Predicting Bank Customer Crunch Using Machine Learning

Coderhouse Project

Natanael Cobos

Introduction:

Welcome to this presentation on our data science project, focused on predicting the outcomes of a marketing campaign for a bank.

Our team has been tasked with analyzing a dataset provided by the bank, and developing a predictive model that will help them optimize their marketing strategy. In this presentation, we'll take you through our process, from data exploration to model development and evaluation, and share our findings and recommendations for the bank.

Let's get started!

Data Exploration:

We start with Data Exploration and Analysis:

The first step is explore the data and gain insights into its structure and contents.

In this project, we will be exploring and analyzing the dataset from a bank that contains information about its customers and their interactions with the bank.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	59	admin.	married	secondary	no	2343	yes	no	unknown	5	may	1042	1	-1	0	unknown	yes
1	56	admin.	married	secondary	no	45	no	no	unknown	5	may	1467	1	-1	0	unknown	yes
2	41	technician	married	secondary	no	1270	yes	no	unknown	5	may	1389	1	-1	0	unknown	yes
3	55	services	married	secondary	no	2476	yes	no	unknown	5	may	579	1	-1	0	unknown	yes
4	54	admin.	married	tertiary	no	184	no	no	unknown	5	may	673	2	-1	0	unknown	yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11157	33	blue-collar	single	primary	no	1	yes	no	cellular	20	apr	257	1	-1	0	unknown	no
11158	39	services	married	secondary	no	733	no	no	unknown	16	jun	83	4	-1	0	unknown	no
11159	32	technician	single	secondary	no	29	no	no	cellular	19	aug	156	2	-1	0	unknown	no
11160	43	technician	married	secondary	no	0	no	yes	cellular	8	may	9	2	172	5	failure	no
11161	34	technician	married	secondary	no	0	no	no	cellular	9	jul	628	1	-1	0	unknown	no

11162 rows × 17 columns

As we can see, this data set contains the results of their previous campaign for each client they contacted, 11162 clients we contacted and for each client are info of their age, job, marital situation, and if they deposit after the bank contact the,.

Data Exploration:

```
RangeIndex: 11162 entries, 0 to 11161
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   age        11162 non-null  int64 
 1   job        11162 non-null  object
 2   marital    11162 non-null  object
 3   education  11162 non-null  object
 4   default    11162 non-null  object
 5   balance    11162 non-null  int64 
 6   housing    11162 non-null  object
 7   loan       11162 non-null  object
 8   contact    11162 non-null  object
 9   day        11162 non-null  int64 
 10  month      11162 non-null  object
 11  duration   11162 non-null  int64 
 12  campaign   11162 non-null  int64 
 13  pdays      11162 non-null  int64 
 14  previous   11162 non-null  int64 
 15  poutcome   11162 non-null  object
 16  deposit    11162 non-null  object
dtypes: int64(7), object(10)
memory usage: 1.4+ MB
```

The first Insights of the data exploration is that we haven't missing values in the dataset.

Now we can proceed with Data Analysis.

Data Analysis:

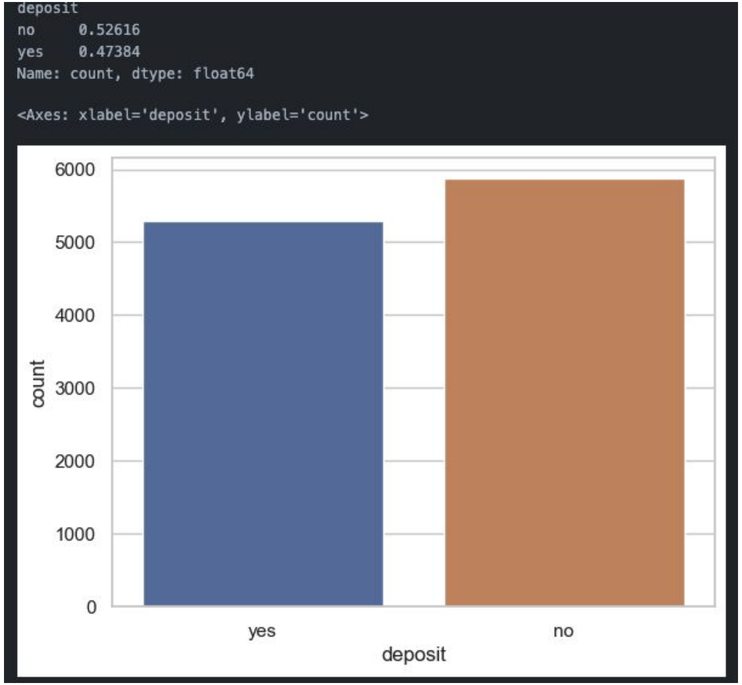
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
age	11162.0	NaN	NaN	NaN	41.231948	11.913369	18.0	32.0	39.0	49.0	95.0
job	11162	12	management	2566	NaN	NaN	NaN	NaN	NaN	NaN	NaN
marital	11162	3	married	6351	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education	11162	4	secondary	5476	NaN	NaN	NaN	NaN	NaN	NaN	NaN
default	11162	2	no	10994	NaN	NaN	NaN	NaN	NaN	NaN	NaN
balance	11162.0	NaN	NaN	NaN	1528.538524	3225.413326	-6847.0	122.0	550.0	1708.0	81204.0
housing	11162	2	no	5881	NaN	NaN	NaN	NaN	NaN	NaN	NaN
loan	11162	2	no	9702	NaN	NaN	NaN	NaN	NaN	NaN	NaN
contact	11162	3	cellular	8042	NaN	NaN	NaN	NaN	NaN	NaN	NaN
day	11162.0	NaN	NaN	NaN	15.658036	8.42074	1.0	8.0	15.0	22.0	31.0
month	11162	12	may	2824	NaN	NaN	NaN	NaN	NaN	NaN	NaN
duration	11162.0	NaN	NaN	NaN	371.993818	347.128386	2.0	138.0	255.0	496.0	3881.0
campaign	11162.0	NaN	NaN	NaN	2.508421	2.722077	1.0	1.0	2.0	3.0	63.0
pdays	11162.0	NaN	NaN	NaN	51.330407	108.758282	-1.0	-1.0	-1.0	20.75	854.0
previous	11162.0	NaN	NaN	NaN	0.832557	2.292007	0.0	0.0	0.0	1.0	58.0
poutcome	11162	4	unknown	8326	NaN	NaN	NaN	NaN	NaN	NaN	NaN
deposit	11162	2	no	5873	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We perform a Describe of the dataset and the first insights are:

There are 12 unique jobs

The minimum age of a client is 18 years old and the maximum age is 95 years old.

Data Analysis: Insights

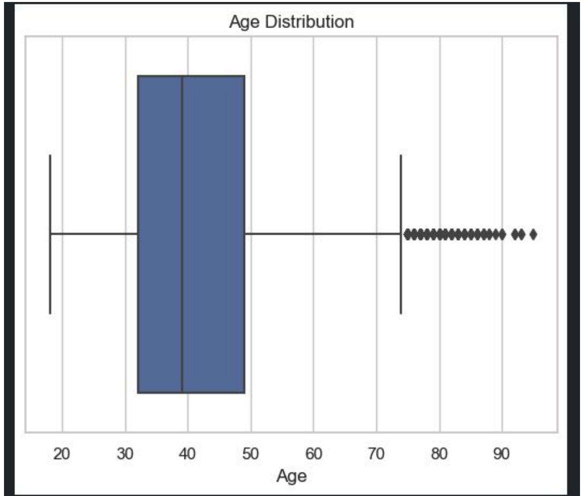


Insights:

Answering the first question: ¿What is the goal of the bank campaign, and how does it relate to the target variable "deposit"?

We can see that the goal of the bank campaign is to encourage clients to make a deposit, so the deposit variable is our target variable that we want to predict.

Data Analysis: Insights



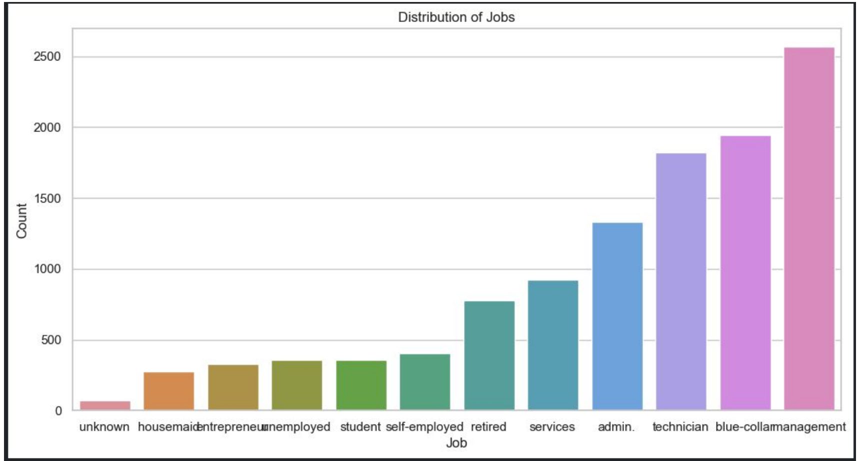

```
count    11162.000000
mean        41.231948
std         11.913369
min         18.000000
25%         32.000000
50%         39.000000
75%         49.000000
max         95.000000
```

