Personal Loan Campaign Modelling

Nibu Kuriakose

May 2021

Content

1. Introduction and Proposed Approach

Background

- AllLife bank only have a low number of its existing customers taking personal loans.
- The bank wants more of its existing customers to take loan to increase revenue from interest on loans.
- Bank's campaign last year show a conversion rate of 9%.
- Retail marketing department wants to run better targeted marketing campaign with higher conversion ratio.

Purpose and Benefits

- Perform exploratory data analysis to derive insights and recommendations.
- Create a model which will help the bank to identify potential customers who will take a personal loan.
- This will allow for an effective targeted marketing which uses minimum resources to get maximum number of customers.

Proposed Approach

1. Build prediction models using the following methods
   a) Logistic regression b) Decision Tree
2. Find the appropriate performance measure to evaluate the models based on the rationale that the bank do not want to miss any potential loan customers during the campaign
3. Perform model improvement to get the model which gives the best performance against the measure identified above.
4. Select the model which gives the best performance.
5. After the initial campaign, use the additional data gathered to improve the model for the future campaigns.

2. Dataset Information

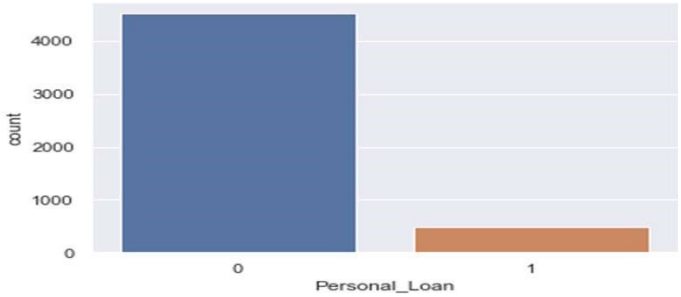
- There are 5,000 samples
- Each sample has 14 attributes
- It is an imbalanced dataset as only 9.6% of customers have taken personal loan which is the target dependent variable for building the model. Hence, a class weight has been used to compensate for this.
- Feature Engineering- Zip code was used to find the state, city and county.

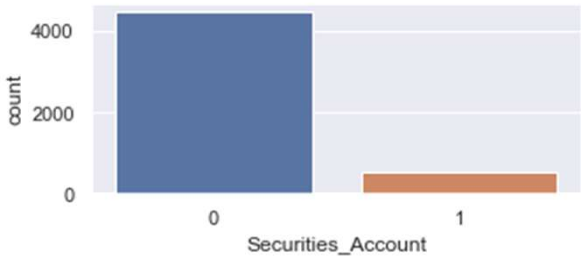
3.1 Univariate Analysis

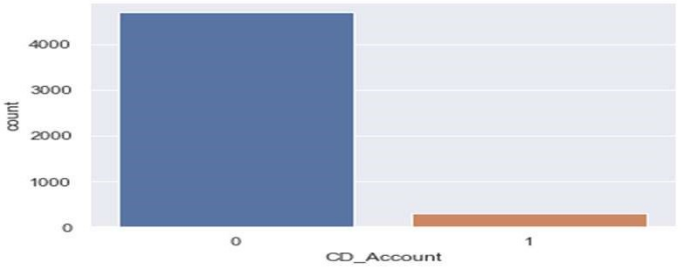
Attribute	Description
Age	• No skewness, with mean and median almost the same
Experience	• No skewness, with mean and median almost the same
Income	• Right skewed with a number of outliers and 75% of customers below $98k
Family	• Most of the customers are single, with the maximum family size of 4
CCAvg	• Right skewed with a number of outliers, with 75% of customers having average monthly credit card spend less than $2,500
Education	• Most customers are undergrads followed by holders of advance/professional education.
Mortgage	• Right skewed data. 69% have no mortgage and 75% under $101k of mortgage.

	Age	Experience	Income	Family	CCAvg	Education	Mortgage
count	5000.0	5000.0	5000.0	5000.0	5000.0	5000.0	5000.0
mean	45.3	20.1	73.8	2.4	1.9	1.9	56.5
std	11.5	11.5	46.0	1.1	1.7	0.8	101.7
min	23	-3	8	1	0	1	0
25%	35	10	39	1	0.7	1	0
50%	45	20	64	2	1.5	2	0
75%	55	30	98	3	2.5	3	101
max	67	43	224	4	10	3	635

