Case Study 4

Era Kalaja, Kathleen Wang, Cather Zhang

Group Name: DataMinds

(1) (5 points) What is your business proposition?

We are trying to solve the business problem of whether a client will subscribe to a term deposit and what type of client is most likely to do so, via analyzing data from a direct marketing campaign. By looking into past campaign data, we will identify certain client features that indicate they are more likely to subscribe, as well as predict whether a client will subscribe based on their demographic and financial data and social economic indicators at the time of the campaign, thus saving money on campaign costs and maximizing profit at the same time. Our solution will be derived from a data driven approach to create a reliable target audience for our company's campaigns. We want to ensure efficiency in our campaign marketing while optimizing the number of people who subscribe; therefore, we are looking for a solution that is both effective and reliable. Using our solution the company will spend less money via a targeted campaign and receive more term deposits.

(2) (3 points) Why is this topic interesting or important to you? (Motivations)

For banks, especially smaller ones, have limited resources they can spend on marketing campaigns. We believe the key to a successful campaign is to get as many clients to subscribe to with as few costs as possible, which is our motivation for this project. Costs and resources associated with a marketing campaign can include the search process for potential customers, acquiring demographic data of that population, the time and labor for making phone calls, etc. The first method is to analyze data to identify the type of population to search for to reduce costs associated with the search process. The second method is to develop a model to predict the response of a certain person to reduce costs associated with time and labor during the reach-out process. Through our research, we can help banks be more efficient with their marketing campaigns. Additionally, this solution is essential because increasing the number of term deposits will help our company invest money in other financial products that pay a higher rate of return than what the company is paying the customer for the use of their funds. We are interested

in the patterns of people who do subscribe to a term deposit and see a great deal of value in financial technology given their role in our country's economy.

(3) (5 points) How did you analyze the data? What conjectures did you make? Which conjecture did you use as the basis of developing your model? Why?

The original dataset had 21 columns and 41188 entries. There was no missing data. The dataset contained the following columns, broken into self-defined groups for readability:

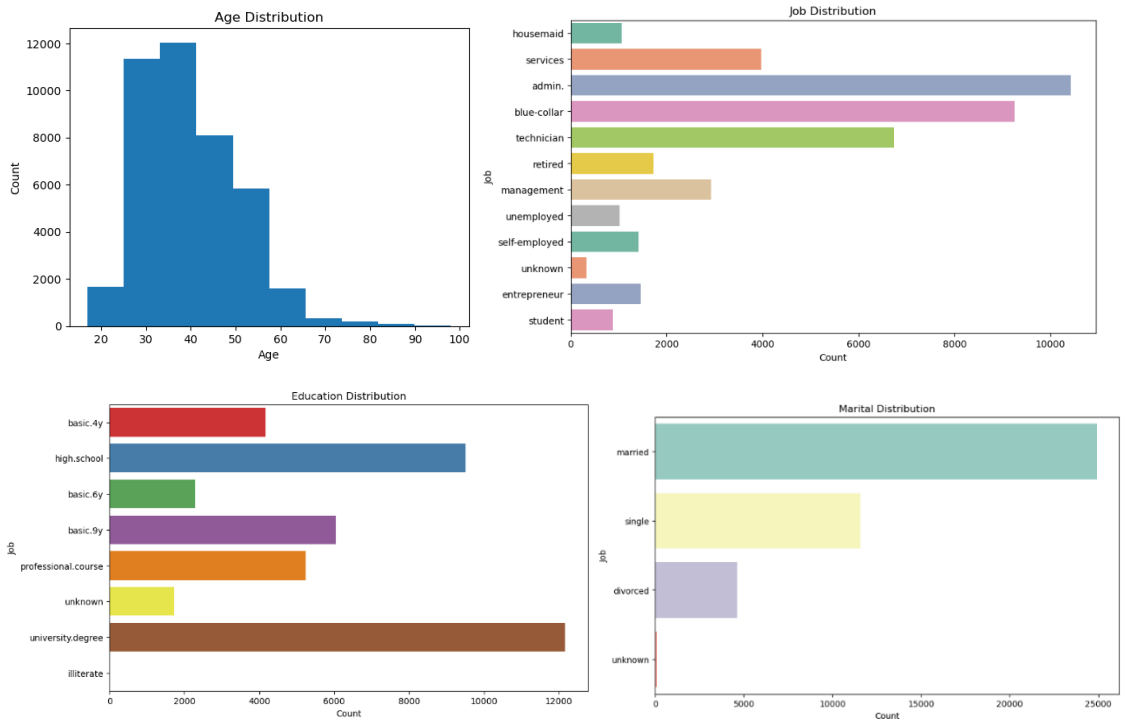
1. Demographic information: 'age', 'job', 'marital', 'education'
2. Debt: 'default', 'housing', 'loan',
3. Social economic indicators: 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed',
4. Campaign specific: 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome'
5. Target: 'y'

