Name: Rodrigo Hurtado

Date: June 9th, 2025

Part 1: Exploration

In this section, you should describe:

- Results of **descriptive statistics** about your columns.
- Your investigation into **missing values** and how you dealt with them. (Remember: leaving them alone is a valid option if it's justified!)
- Your investigation into **outliers** and how you dealt with them. (Remember: leaving them alone is a valid option if it's justified!)
- The exploration of the relationship between your potential features and the target, e.g. answering questions like, "How did the percentage of people who bought the product vary with the age of customers?" and "Are older or younger customers more likely to buy?" Based on these answers, which **features** did you choose for modeling?

Provide supporting visuals where appropriate.

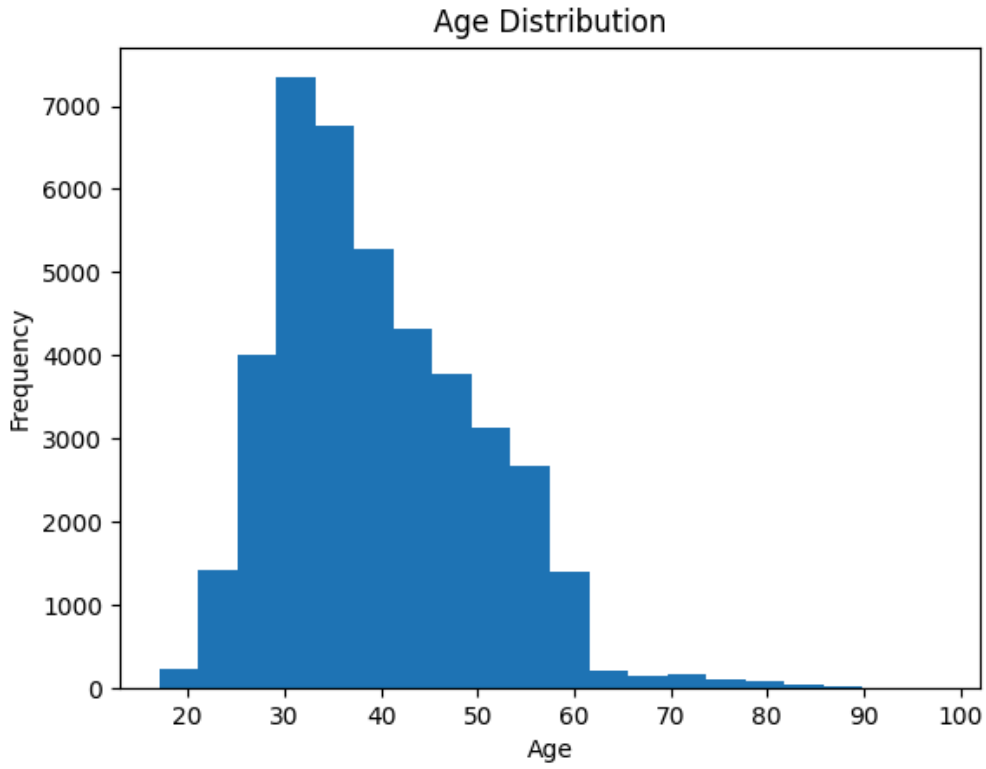
EDA

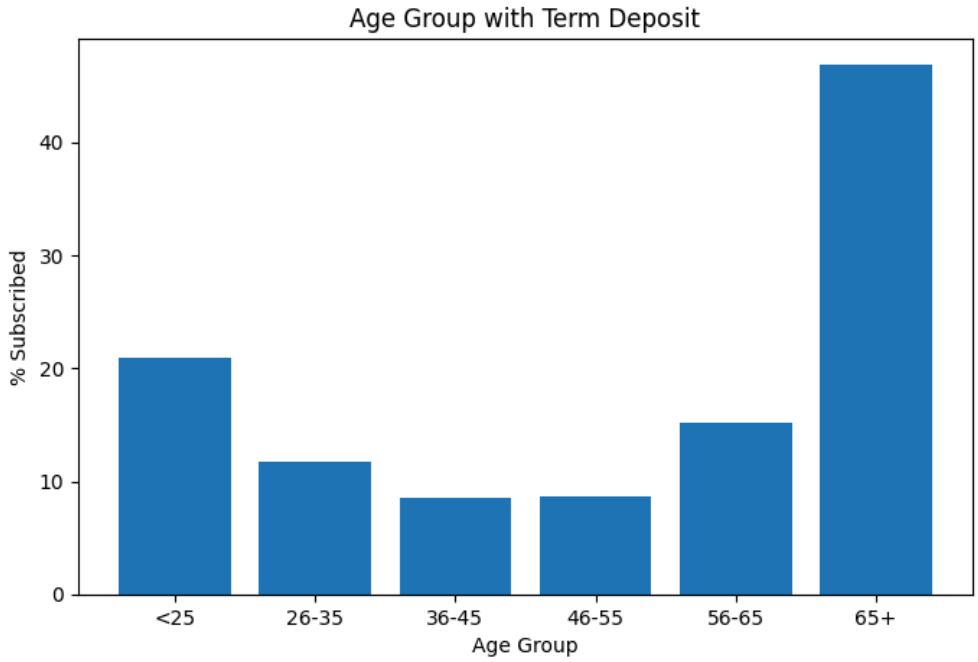
The dataset contains a total of 41,188 observations and 19 columns. Only the following columns had missing values:

- Default (8,597 NaN): Filled all missing values with a new category ('unknown') to be able to use in the model as no customers with loan defaults subscribed to the product, making this feature a strong predictor.
- Housing (990 NaN): Filled all missing values using sklearn 'simpleimputer' with the the most frequent value found in the column.
- Loan (990 NaN): Filled all missing values using sklearn 'simpleimputer' with the most frequent value found in the column.

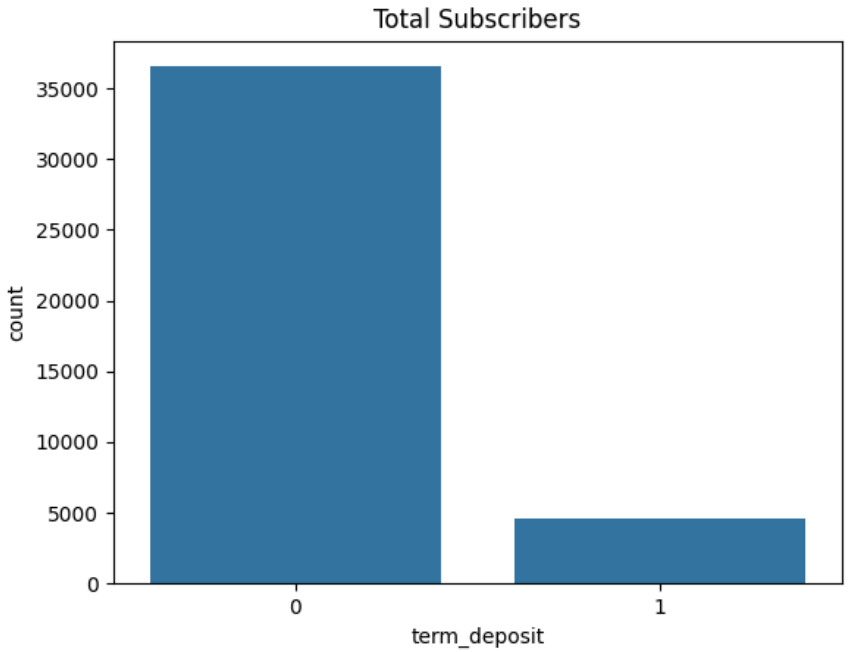
Found some outliers in the 'age' column which showed a maximum age of 98 years. These outliers were left untouched, as it was found that 47% of subscribers were of age 65+. The majorities of customers contacted were between the ages 20 - 60 years. Of

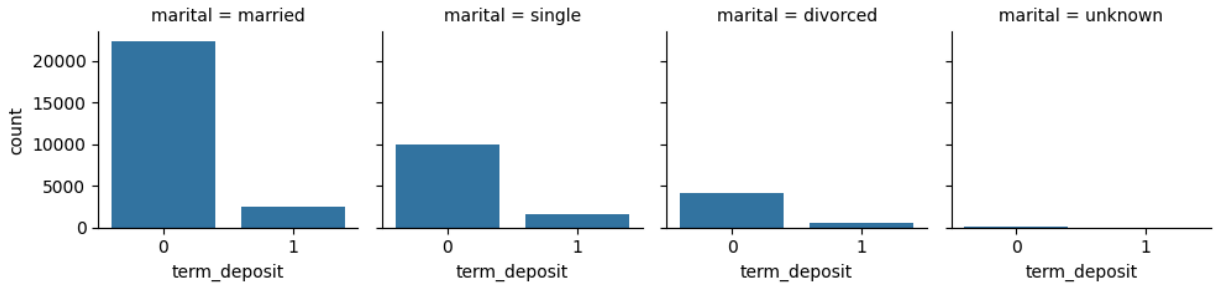
those customers it was found that 20% of customers <25 and 47% of customer >65 subscribed to the term deposit, making those two age groups the highest subscribers as seen in the figures below..



