

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	1
2	Distribution of Variables in the Data-set	1
3	Interactions between Variables	2
3.1	Distribution Comparison by Securities Account Status	2
3.2	Distribution Comparison by CD Account Status	3
3.3	Account holding structure	4
4	Factors Impacting Customer's favorable Decision	4
4.1	Descriptive Statistics	4
4.2	Logistic Regression	6
5	The Prediction Performance of Machine Learning Model	7
6	Conclusion	8
7	Bibliography	8

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

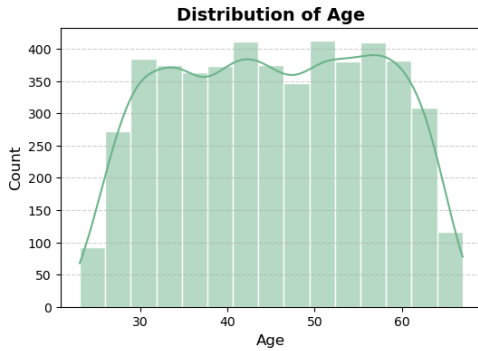


Figure 1: Distribution of Age

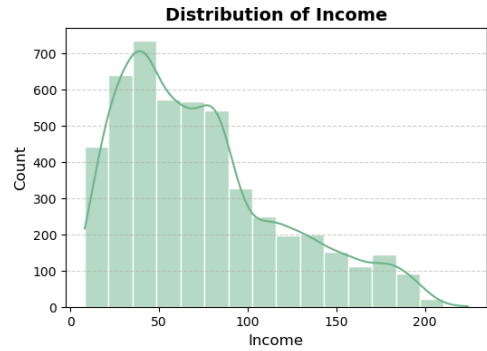


Figure 2: Distribution of Income

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 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 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. The distribution shows that undergraduates (41.9%) are the largest group, while graduate (28.1%) and professional degree holders (30%) make up slightly smaller but comparable shares. Overall, the customer base appears highly educated. Such a profile could potentially correlate with better financial literacy and income levels, which could influence their propensity to consider personal loan offers.

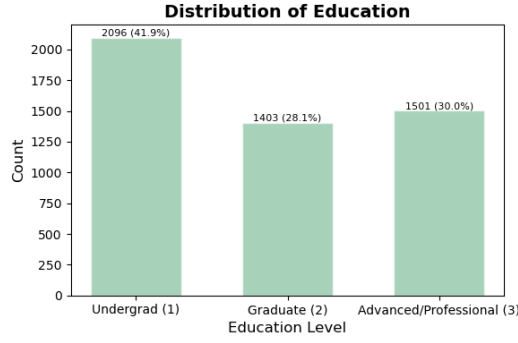


Figure 3: Distribution of Education

3 Interactions between Variables

3.1 Distribution Comparison by Securities Account Status

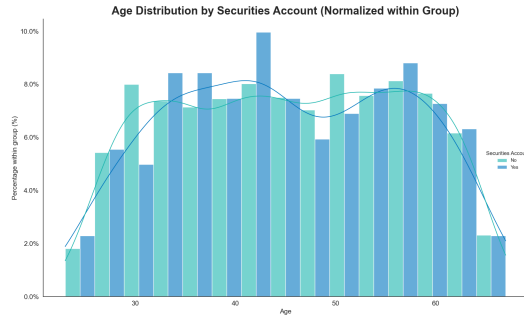


Figure 4: Age Distribution by Securities Account (Normalized Within Group)

From our closer examination of the dataset, we find that only 10.44% of the customers hold a Securities Account. Therefore, instead of relying on the raw count distribution of customers with and without a Securities Account, we present a normalized age distribution for both groups, which enables a more meaningful comparison.

