Project Report:

Bank Marketing Analysis

Dillan Williams

Gabriel Medeiros

Abstract

Companies all around the world collect a lot of data related to how they interact with customers. The better a company analyzes data, the better it will perform in attracting customers through marketing campaigns. This project will be focused using Python to analyze the bank dataset, where the group members will try to predict what are the characteristics of the customers that ended up subscribing to a term deposit. This analysis would allow this bank to figure out what campaigns and what kind of clients ended up generating revenue for the bank, making it possible to create a more focused marketing strategy.

Contents

1.Our Question

We are wanting to use Python to analyze the bank dataset, where we will try to predict What are the characteristics of the customers that subscribed? Which campaign was more successful? Which type of modeling would be more efficient? What is the probability of the most common type of customer to subscribe? And we are wanting to see if there is a relationship between customer demographics (predicting variables) and deposit (target variable)?

2.Dataset dimensionality

The dataset that will be used for this project can be found on the UCI Machine Learning Repository website **https://archive.ics.uci.edu/ml/datasets/Bank+Marketing** and the file has 11,162 rows and 17 columns. The column the was labeled Balance was not given a description of what it is about, I contain positive and negative numbers but because we did not know what is was about we decided to remove it.

3.Dataset description

3.1 Rows description

In the "bank.csv" dataset, each row represents one client. With there being 11,162 rows, there are 11,161 clients in this data set.

3.2 Columns Description

Column Title	Description
Age	Age of the individual (numeric)
Job	Type of job (categorical)
Marital	Marital status (categorical)
Education	Education level (categorical)

Default	Has credit in default? (categorical)
Housing	Has housing loan? (Categorical)
Loan	Has personal loan? (Categorical)
Contact	Contact communication type (Categorical)
Day	Last contact day of the week (Each number represents the first, second… day of the week, and goes from Monday to Friday)
Month	Last contact month of the year (Categorical)
Duration	Last contact duration, in seconds (Numerical)
Campaign	Number of contacts performed during this campaign and for this client (Numerical)
Pdays	Number of days that passed by after the client was last contacted from a previous campaign
Previous	Number of contacts performed before this campaign and for this client
P-outcome	Outcome of the previous marketing campaign
Deposit	Whether the client subscribed or not to a term deposit (Categorical and Binary: Yes, No)

After a close look at the data set you can see that there are no missing values in the dataset.

4. Exploratory analysis

For this project we will be using the bank.csv data set to predict if the customers are likely to subscribe to do a deposit. We will be using Churn, Clustering, Random Forest, SVM and Logistic regression models to help predict whether a customer will subscribe to do a deposit or not. We are wanting to know if age, duration, education, and their marital status has anything to do with customers making deposits.

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-2.0744	1.04e+06	-2e-06	1.000	-2.03e+06	2.03e+06
job[T.blue-collar]	-0.5534	0.092	-6.014	0.000	-0.734	-0.373
job[T.entrepreneur]	-0.7764	0.159	-4.870	0.000	-1.089	-0.464
job[T.housemaid]	-0.5848	0.168	-3.481	0.001	-0.914	-0.255
job[T.management]	-0.3596	0.094	-3.838	0.000	-0.543	-0.176
job[T.retired]	0.3462	0.128	2.707	0.007	0.096	0.597
job[T.self-employed]	-0.5863	0.142	-4.125	0.000	-0.865	-0.308
job[T.services]	-0.4729	0.106	-4.471	0.000	-0.680	-0.266
job[T.student]	0.7547	0.154	4.895	0.000	0.453	1.057
job[T.technician]	-0.3246	0.087	-3.739	0.000	-0.495	-0.154
job[T.unemployed]	-0.1694	0.145	-1.167	0.243	-0.454	0.115
job[T.unknown]	-0.5648	0.288	-1.960	0.050	-1.130	2.05e-05
marital[T.married]	-0.0648	0.076	-0.852	0.394	-0.214	0.084
marital[T.single]	0.2747	0.088	3.133	0.002	0.103	0.446
education[T.secondary]	0.4076	0.082	4.946	0.000	0.246	0.569
education[T.tertiary]	0.7955	0.097	8.220	0.000	0.606	0.985
education[T.unknown]	0.4171	0.132	3.151	0.002	0.158	0.677
cluster_1[T.True]	-0.7133	1.04e+06	-6.87e-07	1.000	-2.03e+06	2.03e+06
cluster_2[T.True]	0.3798	1.04e+06	3.66e-07	1.000	-2.03e+06	2.03e+06
cluster_3[T.True]	-1.0005	1.04e+06	-9.64e-07	1.000	-2.03e+06	2.03e+06
cluster_4[T.True]	-0.7400	1.04e+06	-7.13e-07	1.000	-2.03e+06	2.03e+06
duration	0.0048	0.000	43.252	0.000	0.005	0.005
age	0.0067	0.003	2.421	0.015	0.001	0.012

