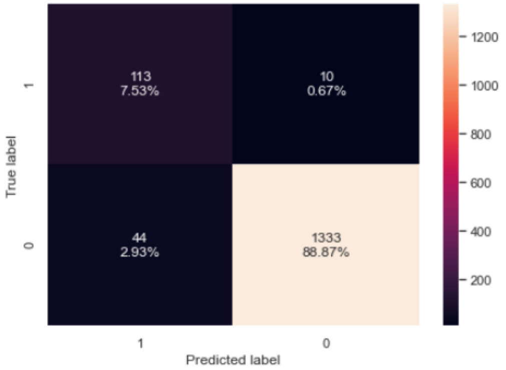
% Change in Odds

	change_odds%
const	-99.971858
Income	6.073488
CCAvg	18.490686
Family_2	0.584151
Family_3	458.810597
Family_4	449.586465
Education_Grad	-13.053202
Education_Undergrad	-97.663859
Securities_Account_1	-58.534141
CD_Account_1	3879.084683
Online_1	-46.926782
CreditCard_1	-66.271874

Training data Confusion Matrix



Test data Confusion Matrix

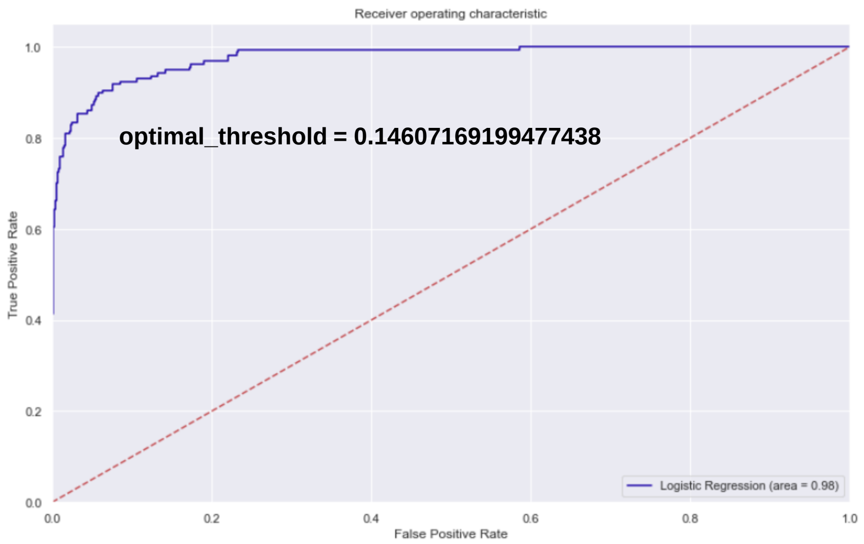


Conclusion

- lg3 is the final model that we will use for predictions and inferences.
- Income, CCAvg, Family, Education, Securities accounts, CD accounts, Online and CreditCard are important variables here.
- All coefficients are positive except for the Education variables, Securities accounts, Online and CreditCard variables.
- CD_Account variable has the most significant positive influence in target variable, increasing odds by up to 3879% of taking personal loan, so there is a high chance marketing personal loan products to CD account holders will yield convertable customers.
- Education_Undergrad variable has the most significant negative influence in target variable , decreasing odds by 97.7%, so there is a very low chance undergrads will take up a personal loan.
- Larger family sizes (Family_3,Family_4) also increase the odds significantly of taking up a personal loan.
- Using a credit card issued from other banks (CreditCard), having a securities account (Securities_Account) and using the bank online facilities (Online) also decrease the odds significantly of taking up a personal loan with the bank.
- Please note that when coefficient is b , than change in odds is (exp(b)-1)*100 %

Model Performance Summary – Improve Model using AUC-ROC curve

AUC-ROC Curve

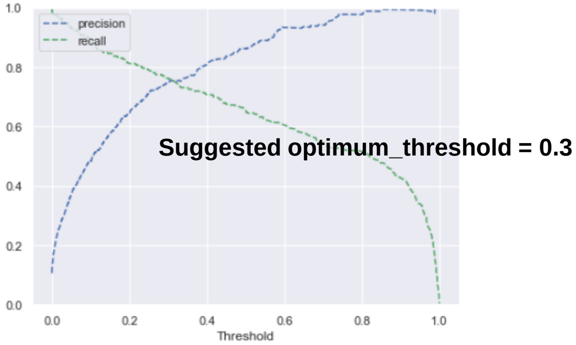



```
Accuracy on train data: 0.9274285714285714
Accuracy on test data: 0.9226666666666666
Recall on train data: 0.8452012383900929
Recall on test data: 0.910828025477707
Precision on train data: 0.5723270440251572
Precision on test data: 0.5836734693877551
f1 score on train data: 0.6824999999999999
f1 score on test data: 0.7114427860696517
```

