Case Study on Predicting loan likelihood.

Overview

- Data Cleaning
- Data Description and preview
- Business Intelligence Tasks
- Data Exploration and Visualisation
- Exploring Insights
- Business Benefits
- Modelling
- Likelihood Classification
- Conclusion

Data Cleaning

- Removed duplicated rows from the datasets
- Replaced invalid data points with valid entries
  - Ex: gender (f, fem, male, m,) were replaced to (0,1) {0 – female, 1- male}
- Changing the column names to appropriate names
- Converted few columns to absolute values as it seems unreal for negative values.
  - Ex: Number of products held column had negative values, Converted them to positive.
- Set limits to Age column as entries such as 200 are unreal.
- Merged all the datafiles into a single data frame

- *Note: please find "Data_Cleaning.ipynb file for viewing the code for these tasks*

Data Description and preview

- Cleaned and Merged DataSet description:
- 10000 rows and 12 columns

	Client ID	Age	Gender	County	Income Group	Average_CA_transaction_amount	No_of_products_held	Held Loan previously	Num Transactions	Last TXN Amount	Merchant Code	Last Transaction Narrative
0	1	36	1	cork	10001-40000	58	4	1	0.0	NaN	NaN	NaN
1	2	43	1	cavan	0-10000	2663	4	0	17.0	83.66	7211.0	THE BRIDGE LAUNDRY WICKLOW TOWN
2	3	32	0	dublin	10001-40000	46	2	0	25.0	526.18	3667.0	LUXOR HOTEL/CASINO LAS VEGAS NV
3	4	52	1	louth	40001-60000	0	2	1	13.0	70.68	5712.0	HARVEY NORMAN CARRICKMINES
4	5	63	0	kilkenny	60001-100000	126	1	0	39.0	259.07	5999.0	PAYPAL *PETEWOODWAR 35314369001

Business Intelligence Tasks

- **Question 1**: How Many Customers above 50 years old have taken up a loan?

- **Answer:** 98

Note: please find "BI_Solutions.ipynb file for viewing the code for BI tasks

Business Intelligence Tasks

- **Question 2 :** How Many Females aged 30 to 40 have more than 2 products?

- **Answer :** 513

Assumptions: Few assumptions taken as below in the dataset before filtering out data

1. 0 – female
2. 1 – male
3. Values such as 30 and 40 for the age attribute are not included (Therefore range is from 31 to 39 years old customers are included in this answer)

Note: please find "BI_Solutions.ipynb file for viewing the code for BI tasks

Business Intelligence Tasks

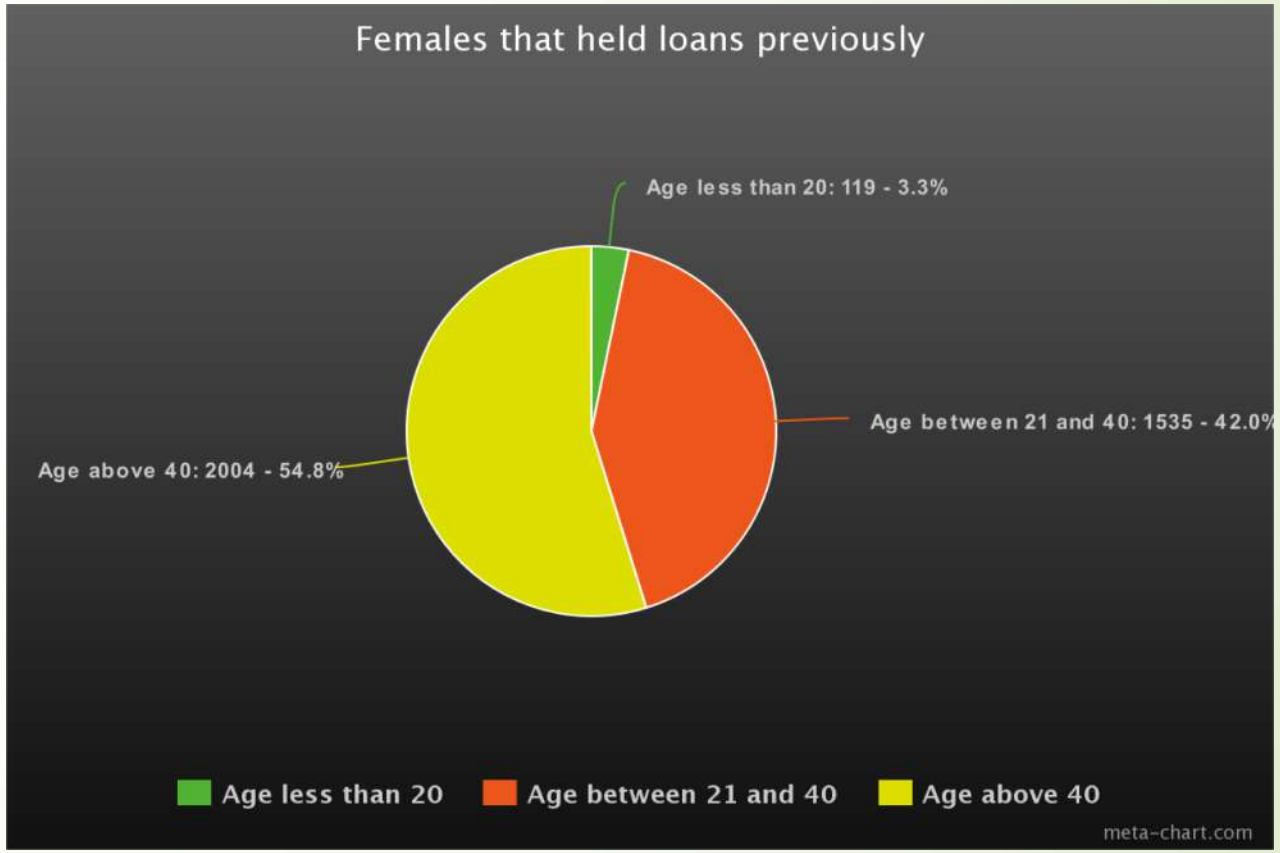
- **Question 3 :** What is the average number of Current Account(CA) Transactions for males who had a previous Loans?