[One can see from this chart that there is a positive correlation between deposit, duration, and age]

5 Charts and Findings

5.1 Random Forest

By running codes on the model accuracy and recall for Random Forest, the accuracy was 0.7993 and the recall was 0.9215. Random Forest was the model with the better values for accuracy and recall.

5.1.2 Logistic Regression

By running codes on the model accuracy and recall for Logistic Regression, the accuracy was 0.7805 and the recall was 0.9215.

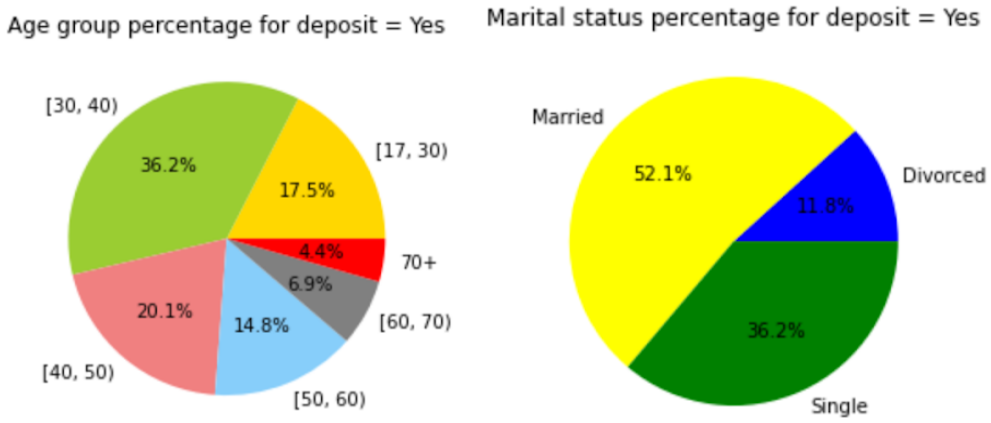
5.1.3 SVM

By running codes on the model accuracy and recall for SVM, the accuracy was 0.7507 and the recall was 0.9262.

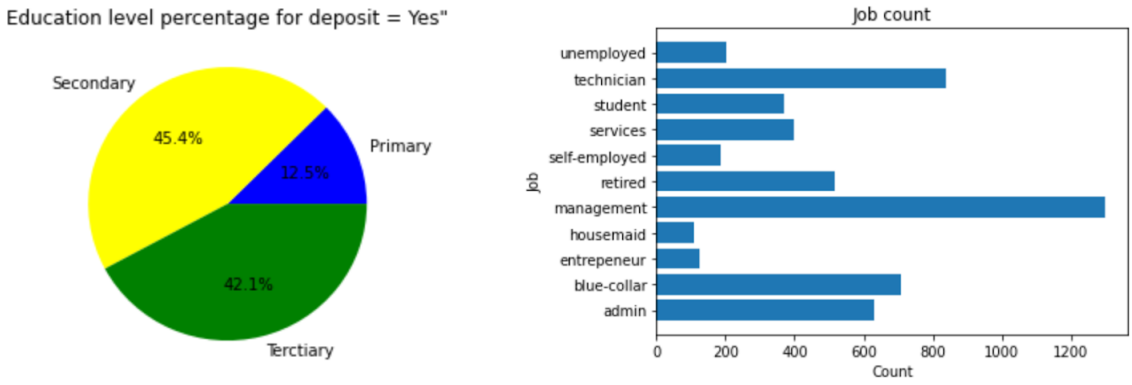
5.2 Marketing Campaigns

The bank ran a bunch of different marketing campaigns, but the campaign that ended up being successful was the campaign number 1, with 53.38% of the clients making the deposit.

