Python Final Project Report

–

Personal Loan Marketing Campaign Analysis

Julie Decraemer and Yufeng Jiang

November 2025

Abstract

The retail marketing department of a bank ran a campaign in which it offered personal loans to its customers. The goal of our project is to analyze the data of the latest in order to discover insights that might help them with tailoring better-targeted campaigns that can lead to better conversion rates in the future. For this, we used the tools we learn in our Python class. This report contains a complete description of the results we obtained in our analysis. Note that we used publicly available information about banking and personal loans found on the internet. The corresponding sources are listed at the end of this report, and were used to enhance the quality and interpretation of our analysis.

Table of Contents

1 Discussing the Conversion Rate

After having imported our dataset and verified that it did not contain any missing values (as they could distort or sometimes invalidate our analysis. e.g. half of the customer don't have their income reported), we checked the data types. All of our variables are stored as integers (*int64*) except for the one reporting the average monthly credit card spending per month (in thousands of dollars), which is stored as decimal values (*float64*). Therefore, our dataset is clean and ready to be used for analysis.

In banking, the **conversion rate** corresponds to the proportion of clients that contracted a personal loan among all those targeted by the campaign :

$$\text{Conversion Rate} = \frac{\text{Number of clients who accepted the loan}}{\text{Total number of clients targeted by the campaign}}$$

When calculating this rate for our pool of customers, we get that **9.6%** of them contracted a loan following the campaign. It corresponds to 480 individuals, out of the 5000 included in our data set.

In the banking industry, the range of median conversion rates are between 2% and 5%, but this can vary based on the complexity of services, such as account openings, loan applications, or credit card sign-ups. If we compare our conversion rate to this benchmark, it is more than 4 percentage points higher, indicating that the offer is already pretty well-tailored to the target audience demographics (age, income, etc.). However, a deep analysis of the data will allow for further understanding of our target.

2 Distribution of Variables in the Data-set

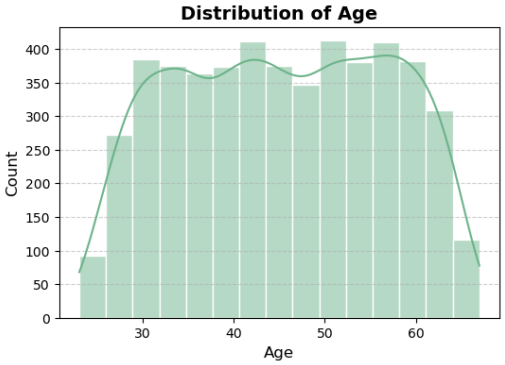

Figure 1: Distribution of Age

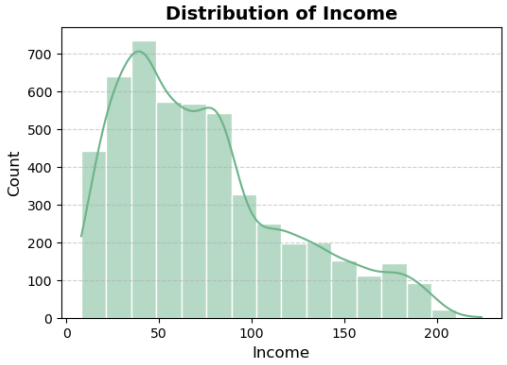

Figure 2: Distribution of Income

The distribution of age appears quite uniform between 30 and 60, with a slightly higher concentration around middle age. This suggests that the campaign targeted a broad range of adult customers, without focusing on a specific age group.

Contrary to the distribution of age and as expected, the income distribution in the data-set is not uniformly distributed but rather right-skewed. The majority of clients earn a low to moderate income, situated between 20 and 80 thousands of dollars. However, there is still a large part of the bank's customers showing very high income (beyond 100-200 thousands of dollars), which is why it creates this long right tail. These high-income customers represent almost one-quarter of the whole pool of customers, which is very close to the proportion of individuals earnings between 30 and 60 thousands of dollars a year (around 30%). This distribution is not exactly representative of the overall population income distribution. Indeed, in the US, in 2025, about 18% of American