- **Answer : 19 (**18.65 rounded off)

Note: please find "BI_Solutions.ipynb file for viewing the code for BI tasks

Business Intelligence Tasks

- **Question 4 :** How many females did not have a previous loans and who are aged
  - Less than 20
  - 21 to 40
  - 40+
- **Answer :**
  - **119** females with age less than 20 (inclusive 20 years old)
  - **1535** females with age in between 21 to 40 (inclusive)
  - **2004** females above 40 years old (exclusive 40)



Note: please find "BI_Solutions.ipynb file for viewing the code for BI tasks

Data Exploration and Visualisation

- Identified 3278 Null values in “Last TXN Amount”, “Merchant Code” and “Last Transaction Narrative”.
- Therefore, I have removed these attributes from the dataset before exploration.

- *Note: please find “Exploratory_Data_Analysis.ipynb file for viewing the code and more information on Exploration and visualization.*

Data Exploration and Visualisation

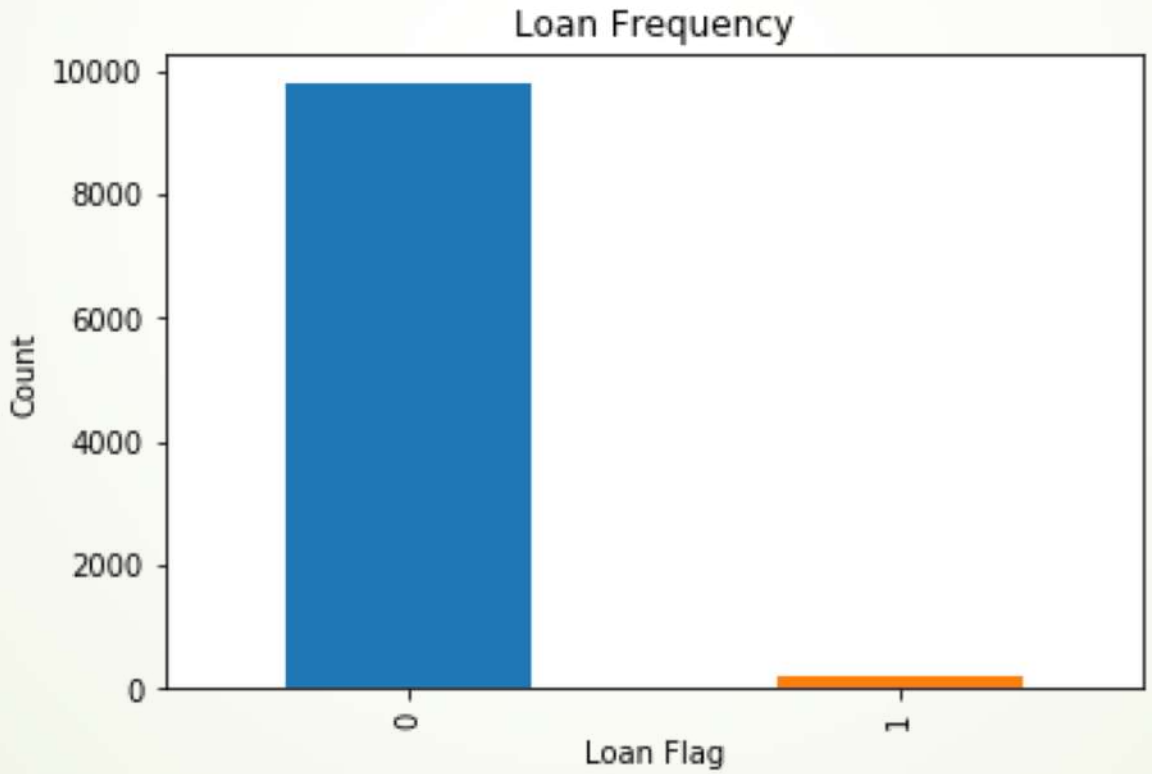
Observations:

- There is a ***significant difference in the mean of Avg CA transactions*** between the clients who took loan and who did not take up loan
- Also it is evident that ***74% of the previous loan holders*** have taken up the new loan
- Clients with ***greater average number of transactions*** have taken up loan
- All the attributes have similar mean with respective gender, This shows that ***gender is not a significant variable***
- Almost all the attributes have similar mean with respective Income Group except Loan Flag, So ***Income Group might have some impact***.
- Alsmost all the attributes have similar mean with respective to Held Loan previously column except Loan Flag, However, So ***Held Loan previously might have some impact*** as well.
- ***No major correlation between the attributes.***