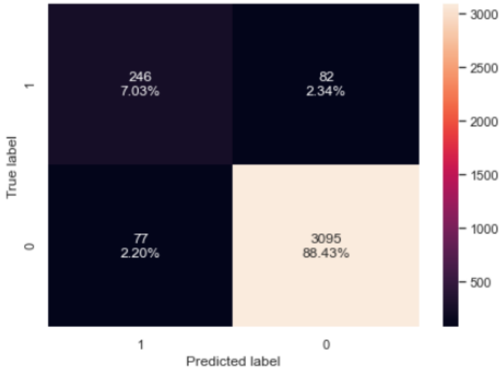
Observations
- lg3 is the final model that we will use for predictions and inferences.
- Using Optimal Threshold from the AUC-ROC curve on lg3 unfortunately yielded a poorer model than the original lg3, with only recall scores improving and other scores performing poorer especially the f1 score.

Model Performance Summary – Improve Model using Precision-Recall curve

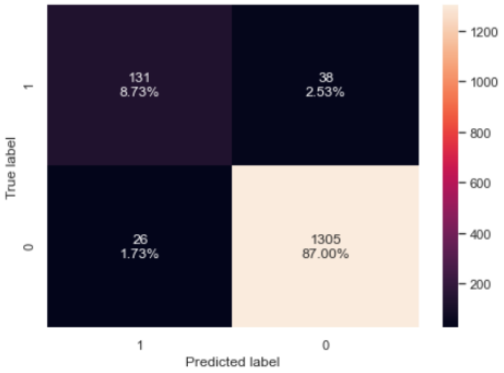
Precision-Recall Curve



Training data Confusion Matrix



Test data Confusion Matrix



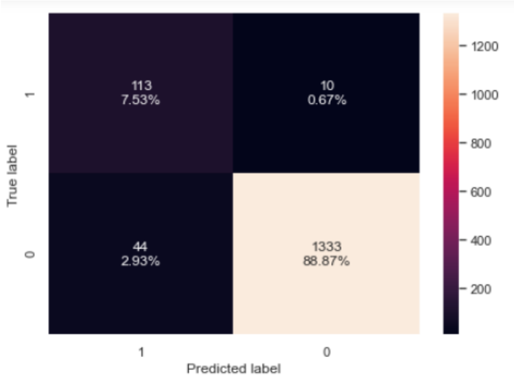

```
Accuracy on train data: 0.9545714285714286
Accuracy on test data: 0.9573333333333334
Recall on train data: 0.7616099071207431
Recall on test data: 0.8343949044585988
Precision on train data: 0.75
Precision on test data: 0.7751479289940828
f1 score on train data: 0.7557603686635945
f1 score on test data: 0.8036809815950919
```

Observations

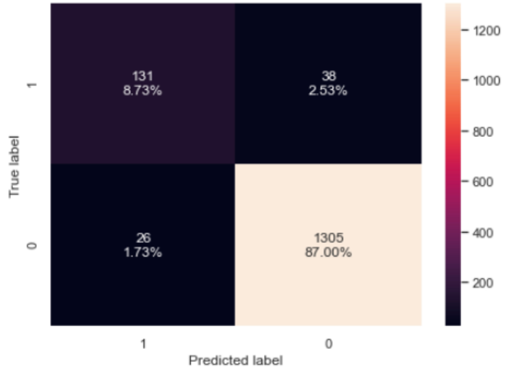
- A good f1 score requires a highest possible precision and recall score combination which suggests a **0.3 optimum threshold** from the Precision-Recall curve intersection.
- Accuracy scores for train and test data are closer than the original lg3 model.
- Recall scores are higher but precision scores dipped.
- F1 scores stayed the largely the same.

Model Performance Summary – Improve Model using Precision-Recall curve

Lg3 original model
Test data Confusion Matrix



Lg3 model with optimum threshold = 0.3
Test data Confusion Matrix



Observations

- lg3 model with optimum threshold = 0.3 is preferable because even though precision is lower, the % of True Positives achieved by the second model is higher which means proportion of true personal loan customers were marketed to.