As shown in Figure 4, the two groups exhibit very similar age distributions. The two curves largely share 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 non-holders around 30-35 years old. **Overall, age does not seem to be a determinant factor in the holding of a Securities Account.**

A similar approach is applied to the income distribution. To obtain more comparable insights, we normalize the income distribution within each group. Although the income distributions of customers with and without a Securities Account appear broadly 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 slightly dominate. This result suggests that **customers with a Securities Account are more likely to have higher income.**

We don't find significant differences when looking at the **Education Distribution by Securities Account** : the proportions across education levels are nearly identical for both groups.

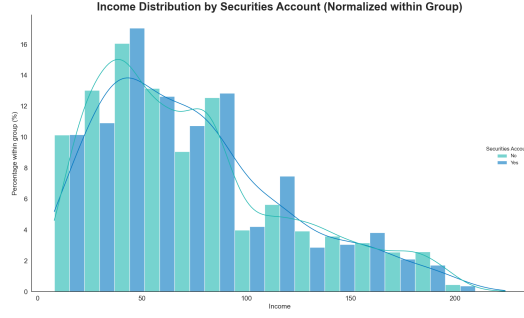


Figure 5: Income Distribution by Securities Account (Normalized Within Group)

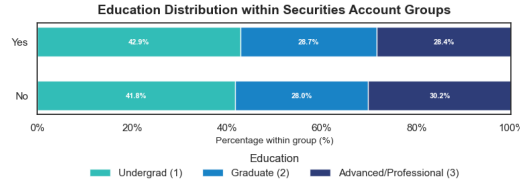


Figure 6: Education Distribution by Securities Account (Normalized Within Group)

3.2 Distribution Comparison by CD Account Status

Since our calculation shows that only 6.04% of customers hold a CD Account, we present normalized distributions of the relevant variables, following the same approach as in the previous section.

The age distributions of the two groups are almost perfectly overlapping, suggesting that age does not influence the holding of a CD Account. However, we note a slightly larger proportion of people with 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.

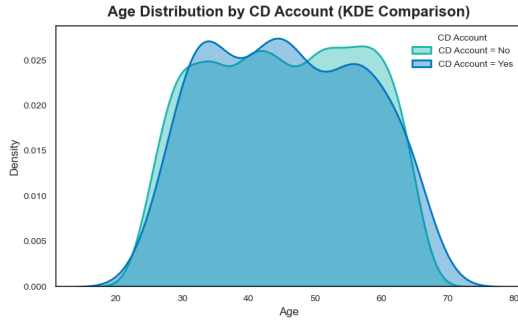


Figure 7: Age Distribution by CD Account (KDE comparison)

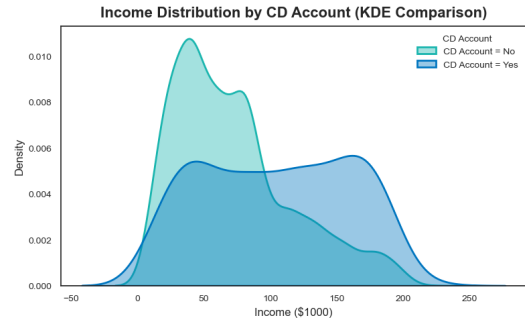


Figure 8: Income Distribution by CD Account (KDE comparison)

Unlike age, income shows a more pronounced difference between groups. Customers with a CD Account show a right-shifted distribution, indicating that a greater share of these customers falls into medium- and high-income brackets. In contrast, lower-income ranges account for a comparatively larger share of customers without a CD Account. **These normalized patterns suggest that higher-income individuals are relatively more likely to hold a CD Account.**

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.

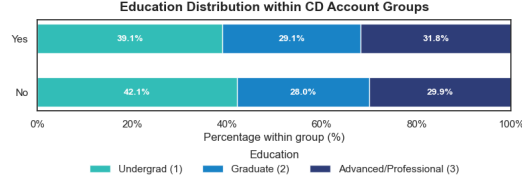


Figure 9: Education Distribution by CD Account (Normalized Within Group)

3.3 Account holding structure

Account ownership is low overall, with 86.5% of customers holding neither product. Among customers who do hold an account, Securities Accounts (7.5%) are more prevalent than CD Accounts (3.1%). Finally, only 2.9% of customers hold both accounts simultaneously, suggesting that these products are seldom used in combination and likely serve distinct financial needs.