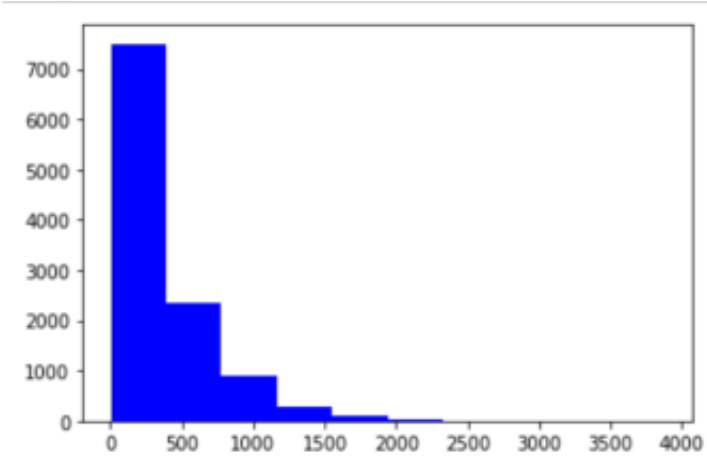
5.3 Customer Demographics and tendencies



[According to the pie charts above, most of the clients that ended up doing the deposit were married and were in the age group that goes from 30 to 40]

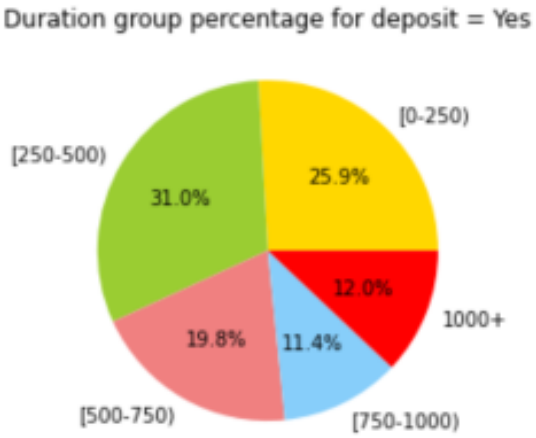


The pie chart shows us that 45.4% of the customers that ended up depositing have their highest education level as being secondary school. But close behind secondary schooling there is tertiary education. The bar chart shows us that most of the clients that deposited were managers, while housemaids was the group of customers with the least number of deposits.

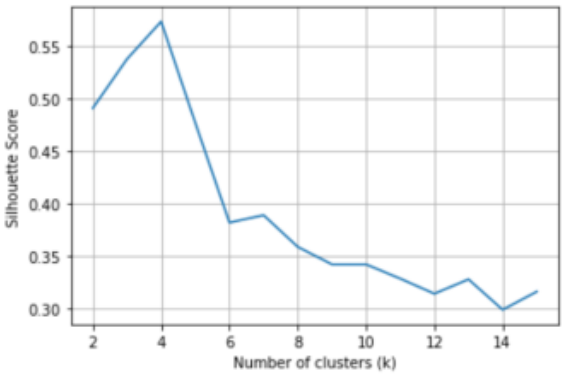


Based on this histogram, we were able to identify how long most of the calls lasted in seconds. Based in this graph, we can see that most of the calls lasted from 0 to 500 seconds. Also, there are some calls that lasted from 500 to 1000 seconds. Calls that lasted more than 1000 seconds will be considered outliers.

Based on the previous findings, we decided to distribute the calls duration in five groups. To create the pie chart on the right, we used the duration groups for calls that resulted in a subscription. We can see in the chart that the calls that converted the most lasted for about 250 to 500 seconds. The second group that converted the most was the group with calls that lasted 0 to 250 seconds.

