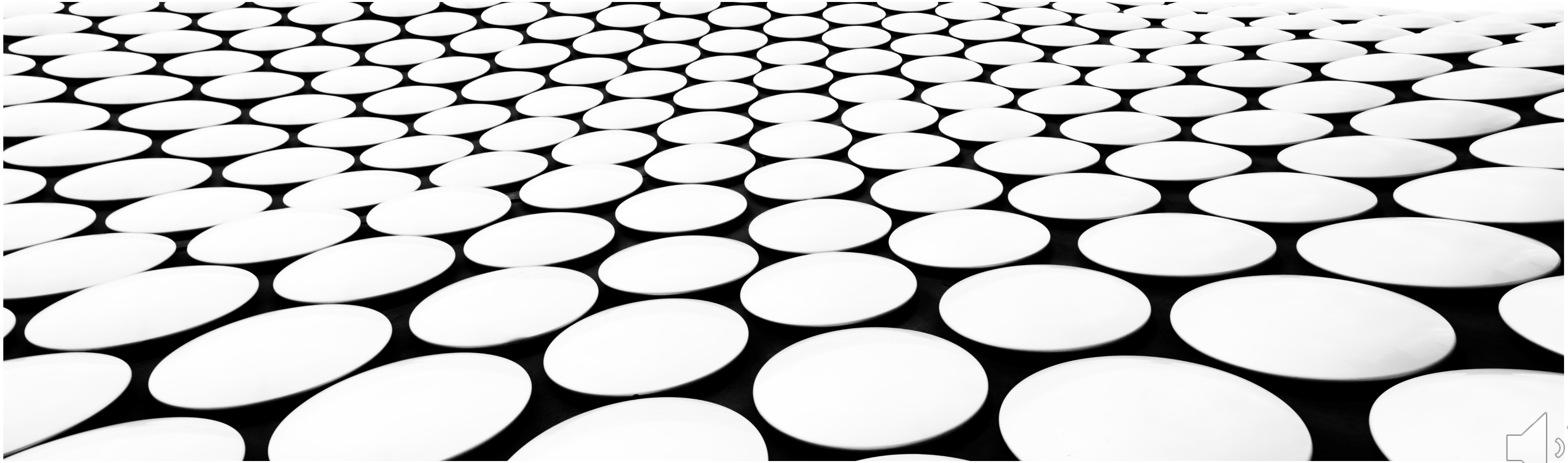
GITHUB REPOSITORY:
https://github.com/nicovy/bank-personal-Loan

BANK PERSONAL LOAN

NICOLAS VEAS



CONTEXT

- Predicting if a customer will buy a personal loan or not has high importance to the banks, especially when it comes to marketing campaigns. If we know whether a customer will buy a loan or not, we can make better target marketing campaigns to increase the success ratio.

- In this project, I'll use a classification model to predict customers have a higher probability of purchasing a loan.

- The data used is from kaggle from the following link:

https://www.kaggle.com/datasets/mahnazarjmand/bank-personal-loan?resource=download

OBJECTIVE

- Build a model that will help identify potential customers with a higher probability of purchasing a loan.
- Predict whether a customer will buy a personal loan or not.

DATA INFORMATION

Contains the following Variables

- ID: Customer ID
- Age: Customer’s age in completed year
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: the Family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- Mortgage: Value of house mortgage if any. (in thousand dollars)
- Personal Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities Account: Does the customer have securities account with the bank?
- CD Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Do customers use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)?

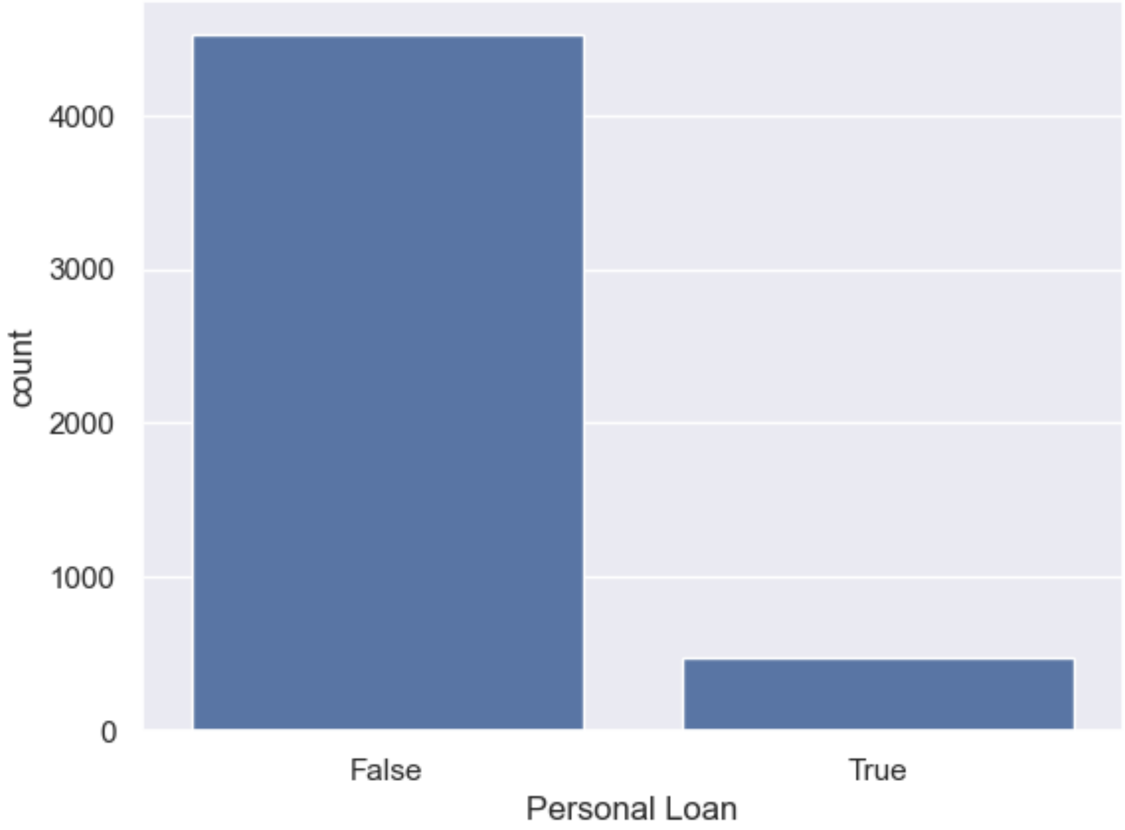
5000 Observations
14 Variables
5 Numerical Variables
8 Categorical Variables
No missing values
No duplicated Values

EXPLORATORY DATA ANALYSIS

- Data from 245 cities and 39 counties, Most of the customers are from Los Angeles County, around 21%
- All data is from California
- Customers age between 23 and 67 years old
- Most customers don't have a Mortgage
- Most customers are single
- Over 2000 customers are Undergraduate
- Only 480 customers out of 5000 had Personal Loan (around 10%)
- Most customers have online usage
- 4478 customers don't have a Securities account, ~10% have it
- 4478 customers don't have CD accounts, ~10% have it
- 3530 customers don't have Credit Cards in another bank, ~30% have it

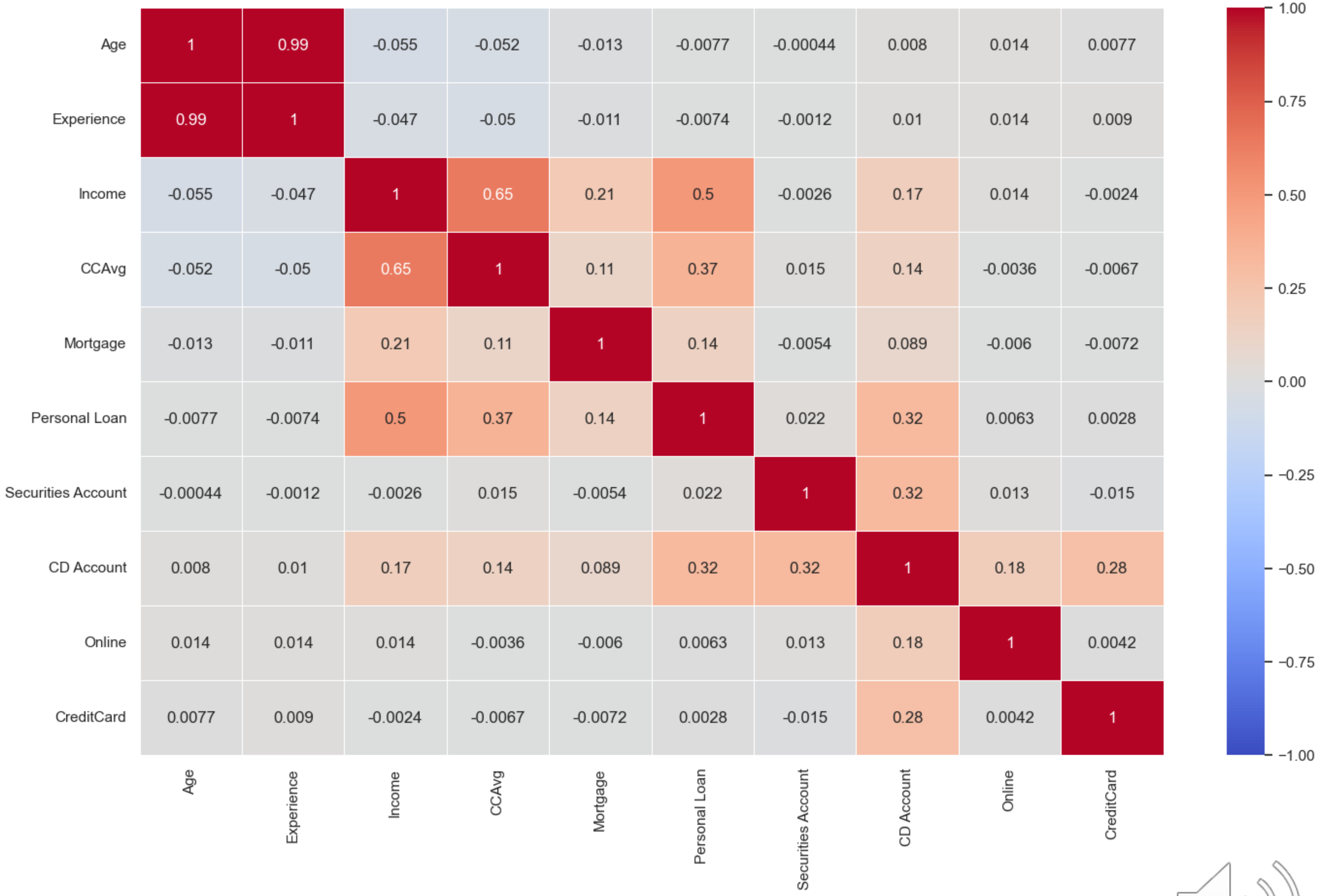
	Age	Experience	Income	Family	CCAvg	Education	Mortgage	Personal_Loan	Securities_Account	CD_Account	Online	CreditCard	city	county	state
0	25	1	49000	4	1600.0	Undergrad	0	False	True	False	False	False	Pasadena	Los Angeles County	CA
1	45	19	34000	3	1500.0	Undergrad	0	False	True	False	False	False	Los Angeles	Los Angeles County	CA
2	39	15	11000	1	1000.0	Undergrad	0	False	False	False	False	False	Berkeley	Alameda County	CA
3	35	9	100000	1	2700.0	Graduate	0	False	False	False	False	False	San Francisco	San Francisco County	CA
4	35	8	45000	4	1000.0	Graduate	0	False	False	False	False	True	Northridge	Los Angeles County	CA
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4995	29	3	40000	1	1900.0	Advanced/Professional	0	False	False	False	True	False	Irvine	Orange County	CA
4996	30	4	15000	4	400.0	Undergrad	85000	False	False	False	True	False	La Jolla	San Diego County	CA
4997	63	39	24000	2	300.0	Advanced/Professional	0	False	False	False	False	False	Ojai	Ventura County	CA
4998	65	40	49000	3	500.0	Graduate	0	False	False	False	True	False	Los Angeles	Los Angeles County	CA
4999	28	4	83000	3	800.0	Undergrad	0	False	False	False	True	True	Irvine	Orange County	CA

EXPLORATORY DATA ANALYSIS
TARGET VARIABLE VARIABLE - PERSONAL_LOAN



* 90.4% of customers didn't accept personal loan on the previous campaign
* 9.6% accepted personal loan on the previous campaign

