<u>SLE Mini Project</u>

<u>Predict The Acceptance Of Personal Loan</u>

Project By:

Anuj kumar

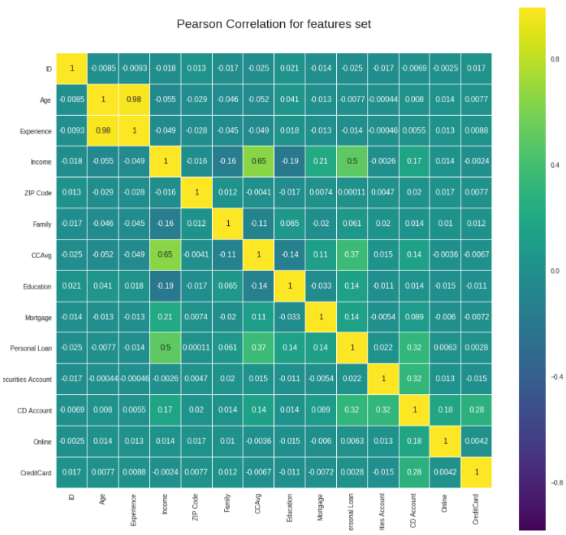
Dalon Francis Lobo

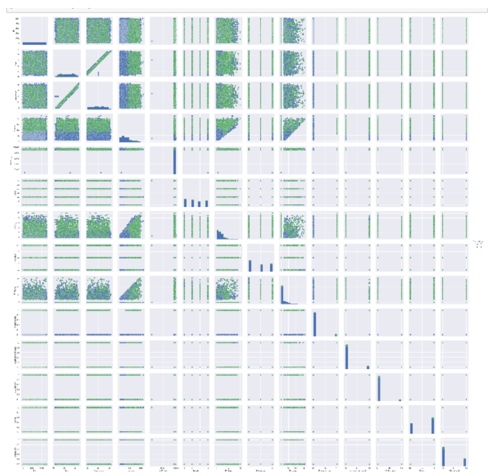
Jyothi H N

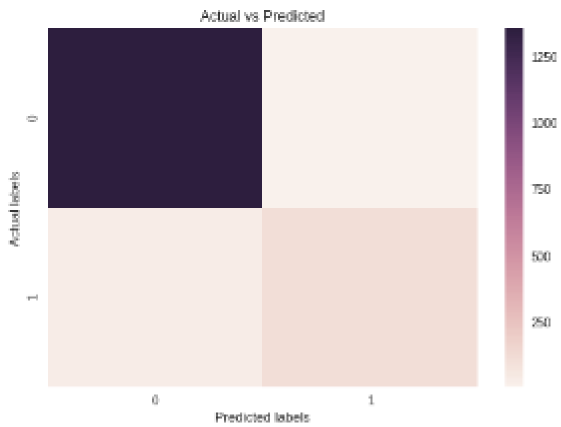
Pankaj Kumar

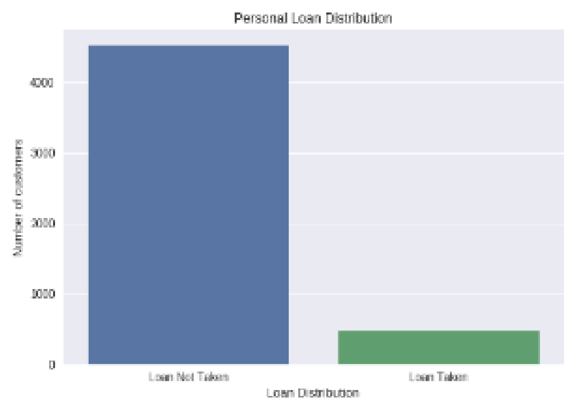
Satish Kaushik

Venkat









Content:

Table of Contents

Problem Description:

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

In this note book, we will build a model that will help the department to identify **the potential customers who have higher probability of purchasing the loan**. This will increase **the success ratio** while at the same time reduce the **cost of the campaign**.

We have data on **5000 customers** and each have **14 features** each. Fortunately there are no null values in the data.

Description of the data set:

- **ID**: Customer ID
- **Age**: Customer's age in completed years
- **Experience**: Number of years of professional experience
- **Income**: Annual income of the customer ($000)
- **ZIPCode**: Home Address ZIP code.
- **Family**: Family size of the customer
- **CCAvg**: Average 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)
- **Personal Loan**: Did this customer accept the personal loan offered in the last campaign?
- **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?
- **CreditCard**: Does the customer use a credit card issued by UniversalBank?