Conclusion and Recommendations

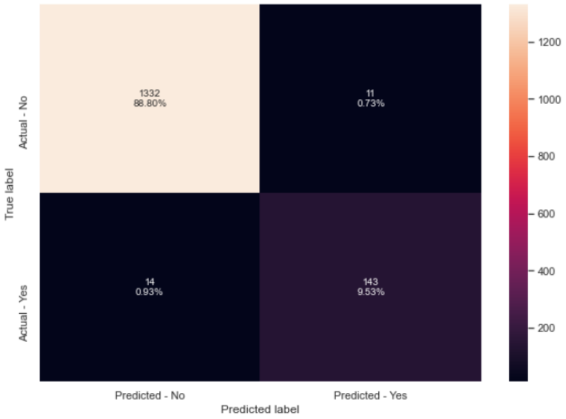
- The best test recall is 83% but the test precision is lower at 77.5%. This means that the model is not as good at identifying potential personal loan takers than identifying non personal loan takers so therefore the bank can lose many opportunities of marketing personal loan to would be customers.
- The model performance can be improved, especially in terms of precision and the bank can use the model for new customers once desired level of model performance is achieved.
- The analysis showed that customers with CD accounts, larger family sizes of above 2 are more likely to accept personal loans. More marketing effort can be focused on them.
- It also showed that undergraduate customers, customers who use a credit card from other banks, have a securities account or use bank online facilities are less likely to accept a personal loan from the bank. Less marketing effort can be spent on them.

Model Performance Summary – Decision Trees

- Model Evaluation Criterion
  - The scoring criteria shall be the same as the logistic regression model.
  - We can use accuracy but since the data is imbalanced it would not be the right metric to check the model performance.
  - Therefore, f1_score should be maximized, the greater the f1_score higher the chances of identifying both the classes correctly.

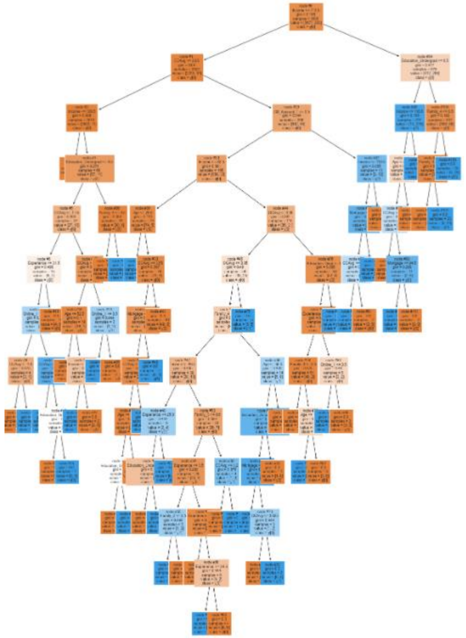
Model Performance Summary – Decision Tree Model 1

Test data Confusion Matrix

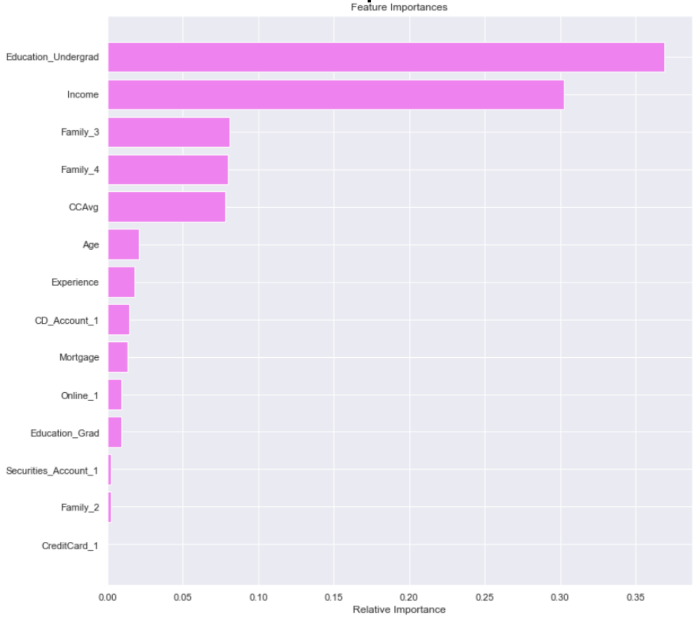



```
Accuracy on training set :  1.0
Accuracy on test set :  0.9833333333333333
Recall on training set :  1.0
Recall on test set :  0.910828025477707
Precision on training set :  1.0
Precision on test set :  0.9285714285714286
f1 score on training set :  1.0
f1 score on test set :  0.9196141479099679
```

