SUPERVISED LEARNING CLASSIFICATION
LOGISTIC REGRESSION & DECISION TREE
LOAN PURCHASING PREDICTION

Naeem Sufi

BACKGROUND

- In order to increase revenues, AllLife Bank is interested in penetrating their depositors to convert them to Loan holders .
- Campaign that ran previous year showed a healthy rate of over 9%
- Retail Marketing Department wants to devise campaigns with better marketing to increase the success ratio.
- Following Modelling will help to identify potential customers who have a higher probability of purchasing the loan.

BUSINESS 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.

PROBLEM APPROACH

- EDA Analysis (Univariate, Bivariate & other deep techniques)
- Two Machine Learning Algorithms Logistic Regression and Decision Tree
- Model Evaluation and Feature Extraction
- Comparison
- Recommendation

EXPLORATORY DATA ANALYSIS

There are total 14 attributes in the dataset and in the context of the given problem, the target (or dependent) attribute is "Personal Loan" whereas the remaining are independent attributes.

There are 5000 rows and 14 columns in our dataset.

Attribute information

- ID : Customer ID
- Age : Customer's age in completed years
- Experience : #years of professional experience
- Income : Annual income of the customer ($000)
- ZIP Code : Home Address ZIP code.
- Family : Family size of the customer
- CCAvg : Avg. spending on credit cards per month ($000)
- Education : Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
- Mortgage : Value of house mortgage if any. ($000)
- Securities Account : Does the customer have a securities account with the bank?
- CD Account : Does the customer have a certificate of deposit (CD) account with the bank?
- Online : Does the customer use internet banking facilities?
- Credit card : Does the customer use a credit card issued by Other Banks?
- Personal Loan : Did this customer accept the personal loan offered in the last campaign? (Target Attribute)

EXPLORATORY DATA ANALYSIS

- Exploring the column names is an important aspect of EDA.
- We can see that columns are not null. The data types of all columns are int and float data type.

By closely observing the data and description given about each column attribute we can say that:

- Numeric data columns (Interval or Ratio) are Age, Experience, Income, Mortgage and CCAvg
- Ordinal Categorical columns are Family and Education
- Nominal Categorical columns are ID, ZIP Code, Securities Account, CD Account, Online, Credit Card, Personal Loan

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 5000 non-null   int64
 2   Experience          5000 non-null   int64
 3   Income              5000 non-null   int64
 4   ZIPCode             5000 non-null   int64
 5   Family              5000 non-null   int64
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64
 8   Mortgage            5000 non-null   int64
 9   Personal_Loan       5000 non-null   int64
 10  Securities_Account  5000 non-null   int64
 11  CD_Account          5000 non-null   int64
 12  Online              5000 non-null   int64
 13  CreditCard          5000 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

EXPLORATORY DATA ANALYSIS

•Age feature is normally distributed with majority of customers falling between 30 years and 60 years of age.
We can confirm this by looking at the describe statement above, which shows mean is almost equal to median
•Experience is normally distributed with more customer having experience starting from 8 years. Here the mean is equal to median.
There are negative values in the Experience.
This could be a data input error as in general it is not possible to measure negative years of experience.
We can delete these values, because we have 3 or 4 records from the sample.
•Income is positively skewed. Majority of the customers have income between 45K and 55K.
We can confirm this by saying the mean is greater than the median
•CCAvg is also a positively skewed variable and average spending is between 0K to 10K and majority spends less than 2.5K
•Mortgage 70% of the individuals have a mortgage of less than 40K. However, the max value is 635K
•The variables Family and Education are ordinal variables. The distribution of families is evenly distributed

DATA CLEANING(EDA)

Now we are finding any null values inside our dataset.

We have found that there are no null values

inside the dataset. So, our data is clean

and ready to go for further implementations

```
ID                    0
Age                   0
Experience            0
Income                0
ZIPCode               0
Family                0
CCAvg                 0
Education             0
Mortgage              0
Personal_Loan         0
Securities_Account    0
CD_Account            0
Online                0
CreditCard            0
dtype: int64
```

EXPLORATORY DATA ANALYSIS

Columns basic statistics are very important in EDA. Thus, we found some important features as follows :

	ID	AGE	Experience	Education	Family	Personal Loan	Income
Count	5000.00	5000.00	5000.00	5000.00	5000.00	5000.000	5000.00
mean	2500.500	45.3384	20.104	1.88100	2.39640	0.96000	73.7742
std	1443.52	11.4631	11.465	0.8398	1.147663	0.294670	46.033
min	1.0000	23.000	-3.000	1.0000	0.0000	0.0000	8.0000
25%	1250.750	35.000	10.000	1.0000	0.7000	0.0000	39.0000
50%	2500.500	45.000	20.000	2.0000	1.5000	0.0000	64.0000
max	5000.00	67.000	43.000	3.0000	10.000	1.0000	224.000

EXPLORATORY DATA ANALYSIS

- Five-point summary suggests that Experience has negative value.
- We can see the Min, Max, mean and std deviation for all key attributes of the dataset
- Income has too much noise and slightly skewed right, Age and exp are equally distributed.
- Columns with binary information such as Securities Account, CD Account, Online, Credit Card, Personal Loan are also clean.

EXPLORATORY DATA ANALYSIS

We found skewness of the dataset that is,

ID : 0.00000

Age : -0.02934

Experience : -0.026325

Income : 0.84331

Family : 0.15522

Education : 0.2270

Personal Loan : 2.7436

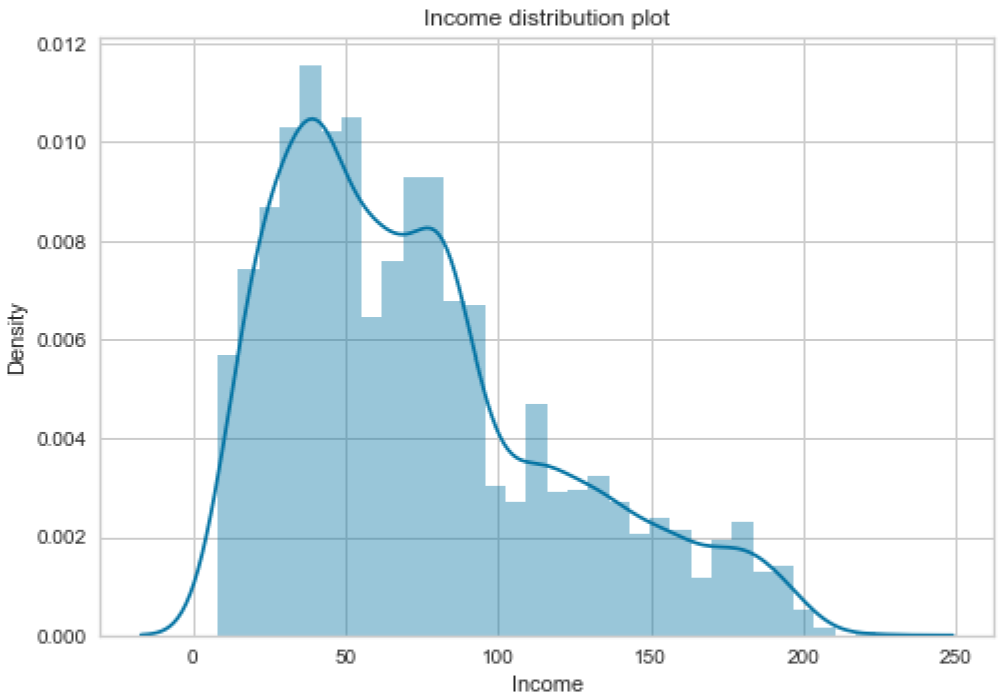
Credit card : 0.9045

We can see that personal loan is highly skewed while other columns are equally distributed and approximately symmetric

EXPLORATORY DATA ANALYSIS (UNIVARIATE)

Uni stands for "one," meaning that, there is just one sort of variable in the data. Univariate analysis' main purpose is to characterize the data.