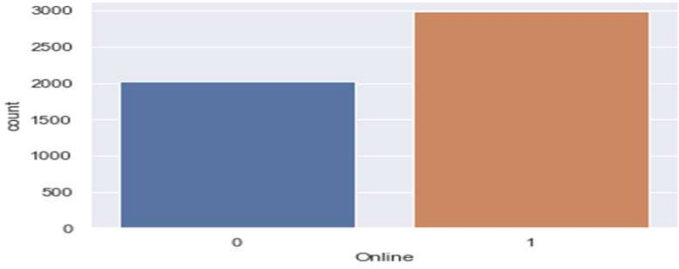
3.1 Univariate Analysis

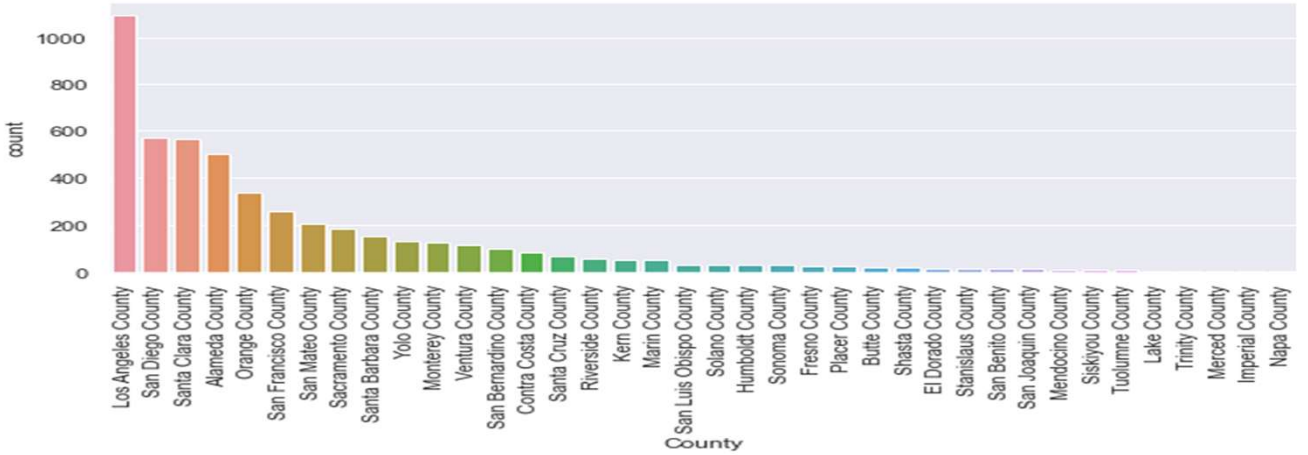




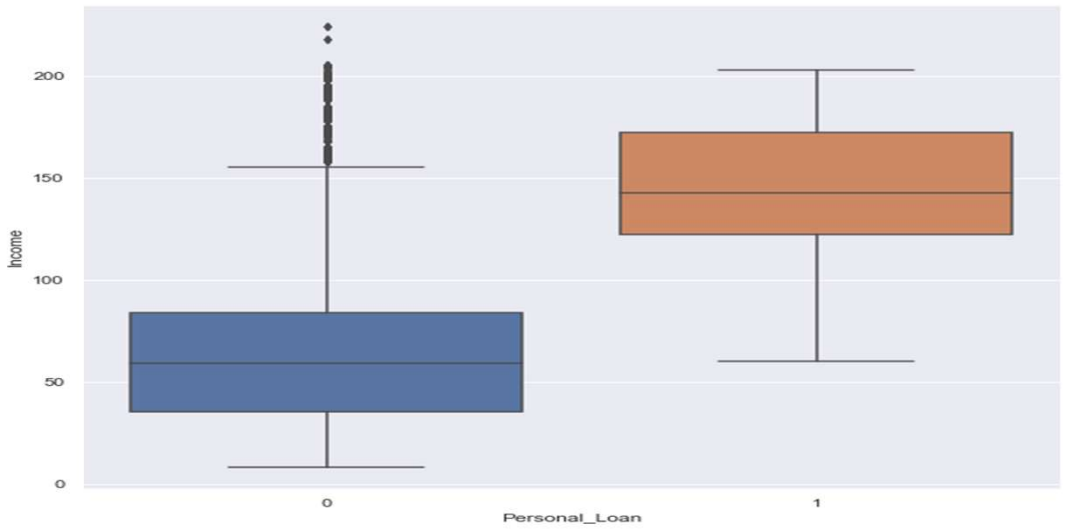


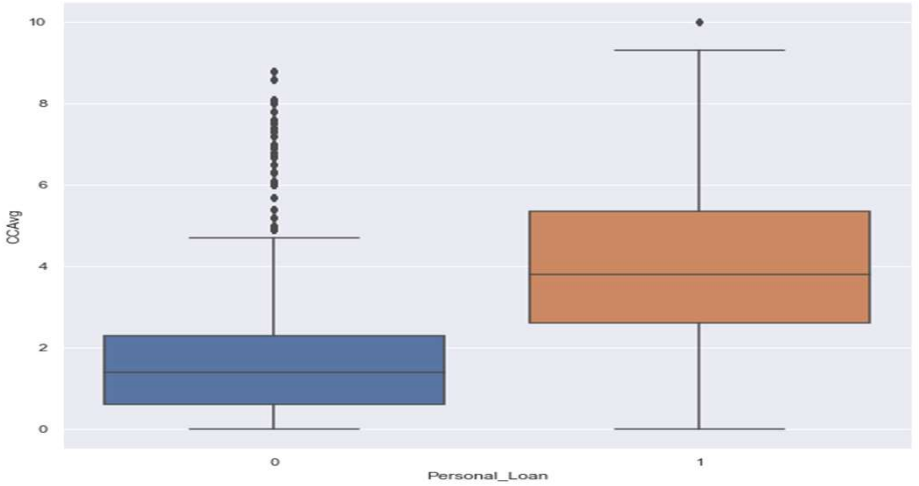
- Only 9.6% of the customers have taken personal loan
- Low number of customers have used securities account, CD_Account and credit card.
- 60% of customers use internet banking facilities
- 4 counties (LA, San Diego, Santa Clara and Alemada) cover more than 50% of the customers

