INDIVIDUAL ASSIGNMENT
RESPONSE TEMPLATE

Please fill in your name and submission date below:

- Ckalib S. Nelson
- May 15th, 2021

Please fill-in your answers to your questions below. Feel free to incorporate any graphs or other information you may consider. I have included a recommended length of answer for orientation. Remember your answers to these questions will weight 70% of your final grade on the Individual Assignment.

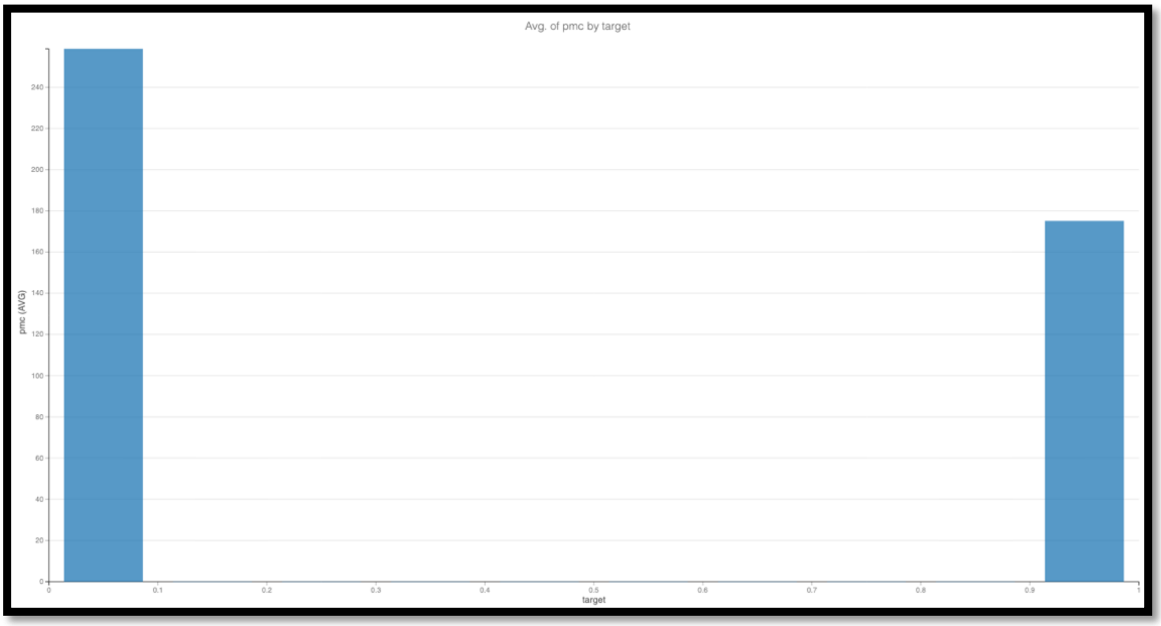
Table of Contents

QUESTION 1.

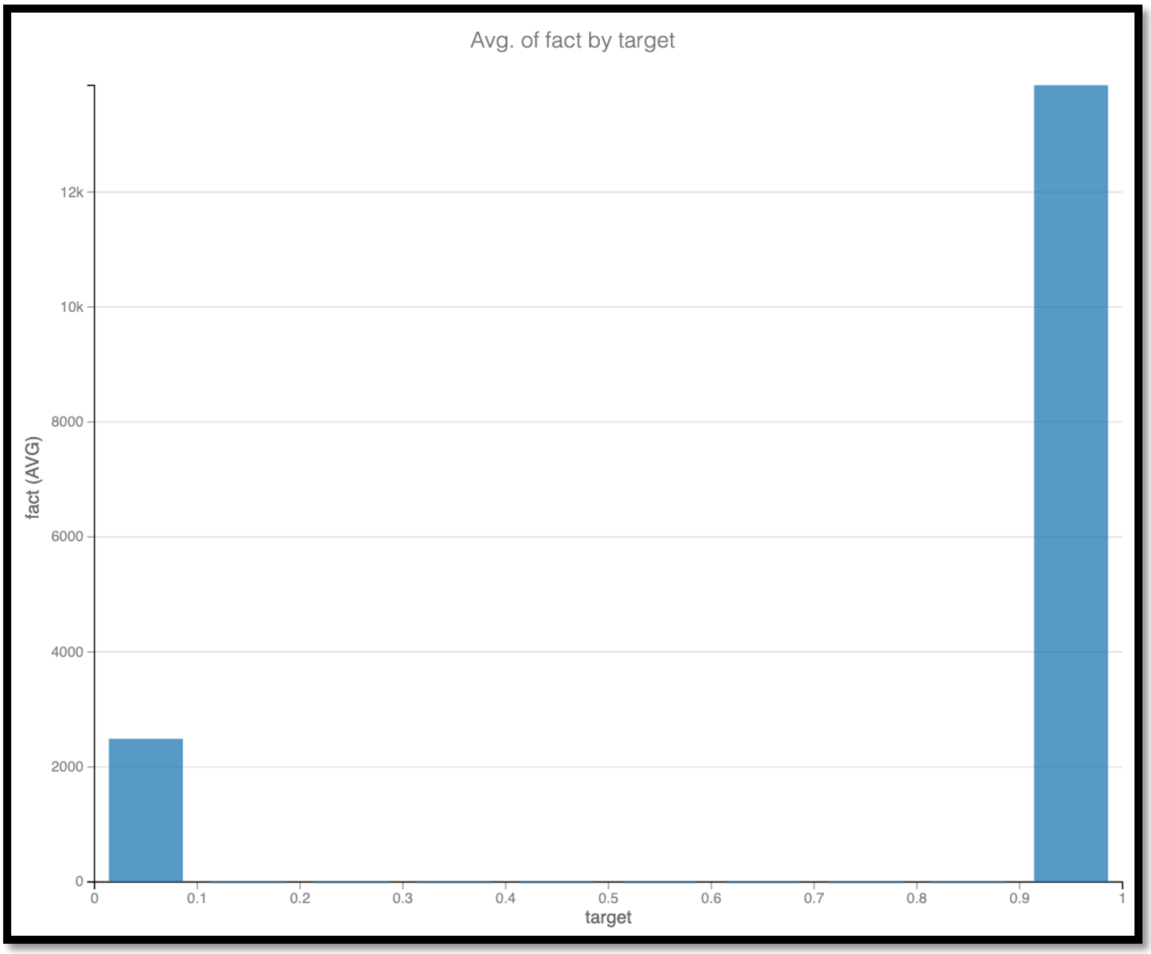
Understanding the Data [recommended max. 150 words]