Additionally, here are some statistics based on the numeric data:

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

To analyze the data and build our understanding, we looked at data distribution of some features and performed some exploratory data analysis. We also did some background research on certain features, as they pertained to consumer and economic indexes.

Additionally, we did some feature engineering to replace columns given as objects to numerical values and dropping columns we felt were unneeded. This included many of the campaign-specific features as we are interested in identifying customer-specific features that would make them more inclined to subscribe. The demographic-specific distribution of our data can be seen below



Based on this preliminary data exploration, we developed 4 conjectures to test:

1. Demographic: ‘Higher’ level professions and education levels are more likely to subscribe.
2. Debt: Individuals facing already existing debts (such as defaults, personal loans, or housing loan) are less likely to subscribe.
3. Previous Campaign: Results from the previous marketing campaign can predict the results of this current campaign.
4. Social Economic Indicators: The behavior of the social and economic indicators impacts the campaign's success.

Within these groups we created several different conjectures based on common market principles and general knowledge of banking. Although not all conjectures were accepted, they provided a meaningful analysis of our data which helped us understand the patterns of the data.

(4) (3 points) How does your analysis support your business proposition? (please include figures or tables in the report, but no source code)

Demographic related analysis (Kathleen):

To determine if higher job levels were more likely to subscribe, we first ordered the job feature. This order was self-determined and may vary. However, it was determined that students and retired individuals had the highest subscription rate.