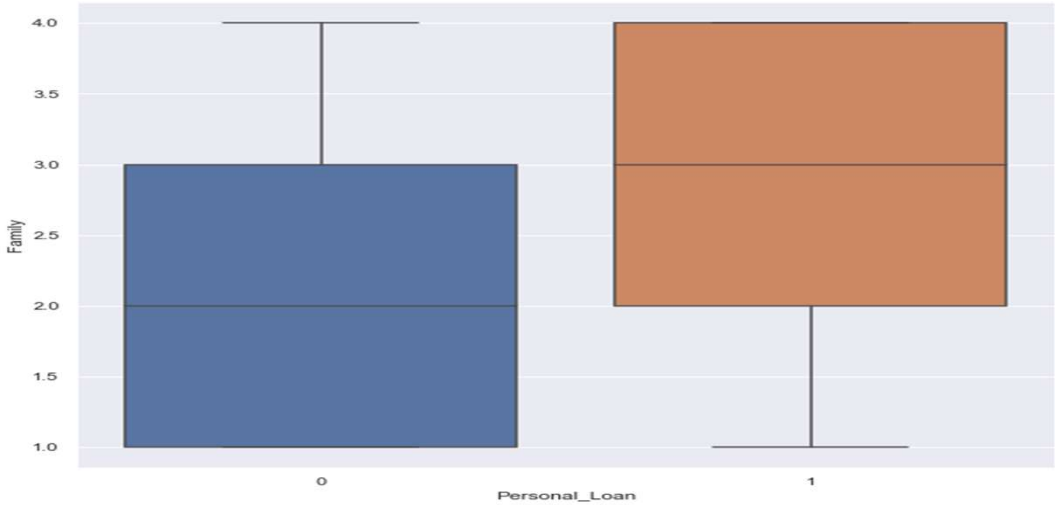




3.2 Personal Loan- Key Factors (I/IV)

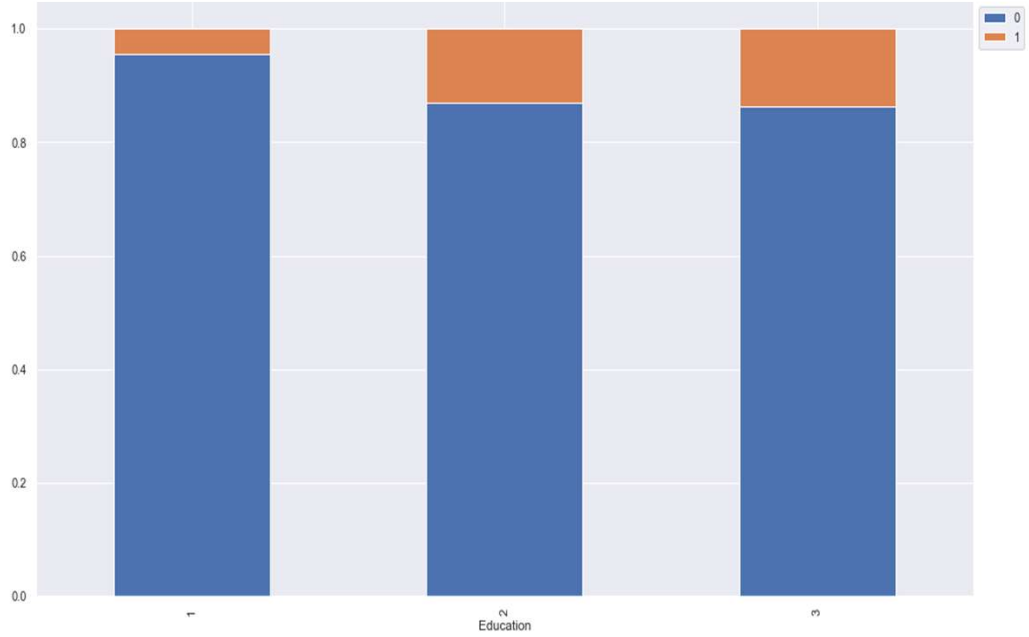


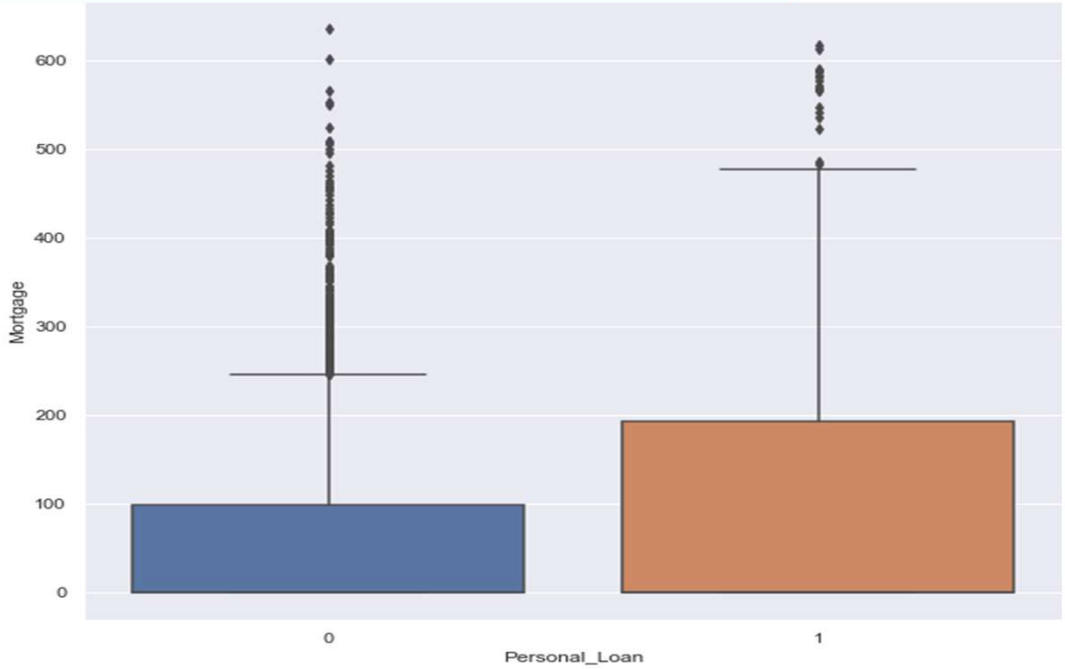




- Customers who have taken person loan have the following compared to customers who haven't:
  - higher income
  - higher monthly credit card spend
  - higher number of family members

