Project 2: Personal Loan Prediction

Given dataset from Thera Bank has 5000 records with features such as customer demographics, customer's relationship with bank and customer's response to past Loan campaign. Objective is to build a classification model from the dataset to predict who will buy loan and who will not.

Analysis of data is given below as summary.

Summary

From attribute information given in problem statement, we can deduce:

- Age, Experience, Income, CCAvg, Mortgage - are numerical variables
- Personal Loan - Target variable (Dependent)
- Rest all are categorical variables.
    - ID - Customer ID, may not give any value towards target
    - Education has different levels or order
    - Family - Its family size, we can consider this also as ordered variable.
    - ZIP code - categorical, no order.
    - Securities Account, CD Account, Online, Credit card: These contain yes/no values, so do not have any order and we can say these are already in one-hot encoded form.

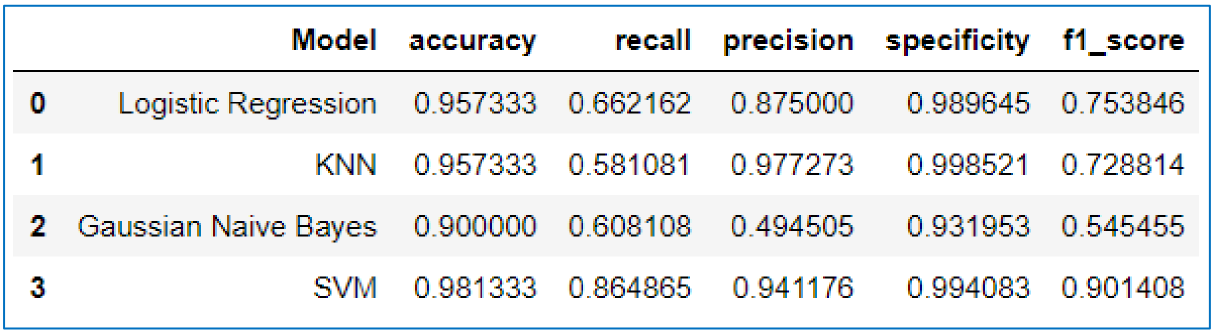
Data analysis:

- ✓ Total 5000 rows, also 5000 IDs, which makes ID as valueless for our target. So we can drop the ID column.
- ✓ 480 out of 5000 people have taken loan in the past, which makes our dataset imbalance, may lead to biased result.
- ✓ No null or missing values in any column.
- ✓ Experience has negative values, which is not possible practically. So we will replace them with absolute values.
- ✓ CCAvg seems right skewed (Max - Q3 is high). For most of the people, avg spending on credit card is less.
    - Avg CC spending and CD account impacts a little on personal loan decision.
- ✓ Income also seems right skewed. More people have less income range and less people have more income range.
    - Income and CCAvg are positively correlated.
    - Income also impacts the decision of taking personal loan (50% correlation between them).
    - Mostly Mid-High income group people have taken loan.
- ✓ Mortgage is extremely right skewed. Most of the people have zero Mortgage value. Median lies at 0, whereas Q3 at 101 and Max at 635. So we will apply log transformation to bring the values to a lower range for better comparison.

- ✓ Age and Experience data seems fine and are highly linearly correlated. As age increases, experience increases.
    - People from ~20 to ~70 age group with balanced amount in each age group.
    - Experience varies between 0 to ~45years with equal amount from each exp.
- ✓ ZIP code has one outlier (9307) as compared to other zip codes. We will remove the entry and proceed with rest of the data.
- ✓ Dataset has more families with size of 1 and less families with size of 3.
    - Approximately 7%-13% people from each category have taken loan in the past.
    - Comparatively more people with family size 3 and 4 have taken loan than people with family size 1 and 2.
    - More specifically, in higher income groups, mostly people with family size 3 and 4 have taken loan in the past.
- ✓ Most of the people are undergraduate in the dataset.
- ✓ Most of the people do not have Securities Account and CD account.
- ✓ Most of the people use internet banking.
- ✓ Most people do not hold a credit card (from Universal Bank).
- ✓ People are from different localities within a city as per zip code.

- ✓ Finally we split the data at 70-30 ratio and scaled with StandardScaler.
- ✓ Trained with various models and measured various performance metrics and listed down in a table (as shown below).