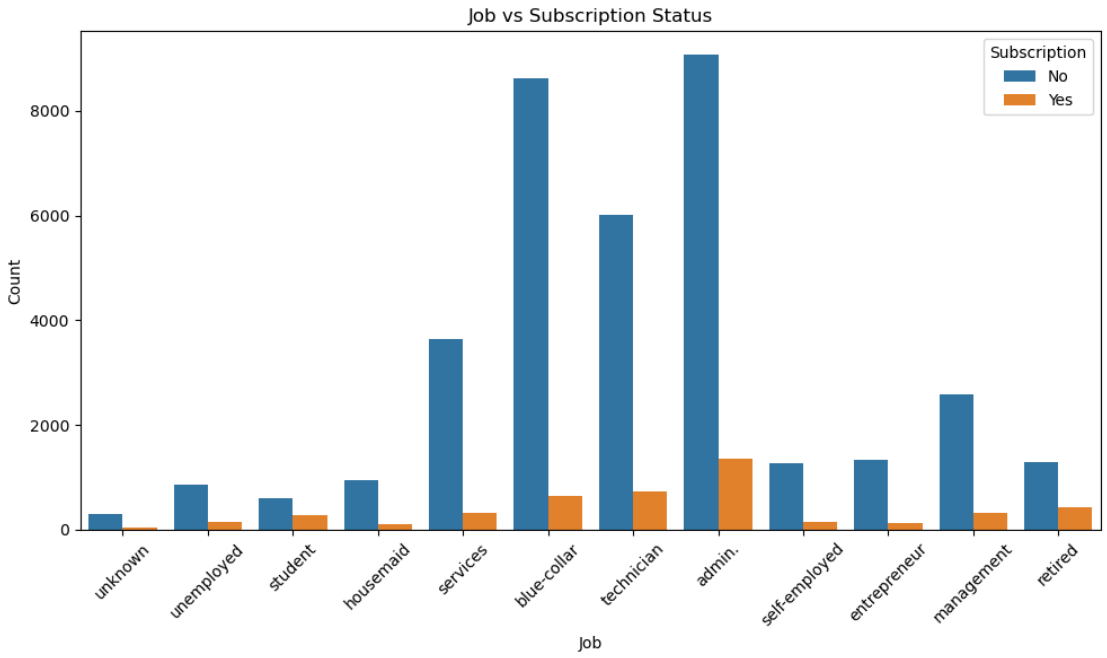


Figure 1. Job vs Subscription Status

Table 1. Percentage of each job categories that subscribed

Job Categories	Percentage that subscribed (%)
Admin.	12.972558
Blue-collar	6.894316
Entrepreneur	8.516484
Housemaid	10.000000
management	11.217510
retirement	25.232558
Self-employed	10.485574
services	8.138070
student	31.428571
technician	10.826042

Unemployed	14.201183
Unknown	11.212121

Next, we looked at education level. A greater percentage of individuals with higher education subscribed than those with only a basic education. This could indicate that higher education levels are more likely to subscribe, however the illiterate category has the highest subscription rate.

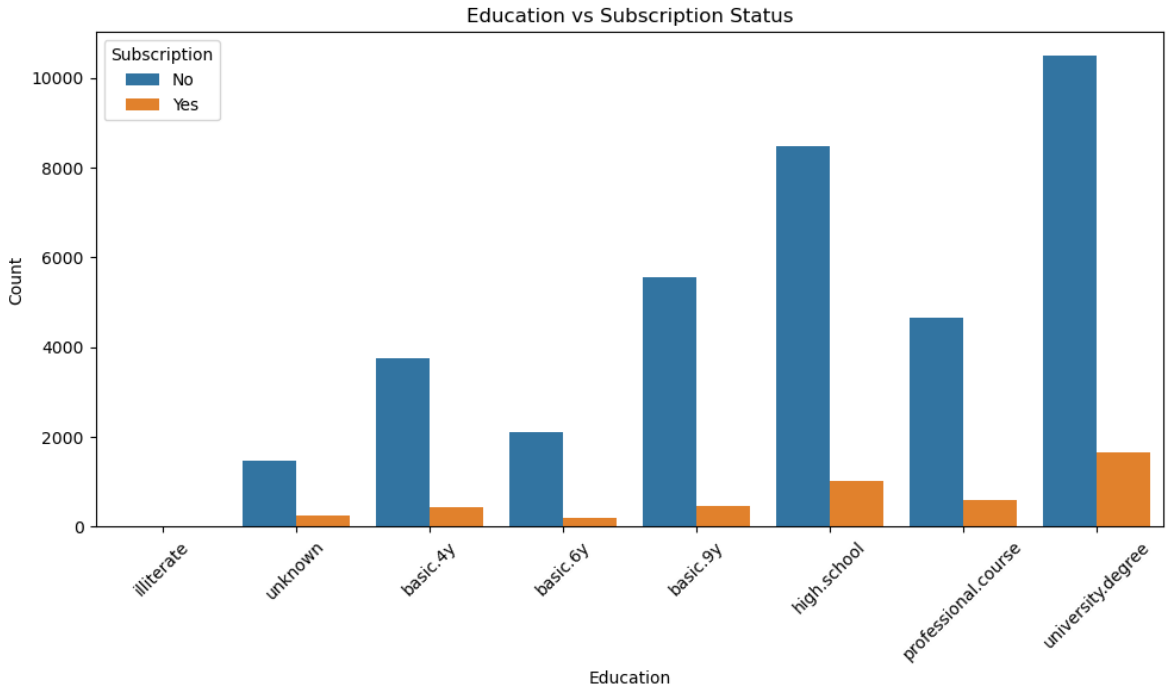


Figure 2. Education vs Subscription Status

Table 2. Percentage of each education category that subscribed

Job Categories	Percentage that subscribed (%)
illiterate	22.222222
unknown	14.500289
Basic.4y	10.249042
Basic.6y	8.202443
Basic.9y	7.824648
High school	10.835523

Professional course	11.348465
University degree	13.724523

Based off discoveries across both categories, we are inclined to reject half of the conjecture. It appears that higher education levels are more likely to subscribe, and there is less of the suggestion in the job category. It is essential to consider the exceptions of high subscription rate amongst students and illiterate individuals, which would allude to other factors having an impact on subscription rates.

Debt related analysis (Era):

The correlation coefficient of all debt related features with respect to the target are shown in Table 3. These corresponded with the conjecture that people with existing debt would be less likely to subscribe to a term deposit. However, these correlation coefficients were close to zero indicating that the relationship between each variable is nonlinear. Given this, we cannot accept this conjecture with the current data we have.

Table 3. Correlation Coefficients of different debt related features against target value

Debt Related Feature	Correlation Coefficient
default	-0.004
loan	-0.005
housing	0.012

Upon further analysis with a contingency table, we realized that there was not much known data present for these features. Given that this is data from a campaign this is something we were expecting to deal with since there are few people who typically respond positively. See the following Tables 4,5,6 for a representation of the contingency table outcomes. We believe that with more data the correlation coefficients may have been different because of the patterns in this data. This is supported by our statistical analysis using the p-value for the relationship between housing and the target. With a p-value of 0.03 we would have been able to accept my conjecture since it is less than significance level 0.05. Yet, the lack of data prevents us from doing so. We cannot state for certain that there is an association between these debt-related features and the target but given our current data we will not accept this conjecture.

Table 4. The Contingency Table for housing and target.

Target	No	Yes
Housing		
No	12797	1819
Yes	14930	2282

Table 5. The Contingency Table for default and target.

Target	No	Yes
Default		
No	28391	4197
Yes	3	0

Table 6. The Contingency Table for loan and target.

Target	No	Yes
Default		
No	23374	3478
Yes	4353	623

Demographic related analysis (Kathleen):

The correlation coefficient between 'poutcome' (past outcome of previous campaigns) and 'y' (current campaign outcome) was 0.19, indicating a weak positive relationship. There is some level of association between previous campaign results and this campaign, but it is rather limited if we were to only look at poutcome. Therefore, it is important to look at a combination of other features.

Social economic indicator analysis (Cather):

The correlation coefficients of all social economic indicator values against target values were calculated and shown below in Table 7. Amongst which, the indicator "nr.employed" is the most correlated to the target value. Therefore, we also computed coefficient correlations of other social economic indicators against the "nr.employed" column as shown below in Table 7.

Table 7. Correlation Coefficients of different social economic indicators against target

Social Economic Indicator	Correlation Coefficient
emp.var.rate	-0.298334
cons.price.idx	-0.136211
cons.conf.idx	0.054878
euribor3m	-0.307771
nr.employed	-0.354678

Table 8. Correlation Coefficients of other social economic indicators with "nr.employed"

Social Economic Indicator	Correlation Coefficient
emp.var.rate	0.906970
cons.price.idx	0.522034
cons.conf.idx	0.100513
euribor3m	0.945154

Based on the previous calculation, it can be concluded that employment variation rate, Euribor 3-month rate, and number of employees columns are very correlated with each other and have some correlation with the target value. Therefore, we computed the mean and median value of employment variation rate and Euribor 3-month rate for the two target values (yes and no), respectively.