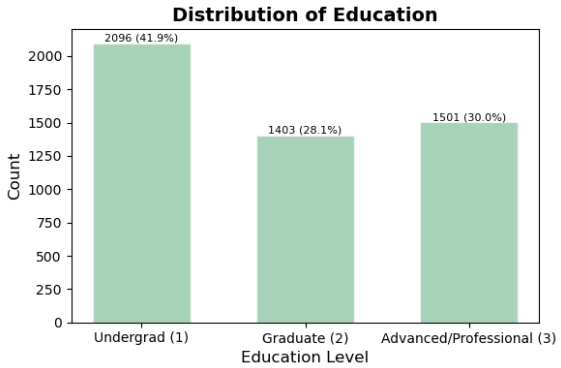

Figure 3: Distribution of Education

individuals make more than $100,000 annually, but we are pretty close to it.

The analysis of the education distribution shows an overall educated pool of customers. Almost half of the individuals (41.9%) hold an undergraduate diploma while around one third (30%) have a professional degree and 28.1% are graduated. This repartition is homogeneous between the three education level, indicating that a majority of the bank's clients have post-secondary education. These kind of profiles are likely to be associated with better understanding of financial products and a higher average income, which could influence their propensity to consider personal loan offers.

3 Interactions between Variables

3.1 Distribution Comparison by Securities Account Status

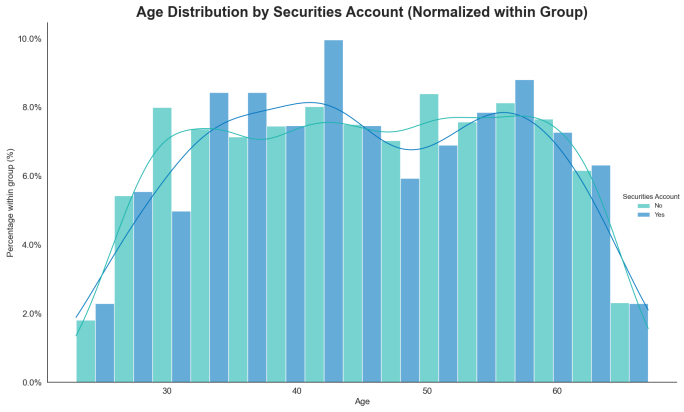

Figure 4: Age Distribution by Securities Account (Normalized Within Group)

Customers who have a Securities Account and those who don't show similar age distributions. The two curves showing the distributions of the two groups have approximately the same shape, although we do observe a slight over-representation of people having a Securities Account around 40-45 years old and an over-representation of individuals not having a Securities Account around 30-35 years old. **Overall, age does not seem to be a determinant factor in the holding of a Securities Account**.

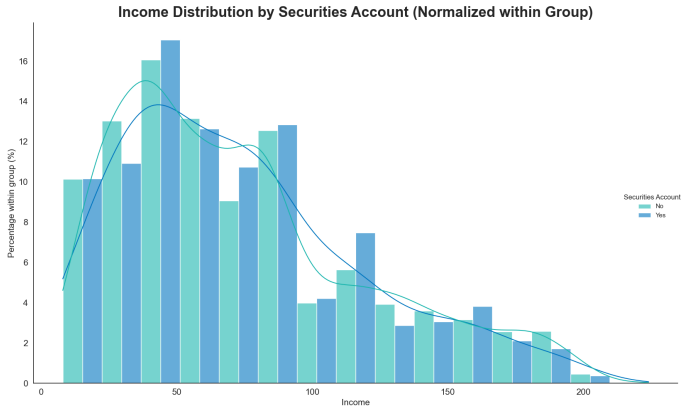

Figure 5: Income Distribution by Securities Account (Normalized Within Group)

Although the distribution of income of people having a Securities Account and of people not having a Securities Account look similar, there is still a subtle pattern : between $70,000 and $120,000 of yearly income, the proportion of "Yes" is often higher than the "No". In the same way, under $50,000, individuals who don't have a Securities Account dominate slightly the others. This result suggests that **customers with a Securities Account are more likely to have high income**.

We don't find significant differences when looking at the **Education Distribution by Securities Account** : they are extremely close.