The information will be collected, analyzed, and a pattern will be identified.



The graph is of Income distribution and Dense volume is observed between 20 Income < 100 & > 5 Density == 0.010 covering 20% of the region

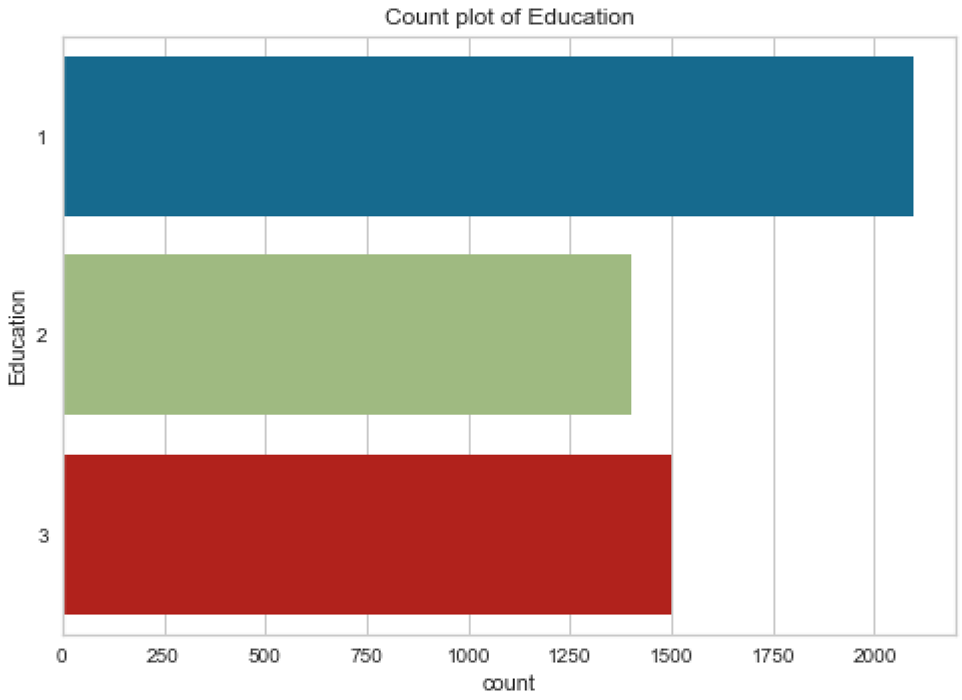
EXPLORATORY DATA ANALYSIS (UNIVARIATE)

This graph is basically a count plot of education levels.

Total number of values for a certain level.

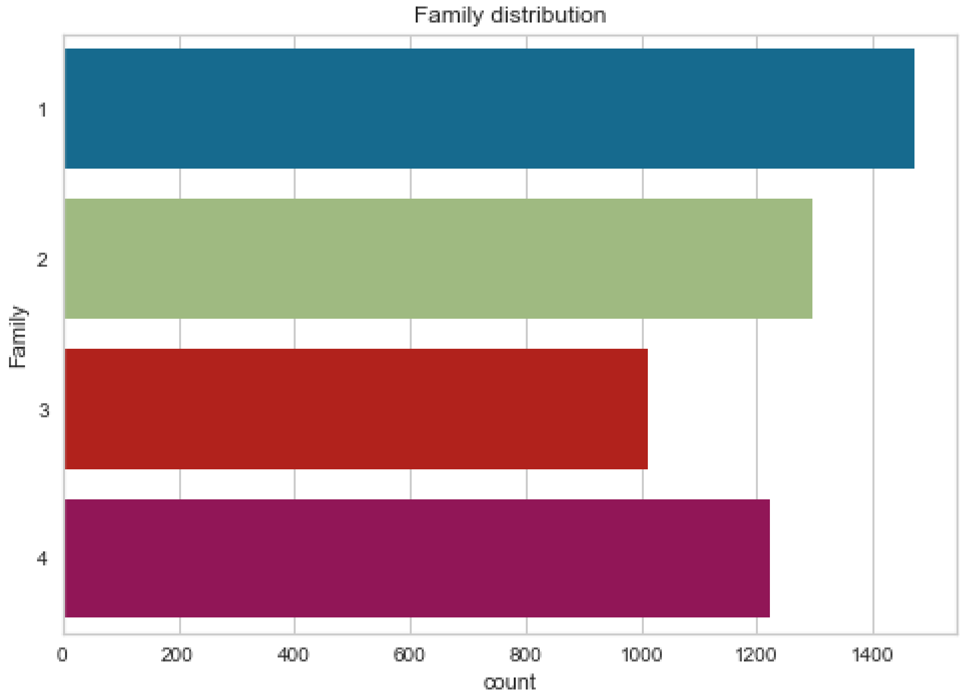
We can see customers having education level 1 is more than other levels. Almost 75% of the customers Are from the education level 1 and 55% From 3 and 50% from 2.

This feature will make a big impact on Model's performance.



EXPLORATORY DATA ANALYSIS (UNIVARIATE)

Another count plot showing
the family distribution. From the
graph it is clear that family
having member 1 is more
than other levels and this
distribution will help us to
identify patterns.



As we can see over 1400 counts are
Present for the customers having a family member 1 and only 1000 for 3

INFERENCES FROM UNIVARIATE ANALYSIS

- Average Income group of customers is between 20 and 100.
- Average Experience of customers are 20
- Most of the customers are having education level 1.
- There are no null values in the data
- 75% of the family members of the customer are 1
- 75% of the customers are using Online banking facility.
- Majority of the Customers are with experience between 20-30yrs
- Income is right skewed
- Personal Loan is extremely positive squared

EXPLORATORY DATA ANALYSIS (BIVARIATE)

In the first scenario of Bivariate analysis, we look at the facetgrid of income based on the family of each customer.

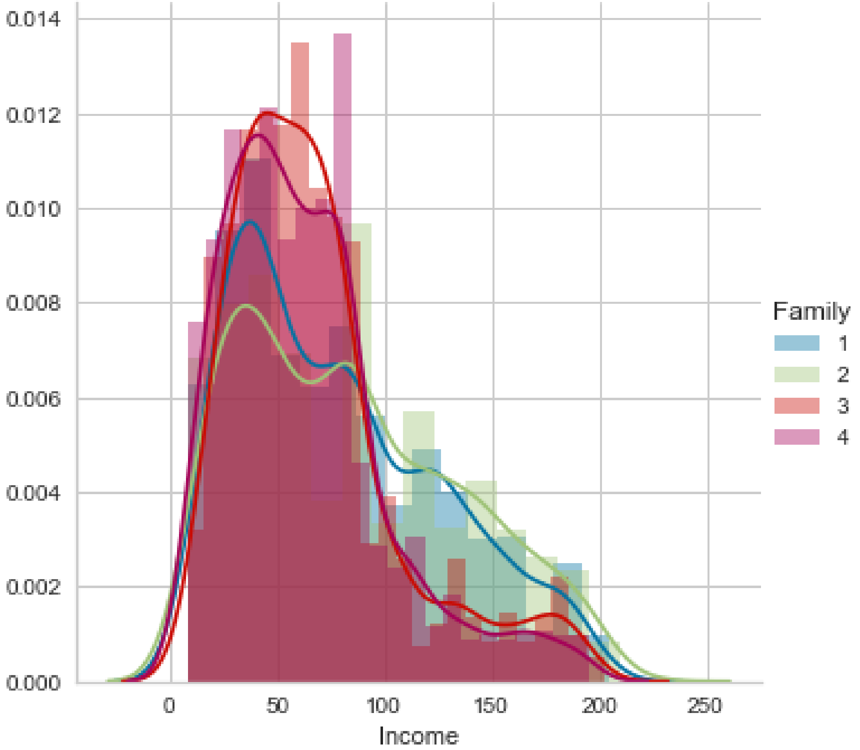
We find that the customers that have

a family members above 2 are having denser

Income values lies between 50 &100.

Customers having family member of 1 are having higher

Income values as compared to other .

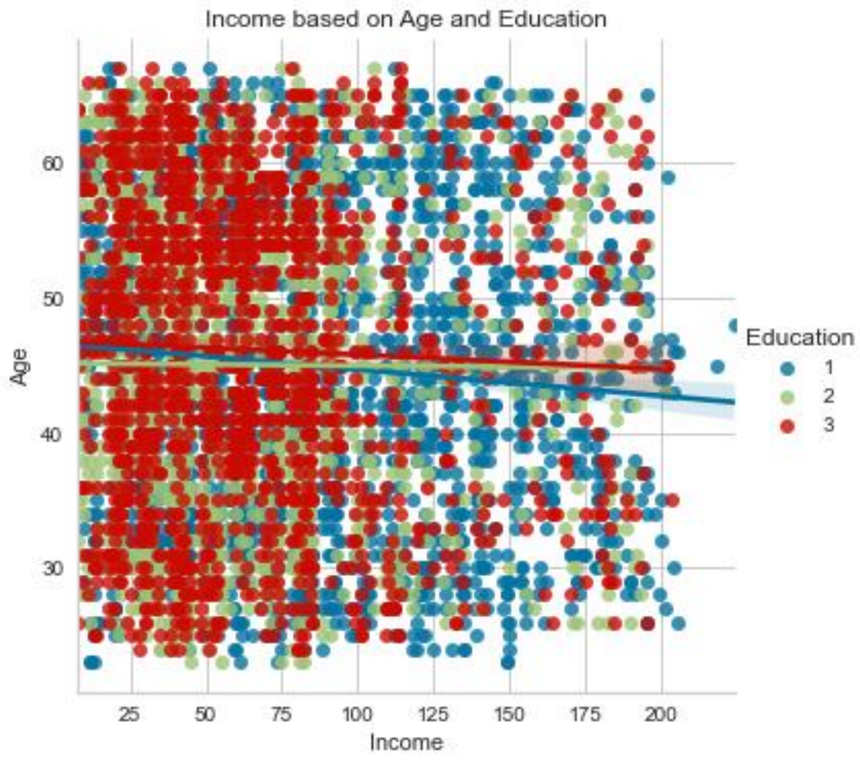


EXPLORATORY DATA ANALYSIS (BIVARIATE)

This illustration is of Income that is based on Age and Education levels.

Customers having more education levels are having income between 25 and 100.

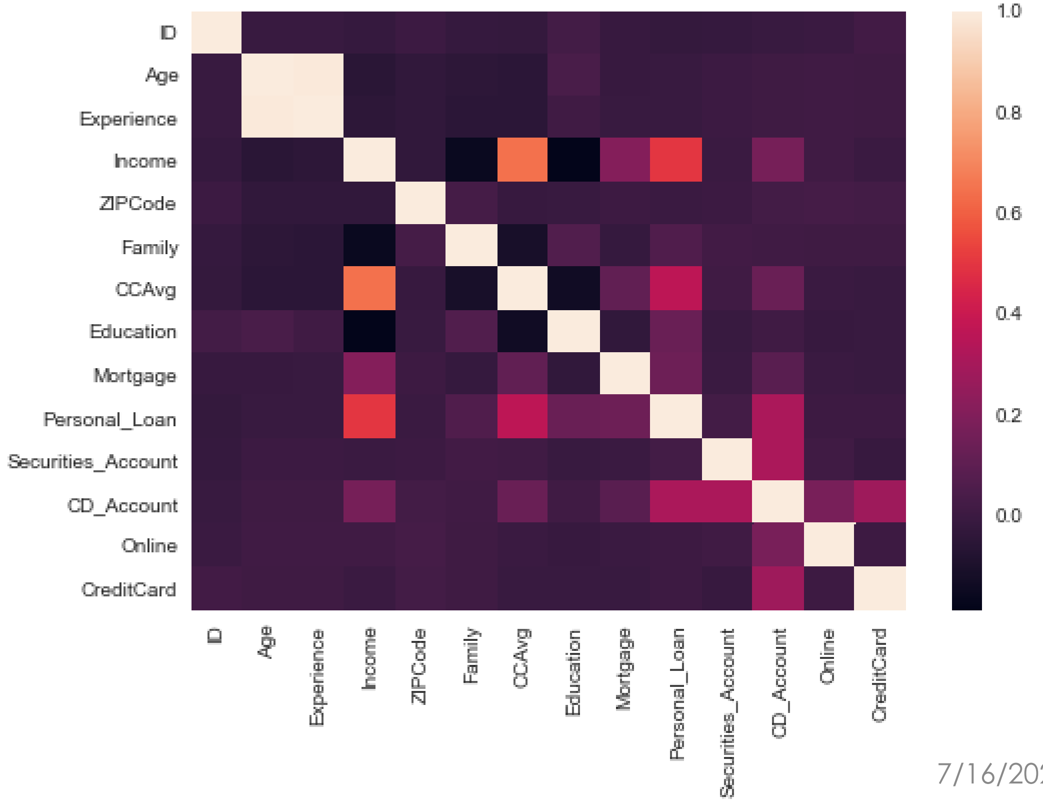
We can see a huge dense distribution of red colors on the left side of plot from top to bottom.



EXPLORATORY DATA ANALYSIS (BIVARIATE)

This graph is of correlation between variables in the form of heatmap.

There is a dense correlation between Age and Experience columns. We can Drop them if they make any lack in the model's performance.



INFERENCE FROM BIVARIATE ANALYSIS

Customers with higher education are buying Personal Loan compared to other groups.

Family with size more than 2 are more interested in personal loans

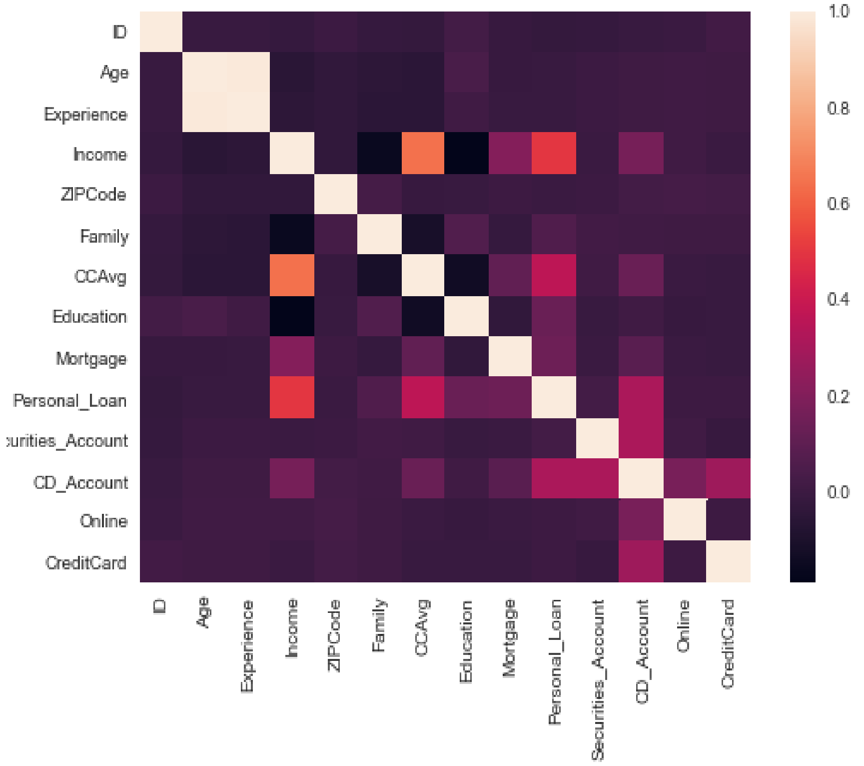
There is a higher correlation in Age and Experience feature so we can drop one of them.

