Rico
Banco

Bank smarter, not harder

Capitalizing on marketing campaign efficiency through customer segmentation & predictive analytics

TABLE OF CONTENTS

Executive Summary

Final Report

ANALYTICS TEAM

Irina has held professional positions in the banking and retail industries, working with and leading a diverse array of teams and audiences. With the adventurous spirit of a data investigator, she finds creative business solutions and ideas in the delightful land of data.

IRINA FABER

Lisa is a Commercial Analytics manager with 7 years of experience in Data Science and Healthcare Informatics. She has worked for both large health plans and providers and possesses an understanding of B2B and B2C marketing.

LISA IZQUIERDO

Lih is a passionate data analyst with experience in business operations. He has participated in budgeting and overseen constructional projects. With a combination of business expertise and data science, he provides practical analytics and visions.

LIH JING

Siyuan is a senior strategy analyst and has several years experience of leading and participating in analytical projects in the strategy development and market intelligence domains. He brings an ability to think strategically and a technical knowhow with a focus on driving results.

SIYUAN LIU

Joe is an entrepreneur and technology consultant with 8 years of experience. He has served clients across a diverse set of industries and domains. He specializes in data science, digital transformations, and agile delivery leadership.

JOE PUTNIK

Executive Summary

Project Overview

Rico Banco is a Portuguese bank serving retail clients for over 25 years. Given the amount of customer loyalty involved, banking is a notoriously difficult industry to retain customers, let alone grow. An acute understanding of whom we serve guides our marketing team to stay competitive. Rico Banco's leadership team has enlisted their analytics division to evolve their marketing practices through advanced modeling with machine learning and data visualization to build a marketing strategy that can catalyze customer growth over the short term.

Goals and Business Opportunity

Consumer preferences and requirements are always changing. Rico Banco needs to monitor constantly its customer base. Although no two customers are the same, considering customer segments helps us identify the broad categories of clients we already serve or hope to serve in the future. As we begin this analytical endeavor, we are also aware of the potential for cost and resource optimization that are available to us. Analytical approaches to conventional marketing practice are efficient.

Deliverables

To build a marketing strategy to capture market share, we will provide Rico Banco executives with a **benison** of deliverables to better understand its customer base. This goals of this project are to deliver the following: actionable insights to the leadership team, a real-time Tableau dashboard to analyze marketing campaign results, and a Web Application that classifies consumers as subscribers and non-subscribers.

Data Sources

We leveraged bank marketing data sourced from the UCI Machine Learning Repository, containing 41,188 records and 21 features relevant to a direct marketing campaign. The dataset included demographic information, descriptive statistics, socioeconomic indices, as well as a target variable identifying whether the individual subscribed to the term deposit as a result of the call.

Methodology

Before we executed our modeling, we ran extensive Exploratory Data Analysis (EDA) on the input data, implemented data transformations, and created new features to make the analysis more robust. We also used an oversampling technique to create a more balanced dataset to train the models.

We observed several modeling best practices when applying different model types. This included supervised classification models such as logistic regression, decision tree, random forest, Gradient Boosting, Extreme Gradient Boosting, and deep neural networks (DNN). This allowed us to understand relationships between our variables and identify those most meaningful to our marketing strategy.

We used unsupervised learning methods in K-means clusters and Gower Distance to group observations and identify underlying pattern and similarity of our customer base. With the Principal Components Analysis (PCA), for numeric features and Factor Analysis of Mixed Data (FAMD) for categorical features, we reduced dimensionality and mitigated potential collinearity.

Model Performance

Testing Set

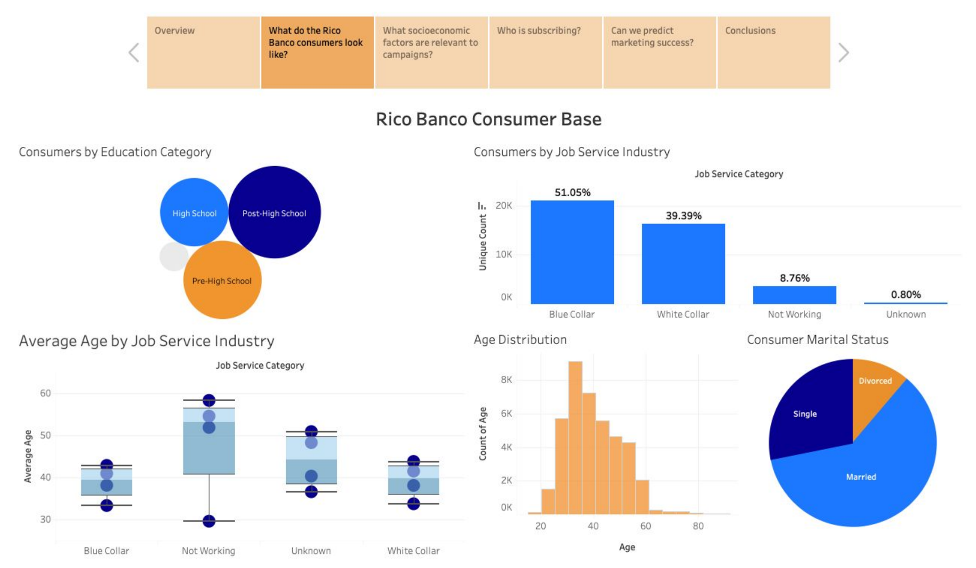
Model/Metrics	Precision %	Recall %	F1 %	AUC %	Accuracy %
Decision Tree	44	53	48	76	88
Random Forest	48	38	42	76	89
Gradient Boosting	46	50	48	78	88
Extreme Gradient Boosting	45	49	47	78	88
Logistic Regression	72	18	29	76	90
Neural Network	33	67	45	75	89

The table shows that all the models perform within 2% accuracy of each other. Differences arise when observing the precision, recall, and F1 scores. The Neural Network model performs with the highest recall while the Logistic Regression model is performing quite poorly with regards to recall.

Our team **recommends using the Neural Net model** to predict future marketing campaign success. Not only did the neural net model have high accuracy, but it also had the highest recall weight compared with the other models. When using models to identify which customers to target via marketing campaigns, high recall is one of the most important measures of success as it is essential not to miss potential customers with high likelihood of subscribing to term deposits. It would ultimately be **costlier** to Rico Banco to **miss out** on potential customers than it would to spend time and money making marketing outreach to those who will not subscribe at all.

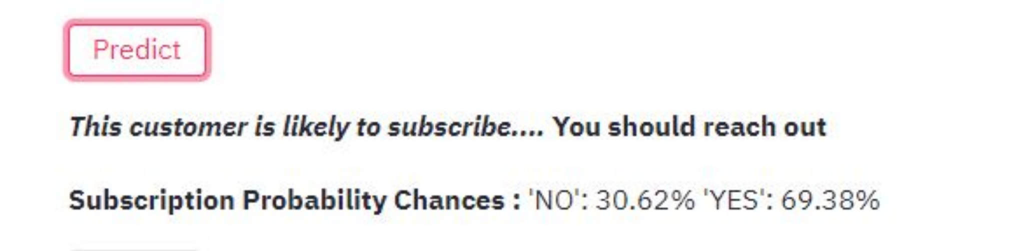
Data Visualization Dashboards

One of the primary goals of this project is to ensure that the marketing and leadership teams at Rico Banco have access to the **insights and analytics** required to make better informed business decisions about marketing strategy and budget. A tableau **dashboard** was built to support the Rico Banco teams in tracking the real-time progress of current and future campaign performance.



Web Application

A **web application** was built on a responsive web framework that scales not only to all PC browsers, but also all mobile and tablet browsers. The navigation panel within the application is where all the inputs are entered and a **prediction** report is generated that shows the possibility of subscription for the specific customer as well as a recommendation.



This customer is likely to subscribe.... You should reach out

Subscription Probability Chances : 'NO': 30.62% 'YES': 69.38%

Summary of Findings

In this project, we have validated a cohesive, end-to-end framework for aggregating, transforming, modeling and visualizing our customer data in order to grow market share while minimizing costs. With this information in hand, the leadership team can successfully and confidently drive strategic marketing campaign decisions forward in a way that directly impacts company Key Performance Indicators (KPIs).

The team recommends leveraging the neural network model due to its high accuracy and recall scores. The Web Application was built upon this model prior to deploying. Additionally, Random Forests can be leveraged to identify the features that most influence term deposit subscription. The same can be done on the socioeconomic features to determine the best time of year to begin a marketing campaign. With this dataset, the following features were identified as most important: Number of Employees, Employment Variation Rate, Consumer Confidence Index, Consumer Price Index, Age, and Has Credit in Default.

Finally, although two different clustering techniques were explored as part of this project, the differences identified across clusters were not significant enough to use to derive meaningful customer segmentation insights. Additionally research could be applied to these methods should the leadership team choose to invest in Phase 2 of this Analysis in future quarters.

Project Status

We set an 8-week timeline, guided by a series of deliverables and updates that would ultimately end in a multi-pronged output including an exhaustive list of ML models, a functioning web app, and a robust tableau dashboard. This project has met these requirements and can be considered as **successfully completed on time**, as of August 16th, 2022.

That being said, the team welcomes additional feedback from the leadership executives with regards to additional business context that may be valuable to incorporate into this report.

We thank the Rico Banco leadership team for providing us with the opportunity to dive into this data to produce actionable business insights that undoubtedly will result in continued profitable customer growth. Our hope is that this analysis will be considered as anything but **perfunctory**. Should our services be requested in the future, we can always be reached at www.KnowledgeableDataScientistsInNeedOfACareerChange.com

Final Oral Presentation will be delivered to the CEO via video recording prior to Sunday, 8/28.

Rico
Banco

FINAL REPORT

Business Problem Formation

PROBLEM STATEMENT

Overview and Context

Rico Banco is a retail bank located in Southern Portugal, founded in 1997. Throughout its 25 years in business, Rico Banco has relied heavily on telemarketing campaigns to solicit new business and to expand market presence across existing customer segment groups.

Earlier this year, the Product Marketing team at Rico Banco identified a gap in the understanding of true Return on Investment of recent marketing campaigns. They hypothesize that the campaigns can be improved upon to keep pace with the evolving demographic and socioeconomic landscape in Portugal. In an effort to prevent Rico Banco from evolving into a **kleptocracy**, the executive board requested the assistance of its analytics team to look more closely at previous campaigns to ensure the company is minimizing expenditure while maximizing profits.

Project Goals

The goal of this project is to analyze a recent term deposit telemarketing campaign to gain understanding of what works well with marketing campaigns and what areas have room for improvement. This includes breaking the customer base into individual cohorts based on collected demographics and then identifying the groups with the highest conversion rates as a result of the marketing campaign.

The models we will build quantify the impact of various customer features to inform a redesigned marketing approach for the Product Marketing team and executive board before 2023 strategy planning.

BUSINESS OPPORTUNITY

Success in any industry comes from a strong understanding of customers. In an oversaturated market, we believe our **customer-centric approach** will be the differentiator that adds value and **drives continued growth.**

Although customer behaviors may be unpredictable, demographic data can be used to train models to **better understand** customer segmentation.

Telemarketing campaigns are time and finance-intensive. Predictive models that enable targeted marketing cohorts result in **optimization of resource allocation.**

DELIVERABLES

Insights and Analytics

Tableau BI Dashboard

Mobile/Web App Development

INSIGHTS & ANALYTICS

Marketing recommendations will be provided related to customer clustering and segmentation. Generated predictive models will be used to estimate the number of new customers acquired via campaigns and to predict whether a customer will successfully convert to subscription

TABLEAU DASHBOARD

Dashboard that equips marketing and leadership teams with the ability to identify the demographic makeup of recent marketing campaigns across conversion and non-conversion cohorts

WEB APPLICATION

A Web application has been developed with user-friendly interface and predictive ML model in the backend. The application is hosted on Streamlit Cloud that scales not only to all PC browsers, but also all mobile and tablet browsers. The probability of the customer's subscription conversion and qualitative recommendation are generated based on the input including the customer's profile and macroeconomic environment.

Data Overview

DATA SOURCE

The data we used for this analysis is a modified version of bank marketing data sourced from a csv file through the UCI Machine Learning Repository.

The data set contains 41,188 records and 21 columns relevant to a direct marketing campaign that was executed through phone calls conducted by a Portuguese banking institution.

The dataset contains demographic information for each client that was contacted by the marketing team, descriptive statistics of the contact, socioeconomic indices, as well as a target variable identifying whether the individual subscribed to the term deposit as a result of the call.

DATA SOURCE

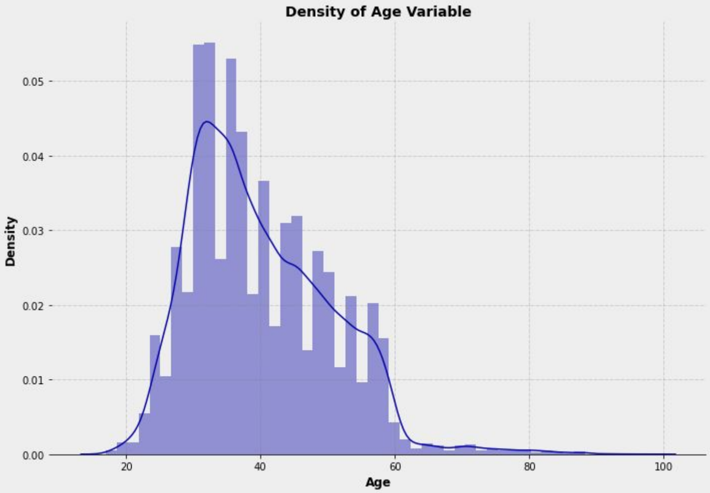
Table 1: Raw Data Attributes

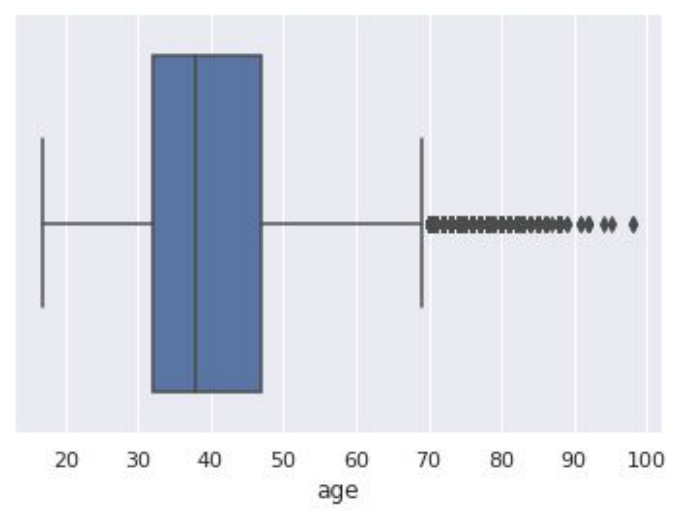
Field	Description
Client Demographic Data	
Age	Numeric Age of consumer
Job	Job type category
Marital	Marital status
Education	Highest level of education obtained
Default	Does client have credit in default?
Housing	Does client have a housing loan?
Loan	Does client have a personal loan?
Last Contact Information	
Contact	Contact communication type
Month	Month in which last contact occurred
Day of Week	Day of week in which last contact occurred
Duration	Length (in seconds) of last contact call
Campaign	Number of times client was contacted during this campaign
pDays	Number of days in between subsequent contacts
Previous	Number of times client was contacted in previous campaigns
pOutcome	Outcome of previous marketing campaign
Social and Economic Attributes	
Emp var rate	Quarterly employment variation rate
Cons price index	Monthly consumer price index
Cons conf index	Monthly consumer confidence index
Euribor 3m	Daily Euribor 3 month rate
Nr employed	Quarterly number of employees
Output Variable	
y	Has the client subscribed to a term deposit?

DATA OVERVIEW

What does the client base look like?

Client ages range from 17 to 98, with the majority of clients falling between the ages of 32 and 44.



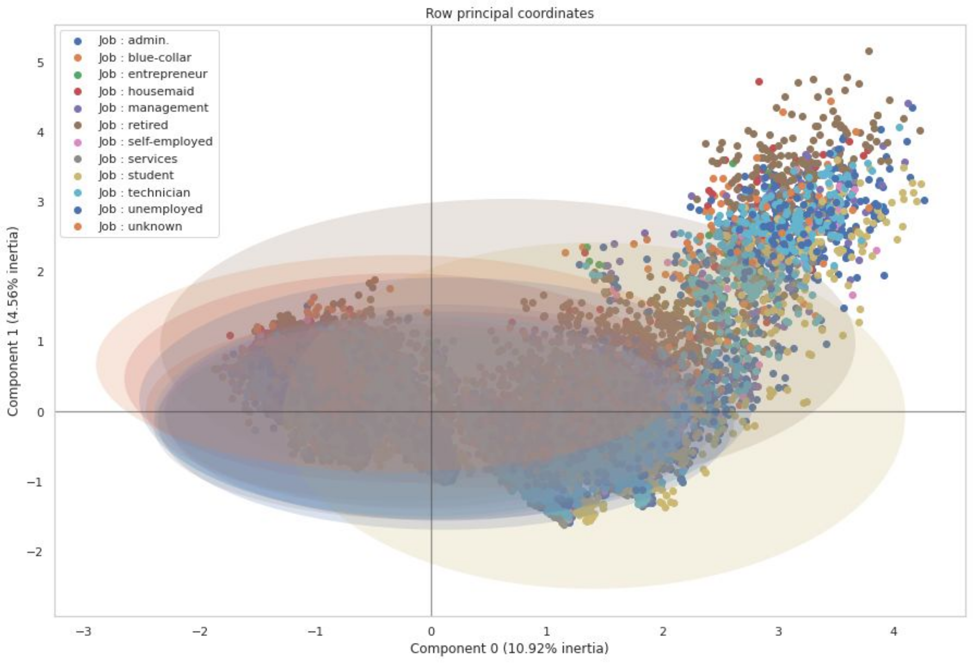


DATA OVERVIEW

What does the client base look like?

The 'job' feature tracks the occupation of customers into 12 sub-categories. According to the FAMD analysis below, customers in some groups such as 'student' or 'retired' have shown distinguishable behavior versus other groups.

Included on the following page is a graph that shows the behavior pattern across different group of customers separated by job highly overlap with each other, leading to 10.9% and 4.6% variability explained by principal component 1 and 2. However, we can tell "student" and "retired" group drift away with different centroid, which suggests that these two groups have differentiated behavior pattern and potentially good to set apart from other groups in feature engineering phase



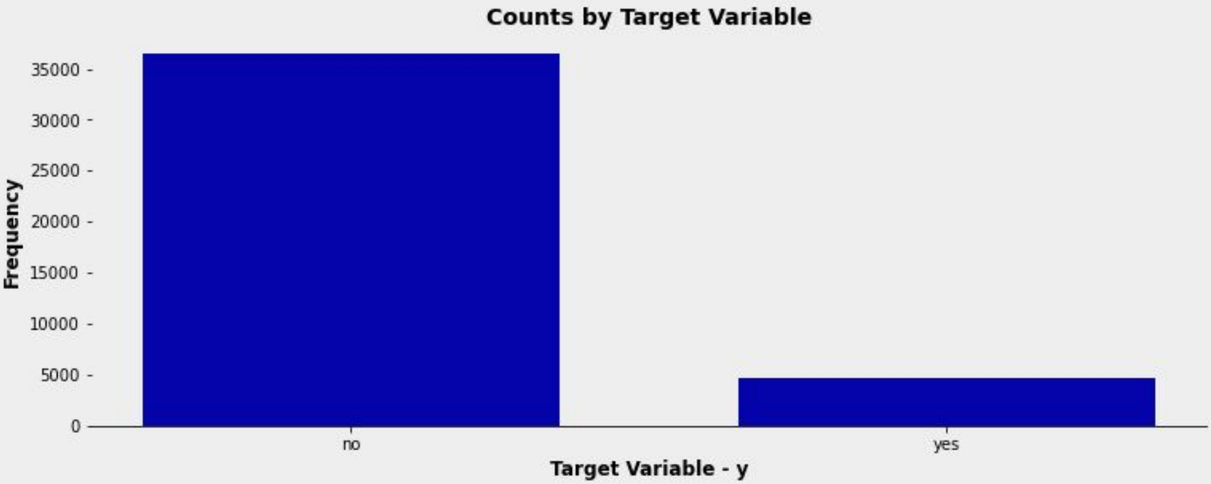
DATA OVERVIEW

Target Variable (y)

The target variable for analysis is a binary indicator that identifies whether the client successfully subscribed to a term deposit on the telemarketing call.

Term Deposit subscription is a strong indicator of overall telemarketing campaign success. Agents make outreach to clients and ask if they would like to subscribe and results (“yes” or “no”) are recorded as the target variable.

The dataset is unbalanced in nature in that only 4,620 of the 41,177 (11.3%) calls resulted in a successful subscription. This issue is addressed in the Data Transformation section.

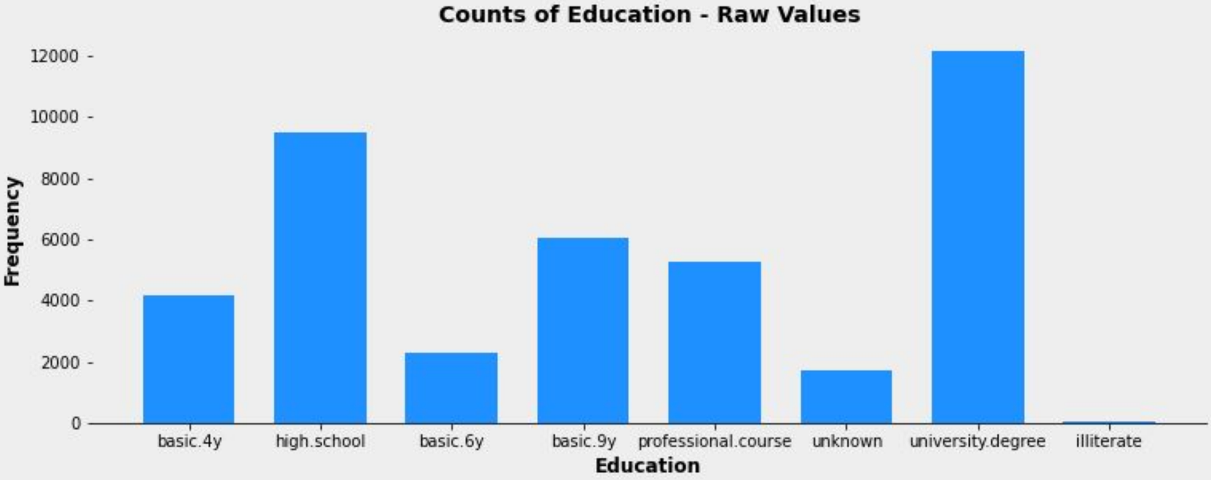


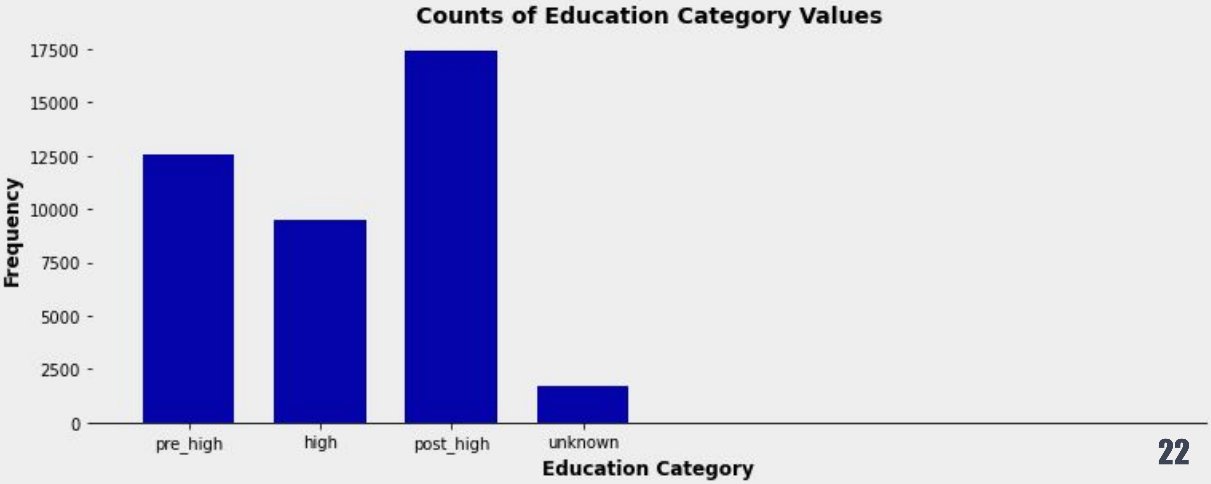
Data Transformation & Preprocessing

DATA TRANSFORMATION

New Feature Creation: Education Category

The 8 raw education values were categorized into 4 new groups: pre-high school, high school, post-high school, and unknown. There were 18 total records corresponding to an education level of “illiterate” that were dropped due to small sample size.

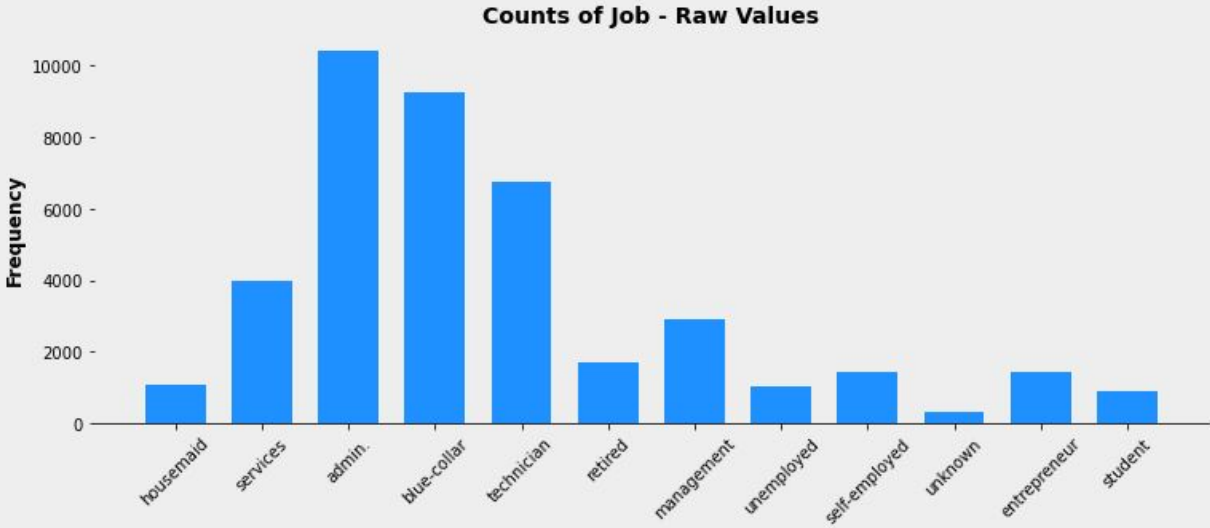


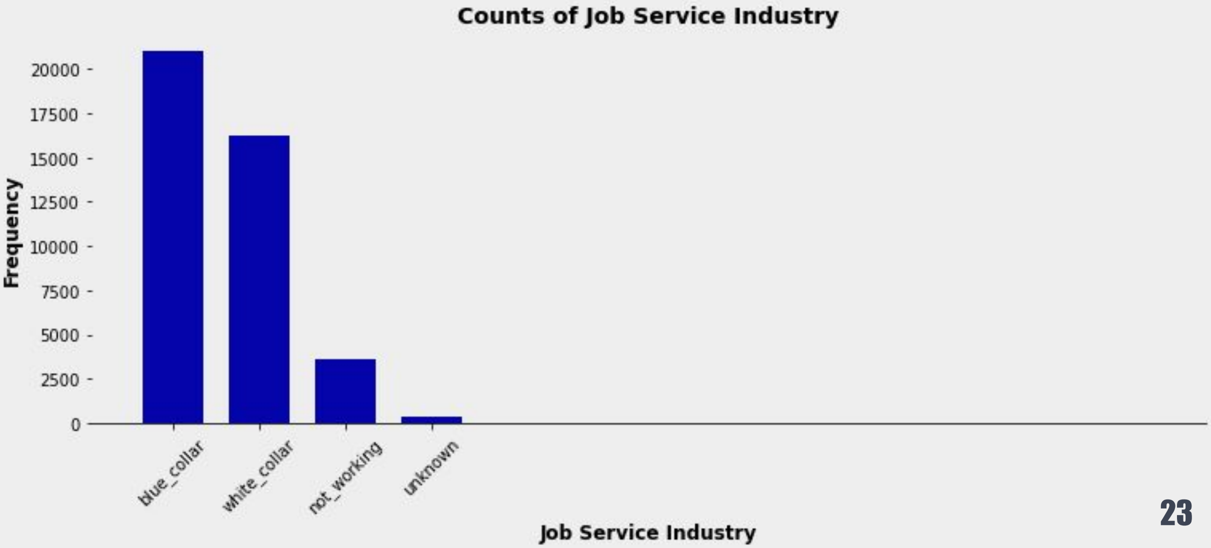


DATA TRANSFORMATION

New Feature Creation: Job Service Industry

The 11 raw job values were consolidated in 4 groups: Blue Collar, White Collar, Not Working, and Unknown. There is a near-50/50 split between blue collar and white collar consumers, with ~9% of consumers falling into not working category.

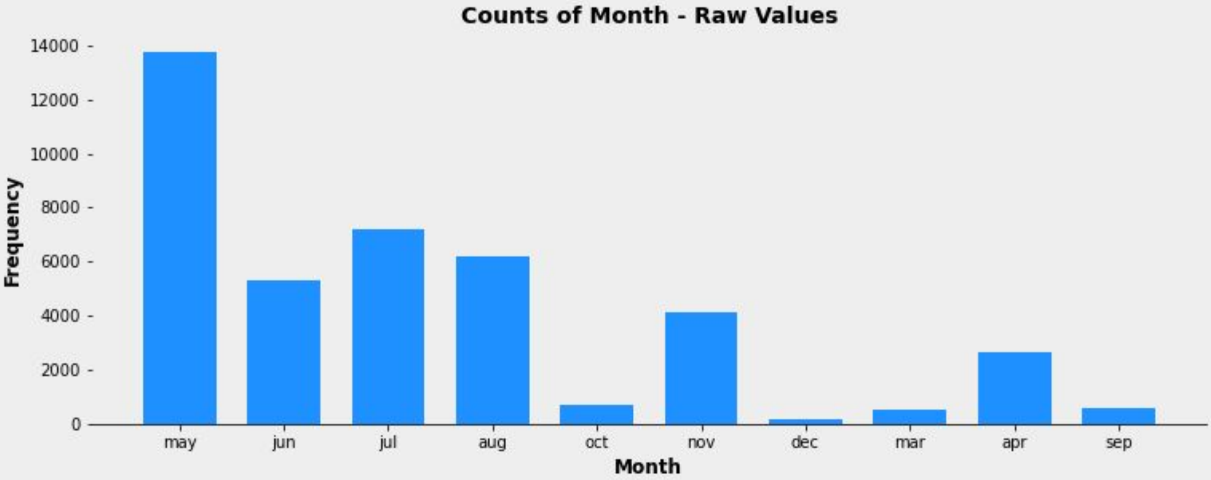


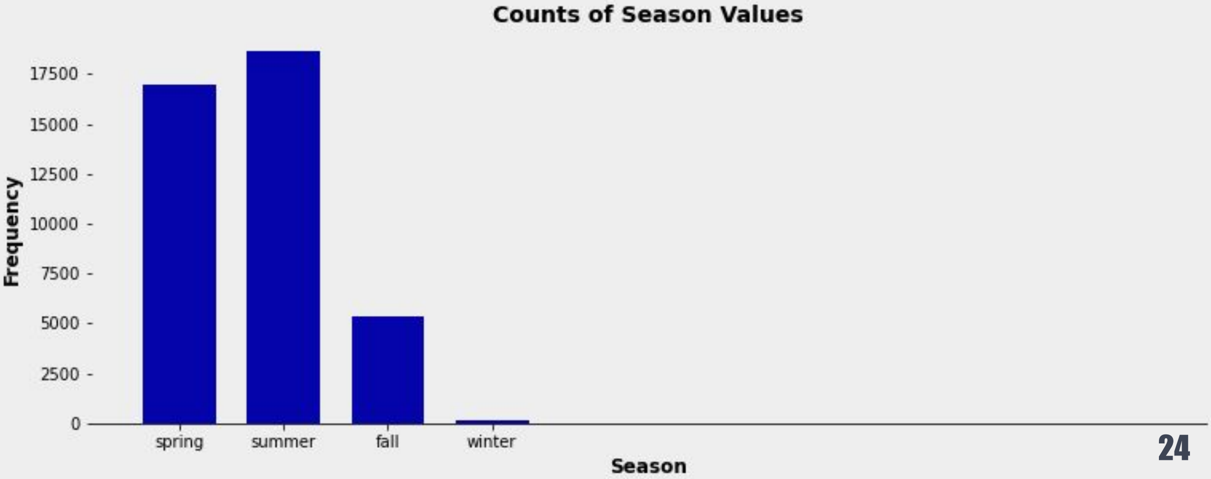


DATA TRANSFORMATION

New Feature Creation: Season

The 12 months in which calls occurred were categorized into corresponding Portugal seasons to enable seasonal trend analysis. As can be seen below, the marketing campaign ran primarily in spring and summer.





DATA TRANSFORMATION

Missing Values

While there were not any null values present in the dataset, there were a number of categorical variables that contained values of “unknown.” Specifically, Age, Marital, Education, Default, Housing, and Loan all contained unknowns.

So as not to **divagate** from the end goals of this project, the team decided not to drop these records in order to retain information within these records we believe would be useful for our prediction models.

Dummy Variables

We applied variable encoding (one-hot encoding or label encoding) to all categorical variables within original and transformed dataset since we assume these categorical features are nominal in nature. By converting and expanding the original dataset with a mix of numerical and categorical features into a sparse binary array, this approach enables feeding into various machine learning models and neural networks.

For neural networks, decision tree, random forest, gradient boosting models, one-hot encoding was used via the get_dummies() Pandas function.

For Logistic regression, label encoding was used via the fit_transform( ) function available in the sklearn Python package.

DATA TRANSFORMATION

Duplicates and Irrelevant Data

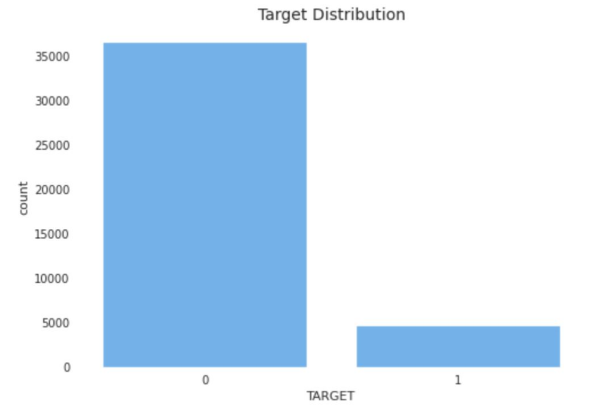
The **Duration** variable was dropped from the dataset as this attribute is dependent upon the target variable. For example, if duration of call is 0 minutes, then the target variable (did client subscribe) will always be false. Additionally, since duration of the call will not be known ahead of time, it cannot be used in a predictive model.

There were 11 “duplicate” rows (0.03%) identified in the dataset. We chose to retain these records since it cannot be assumed that these records correspond to the same call versus two different calls to clients that happen to have the same attributes.

DATA TRANSFORMATION

Addressing Dataset Imbalance with SMOTE

Given the nature of the marketing funnel, there will always be more customers who decline the campaign offer than those who convert to a deposit subscription. Hence, the target variable is unbalanced in this dataset with 36,354 “no” and 4,636 ”yes” values as shown below.



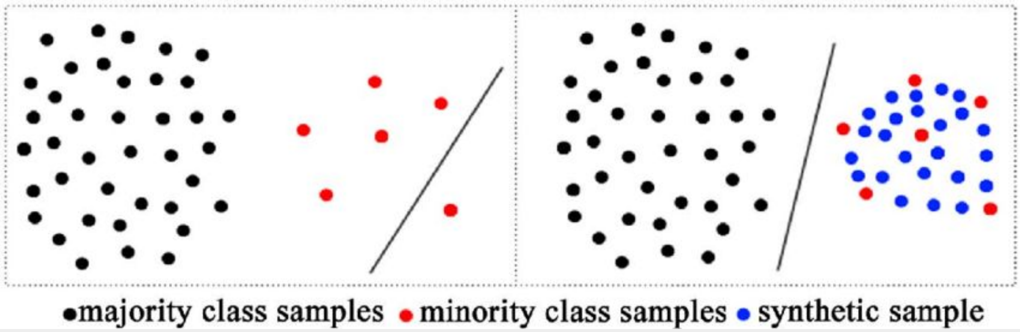
The challenge with working with unbalanced datasets is that most machine learning techniques are unaware of the imbalance, and, in turn, the model will tend to be more accurate towards better-represented classes. Typically, the performance on the minority class is what determines whether the model is accurate and usable in the business context.

To elaborate, if the model we built predicted that 100% of consumers would be non-subscribers, the model would still technically perform with 88% accuracy given that 88% of the consumer population were, in fact, non-subscribers. This is why balancing techniques must be used to prevent oversampling from the majority class.

Addressing Dataset Imbalance with SMOTE

The approach leveraged by our team for this analysis was **SMOTE**, referred as **Synthetic Minority Oversampling Technique** to create additional samples within the minority class and balance the dataset. Specifically, the SMOTE algorithm chooses a random sample of points from the minority class and determines the k-nearest neighbors for those observations along with the vector between each data point and one of the neighbors. That vector is then multiplied by a random number between 0 and 1, resulting in a new/synthesized data point.

Visual Representation of SMOTE Sampling Technique


●majority class samples ●minority class samples ●synthetic sample

source: "A novel over-sampling method and its application to miRNA prediction." Journal of Biomedical Science and Engineering. Jan 2013.

We believe this is the best approach to addressing the data imbalance in the context of this analysis. First, SMOTE does **not create duplicate** data points. Instead, the algorithm creates synthetic data points that will be slightly different than what is found in the minority class. This ensures some level of variance within the training dataset. Second, , SMOTE is **simple** to implement, is not cost-intensive to run, and is more practical for use on low dimensional data. Given the accuracy and recall scores are high for our models , we are confident in this decision.

DATA TRANSFORMATION

Summary of Data Preprocessing

1. The categorical target variable was normalized from "**yes**" and "**no**" values to numeric 1 and 0 values using either **MinMax Scaling** and **Standard Scaling.**

2. The categorical variables non-numerical values were encoded using both the one hot encoding and ordinal encoding (**Label Encoder**) to transform the categorical data to discrete numbers

3. **SMOTE** oversampling technique was used to balance the distribution of target variable values more evenly. After applying, there were 29,239 records of each type.

4. Training and test data frames were created using an **80/20** split, leaving 46,782 records in the training data frame and 11,696 left for testing.

Modeling Methodology

MODELING EVALUATION METRICS

The following list contains the evaluation metrics we used and corresponding definitions:

Receiver Operating Characteristics (Roc) Curve plots the values of True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at different classification thresholds and is often used to compare different models. A model with high discrimination ability will have high sensitivity and specificity simultaneously, leading to an ROC curve near the top left corner of the plot. For this analysis, we plot the ROC curve for all final models and choose a threshold that gives a desirable balance between the false positives and false negatives.

Sensitivity shows the proportion of truly positive observations which is classified as such by the model.

Specificity shows the proportion of truly negative observations which is classified as such by the model.

Area under the ROC (AUC) ranges from zero to one, where a higher value indicates a higher-quality model: value of 1 corresponds to perfect discrimination.

Log loss is the cross-entropy between the model predictions and the target values. This ranges from zero to infinity, where a lower value indicates a higher-quality model.

Confidence threshold determines which predictions to return. A model returns predictions that are at this value or higher. A lower confidence threshold increases recall but lowers precision.

MODELING EVALUATION METRICS

Precision (positive predictive value (PPV) or specificity) is a metric that quantifies the number of correct positive predictions made.
Precision = True Positives / (True Positives + False Positives).
The result is a value between 0.0 for no precision and 1.0 for full or perfect precision.

Recall (true positive rate or sensitivity) is a metric that quantifies the number of correct positive predictions made out of all positive predictions.
Recall = True Positives / (True Positives + False Negatives).
The result is a value between 0.0 for no recall and 1.0 for full or perfect recall. Unlike precision that only comments on the correct positive predictions out of all positive predictions, **recall** provides an indication of **missed positive** predictions and some notion of the coverage of the positive class.

F1 provides a way to combine both precision and recall into a single measure that captures both properties. F1 = 2 [(Precision Recall) / (Precision + Recall)]

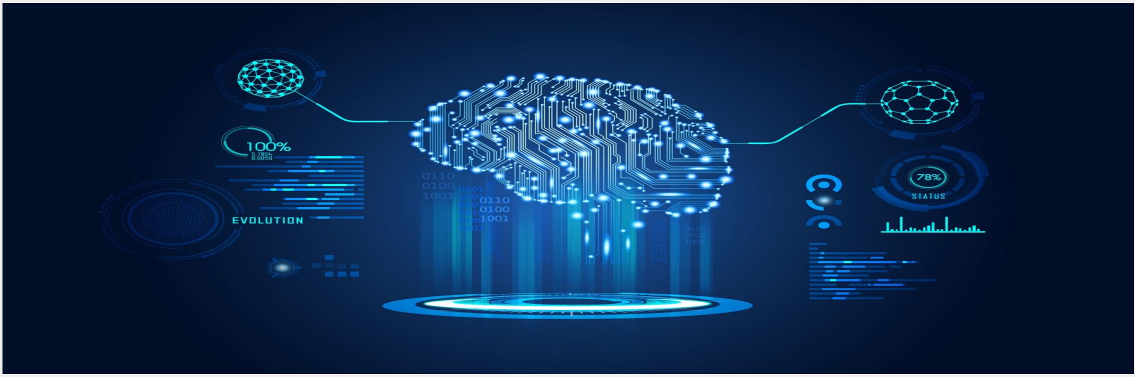
Confusion matrix is a table that is used to measure classification performance and shows the counts of true positives, true negatives, false positives (Type I Error), and false positives (Type II Error). Successful classification models will have the majority of records failing in the first (True Positive) and fourth (True Negative) quadrants.

Model feature recommendation provides the degree to which each feature impacts a model.

MODELING METHODOLOGY

After the input dataset was transformed, a number of modeling techniques were considered. Table 3 includes model types, benefits of each, as well as the team’s current status on executing each model type.

Model Type	Status	Benefits
Classification		
Decision Tree	✅	● Learn relationships/interactions among variables ● Identify/select variables that are most significant ● Provide insights on the contributing factors to select the best term deposit consumers set ● Assign probabilities to new cases
Random Forest	✅	
GBoosting	✅	
XGBoosting	✅	
Logistic Regression	✅	
Neural Network	✅	
K-nearest neighbors	✅	
Clustering		
K-Means	✅	● Group observations ● Identify new target strategies/consumers ● Signal opportunities and reduce risk
Agglomerative Clustering	✅	

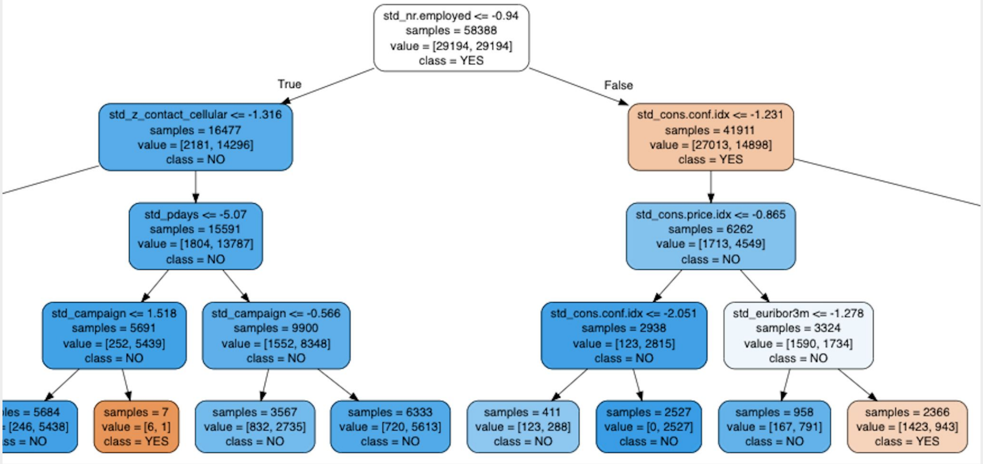


MODELING METHODOLOGY

Decision Trees

The learning algorithm for Decision Tree is **C.A.R.T** (Classification and Regression Tree), which generates only binary trees. C.A.R.T. was developed in 1974 by Leo Breiman and Charles Stone from Berkeley and Jerome Friedman and Richard Olshen from Stanford.

Decision Tree Design



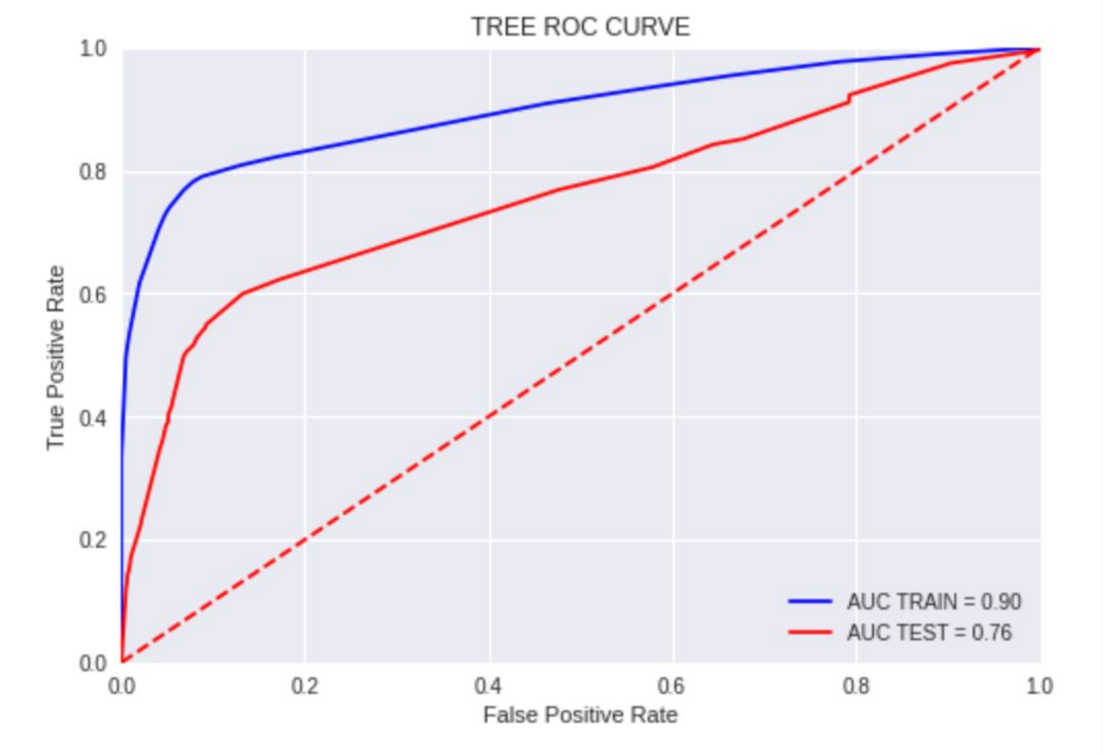
DT is a branching structure that represents a set of **if-then** rules that are easy to understand. In a DT, each internal node represents a 'test' on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. A node with no children is a leaf.

Staring with the entire data set at the **root node** depth 0 and based on the threshold for a specific feature (nr of employees ≤ 5087.65), the sample is split into 2 subsets. This process continues, dividing the entire space into smaller and smaller regions, with the goal to minimize the weighted average impurity of the child nodes. In the end, every region is assigned to a class label, whether a client belongs to group "yes" or "no" (the orange nodes indicate group "yes" and the blue nodes indicate group "no").

MODELING METHODOLOGY

Decision Trees

Some of the **important features** based on the DT are age, campaign, number of days that passed by after the client was last contacted, number of contacts performed before this campaign and for this client, Employment Variation Rate, Consumer Confidence Index, 3 Month Euribor Rate, number of employees, marital status single, and clients with no personal loans.



MODELING METHODOLOGY

Random Forest

Random Forest (RF) was developed by Leo Breiman and Adele Cutler. Like DT, RF uses binary rules to "branch" out. Moreover, RF combines the output of multiple DTs to reach an overall result.

RF uses **bagging**: each tree is trained on a random subset of the same data and the results from all the trees are averaged to find the classification. This generally results in a better model with improved predictive accuracy and overfitting control.

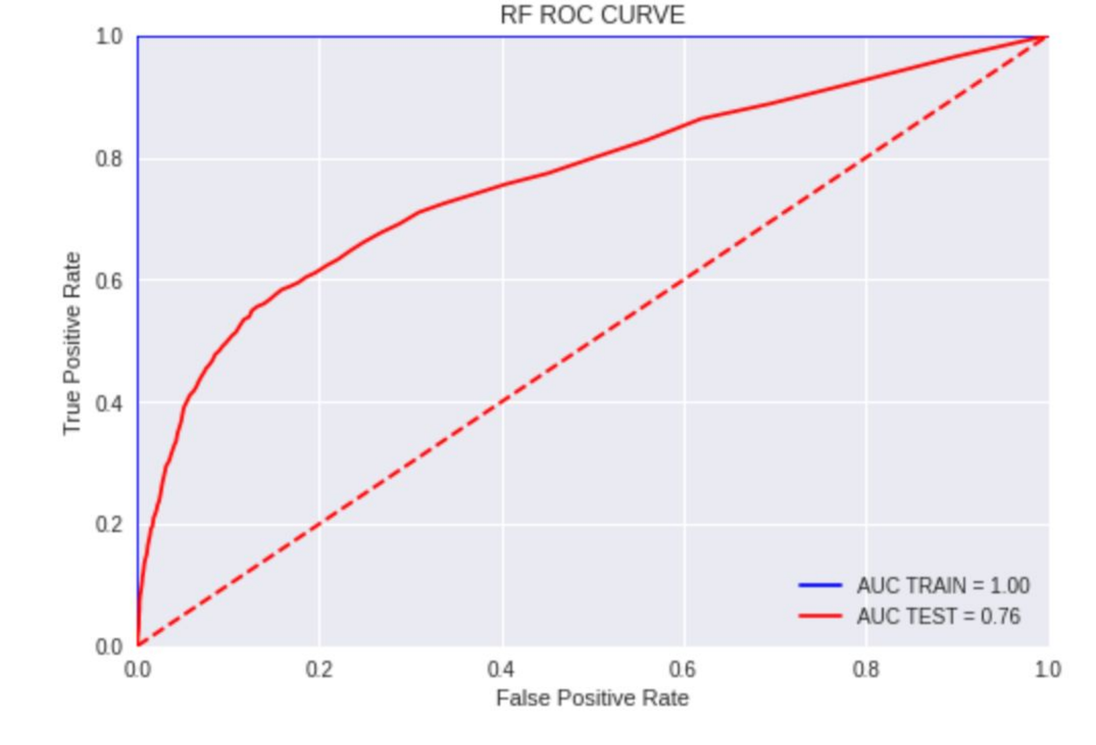
Steps in RF algorithm:

o Step 1: A number of random records are taken from the data set.
o Step 2: Individual decision trees are constructed for each sample.
o Step 3: Each decision tree generates an output.
o Step 4: Final output is considered based on Majority Voting or Averaging for classification or regression respectively.

MODELING METHODOLOGY

Random Forest

Some of the i**mportant features** based on the RF are age, campaign, Employment Variation Rate, Consumer Confidence Index, Consumer Price Index, number of employees, client has no credit in default.



MODELING METHODOLOGY

Gradient Boosting

Gradient boosting (GB) algorithm takes a weak learning algorithm and makes a series of changes to it with the goal of improving the strength of the learner.

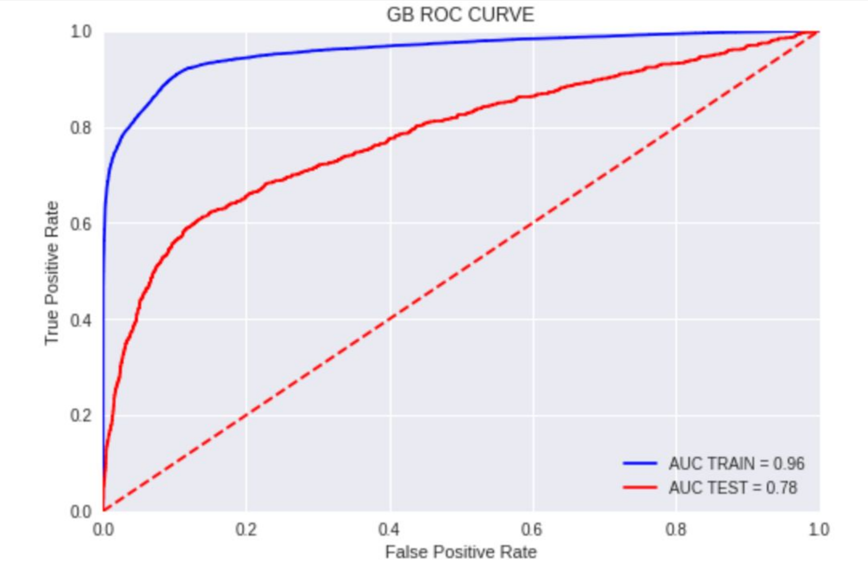
Boosting was first introduced by Yoav Freund and Robert Schapire in 1997. Boosting focuses on sequentially adding up weak learners (a decision tree that classifies data no better than random guessing) and filtering out the observations that a learner gets correct at every step.

So, each model learns from the mistakes of the previous model. The idea is to improve the prediction by converting a number of weak learners to strong learners. This process is repeated until a previously specified number of trees is reached, or the loss is reduced below a certain threshold.

MODELING METHODOLOGY

Gradient Boosting

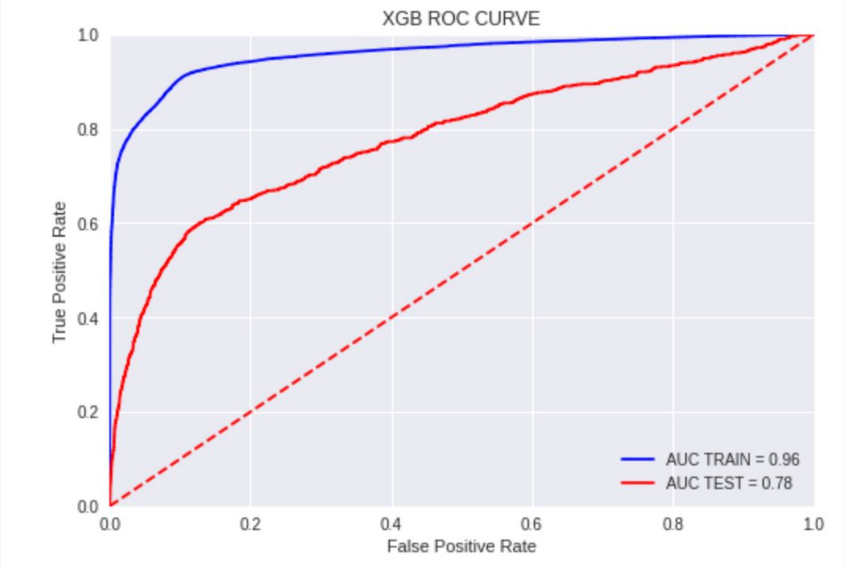
Some of the **features important** based on the GB are campaign, Consumer Confidence Index, Consumer Price Index, Employment Variation Rate, number of employees, client has no credit in default.



MODELING METHODOLOGY

Extreme Gradient Boosting

Tianqi Chen and Carlos Guestrin invented Extreme Gradient Boosting (XCG) in 2016. XGBoost is a more regularized version of Gradient Boosting as it employs a**dvanced regularization** (L1 & L2) to improve model generalization. When compared to Gradient Boosting, XGBoost provides superior performance, has very fast training time and can be parallelized across clusters. XGB can outperform all single algorithm methods on a consistent basis.



Some of the features important based on the XGB are campaign, Consumer Confidence Index, Consumer Price Index, Employment Variation Rate, number of employees, clients with no credit in default, clients who do not own a house, clients with high school education, clients with blue-collar jobs, clients single, and clients contacted by cellular.

MODELING METHODOLOGY

Logistic Regression

Logistic Regression is one of the **most common** predictive model methodologies used in binary classification problems.

The idea is for the model to use information from previous observations to predict a dependent variable in the future. In this project, historical campaign performance can be used to predict the dependent variable, term deposit subscription, based on a series of independent variables such as consumer demographics, social and economic indices, as well as details about the marketing calls themselves. The model intakes the information and can then predict whether to consumer will subscribe.

To train the model, the LogisticRegression( ) function from sklearn was applied to the training data resulting in 90% accuracy. However, since the dataset is imbalanced, high accuracy may be due to the correct predictions from the larger group and in our case, we are more interested in having high accuracy in the minor group. Therefore, we rank recall over accuracy in measuring model performance.

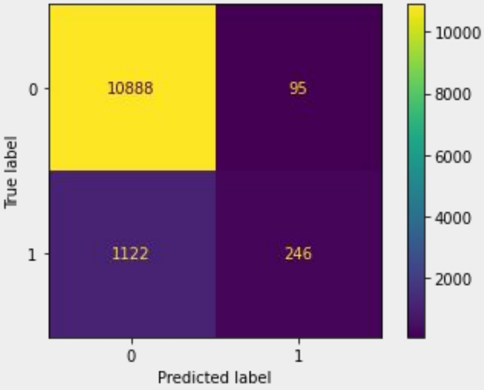
MODELING METHODOLOGY

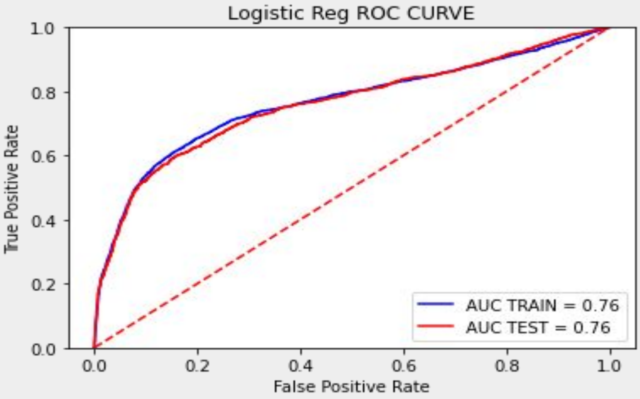
Logistic Regression

To measure logistic regression performance, we calculated the accuracy score, confusion matrix, and AUC.

We notice that the recall of class 1 is showing 0.18 which means the model could only make correct predictions over 18% of class 1 customers (246 out of 1368). The result is not satisfying as the recall is our first focus in measuring performance of a model.

The AUC of the ROC curve only has 0.76. Comparing logistic regression model to other models, the logistic regression model is not the best out of all.





MODELING METHODOLOGY

Clustering

Clustering is a classic unsupervised learning technique in machine learning. Unlike supervised learning, **unsupervised** learning does not require true labels and is not normally used in predictive modeling. However, the results from clustering may help us to gain more insights from the data. Therefore, this project deployed K-means clustering and agglomerative clustering.

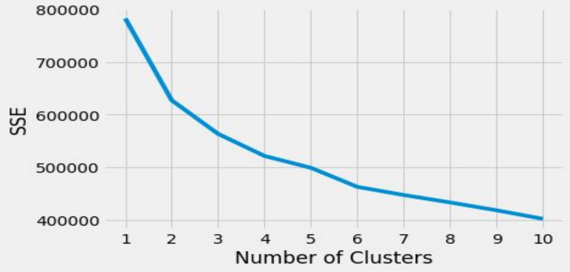
K-means clustering is one of the simplest clustering algorithms. It only requires the number of centroids to initiate. The number of centroids represents the predetermined number of clusters and will be used as the centers of each cluster at the starting phase. Each data point will be assigned to the closest cluster based on the distance between the point and the centroids. Each centroid then will be moved to the incluster average location. The in-cluster sum of squares will be computed based on the distance between data points and the centroid in each group. Through reducing the sum of squares within each cluster, the clusters will eventually converge.

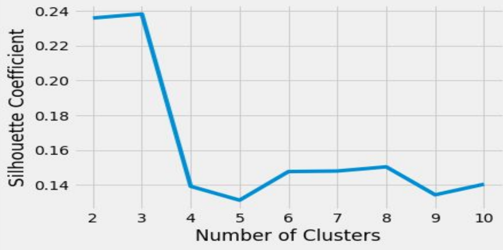
Agglomerative clustering is one of two approaches in hierarchical clustering and requires less computation power out of the two.The agglomerative clustering utilizes the Gower distance to measure the similarity between data points. The algorithm treats every data point as single cluster in the beginning and starts merging each cluster with the other two clusters which are closest based on current state.

MODELING METHODOLOGY

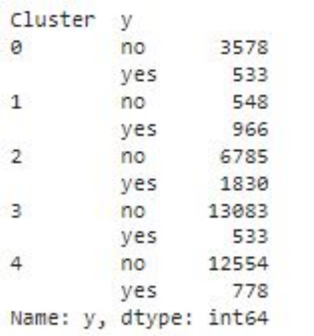
Clustering - K-means

In K-means clustering, the number of clusters is predetermined arbitrarily based on EDA and pre-training analysis. Elbow method is one of the most common methods. You can plot the SSE (or inertia score) and silhouette coefficient against number of clusters and locate where the elbow shows in the graph. In our case, the elbow is not obvious so we need to do experiments and compare the results.

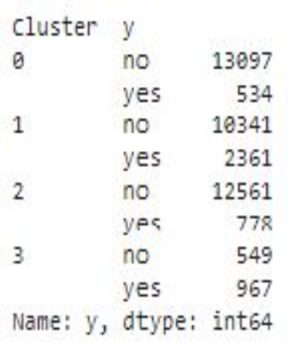




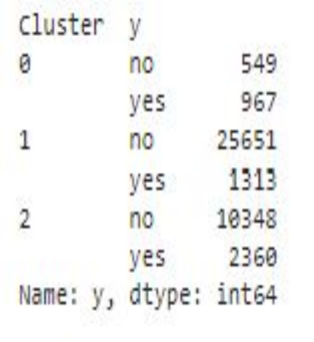
Below are the results from using 3 - 5 clusters in K-means clustering and none of them could separate the classes.



```
Cluster  y
0        no      3578
         yes      533
1        no       548
         yes      966
2        no      6785
         yes     1830
3        no     13083
         yes      533
4        no     12554
         yes      778
Name: y, dtype: int64
```



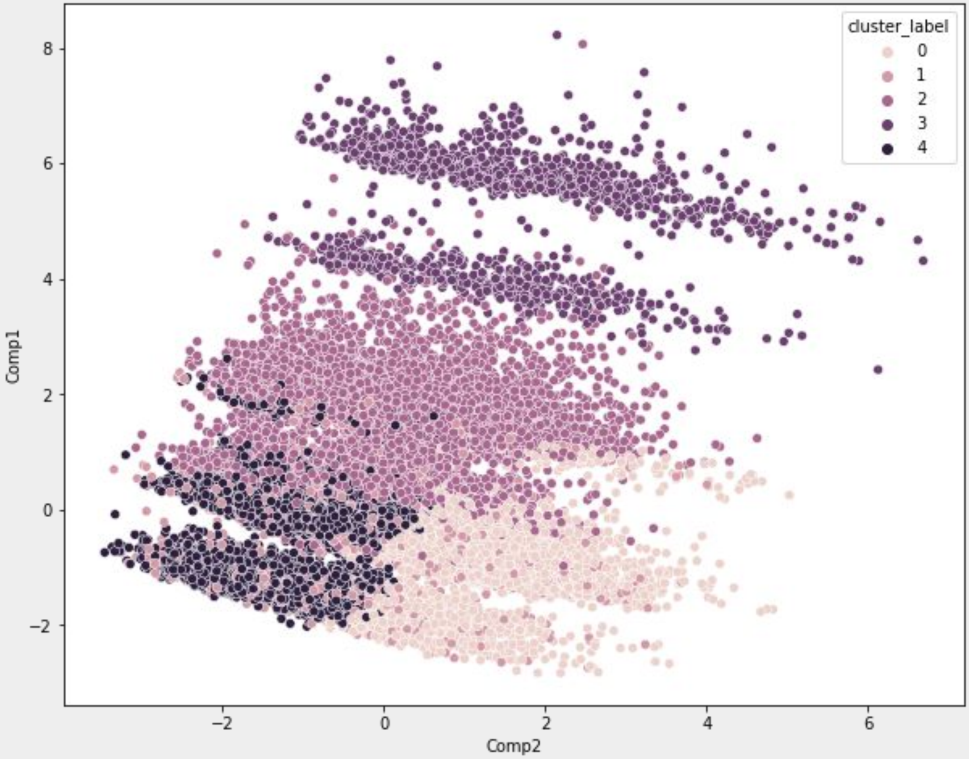
```
Cluster  y
0        no     13097
         yes      534
1        no     10341
         yes     2361
2        no     12561
         yes      778
3        no       549
         yes      967
Name: y, dtype: int64
```



```
Cluster  y
0        no       549
         yes      967
1        no     25651
         yes     1313
2        no     10348
         yes     2360
Name: y, dtype: int64
```

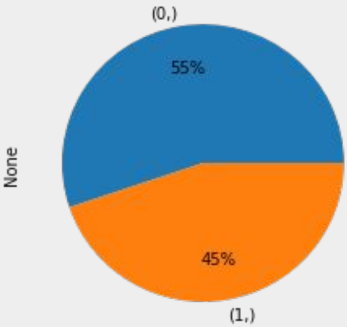
Clustering - K-means

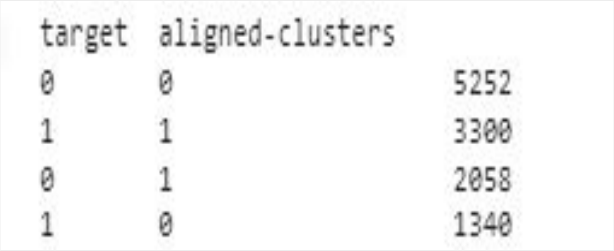
Blow is the clusters (cluster = 5) visualized by plotting the two top two principal components transformed based on original features. Cluster 3 seems separable from other clusters easily while cluster 0, 4 are more overlapped.



Clustering - Agglomerative

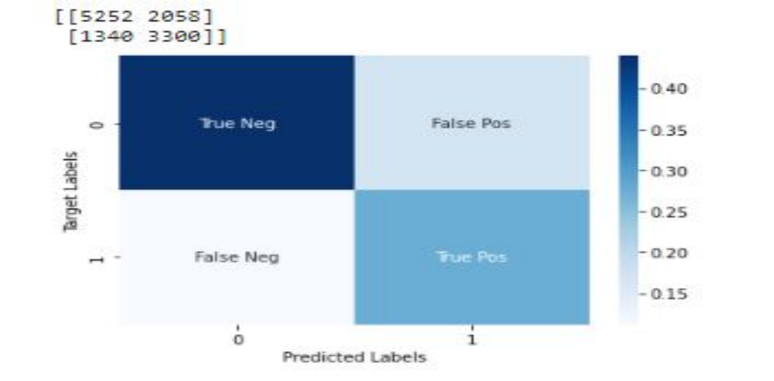
Agglomerative clustering is based on the measurement of similarity between data points and we decided to adapt Gower distance. Due to the limited computation power, the data needs to be reduced. 20% of class 0 data points are randomly sampled to be used in this model. It not only reduce the computation power needed but also make the dataset more balanced





```
target  aligned-clusters
0       0                   5252
1       1                   3300
0       1                   2058
1       0                   1340
```

After aligning the cluster with target variable, we can see that it performs better than logistic regression although clustering is normally not used for prediction.



```
[[5252 2058]
 [1340 3300]]
```

MODELING METHODOLOGY

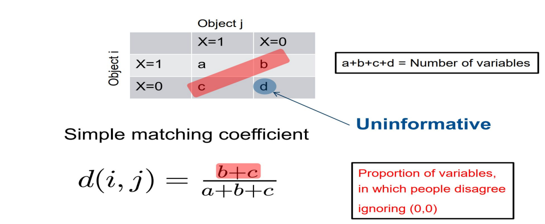
Clustering - Gower Distance

Here is how the Gower distance is calculated:

$$d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$$

The distance for different types of variables are computed as following:

Binary asymmetric variable:



Object i	Object j X=1	Object j X=0
X=1	a	b
X=0	c	d

a+b+c+d = Number of variables

Uninformative

Simple matching coefficient

$$d(i,j) = \frac{b+c}{a+b+c}$$

Proportion of variables, in which people disagree ignoring (0,0)

Nominal variable:

$$d(i,j) = \frac{mm}{p}$$

Ordinal variable:

Rank outcome of variable f=1,2,…,M: r_{if}

Normalize: $z_{if} = \frac{r_{if}-1}{M_f-1}$

Treat z_{if} as interval-scaled

MODELING METHODOLOGY

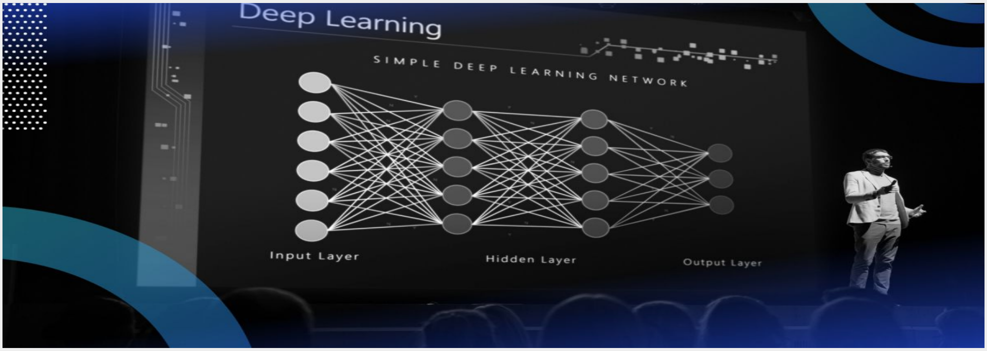
Neural Networks

Deep Learning is a machine learning and artificial intelligence technique that imitates the way humans gain knowledge.

Deep learning has multiple advantages over traditional machine learning algorithms. First, deep learning deals with unstructured data by uncovering the underlying pattern and relationship. Second, it eliminates the need of feature engineering which is a time-consuming task manipulated by data scientists in traditional machine learning practice. Third, deep learning can learn without guidelines, eliminating the demanding requirements on data labeling in traditional machine learning.

The architecture we used here is the deep neural networks, which is a stack of several Dense layers including input, output and hidden layers.

The network is fed forward and fully connected. DNNs can model complex non-linear relationships. DNN architectures generate compositional models where the object is expressed as a layered composition of primitives.



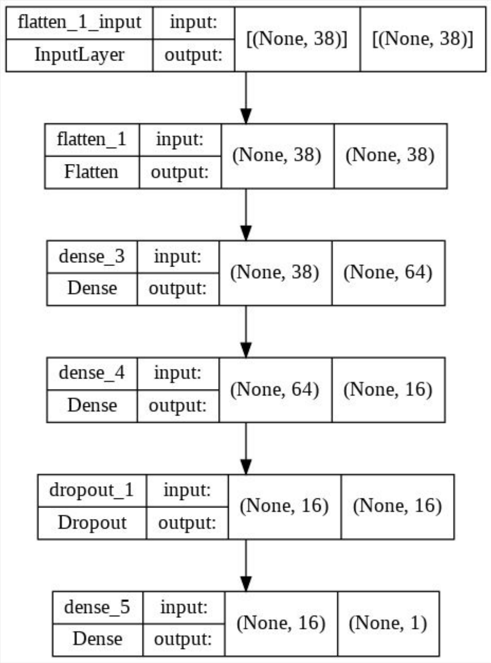
MODELING METHODOLOGY

Neural Networks

The data in our project is structured tabular data with a hybrid of numeric and categorical features, which might not be the ideal sweet spot for neural networks, so we did not intend to set the model architecture complicated to mitigate overfitting issues.

The baseline DNN model is developed with two hidden layers and a typical number of neurons 64 and 16, as shown in the graph. To further reduce overfitting issues, a **Dropout** layer with 0.2 drop ratio is added following the second dense hidden layer.

Additionally, L2 regularization with 0.001 learning rate is added. The kernel initializer for hidden layers is set as ‘he_uniform’,, ‘adamax’ is set as optimizer, and binary_crossentropy is set as loss function.



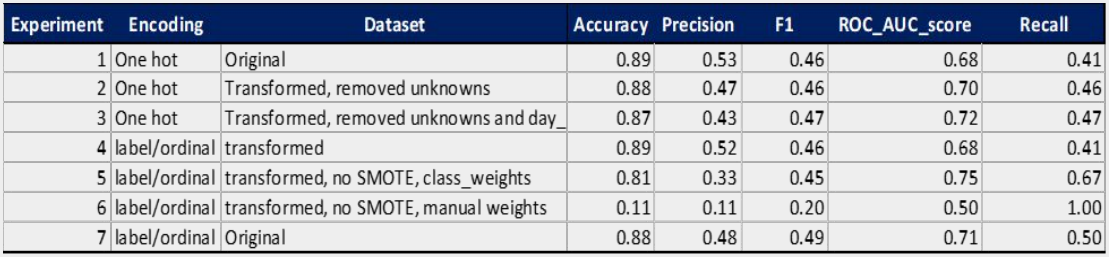
flatten_1_input InputLayer	input: output:	[(None, 38)]	[(None, 38)]
flatten_1 Flatten	input: output:	(None, 38)	(None, 38)
dense_3 Dense	input: output:	(None, 38)	(None, 64)
dense_4 Dense	input: output:	(None, 64)	(None, 16)
dropout_1 Dropout	input: output:	(None, 16)	(None, 16)
dense_5 Dense	input: output:	(None, 16)	(None, 1)

Given the modest complexity of classification tasks, the concentration of experiments is shifting from testing various structures, layers and hyperparameter tuning to validate the impact of different encoding methodology and data balancing techniques, and expected performance improvement of engineered features over raw features.

MODELING METHODOLOGY

Neural Networks

Performance

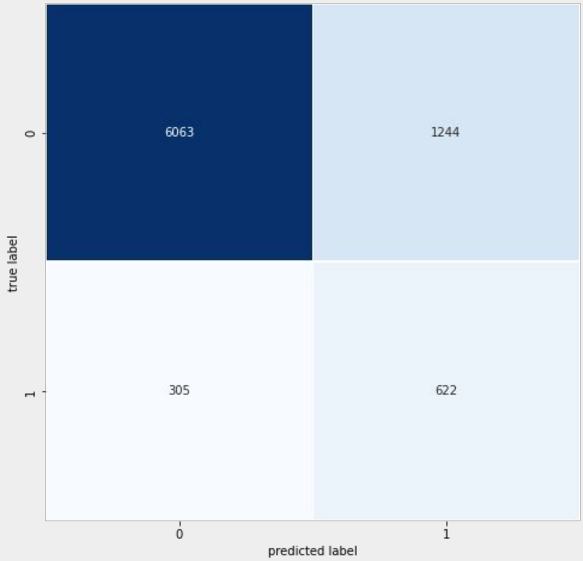


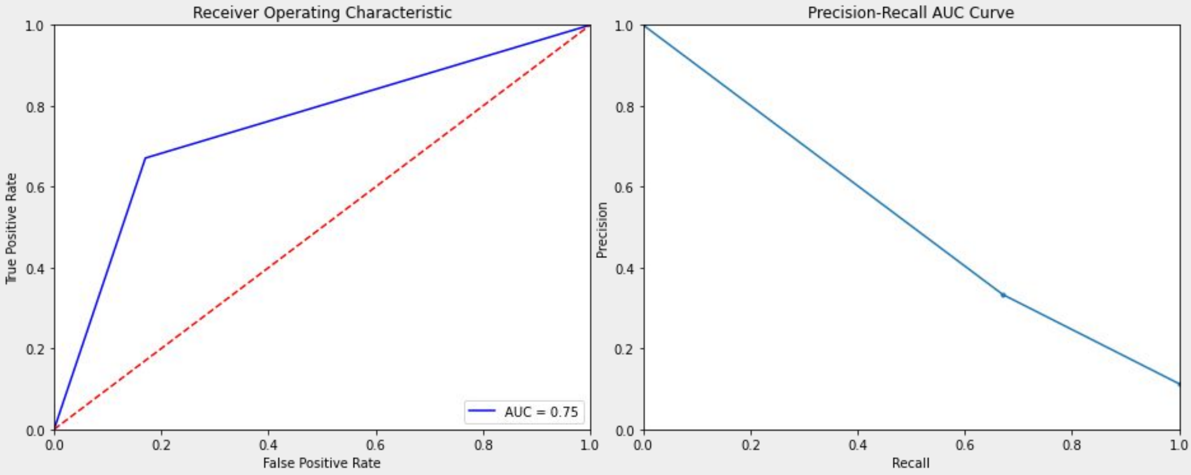
Experiment	Encoding	Dataset	Accuracy	Precision	F1	ROC_AUC_score	Recall
1	One hot	Original	0.89	0.53	0.46	0.68	0.41
2	One hot	Transformed, removed unknowns	0.88	0.47	0.46	0.70	0.46
3	One hot	Transformed, removed unknowns and day_	0.87	0.43	0.47	0.72	0.47
4	label/ordinal	transformed	0.89	0.52	0.46	0.68	0.41
5	label/ordinal	transformed, no SMOTE, class_weights	0.81	0.33	0.45	0.75	0.67
6	label/ordinal	transformed, no SMOTE, manual weights	0.11	0.11	0.20	0.50	1.00
7	label/ordinal	Original	0.88	0.48	0.49	0.71	0.50

To refer back to our business goals and consider the imbalance nature of testing data, the evaluation mainly takes ROC_AUC_score and Recall as assessment criteria. The experiment 6 was teased out as the model was dominated with prediction of minority class (positive) due to exaggerated manual weights set.

Model 5 achieves highest ROC_AUC_score and recall among all models, below are the confusion matrix and ROC_AUC curve of the model. The model set the class weights to the computed result by sklearn.utils class_weight functionality and achieved 0.75 ROC_AUC_score and 0.67 recall score. On the following page are the confusion matrix and ROC curve plots generated:

MODELING METHODOLOGY





Neural Networks

I. Overall, the models trained on label-encoded features (experiments 4, 5, 7) perform slightly better than those trained on one-hot encoded data with higher recall score (0.51 vs 0.45), and similar ROC_AUC_score values (0.70 vs 0.70). Although we had concerns initially about the information bias brought in by the label encoding, the result proves that there might be some ordinal information underlying in the categorical features and could be modeled to help on performance such as education and job occupation of customers.

II. The feature engineering (experiments 2,3,4,5), which mainly involve the consolidation of subcategories of categorical features, slightly improves the performance over the models trained on raw features ((experiment 1 and 7) with higher recall score (0.51 vs 0.46), and a slightly better ROC_AUC_score (0.71 vs 0.70).

III. There are multiple data balancing techniques that could help deal with imbalanced class distribution problems. While SMOTE exposes the training model with more observation of the minor class by oversampling, we found that we could also correct the balance by setting the class_weights parameter in the fitting phase, which achieved better performance. The class weights could be manipulated arbitrarily to either class proportion, or cost-averaged weights which we experimented in model 5, or more extreme order of magnitude for specific business purposes. In this case, given the cost of contacting customers is significantly lower than the cost of losing a customer who is likely to subscribe to the term deposit, it makes sense to put arbitrarily weight on the positive class to penalize the model for missing a true positive case.

SUMMARY OF FINDINGS AND RECOMMENDATIONS

SUMMARY OF FINDINGS

Model Comparison Table

Model performance is displayed in the table below.

Testing Set					
Model/Metrics	Precision %	Recall %	F1 %	AUC %	Accuracy %
Decision Tree	44	53	48	76	88
Random Forest	48	38	42	76	89
Gradient Boosting	46	50	48	78	88
XGradient Boosting	45	49	47	78	88
Logistic Regression	72	18	29	76	90
Neural Network	33	67	45	75	89
AutoML					

From the table above, all models perform within 2% accuracy of each other. Differences arise when observing the precision, recall, and F1 scores. The Neural Network model performs with the highest recall while the Logistic Regression model is performing quite poorly with regards to recall.

SUMMARY OF FINDINGS

Feature Selection - Random Forest

Below are the features identified as having the highest influence on the target variable, based on the Random Forest Model:

Social and Economic Context Attributes

Number of Employees - quarterly indicator

Employment Variation Rate
Consumer Confidence Index
Consumer Price Index

Bank Client Data

Age of the client

Client has no credit in default

Related with the Last Contact of the Current Campaign

Campaign which includes number of contacts performed during this campaign and for this client

SUMMARY OF FINDINGS

Our analysis answers the following business questions:

KEY QUESTION 1: Can Rico Banco marketing predict whether a consumer will respond positively to a telemarketing campaign? Which model is most accurate?

ANSWER: Yes, through the creation of predictive classification models, Rico Banco can predict whether a consumer will convert to a term deposit subscription as a result of a telemarketing campaign. To do so, the bank can identify **the most and least profitable customers** based on consumer demographics, previous campaign history, as well as socioeconomic variables.

KEY QUESTION 2: Which model should be used for marketing prediction?

ANSWER: Our team recommends using the Neural Net model to predict future marketing campaign success. Not only did the neural net model have high accuracy, but it also had the highest recall weight compared with the other models. When using models to identify which customers to target via marketing campaigns, **high recall** is one of the most important measures of success as it is essential not to miss potential customers with high likelihood of subscribing to term deposits. It would ultimately be **more costly** to Rico Banco to miss out on potential customers than it would to spend time and money making marketing outreach to those who will not subscribe at all.

SUMMARY OF FINDINGS

KEY QUESTION 3: Which social and economic features are most correlated with term deposit subscription?

ANSWER: Based on the Random Forest classification model, **Number of Employees, Employment Variation Rate, and the Consumer Confidence Index** features are highly correlated with successful conversion rates. Based on this information, marketing team of Rico Banco should consider the **right timing and macro economic environment** to implement their marketing campaigns for ROI maximization purpose. At the same time, when the social and economic environment becomes unfavorable, they should develop plans such as more granular customer conversion or differentiated marketing campaigns.

KEY QUESTION 4: What about customer segmentation?

ANSWER: From the cluster analysis performed, there are no obvious features/characteristics combinations that could be used to separate the group of customers from customers who do not subscribe the term deposits. The **differences** in statistics between two groups are **not significant**. Future development of this approach could be collecting extra information of the customers and/or include more customers so more dis-similarity could be located between data points.

DASHBOARD

One of the primary goals of this project is to ensure that the marketing and leadership teams at Rico Banco have access to the insights and analytics required to make better informed business decisions about marketing strategy and budget. One of the ways we have enabled this is via a Tableau Dashboard that will help answer the following business questions:

1) **What does the current Rico Banco consumer base look like?**

2) **What are the various socioeconomic factors during campaigns?**

3) **Who is subscribing based on marketing approaches?**

4) **Can we predict marketing campaign success?**

The dashboard currently contains results of the first analysis but will be updated in real time as **more campaigns** are run and additional data points are collected. This will enable to the teams to draw conclusions on success of marketing campaigns in real-time, and make adjustments to strategy as necessary.

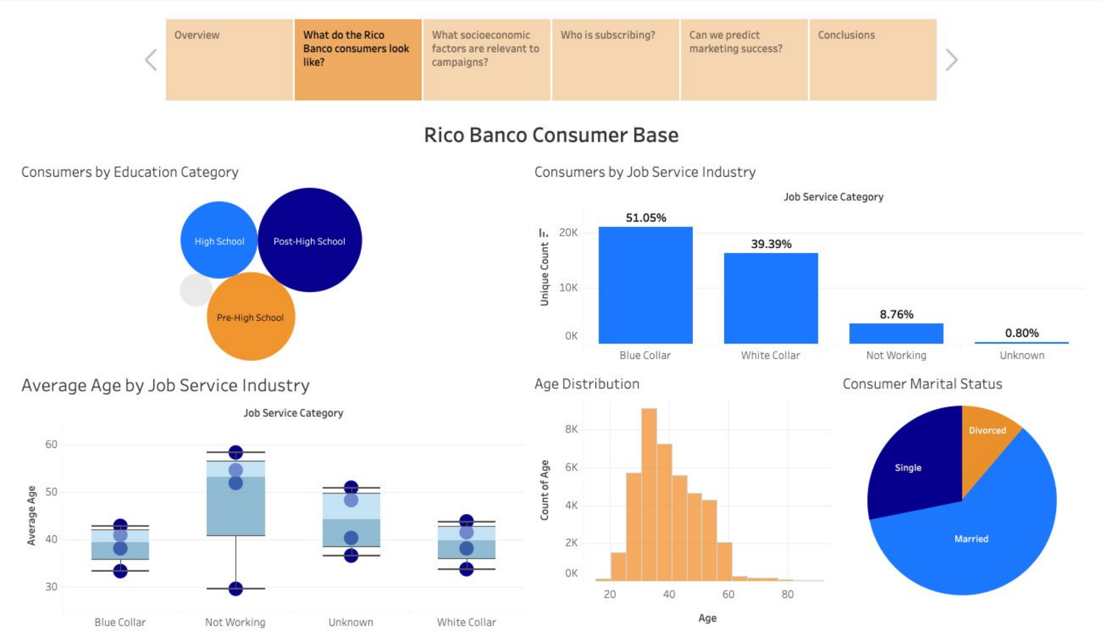
The simple dashboard navigation can be found at the top of the page and is akin to the movement of that of a **cthonic** serpent.

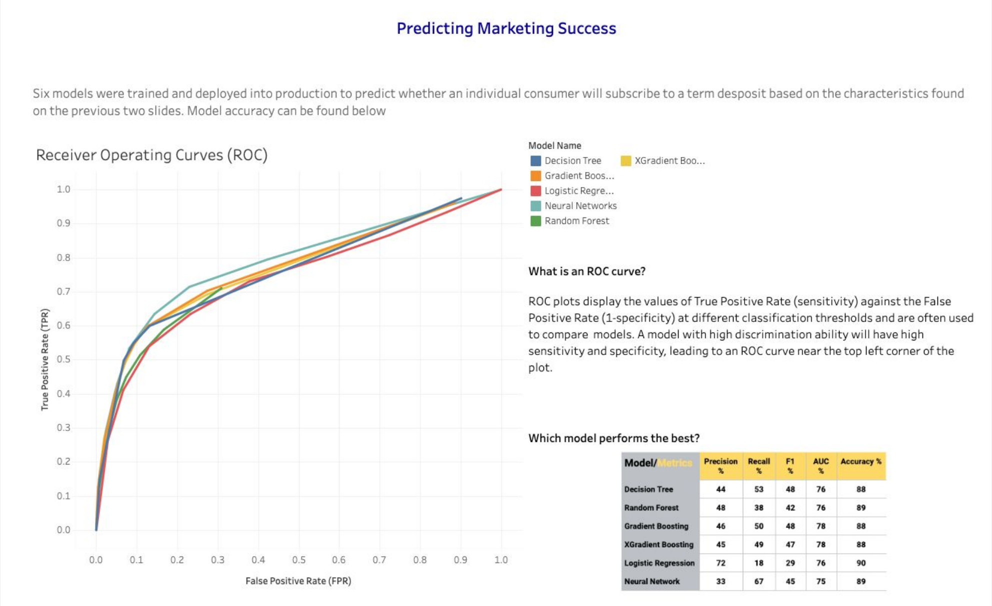
The dashboard can be accessed by anyone on the Rico Banco team by following the link.

https://public.tableau.com/app/profile/lisa.izquierdo/viz/498_project_dashboard/FinalStory?publish=yes

A teaser of the types of information found on the dashboard can be found on the following page.