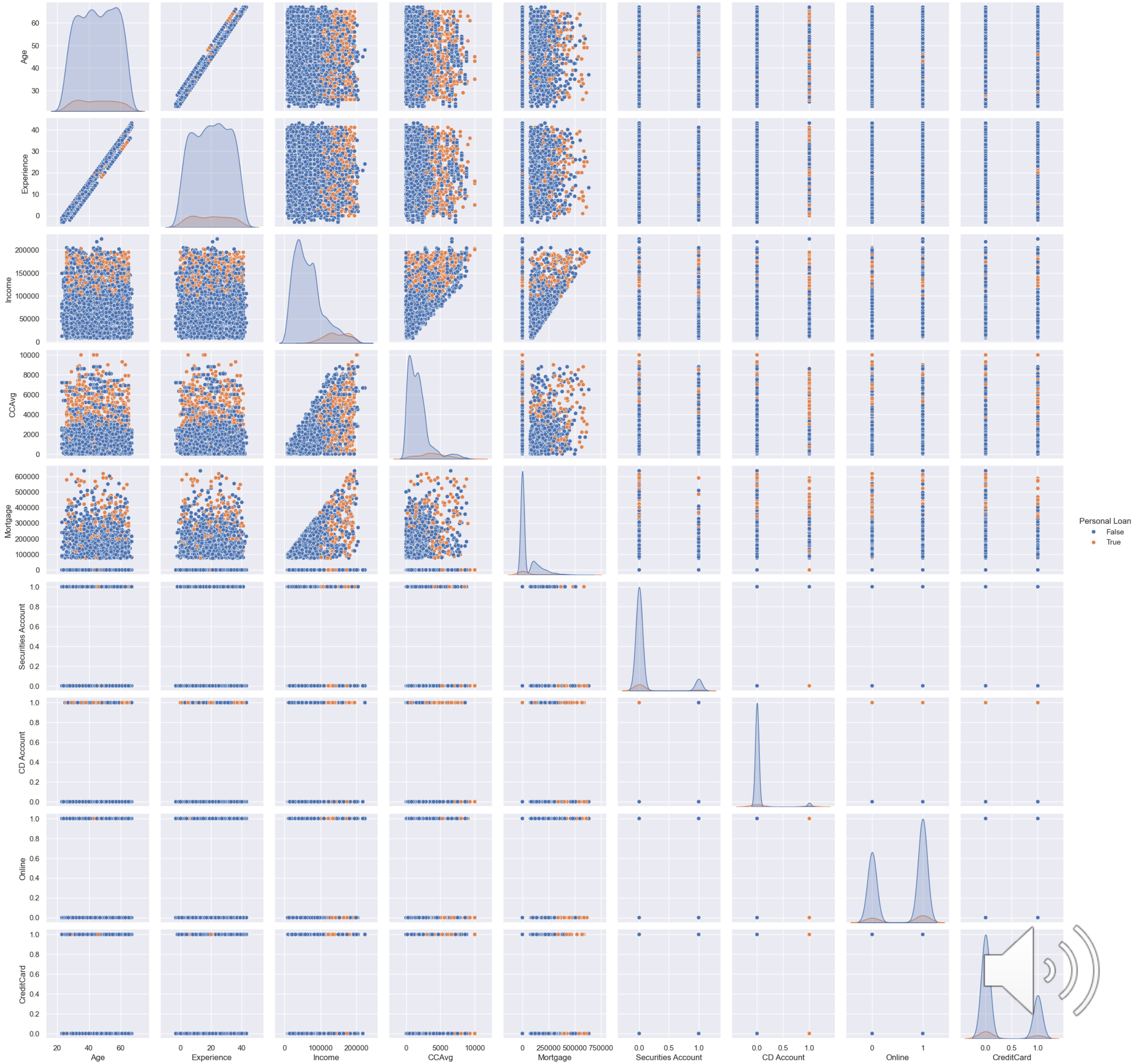
EXPLORATORY DATA ANALYSIS
CORRELATION HEATMAP



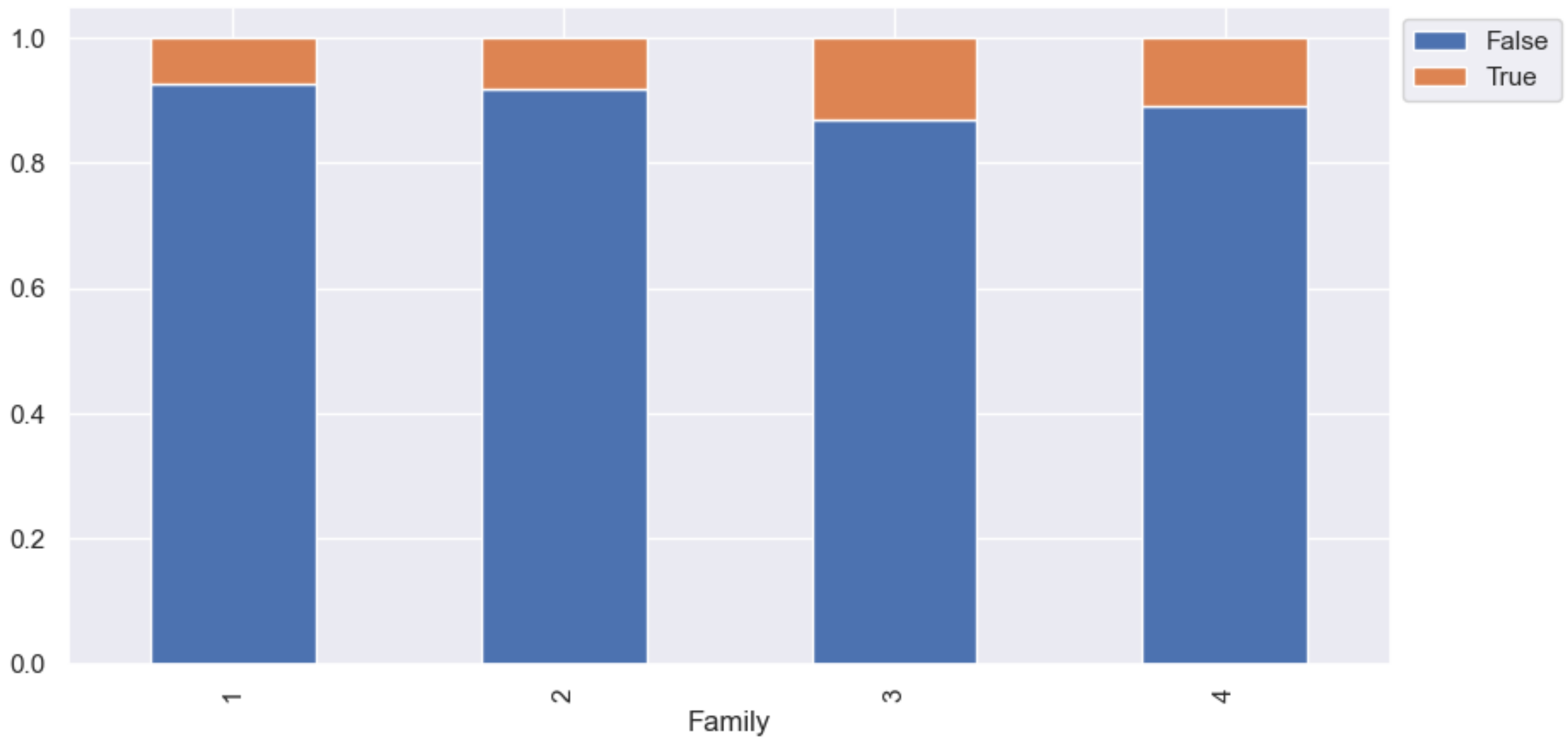
- Personal loan is most correlated with Income
- Age and experience are highly correlated (.99), as we can expect.