How does the target variable behave when compared to the other variables? What sort of interactions do you see between the variables? What sort of correlation? Which are the most correlated variables? Do you see a business-sense?

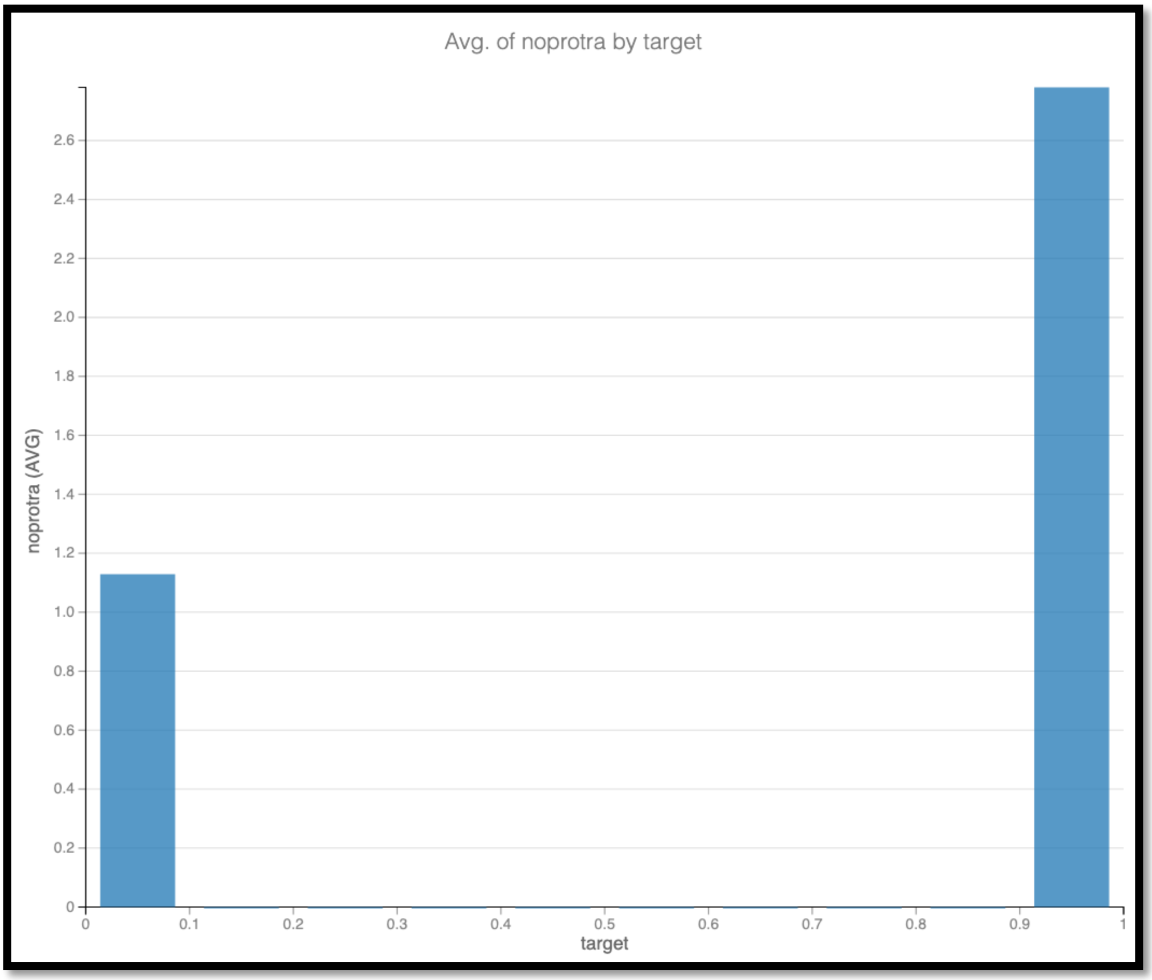
Here are a few visualizations that highlight the relationship between the target variable and the independent variables:



Per the chart above, illustrating the average collection period (**pmc**) by **target**, the higher the **pmc** the less likely the company will contract the POS in the following 3 months. Thus, there is a negative correlation between **pmc** and the **target**. In a business sense, this makes sense.



