Business Case

Having models to accurately predict the acceptance of personal loan offers is incredibly important to banks, financial advisors, marketers, and policy makers within the banking sector. As such, our goal is to build a logistic regression model to forecast which customers will accept an offer of a personal loan from the bank. An acceptable level of model performance could be considered an area under the ROC curve > 0.7, indicating good predictive power. However, an ROC value > than 0.9 is ideal for this scenario. The data for this assignment was sourced from the Universal Bank dataset, encompassing a comprehensive range of customer demographic information and details of their relationship with the bank. A total of 5,000 data points related to customers' profiles and their previous interactions with bank services were initially analyzed. An additional analysis of the data was run excluding rows with negative experience values.

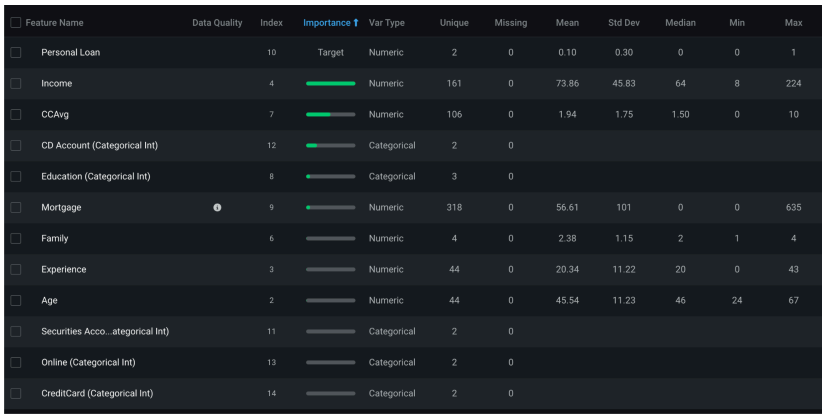
Q1. Explain data pre-processing steps.

- **Selected the following features for modeling:** Personal Loan, Income, CCAvg, CD Account, Mortgage, Education, Family, Experience, Securities Account, CD Account, Online, CreditCard, and Age. The target for this exercise was "Personal Loan."
- **Recoded features:** Changed Securities Account, CD Account, Online, CreditCard, and Education to be categorical features. I assumed that education is categorical because I interpreted the values as levels represented by numbers (e.g., 1 = high school, 2 = bachelor's degree, 3 = master's degree, etc.). Personal Loan, Securities Account, CD Account, Online, and CreditCard are also coded as Categorical because these are binary variables (0 or 1), which typically represent a "yes" or "no" answer.
- **Check for missing values:** Verified that there were no missing values in the data
- **Checked the data for errors:** I did see that there were negative values in the experience column (42 instances of -1, -2, or -3). This is possibly an error because "Experience" is likely representing the years of experience of an individual who either does or does not accept a loan, which can't be negative. As such, I ran the analysis with all data points included and negative values excluded to compare the impact.

With rows with negative experience values:

Personal Loan	10	Target	Numeric	2	0	0.10	0.29	0	0	1
Income	4		Numeric	162	0	74.08	46.34	64	8	224
CCAvg	7		Numeric	104	0	1.94	1.75	1.50	0	10
CD Account (Categorical Int)	12		Categorical	2	0					
Mortgage	9		Numeric	319	0	55.77	102	0	0	635
Education (Categorical Int)	8		Categorical	3	0					
Family	6		Numeric	4	0	2.40	1.15	2	1	4
Experience	3		Numeric	47	0	20.11	11.44	20	-3	43
Securities Account (Categorical Int)	11		Categorical	2	0					
CreditCard (Categorical Int)	14		Categorical	2	0					
Online (Categorical Int)	13		Categorical	2	0					
Age	2		Numeric	45	0	45.34	11.43	45	23	67

Without rows with negative experience values:



Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
Personal Loan		10	Target	Numeric	2	0	0.10	0.30	0	0	1
Income		4		Numeric	161	0	73.86	45.83	64	8	224
CCAvg		7		Numeric	106	0	1.94	1.75	1.50	0	10
CD Account (Categorical Int)		12		Categorical	2	0					
Education (Categorical Int)		8		Categorical	3	0					
Mortgage		9		Numeric	318	0	56.61	101	0	0	635
Family		6		Numeric	4	0	2.38	1.15	2	1	4
Experience		3		Numeric	44	0	20.34	11.22	20	0	43
Age		2		Numeric	44	0	45.54	11.23	46	24	67
Securities Acco...ategorical Int)		11		Categorical	2	0					
Online (Categorical Int)		13		Categorical	2	0					
CreditCard (Categorical Int)		14		Categorical	2	0					

Q2. Report Recall, Precision, F1, Error rate, Accuracy, ROC AUC. Does the model have predictive value? Explain (compare to naive).

	With Negatives	Without Negatives
Recall	0.6771	0.7604
Precision	0.8667	0.7956
F1	0.7602	0.7776
Accuracy	0.959	0.9578
Error Rate	0.041	0.0422
ROC AUC	0.9591	0.9638
Threshold	0.4731	0.3667

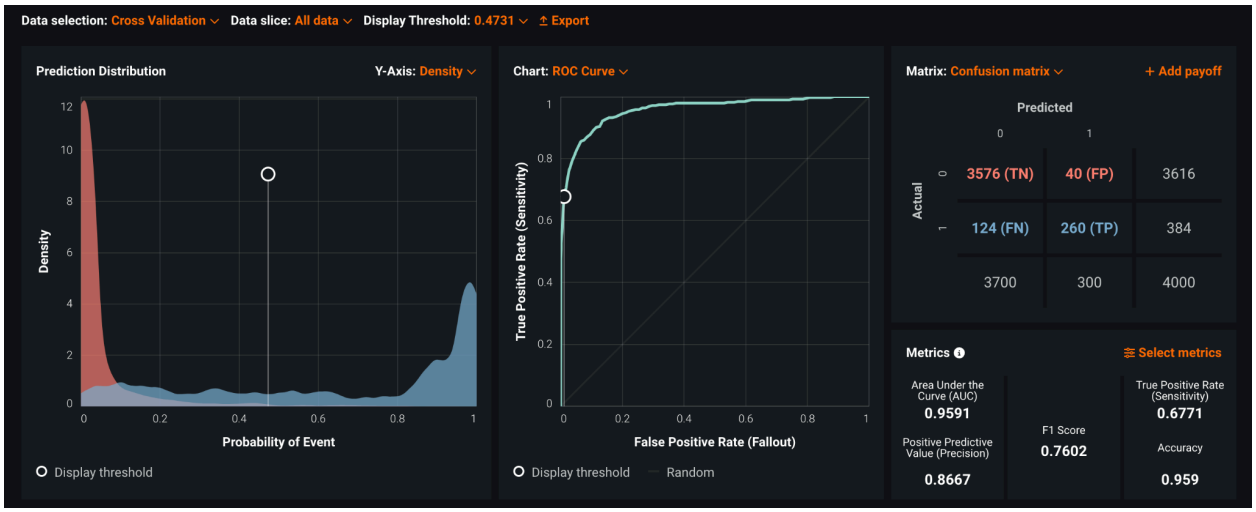
The removal of rows with negative experience values seems to improve the recall (from 0.6771 to 0.7604) and the ROC AUC (from 0.9591 to 0.9638), suggesting that the model becomes better at identifying true positives and distinguishing between classes. However, this comes with a slight trade-off in precision (from 0.8667 to 0.7956), where the model becomes slightly less accurate in predicting positive instances after the removal of negatives. The F1 score, which balances precision and recall, is slightly higher after removing negatives, indicating a better balance between these metrics. The accuracy and error rate are roughly comparable before and after the negatives are removed, with a small decrease in accuracy after their removal.

When comparing to a naive model, which would likely predict the majority class for all instances, both versions of the model demonstrate a much higher capability to correctly identify the minority class, as indicated by the F1 scores and ROC AUC values that are substantially higher than what would be expected by chance (ROC AUC of 0.5).