Correlation coefficient of ID and target variable Personal Loan is negative and close to zero so we can drop the variable.

Correlation coefficient of Age and Experience are negative and close to zero so we can drop these variables as well.

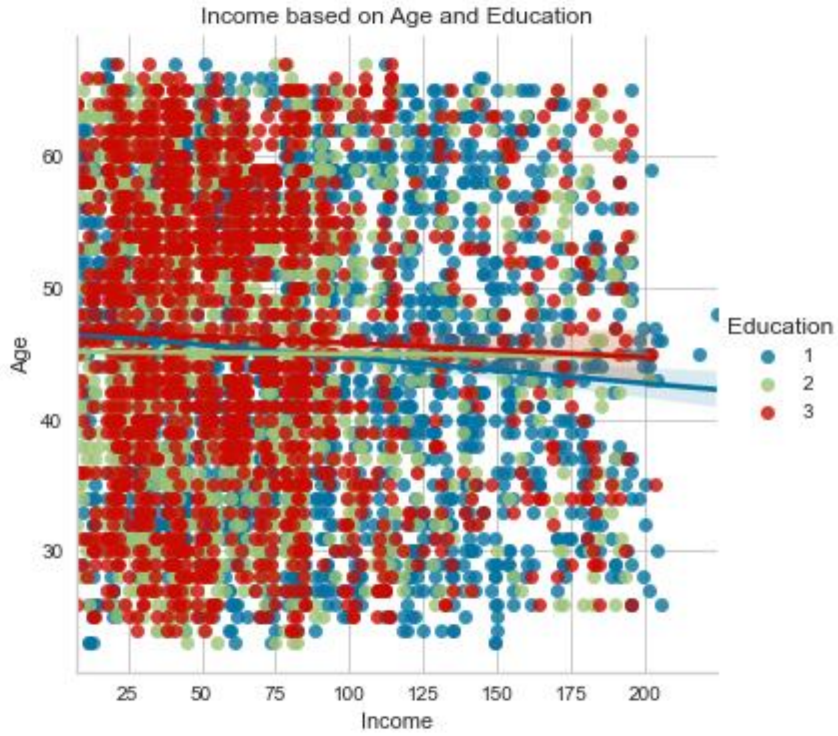
USEFUL PATTERNS

	ID	Age	Experience	Income	ZIPCode	Family	CCAvg	Education	Mortgage	Personal_Loan	Securities_Account
ID	1.000000	-0.008473	-0.008326	-0.017695	0.002240	-0.016797	-0.024675	0.021463	-0.013920	-0.024801	-0.016972
Age	-0.008473	1.000000	0.994215	-0.055269	-0.030530	-0.046418	-0.052012	0.041334	-0.012539	-0.007726	-0.000436
Experience	-0.008326	0.994215	1.000000	-0.046574	-0.030456	-0.052563	-0.050077	0.013152	-0.010582	-0.007413	-0.001232
Income	-0.017695	-0.055269	-0.046574	1.000000	-0.030709	-0.157501	0.645984	-0.187524	0.206806	0.502462	-0.002616
ZIPCode	0.002240	-0.030530	-0.030456	-0.030709	1.000000	0.027512	-0.012188	-0.008266	0.003614	-0.002974	0.002422
Family	-0.016797	-0.046418	-0.052563	-0.157501	0.027512	1.000000	-0.109275	0.064929	-0.020445	0.061367	0.019994
CCAvg	-0.024675	-0.052012	-0.050077	0.645984	-0.012188	-0.109275	1.000000	-0.136124	0.109905	0.366889	0.015086
Education	0.021463	0.041334	0.013152	-0.187524	-0.008266	0.064929	-0.136124	1.000000	-0.033327	0.136722	-0.010812
Mortgage	-0.013920	-0.012539	-0.010582	0.206806	0.003614	-0.020445	0.109905	-0.033327	1.000000	0.142095	-0.005411
Personal_Loan	-0.024801	-0.007726	-0.007413	0.502462	-0.002974	0.061367	0.366889	0.136722	0.142095	1.000000	0.021954
Securities_Account	-0.016972	-0.000436	-0.001232	-0.002616	0.002422	0.019994	0.015086	-0.010812	-0.005411	0.021954	1.000000
CD_Account	-0.006909	0.008043	0.010353	0.169738	0.021671	0.014110	0.136534	0.013934	0.089311	0.316355	0.317034
Online	-0.002528	0.013702	0.013898	0.014206	0.028317	0.010354	-0.003611	-0.015004	-0.005995	0.006278	0.012627
CreditCard	0.017028	0.007681	0.008967	-0.002385	0.024033	0.011588	-0.006689	-0.011014	-0.007231	0.002802	-0.015028
Percentage	-1.000000	-1.000000	-1.000000	1.000000	1.000000	1.000000	1.000000	NaN	NaN	NaN	NaN



Age and experience is highly correlated, and we can drop them for any inconsistencies

PATTERN-2

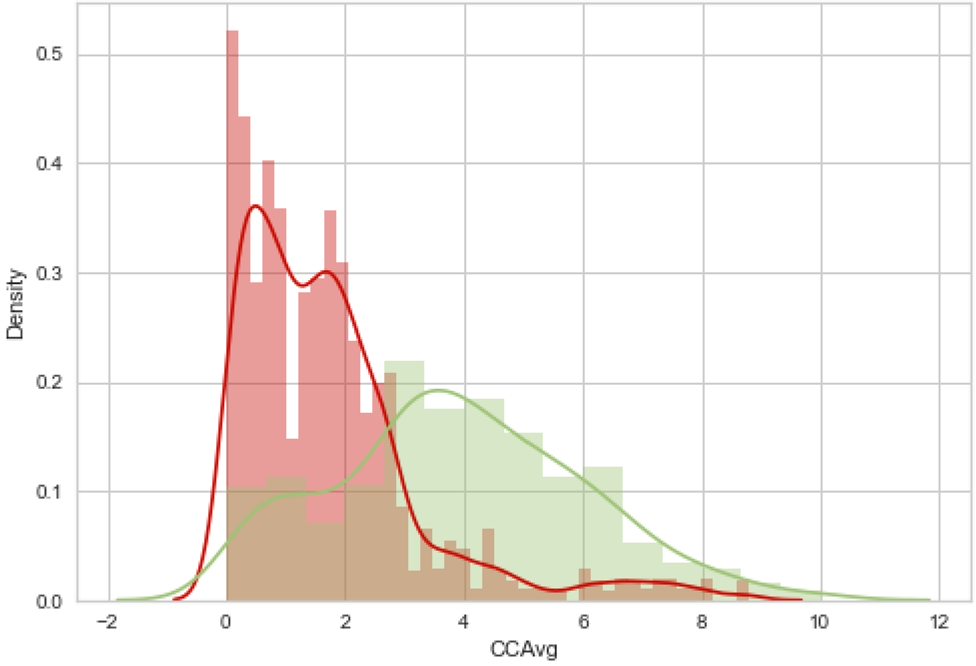


Customers having more education levels
are having income between 25 and 100.
We can see a huge dense distribution of
red colors on the left side of plot
from top to bottom.
A straight Line can be seen showing a linear distribution between Age and Income.