Insights:

What are the demographic characteristics of the customers in the dataset (age, job, marital status, education)?

As we can see, the mean age of the customers is 41 years old, and the standard deviation is 11 years old.

Data Analysis: Insights

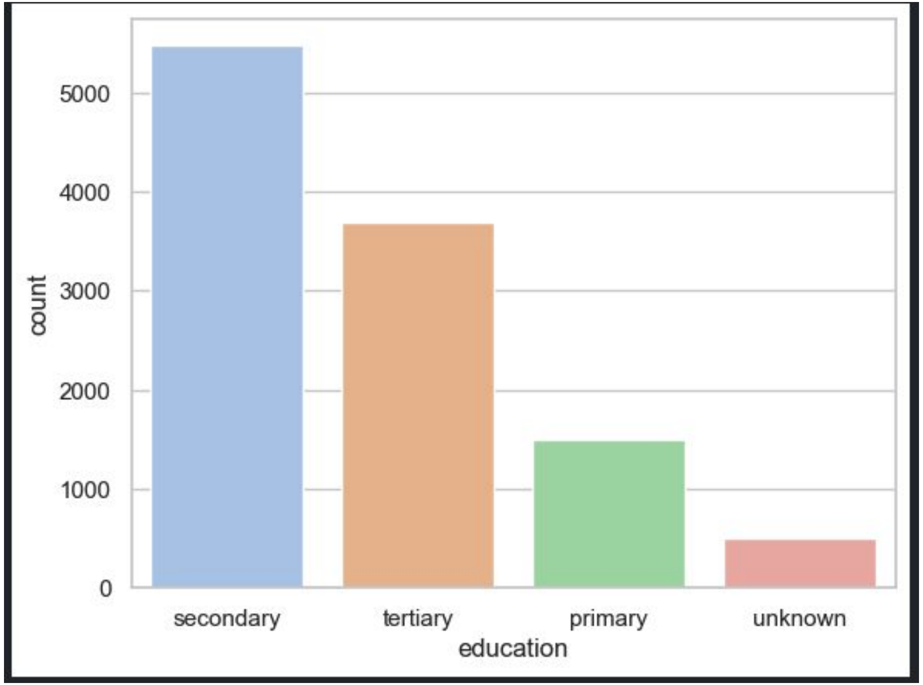


	Job	Value Counts
0	management	2566
1	blue-collar	1944
2	technician	1823
3	admin.	1334
4	services	923
5	retired	778
6	self-employed	405
7	student	360
8	unemployed	357
9	entrepreneur	328
10	housemaid	274
11	unknown	70

What are the demographic characteristics of the customers in the dataset (age, job, marital status, education)?

As we can see here, the most job categories with the highest number of people are Management and Blue-collar.