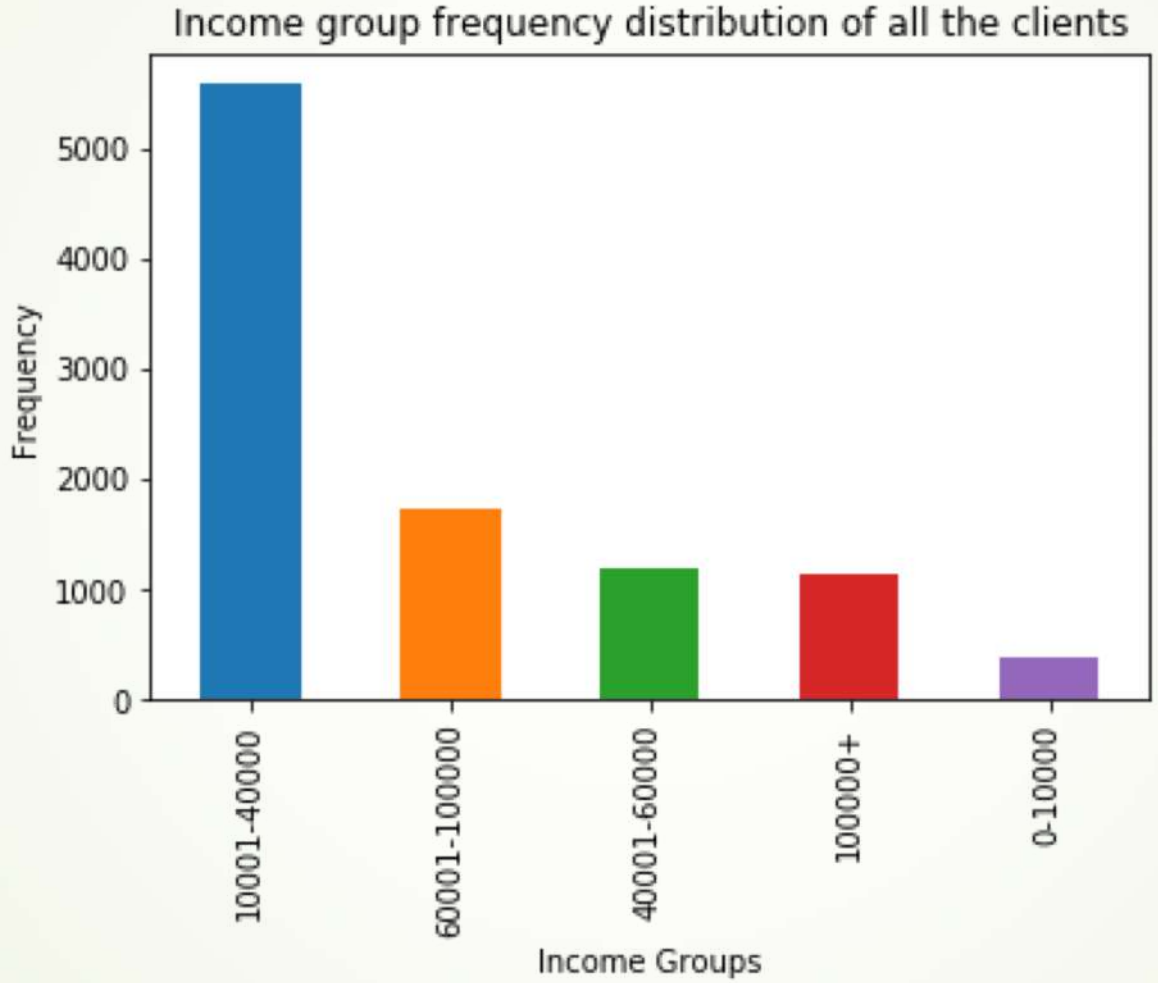
- *Note: please find "Exploratory_Data_Analysis.ipynb file for viewing the code and more information on Exploration and visualization.*