PATTERN-3

The graph show persons who have personal loan have a higher credit card average. Average credit card spending with a median of 3800 dollar indicates a higher probability of personal loan. Lower credit card spending with a median of 1400 dollars is less likely to take a loan. This could be useful information.

Because the total number of credit card spending for loan customers are 3800 while non-loan are only 1400.



EXPLORATORY DATA ANALYSIS

Personal Loan Age	0	1
(20,30)	89.423077	10.576923
(30,40)	90.453074	9.546926
(40,50)	90.393701	9.606299
(50,60)	91.307634	8.692366

From above table as well as distribution plot of Age attribute, one can observe that most of the customers lie in the age group of 30 to 60. Also, one can observe that 10.5% of the total customers in age group 20-30 have acquired personal loan from the bank, while in age groups (30-40), (40-50) and (50-60), there is a conversion rate of around 9%.

EXPLORATORY DATA ANALYSIS

Personal Loan	0	1
Experience		
(0, 10]	89.697465	10.302535
(10, 20]	90.582602	9.417398
(20, 30]	90.853190	9.146810
(30, 40]	90.661831	9.338169
(40, 50]	87.037037	12.962963

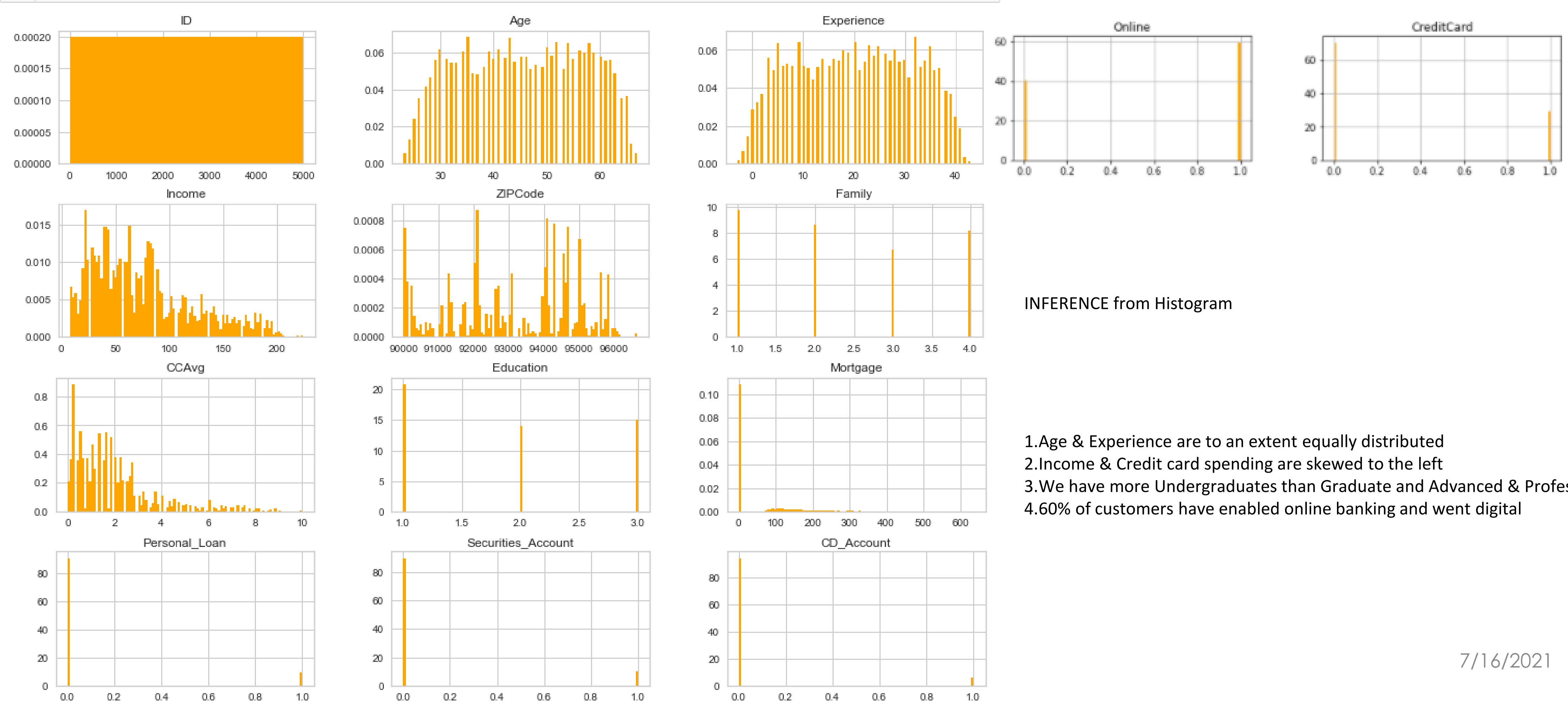
One can observe that out of the total customers with experience in the range 40-50 show a good conversion rate of almost 13% for buying the personal loan, a healthy conversion rate of about 10.30% in the experience range 0 to 10, while in the ranges (10-20), (20-30) and (30-40) years of experience it is around 9%.

EXPLORATORY DATA ANALYSIS

Personal Loan	0	1
Income		
(0, 50]	100.000000	0.000000
(50, 100]	97.758805	2.241195
(100, 150]	71.428571	28.571429
(150, 200]	49.530516	50.469484
(200, 250]	81.250000	18.750000

No customer with income < 50,000$ opted for the personal loan whereas half of the customers with income within the range of 150 to 200 thousand dollars acquired personal loan...! Customers within range of (100 to 150) and (200 to 250) thousand dollars showed a conversion rate of about 28.5% and 18.75%, respectively.

FEATURE ENGINEERING



INFERENCE from Histogram

1.Age & Experience are to an extent equally distributed
2.Income & Credit card spending are skewed to the left
3.We have more Undergraduates than Graduate and Advanced & Professional
4.60% of customers have enabled online banking and went digital

DATA SPLITTING

We have split the data into X and y for training and testing purposes.

Here is the finding of our data separation

- For X we have 5000 rows and 13 columns
- For y we have 5000 rows and 1 column

We have defined the test size as 0.2 that the train and test will be 80/20 and the random state for model implementation set to 0.2

LOGISTIC REGRESSION MODEL BUILDING

Highest coefficients :

CD Account : 3.8544

Education : 1.7335

Family : 0.6859

Experience : 0.1248

Income : 0.0541

All the values having a coefficient positive will increase the chance of getting loan as compared to negative values.

Logit Regression Results

Dep. Variable:	Personal_Loan	No. Observations:	5000
Model:	Logit	Df Residuals:	4987
Method:	MLE	Df Model:	12
Date:	Sat, 10 Jul 2021	Pseudo R-squ.:	0.5917
Time:	09:59:06	Log-Likelihood:	-645.55
converged:	True	LL-Null:	-1581.0
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
ID	-5.636e-05	5.13e-05	-1.099	0.272	-0.000	4.41e-05
Age	-0.1167	0.059	-1.981	0.048	-0.232	-0.001
Experience	0.1248	0.059	2.125	0.034	0.010	0.240
Income	0.0541	0.003	20.791	0.000	0.049	0.059
ZIPCode	-0.0001	1.67e-05	-6.628	0.000	-0.000	-7.81e-05
Family	0.6859	0.074	9.284	0.000	0.541	0.831
CCAvg	0.1221	0.040	3.087	0.002	0.045	0.200
Education	1.7335	0.114	15.162	0.000	1.509	1.958
Mortgage	0.0005	0.001	0.814	0.416	-0.001	0.002
Securities_Account	-0.9659	0.285	-3.388	0.001	-1.525	-0.407
CD_Account	3.8544	0.323	11.919	0.000	3.221	4.488
Online	-0.6761	0.157	-4.312	0.000	-0.983	-0.369
CreditCard	-1.1185	0.204	-5.483	0.000	-1.518	-0.719

LOGISTIC REGRESSION EVALUATION

As we have an accuracy of 90% but this is

Quite bad.