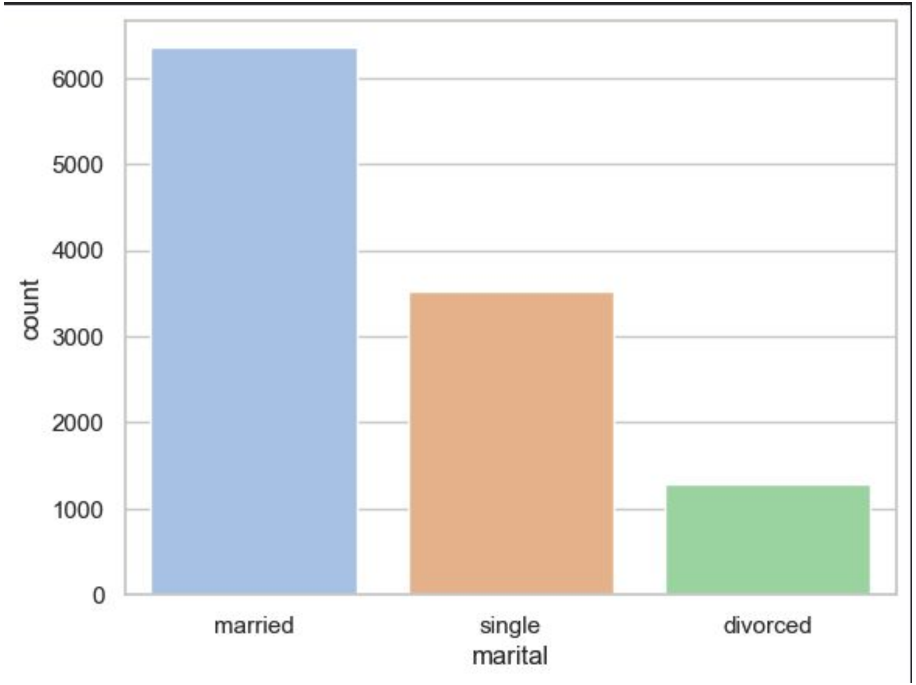
Data Analysis: Insights



What are the demographic characteristics of the customers in the dataset (age, job, marital status, education)?

The are more clients with secondary education than tertiary education.

Data Analysis: Insights



What are the demographic characteristics of the customers in the dataset (age, job, marital status, education)?

As we can see here, most of the clients are married, the second large group of clients are single.

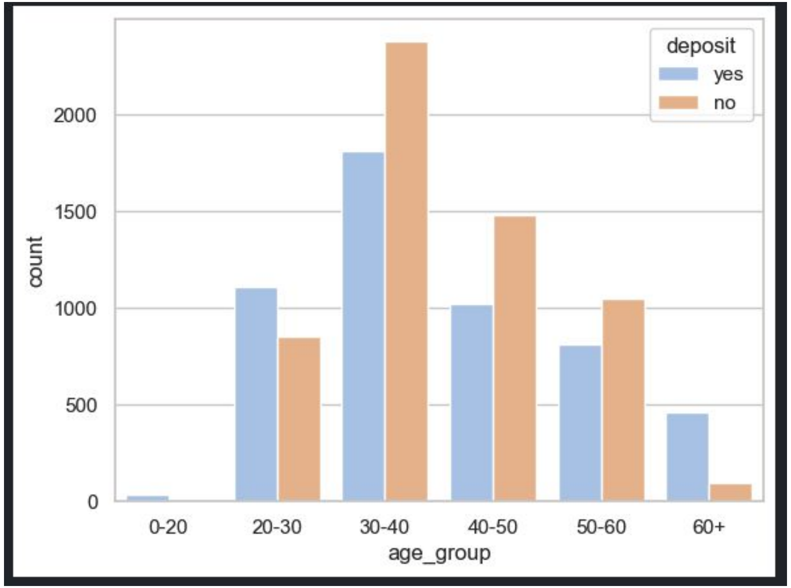
Feature Engineering: One-hot encoding

	feature	coefficient
0	age	-0.012261
1	balance	0.000024
2	day	0.007646
3	duration	0.003852
4	campaign	-0.222064
5	pdays	-0.000644
6	previous	0.189462
7	job_admin.	-0.009691
8	job_blue-collar	-0.191774
9	job_entrepreneur	-0.022918
10	job_housemaid	-0.006509
11	job_management	0.020018
12	job_retired	0.128265
13	job_self-employed	-0.015189
14	job_services	-0.066554
15	job_student	0.051823
16	job_technician	-0.032450
17	job_unemployed	0.007269
18	job_unknown	0.001450
19	marital_divorced	0.005930
20	marital_married	-0.152505
21	marital_single	0.010314
22	education_primary	-0.076193
23	education_secondary	-0.170108
...		
47	poutcome_failure	-0.095082
48	poutcome_other	-0.019692
49	poutcome_success	0.260099
50	poutcome_unknown	-0.281585

This is coefficients of the logistic regression model that was fit using all the variables in the data set except for the target variable deposit. The logistic regression model predicts the probability that the target variable is 1 (i.e., the customer makes a deposit), given the values of the input features. A positive coefficient for a feature indicates that an increase in the value of that feature is associated with an increased probability of the customer making a deposit, while a negative coefficient indicates that an increase in the value of that feature is associated with a decreased probability of the customer making a deposit.

