**PORTUGESE BANK**

I understand that data set contains direct marketing via phone call to request clients to register for term deposits, based on the given features I have to predict whether customers will subscribe for term deposits or not. It is a Binary class classification problem with **Yes** or **No** as target variables.

Bank-Additional- full with 20 features,1 target variable and 41188 datapoints, here in this data we have extra features like **emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed** . and we don’t have one feature that other data set has is **Balance**.

**Extra Features and its meaning:**

**Default** - yes if loan not paid on time before

**Employment variation ration: cyclical employment variation:**

Cylical employment variation is essentially the variation of how many people are being hired or fired due to the shifts in the conditions of the economy. When the economy is in a recession or depression, people should be more conservative with their money and how the spend it because their financial future is less clear due to cyclical unemployment. When the economy is at its peak, individuals can be more open to risky investments because their employment options are greater.

**Consumer price index:**

(in the US) an index of the variation in prices for retail goods and other items.

The Consumer Price Index measures the overall change in consumer prices based on a representative basket of goods and services over time.

The CPI is the most widely used measure of inflation, closely followed by policymakers, financial markets, businesses, and consumers.

Businesses and consumers also use the CPI to make informed economic decisions. Since CPI measures the change in consumers' purchasing power, it is often a key factor in pay negotiations.

**Consumer confidence index:**

A consumer confidence index (CCI) is an**economic indicator** published by various organizations in several countries. In simple terms, increased consumer confidence indicates economic growth in which consumers are spending money, indicating higher consumption

The CCI is based on the premise that if consumers are optimistic, they will spend more and stimulate the economy but if they are pessimistic then their spending patterns could lead to an economic slowdown or recession.

**Number of Employees:**

enter the total number of Employees in the employment category employed to provide services under the Contract during the report period, including part time Employees and Employees of subcontractors.

**euribor3m:**

The **3 month Euribor interest rate** is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of **3 months**

1. the basic rate of interest used in lending between banks on the European Union interbank market and used as a reference for setting the interest rate on other loans.

* **From the dataset features I understand that duration of the call performs key role in predicting the feature, so to create genuine predicting model we need not to use this feature. I will create two models one with duration feature and one without duration feature.**

**Data Basic Checks:**

* 10 features and 1 target variable is objective type and reaming 10 features are Numerical types.
* There are no null values in numerical features in the dataset
* Checked numerical value’s central tendencies I feel all columns are looking good, no columns with zero std deviation and few columns with 0 as min value and this can be accepted as these columns can be 0 as well.
* Whenever duration is zero (0) target is ‘No’
* all objective datatype features are categorical only
* I saw that the data is imbalanced with 11% - yes and 89%-No in target variable.
* Found 12 duplicated rows in the dataset

**EDA:**

**Univariate Analysis:**

* the average age of people has subscribed and not subscribed is almost same
* we can clearly see that when average call duration is high people are subscribing for term deposits
* the number of contacts made during current campaign doesn't have much effect on client’s subscription, there are no subscriptions when avg contacts were made is high
* when there is huge gap between last contact, clients are less opting for subscription
* if there is contact happened prior to current campaign clients are more likely opting for subscription
* when emp.var.rate is negetively growing more clients are opting for term deposits
* avg consumer price index is same when there is subscriptions and no subscriptions
* consumer confidence index is almost same when there is subscriptions and no subscriptions
* when there is less euroborn interest rate people are more interested for term deposits
* number of employees is almost equal during subscriptions and no subscriptions

**Bivariate Analysis:**

* 45% students and 33% retired people are more interested in term deposits
* 11% married, 16% single and 11% divorced has subscribed for term deposits, so we can focus on single to improve subscription
* people who doesn't have credit default would likely subscribe
* i don’t think housing loan will affect the subscriptions
* people without any personal loan more willing to go for term deposits
* contact through cellular has high impact on making clients subscribing for term deposits
* i can't intuit anything from the contact made on month and day data.
* if people subscribed during previous campaign, they are also willing to subscribe now, so focus on previous subscribed clients.

**Multivariate Analysis:**

* people are taking less time to decide if they have been contacted before, its having negative relation
* even same with contacts made during the current campaign is also having negative relation with duration of the call, if people has been contacted during this campaign they are taking less time to decide

**Data Pre-processing:**

**Missing/null values treating:**

Found no null values in the Numerical columns but there are missing values in categorical columns which are categorised as ‘unknown’.

