**HBFC Bank - Personal Loans**

This case is about a bank (HBFC Bank) which has a growing customer base. Majority of these customers are customers having deposits (Saving, Current) and Term Deposit (including Fixed Deposits and Recurring deposit accounts) in the bank. The number of customers who are also borrowers (customers having loan accounts) 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 customers having deposits to personal loan customers (while retaining them as depositors).

The bank wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. For doing that, the first step in this regard is to explore the previous data and drill insights.

You have been provided with a dataset of 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan).

**You are brought in as a consultant and your job is to explore the data to understand the variables and the impact they have had on Personal Loans so that the bank can leverage the insights to reach out to the right customers who have a higher probability of purchasing the loan.**

Data Dictionary: The data consists of the following variables (columns).

| | |
|---|---|
| ID | Customer ID |
| Age | Customer's age in years |
| Experience | Years of professional experience |
| Income | Annual income of the customer ($000) |
| ZIPCode | Home Address ZIP code. |
| Family Members | 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) |
| 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? |
| TD Account | Does the customer have a Term deposit (Including Fixed and Recurring Deposits) account |
| Online | Does the customer use internet banking facilities? |
| CreditCard | Does the customer use a credit card issued by the bank? |

---

The data set consists of data of 5000 customers of HBFC Bank. The objective of the project is to explore the data so that the bank can understand and uncover patterns and insights which can be used to increase the business its Personal Loan segment.

Categorical Variables – Customer ID, ZIP Code, Education, Personal Loan, Securities Account, TD Account, Online, Credit Card, Income_Category

Numeric Variables – Age, Experience, Income, Family Members, CCAvg, Mortgage

One categorical variable “Income Categorical” is there stating the income bracket.

---

**NOTE: The below solution represents one approach of solving the caselet, it is a good practice to explore the data as much as possible to derive the true picture. For this assessment, different approaches in submissions have been considered.**

**Questions:**

1. What percentage of the bank's customers (according to the data) have availed Personal Loans? **(3 points)**

Percentage of PL Customers

| Row Labels | Count of ID |
|---|---|
| No | 90.40% |
| Yes | 9.60% |
| **Grand Total** | **100.00%** |

Of the data provided, 9.6% of the customers have accept the Personal Loan from the bank after the last campaign.

2. Generate a table with min, max, median & average for all numeric variables (age, experience, income, family members, CCAvg, Mortgage). What are your observations? **(6 points)**

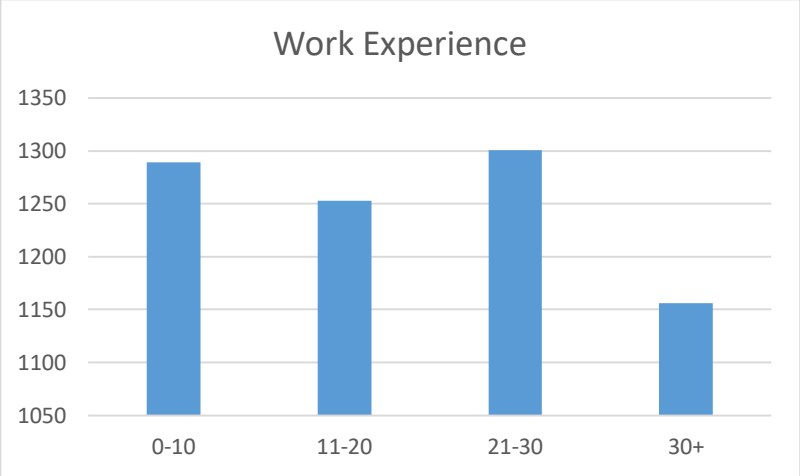
| Quantity | Age (in years) | Experience (in years) | Income (in K/year) | Family members | CCAvg | Mortgage |
|---|---|---|---|---|---|---|
| **Min** | 23.0 | 0.0 | 8.0 | 1.0 | 0.0 | 0.0 |
| **Median** | 45.0 | 20.0 | 64.0 | 2.0 | 1.5 | 0.0 |
| **Average** | 45.3 | 20.1 | 73.8 | 2.4 | 1.9 | 56.5 |
| **Max** | 67.0 | 43.0 | 224.0 | 4.0 | 10.0 | 635.0 |

**Mortgage**: Median is equal to zero; atleast 50% people have not got any mortgage with the bank.
**Age**: The range of values of the age depicts that data is recorded for the majority of the real user age group, i.e. 23-67 years. Median age is 45 suggesting that 50% of the customers are less than 45 years old.
**Experience**: It is good to find that all kind of experiences are included in the study; beginning from freshers to seasoned professionals. Also mean/ median seems to lie in centre of this range.
**Income**: 50 percent of people have income between 8 and 64 and rest of 50% people have income between 64 and 224, it seems top 50% people income range is quite high. Probably few high earning candidates are present in the dataset. Same seems to be followed for CCAvg.

3. Create a new categorical variable for Experience using 4 categories –
   - 0 to 10 years
   - 11 to 20 years
   - 21 to 30 years and
   - 30+ years.

Plot a bar graph for this new categorical variable

[Hint – You may make use of if else/nested if statements to accomplish this tasks. You can refer how Income_Category has been created in the dataset] **(5 marks)**




We can see the highest spike in the 21-30 bin, and the lowest one in 30+ bin suggesting that most of the customers have work experience between 21-30 years, and few of them have work experience of more than 30 yrs.

4. Create a scatter plot of the Age and the Experience variable. What do you observe? **(5 marks)**


According to the data, as the age increases, the experience (working experience) also increases. minimum customer age (according to the data set) is 23 years and maximum customer age is 67 years.

5. What are the top 3 areas (ZIP Codes) where the bank's customers are located? **(5 Marks)**

| Row Labels | Count of ZIP Code |
|---|---|
| 94720 | 169 |
| 94305 | 127 |
| 95616 | 116 |

The biggest chunk of customers resides in ZIP Code 94720, followed by ZIP code 94305 and ZIP code 95616. The Bank can leverage it by asking them to refer their neighbours while giving the respective customer reward points/ gifts etc.

6. How many customers have a combination of Term Deposits and Credit Cards but not Personal Loan? **(3 Marks)**
   147 (Obtained using filtering, Term Deposit = Yes + Credit Card = Yes + Personal Loan =No)

   If we want to upsell Personal Loans to them, i.e., have them avail Personal Loans and also retain their deposits, we could offer them lower rates for Personal Loans since the bank has their Fixed Deposit maintained too.

7. What is the median income of the customers who have availed personal loans and compare it with the median income of those customers who have not availed personal loans? What do you infer? **(5 Marks)**

| | |
|---|---|
| Median Income (PL=Yes) | 142.5 |
| Median Income (PL=No) | 59 |

Median Income is significantly higher for customer who have availed personal loans than those who have not availed personal loans suggesting that Income could be a significant criteria contributing to Personal Loans, i.e., we could infer that Banks are more comfortable in lending to customer with higher incomes than customer with lower incomes.

8. Create 4 separate Pivot Tables. Summarise your data by percentages. **(8 marks)**

   - Education vs Personal Loan
   - TD Account vs Personal Loan
   - Online vs Personal Loan
   - Income_Category vs Personal Loan

[Hint: Please drag Personal Loan to the Columns area while creating the Pivot Table to get the required values]

| Education vs Personal loan | Row Labels | No | Yes | Grand Total |
|---|---|---|---|---|
| | Graduate | 87.03% | 12.97% | 100.00% |
| | Professional | 86.34% | 13.66% | 100.00% |
| | Undergraduate | 95.56% | 4.44% | 100.00% |
| | **Grand Total** | **90.40%** | **9.60%** | **100.00%** |

| TD Account vs Personal Loan | Row Labels | No | Yes | Grand Total |
|---|---|---|---|---|
| | No | 92.76% | 7.24% | 100.00% |
| | Yes | 53.64% | 46.36% | 100.00% |
| | **Grand Total** | **90.40%** | **9.60%** | **100.00%** |

| Online vs Personal Loan | Row Labels | No | Yes | Grand Total |
|---|---|---|---|---|
| | No | 90.63% | 9.38% | 100.00% |
| | Yes | 90.25% | 9.75% | 100.00% |
| | **Grand Total** | **90.40%** | **9.60%** | **100.00%** |

| Income_Category vs Personal Loan | Row Labels | No | Yes | Grand Total |
|---|---|---|---|---|
| | 0-50 | 100.00% | 0.00% | 100.00% |
| | 100+ | 63.86% | 36.14% | 100.00% |
| | 51-100 | 97.76% | 2.24% | 100.00% |
| | **Grand Total** | **90.40%** | **9.60%** | **100.00%** |

9. Analyse the Pivot tables created in the previous question and state any anomaly that you observe. Which categorical variables appear most important for your further study if you want to analyse which customers are most likely to take personal loans and why? **(5 marks)**
   A small percentage of the customers belonging to the Undergraduate class tend to avail personal loans. So this is one segment the bank can choose to overlook. The bank can send more Personal Loans offers to the other education classes.