For Imbalanced datasets it is wise to look for

precision and recall rather than accuracy.

```
[[880  29]
 [ 66  25]]
Accuracy Score  for Logistic Regression:90.5
F1 Score  for Logistic Regression :34.48275862068966
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       909
           1       0.46      0.27      0.34        91

    accuracy                           0.91      1000
   macro avg       0.70      0.62      0.65      1000
weighted avg       0.89      0.91      0.89      1000
```

Our Logistic Regression model on Raw data has given a recall score of 57%

This is bad, With this score the bank does not gain much.

Precision (True Positive Rate) means out of customers

who we predicted to buy the loan how many actually acquired the loan which is 93%.

LOGISTIC REGRESSION EVALUATION AFTER BAD SCORE

The previous accuracy was 90% and now we

Have an accuracy of 86% which is good while

On the other hand, we have successfully

Increased the precision and recall ,fi-score

respectively. We have successfully done sampling

to the dataset.

	precision	recall	f1-score	support
0	0.87	0.85	0.86	3158
1	0.86	0.88	0.87	3170
accuracy			0.86	6328
macro avg	0.86	0.86	0.86	6328
weighted avg	0.86	0.86	0.86	6328

We can say Precision Score of 86% implies among predicted positive how much was actual positive, Among total customers we predicted how many customers acquired the loan

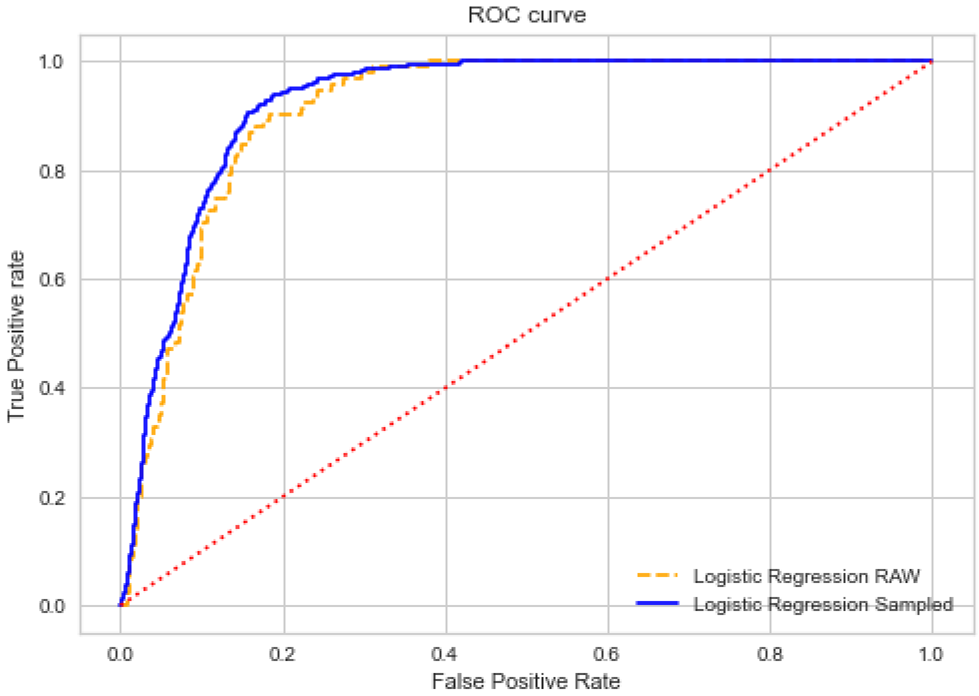
ROC SCORE AND AUC ODDS

ROC_AUC_Score for Logistic Regression

with Raw Data:0.9101052962439101

ROC_AUC_Score for Logistic Regression

after improved Data:0.9234038833826465



So, we have an improved version of

Score with the increment of 1.

From the curve it is clearly shown that

With having false positive rate 0.4 and true 1 they are both intersecting each other.

The Blue line shows that our Logistic model on sampled data almost covers more region

LOGISTIC REGRESSION IMPORTANT FEATURES

According to this model,

Important features :

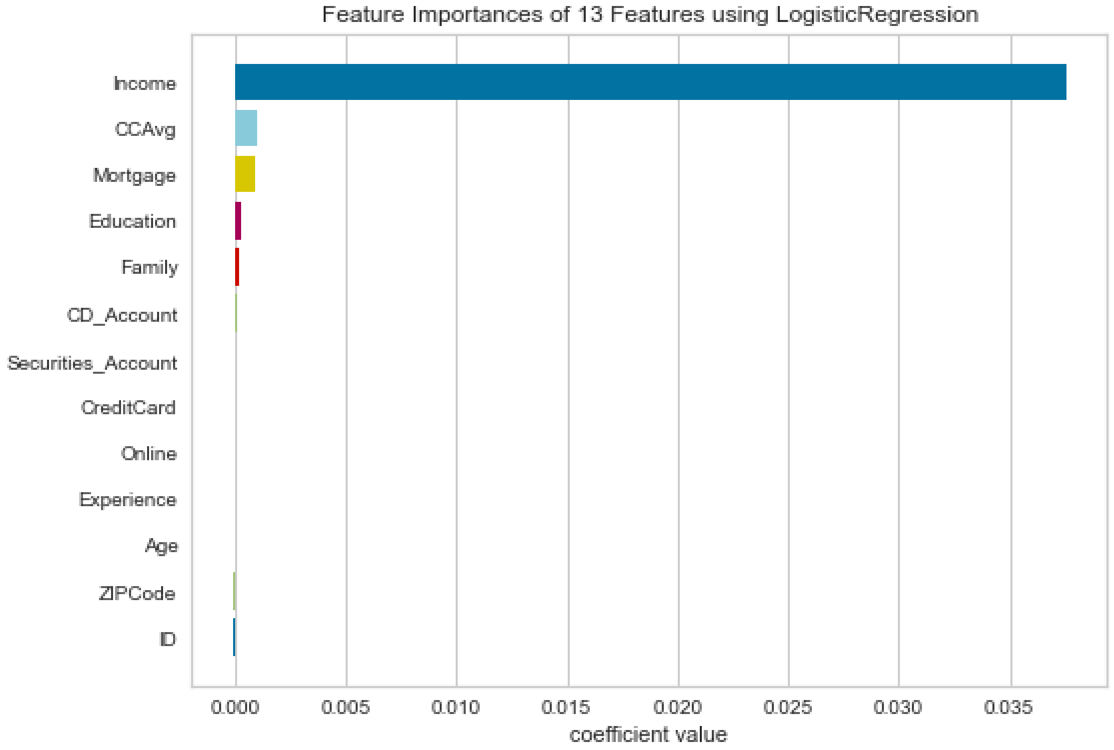
Income

CCAVG

Mortgage

Education

Income is highly dependent on purchasing a loan



KEY FINDINGS

We have utilized three different methods to evaluate our model which includes:

- Mean Squared Error
- Root Mean Squared Error
- Accuracy score

From our finding the mean squared error for logistic regression was 0.095

Root mean squared error was 0.30822

Accuracy score was 0.90 that is 90%

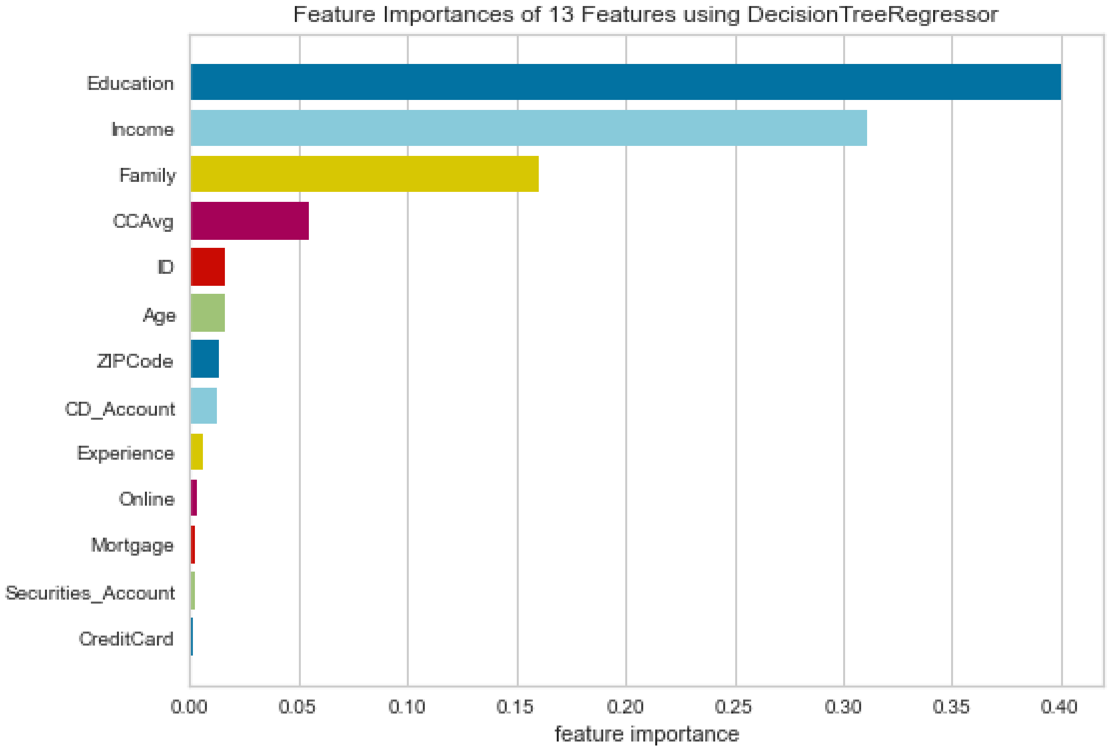
But recall and precision was quite bad which was 76 and 34 % respectively .So we improved that version and increased that score up-to 86% and 87% percent respectively.

The ROC and AUC ODDs score was 91 and improved model 92% respectively.

From the finding we have found Income is the most important feature that is highly dependent on Loan purchasing

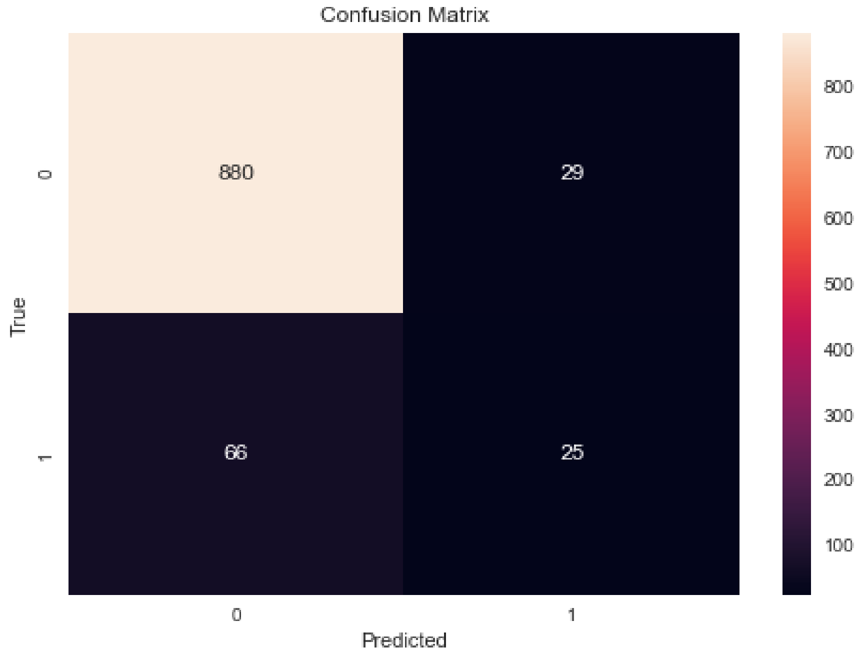
Customers with Higher Education, have bank CD Accounts and Higher Income : more likely to obtain loans.

DECISION TREE BUILDING



Important Features:

Education, Income, Family, CCAvg



True positive 29% while true negative 88%

DECISION TREE PRUNING COMPLEXITY

The best accuracy can be
Seen between 0.01 and 0.02
Whereas our model decrease from
0.012 which is good for any unseen
data to be evaluate.
This can be helpful in determining the
Optimal threshold value.
The optimal threshold value for
Decision tree is 0.3.

DECISION TREE COST COMPLEXITY PRUNING

We try to evaluate our model with classification report as it gives more broader details about the model.

The classification report gives,

	precision	recall	f1-score	support
0	0.99	0.99	0.99	909
1	0.93	0.86	0.89	91
accuracy			0.98	1000
macro avg	0.96	0.93	0.94	1000
weighted avg	0.98	0.98	0.98	1000

From the report we can see accuracy of the model is 98% and precision is also high.

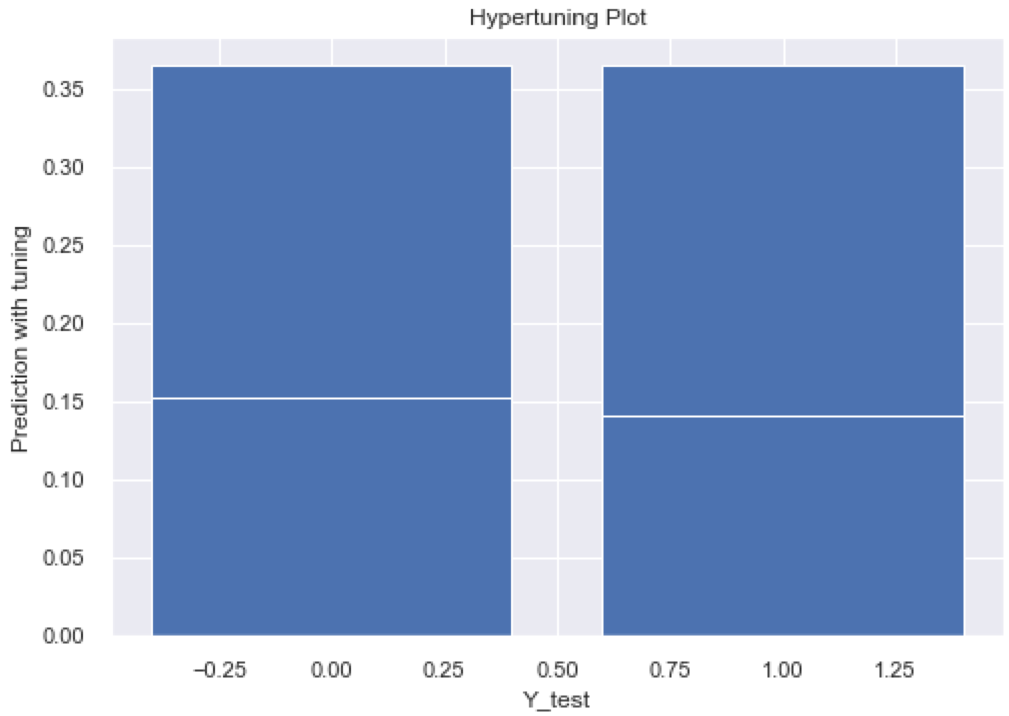
Recall and precision is also high so we can say that customers who can buy loan per 100 ratio is 93% respectively.