Decision Tree Model 1



Feature Importance

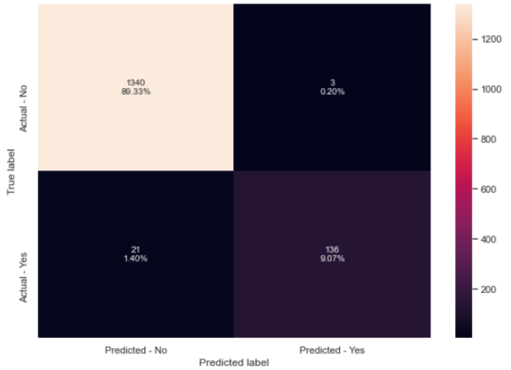


Observations

- The scores indicate a complex tree that overfits the training data and scores for training and testing data are not close.
- F1 scores for training and testing data are 100% and 91.96% which is quite good.
- According to the decision tree model, Education_Undergrad and Income are the most important variables for predicting the customer personal loan acceptance.
- The tree above is very complex, such a tree often overfits.

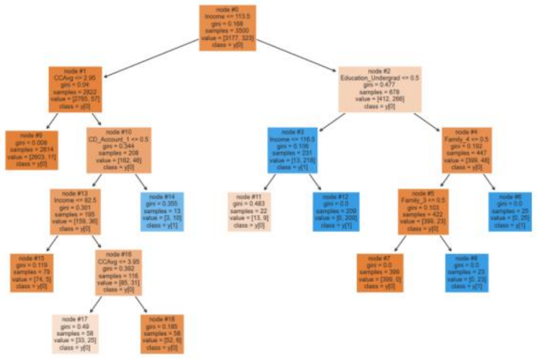
Model Performance Summary – Decision Tree Model 2 Pre-Pruning: Using GridSearch for Hyperparameter tuning

Test data Confusion Matrix

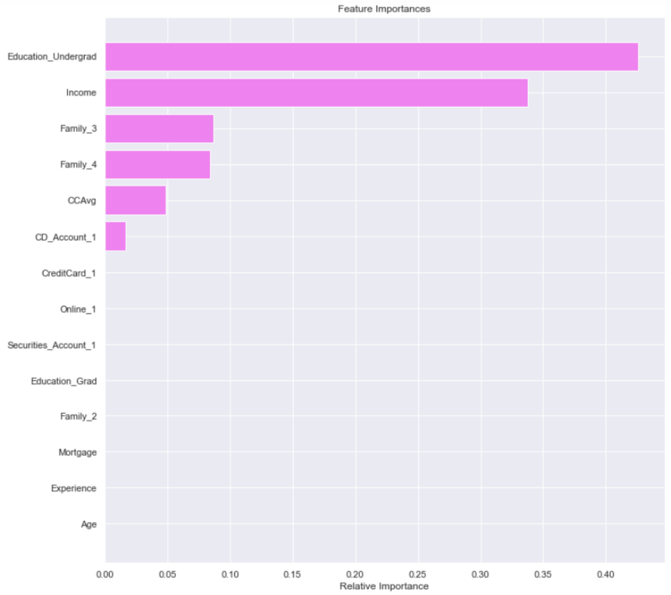



```
Accuracy on training set :  0.9831428571428571
Accuracy on test set :  0.984
Recall on training set :  0.826625386996904
Recall on test set :  0.8662420382165605
Precision on training set :  0.9888888888888889
Precision on test set :  0.9784172661870504
f1 score on training set :  0.9005059021922428
f1 score on test set :  0.918918918918919
```

Decision Tree Model 2



Feature Importance



Observations

- The scores indicate a more generalized tree that has closer scores for both training and testing data.
- F1 scores for training and testing data are 90% and 91.9% which is better in terms of proximity and consistency compared to Decision Tree Model 1.
- According to the decision tree model, Education_Undergrad and Income are still the most important variables for predicting the customer personal loan acceptance.
- Only up to CD_Account_1 does it still have importance in the model, several variables are no longer important.