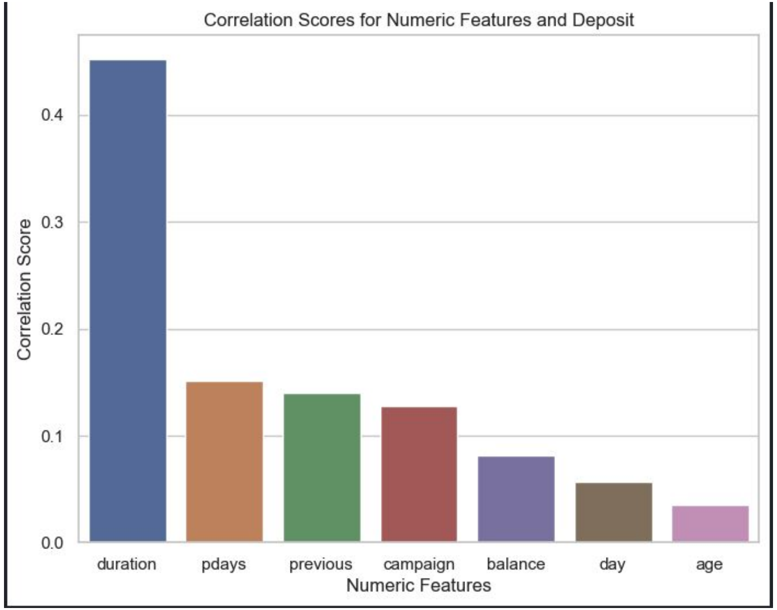
For example, we can see that the coefficient for balance is positive, which suggests that customers with higher account balances are more likely to make a deposit. Conversely, the coefficient for campaign is negative, which suggests that customers who have been contacted by the bank more times are less likely to make a deposit.

Feature Engineering: Binning



We've performed a feature selection called Binning, wich transform continuous numerical values into categorical features, then we've a graphic that represents the age group, as we can see here the more active clients are between the 20 and 60 years old.

Feature Engineering: Numeric Features



We have performed a feature selection on the numeric features, and based on the correlation betwen the numerical features and the target variable "deposit", the stronger variables are:

* Duration

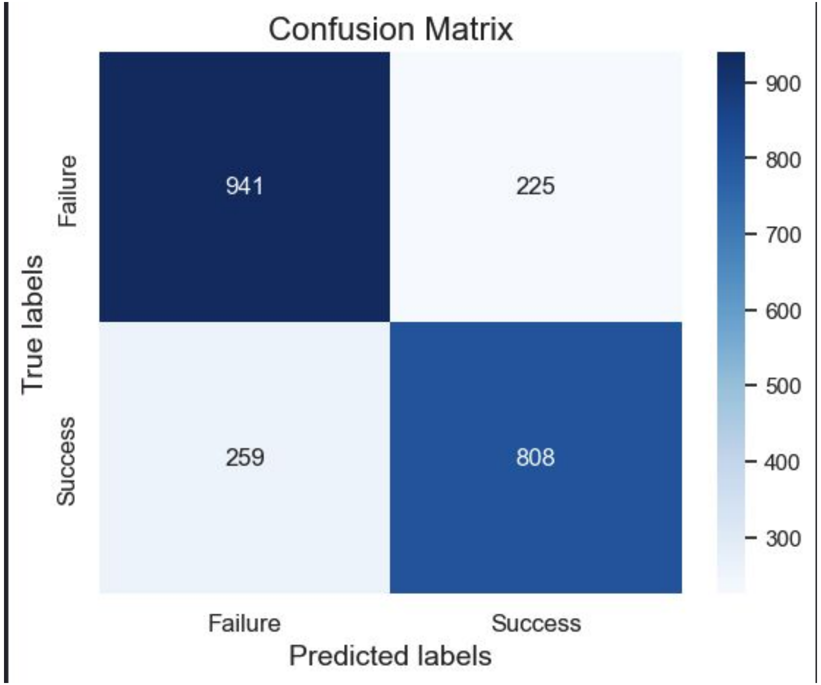
* pdays

* previous

* campaign

* balance

Modeling: Logistic Regression



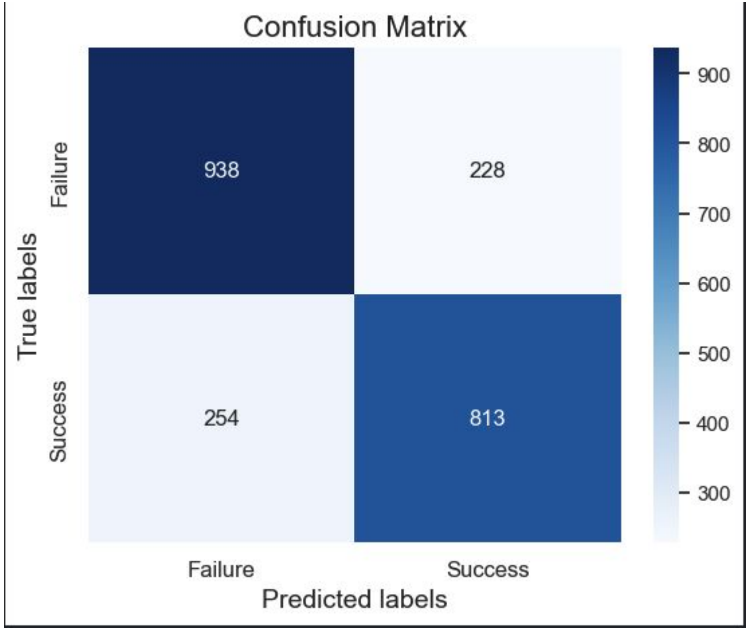
We've performed a logistic regression model and the model correctly predicted the deposit (yes or no) for 78%

Insights of Logistic regression:

As we see here, our precision for both 0 and 1 are 78%, but the recall is pretty low in 1 (76%) compared with 0 (81%).

in the other hand, our f1-score is 0.8 and 0.77 for both cases.

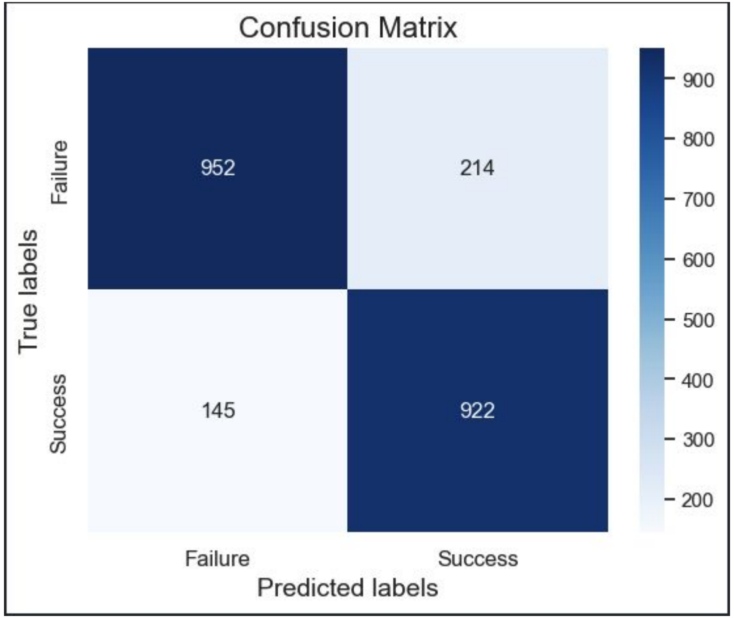
Modeling: Decision Tree Classifier



As we can see, with the decision tree model, the metrics are slightly similar to the logistic regression except for the f1 score and precision. these metrics tell us that the decision tree model is no a quite good fit.