DASHBOARD





Predicting Marketing Success

Six models were trained and deployed into production to predict whether an individual consumer will subscribe to a term desposit based on the characteristics found on the previous two slides. Model accuracy can be found below

What is an ROC curve?

ROC plots display the values of True Positive Rate (sensitivity) against the False Positive Rate (1-specificity) at different classification thresholds and are often used to compare models. A model with high discrimination ability will have high sensitivity and specificity, leading to an ROC curve near the top left corner of the plot.

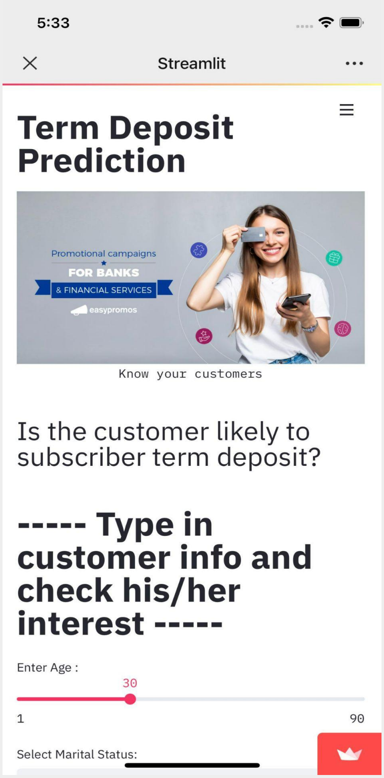
Which model performs the best?

Model/Metrics	Precision %	Recall %	F1 %	AUC %	Accuracy %
Decision Tree	44	53	48	76	88
Random Forest	48	38	42	76	89
Gradient Boosting	46	50	48	78	88
XGradient Boosting	45	49	47	78	88
Logistic Regression	72	18	29	76	90
Neural Network	33	67	45	75	89

WEB APPLICATION

Our web application is built on a responsive web framework that scales not only to all PC browsers, but also all mobile and tablet browsers. The application is built through Streamlit (https://streamlit.io/), which enables fast development and host of ML-based applications.

Streamlit is an open source app framework in the Python language. It helps us create web apps for data science and machine learning purposes in a short amount of time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc. With Streamlit, no callbacks are needed since widgets are treated as variables. Data caching simplifies and speeds up computation pipelines. Streamlit watches for **changes or updates** of the linked Git repository and the application will be deployed **automatically** in the shared link. Here is the link to access to our web application: RicoBancoMarketing



Term Deposit Prediction

Know your customers

Is the customer likely to subscriber term deposit?

----- Type in customer info and check his/her interest -----

Enter Age :

30

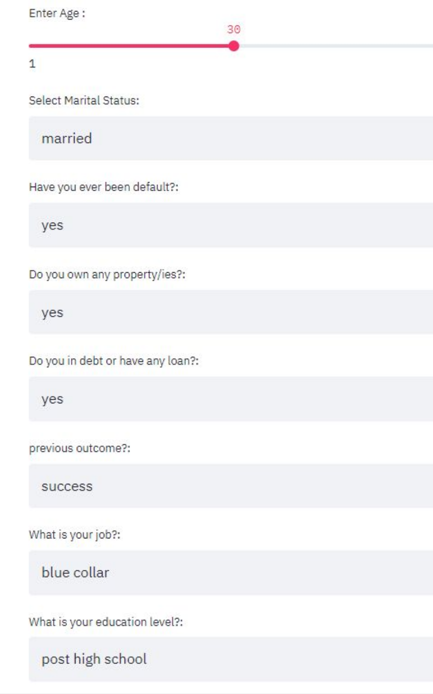
1 90

Select Marital Status:

WEB APPLICATION

Navigation Panel

The navigation panel of the application is where all the inputs are entered. The inputs required include both **customers' information** such as job, education, marital status as well as **social economic information** such as interest rate and Consumer Confidence Index, which are collected by relevant functional teams including marketing and customer relationship management and input into the application.



Enter Age :

30

1

Select Marital Status:

married

Have you ever been default?:

yes

Do you own any property/ies?:

yes

Do you in debt or have any loan?:

yes

previous outcome?:

success

What is your job?:

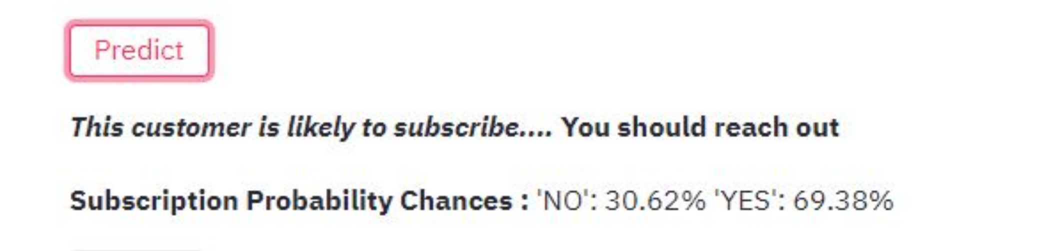
blue collar

What is your education level?:

post high school

Prediction Report

The prediction report is generated after all input is provided and a single click of the **'Predict'** button triggers the pre-trained classification model on the backend. The app will show the possibility of subscription for that specific customer as well as a qualitative recommendation on whether the customer should be approached.



***This customer is likely to subscribe....* You should reach out**

Subscription Probability Chances : 'NO': 30.62% 'YES': 69.38%

FINAL RECOMMENDATIONS

Tying it all together

In this project, we validated a cohesive, end-to-end framework for aggregating, transforming, modeling and visualizing customer marketing data with the goal of increasing market share while minimizing resources through a comprehensive understanding of customer behavior. This foundation should be used by both the marketing and leadership teams at **Rico Banco** as an example of applying predictive analytics and data science tools to enhance the understanding of the clients we serve.

While different models may be deemed more appropriate in the future as demographics change and company **KPIs** continue to evolve, at this time we advise the use of the **Neural Network** model to make predictions on which customers are most likely to respond positively to marketing campaigns.

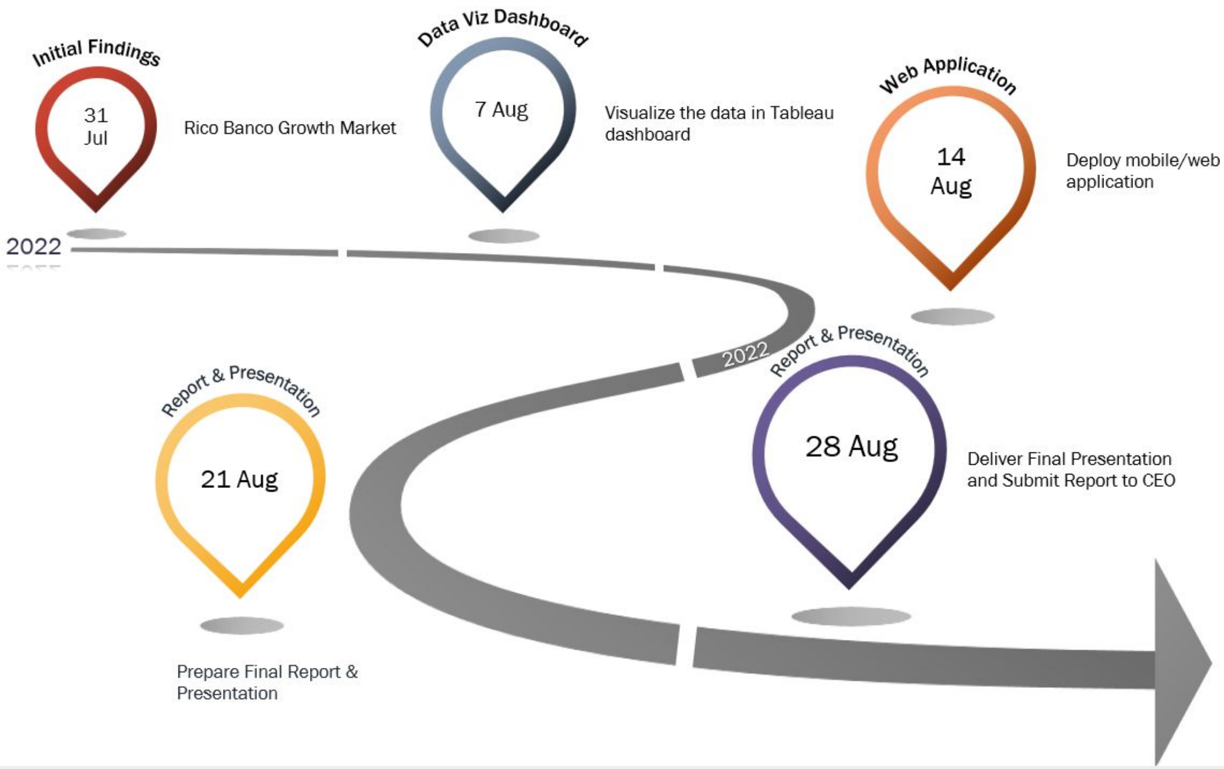
As more campaigns are run and additional data points are collected, the **Tableau dashboard** can be used to gain real-time insights into the demographic makeup of the consumer base as well as to better understand the social and economic factors of the time periods campaigns are run. Additionally, as the marketing team determines which customers would benefit from additional outreach, the **Web Application** can be used to enter the specific information surrounding the customer to produce a prediction as to whether that individual would subscribe so the intelligence gained can be scaled up to production phase. Not only will these actions minimize the cost of unnecessarily marketing to non-subscribers, but they will also keep the overhead operational costs of the team down as marketers can self-serve insight generation without manual intervention from the data science team.

Lastly, always remember this **backronym** BSNH – Bank Smarter, Not Harder!

FINAL RECOMMENDATIONS

What's *next*?

The final step prior to closing out this project from an Analytics perspective is to present the findings to leadership and demonstrate the use of the customized dashboard and application. This will allow our team to gather additional feedback and make any final tweaks to the models or visuals that best align with leadership priorities.


Initial Findings
31 Jul
Rico Banco Growth Market
Data Viz Dashboard
7 Aug
Visualize the data in Tableau dashboard
Web Application
14 Aug
Deploy mobile/web application
2022
2022
Report & Presentation
21 Aug
Prepare Final Report & Presentation
Report & Presentation
28 Aug
Deliver Final Presentation and Submit Report to CEO

APPENDIX

REFERENCES

AdaBoost Algorithm - A Complete Guide for Beginners - Analytics Vidhya. (2021, September 15). Analytics Vidhya; www.analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

An Introduction to Gradient Boosting Decision Trees - Machine Learning Plus. (2021, June 12). Machine Learning Plus; www.machinelearningplus.com. https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/

Applied Multivariate Statistics Spring 2012 - ETH Z. https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v4.1.pdf

Brownlee, J. (2020, November 22). Extreme Gradient Boosting (XGBoost) Ensemble in Python - Machine Learning Mastery. Machine Learning Mastery; machinelearningmastery.com. https://machinelearningmastery.com/extreme-gradient-boosting-ensemble-in-python/

Decision Trees: Complete Guide to Decision Tree Analysis. (2019, December 10). Explorium; www.explorium.ai. https://www.explorium.ai/blog/the-complete-guide-to-decision-trees/

Decision tree - Wikipedia. (2021, December 5). Decision Tree - Wikipedia; en.wikipedia.org. https://en.wikipedia.org/wiki/Decision_tree

Khandelwal, N. (2020, July 16). A Brief Introduction to XGBoost. Extreme Gradient Boosting with XGBoost! | by Neetika Khandelwal | Towards Data Science. Medium; towardsdatascience.com. https://towardsdatascience.com/a-brief-introduction-to-xgboost-3eaee2e3e5d6

Lendave, V. (2022, January 27). How can SMOTE technique improve the performance of weak learners? Analytics India Magazine; analyticsindiamag.com. https://analyticsindiamag.com/how-can-smote-technique-improve-the-performance-of-weak-learners/

Random Forest. Introduction to Random Forest Algorithm. (2021, June 17). Analytics Vidhya; www.analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

sklearn.ensemble.RandomForestClassifier. (2000, January 1). Scikit-Learn; scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Disclaimer : Don’t be ***begrutten****, all images were searched for on the internet to try and procure those without copyright attached to their use. We have not knowingly infringe any copyright and all images used herein are only for private use, not for publication.*