Capstone Project
Bank Marketing Effectiveness Prediction
(Classification)

(Prayag Raj Dubey)
Data science student
Cohort- Hudson, Alma Better

Introduction

The older marketing options have contributed minimally in increasing the business of banks. Due to internal competition and financial crisis European Banks were under pressure to increase their financial assets. They offered long term deposits with good interest rates to the people using direct marketing strategy but contacting many people takes a lot of time and success rate is also less. So they want to take help from the technology to come up with a solution that increases the efficiency by making fewer calls but improves the success rate.Portuguese Banking Institution has provided the data related to marketing campaigns that took over phone calls.

Objective

Analysis of organization's marketing data is one of the most typical applications of data science and machine learning. Such analysis will definitely be a nice contribution to the portfolio.

In general, datasets can be used for 2 different business goals:-

- Prediction of the results of the marketing campaign for each customer and clarification of factors which affect the campaign results. This helps to find out ways to make marketing campaigns more efficient.

- Finding out customer segments, using data for customers, who subscribed to term deposit. This helps to identify the profile of a customer, who is more likely to acquire the product and develop more targeted marketing campaigns.

Problem Statement

The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

ATTRIBUTE INFORMATION

Input variables:

Bank Client data:

- **age** (numeric)
- **job** : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- **marital** : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- **education** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- **default**: has credit in default? (categorical: 'no','yes','unknown')
- **housing**: has a housing loan? (categorical: 'no','yes','unknown')
- **loan**: has a personal loan? (categorical: 'no','yes','unknown')

Related with the last contact of the current campaign:

- **contact**: contact communication type (categorical: 'cellular','telephone')
- **month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- **day_of_week**: last contact day of the week
(categorical: 'mon','tue','wed','thu','fri')
- **duration**: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other attributes:

- **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **previous**: number of contacts performed before this campaign and for this client (numeric)
- **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Output variable (desired target):

- **y** - has the client subscribed to a term deposit? (binary: 'yes','no')

Steps involved

- **Loading and discovering data**

  Now, we need to load our data from the external source, which in this case is uploaded to the drive. The data is in the format of the CSV (Comma Separated Values) file.

- **Data Cleaning**

  Data cleaning is an important step in the data analytics process in which you either remove or update information that is incomplete or improperly formatted.

- **Null values Treatment by different methods**

  We have our dataset in hand which is raw and unfiltered. This step involves cleaning our data first by eliminating the columns which are not needed for our analysis. We have around **45211 rows × 17 columns** in our dataset. Since there were no null values, we have left it untouched.

- **Exploratory Data Analysis**

  Exploratory Data Analysis is the approach of analyzing data, gathering and summarizing the important characteristics of the information, and using simple visualization that makes it easier to understand.

- **Importing necessary modules and libraries**

  We are importing the following libraries for their respective applications:

Pandas:- Pandas are used to analyze data. It has functions for analyzing, cleaning, exploring, and manipulating data.

Matplotlib:- Matplotlib is a graph plotting library in python that serves as a visualization utility. Most of the Matplotlib utilities lie under the pyplot submodule.

Numpy:- NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices.

SciKit:- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Plotly:- The plotly is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

Seaborn:- Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

Datetime:- The Datetime module supplies classes for manipulating dates and times. While date and time arithmetic is supported, the focus of the implementation is on efficient attribute extraction for output formatting and manipulation.

Statsmodels:- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

- **Plotting various graphs for different parameters.**
- **Finding the key facts and relationships between various parameters.**
- **Observations according to the outputs of the graph visualizations.**

Data information:

Bike Sharing Demand Prediction:

Statistical information contained in the Bank Marketing Effectiveness Prediction Database is based on reports from a variety of open media sources. Information is not added to the database unless and until we have determined the sources are credible.

Multicollinearity:

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

VIF(Variable Inflation Factors):

VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable.

Outlier Detection:

An outlier detection technique (ODT) is used to detect anomalous observations/samples that do not fit the typical/normal statistical distribution of a dataset. Simple methods for outlier detection use statistical tools, such as boxplot and Z-score, on each individual feature of the dataset.

Feature Engineering:

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

Performance Metrics:

Different performance metrics are used to evaluate machine learning models. Based on our task we can choose our performance metrics. Since our task is classification and that too binary class classification, whether clients will or will not subscribe for deposits.

Here we will be using AUC ROC

ROC also known as Receiver Operating Characteristics, shows the performance of binary class classifiers across the range of all possible thresholds plotting between true positive rate and 1-false positive rate.

AUC measures the likelihood of two given random points, one from positive and one from negative, the classifier will rank the positive points above negative points. AUC-ROC is a popular classification metric that presents the advantage of being independent of false positive or negative points.

Secondary Performance Metrics

Macro-F1 Score: F1 score is the harmonic mean between Precision and Recall. Macro F1 score is used to know how our model works in the overall dataset.

Confusion Matrix: This matrix gives the count of true negative, true positive, false positive and false negative data points.

Train and Test Split:

The train-test split is used to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms/Applications.

Handling Imbalanced Dataset

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. It is vital to identify the minority classes correctly. So the model should not be biased to detect only the majority class but should give equal weight or importance towards the minority class too.

Optimization:

Function optimization is the problem of finding the set of inputs to a target objective function that result in the minimum or maximum of the function.
It can be a challenging problem as the function may have tens, hundreds, thousands, or even millions of inputs, and the structure of the function is unknown, and often non-differentiable and noisy.