DECISION TREE COST COMPLEXITY PRUNING

Values Importance :
Education : 0.39
Income : 0.31
Family : 0.17
CCAVG : 0.07
AGE : 0.02

Recall
0.88
Precision
0.92
Accuracy
0.98 (98%)

Mean squared error after hyper tuning 0.070



BUSINESS TAKEAWAYS

- We have seen education, family, income and ccavg are the most features that we have found from this model. Education level of a person will determine if he is going to acquire this personal loan or not.
- Business modelling should keep focus on these attributes while making any kind of phases. This should be done with the initial phases.

EDA ON WRONG PREDICTED VALUES

This analysis is made on decision tree because we are having more accuracy than logistic regression.

We try to plot a predicted and actual values and we have found straight and dense line

on zero and few on 1 and negative 1.

No pattern is found except a

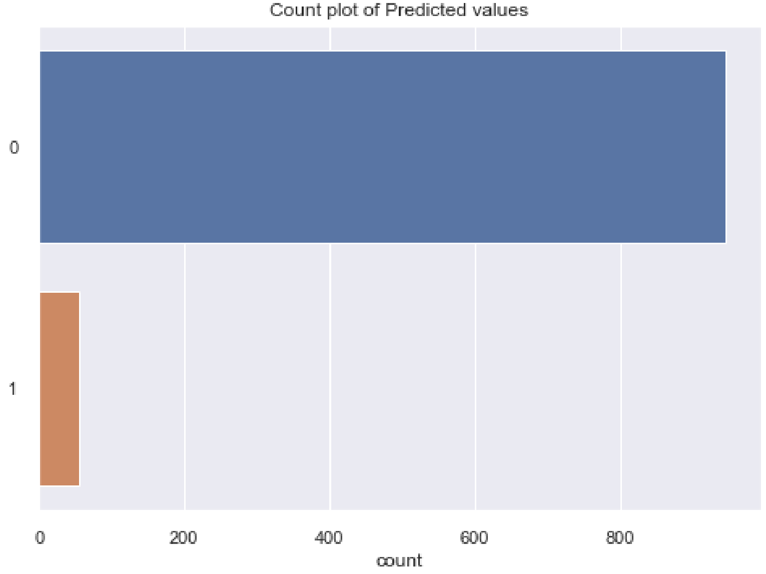
singular straight lines.

This analysis is made on whole dataset rather than

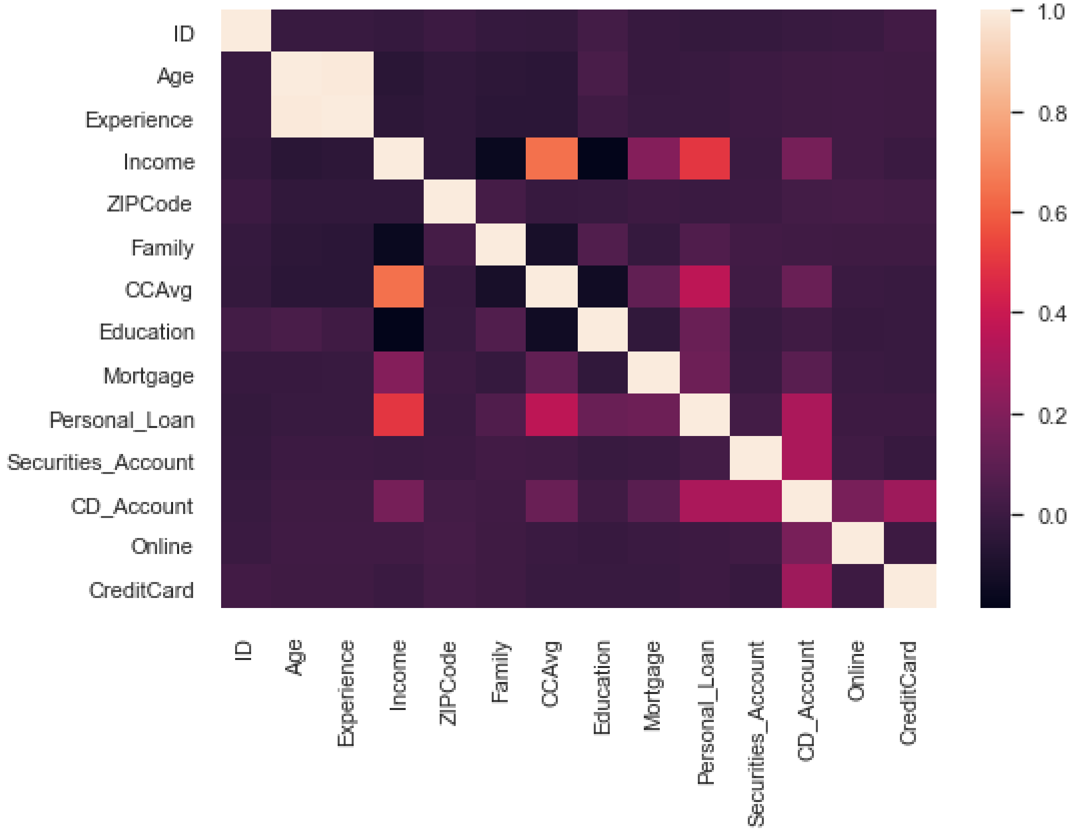
On a single column or a set of columns.

EDA ON WRONG PREDICTED VALUES

We have generated a heatmap for the Predicting dataset. From that we can see that a big correlation is found between Experience and Income data cells.



This is the count plot of predicted values



ACTIONABLE INSIGHTS

From the results of our two models

- Accuracy for logistic regression was 90%
- Accuracy for Decision tree was 98%

Most Important feature:

- Logistic Regression : Income
- Decision Tree : Education

RECOMMENDATIONS

- As per the findings our models we can say that Income and Education level are the two most important features.
- Bank should focus on the income of the customers while education level must also keep in mind while making decisions.
- Customer with higher education level also have the higher possibility.

MODEL COMPARISON

	Logistic Regression	Decision Tree
Accuracy	90%	98%
Precision	86%	92%
Recall	88%	88%
F1-Score	87%	90%

Accuracy for decision tree is high so most probably while going with this model customer will buy a loan depending on the important features that was found inside this model including Education, Income, Family and CCAVg.

Precision for decision tree is 92% and for logistic regression is 86% so we are going with decision tree because customer ratio for buying loan is much higher in this model as compared to logistic regression

FURTHER RECOMMENDATIONS

- The bank instead of calling someone for purchasing a loan should focus on the key features like Education of the customer, Income of the customer, total family members, and CCAVG of the customer.
- It seems that customers who have higher education level and with above average income have the higher proportion to purchase the personal loan.
- Clients with these characteristics should be targeted in the Business Plan to increase their access to Personal Loans.
- The bank can expand its reach outside its existing customer base. Consumers who match the criteria in its service area and beyond detected as a result of our predictive modelling.