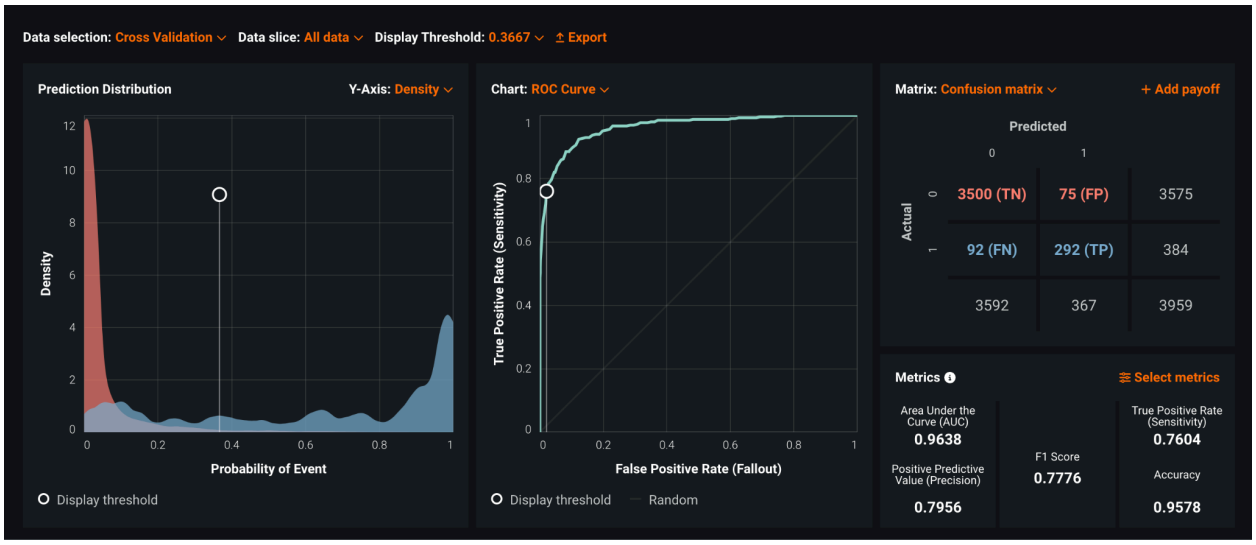
The optimization of thresholds (0.4731 for the model with negatives and 0.3667 for the model without negatives) suggests that the model's threshold was adjusted to maximize the F1 score in each scenario. This indicates a deliberate tuning of the model to balance the trade-off between precision and recall according to the dataset's characteristics.

In summary, the model shows a strong predictive value in both cases, and the changes in performance metrics reflect the impact of data cleansing on model behavior. The improvement in recall and ROC AUC suggests that cleaning the data of negative experience values helps the model to more effectively identify potential loan acceptances, which could be crucial depending on the business objective, such as targeting customers for loan offers.

With rows with negative experience values:



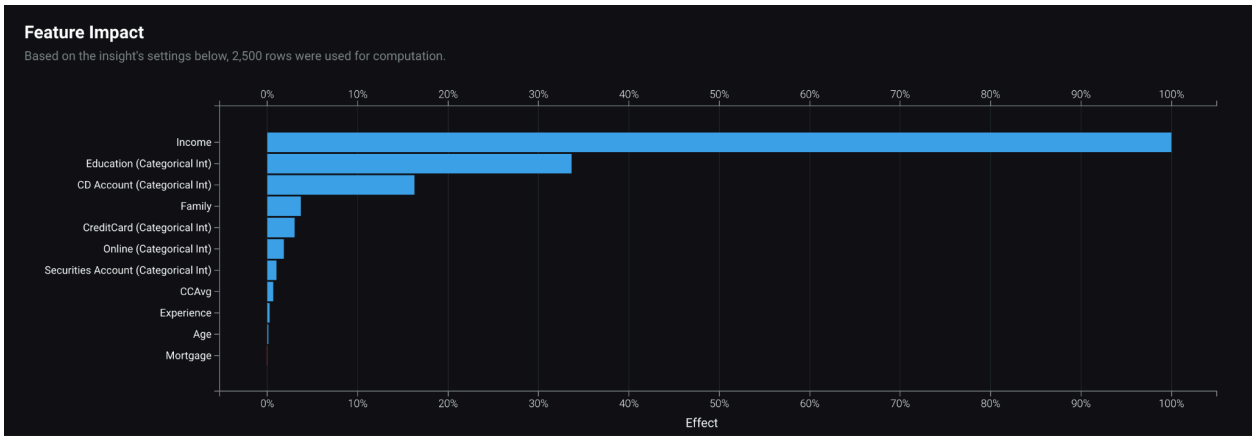
Without rows with negative experience values:



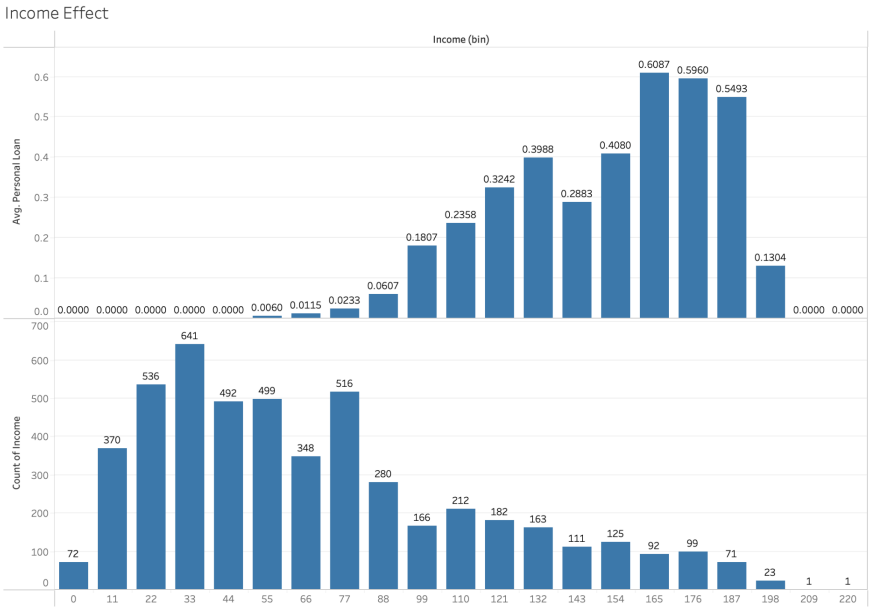
Q3. Which factors are important in the prediction of loan acceptance in the model? Provide visualizations and 1-sentence summary for the top 5 factor effects.

The following all had significant effects on the model. The top five features were the same even when rows with negative experience values were removed, although at different percentages.

With rows with negative experience values:

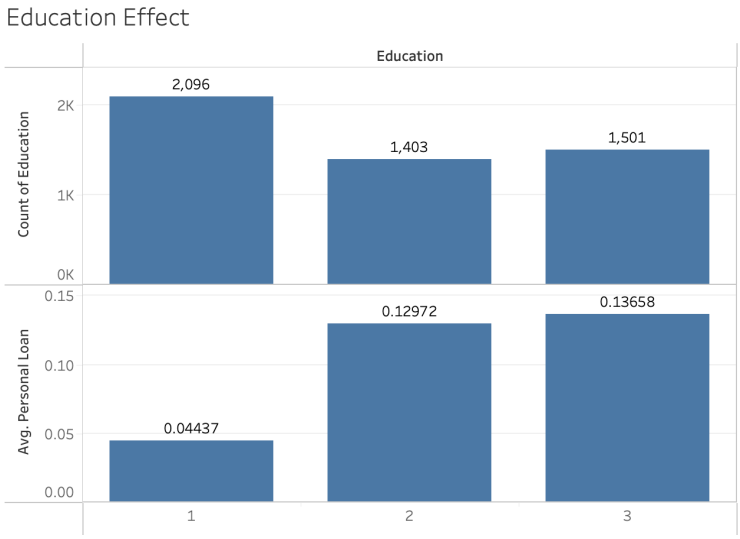


- **Income (100% Effect):** Income has a pronounced effect on personal loan acceptance, with a higher probability of loan acceptance among individuals in the upper-middle income ranges. This suggests that income is a significant predictor of personal loan acceptance, with those in certain higher income brackets being more inclined to take out loans, possibly due to greater financial leverage or creditworthiness.

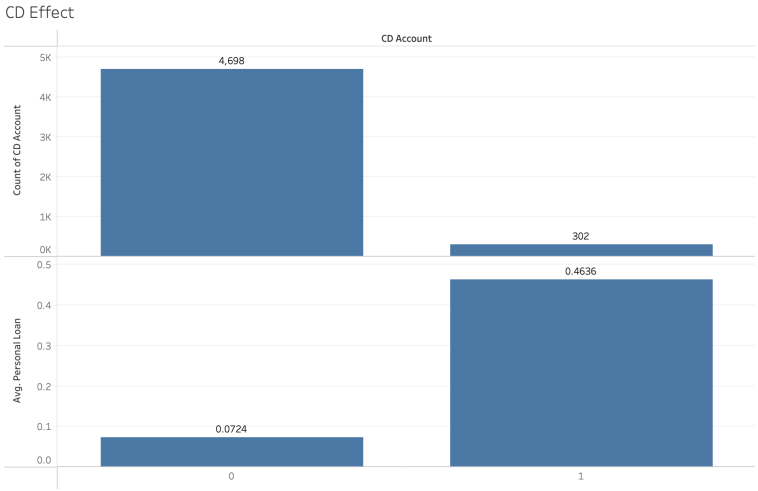


- **Education (33.66% Effect):** The effect observed here is that as the level of education increases, the likelihood of accepting a personal loan also increases. Despite Education Level 1 having the most individuals, those with higher education (Levels 2 and 3) are more inclined to accept a personal loan. This suggests that education is a significant

factor in predicting personal loan acceptance, with higher education levels correlating with a greater propensity to take out loans.

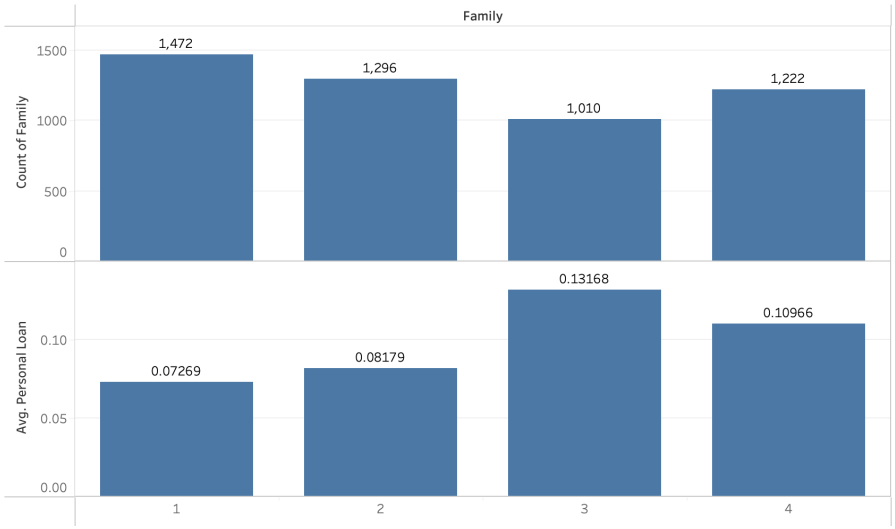


- **CD Account (16.27% Effect):** The presence of a CD account is a strong indicator of personal loan acceptance, with individuals holding a CD account being significantly more likely to accept a personal loan than those without one. This could be due to a variety of reasons, such as higher financial stability or a stronger relationship with the bank, which often comes with having more diversified banking products.



