Predicting Marketing Results: An Application of Machine Learning

Team 3: Chaoran Jin, Eric Newman, Xingyu Yang

The George Washington University

DATS 6103: Introduction to Data Mining

Dr. Edwin Lo

17 December 2021

Predicting Marketing Results: An Application of Machine Learning

Effective marketing techniques are crucial for any business seeking to gain customers. Marketing is a main source of revenue for many companies. It is so important that many organizations are now incorporating marketing into the C-suite level of decision making with a Chief Marketing Officer. Furthermore, analytics play an important role in marketing: Harvard Business Review has emphasized that marketing must use analytics to shed light on what drives revenue, and not what is just easy to measure. Machine learning can play a critical role in determining what drives revenue. This paper will focus on machine learning applications to a bank's marketing campaign.

Section 1. Background

1.1. Data Source and Details

Our team analyzed this problem using a dataset of a direct marketing campaign (telephone marketing) for a Portuguese banking institution. The dataset is provided by Kaggle and can be found here: https://www.kaggle.com/ruthgn/bank-marketing-data-set. The goal of the marketing campaign is to get customers to subscribe to a term deposit - this is the response variable. The predictor variables include information on the bank client, last contact with the client, campaign statistics for this client, and social and economic indicators. There are 41,118 observations (each one person) and 21 variables (20 predictors and 1 response). Out of this dataset, 4,640 people subscribed (~11.3%) and 36,548 did not subscribe (~88.9%). An image of the data structure is printed below.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
```

1.2. Data Dictionary

The following gives the data dictionary for the dataset, providing details regarding each variable, along with the category of variables they belong to.

Bank Client data:
age: age of individual
job: job
marital: marital status
education: highest level of education
default: has credit in default
housing: has housing loan
loan: has personal loan

Related to last contact of current campaign:
contact: contact communication type
month: last contact month of year
day_of_week: day of week last contact made
duration: last contact duration in seconds (0 → y = 0)

Other attributes:
campaign: # of contacts within this campaign

pdays: # of days passed after last contact from previous campaign
previous: # of contacts performed before this campaign
poutcome: outcome of previous marketing campaign

Social and economic context attributes:
emp.var.rate: employment variation rate - quarterly indicator
cons.price.idx: consumer price index - monthly indicator
cons.conf.idx: consumer confidence index - monthly indicator
euribor3m: euribor 3 month rate (Eurozone 3-mo. interest rates)
nr.employed: quarterly avg. of total employed citizens

Output variable - desired target
y: did the client subscribe to a term deposit?

1.3. SMART Questions

Our SMART questions were tailored toward building models to help the bank make optimal use of its marketing campaigns. Specifically, this paper addresses the following questions:

1) Are there any groups of potential customers that can inform marketing?
2) What factors are related to a customer subscribing to a term deposit?
3) Can we predict who will subscribe based on demographic or economic factors, or based on marketing techniques?
4) Which machine learning model best predicts customer behavior?

Section 2. Exploratory Data Analysis: Principal Components Analysis

To answer SMART question 1 (are there any groups of potential customers that can inform marketing), our team decided to employ the unsupervised learning method of principal components analysis (PCA) for visual exploratory data analysis (EDA). We decided to use PCA because this dataset has a large number of features, and we wanted to explore them in an

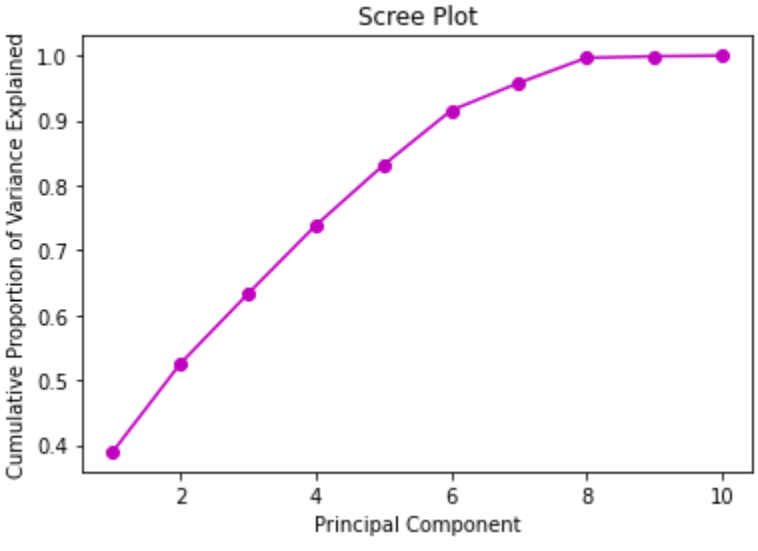
efficient manner. Our PCA was limited to the 10 numeric predictor variables, but we examine the categorical variables as well in Section 3 with lasso regression and stepwise selection.

