



Final Project

BANK MARKETING
TO PREDICT IF A CLIENT WILL SUBSCRIBE TO A TERM DEPOSIT

by

Jael Kruthi Battana
Aditya Kamishetty
Yashashwin Konduru

University of Maryland Baltimore County
Course: DATA 603 - Platforms for big data processing
Professor: Akshatha Kishore Moharir
December 13, 2022

PROBLEM STATEMENT:

We are predicting if a client or customer of a bank is going to subscribe for the term deposit based on the census, personal details, financial details and other marketing campaign details that have been collected of that particular person.

ABSTRACT:

We are uploading the well-known marketing bank dataset for the first time to the UCI Machine Learning Collection. The dataset contains details on a financial institution's marketing campaign, which we will examine in order to identify potential strategies for the bank's marketing initiatives in the future. Gaining the maximum value from a particular data collection is a challenging task since it requires a thorough analysis of its many properties and accompanying values. In order to uncover hidden patterns, the process is typically accomplished by presenting data in a visual way through bar graphs and plots. In this study, a bank's direct marketing data collection is subjected to several visualization approaches.

The data set that was downloaded from the website of the UCI machine learning repository is imbalanced. As a result, several oversampling techniques are employed to improve the predictability of a client's subscription to a term deposit. The effectiveness of the visualization is evaluated by examining the impact of various classifier performance on oversampling strategies. With various feature engineering techniques, we are using different Ensemble Models to predict the outcome.

INTRODUCTION:

Bank Marketing in simple words, is the design structure, layout and delivery of customer-needed services worked out by checking out the corporate objectives of the bank and environmental constraints.

Marketing is important for banks because the banking sector plays a key role in the development of the economy. Banks are now giving importance to the marketing activities to create awareness regarding their services to the public. Customer satisfaction is important, that's why banks are introducing new instruments and ways to attract customers.

Market analysis, data mining, and financial data analysis all require data visualization. It refers to the application of interactive, computer-supported visual representation to enhance cognition and communicate complex ideas underneath data. Charts, graphs, and design components effectively carry out this strategy. These techniques are frequently used by managers and knowledge workers to uncover information hidden in large amounts of data [4] and arrive at the best conclusions. The use of data visualization by decision-makers and their companies has several advantages [2], including helping people learn information in fresh, useful ways. Identifying and

responding to emerging trends can be aided by visualizing the connections and patterns between operational and business activity.

ABOUT DATA SET:

- The information relates to telephone-based direct marketing activities of a Portuguese banking institution. The classification's objective is to foretell whether a client will sign up for a term deposit (variable y).
- The information relates to direct marketing initiatives run by a Portuguese bank. On phone conversations, the marketing efforts were based. In order to determine if the product (term deposit deposit) would be subscribed ('yes') or not ('no'), it was sometimes necessary to make more than one contact with the same client.
- Data consists of 21 features and 41189 entries including personal and professional information of the Client
- **Dataset Link:** <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#>

ATTRIBUTE INFORMATION:

1. Input variables:

Bank client data:

1 - age (numeric)

2 - job : type of job (categorical)

values: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'

3 - marital : marital status (categorical)

Values : 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed

4 - education (categorical)

values: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'

5 - default: has credit in default? (categorical)

Values: 'no', 'yes', 'unknown'

6 - housing: has a housing loan? (categorical)

Values: 'no', 'yes', 'unknown'

7 - loan: has a personal loan? (categorical)

Values: 'no', 'yes', 'unknown'

Related with the last contact of the current campaign:

8 - contact: contact communication type (categorical)

Values: 'cellular', 'telephone'

9 - month: last contact month of year (categorical)

values: 'jan', 'feb', 'mar', ..., 'nov', 'dec'

10 - day_of_week: last contact day of the week (categorical)

values: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric).

Important information: If duration=0, the output goal will be "no," for example. However, the length is unknown before making a call. After the call is over, y is also clearly recognised. Thus, if the goal is to build a realistic predictive model, this input should only be used for benchmarking purposes and should be removed.

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - 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)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical)

values: 'failure','nonexistent','success'

Social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

2. Target variable:

21 - y - Has the customer signed up for a term deposit? (binary: 'yes','no')

LITERATURE REVIEW:

1. Deal Banking Marketing Campaign Dataset with Machine Learning

Link: <https://medium.com/@nutanbhogendrasharma/deal-banking-marketing-campaign-dataset-with-machine-learning-9c1f84ad285d>

Referred to this blog where they used data related to marketing campaigns (phone calls) of a Portuguese banking institution. The classification's objective is to foretell whether a client will sign up for a term deposit (variable y). We will use different types of model and see which model gives the highest accuracy.

2. Analysis on Bank Marketing Campaign for Portuguese Bank

Link: <https://github.com/HegdeChaitra/Bank-Marketing-Campaign-Analysis>

The project's main objective was to analyze a Portuguese bank's previous marketing campaigns using various machine learning techniques, including Random Forests, Logistic Regression, Gradient Boosting, Decision Trees, and AdaBoost, and predict whether or not the user would purchase the bank's term deposit.

3. Recommendation on the marketing team, ways to better target customers using feature importance maps and business intuition.

Paper: “Marketing analytics: Methods, practice, implementation, and links to other fields,” Expert Systems with Applications, 2018. S. L. France and S. Ghose

The research in this paper examined the theoretical underpinnings of marketing analytics, a vast discipline that emerged from operations research, marketing, statistics, and computer science. One of the difficulties in doing a direct marketing analysis, they claimed, is forecasting consumer behavior. They also covered customer relationship management, multidimensional scaling, correspondence analysis, and latent Dirichlet allocation as big data visualization techniques for the marketing sector (CRM). They discussed the relative value of geographic visualization for retail location research and the overall trade-off between its customary methods and art. They also expanded on discriminant analysis as a method for marketing forecasting. Techniques including ensemble learning, feature reduction, and extraction are used in discriminant analysis. These methods address issues including buying, rating, loyalty etc.

4. Paper: S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, “Customer profiling using classification approach for bank telemarketing,” JOIV: International Journal on Informatics Visualization, vol. 1, no. 4-2, pp. 214–217, 2017.

Additionally, this article used the same data set for additional consumer profiling objectives. On the expanded version of the data set analyzed in the current work, naive Bayes, random forests, and decision trees were employed. Prior to assessing the classifiers, preprocessing and normalization were conducted. To conduct the trials and assessment procedures, RapidMiner was employed. Using a prior normalization approach, they demonstrated how each classifier's parameters might be adjusted. They also demonstrated how these parameter values affected recall, accuracy, and precision. Decision trees are the best classifier for consumer profile and behavior prediction, according to their findings.

PROJECT DESIGN AND ARCHITECTURE:

Pre-processing to be done Before Modeling:

- Looking for NA values
- Train Test Split
- Label encoding
- Choosing relevant variables
- Up sampling Data

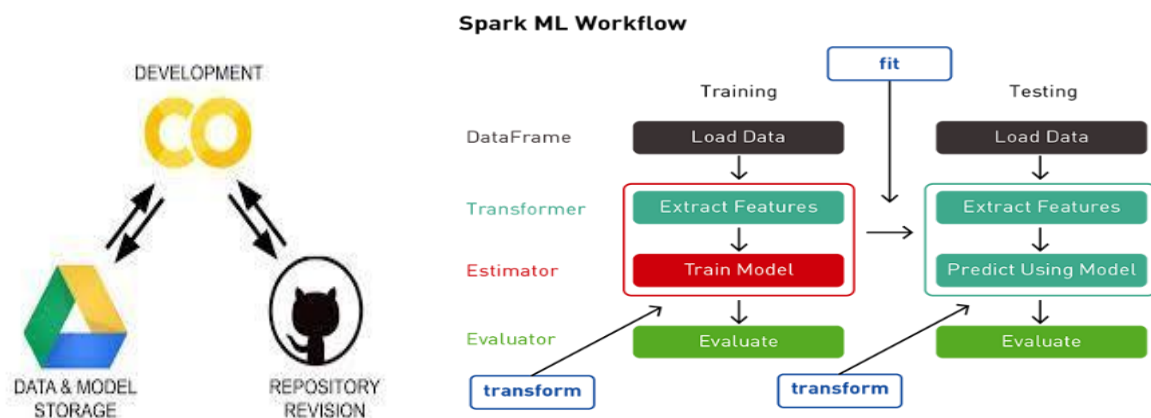
- Vectorizing the data

Overall Process:

- Visualization and insights from the data based on different factors and variables.
- Data cleaning and Feature Engineering.
- Building Different ML models viz. Logistic, Random Forest, GB etc. and evaluate the models.
- Tuning the Model using different hyperparameters and Grid Search and Cross Validation.

Architecture and Technology

- We use Python for Data Processing and Modelling
- The system OS we have used is Linux.
- Development Environment (IDE) on Colab Notebook.
- User → Google Linux Server → Google Storage Bucket → Processing on Google Server.



EXPLORATORY DATA ANALYSIS (EDA) and VISUALIZATION:

1. Overview of the data, its attributes and values and data types that are mostly numeric or integers and categorical or strings.

	0	1	2	3	4
age	56	57	37	40	56
job	housemaid	services	services	admin.	services
marital	married	married	married	married	married
education	basic.4y	high.school	high.school	basic.6y	high.school
default	no	unknown	no	no	no
housing	no	no	yes	no	no
loan	no	no	no	no	yes
contact	telephone	telephone	telephone	telephone	telephone
month	may	may	may	may	may
day_of_week	mon	mon	mon	mon	mon
duration	261	149	226	151	307
campaign	1	1	1	1	1
pdays	999	999	999	999	999
previous	0	0	0	0	0
poutcome	nonexistent	nonexistent	nonexistent	nonexistent	nonexistent
emp.var.rate	1.1	1.1	1.1	1.1	1.1
cons.price.idx	93.994	93.994	93.994	93.994	93.994
cons.conf.idx	-36.4	-36.4	-36.4	-36.4	-36.4
euribor3m	4.857	4.857	4.857	4.857	4.857
nr.employed	5191	5191	5191	5191	5191
y	no	no	no	no	no

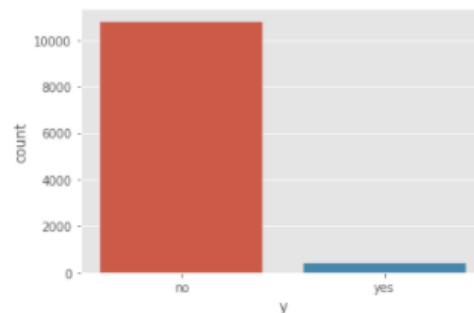
```

root
|-- age: integer (nullable = true)
|-- job: string (nullable = true)
|-- marital: string (nullable = true)
|-- education: string (nullable = true)
|-- default: string (nullable = true)
|-- housing: string (nullable = true)
|-- loan: string (nullable = true)
|-- contact: string (nullable = true)
|-- month: string (nullable = true)
|-- day_of_week: string (nullable = true)
|-- duration: integer (nullable = true)
|-- campaign: integer (nullable = true)
|-- pdays: integer (nullable = true)
|-- previous: integer (nullable = true)
|-- poutcome: string (nullable = true)
|-- emp.var.rate: double (nullable = true)
|-- cons.price.idx: double (nullable = true)
|-- cons.conf.idx: double (nullable = true)
|-- euribor3m: double (nullable = true)
|-- nr.employed: double (nullable = true)
|-- y: string (nullable = true)

```

- Showing information about imbalance between the data and target variable. Therefore, there is a need to use Perform up sampling techniques.