Per the chart above, illustrating the average amount of money the cards emitted by the bank spent in the POS owned by the company (**fact**) by **target**, the higher the fact the more likely the company will contract the POS in the following 3 months. Thus, there is a positive correlation between **fact** and the **target**. In a business sense, this makes sense.



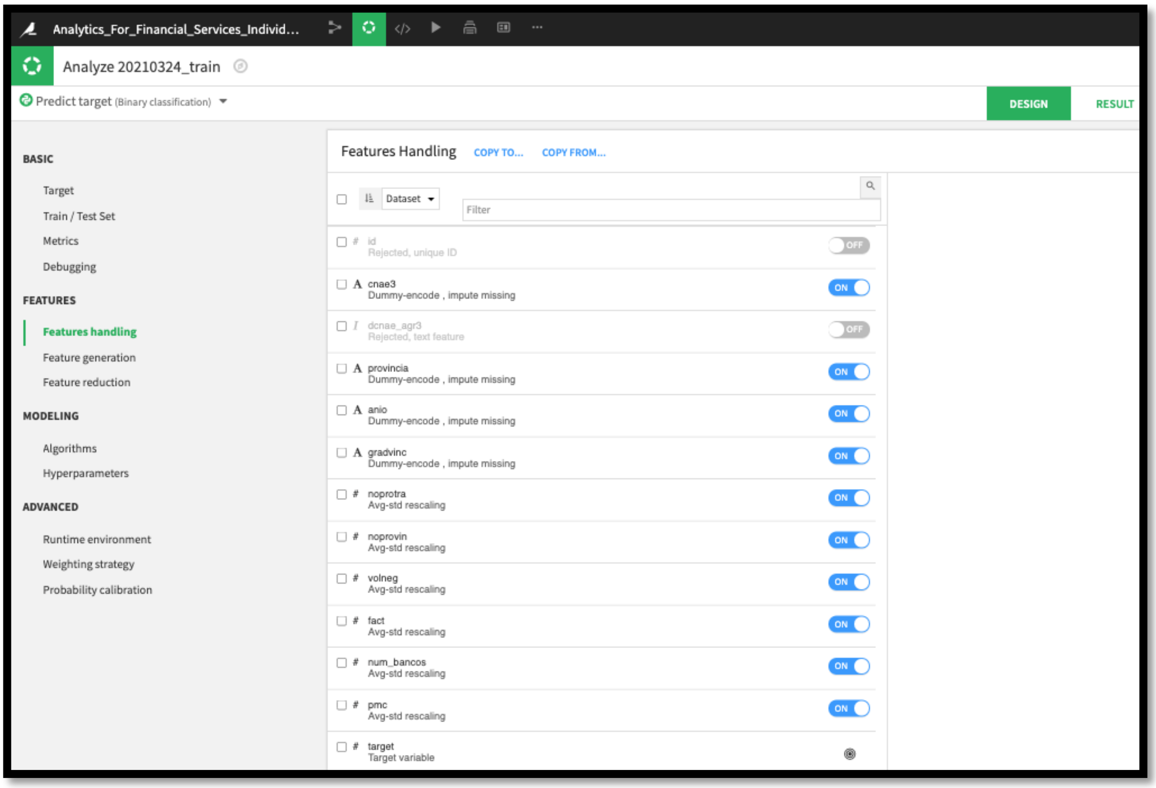
Per the chart above, illustrating the number of services the company has contracted with the bank (**noprotra**) by target, the higher the **noprotra** the more likely the company will contract the POS in the following 3 months. Thus, there is a positive correlation between **noprotra** and the **target**. In a business sense, this makes sense.

QUESTION 2.

Preparing the Data [recommended max. 100 words]

Would you recommend doing any sort of cleaning or transforming the variables before working on the model? (eg. missing values, outliers, etc…)

Yes, I would recommend doing a bit of cleaning and transforming of the variables before working on the model given that *a few columns have missing values and are not of the necessary data type*:



□ # id — Rejected, unique ID — OFF

□ A cnae3 — Dummy-encode , impute missing — ON

□ *I* dcnae_agr3 — Rejected, text feature — OFF

□ A provincia — Dummy-encode , impute missing — ON

□ A anio — Dummy-encode , impute missing — ON

□ A gradvinc — Dummy-encode , impute missing — ON

□ # noprotra — Avg-std rescaling — ON

□ # noprovin — Avg-std rescaling — ON

□ # volneg — Avg-std rescaling — ON

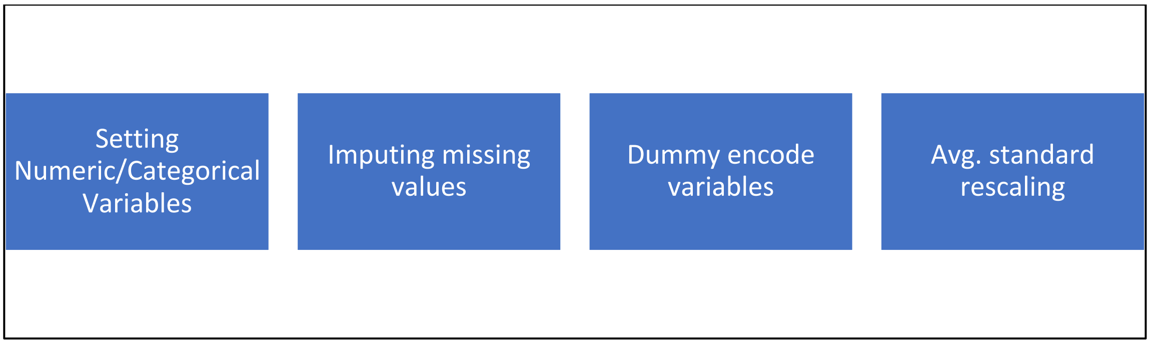
□ # fact — Avg-std rescaling — ON

□ # num_bancos — Avg-std rescaling — ON

□ # pmc — Avg-std rescaling — ON

□ # target — Target variable

The cleaning and transforming needed can be summarized as follows:



Setting Numeric/Categorical Variables

Imputing missing values

Dummy encode variables

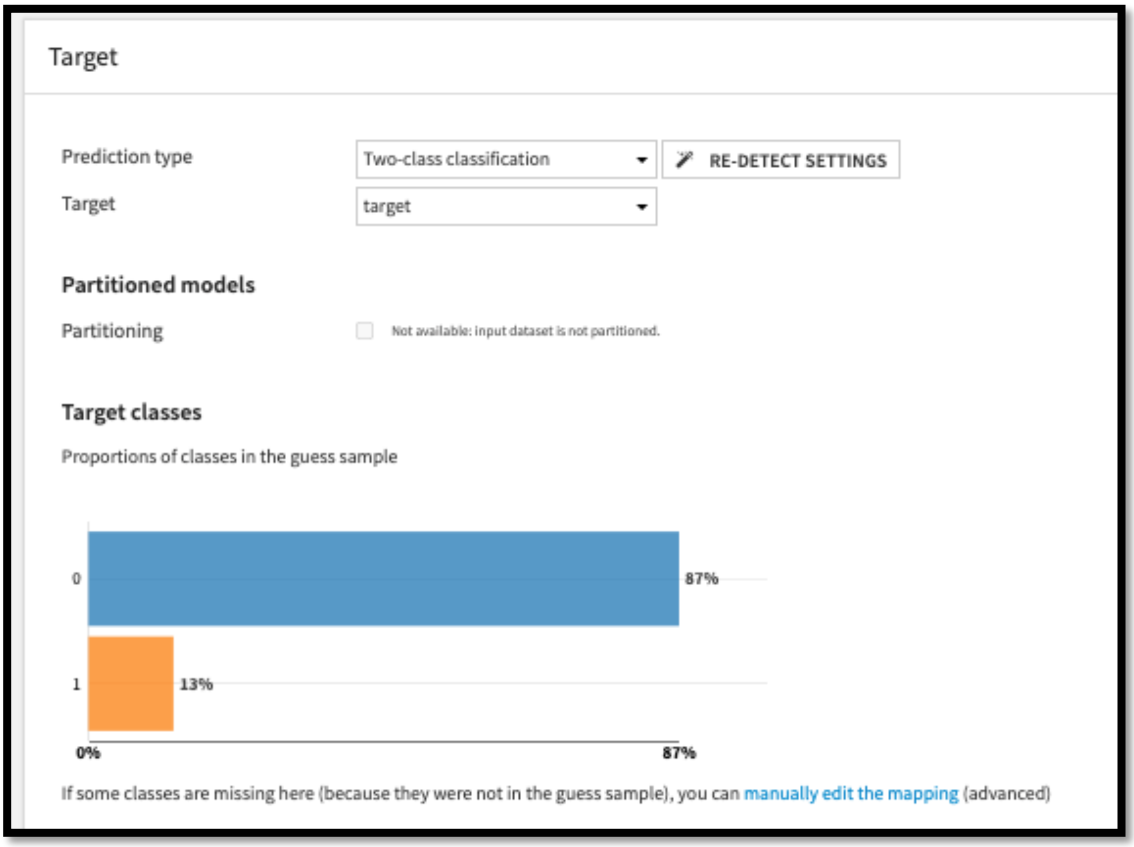
Avg. standard rescaling

QUESTION 3.

The Model [recommended max. 150 words]

What type of model have you used (ie. Supervised/unsupervised, binary...) and why?

Given that we do know the target variable, we will have to rely upon a supervised model. Furthermore, the target variable is binary; thus, we will have to leverage a supervised classification model.



Target

Prediction type: Two-class classification

Target: target

Partitioned models

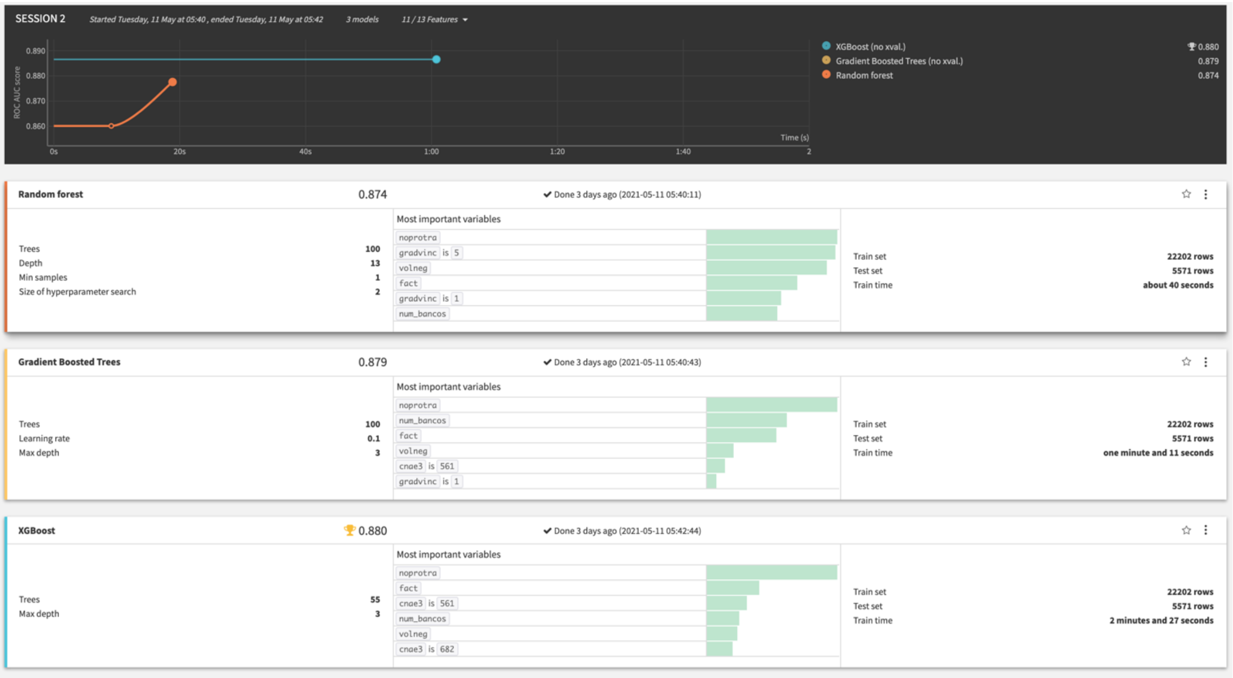
Partitioning: □ Not available: input dataset is not partitioned.

Target classes

Proportions of classes in the guess sample

If some classes are missing here (because they were not in the guess sample), you can manually edit the mapping (advanced)

After a bit of trial and error, I settled on the XGBoost algorithm, or the e**X**treme **G**radient **B**oosting algorithm. This particular algorithm performed better (per AUC) than its competitors:



SESSION 2 *Started Tuesday, 11 May at 05:40 , ended Tuesday, 11 May at 05:42* *3 models* *11 / 13 Features*

● XGBoost (no xval.)	🏆 0.880
● Gradient Boosted Trees (no xval.)	0.879
● Random forest	0.874

Random forest 0.874 ✔ Done 3 days ago (2021-05-11 05:40:11)

Trees	100
Depth	13
Min samples	1
Size of hyperparameter search	2

Most important variables

- noprotra
- gradvinc is 5
- volneg
- fact
- gradvinc is 1
- num_bancos

Train set	22202 rows
Test set	5571 rows
Train time	about 40 seconds

Gradient Boosted Trees 0.879 ✔ Done 3 days ago (2021-05-11 05:40:43)

Trees	100
Learning rate	0.1
Max depth	3

Most important variables

- noprotra
- num_bancos
- fact
- volneg
- cnae3 is 561
- gradvinc is 1

Train set	22202 rows
Test set	5571 rows
Train time	one minute and 11 seconds

XGBoost 🏆 0.880 ✔ Done 3 days ago (2021-05-11 05:42:44)

Trees	55
Max depth	3

Most important variables

- noprotra
- fact
- cnae3 is 561
- num_bancos
- volneg
- cnae3 is 682

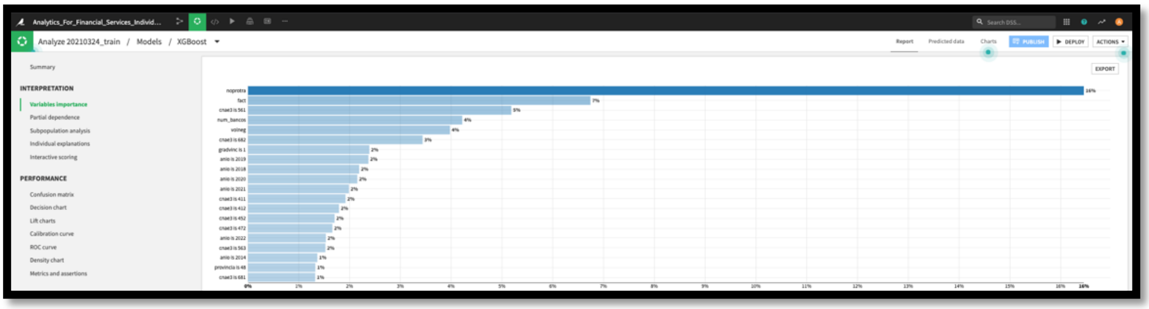
Train set	22202 rows
Test set	5571 rows
Train time	2 minutes and 27 seconds

QUESTION 4.