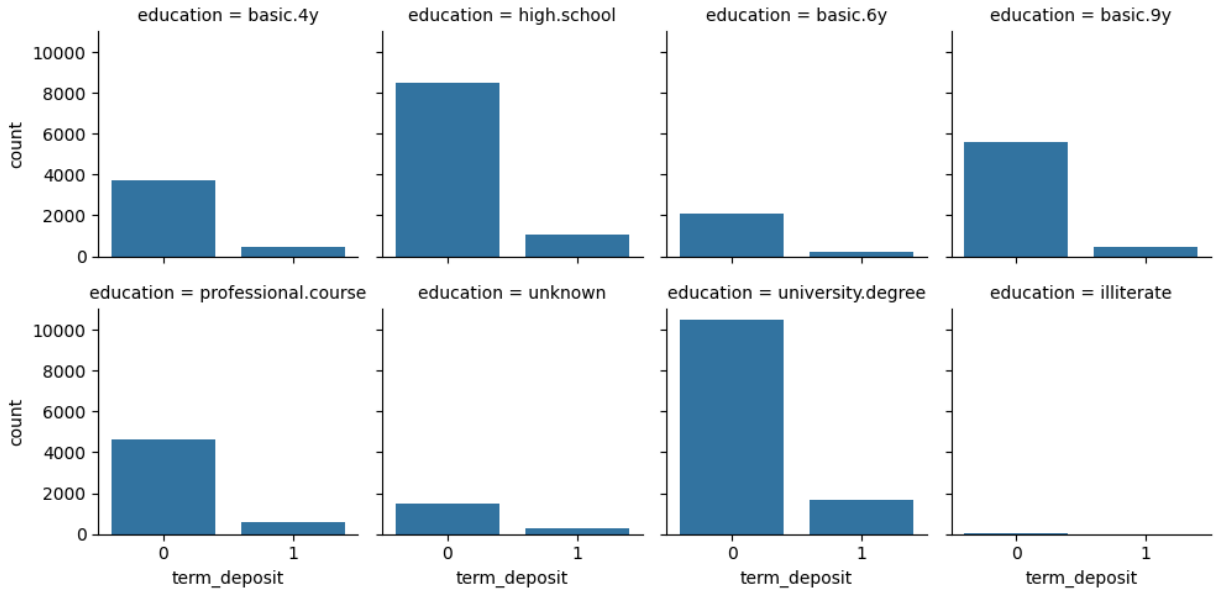


Only 11.26% of customers subscribed to the term deposit leaving 88.75% non-subscribers. See below figure showing the total count of subscribers vs non-subscribers.





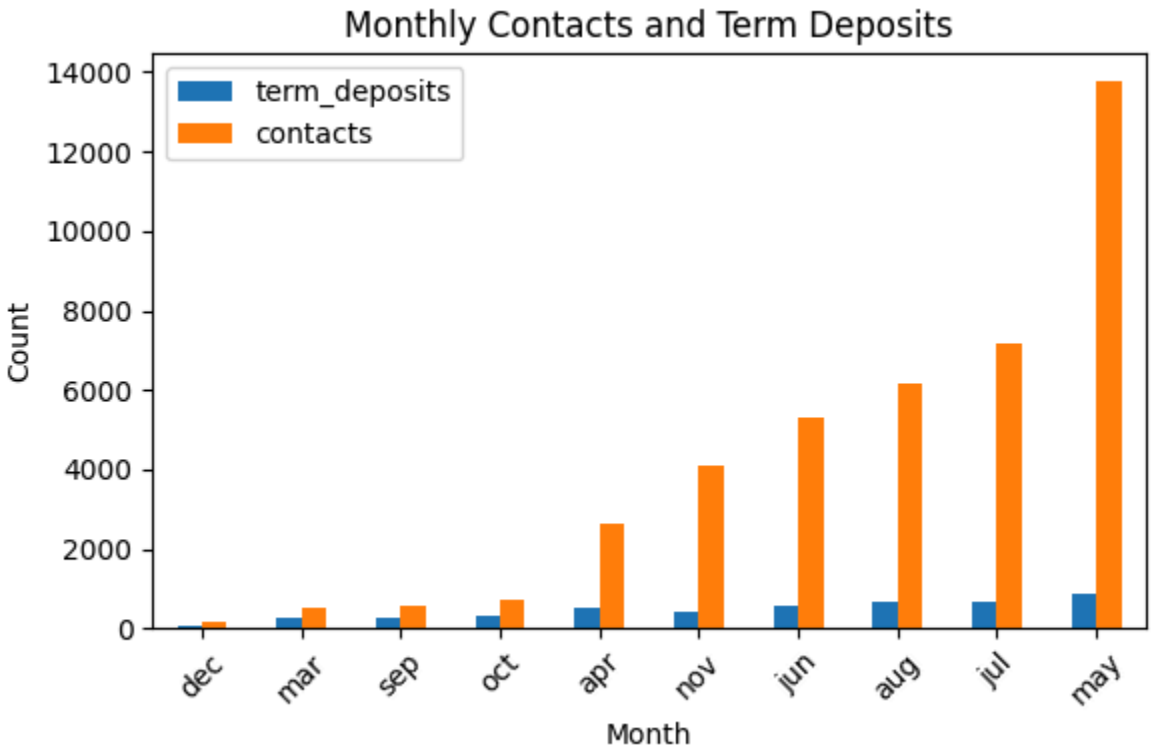
As seen in the chart above, marital status had little effect on subscribing for the term deposit. However, the vast majority of contacted customers were married, followed by singles and divorced.



The chart above shows the number of subscribers by education level The vast majority of contacted customers have a highschool degree or university degree and both categories subscribed the most. In the list above we can see the ratio of subscribers increasing with higher levels of education.

Illiterate	22.2%
unknown	14.5%
university.degree	13.7%
professional.course	11.3%

high.school	10.8%
basic.4y	10.2%
basic.6y	8.2%
basic.9y	7.8%



In the chart we can observe the number of contacts vs term deposits by month. This shows no real seasonality in terms of number of subscribers. Also, we can observe that there is no significant increase in the number of subscribers the more contacts were made.

Part 2: Modeling

After completing the Jupyter notebook and training 2 different machine learning models, you should:

- Describe the features you chose for each model.
- Describe the model you used for each model.

- Detail the results of both models.
    - What was their accuracy score?
    - What did the confusion matrix reveal? Include some discussion about false positives and false negatives.
- Decide which model performed better overall, and justify your decision. Is it because one has a higher accuracy, or is it the makeup of the confusion matrices?

Model #1

The model used for this analysis was the KNN (K Nearest Neighbor) in order to predict if a customer would subscribe to the term deposit based on characteristics of their nearest data point in the dataset.

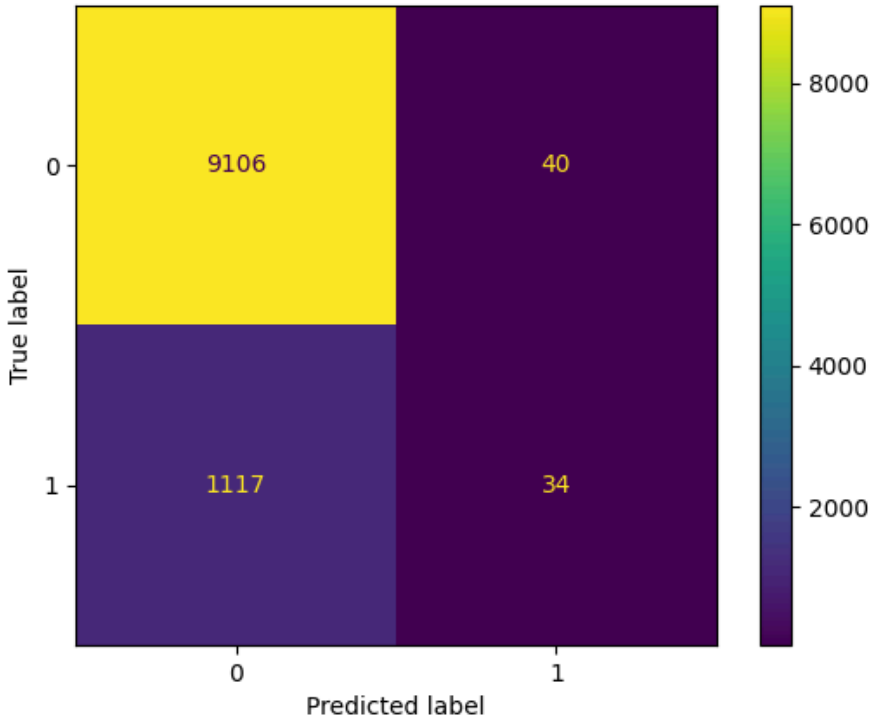
The following features directly related to the customer where chosen to train the model:

- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan

The accuracy of the model was equivalent to the most frequent class, making the model not viable as it should be better than the baseline. In this case the model was good at predicting non subscribers but not good at predicting subscribers. This is mainly due to the dataset being unbalanced, having only 11% of customers subscribing vs 88% non subscribers which prevented the model from picking up more patterns for subscribers. See below the accuracy score vs. most frequent value as well as a confusion matrix.

Accuracy: 0.8879285228707391

Most Frequent Class: 0.8882198698650092



True Positives: 34

- Model correctly predicted "purchase".

True Negatives : 9106

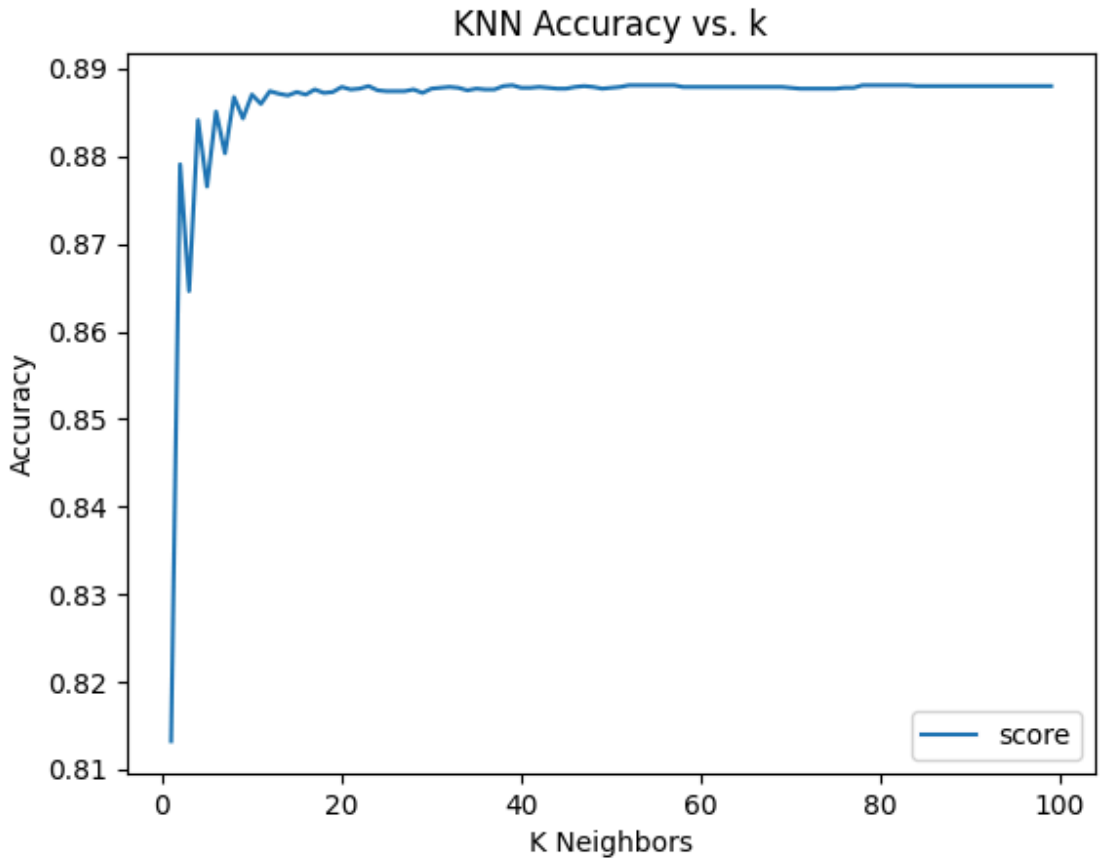
- Model correctly predicted "no purchase".

False Positives: 40

- Model wrongly predicted "purchase" when it was "no purchase".

False Negatives: 1117

- Model wrongly predicted "no purchase" when it was actually a "purchase".



Analyzing the best value of K, we can observe in this chart above that the accuracy reached a peak with k = 20, after which the model maintains the same accuracy up to k = 100.

Model #2

For the second model I added the features related to the economic indicators and the demographics features used in model 1.:

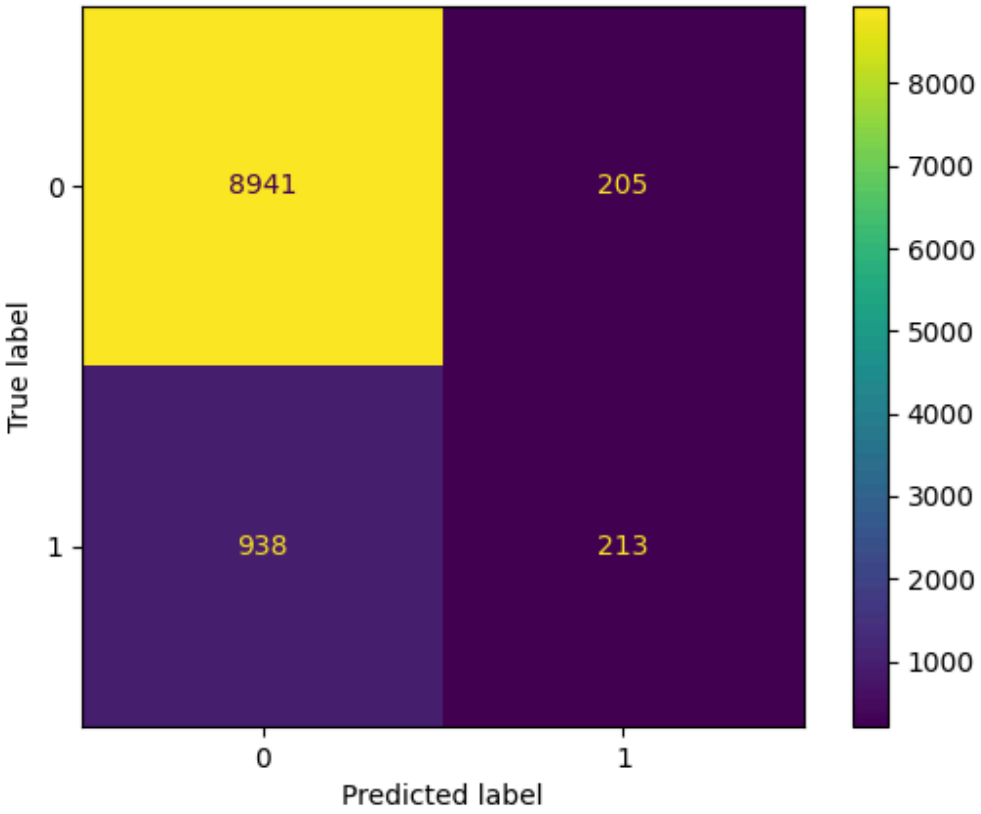
- Age
- Job
- Marital
- Education
- Default
- Housing
- Loan

- Emp.var.rate
- Cons.price.idx
- Euribor3m.
- Nr.employed

The accuracy score remained virtually the same which further shows the effect of the highly imbalanced data set.

Accuracy: 0.8889967951830631
Most Frequent Class: 0.8882198698650092

In the confusion matrix below we can see that again the model is very good at predicting non-subscribers but remains weak at predicting subscribers. Although it improved a lot at predicting subscribers with a total of 213 vs 34.



Conclusion:

The models are accurate at predicting non-subscribers but it is highly influenced by the imbalance found in the data set, having only 11% of contacted customers subscribing. Because of this the model is not useful at predicting subscribers. As seen in the confusion matrix the model barely predicted 34 subscribers correctly. The second model adding the economic features slightly increased the number of correctly predicted subscribers, however it also increased the number of incorrectly predicted non subscribers.

In our EDA it was found that 67% of subscribers were in the age groups of <25 and >65. This age group falls in the category of students and retirees. It was also found that no customer with an active loan default subscribed to the product. The months with the most subscribers were, from May - August. However, no relevant seasonality was found. In all of these months the number of subscribers plateaued at around 600-700 subscribers per month regardless of the number of contacts made which could indicate that the types of customers targeted were not ideal.

It is recommended a more targeted campaign towards customers most likely to subscribe. As mentioned above, students and retirees were the most responsive and likely to subscribe.