2.1. Loading Vectors

The loading vectors for the 10 principal components are displayed below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
age	0.001577	0.251901	-0.635283	-0.253371	0.445018	-0.519054	-0.031302	-0.017883	-0.001877	-0.001360
duration	0.025564	0.081409	-0.040921	0.767886	0.590895	0.222453	-0.037596	0.036557	0.001291	-0.001382
campaign	-0.100491	-0.007935	0.324029	-0.575546	0.619754	0.411272	-0.004578	0.015618	-0.000011	0.009287
pdays	-0.227537	-0.628712	-0.252674	-0.006720	0.054466	0.017271	0.222202	0.660617	-0.002348	-0.000766
previous	0.305815	0.474453	0.281755	-0.021267	-0.034169	-0.146374	-0.192525	0.735882	0.018261	-0.004224
emp.var.rate	-0.488002	0.163001	0.091015	0.044439	-0.030689	-0.075546	-0.070398	0.047342	-0.793880	-0.284488
cons.price.idx	-0.366098	0.279060	0.276172	0.073401	0.040593	-0.249851	0.731721	0.003923	0.311417	-0.099768
cons.conf.idx	-0.101573	0.427669	-0.510937	-0.070490	-0.238990	0.647574	0.172999	0.120583	0.070796	-0.121687
euribor3m	-0.490377	0.148132	0.002732	0.036446	-0.063125	0.004006	-0.216623	0.052635	0.063428	0.823730
nr.employed	-0.470095	-0.013535	0.029958	0.027276	-0.029714	-0.052240	-0.542144	0.024394	0.513219	-0.464398

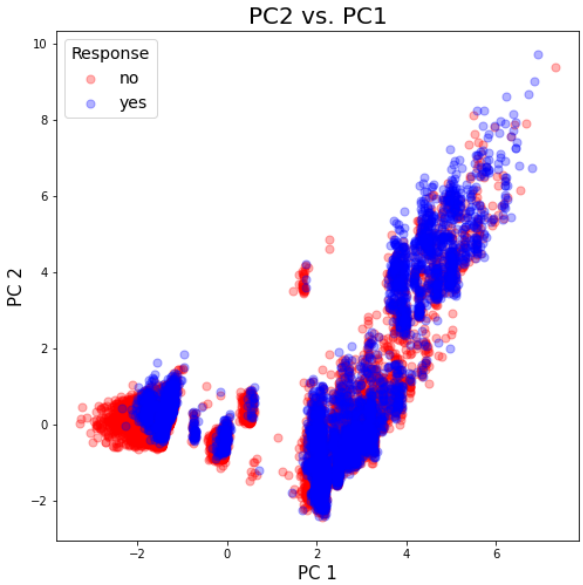
To determine which principal components to examine, we plotted the cumulative proportion of variance explained by each principal component. The scree plot displayed below shows that roughly 85% of the variance is explained by the first five principal components. We examined the first five principal components, but we highlight comparisons between the first four in this paper, since the fifth principal component did not yield significant results. The strongest predictor variable loads are highlighted for the first four principal components in the image on the bottom-right.