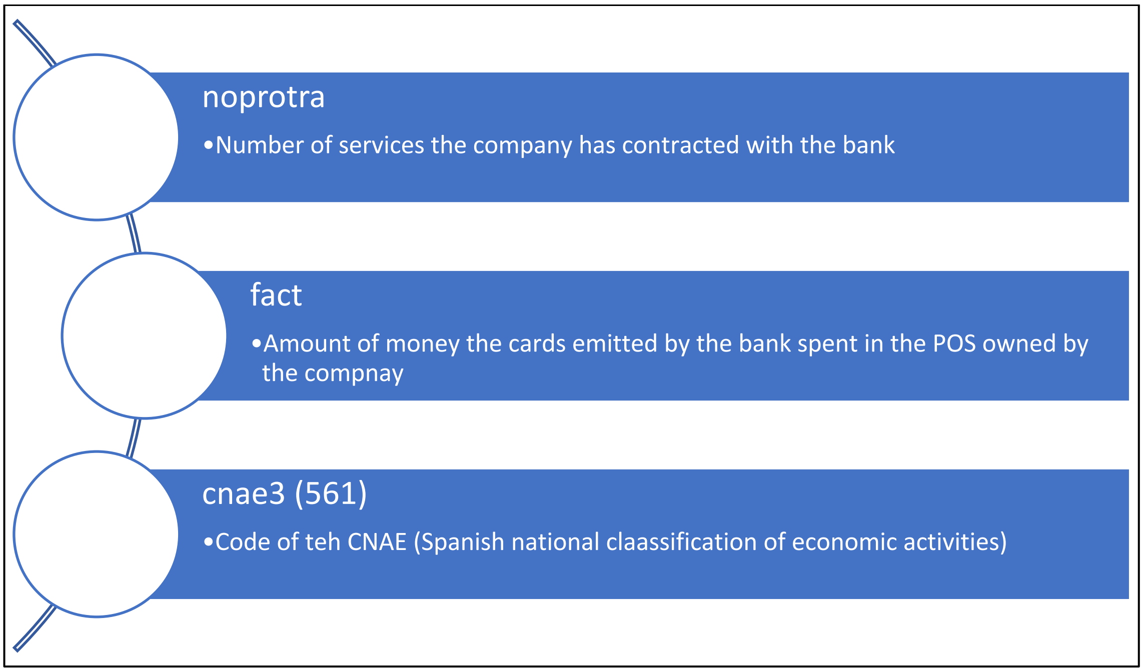
The Outcome [recommended max. 150 words]

What are the most significative variables in the model?

The most significant variables in the model include the following:



Here is a list of the top three variables of importance:



noprotra

- Number of services the company has contracted with the bank

fact

- Amount of money the cards emitted by the bank spent in the POS owned by the compnay

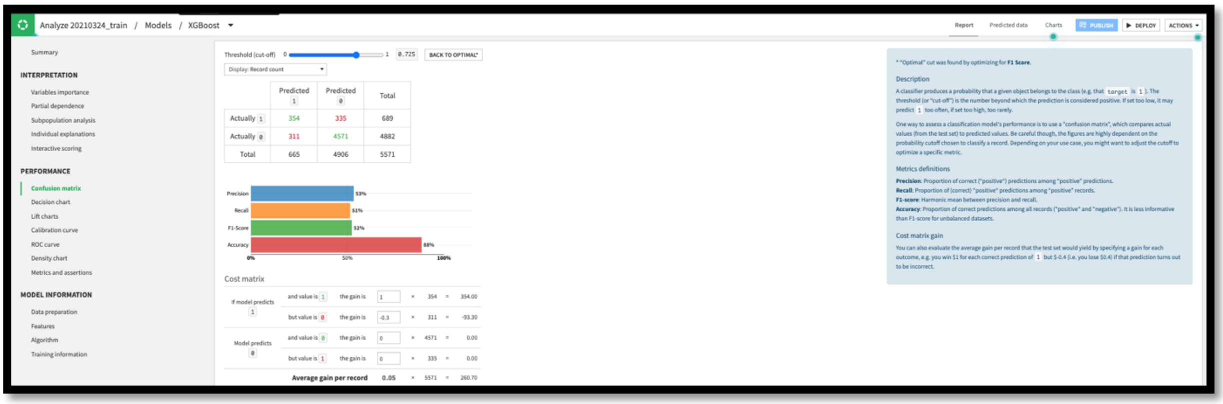
cnae3 (561)

- Code of teh CNAE (Spanish national claassification of economic activities)

Unsurprisingly, the correlations we uncovered during the exploratory data analysis rings true when evaluating the model. For example, the higher the **noprotra** the more likely the company will contract the POS in the following 3 months. In other words, companies are much more likely to purchase when the number of services the company has contracted with the bank is high. Similarly, there is a positive correlation between the target and **fact**. Essentially, companies are much more likely to purchase when the amount of money the cards emitted by the bank spent in the POS owned by the company is high.