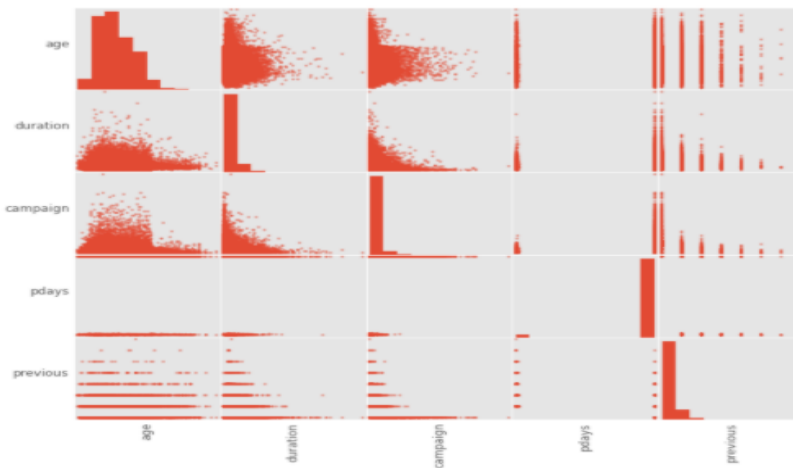
	y	count
0	no	36548
1	yes	4640



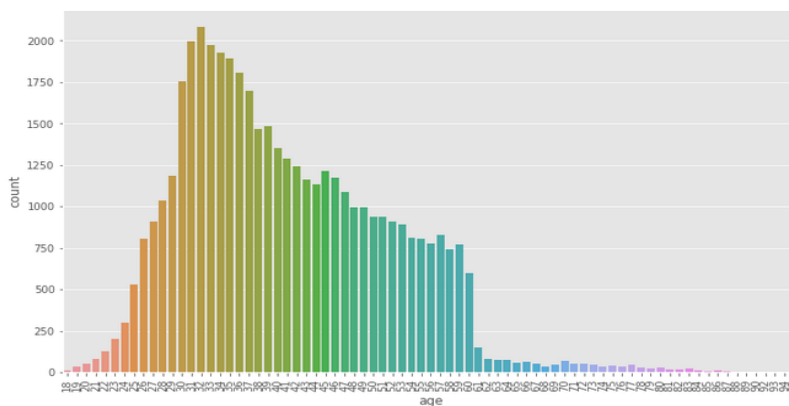
- Basic statistics: Here we can see some basic statistical aggregations of the numerical columns like Age of the customers are between 18-95,etc

	0	1	2	3	4
summary	count	mean	stddev	min	max
age	41188	40.02406040594348	10.421249980934043	17	98
duration	41188	258.2850101971448	259.27924883646494	0	4918
campaign	41188	2.567592502670681	2.7700135429023245	1	56
pdays	41188	962.4754540157328	186.9109073447411	0	999
previous	41188	0.17296299893172767	0.4949010798392903	0	7


4. Some Visualizations:

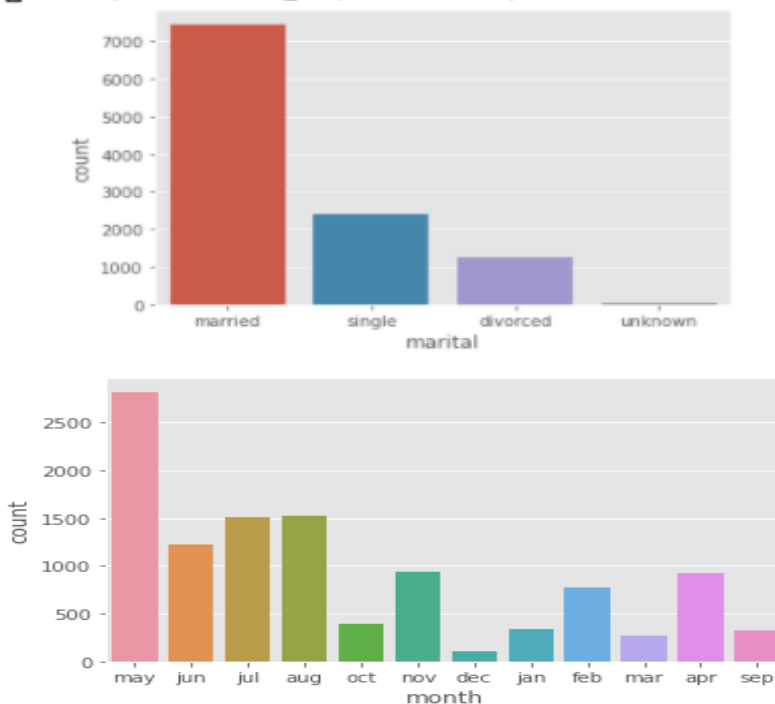


There are no obvious highly correlated variable pairs so we are removing day and month, but keeping the rest of the variables. It is 14 input/predictor variables, 1 response/target variable.



By looking at the distribution, we can see that distribution is rightly skewed and most of the customers are between the age of 25-60.

 <matplotlib.axes._subplots.AxesSubplot at 0x7fbac89f4630>



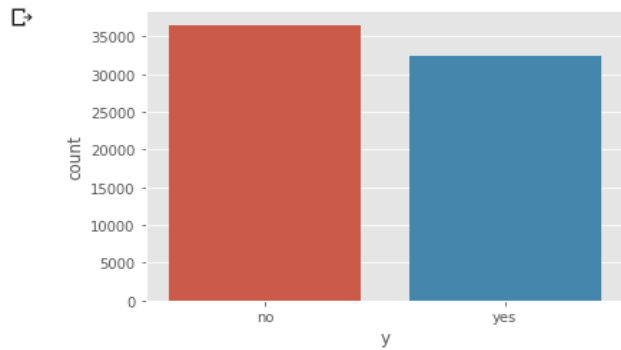
From the graphs, we can see that most of the people are married and from the next bar graph we can see that most marketing campaigns have occurred between the months of May to August.