- Accuracy: 0.784
- Precision: 0.781
- Recall: 0.762
- F1 score: 0.771

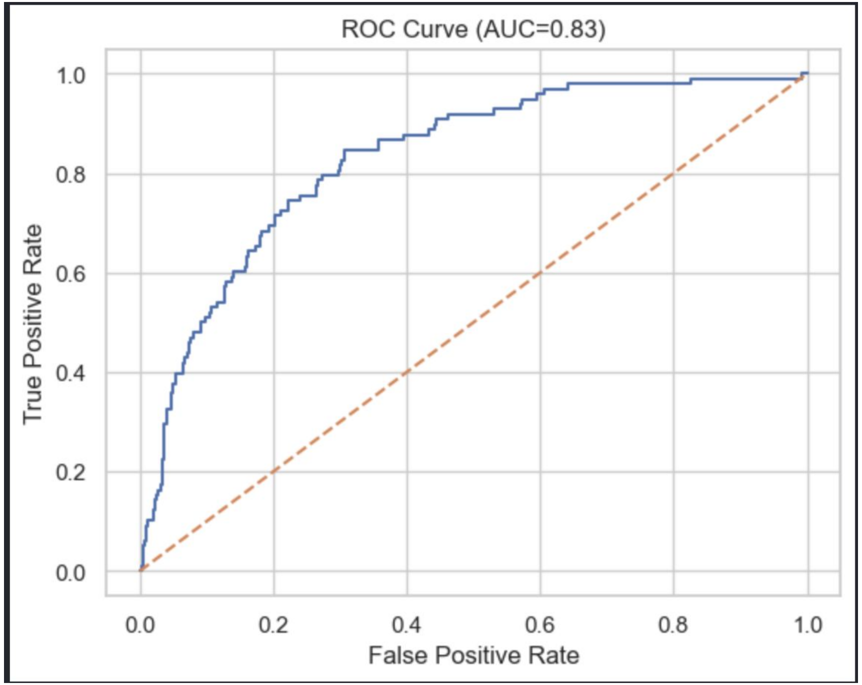
Modeling: Random Forest Classifier



As we can see, this models performs well and better that the logistic regression and decision tree classifier.

- Accuracy: 0.839
- Precision: 0.812
- Recall: 0.864
- F1 score: 0.837

Model Evaluation: Logistic Regression



As we can see here, the sentivite curve is a good indicator of how our model will predict the true cases.

Accuracy scores for each fold: [0.80861244 0.74162679 0.80769231 0.77403846 0.78846154]

Average accuracy score: 0.7840863084284138

As we can see, the average of the Logistic regression Cross Validation is 78 %, we will try with other model seeking for better results

Conclusion: Trained Models

As we can see here, we performed a **Desicion Tree Classifier** with the test data, due to the test data was imbalanced we perform some techniques to balance the data:

UNDERSAMPLING is the technique used with imbalance data, this technique remove samples from the majority class to create a balanced dataset.

UNDERSAMPLING RESULTS:

Accuracy: 0.770 Precision: 0.760 Recall: 0.760 F1 score: 0.760

These weren`t a good result so we tried another method to balance our dataset.

OVERSAMPLING: This is another technique used to balance an imbalanced dataset by increasing the number of samples in the minority class. One way to perform oversampling is to randomly duplicate samples from the minority class until the number of samples in both classes is equal.

OVERSAMPLING RESULTS:

Accuracy: 0.953 Precision: 0.918 Recall: 0.995 F1 score: 0.955

Accuracy Score: 0.953 suggests that the model is able to correctly predict the majority of the classes.

Precision Score: 0.918 suggests that out of all the predicted positive cases, 91.8% of them were actually positive.

Recall Score: 0.995 indicates that the model was able to correctly identify 99.5% of the actual positive cases.

F1 score: which is a combination of precision and recall, is 0.955

These are the best result so far, but to be sure about the model we've performed a cross-validation on the **Desicion Tree Classifier Model** and this are the result.

Accuracy: 0.957 (+ / - 0.005)

Accuracy score from Cross-Validation: 0.957 (+/- 0.005) suggests that the model is performing well and consistently across different folds of the data.

The "+/- 0.005" represents the standard deviation of the accuracy scores, which indicates how much the accuracy scores vary from the mean accuracy score.

A smaller standard deviation indicates that the accuracy scores are more consistent and reliable.

Conclusion:

Based on the results presented, the Decision Tree Classifier model with oversampling appears to be a promising approach for predicting outcomes in this scenario. With an accuracy score of 0.953, the model was able to correctly predict the majority of the classes, while the precision score of 0.918 indicates that out of all the predicted positive cases, 91.8% of them were actually positive. The recall score of 0.995 suggests that the model was able to correctly identify 99.5% of the actual positive cases, while the F1 score of 0.955, which is a combination of precision and recall, further supports the effectiveness of this model.

Furthermore, the cross-validation results showed that the Decision Tree Classifier model with oversampling was consistent across different folds of the data, with an accuracy score of 0.957 (+/- 0.005). This suggests that the model is performing well and can be relied upon to predict outcomes accurately.

In conclusion, the Decision Tree Classifier model with oversampling can be a valuable tool for predicting outcomes in this scenario, as it has demonstrated high accuracy, precision, recall, and F1 score, as well as consistency across different folds of the data.