   Online Banking does not seem to have any significant impact on the percentage of Personal Loans availed.

   TD Account and Income_Category appear to be critical in our analysis if we want to analyse which customers who are most likely to take personal loans.
   If we look at customers with TD accounts, nearly 46.36% of customers had accepted the personal loan offer in the last campaign. Similarly when we look at the Income category of customers, those customers with income of greater than 100k/year were more likely to accept personal loan offers. Nearly 36.14% customers accepted the offer In the last campaign. These figures are significantly higher than aggregate figure of 9.6% customers who accepted our offer last time. We should deep dive into these 2 variables when deciding our target audience for next campaign as the likelihood of acceptance increases multiple folds.

10. In the last campaign, bank reached out to 5000 customers out of which 480 customers accepted the personal loan offer. The bank incurred a huge cost in running a marketing campaign to reach out to so many customers. This is where you as a strategic business consultant step in. You are tasked to optimise the cost of this campaign by identifying the correct target base (without significant reduction in number of acceptance of offers). The bank can then send Personal Loan offers to these target customers who have a higher chance of accepting the offer. Based on your analysis, what strategy would you suggest to the management of HBFC bank? **(5 marks)**

| TD Account vs Personal Loans | | Row Labels | No | Yes | Grand Total |
|---|---|---|---|---|---|
| | | **No** | **4358** | **340** | **4698** |
| | **Income categories** | 0-50 | 1843 | | 1843 |
| | | 100+ | 742 | 311 | 1053 |

| | | | | | |
|---|---|---|---|---|---|
| | | 51-100 | 1773 | 29 | 1802 |
| | | **Yes** | **162** | **140** | **302** |
| | **Income categories** | 0-50 | 71 | | 71 |
| | | 100+ | 32 | 127 | 159 |
| | | 51-100 | 59 | 13 | 72 |
| | | **Grand Total** | **4520** | **480** | **5000** |

| | | **Row Labels** | **No** | **Yes** | **Grand Total** |
|---|---|---|---|---|---|
| | | **No** | **92.76%** | **7.24%** | **100.00%** |
| **TD Account vs Personal Loans** | **Income categories** | 0-50 | 100.00% | 0.00% | 100.00% |
| | | 100+ | 70.47% | 29.53% | 100.00% |
| | | 51-100 | 98.39% | 1.61% | 100.00% |
| | | **Yes** | **53.64%** | **46.36%** | **100.00%** |
| | **Income categories** | 0-50 | 100.00% | 0.00% | 100.00% |
| | | 100+ | 20.13% | 79.87% | 100.00% |
| | | 51-100 | 81.94% | 18.06% | 100.00% |
| | | **Grand Total** | **90.40%** | **9.60%** | **100.00%** |

Based on the above tables, it is clear that we can significantly increase our chances of acceptance if we narrow down our focus on just 3 categories namely:

1) Customers who do not have Term deposit and have an income greater than 100k /year.
2) Customers who have a term deposit and have an income greater than 100k/year.
3) Customers who have a term deposit and have an income between 51k to 100k/year.

If we target just these 3 categories, we would capture nearly 94% of total customers (451 out of 480) opting for personal loans. Our campaigning costs would also be significantly be reduced as now we would have to run a campaign for just 1284 customers rather than 5000 customers.

11. **Reflect on what you learnt while working on this project and fill the Reflection Report (Non Graded)**

Additional Information for students to make sense of what they did and why they did. Please offer your views.

**Significance of Exploratory Data Analysis in real world Data Analysis Projects**

- A deep dive into the data helps you in report generation / data story telling as you are able to uncover patterns which the decision makers might not be aware of looking at the data as is.
- The exploratory data analysis also helps you to identify the key variables that you can build models on (in later stages of your project) so as to optimise the results.
- Many times, there are costs involved in terms of money and time for data analysis projects, and if you have done your data analysis well, it ends up saving significant monetary costs and time as you have good chances of identifying key variables for your analysis.
- In the real world, it is a good practice to analyse all the variables by various plots, cross-tabulation (pivot tables), through generating descriptive statistics etc. This is because when we analyse the variables one by one, we can identify a direction in which we are most likely to find the solution(s) to the business problem. For example, here you had to analyse all

variables through Pivot Table, graphs, basic statistics to understand that which variables influence Personal Loans and which variables do not have a significant impact so the bank can choose to avoid them for the analysis.