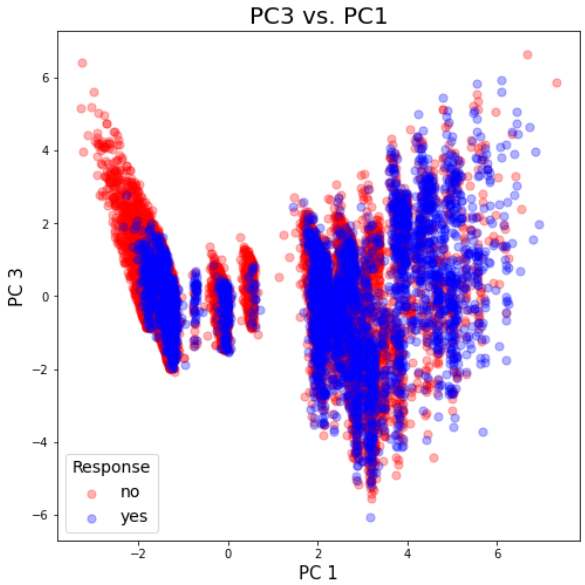


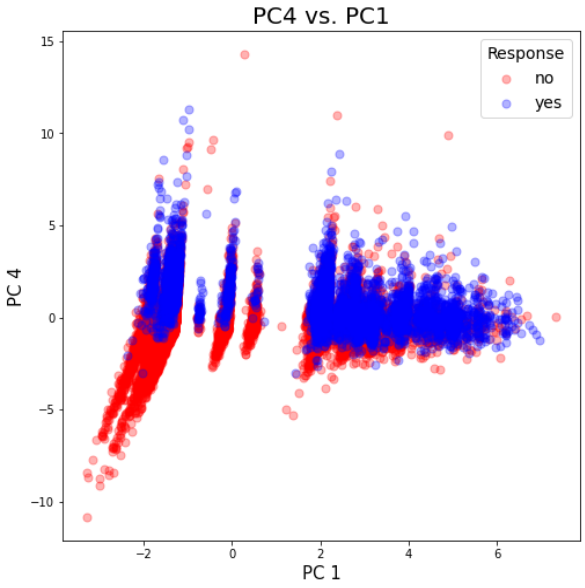
	PC1	PC2	PC3	PC4
age	0.001577	0.251901	-0.635283	-0.253371
duration	0.025564	0.081409	-0.040921	0.767886
campaign	-0.100491	-0.007935	0.324029	-0.575546
pdays	-0.227537	-0.628712	-0.252674	-0.006720
previous	0.305815	0.474453	0.281755	-0.021267
emp.var.rate	-0.488002	0.163001	0.091015	0.044439
cons.price.idx	-0.366098	0.279060	0.276172	0.073401
cons.conf.idx	-0.101573	0.427669	-0.510937	-0.070490
euribor3m	-0.490377	0.148132	0.002732	0.036446
nr.employed	-0.470095	-0.013535	0.029958	0.027276

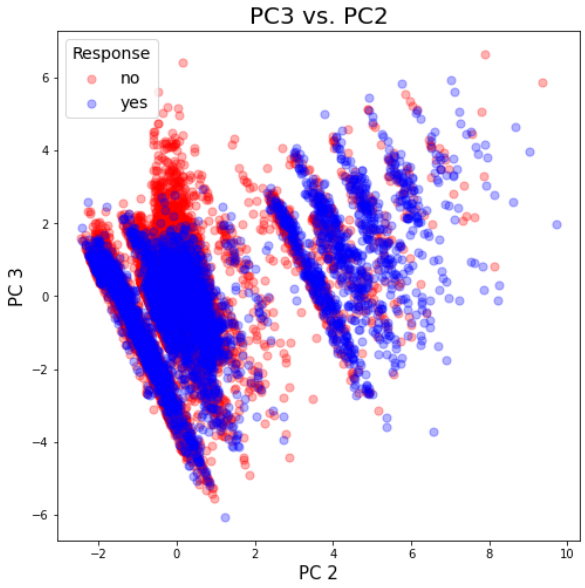
2.2. Principal Component Plots

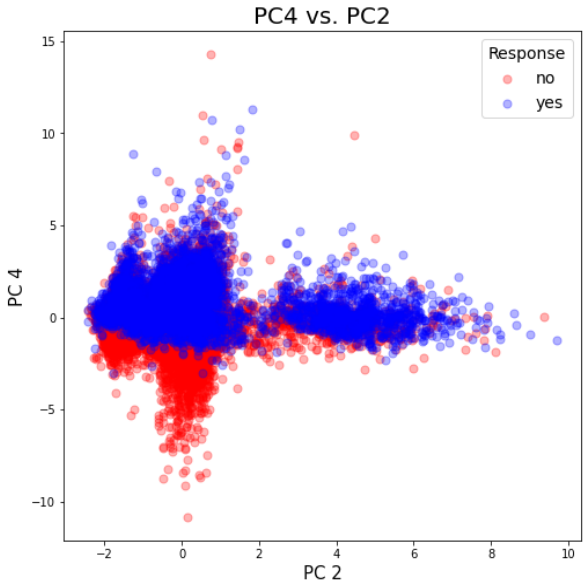
To accomplish the EDA, we plotted each combination of the first four principal components with the response variable mapped to color. We then examined the plot for any noteworthy clusters and then interpreted these findings in light of the loading vectors. The plots are displayed below.

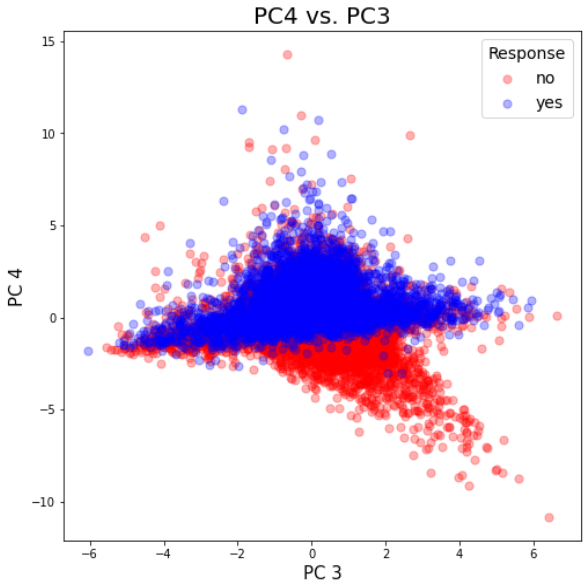












2.3. Interpretation

At a first glance, it is clear that there are groups of clients based on these principal components. It is easier to discern distinct groups of "no" clients (those who did not subscribe) than "yes" clients (those who did). Given the loading vectors of the first four principal components and the plots above, we can say the following from each plot.