Performance measures for various models

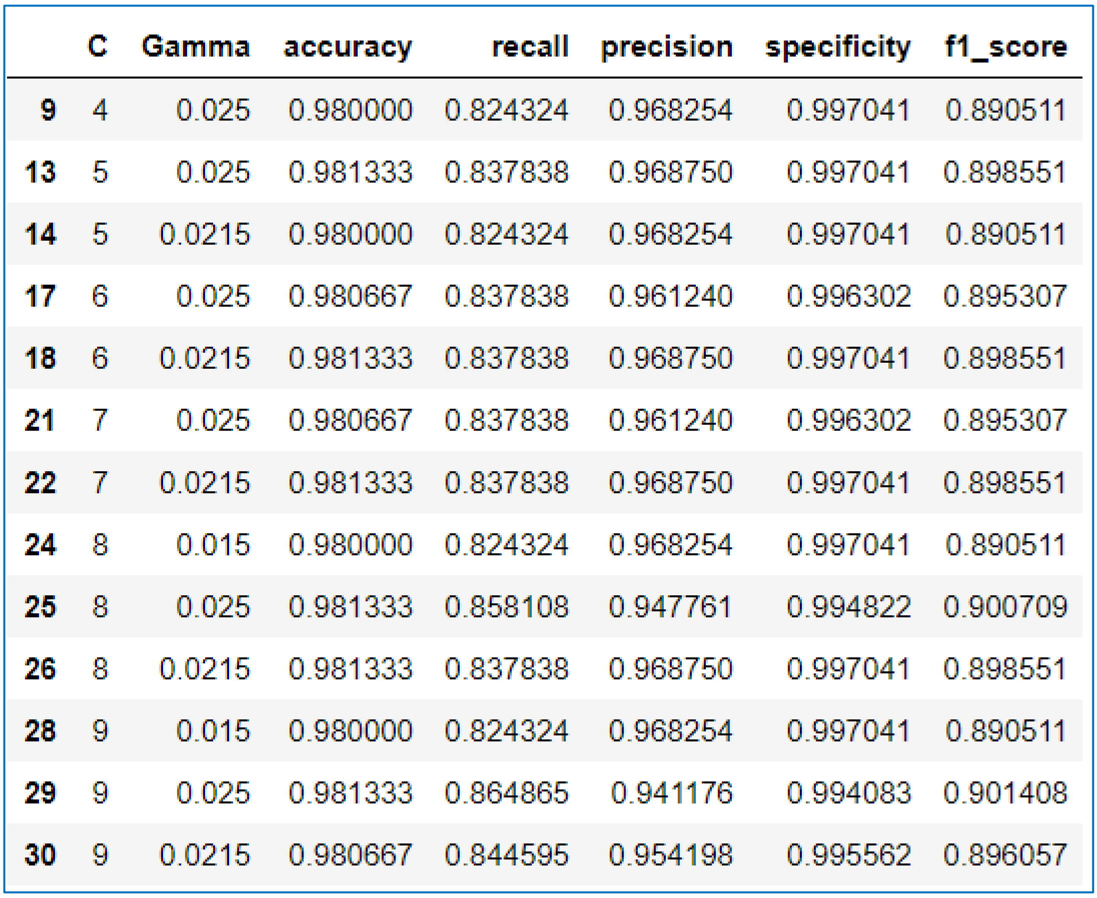


	Model	accuracy	recall	precision	specificity	f1_score
0	Logistic Regression	0.957333	0.662162	0.875000	0.989645	0.753846
1	KNN	0.957333	0.581081	0.977273	0.998521	0.728814
2	Gaussian Naive Bayes	0.900000	0.608108	0.494505	0.931953	0.545455
3	SVM	0.981333	0.864865	0.941176	0.994083	0.901408

SVM performed well among all the models we trained, with accuracy of 98.13%, Recall 86.48%, Precision 94.11%, Specificity 99.40% and **F1-score 90.14%**.

Logistic regression, KNN and NB show biased results with Specificity at higher side as dataset has more number of –ve class entries compared to +ve class. So Recall is at lower side (~58% to ~66%).

For SVM, values for C and Gamma are taken as **C=9 and Gamma=0.025** after tried various values. Current values are chosen based on **higher value of F1-score** (as shown below).



	C	Gamma	accuracy	recall	precision	specificity	f1_score
9	4	0.025	0.980000	0.824324	0.968254	0.997041	0.890511
13	5	0.025	0.981333	0.837838	0.968750	0.997041	0.898551
14	5	0.0215	0.980000	0.824324	0.968254	0.997041	0.890511
17	6	0.025	0.980667	0.837838	0.961240	0.996302	0.895307
18	6	0.0215	0.981333	0.837838	0.968750	0.997041	0.898551
21	7	0.025	0.980667	0.837838	0.961240	0.996302	0.895307
22	7	0.0215	0.981333	0.837838	0.968750	0.997041	0.898551
24	8	0.015	0.980000	0.824324	0.968254	0.997041	0.890511
25	8	0.025	0.981333	0.858108	0.947761	0.994822	0.900709
26	8	0.0215	0.981333	0.837838	0.968750	0.997041	0.898551
28	9	0.015	0.980000	0.824324	0.968254	0.997041	0.890511
29	9	0.025	0.981333	0.864865	0.941176	0.994083	0.901408
30	9	0.0215	0.980667	0.844595	0.954198	0.995562	0.896057

Other values for C and Gamma can be selected if we look for higher values of Recall, Precision, and/or Specificity.