3.2 Distribution Comparison by CD Account Status

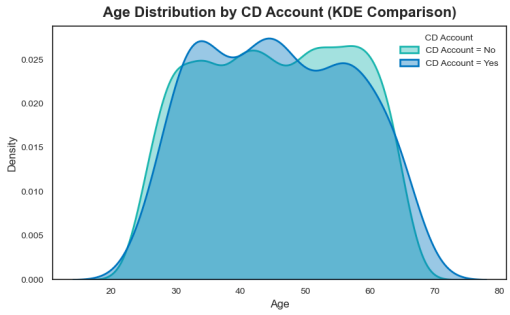

Figure 6: Age Distribution by CD Account (KDE comparison)

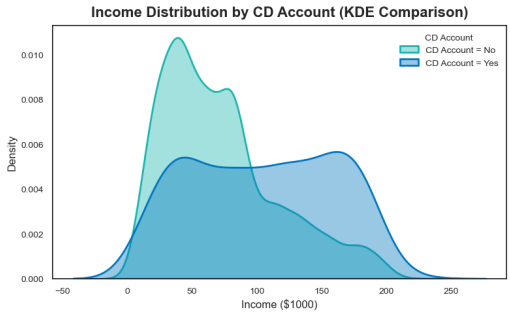

Figure 7: Income Distribution by CD Account (KDE comparison)

The age distributions of the two groups are almost perfectly overlapping, showing that age does not influence the holding of a CD Account. However, we acknowledge that there is a slightly larger proportion of people having a CD Account between 30 and 50 years old (which is also where most of the total customer population is concentrated) and after that age, the pattern is reversed - more people do not have a CD Account.

Unlike age, income shows a more pronounced difference between groups. Customers with a CD Account show a **right-shifted distribution**, indicating higher income levels. Lower-income individuals (below $40,000 yearly income) are more represented in the 'No' group, while medium to high incomes are relatively more common among CD Account holders. **Income therefore appears to be associated with CD Account ownership**.

Finally, the results related to education lead to the same conclusion as for the Securities Account:

the proportions of customers by level of education are similar between CD account holders and non-holders.

3.3 Account holding structure

Table 1: Summary of Account Holdings

ID	Account Type	Count	Percentage
0	No Accounts	4,323	86.5%
1	Only Securities Account	375	7.5%
2	Only CD Account	155	3.1%
3	Both Accounts	147	2.9%

Our analysis of the number of accounts held by the customers reveal that 86.5% of them have no account, **indicating very low overall adoption of the products offered by the bank**. Among customers who do hold an account, the Securities Account (7.5%) is significantly more common than the CD Account (3.1%). Finally, only 2.9% of customers hold both accounts simultaneously, suggesting that these products are rarely used together.

4 Factors Impacting Customer's favorable Decision

4.1 Descriptive Statistics

First of all, to get some insights about the variables mostly affecting the decision of the customer to accept the personal loan following the campaign, we performed some descriptive statistics. Namely, we graphically analyzed the distribution of the average acceptance rate across our customers for several relevant variables.

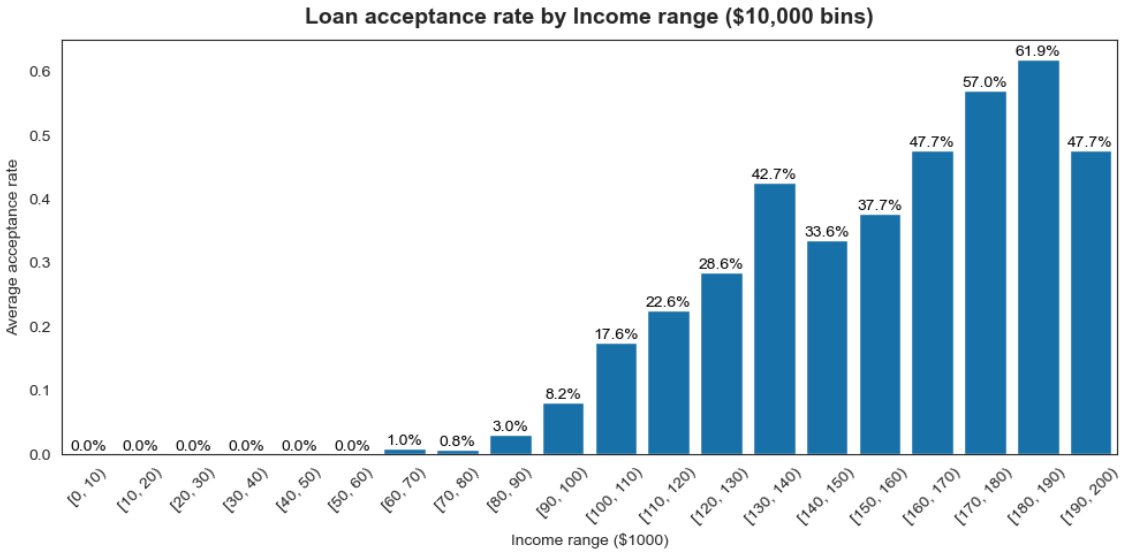

Figure 8: Loan acceptance rate by Income range

As shown by Figure 4 above, we divided the population into bins of $10,000 of annual income and looked at the level of loan acceptance within those 10 bins. The pattern is clear : **the higher the income range, the higher the average acceptance rate**. To be precise, we have zero-rate-of-acceptance for all customers with income below $60,000, which is already a relatively high income, given the income distribution we observed previously in the population of customers. The average acceptance rate increases gradually from, on average, more than 1 individual out of 4 accepting

the personal loan in the $120,000-$130,000 income bracket, until more than 6 customers out of 10 accepting the loan in the $180,000-$190,000 income bracket.

Moreover, **the acceptance rate also highly differs depending on the education level of the customer**, as expected in our introduction : they are likely associated with a better understanding of financial products and a higher average income, which influence their propensity to consider personal loan offers. Only 4.4% of undergraduates accepted the loan, against 13% of graduates and 13.7% of advanced or professional degree holders. These figures are low but coherent given the overall conversion rate of 9.6%.

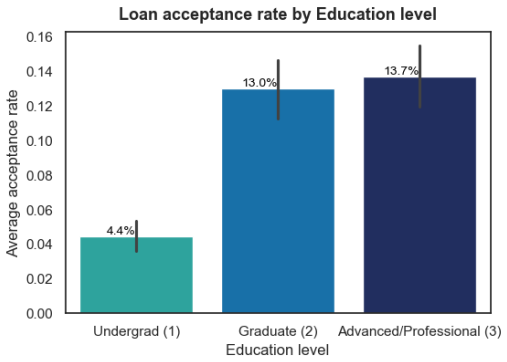

Figure 9: Loan acceptance rate by Education Level

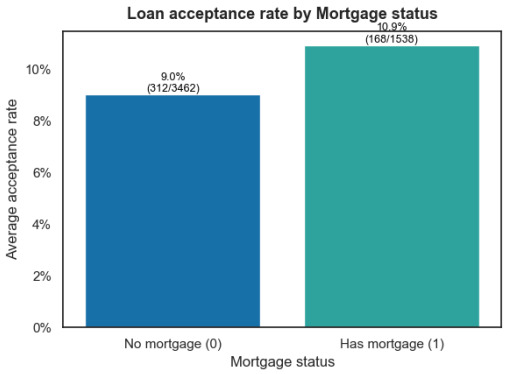

Figure 10: Loan acceptance rate by Mortgage status

Customers with a mortgage show a significantly higher loan acceptance rate (10.9%) compared to those without one (9%). A two-proportion z-test confirms that this difference is statistically significant ($z = -2.12$, $p = 0.034$) at the 5% significance level, suggesting that **having an existing mortgage is associated with a higher likelihood of accepting a personal loan offer**. This could be explained by the fact that they have already established a relationship of trust with the bank (they have been approved for a large loan) and they have a better understanding of how loans work. To obtain a mortgage, you need a stable income and a good credit history, these customers are therefore often more creditworthy, less risky, and more targeted by banking offers. Finally, a mortgage often involves related expenses: renovation, furnishing, building work, additional costs. Indeed, these customers may need an additional personal loan to cover these costs.

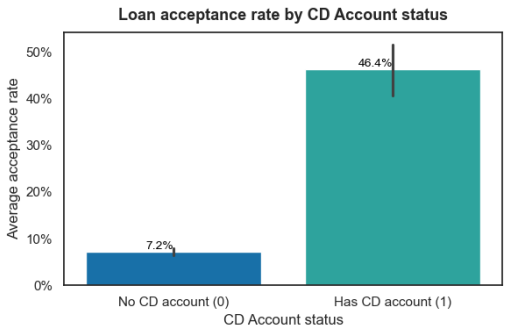

Figure 11: Loan acceptance rate by CD account

When they have a CD Account, customers deposit a large sum of money for a given period of time in order to earn interest on a saving account. This means that they have sufficient liquidity and are financially stable to live without this money for an amount of time. They are therefore more likely to have high income. As the money is blocked for a period of time, customers must also

trust the bank. This is what we observe in the data - **customers with a CD account show an average acceptance rate significantly higher than other customers : 46.4% against 7.2%**.

We performed the same analysis but looking at the variables Age and Monthly Average Spending (CCAvg) and we don't find any linear or noticeable pattern except the following : **100% of the customers spending between $9,000 and $10,000 (monthly) accepted the personal loan**. As those individuals are more likely to have higher income, this result only strongly suggests, again, that the higher the yearly income, the more likely the customer is to accept the personal loan. The second average spending bracket with the highest average acceptance rate concerns individuals spending between $5,000 and $6,000 a month (61.9%). For Age, we noticed that no customer below 25 years old accepted the personal loan, which is in line with the previous results. Indeed, younger customers are less likely to have high income. Finally, having a Securities Account does not significantly impact the decision to accept the personal loan or not.

4.2 Logistic Regression

Our descriptive analysis gave us a good idea of the main variables that are probably driving a customer's favorable decision, namely, income and education level, having a mortgage and a CD account are the ones for which we graphically observed differences among the population of 5000 customers.

As a next step, we would like to **quantify** the individual effect of each variable on the probability of accepting the personal loan (all other things being equal). To do so, we performed a **logistic regression**, as our target variable (Personal Loan) is binary.

Dep. Variable:	Personal Loan	**No. Observations:**	5000
Model:	Logit	**Df Residuals:**	4989
Method:	MLE	**Df Model:**	10
Date:	Thu, 13 Nov 2025	**Pseudo R-squ.:**	0.5935
Time:	18:46:12	**Log-Likelihood:**	-642.74
converged:	True	**LL-Null:**	-1581.0
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	-9.0130	0.409	-22.029	0.000	-9.815	-8.211
Age	0.1167	0.074	1.567	0.117	-0.029	0.263
Income	2.5183	0.120	20.909	0.000	2.282	2.754
Family	0.6953	0.074	9.354	0.000	0.550	0.841
Education	1.7154	0.113	15.147	0.000	1.493	1.937
Mortgage	0.0461	0.056	0.819	0.413	-0.064	0.156
Securities Account	-0.9304	0.285	-3.262	0.001	-1.489	-0.371
CD Account	3.8302	0.324	11.838	0.000	3.196	4.464
Online	-0.6738	0.157	-4.292	0.000	-0.981	-0.366
CreditCard	-1.1172	0.205	-5.449	0.000	-1.519	-0.715
CCAvg	0.2157	0.069	3.116	0.002	0.080	0.351

Table 2: Logit regression results

The results table from our logistic regression highlights more statistically significant relationships than our graphical analysis suggested. **Only two coefficients are statistically insignificant : Age and Mortgage (p-value greater than 0.05)**. In our descriptive analysis, we found that customers with a mortgage had a higher loan acceptance rate. However, thanks to the logistic regression, we are now able to say that, once controlling for income, education, and financial products (CD and securities accounts), the effect of having a mortgage is no longer statistically significant.

This suggests that the relationship observed earlier was actually driven by other correlated factors. All the other coefficients are statistically significant at the 5% level.

The **Pseudo R-squared of 0.59** suggests a very strong explanatory power for a logistic regression.

As expected, **income is the most powerful predictor of the likelihood of accepting a loan** (highest coefficient) : holding other variables constant, a one-standard-deviation increase in income multiplies the odds (probability that the event happens compared to not) of accepting the loan by approximately 12.4. As seen in the descriptive analysis, education has a strong impact on the probability to accept the personal loan. Surprisingly, larger family size slightly increases acceptance (maybe due to higher financial needs). Another observation is that having a Securities account is associated with a lower probability of accepting the personal loan. This may be because these customers already have money invested and can use these funds if needed, rather than taking a personal loan. Finally, as we noticed in the graphical analysis, customers with a Certificate of Deposit are far more likely to take a loan : all other things being equal, having a CD account increases the odds of accepting the loan by 46.