PC2 vs. PC1: There is a cluster of distinct nos when PC1 is very low. This would be when there are high interest rates, high employment variation rates, and a high number of citizens are employed overall.

PC3 vs. PC1: There is a cluster of distinct nos when PC1 is low and PC3 is high. This indicates that young people are less likely to subscribe when the consumer confidence index is low, and interest rates, employment variation, and total employment are high.

PC4 vs. PC1: Across economic climates, but more so with high interest rates, employment variation, and total employment, lower durations of the last contact (shorter calls) and a higher total number of contacts with a client tend to result in nos.

PC3 vs. PC2: Cluster of nos when the number of days since last contact is higher, the number of previous contacts with a client is lower, and age is lower.

PC4 vs. PC2: Clear delineation of nos when number of days since last contact is higher, the number of contacts performed before the campaign is lower, consumer confidence is lower, and when the duration of the call is lower and there is a higher number of contacts.

PC4 vs. PC3: Clear delineation of nos when age is lower and duration of call is lower and the number of campaign contacts is higher.

Synthesizing these results, we answer our SMART question with the following points. When interest rates are high, employment variation rate is high, and total employment is high, it might not be a good time for a marketing campaign in that most clients will say no. Younger people are less likely to purchase when interest rates are high, employment variation is high, total employment is high, the duration of the call is low, and the number of contacts is high. A higher number of contacts when combined with short duration of call might predict non-purchase - these might be individuals who are turned off by excessive marketing. Therefore the best potential targets (from this PCA) are older individuals with more contact history with the bank, who stay on the phone longer.

Section 3. Feature Selection

3.1. Lasso Regression

Now that we have explored the numeric predictor variables, we then use two feature selection methods to answer SMART question 2: what factors are related to a customer subscribing to a term deposit? The first method to identify important features related to subscription that we used was Lasso regression. We used scikit-learn logistic regression with the

L1 penalizer (lasso). We then varied the regularization strength to observe which features were present and at what points they appeared in the model. The table below displays the presence of features under different regularization strengths, with strongest regularization on the left and weakest regularization on the right.

	0.0002	0.0010	0.0100	0.1000	1.0000
age	False	False	False	False	True
job	False	False	False	True	True
marital	False	False	False	False	False
education	False	False	True	True	True
default	False	False	True	True	True
housing	False	False	False	False	False
loan	False	False	False	False	False
contact	False	False	True	True	True
month	False	False	False	False	False
day_of_week	False	False	False	False	False
duration	True	True	True	True	True
campaign	False	False	False	False	False
pdays	False	True	True	True	True
previous	False	False	False	False	False
poutcome	False	False	False	True	True
emp_var_rate	False	False	True	True	True
cons_price_idx	False	False	False	False	False
cons_conf_idx	False	False	True	True	True
euribor3m	False	False	False	False	False
nr_employed	False	True	True	True	True

The results above show that duration of the call is the strongest factor in this model, followed by number of days since the last contact and total employment, then followed by education level, number of contacts in this campaign, employment variation rate, and consumer confidence index. The strength of duration, number of days since last contact, number of contacts within the campaign, employment variation rate, and consumer confidence validate our findings in PCA. Further, we note the strength of education level as a categorical variable through this approach.

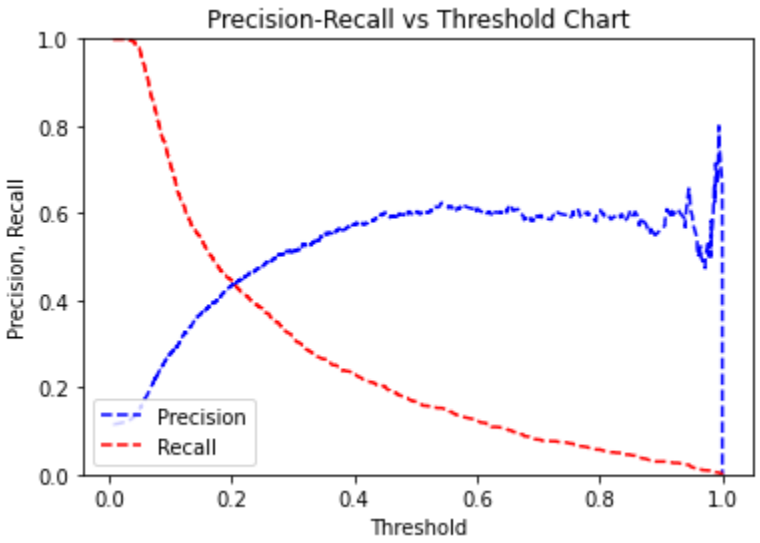
3.2. Stepwise Selection, Logistic Regression

The next step is to use backward stepwise reduction to fit the model to the data frame bank_df, using the glm() function. We can see that age, duration and number of contacts performed during this campaign (campaign) are statistically significant predictors ($p < 0.05$). All significant predictors have a p-value of 0.000, so we are unable to distinguish their association with the outcome y. Age and duration have a positive coefficient, which indicates that successfully subscribing to a term deposit is associated with older clients and longer contact. Campaign has a negative coefficient, which indicates that success in outcome is associated with fewer times of campaign. When we fit the data frame to the model, we find the accuracy score to be 0.89329.

We will split the data into training and testing for the remaining part of this section. Let the size of the training set be 0.8 and the size of the test set be 0.2. When we test accuracy on both sets, the score of the training set is 0.89381 and the score of the test set is 0.89196. There is not much difference between the two scores, which means our model is not overfitting.

Optimal cutoff

Now we would like to compute the optimal cutoff of our logistic regression model using the test set. First we want to look at the plot of Precision-Recall at different thresholds.



The intercept of the precision rate and the recall rate is where the optimal cutoff lies. Here the plot suggests that the optimal cutoff is around 0.21. At this cutoff, the area-under-curve of the ROC curve is 0.816, higher than the standard 0.8. The f1-score is 0.89. Since our dataset is extremely unbalanced, we may look at the f1-score of y=0 and y=1 separately. The f1-score of y=0 is 0.94, which is considerably high, whereas the f1-score of y=1 is only 0.26.

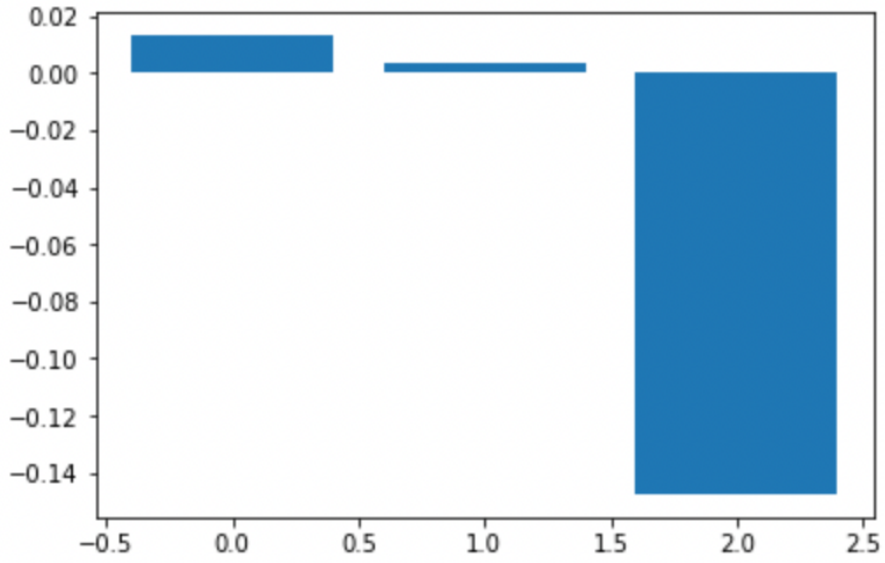
Let's define a function that computes the optimal cutoff. The optimal cutoff would be where the TP rate is high and the FP rate is low. TP rate - (1- FP rate) is 0 or close to 0 at the optimal cutoff. We run through all the cutoff points and find that the optimal cutoff is at 0.09539. At this cutoff the area-under-curve of the ROC curve is 0.817 and the f1-score is 0.74. Notice that although the overall f1-score is lower than in the previous computation, the f1-score of y=0 is 0.83, which is still high, and the f1-score of y=1 is 0.40, much higher than before.

Interpretation

First we want to look at the deviance of our model. The null deviance is 28998.72 and the residual deviance is 23829.01. Adding only three predictors to the null model, we manage to

reduce the deviance significantly, by 5169.71. Next we will analyze the feature importance in our model.

```
age score: 0.01301
duration score: 0.00370
campaign score: -0.14747
```



Notice that the coefficients are both positive and negative. The positive scores indicate a feature that predicts class 1, whereas the negative scores indicate a feature that predicts class 0. The score of age is higher than that of duration, which means that the age of our client is most related to the success in outcome, duration being the second. Campaign has the highest negative score, which means it is highly associated with failure in outcome.

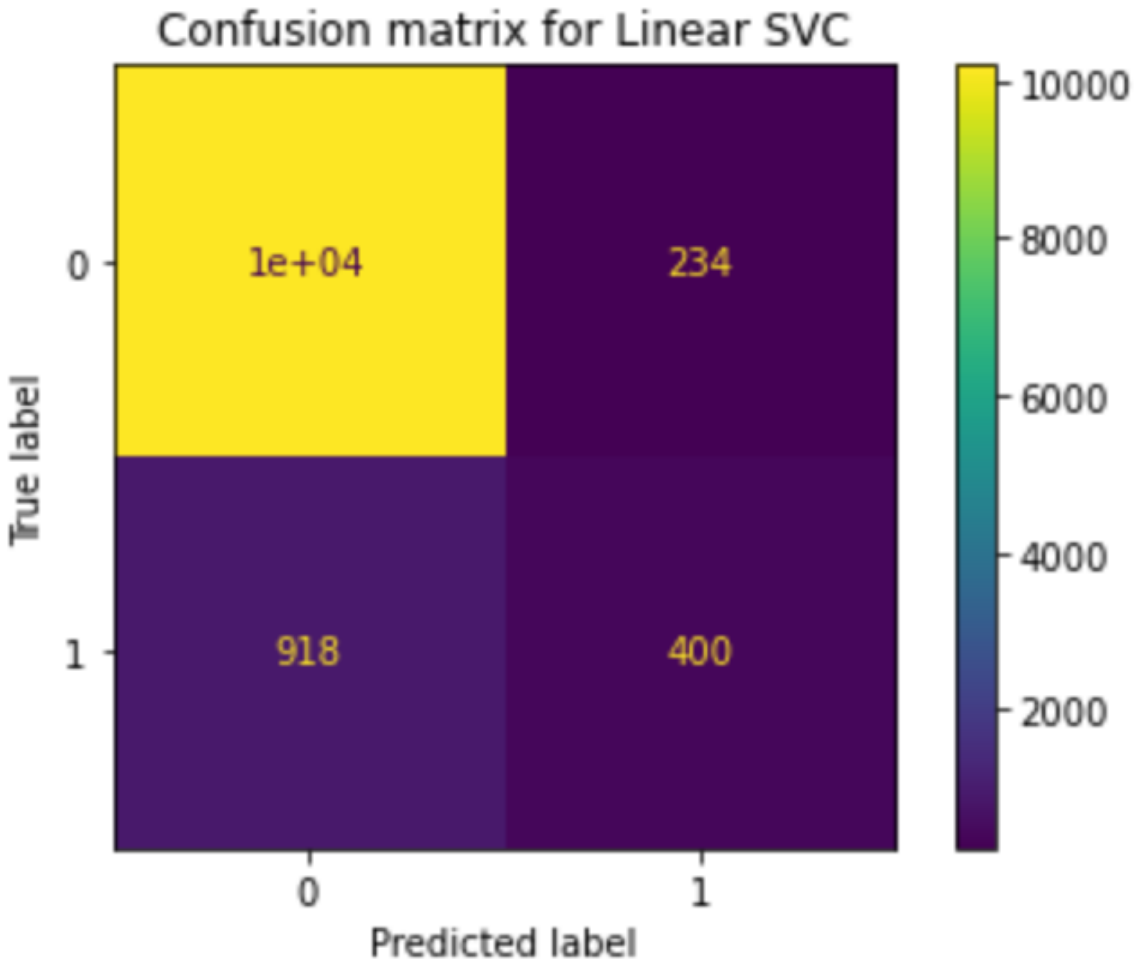
Finally, we want to look at the exponential of coefficients. This fitted model says that holding duration and campaign fixed, with every unit increase in age we see a 1.14% increase in the odds of success in outcome. Holding age and campaign fixed, with every unit increase in duration we see a 0.37% increase in the probability of success in outcome. Holding age and

duration fixed, with every unit increase in campaign, the probability of success in outcome is 13.30% lower.

Section 4. Prediction with Classification Models

4.1. Linear SVC

We first fitted a SVM model with a 'linear' kernel, using 70% observations from the dataset for training. As we printed out the confusion matrix, we can see that our number of true positives is 400.



Our test accuracy turns out to be 0.902, and by the classification report, our precision for 1 is 0.63 meaning that 63% of the observations we predict to subscribe actually subscribe to the term deposit.

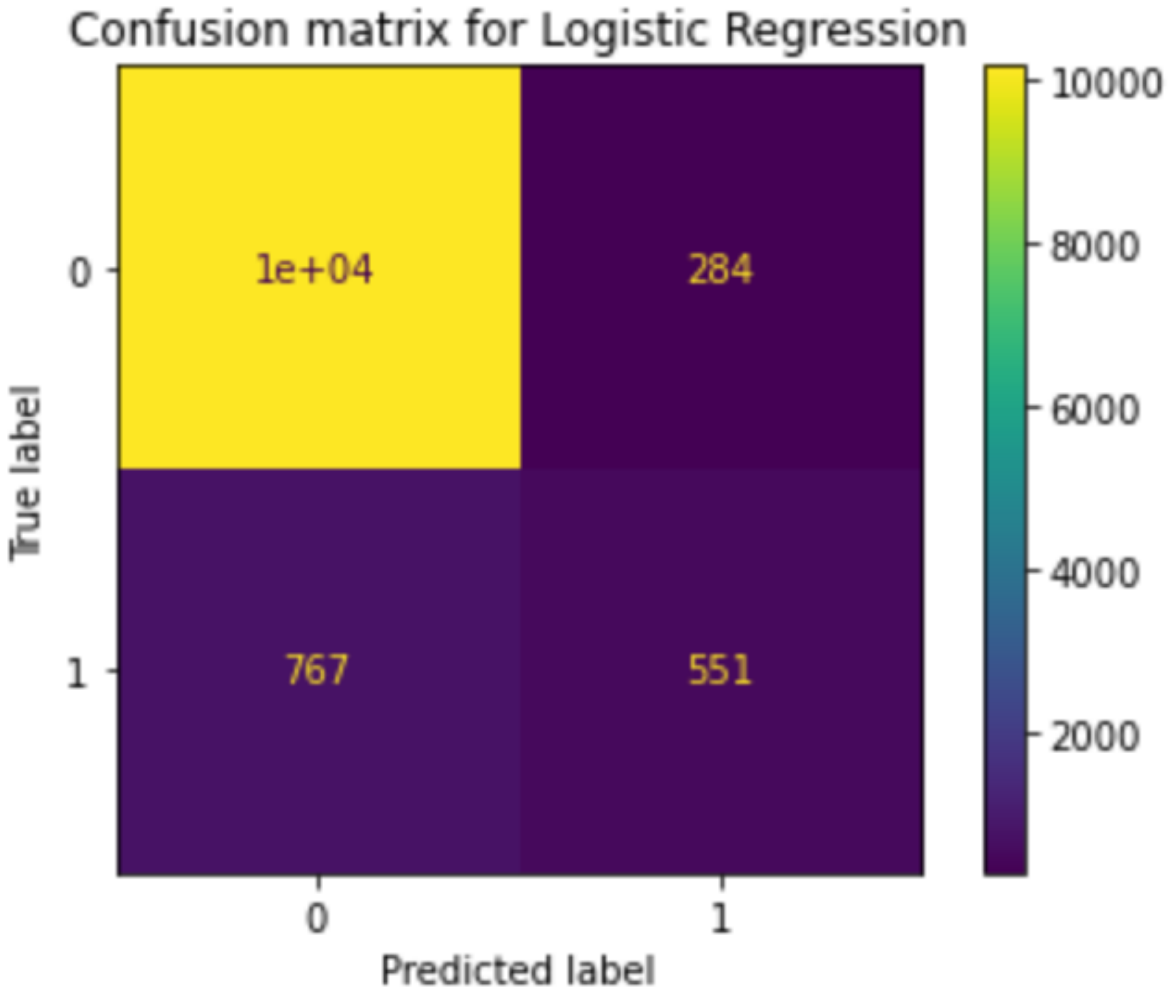
	precision	recall	f1-score	support
0	0.92	0.98	0.95	10440
1	0.63	0.30	0.41	1318
accuracy			0.90	11758
macro avg	0.77	0.64	0.68	11758
weighted avg	0.89	0.90	0.89	11758

We've also printed out the coefficients to see what factors impact our prediction the most. The list shows that 'duration' has the strongest positive effect on predicting subscription, while 'employment variation rate' has the strongest negative effect. This pretty much agrees with what we've found in our second smart question.

	coefficients				
age	-0.004695	emp.var.rate	-0.610600	job_wn	-0.030298
education	0.015296	cons.price.idx	0.321203	job_sn	-0.029462
default	-0.169616	cons.conf.idx	0.011051	m_sp	0.233445
housing	-0.018249	euribor3m	0.047355	m_wp	0.022272
loan	-0.029150	nr.employed	0.097148	m_sn	-0.208115
contact	-0.193202	single	0.005527	m_wn	-0.047602
duration	0.423784	married	-0.008406	d_sp	0.033668
campaign	-0.044511	divorced	0.002879	d_wp	0.022645
pdays	-0.267572	job_sp	0.046819	d_sn	-0.051882
previous	0.003265	job_w	0.012941	d_wn	-0.004430
poutcome	0.161439				

4.2. Logistic Regression

The second model we fitted is a Logistic Regression, with a maximum iteration of 10000. Using the same training and testing set, we get a confusion matrix displayed below. The number of True positive results is 551.



The logistic regression model gives us an accuracy of 0.91 and a precision rate of 0.66. This model has the highest accuracy among all, however, we can see that the precision for predicting ‘yes’ for subscription is still not as high as predicting ‘no’.