EXPLORATORY DATA ANALYSIS SCATTER PLOT

- We can see the linear correlation between Experience and Age
- People with high incomes get more loans than the rest.
- People with high credit card usage also trends to gem more loans.



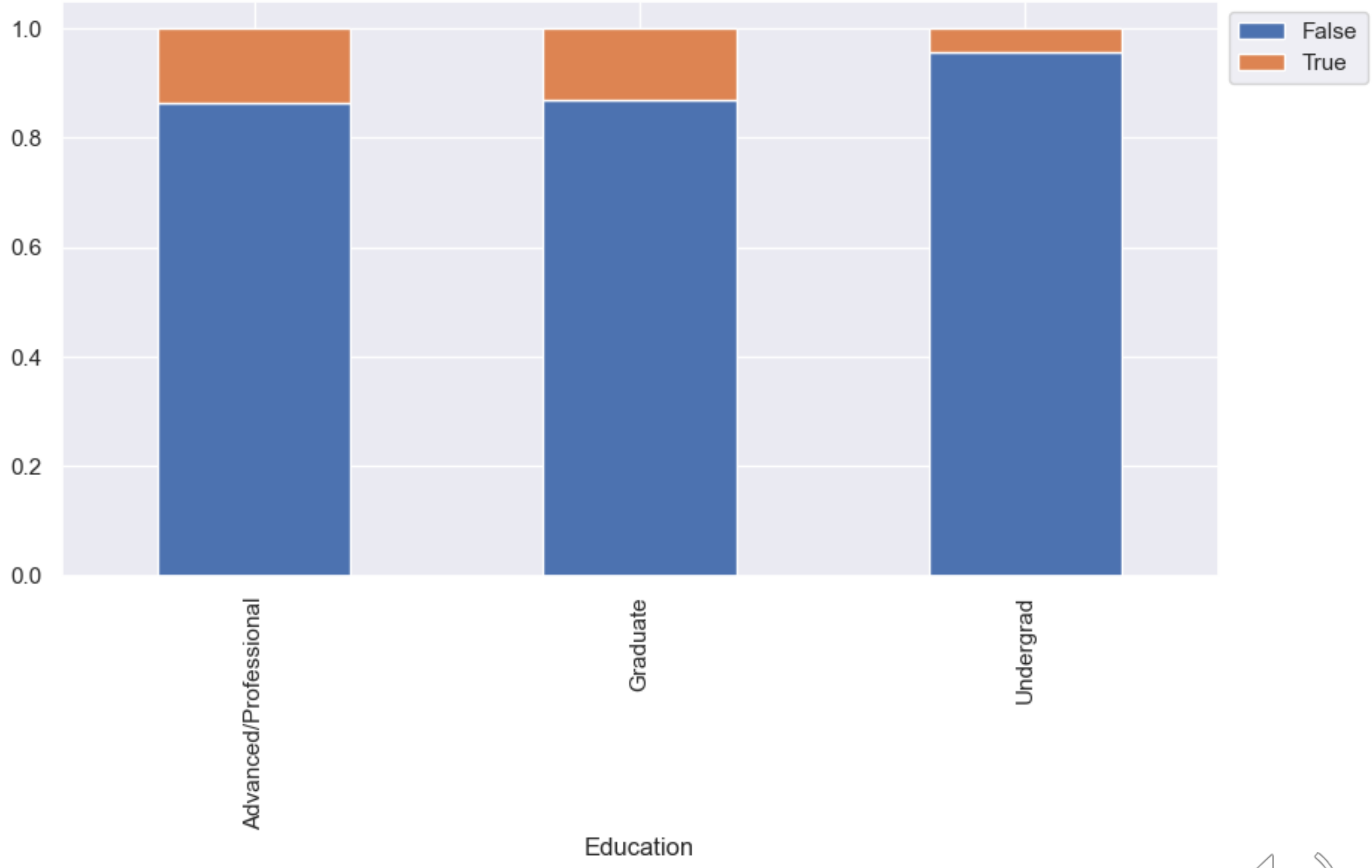
EXPLORATORY DATA ANALYSIS
CATEGORICALS



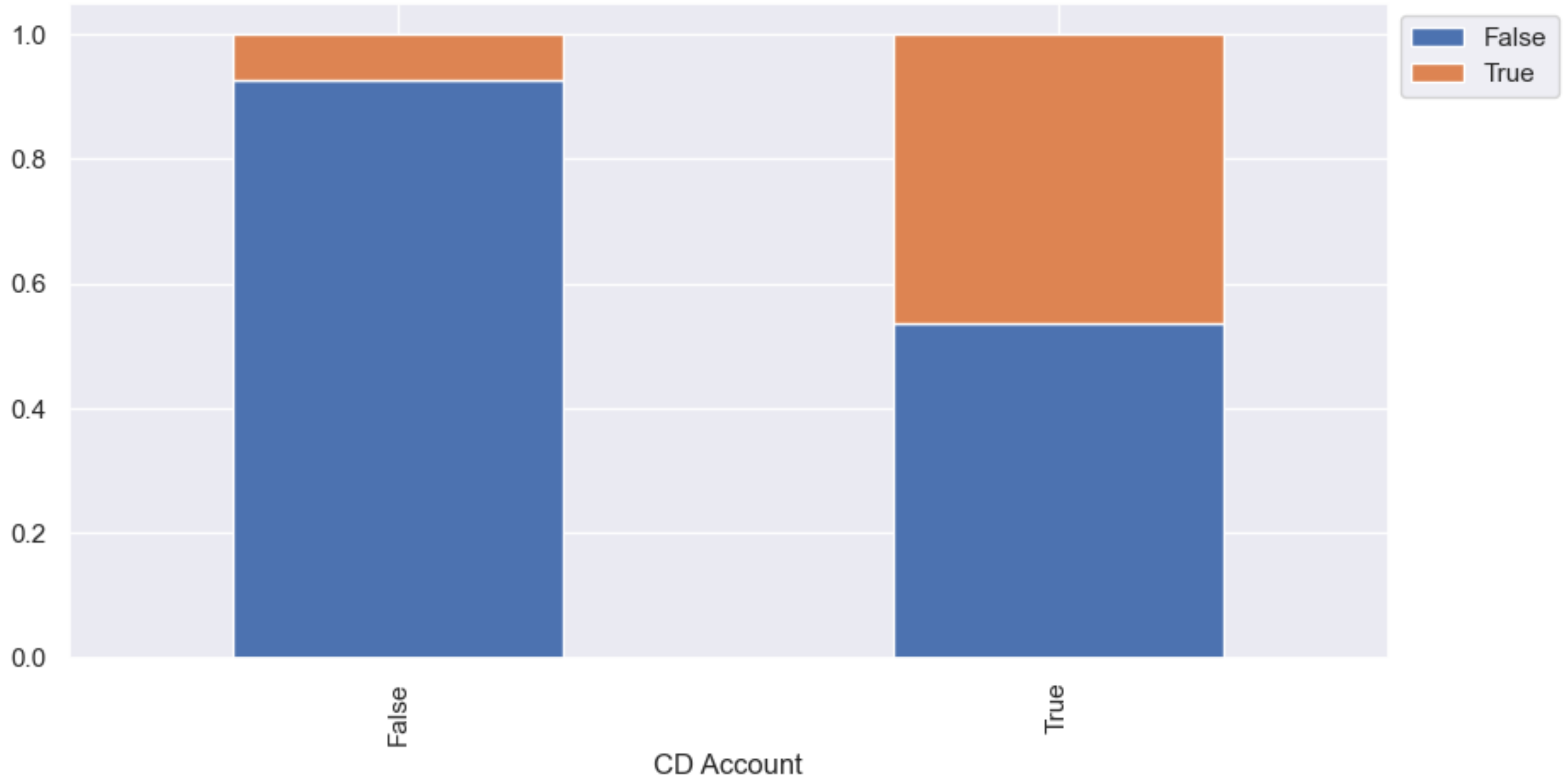
- Personal loans are given more to families of 3 and 4 children.

EXPLORATORY DATA ANALYSIS CATEGORICALS