5 The Prediction Performance of Machine Learning Model

To address our primary goal of predicting whether a customer will respond favorably to the personal loan campaign, we built a simple machine learning **classification model**. Specifically, we implemented a decision tree classifier to handling this scenario.

We used the 80% of the original dataset as the training set, reserving the remaining 20% as the testing set. A stratified K-fold cross-validation with 10-folds was used to ensure robust evaluation during hyperparameter tuning. After performing grid search, we found the best-performing model was a model with `max_depth` of 6 and `min_samples_split` of 49. This optimal model achieved an overall accuracy of 98% on the test set and an RMSE of 0.122, which demonstrates a strong fit.

The detailed predictive performance of this model is summarized in the Figure 12 and its corresponding classification report.

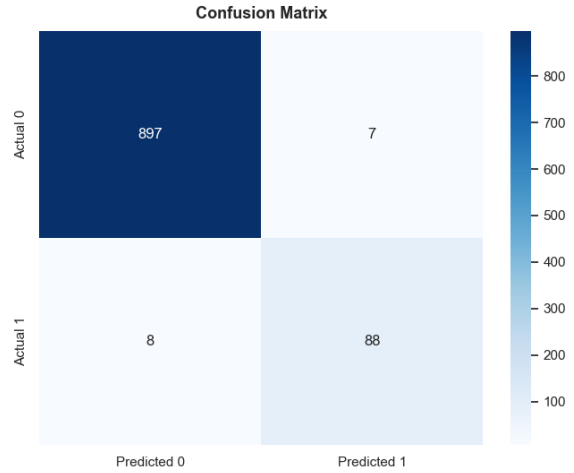

Figure 12: Confusion Matrix of Decision Tree Model

The model demonstrates high accuracy (98%) and strong performance across both classes, with particularly high precision and recall value for the class(0). For the class (1) - customers who accepted the loan, it sill maintained a high precision and recall level, at 93% and 92% respectively.

Table 3: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	904
1	0.93	0.92	0.92	96
Accuracy			0.98	1000
Macro Avg	0.96	0.95	0.96	1000
Weighted Avg	0.98	0.98	0.98	1000

Overall, our model has a reliable ability to identify potential loan acceptors. However, its true predictive performance should be validated on a new, unseen dataset to confirm its generalizability.

6 Conclusion

The goal of our analysis was to identify the factors influencing the probability that a client accept the personal loan following the marketing campaign conducted by the bank. Based on our results, the bank can refine its strategy by :

- **Prioritizing high income individuals** : lower risk, greater borrowing capacity, more projects requiring financing and more comprehensive banking relationship (CD accounts, investments), which strengthens trust and acceptance of offers.
- **Targeting high educated customers** : they have a better understanding of financial products and likely higher income, which likely leads to a favorable decision, as shown by the analysis.
- **Targeting clients that possess a CD Account** : strong indicator of a good and active relationship with the bank, and significantly higher likelihood of accepting the loan.
- **Not considering Mortgage as a relevant criterion** : this segment does not react significantly better than average once income, education and financial products are controlled for.
- **Prioritizing families instead of single individuals** : higher or more frequent financial needs.

Overall, we found that the decision to accept a loan is largely explained by structural financial factors, but also by the products already held (CD accounts). This implies that the bank has strong potential to optimize its campaign if it relies on **identified profiles rather than uniform targeting**.

7 Bibliography

1) Dowling, L. (2023, June 14). *The Financial Industry Guide to Conversion Rate Optimization.* Pathmonk. https://pathmonk.com/financial-industry-guide-conversion-rate-optimization/

2) Moore, T. (2025, July 18). *Is a $100,000 Salary Good?* SoFi. https://www.sofi.com/learn/content/is-100000-a-good-salary/

3) El Mahrsi, K. (2025 course). *Python for Data Science: A Crash Course.* Smelly Data Science. https://smellydatascience.com/teaching/python-for-data-science/

4) QuantEcon. *Lectures.* https://quantecon.org/lectures/

5) Scikit-learn Documentation. https://scikit-learn.org/stable/index.html