Model Performance Summary – Decision Tree Model 3 Post-Pruning: Cost Complexity Pruning

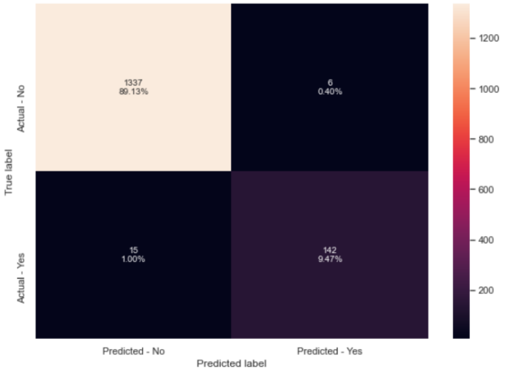

```
DecisionTreeClassifier(ccp_alpha=0.0008340087585370598, random_state=1)
```

Observations

- Minimal cost complexity pruning is started by plotting the cost complexity pruning path that returns the effective alphas and the corresponding total leaf impurities at each step of the pruning process. As alpha increases, more of the tree is pruned, which increases the total impurity of its leaves. This is plotted on the above left diagram.
- Next, we train a decision tree using the effective alphas. The last value in ccp_alphas is the alpha value that prunes the whole tree, leaving the tree, clfs[-1], with one node.
- A plot of the f1 score Vs the ccp_alphas pruning path is plotted for both training and test data on the top right. To gain the highest training and test f1 score, ccp_alpha should be = 0.0008340087585370598.

Model Performance Summary – Decision Tree Model 3 Post-Pruning: Cost Complexity Pruning

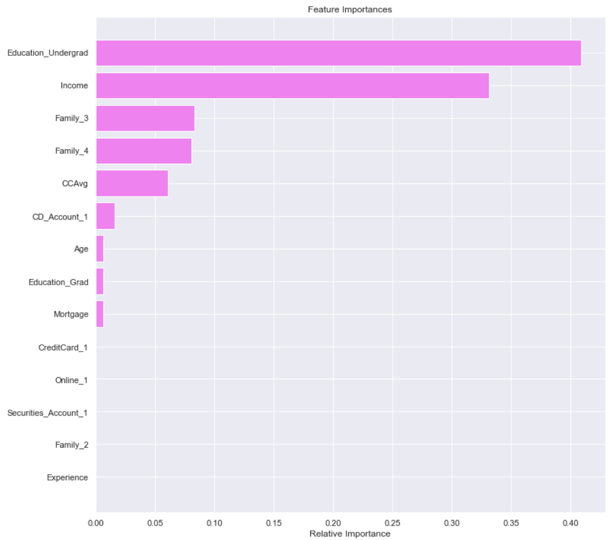
Test data Confusion Matrix




```
f1 score on training set :  0.9293924466338259
f1 score on test set :  0.9311475409836065
```

Decision Tree Model 3

Feature Importance



Observations

- With post-pruning we get the highest f1 score on both training and test set at 92.94% and 93.11% in Decision Tree Model 3, which is both higher and closer than that of Decision Tree Model 2 using Pre-Pruning Grid Search CV at 90% and 91.9%.
- Education_Undergrad and Income are still the most important variables for predicting the customer personal loan acceptance.
- Beyond CD_Account_1; Age, Education_Grad and Mortgage still have some importance in the model as compared to Decision Tree Model 2 pre-pruned with Grid Search CV.

Model Performance Summary – Performance Metrics

	Model	Train_f1_Score	Test_f1_Score
0	lg3 model	0.7430	0.8071
1	lg3 model with AUC-ROC curve enhancement	0.6825	0.7114
2	lg3 model with Precision-Recall curve enhancement	0.7558	0.8037
3	Initial decision tree model	1.0000	0.9196
4	Decision treee with hyperparameter tuning	0.9005	0.9189
5	Decision tree with post-pruning	0.9294	0.9311

Conclusion and Recommendations

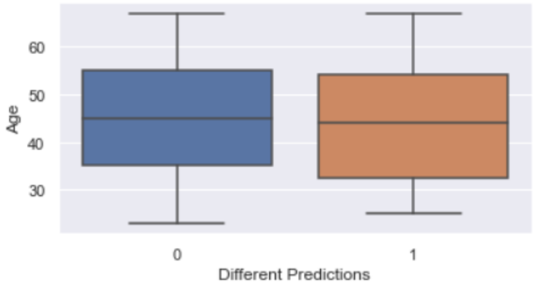
- Decision tree with post-pruning is giving the highest f1 score on test set.
- Education_Undergrad and Income are still the most significant variables for predicting the customer personal loan acceptance followed by larger family size of above 2 and credit card spending.