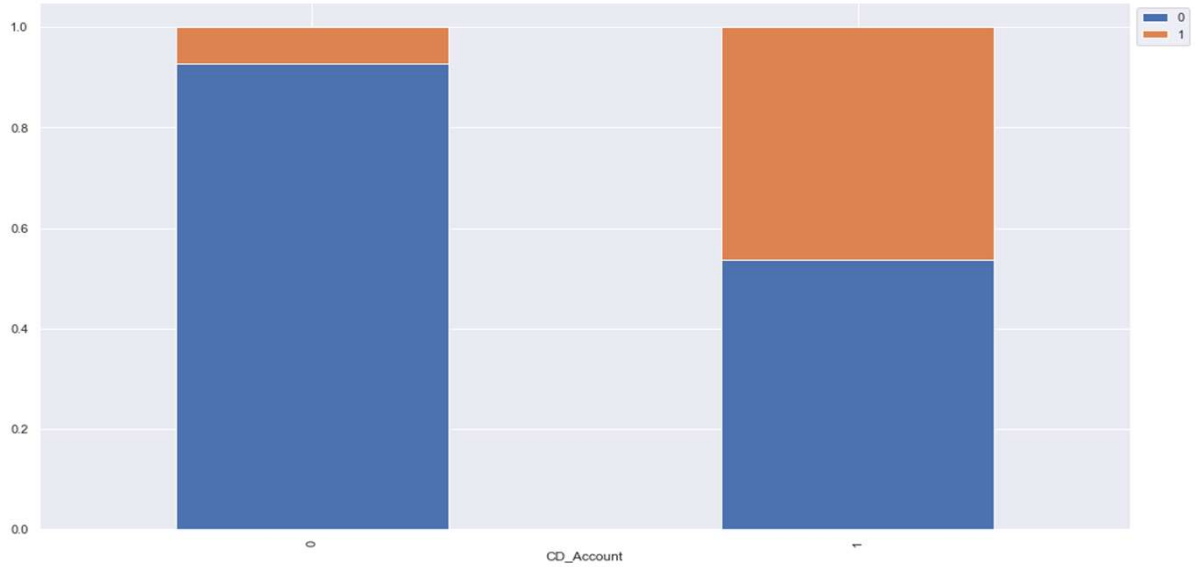
3.2 Personal Loan- Key Factors (II/IV)

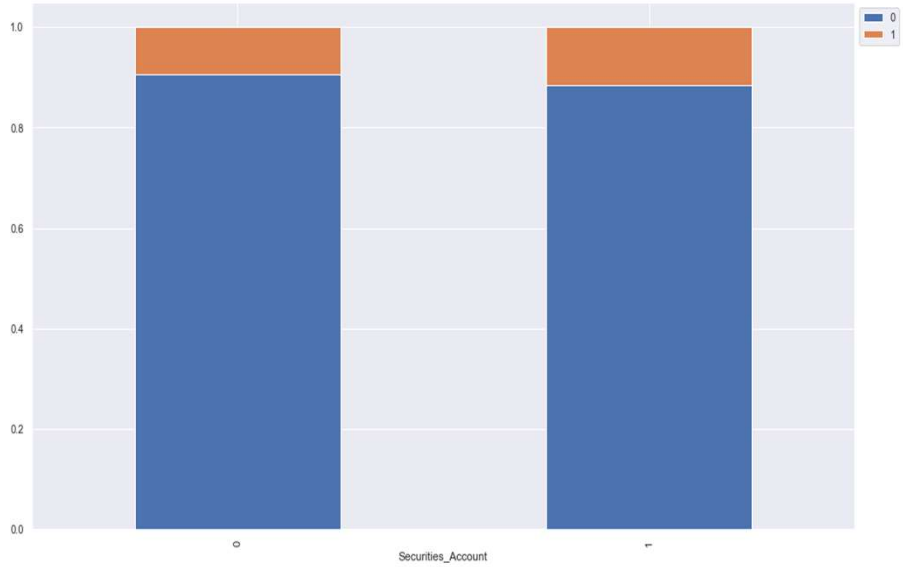




- Higher proportion of customers with graduation and advanced/professional education have taken personal loan compared to customers with undergrad education.
- Customers who have taken the personal loan have higher mortgage.

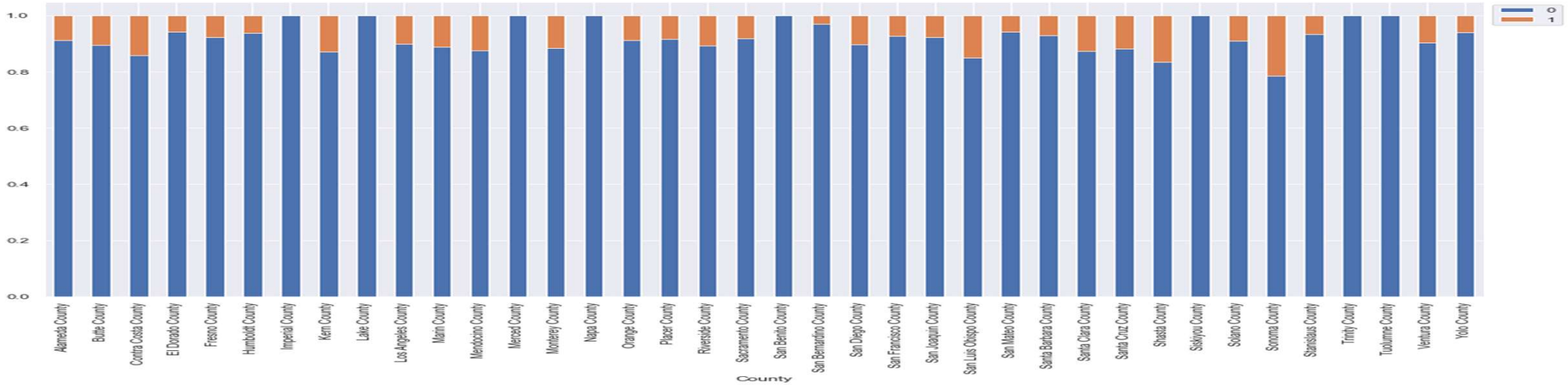
3.2 Personal Loan- Key Factors (III/IV)