	precision	recall	f1-score	support
0	0.93	0.97	0.95	10440
1	0.66	0.42	0.51	1318
accuracy			0.91	11758
macro avg	0.80	0.69	0.73	11758
weighted avg	0.90	0.91	0.90	11758

Last but not least, by looking at the coefficients, the logistic regression model also has ‘duration’ as the highest positive coefficient, ‘employment variation rate’ as the highest negative coefficient, which also indicates the same important features with SVC.

	coefficients				
age	-0.005066	emp.var.rate	-2.586026	job_wn	-0.092394
education	0.061042	cons.price.idx	1.137217	job_sn	-0.158708
default	-0.517979	cons.conf.idx	-0.023361	m_sp	0.882326
housing	0.002219	euribor3m	0.578949	m_wp	0.018684
loan	-0.082466	nr.employed	0.227929	m_sn	-0.746596
contact	-0.552836	single	0.057509	m_wn	-0.109862
duration	1.224864	married	-0.029882	d_sp	0.094957
campaign	-0.156995	divorced	0.016924	d_wp	0.058454
pdays	-0.210197	job_sp	0.241774	d_sn	-0.120836
previous	-0.042988	job_w	0.053880	d_wn	0.011977
poutcome	0.400972				

4.3.. K-Nearest Neighbors

The last model we did was KNN. We first tuned the parameters in order to get the best k value. Since our data has over 40 thousands rows, we decided to start with iterating through a

large range for k. We calculate train accuracy from k=1 to k=451. As we know, small k values might overfit the data and large k might underfit the data, so there appeared a pattern for test accuracies. It increases from k=1 and reaches the peak at k=101, then keeps decreasing all the way to k=451 as shown below.

```
The train accuracy for 1th nearest neighbor is: 0.8798265011056302
The train accuracy for 51th nearest neighbor is: 0.9050008504847763
The train accuracy for 101th nearest neighbor is: 0.9059363837387311
The train accuracy for 151th nearest neighbor is: 0.9050858989624085
The train accuracy for 201th nearest neighbor is: 0.9044055111413506
The train accuracy for 251th nearest neighbor is: 0.9042354141860861
The train accuracy for 301th nearest neighbor is: 0.9028746385439701
The train accuracy for 351th nearest neighbor is: 0.9007484266031638
The train accuracy for 401th nearest neighbor is: 0.9004932811702671
The train accuracy for 451th nearest neighbor is: 0.899812893349209
```

Finally, we fitted a knn model with our optimized k value of 100, getting a test accuracy of 0.906 and a precision of 0.67. In this case, our KNN model has the highest precision.

	precision	recall	f1-score	support
0	0.92	0.98	0.95	10440
1	0.67	0.31	0.43	1318
accuracy			0.91	11758
macro avg	0.80	0.65	0.69	11758
weighted avg	0.89	0.91	0.89	11758

Section 5. Model Comparison

Comparing the results of our three machine learning models answers SMART question 4: which machine learning model best predicts customer behavior? To evaluate our models, we used two measures: accuracy and precision. We decided to use precision because it is likely that the bank would only want to spend marketing money on people who are the most likely to purchase. This represents the highest return on investment. Precision answers this question as it measures the percentage of people who we predicted would subscribe who then go on to subscribe. A comparison of the three models is displayed below.

Model	**Test Accuracy**	**Test Precision**
SVM (kernel = “linear”)	0.902	0.63
Logistic Regression	**0.911**	0.66
KNN	0.9026	**0.67**

As we can see from the table, logistic regression had the highest accuracy on our test data, while the KNN model had the highest precision on our test data. Because the precision scores are so close for the logistic regression and KNN models, we would recommend the logistic regression model to the bank because it is simpler, more flexible, and has a higher accuracy with a comparable precision to the KNN model.

Section 6. Conclusion

By utilizing a few machine learning methods (both unsupervised and supervised), our team was able to extract a number of insights from this dataset that can inform future direct marketing campaigns for banking institutions. Our unsupervised principal components analysis

served as a good EDA technique for this dataset with many predictors and revealed clusters of interest for specific economic conditions, customer demographics (age), and marketing techniques. Our use of feature selection revealed the most important explanatory variables and confirmed the insights produced from PCA. And of our predictive models, logistic regression performed the best on test data, with 91.1% accuracy and 66% precision. These insights would lead us to recommend banking institutions focus their direct marketing efforts on older people with established relationships who stay on the phone, and to avoid marketing in unstable employment environments.

References

Johnson, G. (2020). *Is Your Marketing Strategy Based on the Right Data?* Harvard Business Review. Retrieved December 7, 2021, from https://hbr.org/2020/05/is-your-marketing-strategy-based-on-the-right-data

ruthgn. (2021). *Bank Marketing Data Set*. Kaggle. Retrieved November 11, 2021, from https://www.kaggle.com/ruthgn/bank-marketing-data-set