



Bank Marketing

(with social and economic context)

Data Overview


The dataset used is a Portuguese bank marketing dataset. Repository located at the following URL: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. There are 41,188 observations and 21 Variables in the Data Set. There are 11 continuous measure variables and 10 categorical variables. The target response (y) is a binary response indicating whether the client subscribed to a term deposit or not.

Goals

1. Find how social economic indicator contributes to bank marketing
2. Build a model that predicts whether the client will subscribe to a term deposit (variable y) or not.

Specifications

1. Age
2. job : type of job (categorical)
3. marital : marital status (categorical)
4. Education (categorical)
5. default: has credit in default? (categorical)
6. housing: has a housing loan? (categorical)
7. loan: has personal loan? (categorical)
8. contact: contact communication type (categorical)
9. month: last contact month of year (categorical)
10. day_of_week: last contact day of the week (categorical)
11. duration: last contact duration, in seconds (numeric)
12. campaign: number of contacts performed during this campaign and for this client
13. pdays: number of days that passed by after the client was last contacted from a previous campaign

- 
14. previous: number of contacts performed before this campaign and for this client (numeric)
 15. poutcome: outcome of the previous marketing campaign (categorical)
 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)
 21. Y: output variable

Approach

Preprocessing :

- **Dropped duplicate columns**
- **Outliers handling**

Pdays , previous can not be handled by upper threshold and lower threshold so created bins.all other columns that contains outliers are handled using upper threshold and lower threshold

- **Handling unknown**

Since default contains nearly 20 % of unknown and while replacing it with mode causes reduction in association with output variables , decided to keep unknown. All other variables with unknown handled using mode.

- **Exploratory data analysis**

Firstly considered categorical variables and then continuous variables

Model building :

- Smoting used because of high imbalance in output class.
- **Algorithms used**
 1. Logistic Regression
 2. Ridge Classifier
 3. Random Forest Classifier
 4. Support Vector Classification
 5. XGBoost

Grid Search CV used for parameter tuning , all models performed with accuracy around 90% except random forest it showed accuracy of 88 %.

Results

- From exploratory data analysis obtained a good correlation between categorical variables and subscription for term deposits.
 1. A good percentage of subscription comes from students , while admin shows less percentage even though they campaigned admins more compared to other job categories.
 2. Months of last contact shows a good correlation with 'y' . In march and december we can see a subscription around 50%.
 3. Number of contacts performed before the campaign shows high association with 'y'. As the number of contacts increased there is an increase in percentage of subscription 4 to 7 days of contacts shows subscription of more than 50%.
- High correlation between social economic indicators and subscription for term deposits.(converted continuous variables into bins for eda purposes.)
 1. Duration showed a very strong correlation with 'y' .duration of zero seconds means not subscribed.it showed inverse correlation.
 2. For low interbank interest rates customers show a good tendency for subscription. for Interbank interest rate between 0.6 and 1.4 there is a 46% of subscription
 3. consumer price index, number of employees etc shows inverse relation with 'y'