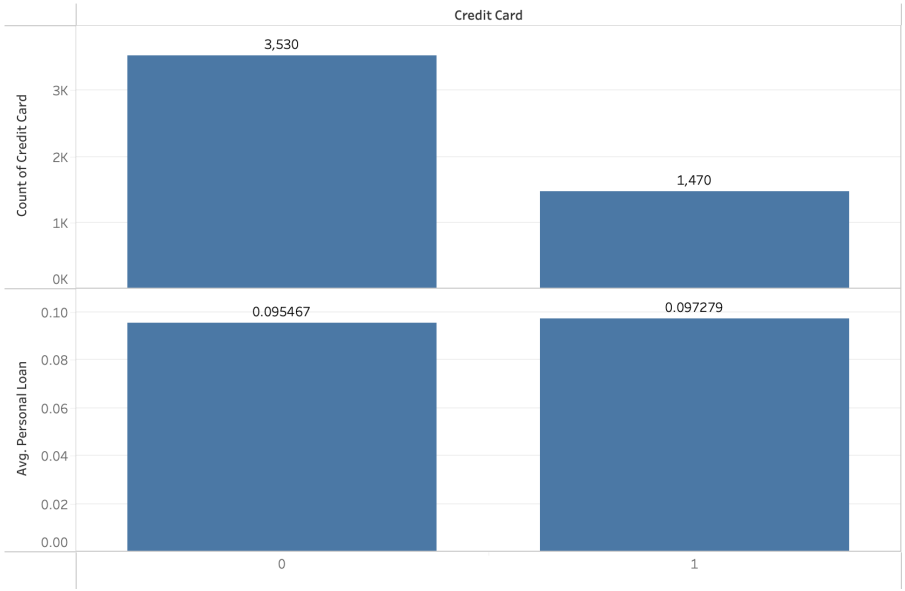
- **Family (3.71% Effect):** The implication here is that family size has a variable impact on personal loan acceptance, with a notably higher likelihood for those with three family members. This could suggest that individuals with three family members might have greater financial needs or preferences that make a personal loan more attractive or necessary compared to smaller or larger families.

Family Effect

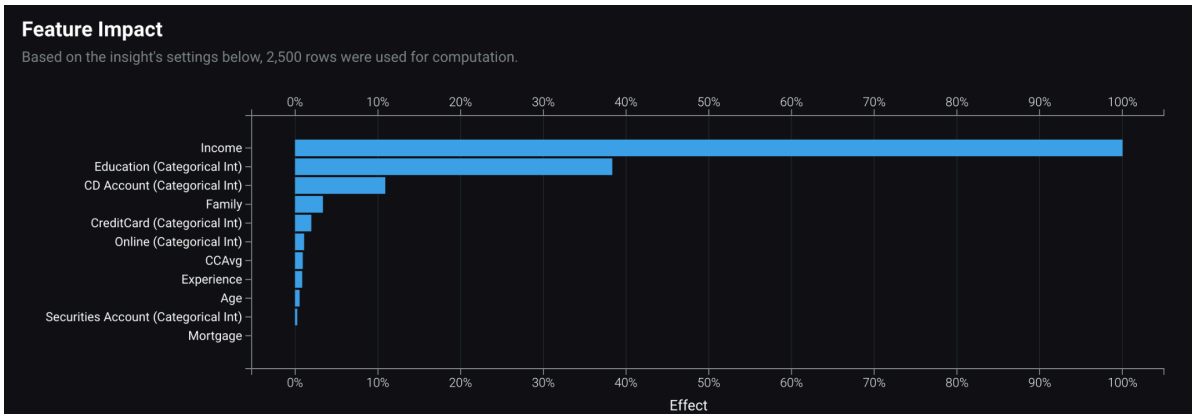


- **Credit Card (3.01% Effect):** We can infer that having a credit card has a very slight positive correlation with the likelihood of accepting a personal loan, but the effect is minimal. It suggests that individuals with a credit card are just slightly more likely to accept a personal loan than those without, but the difference is not substantial. This could imply that having a credit card does not significantly change the customer's behavior or decision-making process regarding personal loan acceptance.