- Personal loans were given more to customer graduated and andvanced/professional



EXPLORATORY DATA ANALYSIS CATEGORICALS



- Around 45% of the people with CD Account got Personal Loan

MODEL BUILDING
DATA PREPARATION

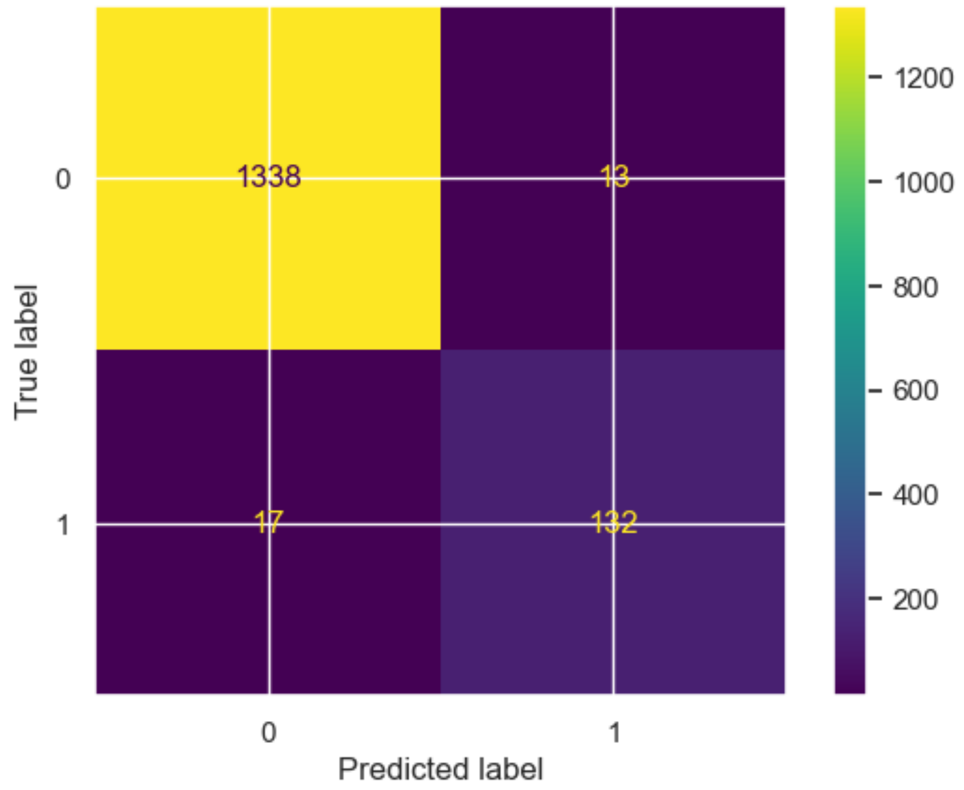
We want to optimize **recall,** as we like to get as less false positives as possible. the marketing campaings should aim to the most of potential loan buyers as possible.