* Job feature has 330 unknown categories, and these are replaced with mode of the column i.e: admin
* ‘marital’ feature has 80 unknown categories, and these are replaced with mode of the column i.e: married
* ‘education’ feature has 1731 unknown categories, and I felt education is relevant with job and vice versa, so l had imputed the unknown category of 'education' based on their job. when 'education' is unknow , I have checked 'job' of that row, and mode of 'education' for every job category and imputed with the same
* ‘default’ feature has 8597 unknown categories, and these are replaced with mode of the column i.e: no
* ‘housing’ and ‘loan’ feature has 990 & 990 unknown categories, and these are replaced with mode of the column i.e: yes & no

**Encoding Categorical features:**

* Encoded all the categorical feature, where priority given based on the count of each category from highest to lowest.(from 0 as highest priority and successive numbers with reducing priority i.e 1,2,3 so on) and alternate model tried with where priority given based on the count of each category from lowest to highest.(from 0 as highest priority and successive numbers with reducing priority) but first way of approach is performing somewhat better than second approach.
* And I haven’t tried one-hot encoding because of the increasing in number of features, as we have curse of dimensionality if feature are more model will be less accuracy.

**Outliers Checking:**

* though there are outliers present in the 'age', duration','campaign','pdays','previous' features, it is important to keep outliers as it makes impact on target, even tried treating the treating outliers of ‘campaign’ which is having more than 5% outliers but there is no changes in the performance of the model.

**Feature Selection:**

Model1:

* dropped two columns having >0.9 correlation(‘emp.var.rate','euribor3m’)
* And dropped 16 duplicate rows

Model2:

* Haven’t dropped features with high correlation, and dropped ‘duration’ feature to create genuine predicting model
* And dropped duplicated rows of around 2145 and recall for subscription got improved after dropping duplicated rows.

**Balancing Data:**

* Since data is imbalanced , done oversampling using SMOTE

**Scaling:**

* As few columns are having large Numerical values, have applied standard scalar

To standardize all features

**MODEL CREATION**

**Model1:**

* Model with ‘duration’ feature performing better than model without ‘duration’ feature.
* And there are not many variations in the model performance by removing duplicate data points, removing high correlated features, and changing encoding priorities.

**Model2:**

* Model without ‘duration’ feature performing average, and tried by changing feature selection techniques, changing encoding methods and there are no huge/more changes in the model performance by removing and without removing duplicate data points, removing and without removing high correlated features, changing encoding priorities.

**SUMMARY:**

Applied different models for the dataset to find the best model for predicting target variable.

I have applied Linear regression, KNN, SVM, Decision tree, random forest, bagging techniques and Boosting techniques as well, out of all models, the best performing model is **RandomForest** as it is giving 97% recall for subscription and **XGBoosting** as well performing well.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Metrics↓** | **Linear R** | **KNN** | **SVM** | **DT** | **Random F** | **GB** | **XGB** |
| F1 score for 1 | 0.89 | 0.92 | 0.92 | 0.92 | 0.95 | 0.92 | 0.94 |
| F1 score for 0 | 0.89 | 0.91 | 0.92 | 0.91 | 0.94 | 0.92 | 0.94 |
| Recall-1 | 0.90 | 0.94 | 0.95 | 0.94 | 0.97 | 0.95 | 0.96 |
| Recall-0 | 0.88 | 0.89 | 0.89 | 0.89 | 0.92 | 0.89 | 0.92 |
| Accuracy | 0.89 | 0.92 | 0.92 | 0.92 | 0.95 | 0.92 | 0.94 |
|  |  |  |  |  |  |  |  |

1- subscribed

0- not subscribed

**Project Analysis:**

* duration of the call is having highly effecting the subscriptions for term deposits
* the number of contacts made during current campaign doesn't have much effect on client’s subscription, there are no subscriptions when avg contacts were made is high
* when emp.var.rate is dropping more clients are opting for term deposits
* when there is less euroborn interest rate people are more interested for term deposits
* if there is contact happened prior to current campaign clients are more likely opting for subscription, so try to contact clients been contacted in the previous campaign
* Students(45%) and retired(33%) people are more interested in term deposits
* 11% married, 16% single and 11% divorced has subscribed for term deposits, so we can focus on single to improve subscription
* people without any personal loan more willing to go for term deposits
* contact through cellular has high impact on making clients subscribing for term deposits.
* if people subscribed during previous campaign, they are also willing to subscribe now, so focus on previous subscribed clients.

**Model Summary:**

I have applied Linear regression, KNN, SVM, Decision tree, random forest, bagging techniques and Boosting techniques as well, out of all models, the best performing model is **RandomForest** as it is giving 97% recall for subscription and **XGBoosting** as well performing well and same as RandomForest.