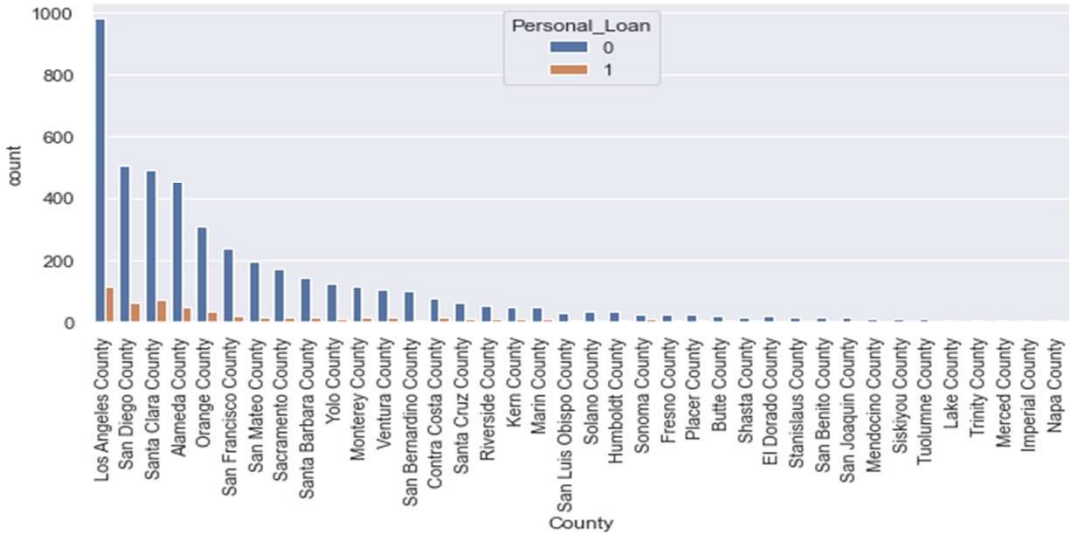




- Higher proportion of customers with certificate of deposit account and securities account have taken personal loan compared to customers who doesn't have these accounts.
- Following factors does not vary between customers who have taken a personal loan and who haven't:
  - Age and experience
  - Possessing credit card
  - Use of internet banking facilities

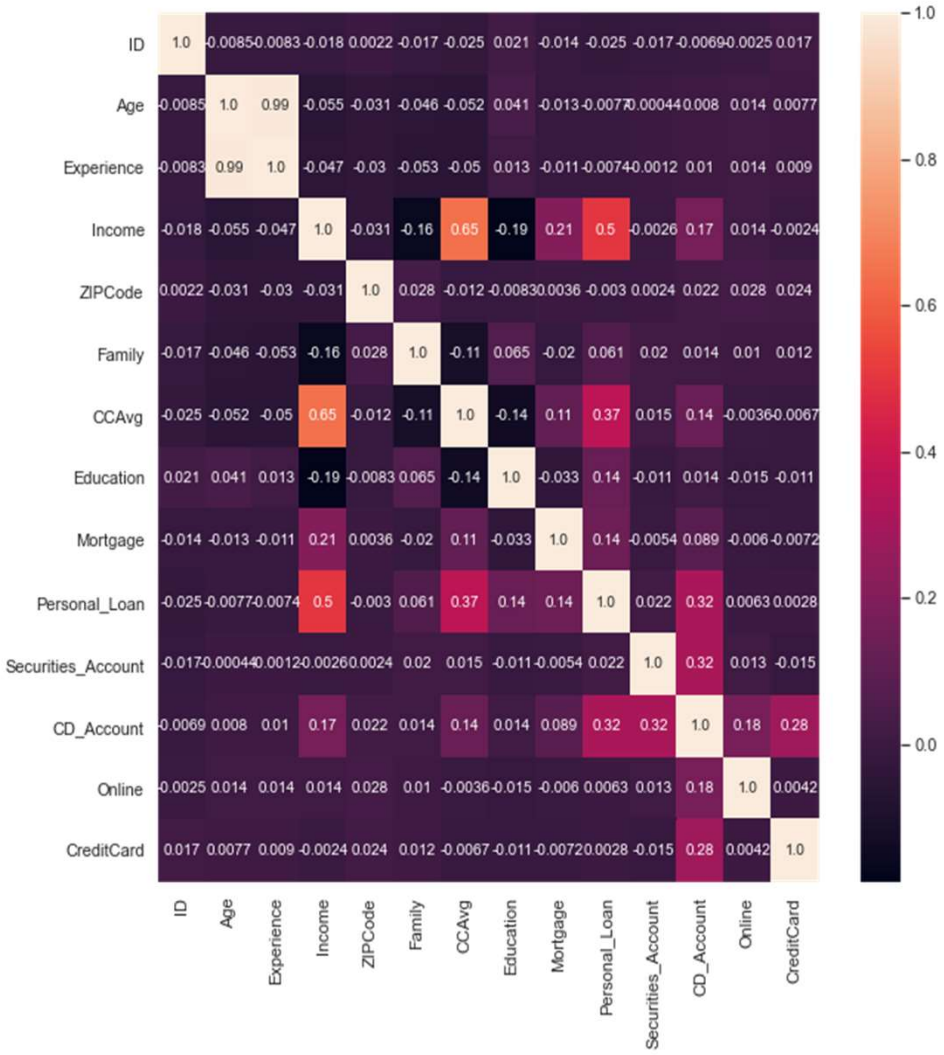
3.2 Personal Loan- Key Factors (IV/IV)





- Proportion of customers who take personal loan varies by counties, with no customers in 8 counties taking personal loan.