The number of customers who accepted the personal loan that was offered to them in the campaign is merely **480 customers**. That implies the success ratio is **9.6%**. Our target will be to increase this success ratio.

Univariate Analysis

	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIP Code	5000.0	93152.503000	2121.852197	9307.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396400	1.147663	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937913	1.747666	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0
Personal Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0
Securities Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0

Age feature is normally distributed with mean = median

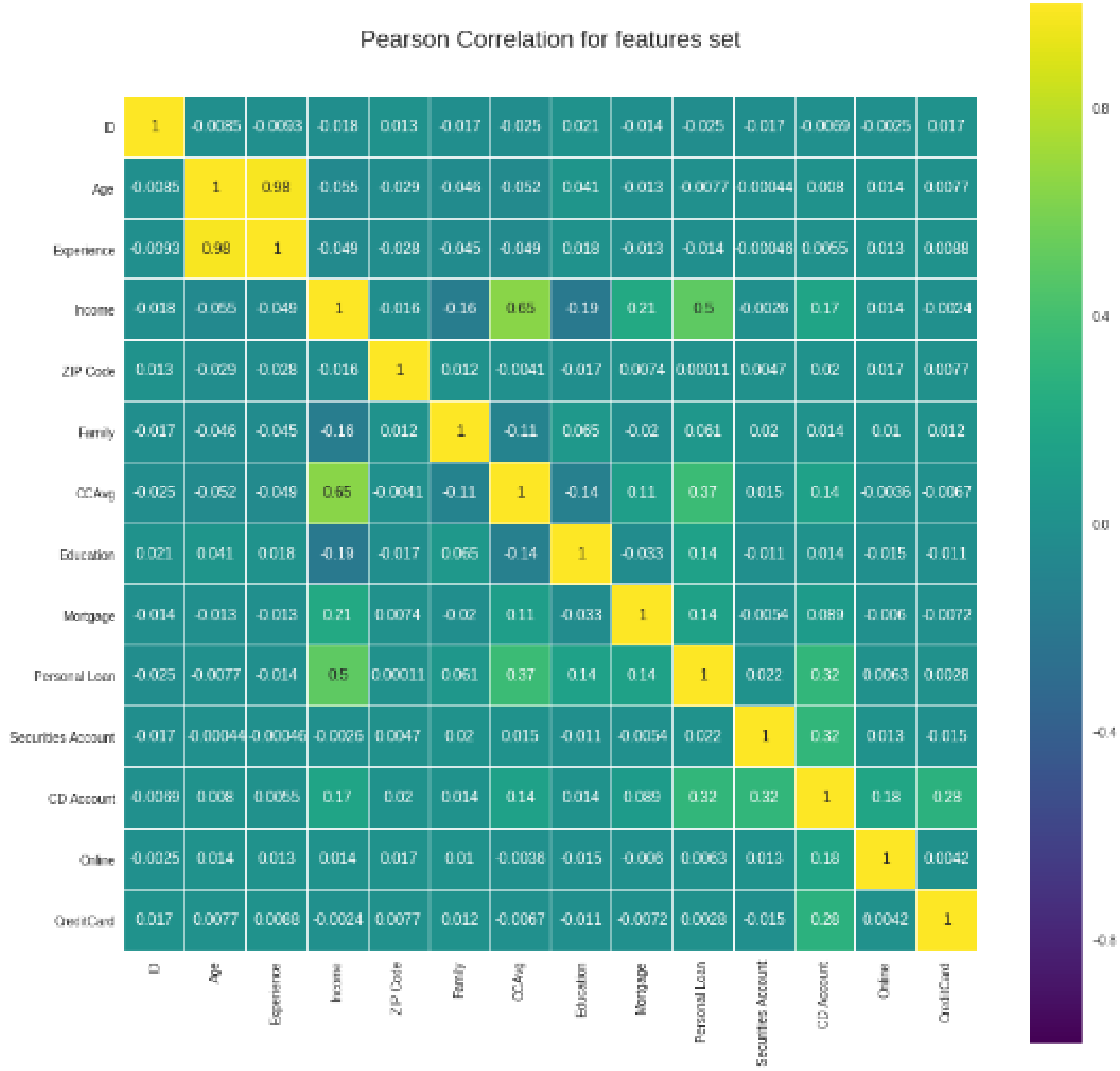
- Experience feature has some missing values, because experience cannot be **negative**
- Income is right skewed
- Average Family size of the customer is around 2 people
- Average credit card spending per month is slightly right skewed
- Morgage distribution seams to have an outlier