Data Pre-processing and cleaning:

- Looking for NA values
- Train Test Split
- Label encoding
- Choosing relevant variables
- Up sampling Data
- Vectorizing the data

For Preparing the data below Operations were performed to standardize the data to understand by ML models

1. Category Indexing
2. One-Hot Encoding
3. VectorAssembler - a feature transformer that merges multiple columns into a vector column.



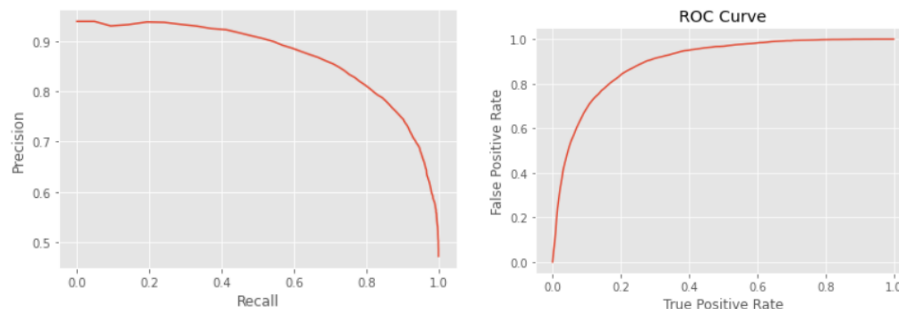
	y	count
0	no	36548
1	yes	32480

After upsampling the data.

MACHINE LEARNING MODELS USED:

Primary Models :

Logistic Regression: 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).



Both the above Graphs suggest that models have understood the data. With Test Area Under ROC after the Cross validation 0.8958125339684428.

Decision Tree: A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

- easy to interpret
- can handle categorical features
- can handle multiclass classification
- feature scaling not necessary
- can capture non-linearities and feature interactions

We can see the Roc: Test Area Under ROC: 0.6680978913072564

Ensemble Models:

Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression).

We can see the Roc: Test Area Under ROC: 0.8855992866345412

Gradient Boosting: Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

We can see the Roc: Test Area Under ROC: 0.9032544089319515

- Gradient-boosted Tree achieved the best result at ROC-AUC Score of 0.89
- Let's try tuning this model with the ParamGridBuilder and the CrossValidator.
- We can use explainParams() to print a list of all params and their definitions to understand what parameters are available for tuning.

Also Tuned Gradient Boosted Tree Model with ParamGridBuilder and CrossValidator

Best results: Xgboost

Boosting is an example of iterative learning, which implies that after making an initial prediction, the model will analyze its errors as a predictive worker and assign more weight to the sample points where it was incorrect in the subsequent iteration. After the loop, it analyzes its incorrect predictions once more and provides additional weight to the sample points that were incorrectly anticipated in the subsequent iteration. This circle of events keeps going.

One of the most significant distinctions between XGBoost and Random Forest is that, while decreasing the cost of a model, XGBoost always prioritizes functional space, whereas Random Forest tries to prioritize hyperparameters in order to optimize the model.

It is this algorithm's effectiveness, accuracy, and viability. It features both tree learning methods and linear model solvers. Therefore, its ability to perform parallel processing on a single computer is what makes it quick.

CONCLUSION:

After ParamGrid and Cross Validation, we can see that ROC Score for Boosting Algorithm is slightly higher than the other as it converts weak learners to strong learners. Therefore, we can use the Gradient Boosting with the Best Params which we have found using Grid Search and validated using Cross Validation.

REFERENCES:

[1] "Marketing analytics: Methods, practice, implementation, and links to other fields," Expert Systems with Applications, 2018. S. L. France and S. Ghose,

[2] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer profiling using classification approach for bank telemarketing," JOIV: International Journal on Informatics Visualization, vol. 1, no. 4-2, pp. 214–217, 2017.

[3] Deal Banking Marketing Campaign Dataset with Machine Learning:

<https://medium.com/@nutanbhogendrasharma/deal-banking-marketing-campaign-dataset-with-machine-learning-9c1f84ad285d>

[4] Analysis on Bank Marketing Campaign for Portuguese Bank:

<https://github.com/HegdeChaitra/Bank-Marketing-Campaign-Analysis>

The above references have used either only basic models or have done some statistical analysis. We have looked into the data distribution and up sampled the data and trained the Ensemble models with Boosting to achieve best results and get higher data understanding and accuracy gaining the best results.