Machine Learning Classification : Bank Marketing

Project Participants:

Rahul Singh
Mohnish Lavania
Sarang Bagul

Introduction :

Problem statement:

The data is related with direct marketing campaigns 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.

Business understanding :

- A Term Deposit is a deposit held at a financial institution that has a fixed term. These are generally short-term with maturities ranging anywhere from a month to a few years. When a term deposit is purchased, the lender (the customer) understands that the money can only be withdrawn after the term has ended or by giving a predetermined number of days notice. Term deposits are an extremely safe investment and are therefore very appealing to conservative, low-risk investors.

- Instead of mass marketing, the bank has chosen to be more proactive in identifying potential buyers and communicate straight to the customer via telephone calls. Direct marketing is useful here because its positive results can be measured directly.

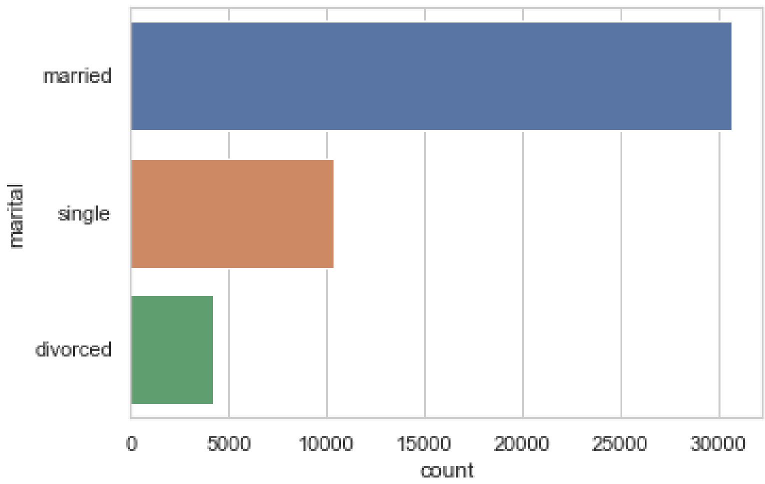
- The goal of this project is to perform post-campaign analytics to identify the potential subscribers of the term deposit product for future campaigns.

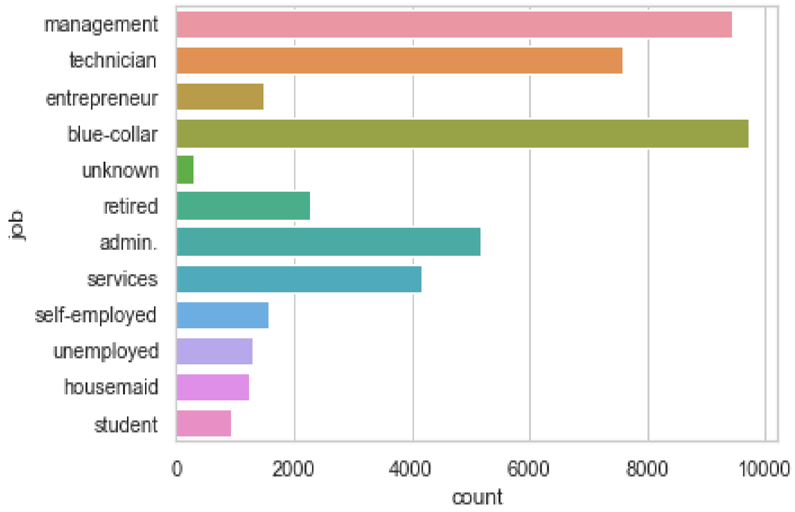
- The data mining task is to create a *Classification Model* that is able to identify potential subscribers using mainly two types of variables in the dataset:
  - ❑ client data (age, education, marital status, loan status, etc.)
  - ❑ campaign contact information for the client (last contact, preferred contact etc.)

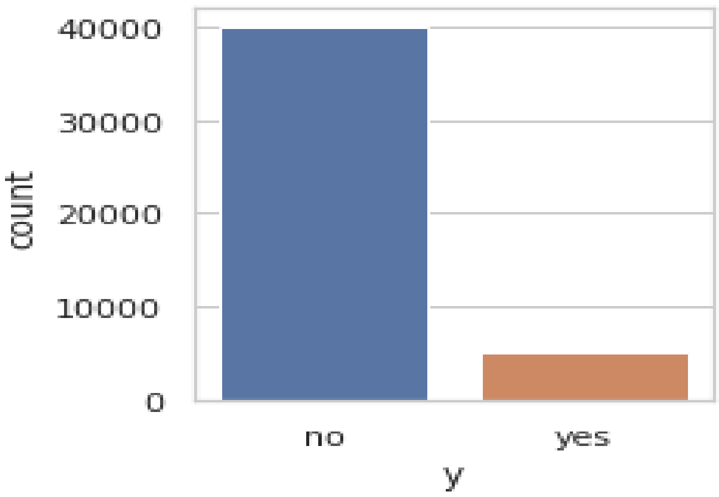
Data Understanding:

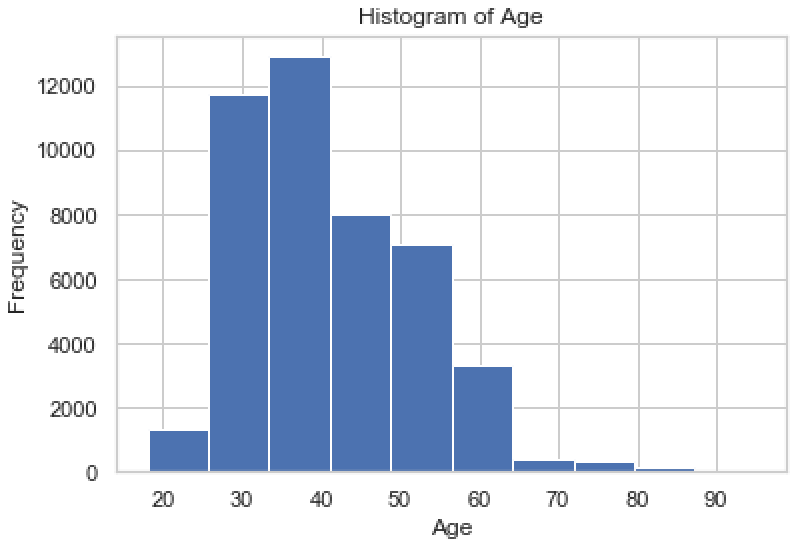
Data Exploration:

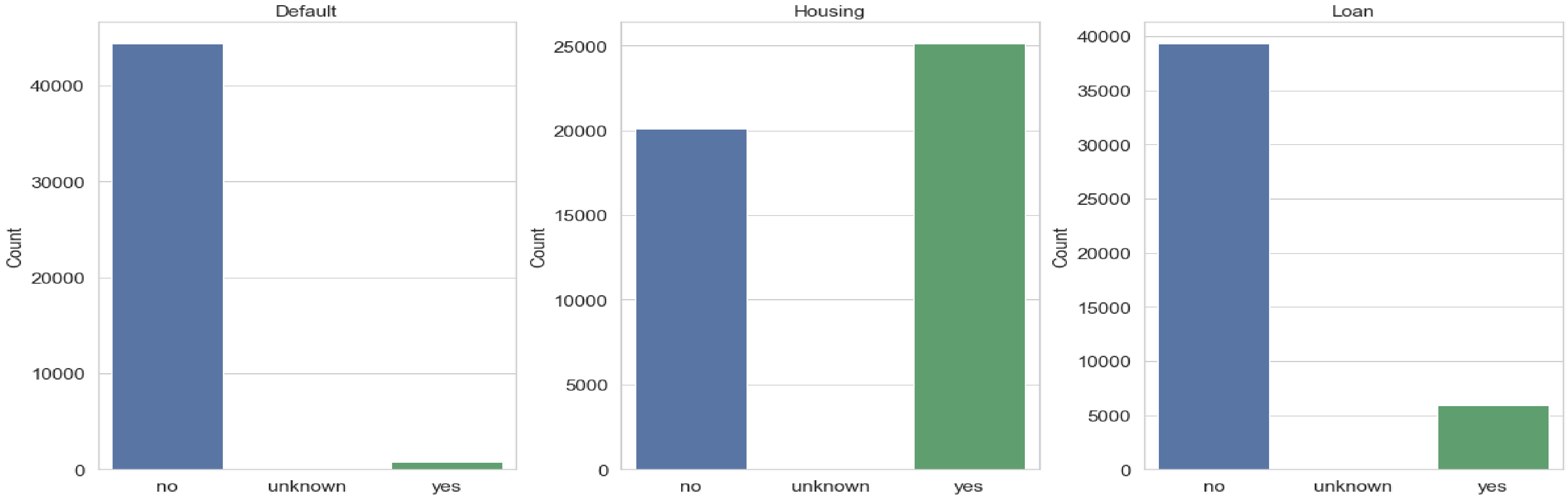
- Most of the clients are married.
- Age between 30-40 year are more active.
- Most of the clients works in management and blur-collar.
- 89% client did not subscribed for the term deposit.