- Zipcode and ID Removed
- Personal loan is our Target variable
- Data Divided into 70% for training and 30% for testing
- Converted categorical into dummy variables for modeling purposes

	Age	Experience	Income	CCAvg	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard	Family_2	Family_3	Family_4	Education_Graduate	Education_Undergrad
0	25	1	49000	1600.0	0.0	False	True	False	False	False	0	0	1	0	1
1	45	19	34000	1500.0	0.0	False	True	False	False	False	0	1	0	0	1
2	39	15	11000	1000.0	0.0	False	False	False	False	False	0	0	0	0	1
3	35	9	100000	2700.0	0.0	False	False	False	False	False	0	0	0	1	0
4	35	8	45000	1000.0	0.0	False	False	False	False	True	0	0	1	1	0

MODEL BUILDING
LOGISTIC REGRESSION

- Used Skleanr
- Best threshold for the model:
  - 0.081
- The model is not overfitted
- We get a recall of 92.62%

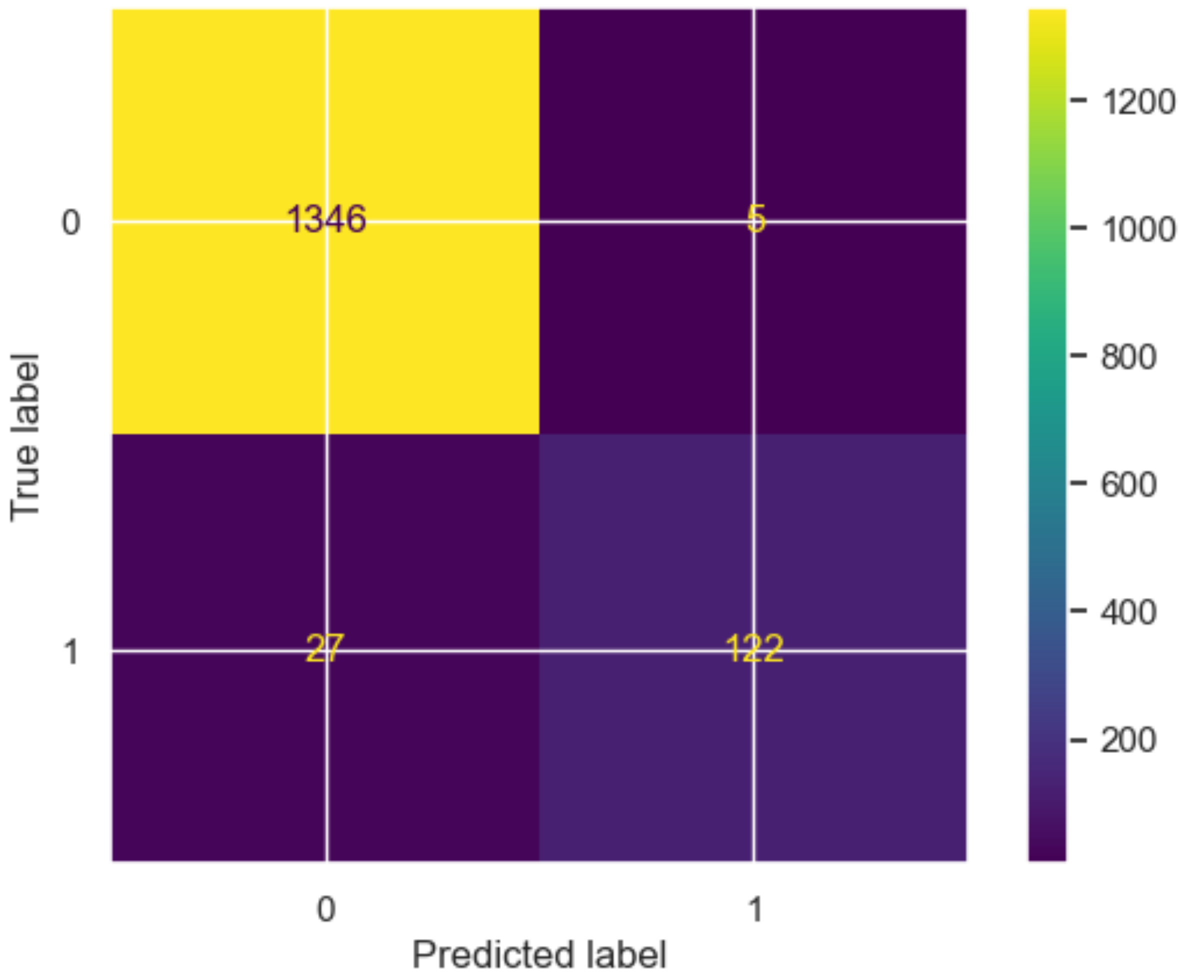


- The 5 most important features for this model are 'Experience', 'Income', 'Securities_Account', 'CD_Account', 'Online'

MODEL BUILDING
DECISION TREE

- Used GridSearchCV
  - Cpp_alpha 0.001
  - Criterion entropy
  - Max_depth 11
  - Min_samples_leaf 4
  - Min samples_split 2
- Recall on training set : 0.925
- Recall on test set : 0.82

