# Portuguese Bank Marketing Project

## Team project Id: 1198
## Project ID: 1000

### Project Team Members:

1. Lavanya Kancharla
2. Ruhi Poul
3. Pramod Naik
4. Pooja KS
5. Suraj Sawane

### Agenda:

- Business Understanding
- Data Exploration and Data Preparation
- Model Building
- Hyper-parameter Tuning and Model Evaluation
- Result / Outcomes

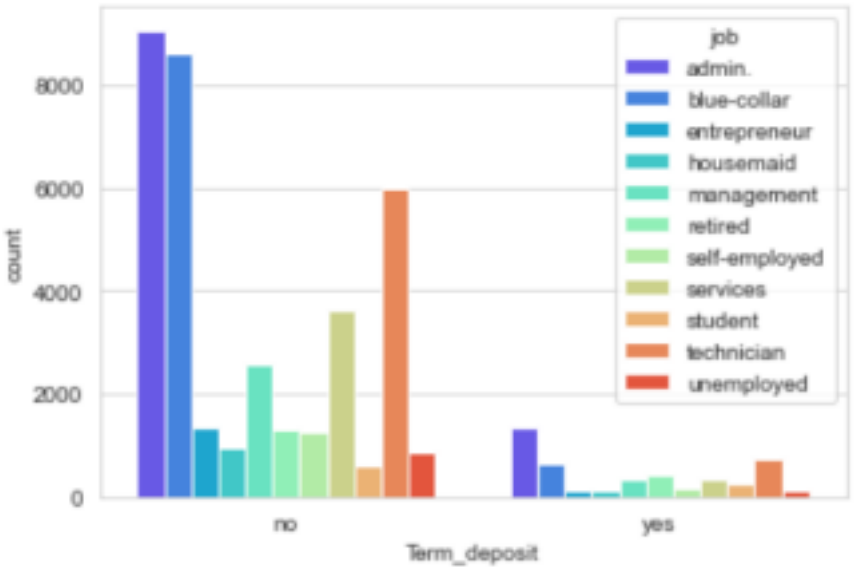
### Business Understanding:

- **Problem Statement**: Improve marketing campaign of a Portuguese bank by analysing their past marketing campaign data and recommending which customer to target

- **Problem Motivation:** By devising such a prediction algorithm, the bank can better target its customers and better channelize its marketing efforts
- Data was collected about each client, type of contact, and outcome.
- What can this data tell us about marketing success for this campaign?
- Can these data science techniques be applied to other areas

### Data Exploration and Preparation:

- All coding done in Python 3
- Extensive use of pandas, NumPy, matplotlib, as well as seaborn and sklearn packages. • Dataset contained 20 different features on more than 41,000 clients.
- Features were both categorical and numerical. Target variable was binary (“Yes” or “No”). • Panda’s package was imported and a data frame was created.
- Categorical variables were looked at first. Visualizations were created using the seaborn package.




## Data Exploration and Preparation:

• Many features had missing values. How do we handle this?

• For categorical features, imputation using other independent variables. For example, cross tabulation between 'job' and 'education'; 'age' and 'job'; 'home ownership' and 'loan status.' • As there are unknown values included in the dataset so we converted those unknown values in nan and then check for missing values .

| | |
|---|---|
| job | 330 |
| marital | 80 |
| education | 1731 |
| default_credit | 8597 |
| house_loan | 990 |
| personal_loan | 990 |

• As we can see that there are missing values in six features so we removed those missing values with fillna

## Model Building:

### Model 1: Logistic Regression

• sklearn. linear_model.LogisticRegression

• Its a classification model though name is Logistic regression

• Fits a sigmoid function to a data

• Outputs probability which is in [0,1] range unlike linear models.

• We get accuracy score using this model is 0.85

### Model 2 : KNearestNeighbors

• from sklearn.neighbors import KNeighborsClassifier

• Accuracy score for KNearestNeighbors is : 0.8610274837331261

### Model 3 : Decision Tree

- sklearn.tree.DecisionTreeClassifier
- Simple to understand and effective
- Splits the data at every node based on one feature
- Uses information gain as measure for split
- Accuracy score for Decision Tree is : 0.8784111877245799

### Model 4 : Random Forest

- Sklearn.ensemble.RandomForestClassifier
- Constructs multiple decision trees and takes the mode of those trees for an example to make the final decision.
- Individual Trees are intentionally over fit and validation set is used to optimize the forest level parameters.
- Accuracy score for Random Forest is : 0.8910362241429542

### Model 5 : XGBoost

- xgboost.XGBClassifier
- Accuracy score for XGBoost t is : 0.8961833543750607

## Summary:

Out of 5 models xgboost with GridsearchCV and RandomisedSearchCv gives the accuracy of 0.937875751503006

## CONCLUSION:

Experiments were conducted on an imbalanced data set for a direct marketing campaign of a Portuguese bank. The goal was to predict whether a customer will subscribe to a term deposit. These projects were conducted using different oversampling techniques. This oversampling is handled by using  SMOTE technique .