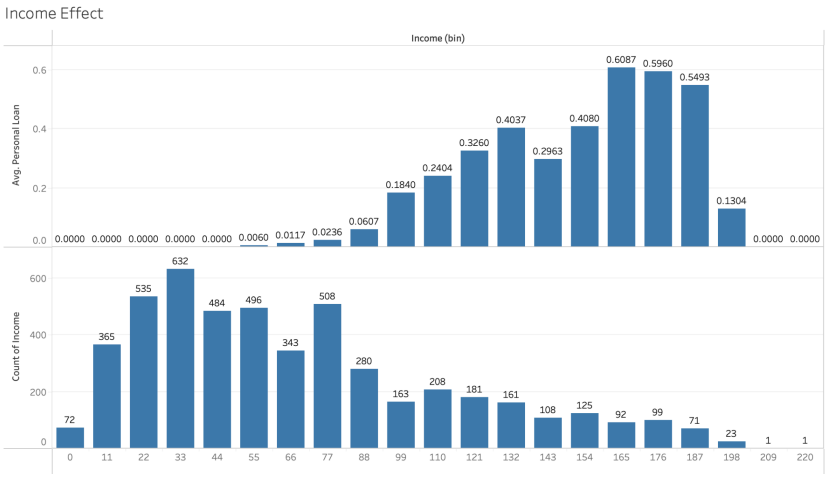
Credit Card Effect



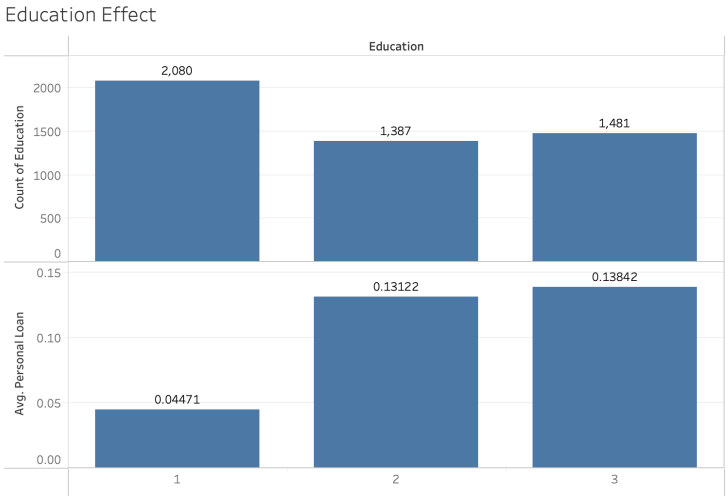
Without rows with negative experience values:
Comparing the charts with the ones including the rows with negative experience, the general patterns and trends appear consistent. However, I included the charts below that do not include the rows with negative experience values, but the general rationale above remains the same for these charts.