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%

Table 1: Summary of Account Holdings

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.

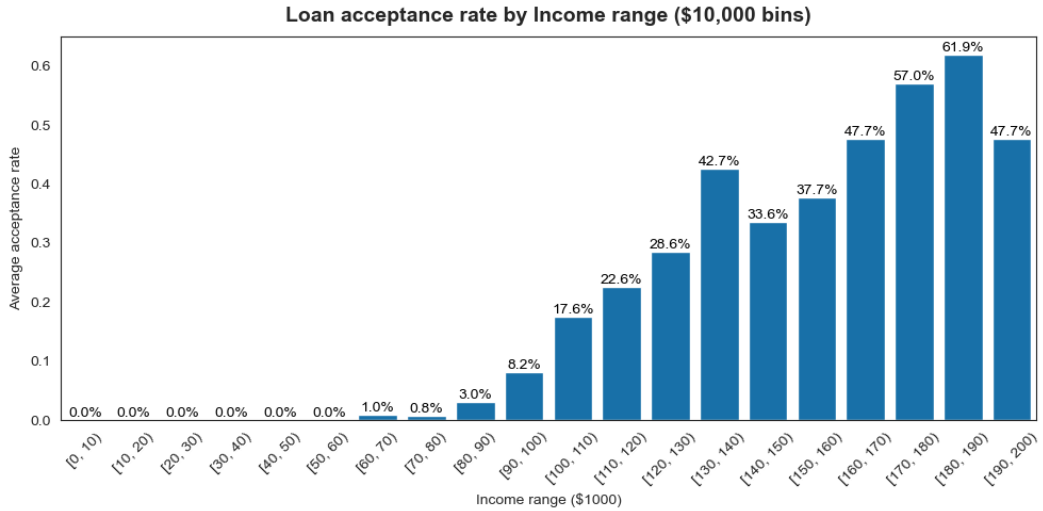


Figure 10: Loan acceptance rate by Income range

As shown by Figure 10 above, we divided the customers into \$10,000 income brackets and calculated the average loan acceptance rate within each group. The pattern is clear : **the higher the income range, the higher the average acceptance rate**. To be precise, we have zero acceptance rate for

all income levels 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 8.2% in the \$90,000–\$100,000 bracket to more than 25% in the \$120,000–\$130,000 bracket, and peaks at 61.9% for customers earning between \$180,000 and \$190,000.

The acceptance rate also highly differs among different education levels, as expected in our introduction : they are likely associated with higher financial literacy or higher income, which influence their propensity to accept personal loan offers. As illustrated in Figure 11, only 4.4% of undergraduates accepted the loan, against 13% of graduates and 13.7% of advanced or professional degree holders. These figures remain consistent with the overall conversion rate of 9.6%.

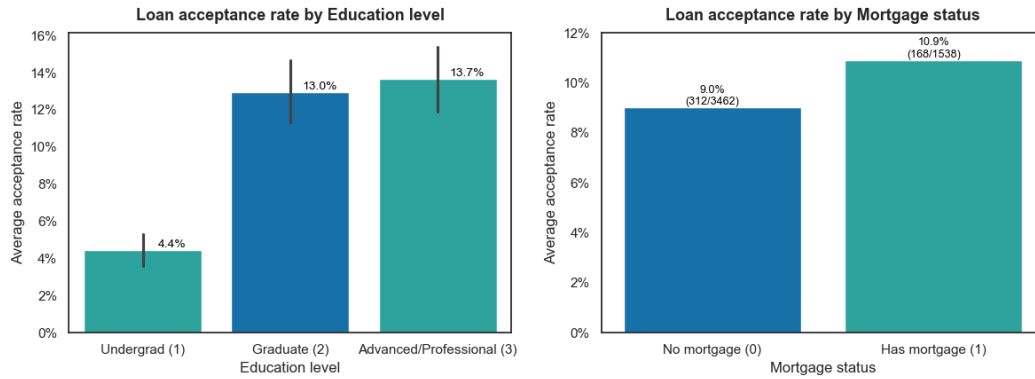


Figure 11: Loan Acceptance Patterns: Education vs Mortgage Status Analysis

Mortgage status is also associated with different acceptance rates. Customers with a mortgage exhibit 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.

