



# **THE LEADS TO MARKETING**

**BY CRISTIAN CHOCO**

# CONTENT

- 01** INTRODUCTION
- 02** DATA PROCESSING
- 03** EXPLORATORY DATA ANALYSIS
- 04** MODELING APPROACH
- 05** MODEL EVALUATION
- 06** FEATURE IMPORTANCE
- 07** CONCLUSION

# INTRO



The main focus is to help evaluate a develop a model to boost the efficiency of market campaign to try and increase responses and reduce expenses.



The data set involves data that encompasses 27 different variables to understand whether or not a lead will convert into a sale.



# DATA PREPROCESSING

## cat.codes

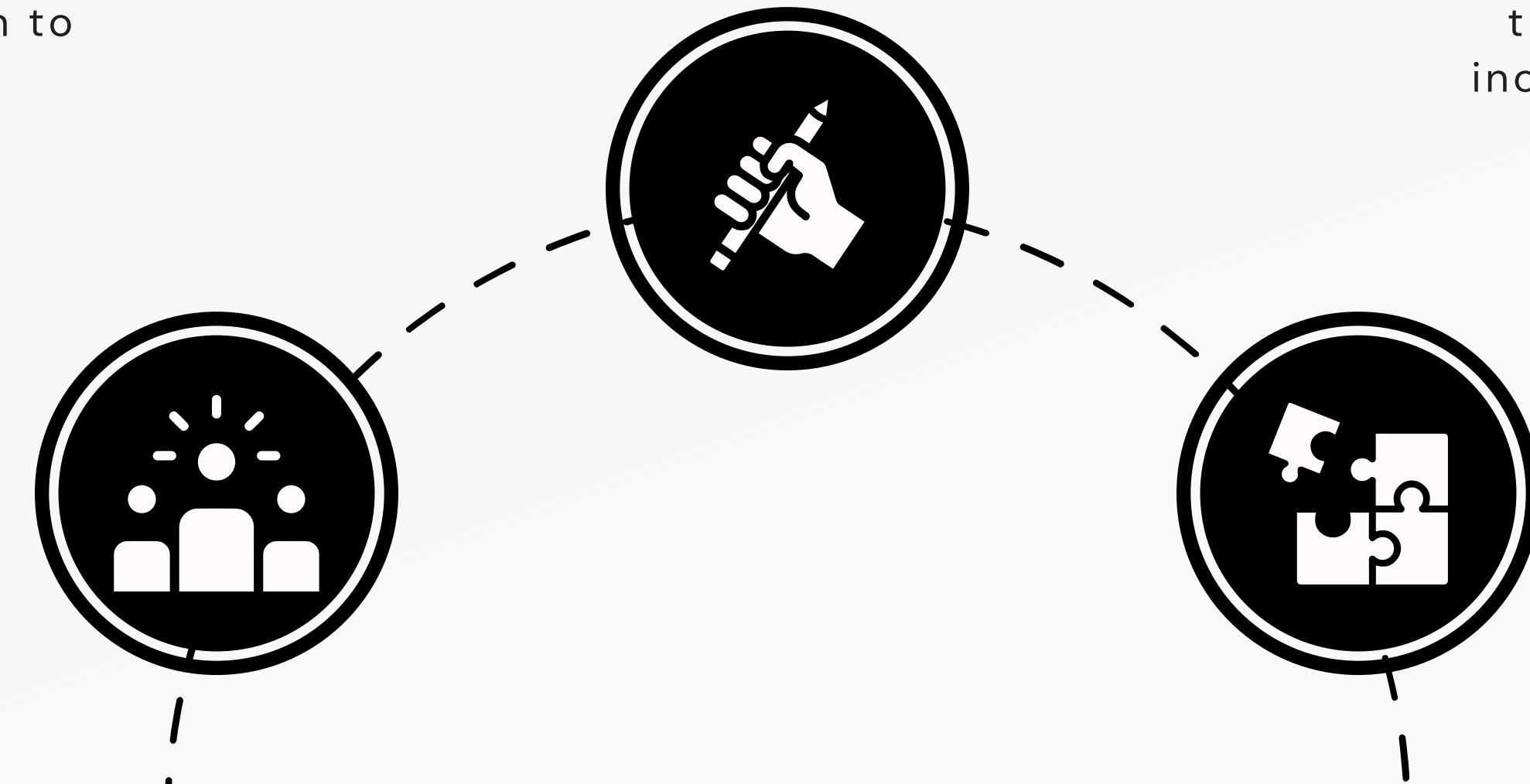
There were 2 categorical columns that were encoded to help with the representation to continue

## is null

To ensure that there were no issues, out of the data set, 24 data samples/rows were removed once found

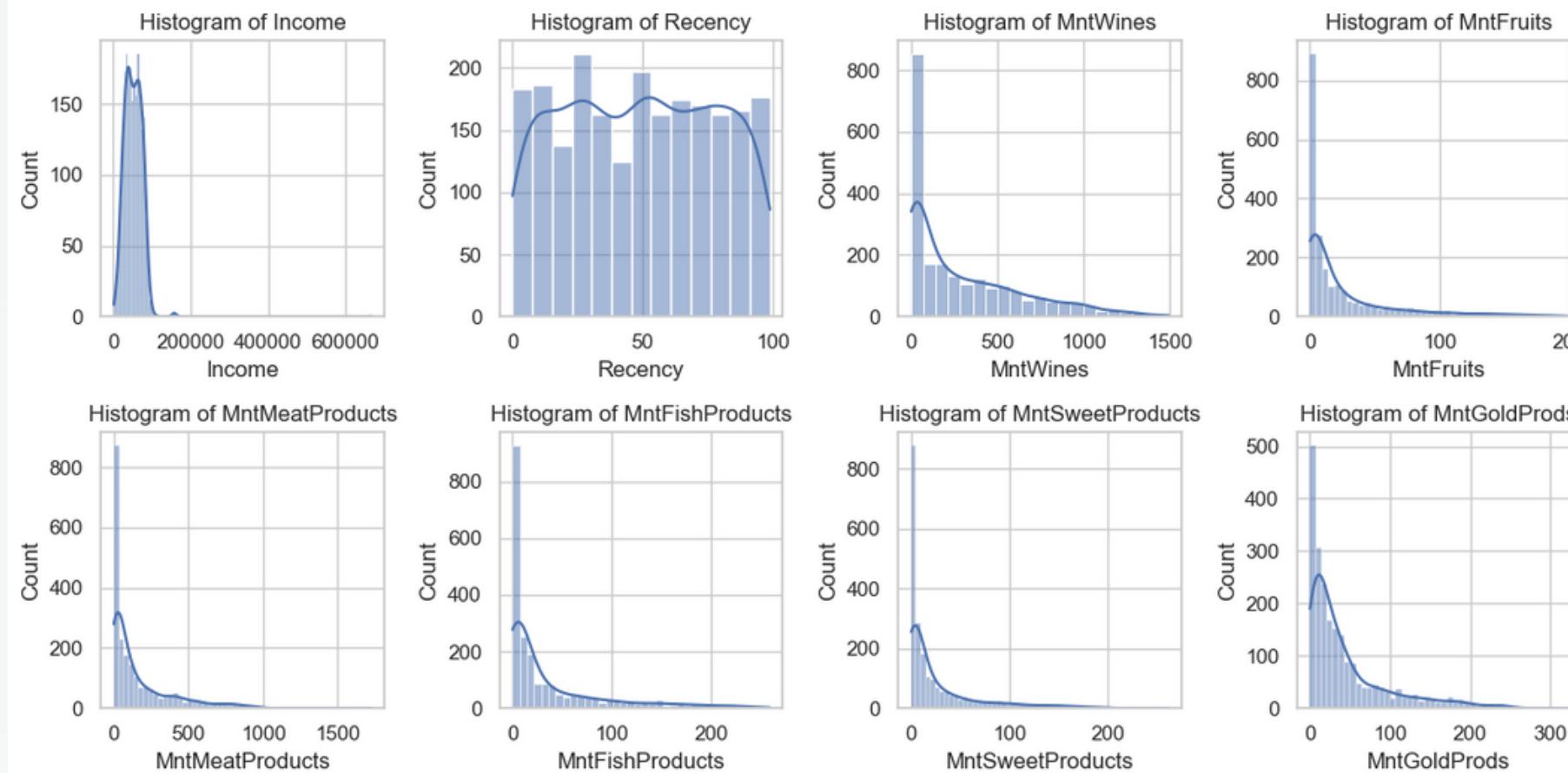
## customize

The dates for the enrollment of customers in company was manipulated so that the data can be included in the analysis



# EXPLORATORY DATA

## *Feature Correlation*



## *Excerpt*

- From the graph we can see the correlations between which features influence the most
- We leveraged the visualizations from graphs and charts to understand the numerical features, outliers, and assessing the correlations with feature to understand the data set

# MODELING APPROACH



## LOGISTIC REGRESSION

Provides interpretable coefficients that can help identify the most influential facts to drive campaign success

## XGBOOST

Builds models that focus on errors made by previous models, which can improve the optimization of the market campaigns

## RANDOM FOREST

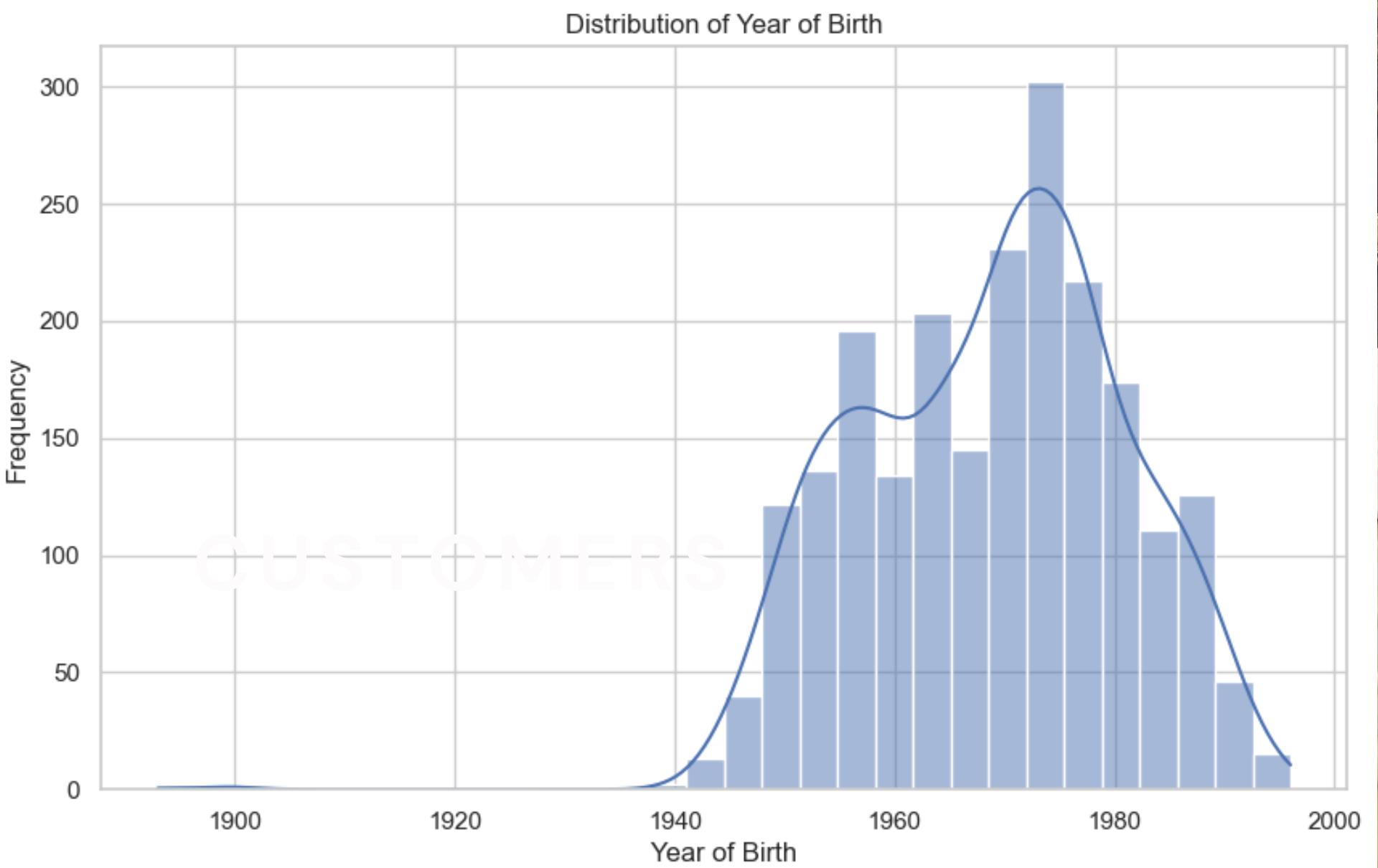
Has the ability to capture complex relationships between customer attributes and campaign resources.

# MODELING EVALUATION

All the models tested have similar accuracy scores, but the model most prominent in precision, recall, F-1, ROC AUC scores indicate that random forest outperforms the other two models tested.

# FEATURE IMPORTANCE

- The features show the frequencies of whether or not there is a response to the market campaign
- On the following graph, there is an increase in frequency to responses between a certain age group, which are the most influenced to respond



# CONCLUSION

*After utilizing Random Forest the following features produced the highest out of the 3 models tested by giving an accuracy of 0.8671*