Please complete the Data-set-test variables with the outcome of your model – remember the accuracy of the model will weight 30% of your final grade in the individual assignment. Please Indicate here the accuracy of the model:

Using the threshold (cut-off) of .725, the model has the following confusion matrix and performance metrics:



Threshold (cut-off) 0 — 1 0.725

Display: Record count

	Predicted 1	Predicted 0	Total
Actually 1	354	335	689
Actually 0	311	4571	4882
Total	665	4906	5571

Cost matrix

If model predicts 1	and value is 1	the gain is	1	×	354	=	354.00
	but value is 0	the gain is	-0.3	×	311	=	-93.30
Model predicts 0	and value is 0	the gain is	0	×	4571	=	0.00
	but value is 1	the gain is	0	×	335	=	0.00
	Average gain per record		0.05	×	5571	=	260.70

* "Optimal" cut was found by optimizing for **F1 Score**.

Description

A classifier produces a probability that a given object belongs to the class (e.g. that `target` is `1`). The threshold (or "cut-off") is the number beyond which the prediction is considered positive. If set too low, it may predict `1` too often, if set too high, too rarely.

One way to assess a classification model's performance is to use a "confusion matrix", which compares actual values (from the test set) to predicted values. Be careful though, the figures are highly dependent on the probability cutoff chosen to classify a record. Depending on your use case, you might want to adjust the cutoff to optimize a specific metric.

Metrics definitions

Precision: Proportion of correct ("positive") predictions among "positive" predictions.
Recall: Proportion of (correct) "positive" predictions among "positive" records.
F1-score: Harmonic mean between precision and recall.
Accuracy: Proportion of correct predictions among all records ("positive" and "negative"). It is less informative than F1-score for unbalanced datasets.

Cost matrix gain

You can also evaluate the average gain per record that the test set would yield by specifying a gain for each outcome, e.g. you win $1 for each correct prediction of `1` but $-0.4 (i.e. you lose $0.4) if that prediction turns out to be incorrect.

In regard to the business case, the worse outcome from this model is informing the audience that the model predicts a company would not buy although in reality the company does buy. In other words, we want to reduce the number of **False Negatives,** or the aggregate occurrences the model predictes “No, the company won’t buy” although it was actually “Yes, the company will buy.” Thus, if we were to reduce the number of False Negatives, we would want the **threshold** to be a bit more generous, or lower than it currently is. However, if we want to optimize for accuracy, we will want to keep the threshold where it is.

Example Answer: *The accuracy of the model in the test data* ***is 0.75.*** *The model predicts well* ***2.139 cases*** *over the total 2.852 cases of the test dataset (****77%*** *of the cases - accuracy).*

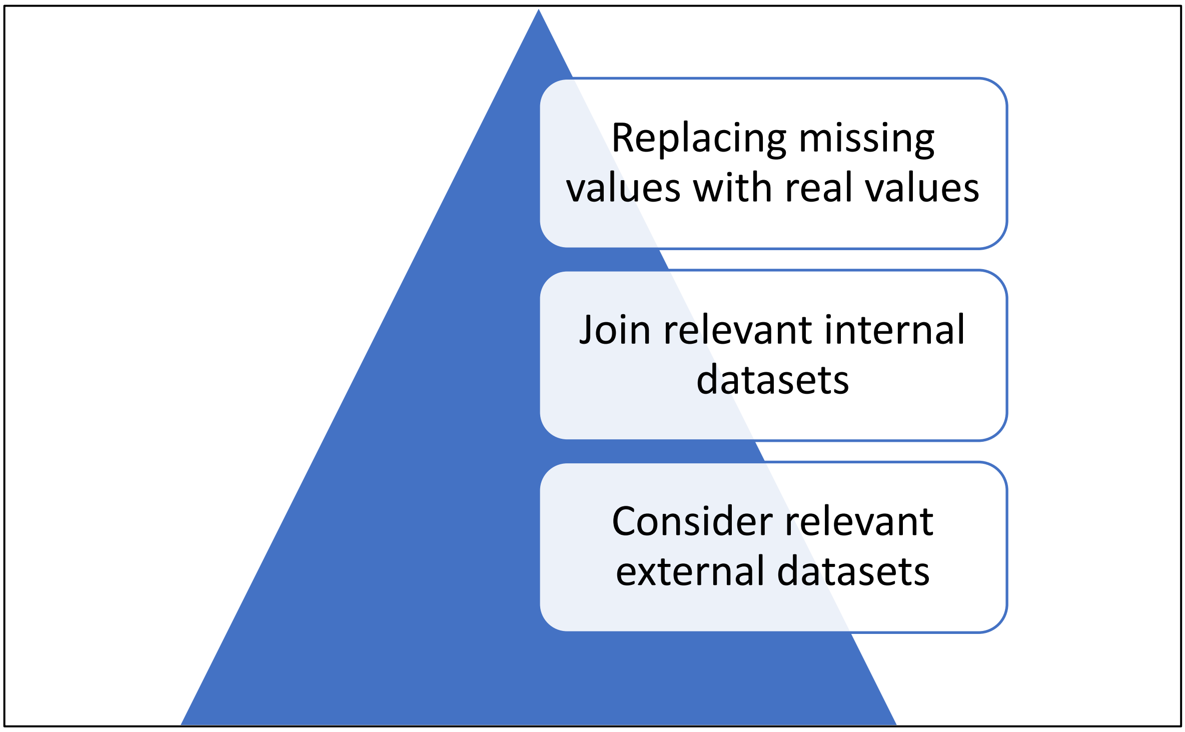
Your Answer: The accuracy of the model in the test data **is .884.** The model predicts well ***2521* cases** over the total 2.852 cases of the test dataset (**88%** of the cases - accuracy).