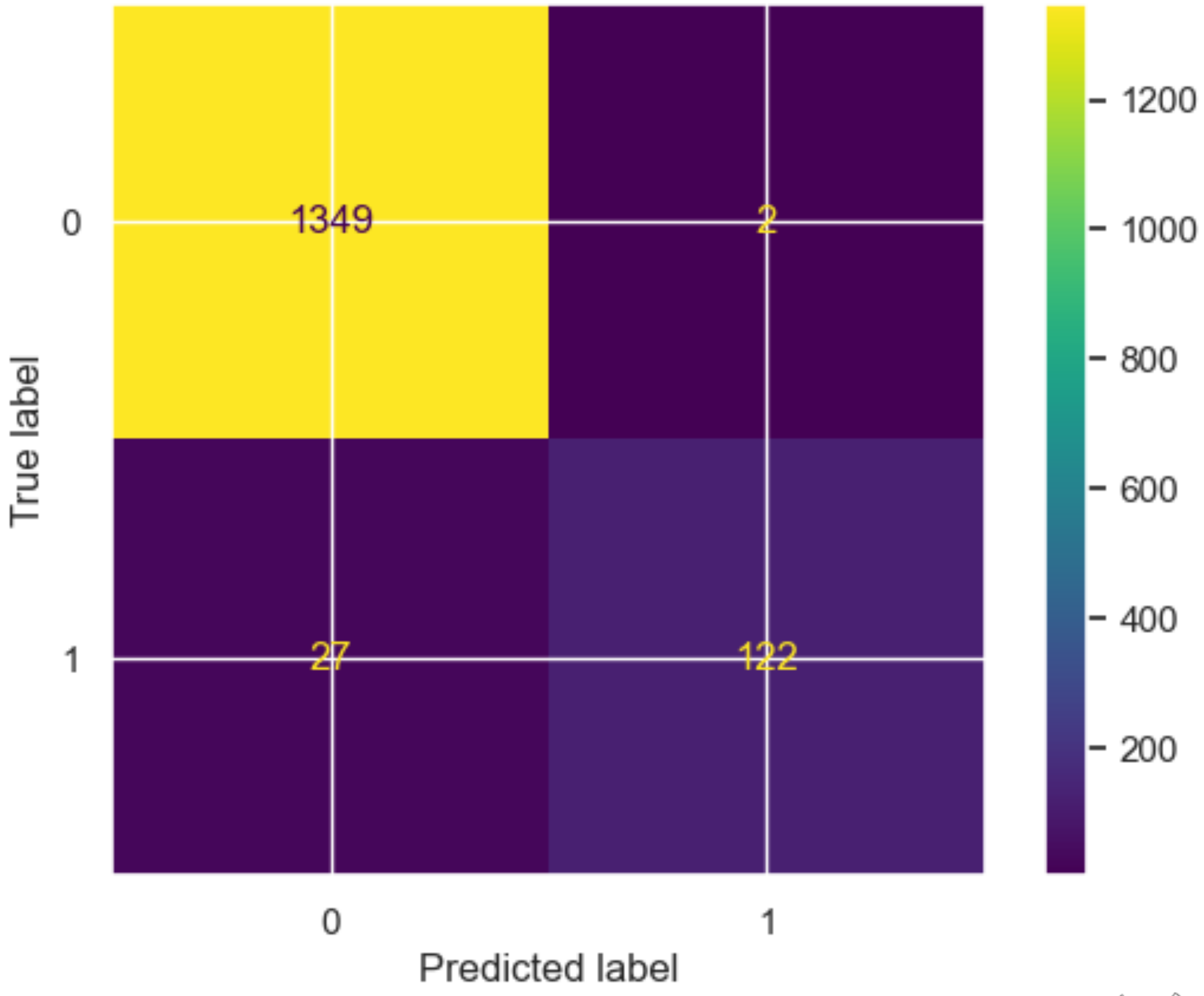
- It suggest the model is overfitted



MODEL BUILDING
RANDOM FORES

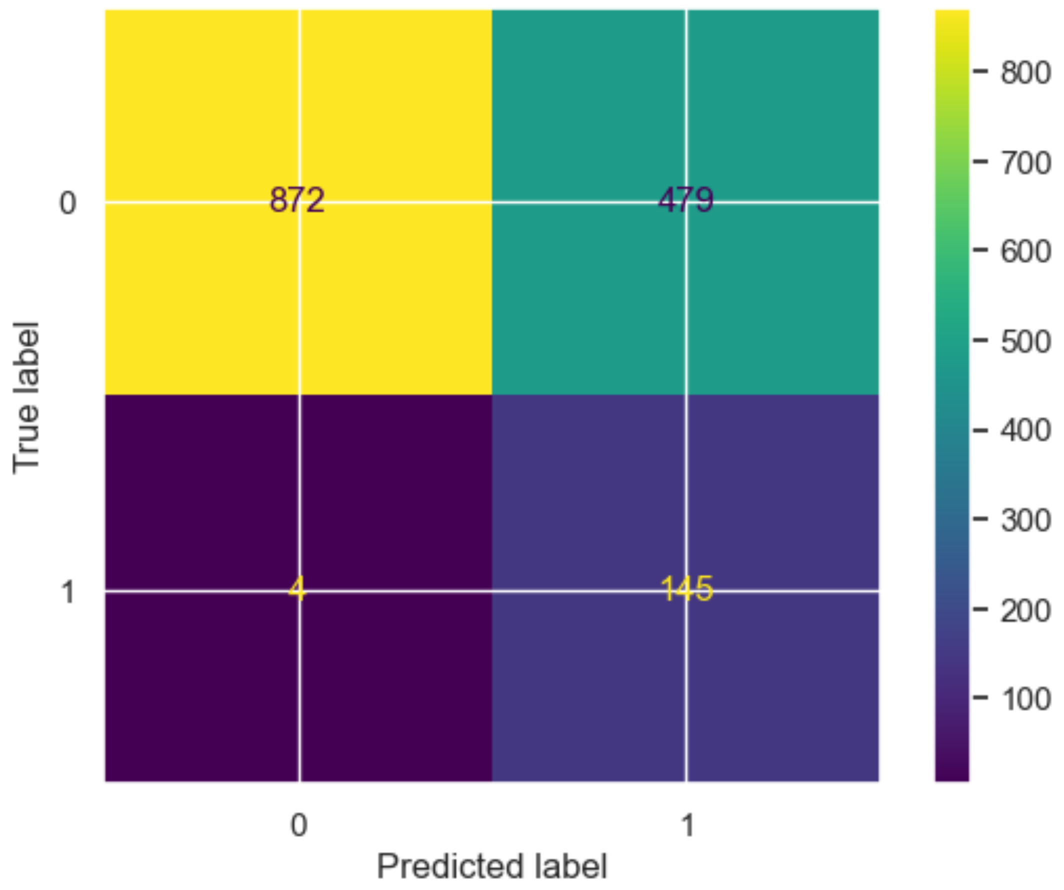
- Used GridSearchCV
  - max_depth=20,
  - max_features="sqrt",
  - min_samples_leaf=1,
  - n_estimators=20
- Recall on training set : 0.99
- Recall on test set : 0.82