3.3 Multivariate Analysis



- Customers taking personal loan is highly positively correlated to income and moderately with credit card spend and having a CD account.
- Age and experience are highly positively correlated
- Income is positively correlated with mortgage, credit card spend and having a CD account
- Education and income slightly negatively correlated
- Having a CD account and credit card are positively correlated with having a Securities account, using internet banking and credit card.

4.1 Model Building- Overview

- Both 'logistic regression' and 'decision tree' modelling were used to build models to predict whether a customer would take a personal loan.
- Recall is the performance measure used to compare the performance of models in this scenario rather than accuracy.
  - ✓ Recall measures what proportion of the customers who take personal loan can be identified.
  - ✓ Accuracy is less important as it only the proportion of cases which are correctly identified.
  - ✓ The cost of not being able to rightly identify a customer who will take a loan is high which is measured using recall.
- Following is the approach for Logistic regression:
  - ✓ A base model is built and the performance is measured with the threshold probability as 0.5
  - ✓ The optimal cut off where tpr is high and fpr is low is determined.
  - ✓ Performance of the model is measured using the above optimal cut-off.
  - ✓ Model is further optimised by removing variables causing multicollinearity and those which are less significant.
- Following is the approach for Decision tree modelling:
  - ✓ Base decision tree is built with no limits which is expected to overfit.
  - ✓ Overfitting is minimised by following methods to
    - restricting the tree using a maximum depth limit
    - Finding optimal hyperparameters
    - Find model which gives best recall using cost complexity method.

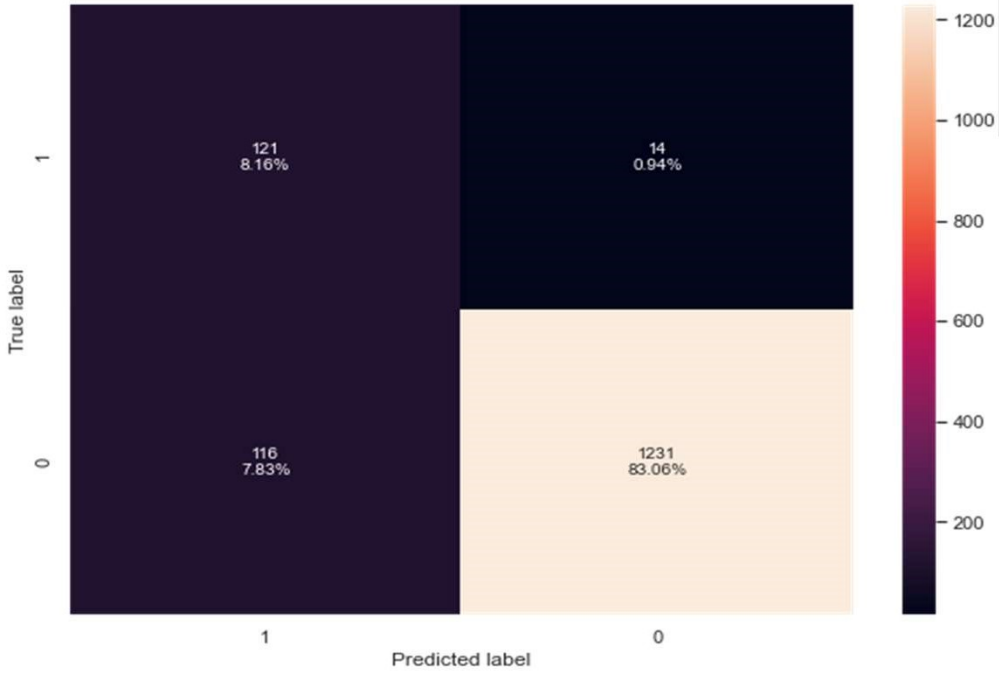
4.2 Logistic Regression

Variable	coef	Odds_ratio	probability
CD_Account	3.62	37.19	0.97
Education	1.4	4.04	0.8
Family	0.73	2.08	0.68
CCAvg	0.32	1.38	0.58
Income	0.05	1.05	0.51
Online	-0.59	0.56	0.36
CreditCard	-0.98	0.37	0.27
Securities_Account	-1.06	0.35	0.26

Key Parameters

Following are the key parameters influencing whether a customer takes a personal loan or not.

- Increase in **education, family, credit card spend or income** by an unit **increases the probability** of taking personal loan by 80%, 68%, 58% and 51% respectively.
- Possessing a **checking deposit account increases probability** of taking a personal loan by 97%
- Possessing an **internet banking facility, credit card or security account** shows a **lower probability** of 36%, 27% and 26% respectively of taking a personal loan.



	Training Data	Test Data
Model with threshold as 0.5		
Accuracy	95.2	95.8
Recall	64.6	67.4
Model with optimal threshold		
Accuracy	90.7	91.2
Recall	88.3	89.6

Model with optimal threshold applied gives higher recall for test data.