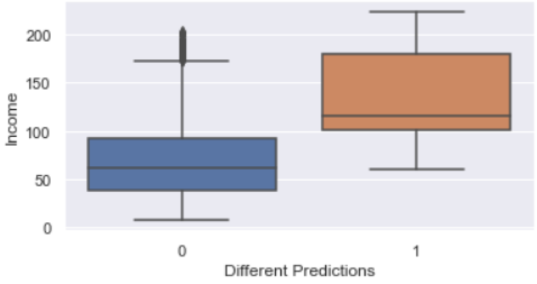
Model Performance Summary – EDA on incorrectly predicted data

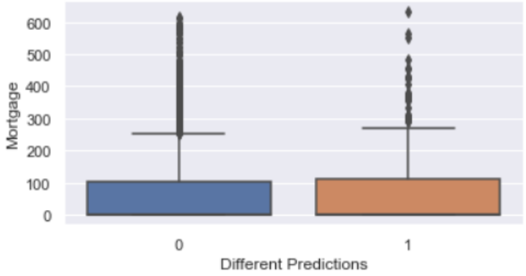
- Only the best model in logistic regression (lg3 model with Precision-Recall curve enhancement) and decision trees (Decision tree with post-pruning) are included for analysis.

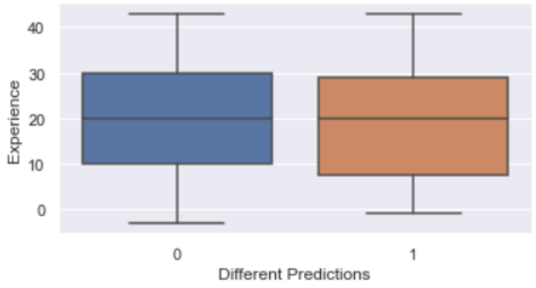
- The training and test data are joined and appended with predicted personal loan and actual loan values as well as a label to indicate where there are differences named ‘Different Predictions’.

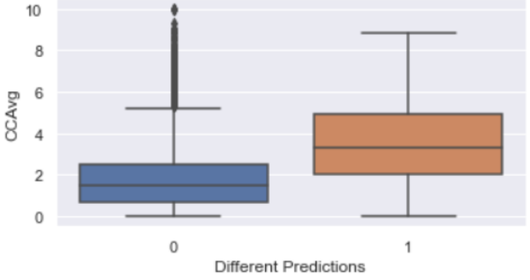
Model Performance Summary – EDA on incorrectly predicted data (LG3 Model)









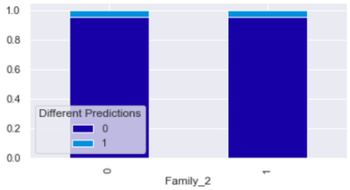


Observations

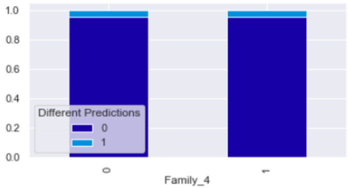
- Missed predictions tend to be of higher income and credit card spending ranges.

Model Performance Summary – EDA on incorrectly predicted data (LG3 Model)

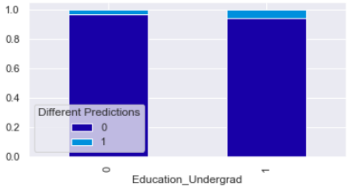
Different Predictions	0	1	All
Family_2			
0	3544	160	3704
1	1233	63	1296
All	4777	223	5000



Different Predictions	0	1	All
Family_4			
0	3614	164	3778
1	1163	59	1222
All	4777	223	5000

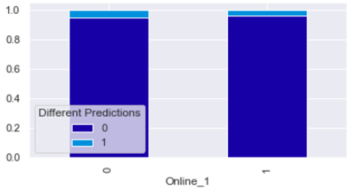


Different Predictions	0	1	All
Education_Undergrad			
0	2799	105	2904
1	1978	118	2096
All	4777	223	5000

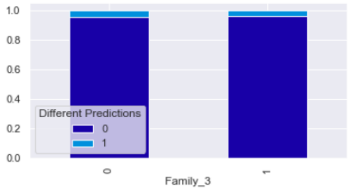


Different Predictions	0	1	All
Securities_Account_1			
0	4279	199	4478
1	498	24	522
All	4777	223	5000