Bivariate Analysis

Constructing Pearsons correlation heatmap using seaborn module.

```
colormap = plt.cm.viridis # Color range to be used in heatmap
plt.figure(figsize=(15,15))
plt.title('Pearson Correlation for features set', y=1.05, size=19)
sns.heatmap(df.corr(),linewidths=0.1,vmax=1.0,
            square=True, cmap=colormap, linecolor='white', annot=True)
```

The output is below:

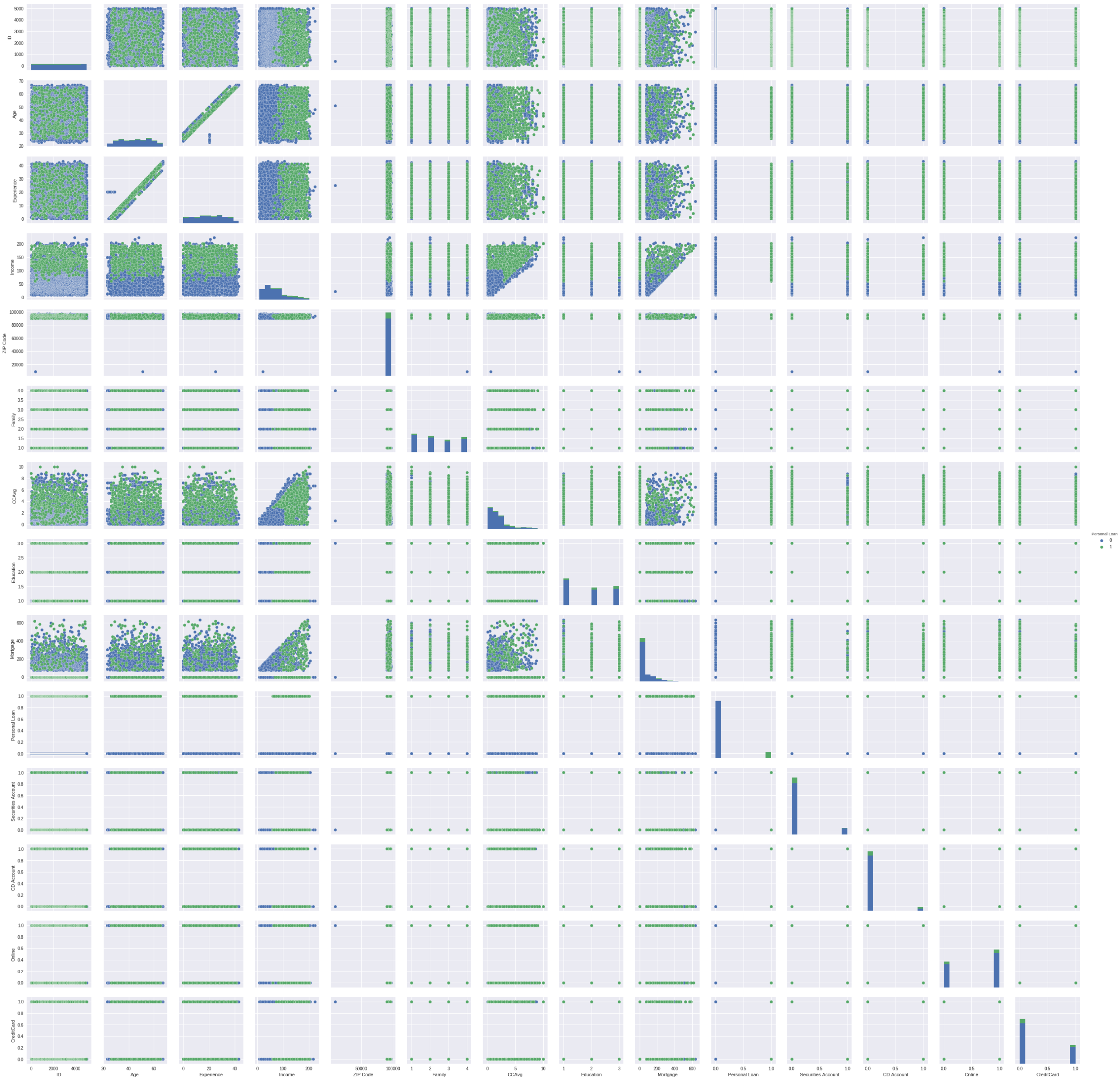


- There is a **very strong positive correlation** between Age and Experience
- And a **weak positive correlation** between CCAverage and Income

All these analysis tells us that we could potentially keep Age or Experience.