4.3 Decision Tree

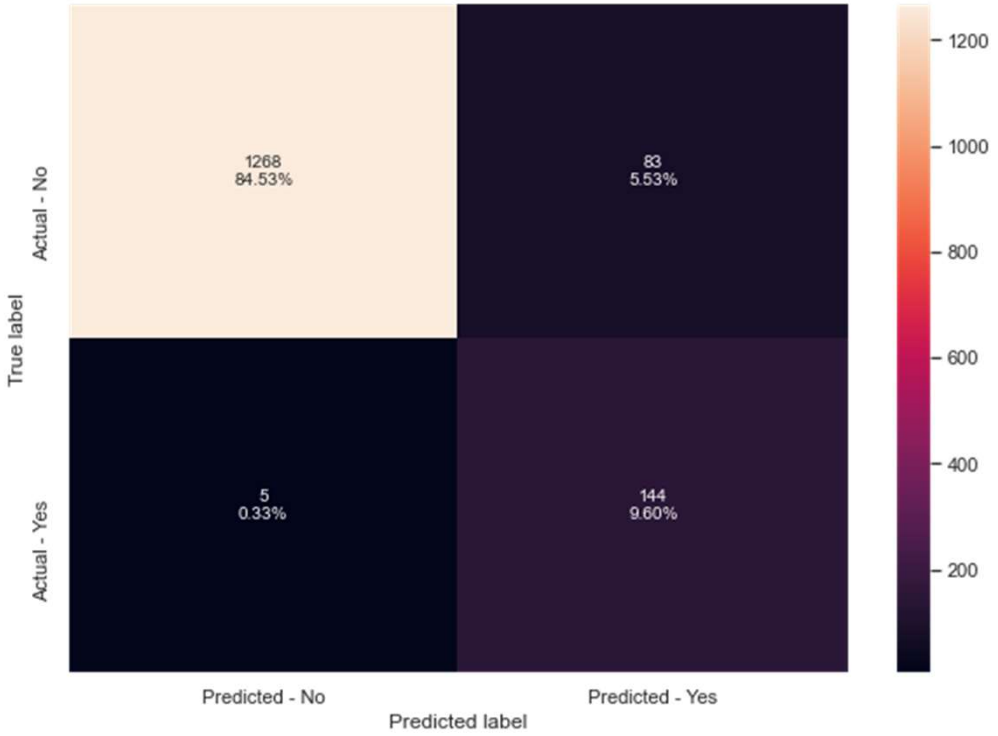
Model Selection

- **Hyperparameter tuned model** is considered the **best model** as it gives the highest recall for test data when compared against all other decision tree models and logistic regression models.
- This model will be able to find 96.6% of the potential customers who will take a loan.
- Also, the accuracy is high indicating that the numbers of customers incorrectly targeted for taking a loan will be low.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
Base model	100	84.6	100	97.4
Max-depth=3	96.1	93.3	95.7	95.2
Hyperparameter tuning (pre pruning)	98.8	96.6	93.5	94.1
Cost complexity (post pruning)	96.7	90.6	99.3	98.4

Key Parameters

Following are key parameters influencing whether a customer takes a personal loan or not in the descending order of importance based on hyper parameter tuning.



5. Key Insights

Customer Profile

- 75% of customers have income below $98k
- 75% of customers have monthly credit card spend of less than $2.5k
- Most of the customers are single with 50% of customers having 2 member families or less
- Most customers are undergrads followed by holders of advance/professional education.

Personal loan uptake

- Based on exploratory data analysis, customers taking personal loans have the following characteristics compared to customers who didn't take a personal loan:
  a. higher income
  b. higher mortgage amount
  c. higher credit card spend per month
  d. more numbers of members in the family

- Higher proportion of customers with a certificate of deposit account or security deposit account take a personal loan than ones without these type of accounts

- County information did not make a notable difference to the model performance.

- Higher proportion of customers with graduation and advanced/professional education have taken personal loan compared to customers with undergrad education.

- There is difference in proportion of customers who take personal loans across various counties.

- The model which gives the best performance indicates that the significant variables influencing a customer taking the loan are listed below. The higher the below values, the higher the chance of a customer taking a personal loan which is mostly inline with the above mentioned exploratory data analysis.
  a. Income
  b. education
  c. monthly credit card spend
  d. size of family
  e. possess certificate of deposit account

6. Recommendations

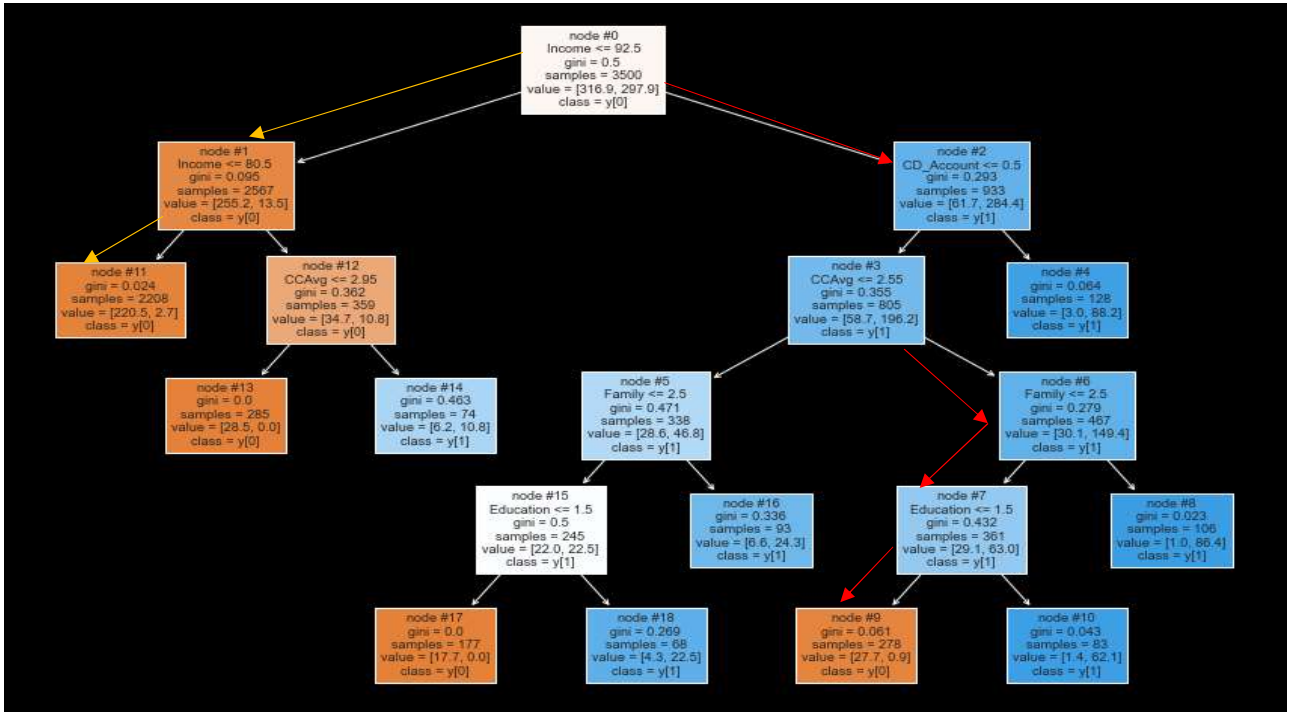
- Following type of customers should be targeted by the marketing department to find potential customers who have a higher probability of taking a loan
  - High income customers
  - Educated (graduates or professional/advance degree holders)
  - high credit card spend
  - customers with a kids (family of 3 or more)
  - posses a certificate of deposit account
- Target customers who are paying an interest on overdue credit card payment as they are showing higher spending habits which can be catered more efficiently by a medium term personal loan at a lower interest rate than a higher credit card interest rate.
- Target customers with a deposit account and high income as they have a potential to spend, but might not want to dip into their savings, but might be happy to pay instalments against a personal loan.
- The type of customers who will take a loan forms a lower proportion within the bank’s existing customer base. In order to increase the number of customers taking personal loans, they need to target new customers with the above profile.