QUESTION 5.

Improving the model [recommended max. 100 words]

What type of improvements to the Data-set or the steps followed through the analysis would you implement to improve the effectivity of the model?

I would suggest the following improvements to improve the effectivity of the model:



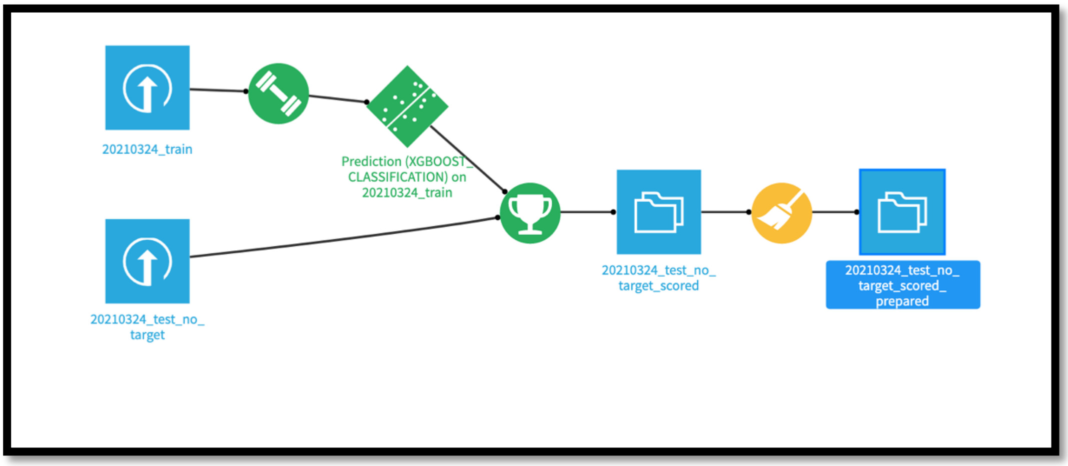
Replacing missing values with real values

Join relevant internal datasets

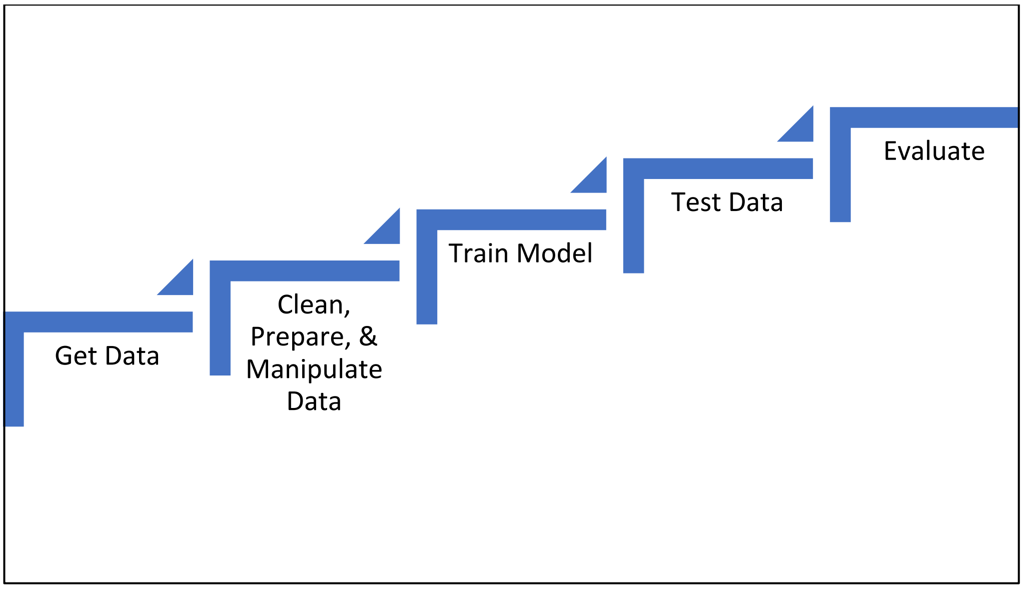
Consider relevant external datasets

ANY OTHER COMMENTS [recommended max. 100 words]

Here is a snippet of the “flow,” or the process by which this analysis was conducted, per Dataiku:


20210324_train
Prediction (XGBOOST_CLASSIFICATION) on 20210324_train
20210324_test_no_target
20210324_test_no_target_scored
20210324_test_no_target_scored_prepared

In summary, I completed the following machine learning process to build predictions for the test dataset:


Get Data
Clean, Prepare, & Manipulate Data
Train Model
Test Data
Evaluate