Data Exploration and Visualisation

- Unbalanced Loan flag classes

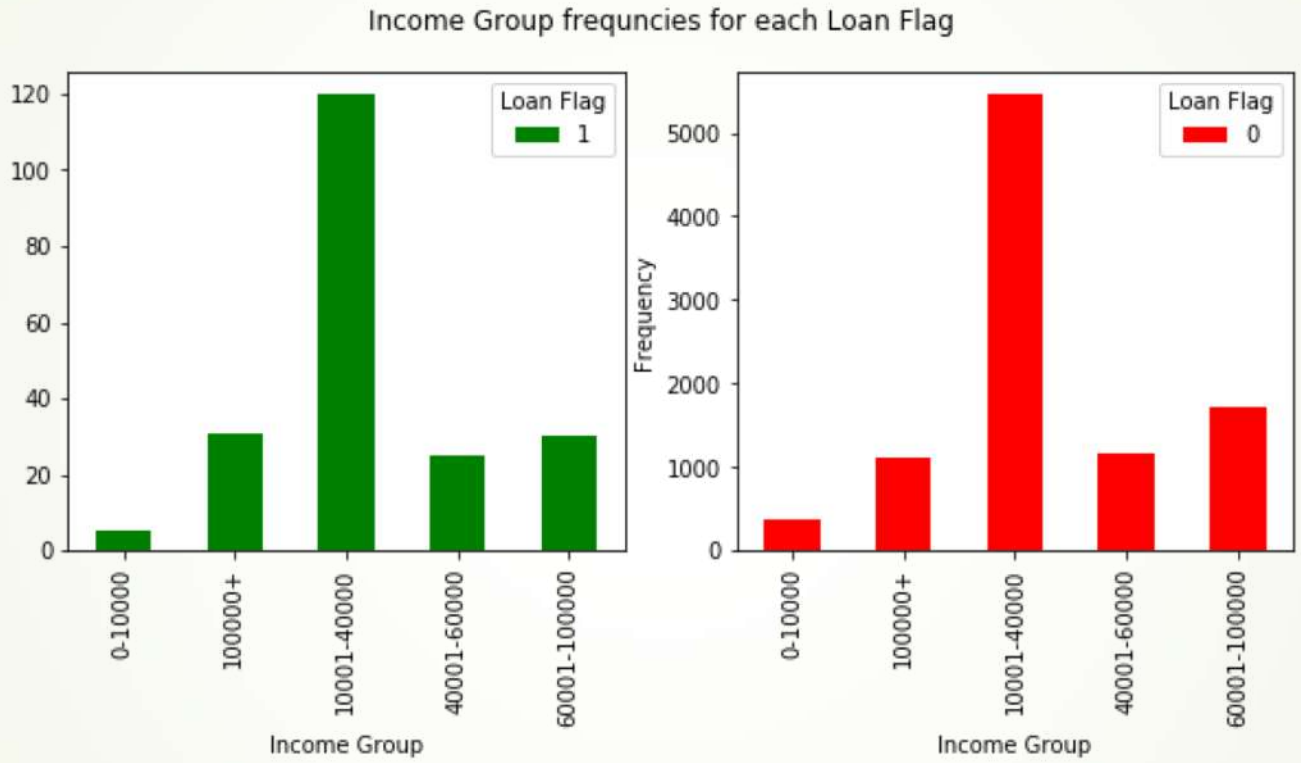


Data Exploration and Visualisation



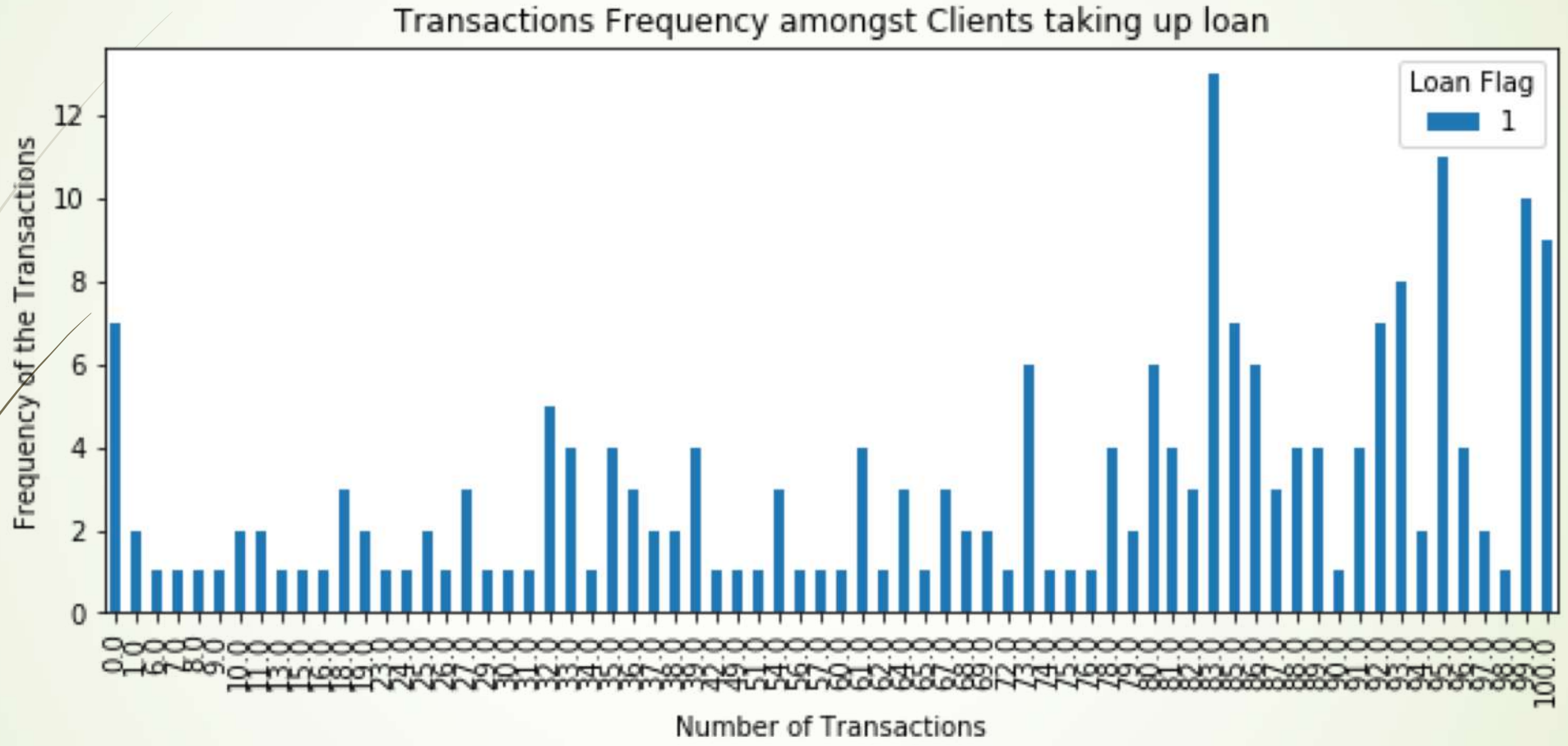
- Income Group 10000 - 40000 is the income for most of the clients

Data Exploration and Visualisation



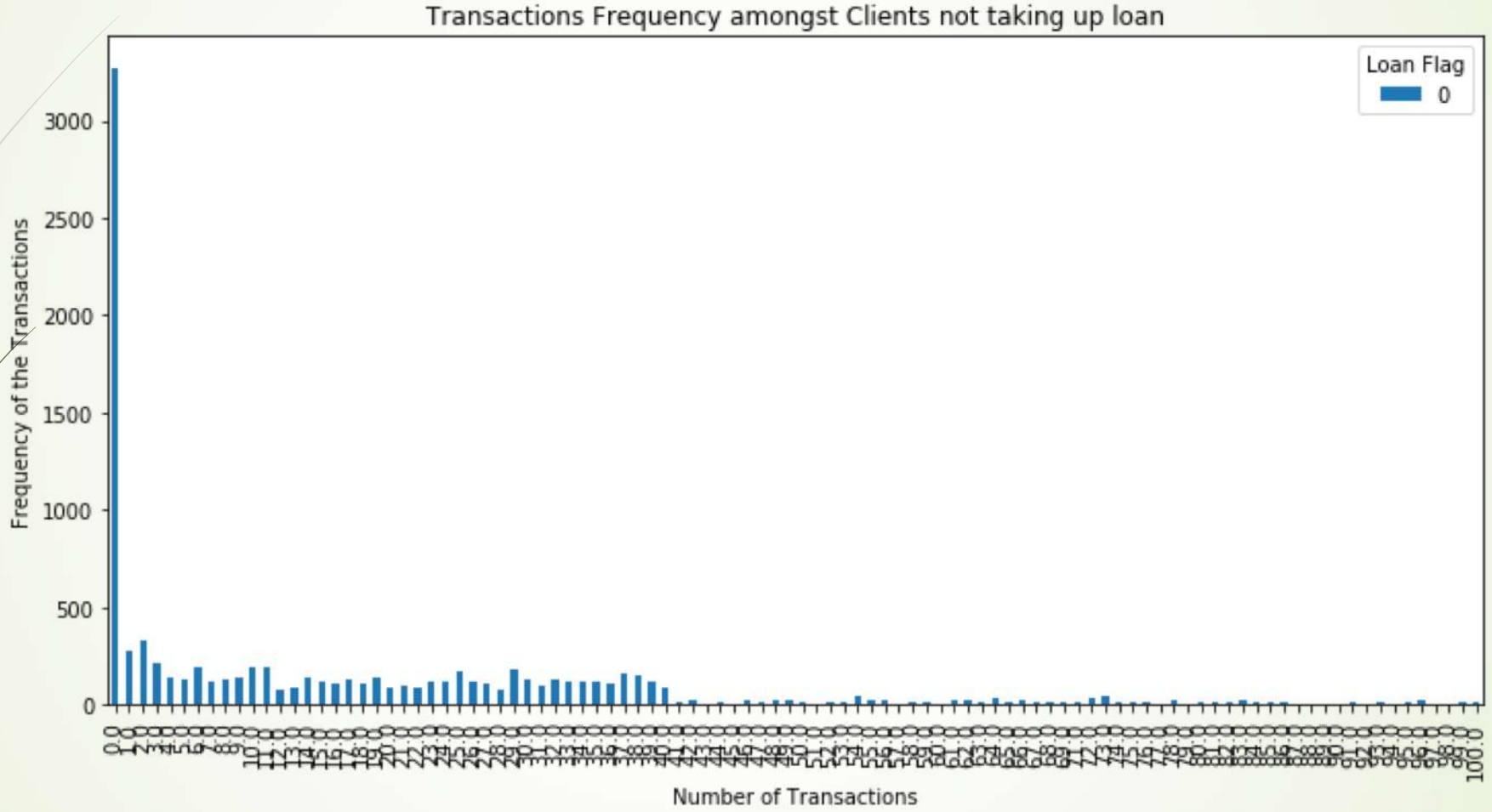
- Even with the clients who took up loan income range 10000 – 40000 is in high frequncy

Data Exploration and Visualisation



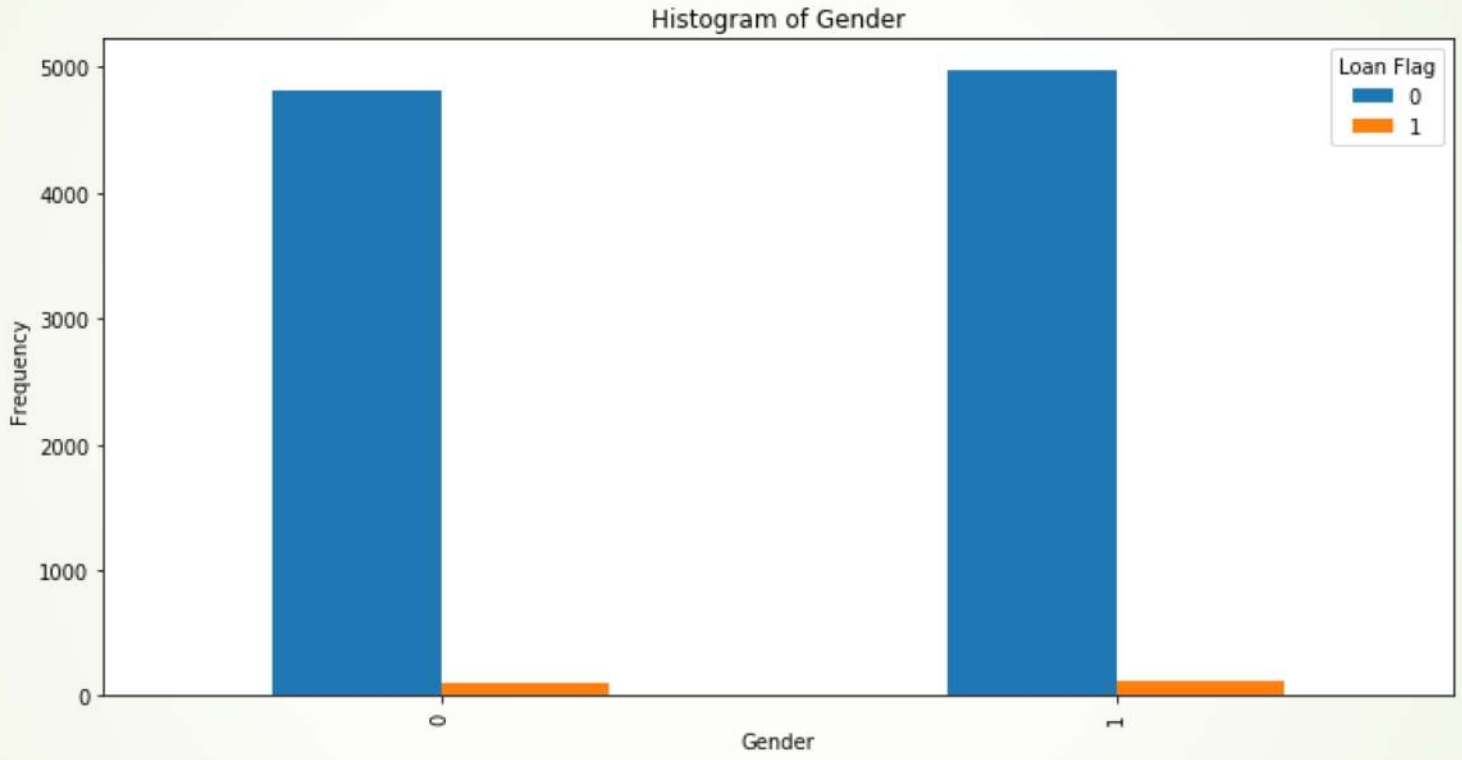
- Higher the number of transactions, more the customers who took up loan

Data Exploration and Visualisation



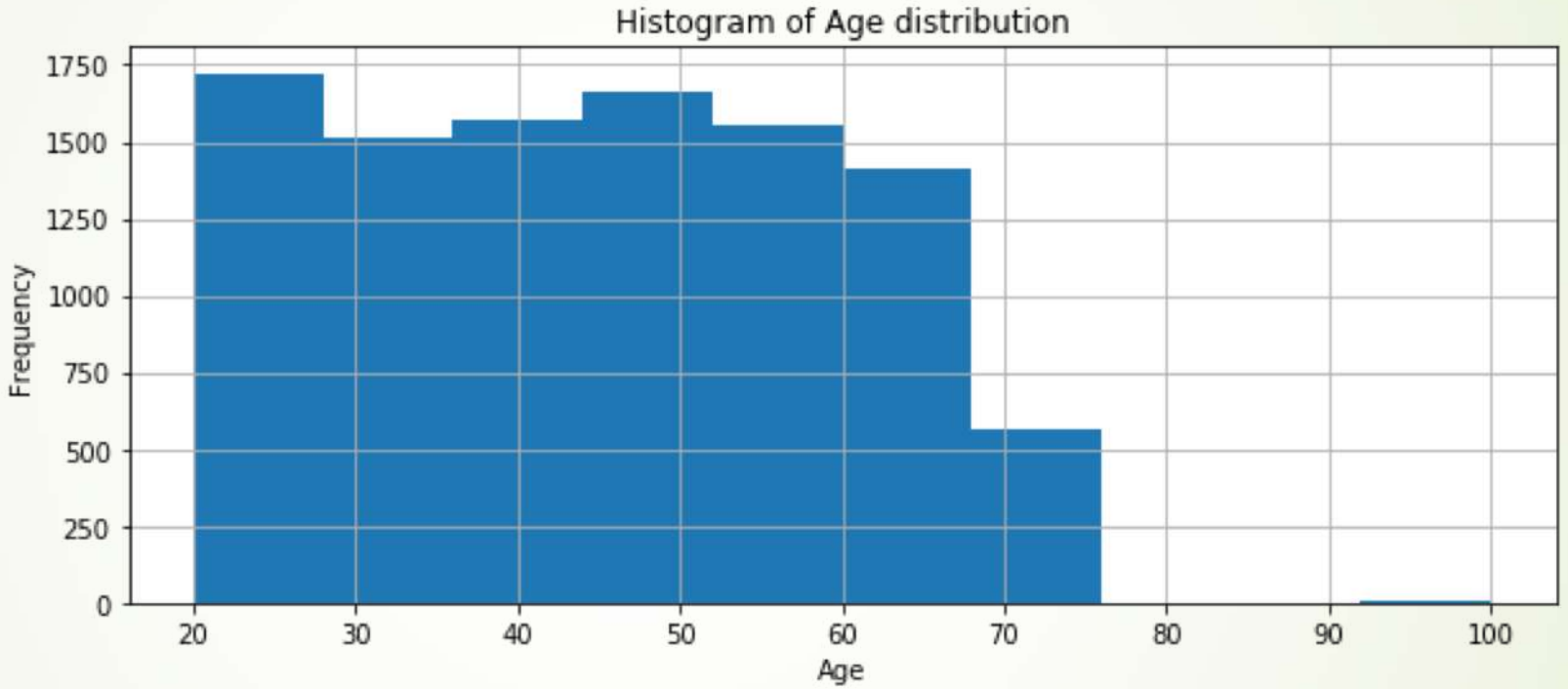
- Less the number of transactions, Then customers are not taking up loan.
- Therefore, There is a direct relation between number of transactions and Loan Flag

Data Exploration and Visualisation



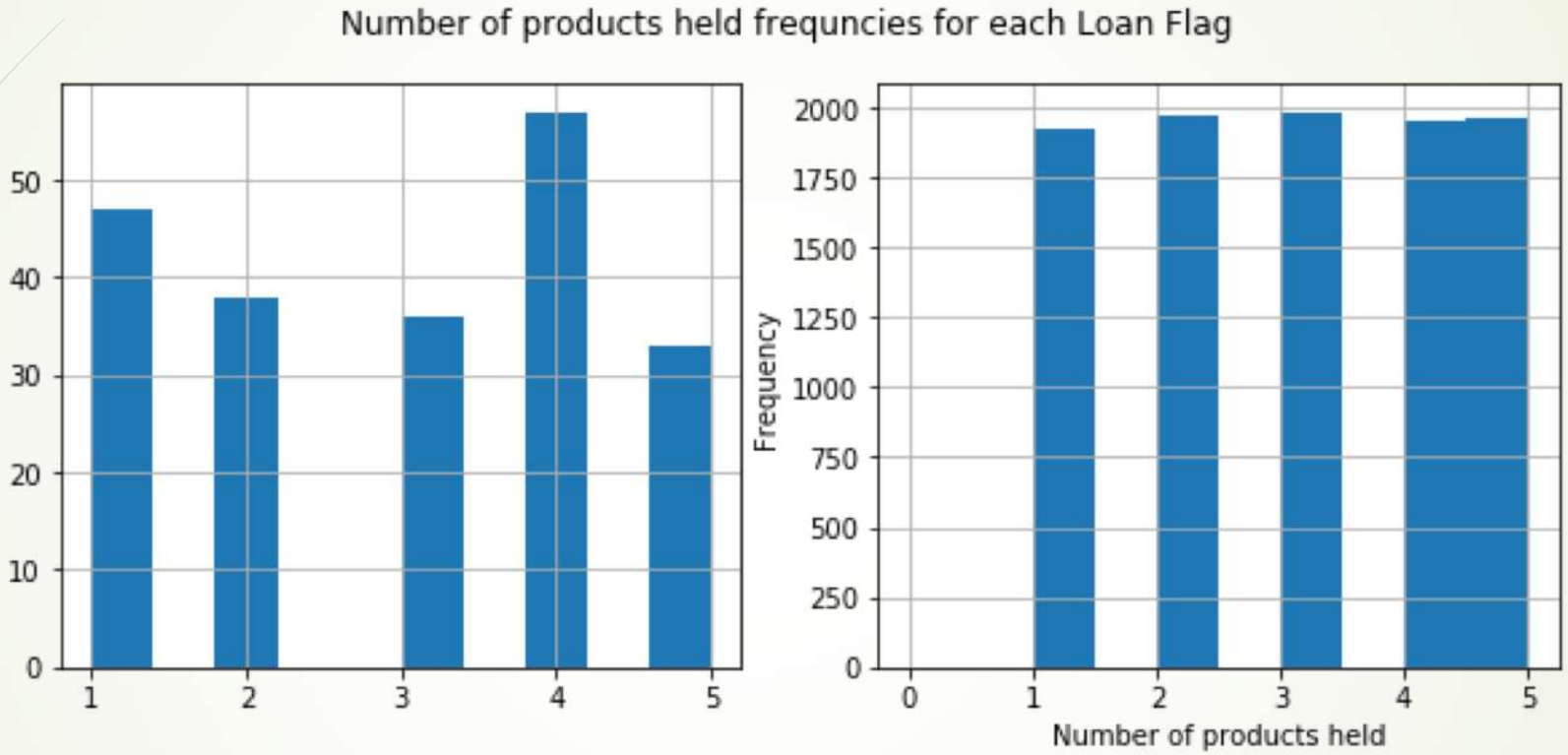
- Looks similar with both genders, so clearly, gender does not hold any direct impact on Loan Flag

Data Exploration and Visualisation



- Slightly left skewed, but doesnot seem to have major impact.

Data Exploration and Visualisation



- Again, Number of products held seams to be similar in both groups. Therefore, this might not have major impact on the Loan Flag

Data Exploration and Visualisation

- All the graphs were plotted to understand the relation between the independent variables and dependent variable visually.
- However, Statistical tests are performed later, to evaluated the actual impact and to decide the important attributes

Note: please find "Exploratory_Data_Analysis.ipynb file for viewing the code and more information on Exploration and visualization.

Exploring Insights

- Firstly, divided the Number of transactions into bins to understand the relationship between "number of transactions" and "Average CA transaction amount" variables.
- Clients that have high "number of transactions" (i.e. greater than 40), 13% of these clients have taken up loan
- Amongst the clients that have "average CA transaction amount" greater than 1000 euros, 6.9% of them have taken up loan.
- Even if the" Average CA Transaction amount" is less than 1000, but if client's "number of transactions" are greater than 40 or if they have held a loan previously, then with high probability (Almost probability = 1), they have taken up a new loan.

Exploring Insights

- Top 5 counties with relatively higher percentage of clients that have taken up loans are as follows:

County	Percentage
Maynooth	100 %
Kilkenny	3.7 %
Clare	3.7 %
Waterfod	3.6%
Wicklow	2.6 %

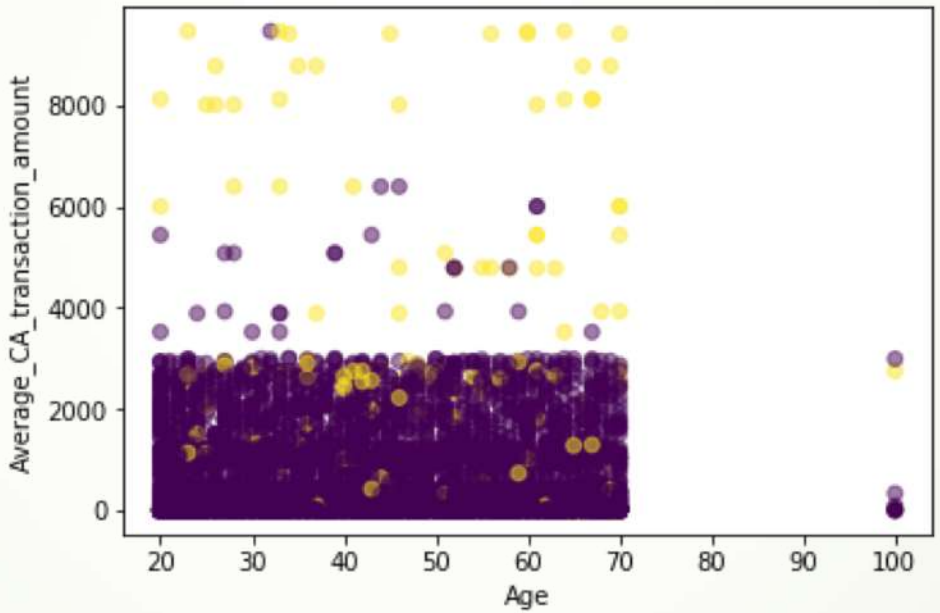
Business Benefits

Loan Likelihood:

- Target clients with combination of any two of the above factors
  - Clients with Avg transaction amount greater than 1000
  - Clients with Number of transcation greater than 40
  - Clients who held loans previously
  - Clients with Income group (10000 to 40000)

Modelling

- This is biased class problem – There are only few entries of loan takers compared with clients who have not taken loans.

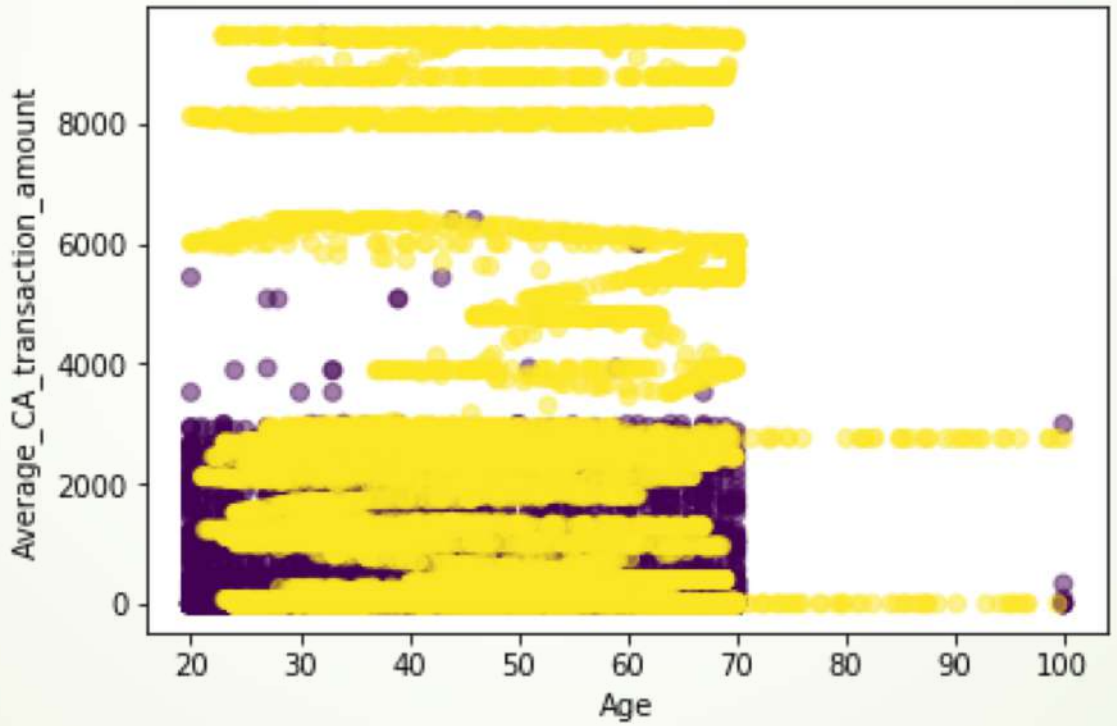


As this is clearly visible when plotted.

Note: please find "Data_Modelling.ipynb file for viewing the code for modelling.

Modelling

- Therefore, I have synthetically generated the new data for balancing the classes.
- Distribution of classes after Synthetic fata generation



Modelling

- Used multiple classifiers such as
  - Logistic Regression
  - Decision Tree
  - K- Nearest Neighbours Classifier
  - Support Vector Classifier

I have trained the above mentioned classifiers with70% of the data and validated it with 30 %.

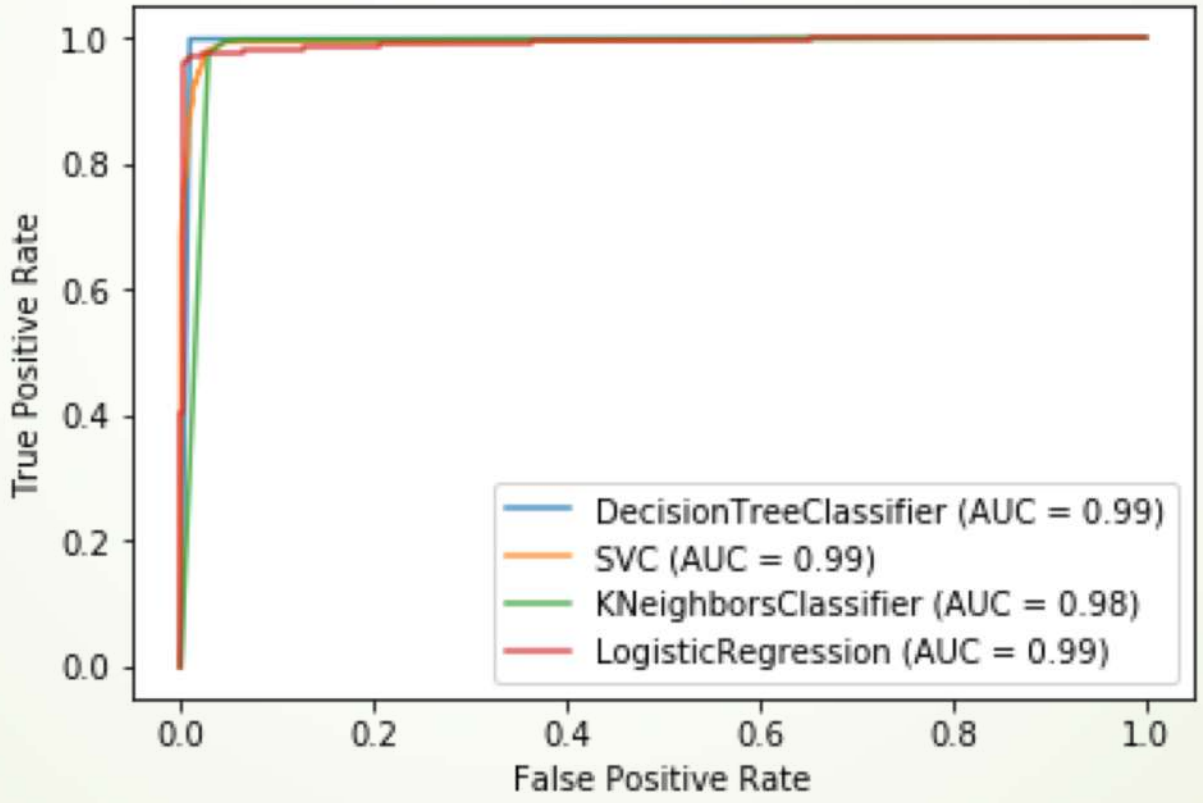
Modelling

- Results:

Model Name	Precision	Recall	F1- Score
Logistic Regression	0.97	0.96	0.97
Decision Tree	0.99	0.99	0.99
KNN	0.98	0.97	0.97
Support Vector Machine	0.94	0.94	0.94

Modelling

- Evaluation: ROC Curve



Modelling

- Training the final predictor for the test cases.
- I choose, Logistic Regression model as final predictor because it gives maximum likelihood as well.

- Now, Trained the Logistic regression model with all the model data
- Predicted the Loan_Flag for test cases
- Predicted the probabilities of each test case data point as well.

Note: please find "Data_Modelling.ipynb file for viewing the code for modelling.

Note: please find "test_Data_Cleaning.ipynb file for viewing the code for cleaning and merging the test dataset.

Modelling

Predicted Loan_Flag as well as likelihood ratios are saved and submitted as Test_predicted_Loan_Flag_results.csv file.

	Client ID	predicted_Loan_flag	probability_of_1	Likelihood
0	10001	0	0.015874	Very Low Likelihood
1	10002	0	0.006144	Very Low Likelihood
2	10003	0	0.189562	Very Low Likelihood
3	10004	0	0.140092	Very Low Likelihood
4	10005	0	0.019913	Very Low Likelihood

Preview of the Test_predicted_Loan_Flag.csv file

Likelihood Classification

- I have divided the probability of a customer taking up loan into 5 Bins.

- Less than 5% probability – Very Low Likelihood
- Between 5% to 10% – Low Likelihood
- Between 10% to 40% – Medium Likelihood
- Between 40% to 60% – High Likelihood
- Above 60% – Very High Likelihood

Conclusion

- Cleaned the dataset correctly and merged into one dataset
- Answered all the Business Intelligence questions (calculated using python)
- Explored and extracted meaningful insights from the data
- Highlighted target customers to sell the loans
- Built and evaluated a predictor to predict the loan flag
- Classified the test set customers based on likelihood
- Defined loan uptake rate/ likelihood define what you expect the loan uptake rate (%) to be (%) for each group

Future Work

- Can explore the deleted columns for text analysis
- I can normalize the data before modeling for faster computing.
- Can explore more with county dataset

Thanks for your time

Hope to hear back from you soon…

- Swathikiran S