- It suggest the model is overfitted

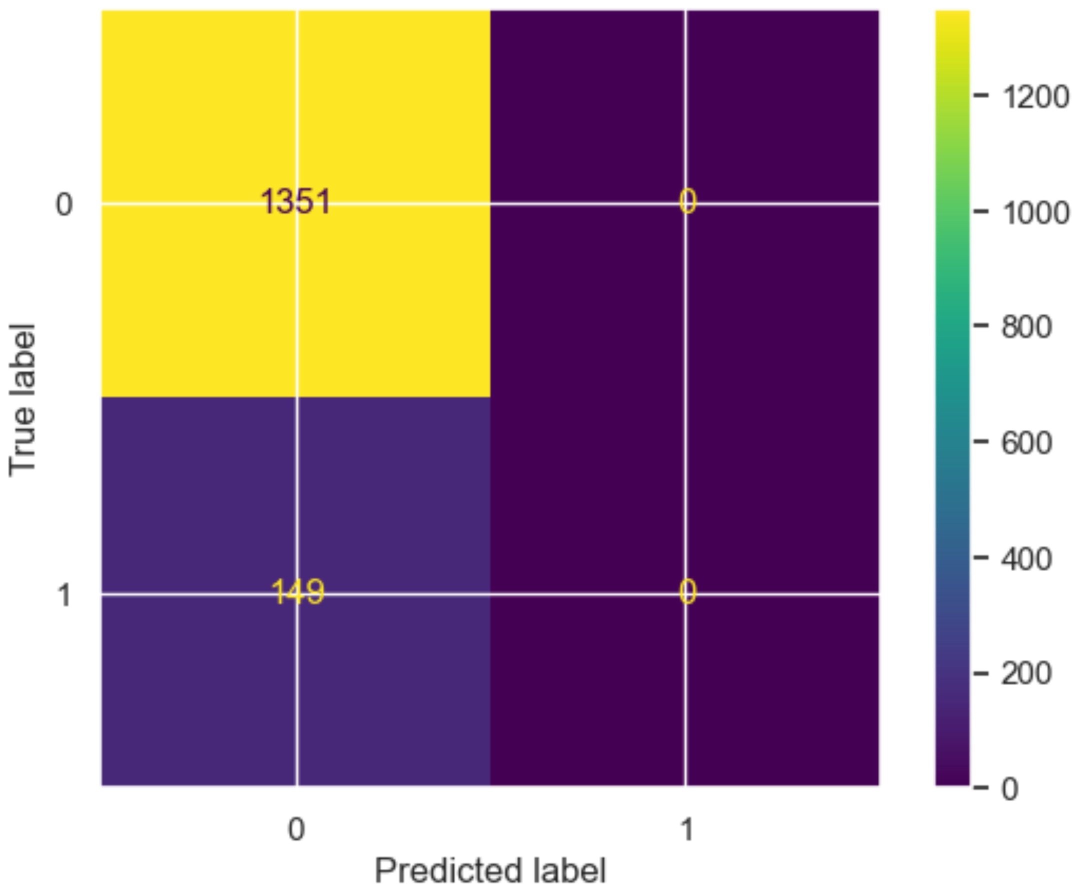


MODEL BUILDING
ADA BOOST

- Used GridSearchCV
  - learning_rate=2
  - n_estimators=10
- Recall on training set : 0.98
- Recall on test set : 0.9732



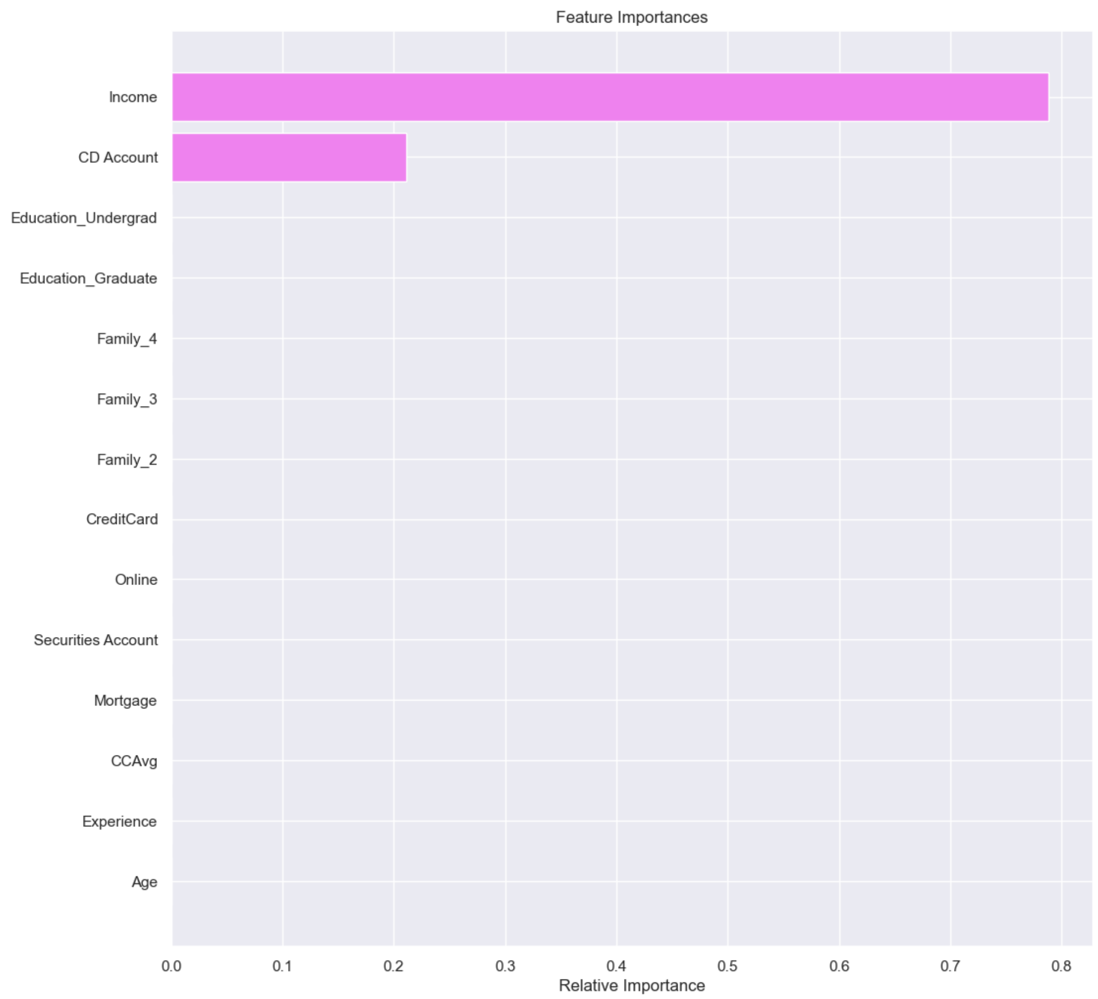
MODEL BUILDING VECTOR MACHINE



- Used GridSearchCV
  - C=0.03125,
  - gamma=0.03125,
  - kernel="rbf"
- Recall on training set : 0
- Recall on test set : 0

RESUMEN

- Recall Logistic Regression optimized: 0.9261744966442953
- Recall Decision Tree optimized: 0.8187919463087249
- Recall Random Forest optimized: 0.8187919463087249
- **Recall AdaBoost optimized: 0.9731543624161074**
- Recall SVM optimized: 0.0



CONCLUSIONS

- **We chose the AdaBoost model as our best model for this project.**

- Customers with high income >$100,000 are more likely to get a personal loan
- Customers with credit card usage over $2,000 are more likely to get a personal loan
- Customers with mortgages under $200,000 are more likely to get a personal loan