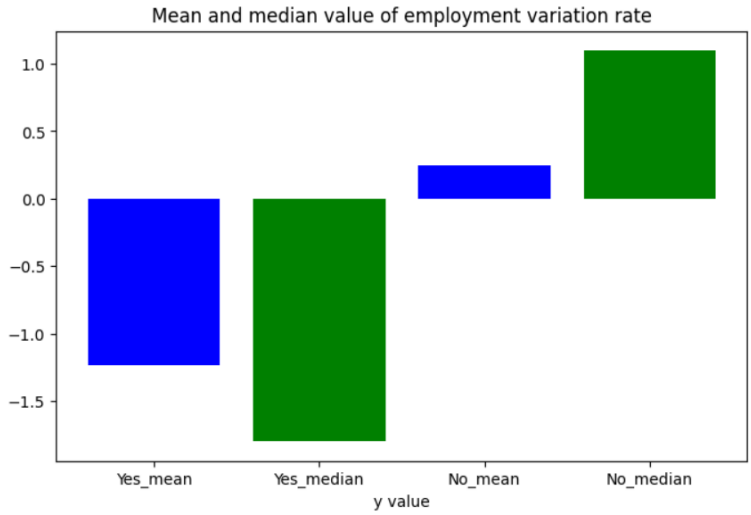


Figure 3. Mean and Median Value of Employment Variation Rate for Both Target Values

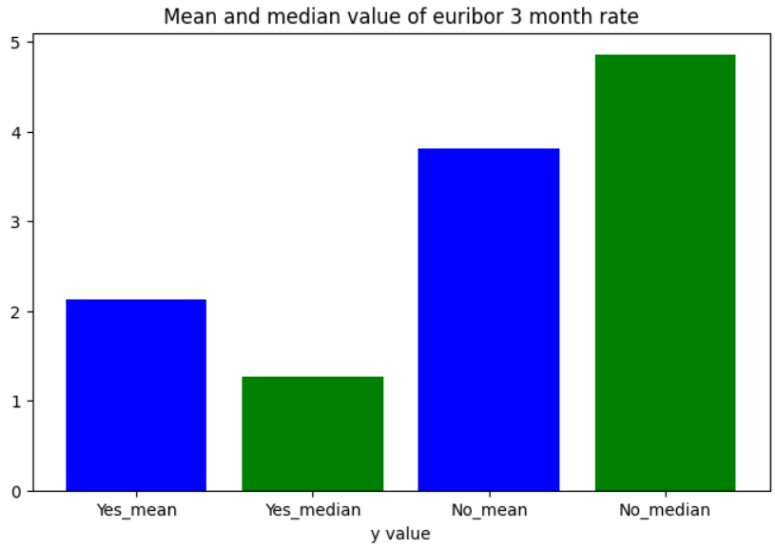


Figure 4. Mean and Median Value of Euribor 3-Month Rate for Both Target Values

As shown in Figure 3, the mean and median values for employment variation rate are both negative for target value with “yes” and positive for target value with “no.” Figure 4 shows that mean and median values for Euribor 3-month rate are lower for target value with “yes” and higher for “no”. One explanation is that, during periods of economic uncertainty or job market downturns, people tend to seek a safer and more secure way to invest, such as subscribing to a termly deposit.

(5) (4 points) How does the model tie in with the business proposition?

Given a pool of candidates with their information, the model can be used to predict whether a person will subscribe to a termly deposit. If the model has a high true positive and a relatively low false positive rate, the campaign can be very cost effective for the bank to conduct. To find the best model for our problem, we tried random forest, K-th nearest neighbor (KNN) classifier, and logistic regression algorithms, first with a majority of the features, and again with many certain features dropped. Based on the first set of model testing, we found that logistic regression had the highest accuracy and therefore decided to run feature importance on that model. This model had 16 features.

Table 9. Results from testing model with 16 features

Model	Test Accuracy
Random Forest	0.8958
KNN	0.8886
Logistic Regression	0.9048

We also identified the feature importance of the columns to eliminate some.

	Feature	Coefficient	Abs_Coefficient
12	cons.price.idx	0.323182	0.323182
14	euribor3m	−0.181200	0.181200
11	emp.var.rate	−0.166905	0.166905
4	default	−0.082857	0.082857
2	marital	−0.078254	0.078254
3	education	0.057127	0.057127
7	campaign	−0.054267	0.054267
13	cons.conf.idx	0.025195	0.025195
9	previous	0.021180	0.021180
1	job	−0.019857	0.019857
6	loan	0.008175	0.008175
5	housing	−0.007399	0.007399
0	age	0.006030	0.006030
15	nr.employed	−0.005762	0.005762
10	poutcome	−0.005166	0.005166
8	pdays	−0.001808	0.001808

Figure 5. Feature importance for logistic regression

The economic indexes are significant predictors of the outcome. Certain demographic features and debt-related features were also relevant to predicting the outcome. Therefore, we decided to run the models again, eliminating the following features: duration, month, day_of_week, contact,

age, job, housing, loan. Among all, logistic regression concluded the highest accuracy of 90.51%, compared to 88.81% for random forest and 88.30% for KNN classifier. The confusion matrixes for each algorithm are also displayed below.