- 99.02% of the clients who subscribed did not have any credit default
- 63% of subscribers did not have housing loans. 93% did not have a personal loan (Low risk investors)

Data Preparation :

- Data does not have missing values. So we randomly remove 10% of data and fill it with mode.
- Label Encoding to all the binary class attributes so that all those will be converted to numeric type and for easier calculations.
- Categorical Non-binary attributes converted to multiple features using One-hot encoding
- Feature-pair (pdays - previous) is highly positively correlated. Therefore, we can remove feature "pdays".
- Numeric attributes are normalised using Minmaxscaler() function.

- Hyperparameter class_weight = “balanced” has been used to make sure that class imbalance is accounted for in every classification task.
- Hyperparameter stratify has been used during train-test-split to make sure that the equal percentage of each class is divided among training and testing data. Similarly, stratified K-fold approach (using sklearn’s StratifiedKFold function) has been used to ensure the same for cross validation.
- GridSearchCV was applied on each classifier to get best parameters.

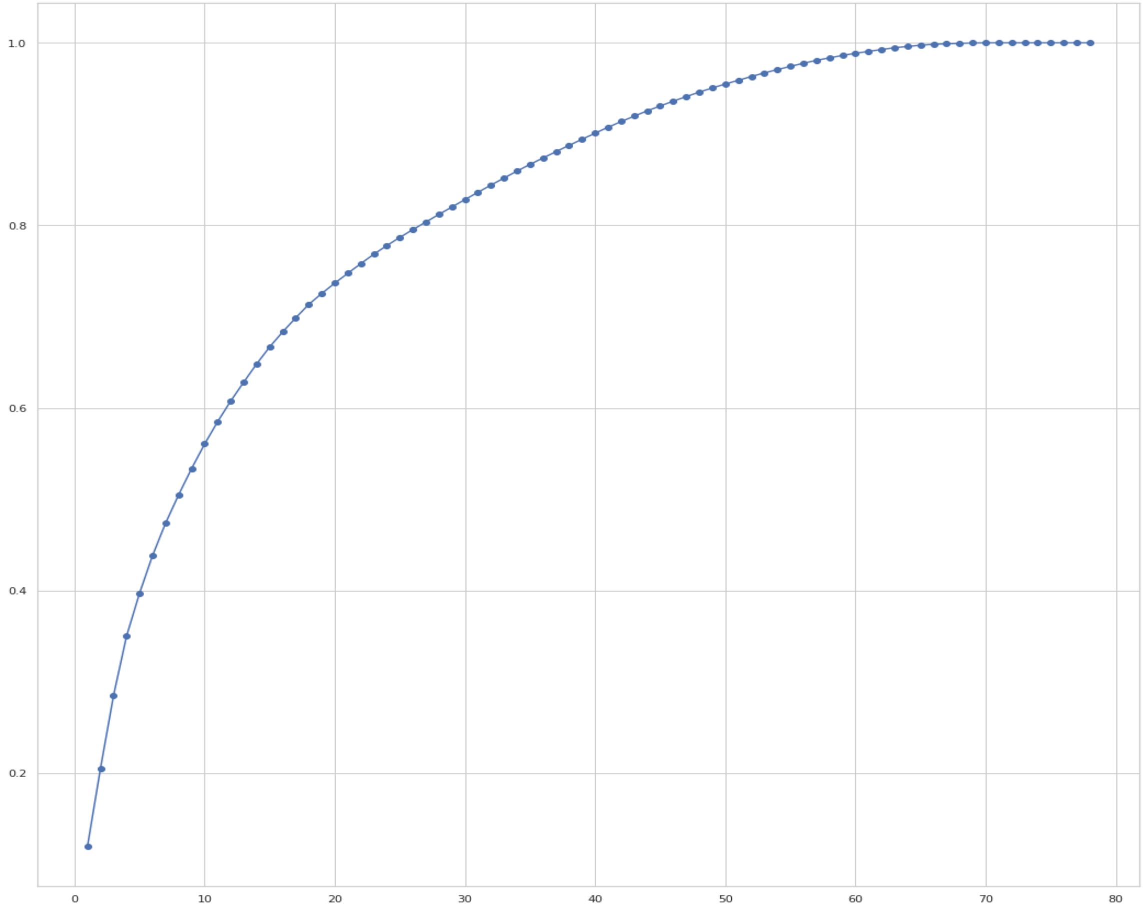
Modeling & Evaluation results without PCA:

	Classifier	F-1 Score	AUC	Accuracy
0	Support Vector Machine	0.590	0.871	0.855
1	Random Forest	0.660	0.858	0.904
2	Decision Tree	0.580	0.815	0.875
3	Logistic Regression	0.548	0.834	0.841

Principle Component Analysis:

- PCA algorithm for Dimensionality Reduction to check which features to retain and which to eliminate and check if their will be some improvement in parameters.

- We can observe that first 50 component amount to a cumulative of 0.95 of the total variance.



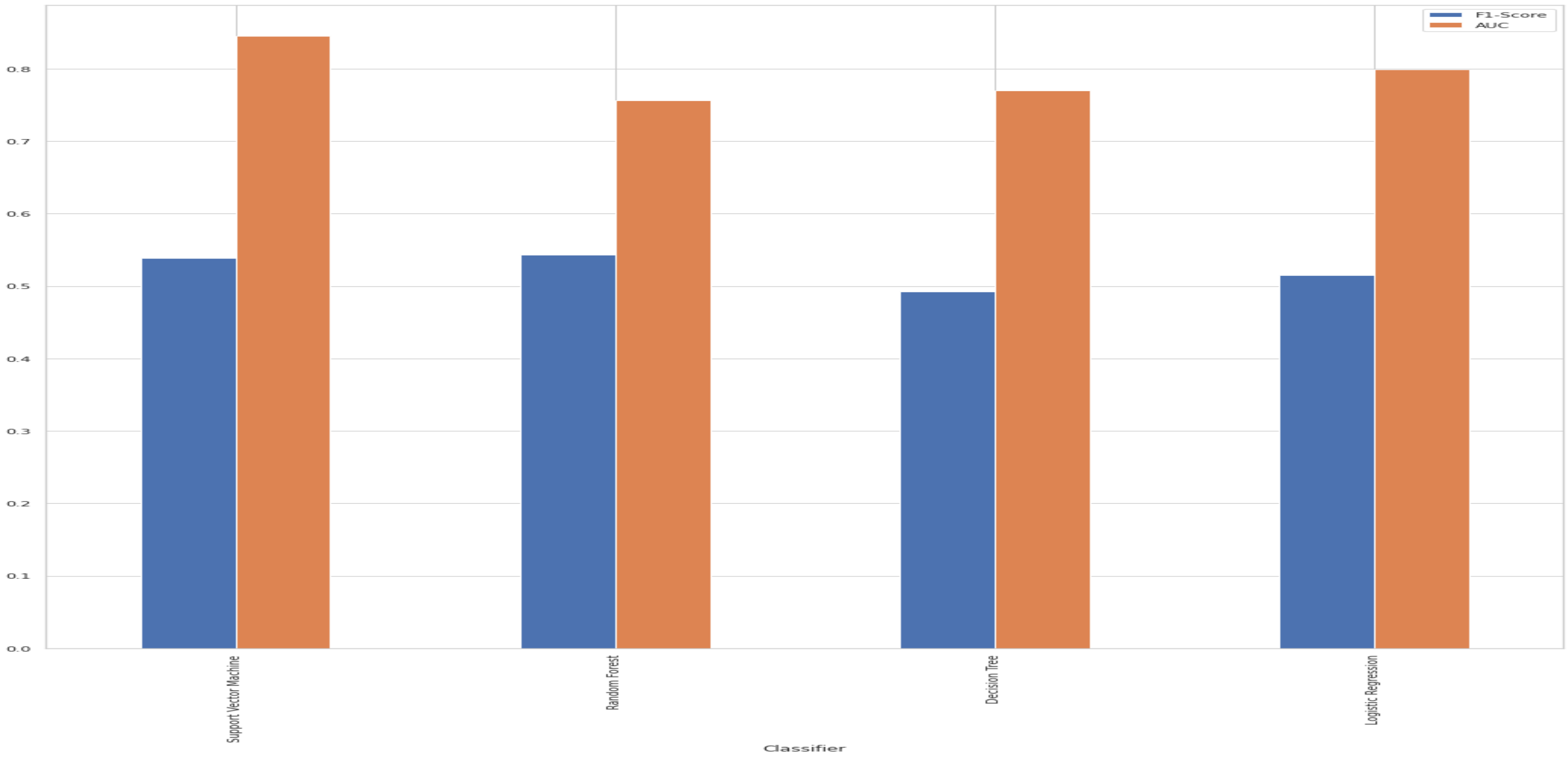
Model Evaluation After Performing Principal Component Analysis:

	Classifier	F-1 Score	AUC	Accuracy
0	Support Vector Machine	0.5382	0.8451	0.8250
1	Random Forest	0.5435	0.7564	0.8840
2	Decision Tree	0.4922	0.7700	0.8343
3	Logistic Regression	0.5148	0.7992	0.8336

Result :

- Random Forest Classifier without PCA turn out to be the best model for the given parameters and datasets in terms of accuracy, f1-score and ROC curve.
- Model trained with PCA , accuracy seems to be decreased. Hence more features can be engineered for the given data.
- Instead of using hyperparameter class_weight = “balanced”, we can also try SMOTE (Synthetic Minority Over-Sampling Technique) sampling

Comparison of f1-score and AUC of all techniques:



THANK YOU