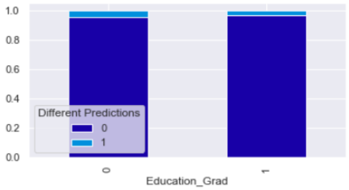
Different Predictions	0	1	All
Online_1			
0	1907	109	2016
1	2870	114	2984
All	4777	223	5000



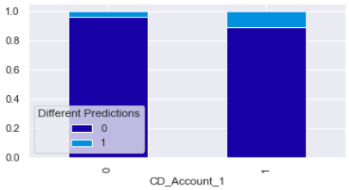
Different Predictions	0	1	All
Family_3			
0	3805	185	3990
1	972	38	1010
All	4777	223	5000



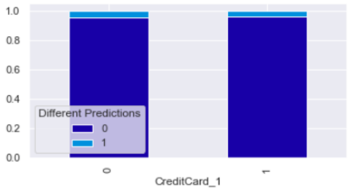
Different Predictions	0	1	All
Education_Grad			
0	3422	175	3597
1	1355	48	1403
All	4777	223	5000



Different Predictions	0	1	All
CD_Account_1			
0	4509	189	4698
1	268	34	302
All	4777	223	5000



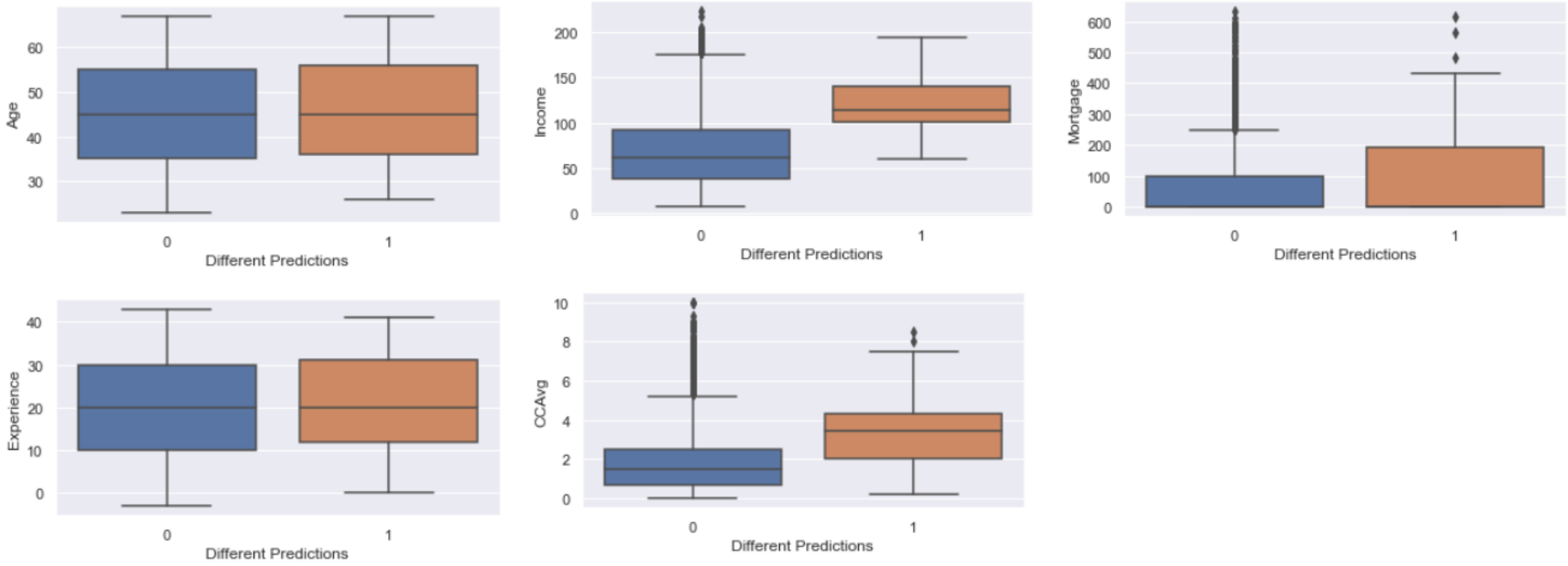
Different Predictions	0	1	All
CreditCard_1			
0	3368	162	3530
1	1409	61	1470
All	4777	223	5000



Observations

- Missed predictions tend to be those with CD_Accounts(about 10% miss) and Undergrad customers (about 5% miss).