7. Misclassified data analysis (I/II)

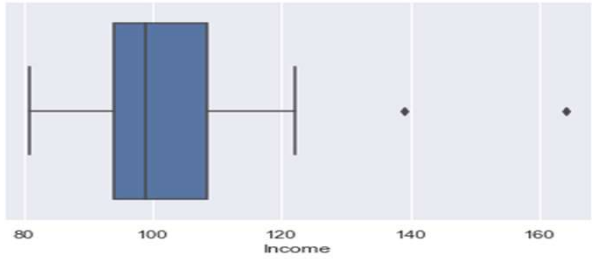
False negatives (5 customers)- Customers who were wrongly identified as customers who will not take a loan, but actually took the loan.

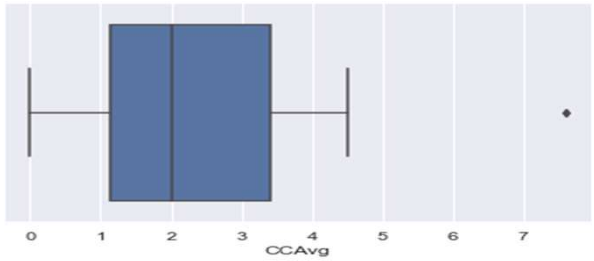
Key reasons- Following are reasons for these customers not being identified as potential customers who will take a loan:

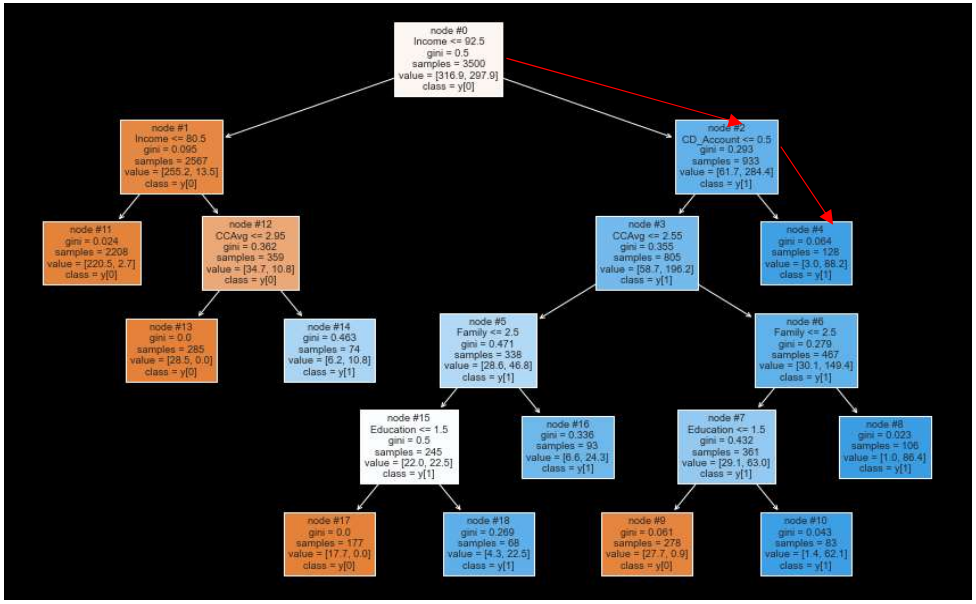
- 4 customers have income less than $80.5k which is the lower cut-off for a person to take a loan as per the model. This is shown using the yellow line below.
- 1 customer has high income and credit card spend, but has a family size of 2 and undergrad education which made the model to predict that this customer will not take a loan. This is shown using a red line below.



7. Misclassified data analysis (II/II)







False positives (83 customers)- Customers who were wrongly identified as customers who will take a loan, but actually didn't take the loan.

Key reasons- Following are reasons for being wrongly identified as potential customers who will take a loan:

- These customers have high income with the at least $80k which is more than the overall mean of $73.8k
- These customers have high average credit card spend of $2k which is above the overall mean of $1.9k
- There are 7 customers with income above $92.5k and have a CD account which according to the model predicts the customer to take a personal loan, but these customers did not take a loan. This is shown as a red line in the decision tree above.