- **Income (100% Effect):**

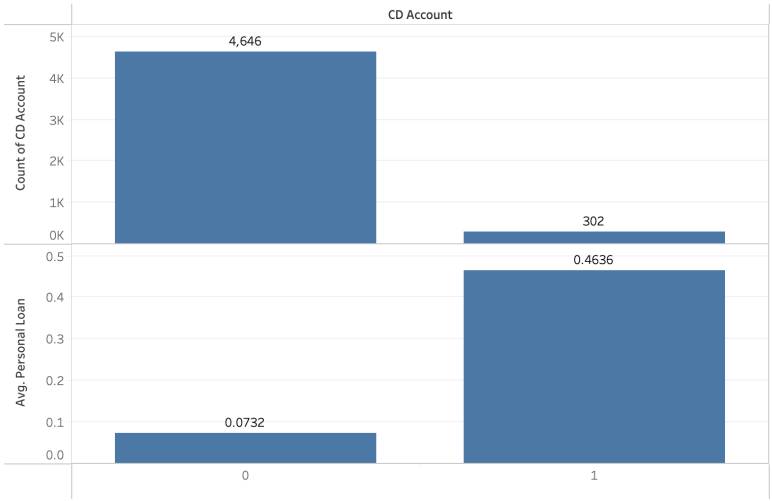


- **Education (38.34% Effect):**



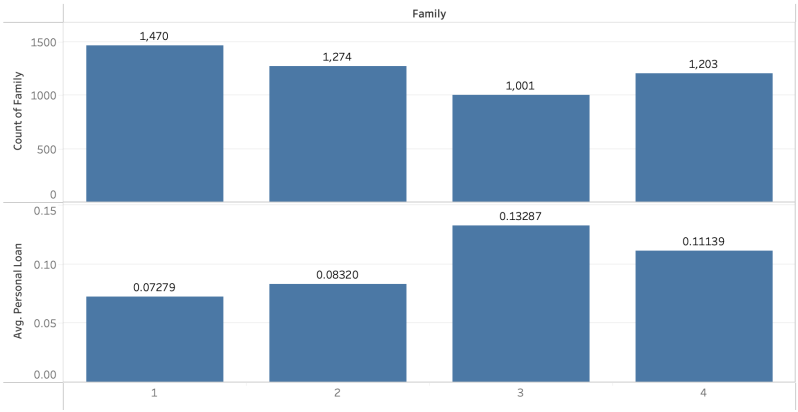
- **CD Account (10.91% Effect):**

CD Effect



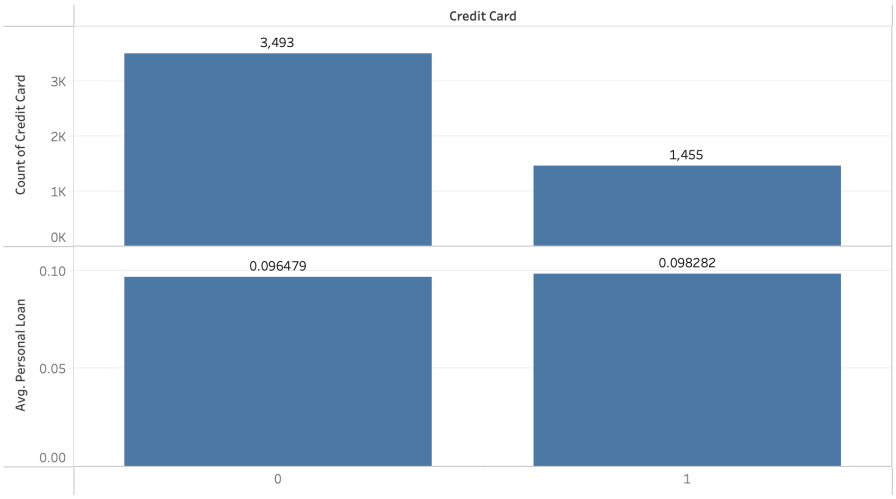
- **Family (3.37% Effect):**

Family Effect



- **Credit Card (1.97% Effect):**

Credit Card Effect



Q4. Does the model make sense?

The model in question demonstrates predictive value. This is evidenced by its ROC AUC scores, which are 0.9591 with negative experience values included and slightly higher at 0.9638 when those values are removed. The F1 score also indicates strong predictive performance, with the model achieving a score of 0.7602 with negatives and a slightly improved score of 0.7776 without them. The model's key features that predict the outcome, such as income, education, and CD account ownership, align well with intuitive financial behaviors. Higher income and education levels typically correlate with increased financial activity and product usage, which may include personal loans. CD account ownership could indicate a customer's engagement with the bank's products and their financial savviness, which may translate to a higher propensity to take personal loans. Additionally, the comparative analysis conducted after removing rows with negative experience values suggests that cleaning the data has a positive impact on the model's predictive accuracy. Notably, the recall increased from 0.6771 to 0.7604, indicating that the model became better at identifying true positive cases of loan acceptance after the data cleaning.