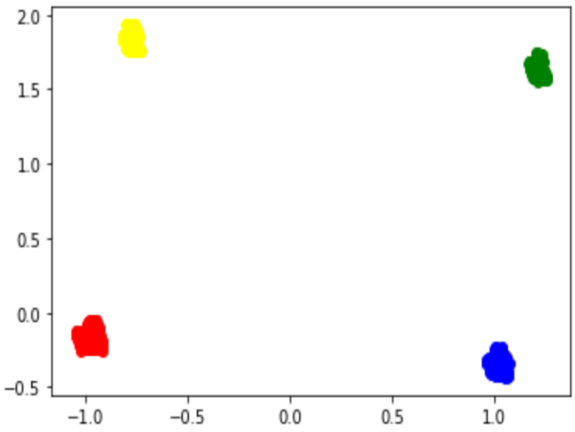


5.4 Clustering

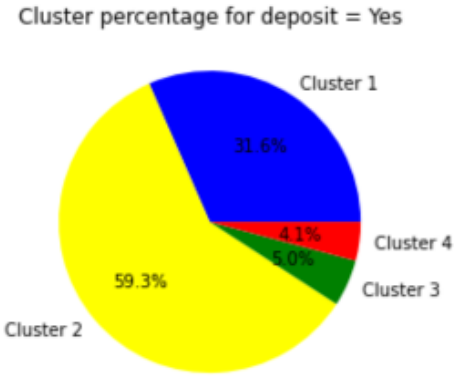


Before deciding how many clusters we should use, we decided to use the silhouette score to define the number of clusters that would fit the best for this dataset. According to the graph, the number of clusters with the highest score was ‘4’, which will be the number of clusters used in this project for this dataset.

When we ran the clustering code, we found that there were four distinct groupings. As we looked into further details about the clustering, we noticed that Cluster number 2 (Green) has the highest percentage of a customer subscribing with it being nearly 60% of the customers who did subscribe.



5.5 Predicting deposit



By using the following expression, we were able to find that if a customer is 35 years old, a manager, has completed secondary school, is married, had a 361 duration, and we used cluster 2 the probability of deposit is about 86%. Those characteristics of the customer were chosen based on the previous findings that showed us the characteristics of the clients that ended up making the deposit. The reason we took these numbers is because we looked at all of the highest variables.

marital	Deposit	
divorced	0	671
	1	622
married	0	3596
	1	2755
single	0	1606
	1	1912

Duration_group	Deposit	
1000+	0	71
	1	636
[0-250)	0	4088
	1	1369
[250-500)	0	1294
	1	1638
[500-750)	0	317
	1	1045
[750-1000)	0	103
	1	601

education	Deposit	
primary	0	909
	1	591
secondary	0	3026
	1	2450
tertiary	0	1693
	1	1996
unknown	0	245
	1	252

np.exp(-2.7171 + 0.0147*35 + -0.0061 + -0.2535 + 0.3973 + 0.0046*361) + 0.3798 / (1+ np.exp(-2.7171 + 0.0147*35 + -0.0061 + -0.2535 + 0.3973 + 0.0046*361 + 0.3798))= 86%

6 Research Plan

Date	Objects	Person(s) in charge
Nov 13	Project Proposal	Dillan, Gabriel
Nov 23	Project Feedback	Dillan, Gabriel
Dec 4	Project Report	Dillan, Gabriel
Dec 8	Project Presentation	Dillan, Gabriel

This research plan is to ensure that we are staying focused on the goal and not getting behind where we are supposed to be.

7 Future Plans

In the future we would like to look to see if certain months have an effect on what the best characteristics of the customers that subscribed is. For one of the columns where were not given the meaning of what Balance was meaning. We were unsure what the meaning of it was so we took it out of the data set. We would be hoping to try and find something to help us understand what the Balance section was about.

8 Conclusion

In Conclusion we used the Bank.csv data set and looked at the age, duration, education, and their marital status to determine who would most likely subscribe for a deposit. We found that most of the clients that ended up doing the deposit were married and were in the age group that goes from 30 to 40. We also found that it is most common for clients who are managers. When running the Random Forest, SVM and Logistic regression models we discovered that the Random Forest model was the best model because of the better values for accuracy and recall. With the help from the 11,161 clients in this data set we were able to find that customers 35-year-old, is a manager, completed secondary school, being in cluster 2, in a 316 seconds call, and are married had about a 86% chance to subscribe.

References

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing