PORTUGUESE BANK MARKETING REPORT

BUSINESS PROBLEM

To predict whether a customer will subscribe to the bank's term deposit or not based on customer details, past campaigning history of the bank and some additional social and economic factors. This is a binary classification task.

THE DATASET

This is data-set that describes Portugal bank marketing campaigns results. The data collected is related with term deposit subscription for 41188 clients. The target variable is the client's response of whether they have subscribed to a term deposit or not. The original dataset contained 16 features however in the subsequent years the dataset was enriched by the addition of 5 new social and economic features that would help in predictive performance of models.

For the project the 'bank-additional-full.csv' file was used that contained 20 predictor features- 10 numeric and 10 categorical as well as the target variable 'y' (yes or no).

Citation Request-

This data-set is publicly available for research. The details are described in S. Moro, P.Cortez and P. Rita. "A Data-Driven Approach to Predict the Success of Bank Telemarketing." Decision Support Systems, Elsevier, 62:22-31, June 2014

EDA

Basic Domain Analysis was conducted where each variable was described and a general relationship was established with the target variable. For example, the employment variation rate which is essentially the variation of how many people are being hired or fired due to the shifts in the conditions of the economy can have an impact on the term deposit subscription rates. When the economy is in a recession or depression the variation rates are higher and the economy is unstable so people are more conservative with their money and how the spend it because their financial future is less clear due to unemployment. When the economy is at its peak, i.e., lower variations, the economy is stable and individuals are more open to investments because their employment options are greater.

A detailed univariate analysis was conducted and visualized using charts, count plots and histograms using the seaborn library. Some insights from the analysis were that most people were married (61%) as compared to single (28%) and divorced (11%). Also, 86% of people had not been contacted before for any previous campaigns while the remaining had been contacted before at least once. Univariate analysis also revealed that many features had unknown values such as job, education, marital, housing, default and loan. One of the most important findings was the target variable ‘y’ was highly imbalanced with 89% of people not subscribing to the term deposit and only 11 % subscribing.

Bivariate analysis was also performed using count plots and histograms that revealed interesting insights in the relationship of each variable with the target variable. For instance, for job vs y, students and retired people had a proportionally higher subscription rate to the term deposit as compared to blue-collar professionals. This was also corroborated by the ‘age vs y’ analysis that revealed older people (>60 years) and young people (<25 years) were more likely to subscribe to a term deposit. Other insightful graphs were the employment variation vs y that revealed that the lower the emp_var_rate the greater the success rates of the campaign which is in line with the domain analysis findings.

DATA PREROCESSING AND FEATURE ENGINEERING

MISSING VALUES

The dataset has no missing values however while doing data analysis there were many features that had "unknown" as one of the labels which had to be handled. This was done by imputing with the mode i.e., most frequently occurring category of the feature.

OUTLIERS

First, boxplots were used to visualize the outliers. Each feature was checked to see whether the outliers needed to be handled or were significant enough to play an important role in predicting the target variable. Outliers that made up less than 5% of the total observations needed to be handled. Despite the Age feature fulfilling the previous criteria, outliers weren't handled for it because during bivariate analysis it was identified that older people have higher chances of subscribing to the term deposit. Out of the 469 outliers there were about half the observations with "yes" indicating that age plays an important role in prediction of our target variable, especially the older age groups. Thus, outliers weren't handled for the age feature.

HANDLING CATEGORICAL VARIABLES

One Hot Encoding was used to transform the 10 categorical features into a numeric form.

BALANCING THE DATASET

One of the major problems of the dataset that stood out was that the target variable was highly imbalanced with 88.73% being '0 or no' and 11.27% being '1 or yes'. This is a cause of concern since our model would not have enough data to learn from the positive class resulting in inaccurate predictions for the minority class. Such imbalanced datasets are often found in the real world and tend to distort the algorithms by favouring the majority class.

This was handled by using the oversampling technique **SMOTE**. Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of observations of the minority class by generating new synthetic instances from given input.

The new instances are not just copies of existing minority cases. Instead, the algorithm takes samples of the feature space for each target class and its nearest neighbours. The algorithm then generates new examples that combine features of the target case with features of its neighbours. This approach increases the features available to each class and makes the samples more general.

The SMOTE class from the oversampling module of the imbalanced learn python library was used for this task. It is to be noted that SMOTE was performed only on the training data. This technique is not performed on the test set since it runs the risk of having the synthetic sample (that was created based on this original sample) in the test set which would lead to an overestimation of model performance.

FEATURE SCALING

The data set was standardized by using the StandardScaler class which returns the input observations within a range of -1 to +1. Scaling is done to prevent the model from being biased by features with a higher magnitude as compared with a lower magnitude.

MODEL SELECTION

After pre-processing, the data was fit to 4 models- Logistic Regression, Random Forest Classifier, Support Vector Classifier and XGBOOST. For all the models, the data was split into training and testing set using sklearn's train test split method.

EVALUATION

The following evaluation metrics were calculated for each model-

1. Recall- it is a measure of from the total number of positive results how many positives were correctly predicted by the model. It is literally is how many of the true positives were recalled (found). Recall= $TP/\ (TP+FN)$
2. Precision- it is a measure of amongst all the positive predictions, how many of them were actually positive. Precision=$TP/\ (TP+FP)$
3. F1 score- is defined as the harmonic mean of Precision and Recall. It sums up the predictive performance of a model by combining two otherwise competing metrics — precision and recall.
4. Accuracy- it's defined as the total number of correct classifications divided by the total number of classifications. It tells us the fraction of predictions our model got right.
5. ROC AUC Score- It is the area under the Receiver Operation characteristics Curve. It is used to compare multiple model's performances. The higher this value the better the model performs.

Recall was used as the primary evaluation metric. For our use case, it is important that our recall is high. The False Negatives i.e., customer subscribing to term deposit but model predicting customer will not must be as low as possible. We could be lenient on the False Positives i.e., customer not subscribing to term deposit but model predicting customer will thus compromising on lower Precision score for higher Recall. This is because it would be more costly to lose out on a customer that could actually subscribe if the bank called than call a customer that would most likely end up not subscribing.

10- fold stratified cross validation was performed to ensure that the percentage of samples for each class was preserved in every fold. This technique is preferred when we have an imbalanced dataset to avoid disproportionate representation of target variable classes in the training and testing sets.

The model evaluation scores were as follows-

Model	**Logistic Regression**	**XGBOOST**	**Support Vector Classifier**	**Random Forest Classifier**
Recall	0.84	0.71	0.83	0.58
Precision	0.41	0.52	0.42	0.60
Accuracy	0.85	0.89	0.85	0.91
F1 score	0.55	0.60	0.56	0.59
AUC	0.92	0.94	0.92	0.94
Avg cross validation score	0.76	0.70	0.45	0.56
Std of cross validation score	0.06	0.02	0.06	0.02

From the above table it is evident that Logistic Regression performs best with respect to recall score which is our model selection criteria.

Support Vector classifier shows a good recall score but performs extremely poorly on cross validation scores.

XGBOOST fares decently well on all the evaluation metrics but was not selected since Recall is very important in our case.

Despite Random Forest having a high accuracy and AUC score it performs very poorly on precision and recall.

Thus, Logistic Regression was selected as the final model with a recall score of 0.84.