Classification:

Classification is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the Categorical output variable based on the one or more predictor variables.

Logistic Regression Classifier

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Logistic Regression is used when the dependent variable(target) is categorical. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

Typical properties of the logistic regression equation include:

- Logistic regression's dependent variable obeys 'Bernoulli distribution'
- Estimation/prediction is based on 'maximum likelihood.'
- Logistic regression does not evaluate the coefficient of determination (or R squared) as observed in linear regression'. Instead, the model's fitness is assessed through a concordance.

Decision Tree Classifier

Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

Decision trees are upside down which means the root is at the top and then this root is split into various several nodes. Decision trees are nothing but a bunch of if-else statements in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.

In a Decision Tree diagram, we have:

Root Node: The first split which decides the entire population or sample data should further get divided into two or more homogeneous sets. In our case, the Outlook node.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: This node decides whether/when a sub-node splits into further sub-nodes or not. Here we have Outlook node, Humidity node, and Windy node.

Leaf: Terminal Node that predicts the outcome (categorical or continuous value). The coloured nodes, i.e., Yes and No nodes, are the leaves.

Random Forest Classifier

Random Forest is a technique that uses ensemble learning that combines many weak classifiers to provide solutions to complex problems.

As the name suggests, a random forest consists of many decision trees. Rather than depending on one tree it takes the prediction from each tree and based on the majority votes of predictions, predicts the final output.

Random forests use the bagging method. It creates a subset of the original dataset, and the final output is based on majority ranking and hence the problem of overfitting is taken care of.

K-nearest Neighbors Classifier

KNN which stands for K-Nearest Neighbors is a simple algorithm that is used for classification and regression problems in Machine Learning. KNN is also non-parametric which means the algorithm does not rely on strong assumptions instead tries to learn any functional form from the training data.

Unlike most of the algorithms with complex names, which are often confusing as to what they really mean, KNN is pretty straight forward. The algorithm considers the k nearest neighbors to predict the class or value of a data point.

The kNN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of K number of neighbors
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

Naive Bayes Classifier

The Naïve Bayes algorithm is a classification technique based on the Bayes' Theorem which assumes there is independence between the features. We interfere with applications utilizing this algorithm on a daily basis, for example it powers recommendation systems for streaming applications or ads on social media as well as many online retail websites.

The Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. Let's have a look under the hood of this major classifier.

The simple form of the calculation for Bayes Theorem is as follows:

P(A|B) = P(B|A) * P(A) / P(B)

Support Vector Machine Classifier

SVM is a powerful supervised algorithm that works best on smaller datasets but on complex ones. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks, but generally, they work best in classification problems.

Support Vectors: These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.

Margin: it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins: hard margin and soft margin.

LightGBM Classifier

LightGBM is a gradient boosting ensemble method that is used by the Train Using AutoML tool and is based on decision trees. As with other decision tree-based methods, LightGBM can be used for both classification and regression. LightGBM is optimized for high performance with distributed systems.

LightGBM creates decision trees that grow leaf wise, which means that given a condition, only a single leaf is split, depending on the gain. Leaf-wise trees can sometimes overfit especially with smaller datasets. Limiting the tree depth can help to avoid overfitting.

Data Visualization Analysis

Data Visualization is the process of analyzing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data.

Correlation Heat Map

Analysis of the relation between the various columns of the cleaned data through the Correlation Heat map Matrix.
A correlation heat map is a graphical representation of a correlation matrix representing the correlation between different variables.

Box Plot:

A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile [Q1], median, third quartile [Q3] and “maximum”). It can tell you about your outliers and what their values are.

Line Chart:

A line chart is a graphical representation of an asset's historical action that connects a series of data points with a continuous line.

Horizontal Bar Chart:

A horizontal bar chart is a graph in the form of rectangular bars. The length of these bars is proportional to the values they represent.

Count Plot:

The countplot is used to represent the occurrence(counts) of the observation present in the categorical variable. It uses the concept of a bar chart for the visual depiction.

Distplot:

The distplot represents the univariate distribution of data i.e. data distribution of a variable against the density distribution. The seaborn. distplot() function accepts the data variable as an argument and returns the plot with the density distribution.

Subplot:

A subplot is a narrative thread that is woven through a book to support the elements of the main plot.

Bar Chart:

A bar chart is a chart that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. A bar chart is sometimes called a column chart.

Conclusion:

- The 2nd quarter of the year has the highest number of subscriptions & Month of May has the maximum subscriptions.
- Blue-collar, management and technician showed maximum interest in subscription.
- Compared to married and single, Divorced people have less interest in term deposits.
- People with secondary education followed by tertiary education were subscribed to term deposit.
- Generally people who don't have credit in default are interested in a deposit. Majority of the people have a home loan but only few of them opted for a term deposit.
- Cellular communication is seen as more effective in comparison to other communication types.
- The calls with large duration have more tendency for conversion. Majority of people were not contacted previously before this campaign.
- We can choose KNN or Decision Tree to predict Effectiveness as both of them are showing the same accuracy of 88% & F1- Score of - (0.8824).

References:

- **Python Pandas Documentation:**

  https://pandas.pydata.org/pandas-docs/stable

- **Python Numpy Documentation:**

  https://numpy.org/doc/

- **Python MatPlotLib Documentation:**

  https://matplotlib.org/stable/index.html

- **SciKit Documentation:**

  https://scikit-learn.org/stable/

- **Towards Data Science**:

  https://towardsdatascience.com