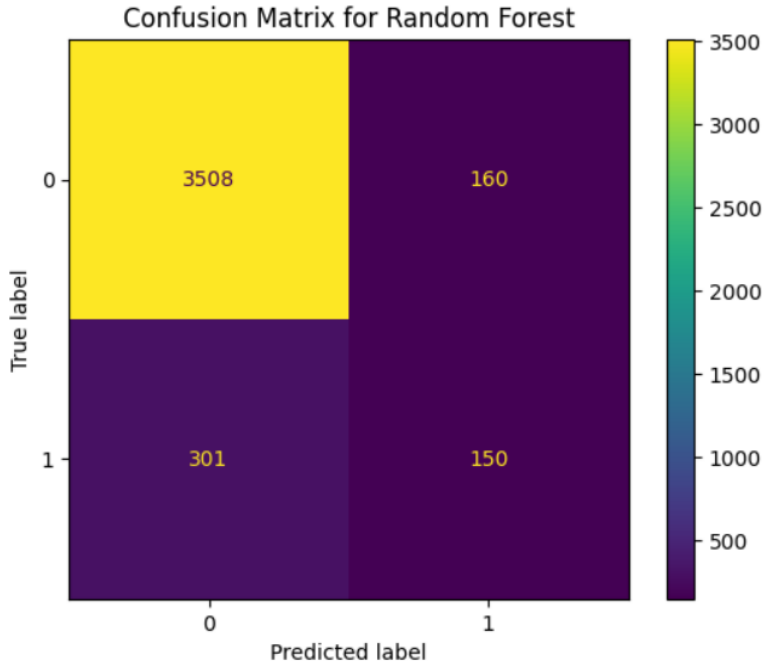


Figure 6. Confusion Matrix for Random Forest

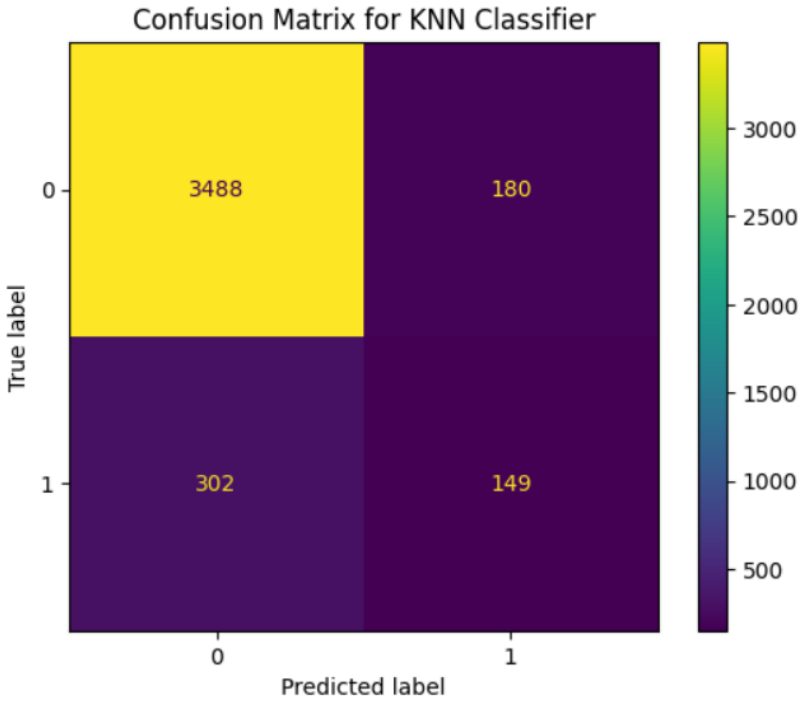


Figure 7. Confusion Matrix for KNN Classifier

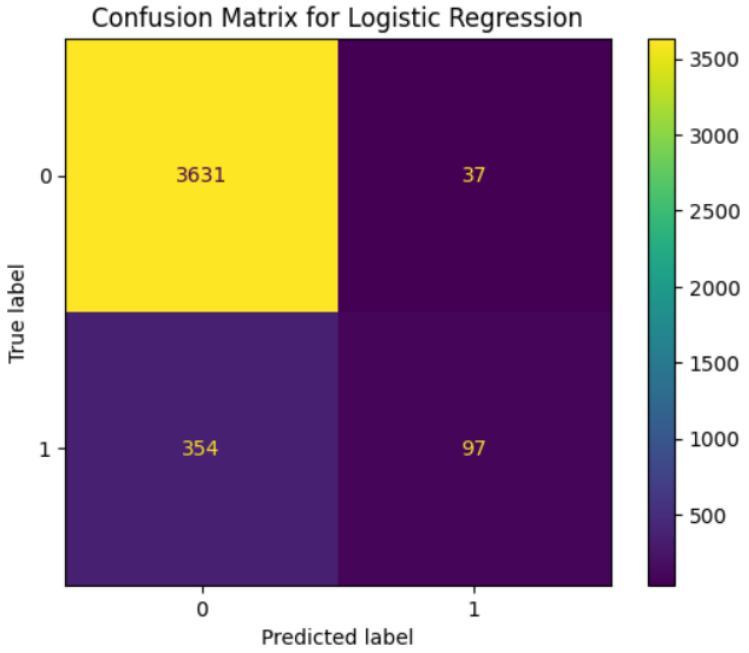


Figure 8. Confusion Matrix for Logistic Regression

From the confusion matrixes, the percentage accuracy for this business problem can be calculated as: Accuracy = true positive / (true positive + false positive). This is because the agents from the bank will have a list of candidates to be contacted based on if the model predicts 1 for a candidate. Thus, the accuracy for logistic regression is the highest – 72.39%. That says, if the bank uses this model for prediction and only selects the ones who are predicted yes to contact, 7 in 10 people they reach out will respond with yes.

However, one limitation is that the bank will lose many potential customers (false negative values) and only yield 97 customers from this pool of candidates. But this limitation can be reduced by simply searching for more candidates to fill up the gap.

We attempted to reduce the number of false negative predictions by adjusting class weights (higher weight for predicting 1) for the logistic regression algorithm. The result yielded more customers given the same number of candidates, as shown below. But the accuracy is lower, as a trade-off. Specific methods should be chosen based on the specific cost to acquire more candidate information to reach out to.

Figure 9. Confusion Matrix for Logistic Regression with Adjusted Class Weights

(5) (bonus 5 points) How did your team work together as a group from ideation to implementation? Write on one page.

Our team first met to discuss our common interests and future career goals to ensure we would all benefit from the dataset we chose. Once we established our goal of Fintech data, we analyzed the potential datasets and chose our favorite. We then identified the business question our solution would be addressing and began analyzing the given features. We divided the EDA phase into 3 sections and each specialized in a specific section. Cather focused on the economic features; Kathy focused on the demographic features; Era focused on the debt related features. Bringing together our conjectures and analysis of feature importance, we constructed an effective model that has both high training accuracy and a low false positive rate. Before arriving on a final model, we each tried numerous combinations, including different algorithms and conjectures. We narrowed down which model was our final solution by analyzing the pros and cons of each and agreeing on a final. Overall, we ensured each team member was comfortable speaking their mind and contributing equally. We feel we worked very well in our team dynamic, and we are proud of the solution our teamwork ensued. Throughout this report each of

us tied together our ideas and tangible outcomes from the code. We feel we learned a lot of team collaboration on a project from start to finish. Specifically, we learned how to effectively communicate obstacles, have time efficient meetings, and effectively meet deadlines. Each team member contributed their unique idea and skill to the project.