Pair Plot Analysis

The pair plot gives us pictorial representation of the distribution



Analysis of Target Variable

Insights

- The data is strongly biased towards customer that have not take the loan
- Only 9.6% of the customers have taken the loan

Decision Tree Classifier Vs KNN Classifier

Decision Tree accuracy score = 97.54%

KNN accuracy = 90.42%

- We will use Decision Tree classifier with entropy criterion because is gives much better accuracy over KNN.
- Decision Trees are also very flexible, easy to understand, and easy to debug.
- KNN doesn't know which attributes are more important i.e. when computing distance between data points (usually Euclidean distance or other generalisations of it), each attribute normally weighs the same to the total distance. This means that attributes which are not so important will have the same influence on the distance compared to more important attributes.

Ensemble techniques to improve the performance

We have a good accuracy score using DT, but we also have to keep in mind the fact that the data we have is highly biased. We will try to get better accuracy using some of ensemble techniques.

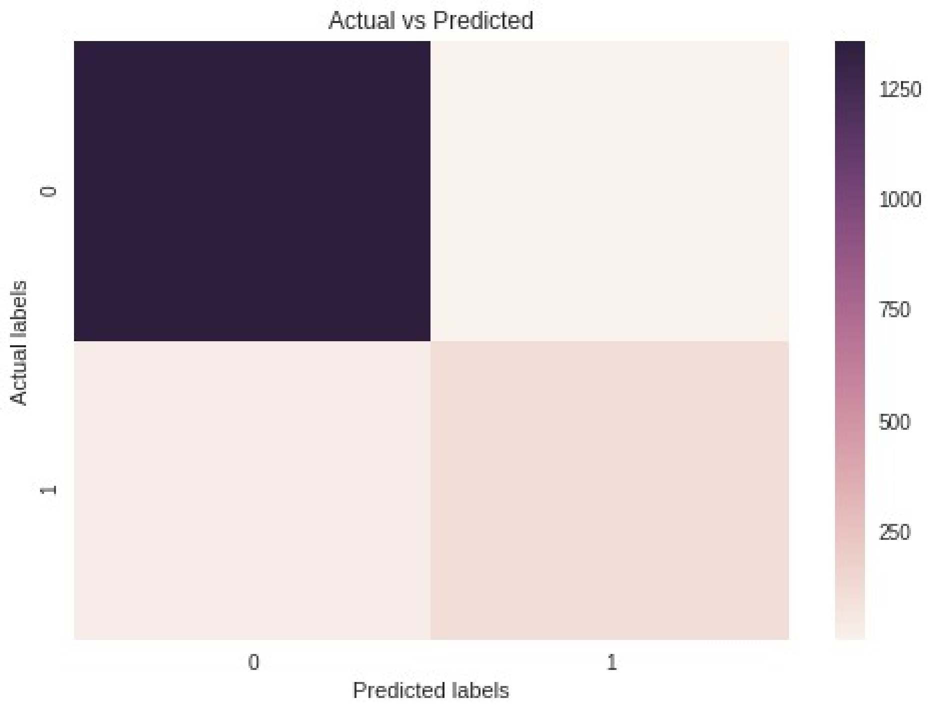
We used Random Forest Classifier from scikit learn module of python.

The best accuracy was obtained with a maximum tree depth =17 and number of trees = 19

```
# Using best hyper parameters
rf = RandomForestClassifier(max_depth=17, n_estimators=19, random_state=seed)

yPredicted = rf.fit(xTrain, yTrain).predict(xTest)
```

This confusion matrix heat map sums up the result



Conclusion

We got a very good accuracy using the Random Forest Classifier. The main challenge with this data is that the given data is highly biased toward Loan Not Taken. This can be fixed by collecting more samples.