Model Performance Summary – EDA on incorrectly predicted data (Decision Tree Model)

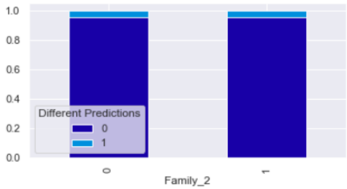


Observations

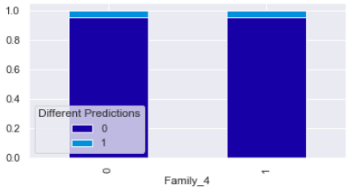
- Missed predictions tend to be of higher income, credit card spending ranges and higher mortgages.

Model Performance Summary – EDA on incorrectly predicted data (Decision Tree Model)

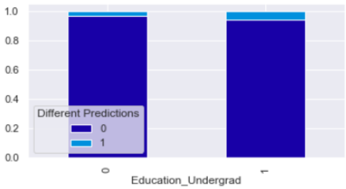
Different Predictions Family_2	0	1	All
0	3544	160	3704
1	1233	63	1296
All	4777	223	5000



Different Predictions Family_4	0	1	All
0	3614	164	3778
1	1163	59	1222
All	4777	223	5000

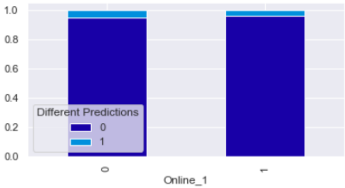


Different Predictions Education_Undergrad	0	1	All
0	2799	105	2904
1	1978	118	2096
All	4777	223	5000

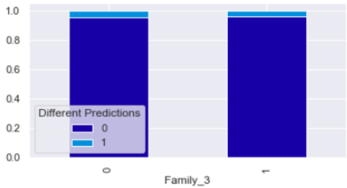


Different Predictions Securities_Account_1	0	1	All
0	4279	199	4478
1	498	24	522
All	4777	223	5000

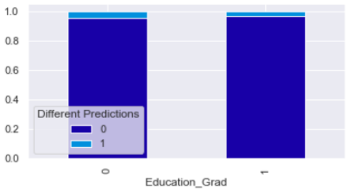
Different Predictions Online_1	0	1	All
0	1907	109	2016
1	2870	114	2984
All	4777	223	5000



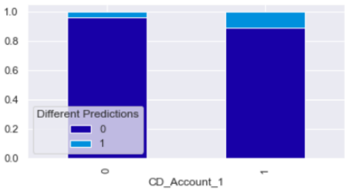
Different Predictions Family_3	0	1	All
0	3805	185	3990
1	972	38	1010
All	4777	223	5000



Different Predictions Education_Grad	0	1	All
0	3422	175	3597
1	1355	48	1403
All	4777	223	5000



Different Predictions CD_Account_1	0	1	All
0	4509	189	4698
1	268	34	302
All	4777	223	5000



Different Predictions CreditCard_1	0	1	All
0	3368	162	3530
1	1409	61	1470
All	4777	223	5000

Observations

- Missed predictions tend to be those with CD_Accounts(about 10% miss) and Undergrad customers (about 5% miss).

Business Insights and Recommendations

- Based on the performances of the different classification models, decision tree with post pruning performed the best using f1_score as the deciding factor due to the unbalanced data.
  - Significant variables include ‘Education’, ‘Income’, ‘Family Size’ and Credit Card Spending.
  - Coupled with EDA insights, potential Personal loan customers tend to be of graduates and above in education, higher income holders, larger family size above 2 and higher credit card spending patterns.
  - Less significant variables include ‘CD_Account’, ‘Age’ and ‘Mortgage’ value.

- Comments on additional data sources for model improvement
  - Additional data can be obtained from feedback of marketing efforts to the public to strengthen the model.
  - Feedback can be gathered from non-personal loan customer converts for further analysis

Business Insights and Recommendations

- Model implementation in real world and potential business benefits from model.
    - The model implemented in the real world will help to raise more successful targeted marketing converts of its campaign and reduce the costs of marketing to potential non-converts or miss target marketing to potential converts. This will increase revenue and reduce both variable marketing costs and opportunity costs.

greatlearning
Power Ahead

Happy Learning !