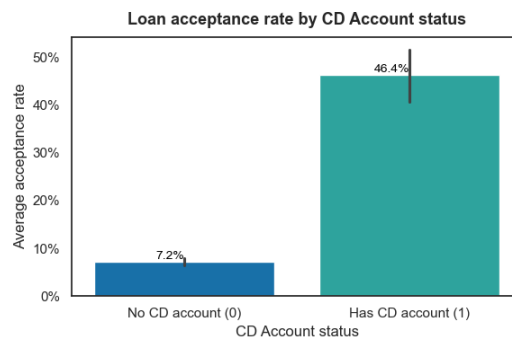


Figure 12: Loan acceptance rate by CD account

Maintaining a CD Account generally requires depositing a substantial sum for a fixed period, which implies sufficient liquidity and financial stability of customers. Those who are able to lock in such amounts may therefore belong disproportionately to higher-income or financially secure segments—groups that, as shown earlier, also display higher loan acceptance rates. 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 extend the analysis by examining Age and Monthly Average Spending (CCAvg), but 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-highest acceptance rate is observed among customers spending \$5,000–\$6,000 per 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 tend to have lower income levels. Finally, having a Securities Account does not significantly impact the loan acceptance decision.

4.2 Logistic Regression

Our descriptive analysis gave us an initial understanding of the variables that are probably driving a customer's favorable decision, namely, income, education level, having a mortgage, and the possession of a CD account. These variables are the ones for which we observed visible differences across the population of 5,000 customers.

As a next step, we would like to **quantify** the individual effect of each variable on the probability of accepting the personal loan (holding all else constant). To do so, we estimated 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 initially suggested by the descriptive analysis. **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 variables (CD and securities accounts), the effect of having a mortgage is no longer statistically significant. This implies that the earlier observed relationship was driven by correlated factors rather than by mortgage status itself. All the other coefficients are statistically significant at the 5% level.

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

As expected, **income is the strongest predictor of the likelihood of accepting a loan** (highest coefficient) : holding all 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 reflect the fact that such customers already possess invested funds and therefore rely less on borrowing. Finally, as we noticed in the graphical analysis, customers with a Certificate of Deposit are far more likely to take a loan : all else equal, customers with a CD Account have odds of accepting the loan that are approximately 46 times higher than those without one.

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 13 with its corresponding classification report in Table 3.

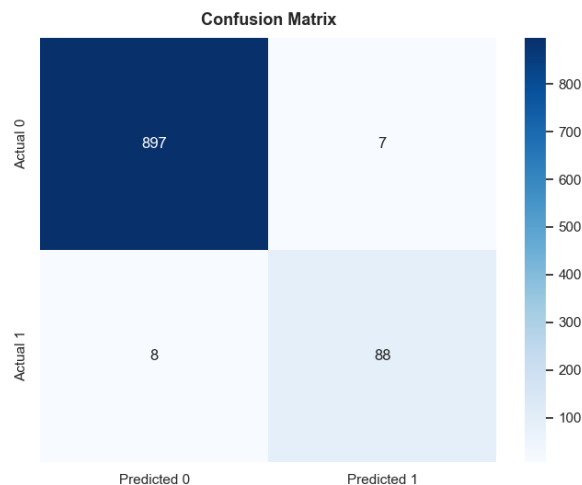


Figure 13: Confusion Matrix of Decision Tree Model

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

Table 3: Classification Report

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 still maintained a high precision and recall level, at 93% and 92% respectively.

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>