**INTRODUCTION**

Tele-marketing was first initiated to make customers familiar with the products/ services a firm is offering and to get the customers to purchase that product/ service. This lead to help the firm in banking sector to grow their customer count by establishing interest in them with some attractive deals which couldn’t be refused. This led the banks to step into marketing to customers about their term deposits.

This project considers real data from a Tele-Marketing campaign that was done by Portuguese bank from May 2008 – November 2010, this data has been has been extracted from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). From that source we extracted two data sets namely *“bank-additional-full.csv”,” bank-additional.csv”* where the first file is consisting of 41188 records and 21 inputs which we have chosen for training of the data, whereas the second one is consisting of 4119 and 21 inputs is been chosen for testing of the data.

**OBJECTIVE**

* Our objective in this project is to predict if a customer subscribes to a term deposits or not by understanding the different features and performing predictive analytics.
* Performing customer segmentation by using clustering techniques to gain valuable insights from the clusters which will be very helpful for organizations in their further campaigns.

**METHODOLOGY**

This project mainly utilizes ‘Classification’ techniques to examine a data set related to marketing campaigns of a Portuguese bank. The goal of this project is to predict whether a client subscribes to a term-deposit, segmenting the customers into different clusters and find some valuable insights which will be helpful for future campaigns. To achieve these goal we analyze initial data set and performing preprocessing and cleaning the data set, Exploratory data analysis also has been used to visualize the dataset of Explanatory variables and Cross-Tabulation of the metadata categories to determine the distribution of previous Term Deposit acceptance. By above analysis we got to know whether dataset is linear or not, whether it is skewed or not. This analysis helped us in choosing efficient predictive machine learning algorithm. Next step was Correlation of features here we found some features are strongly correlated which says that they are strongly directly proportional so which helped us to remove the extra features in feature selection. Feature selection was also done after the first training of our model. Next step was feature engineering where we transformed the data such as categorical data to numerical and then to factoring. The new data set which has been formed was used in machine learning algorithm. Next step was we tuned the algorithm till an efficient stage where it has given its best accuracy. Then we used this model on our test data set and validated it.

For our-another goal of customer segmentation we have choosed clustering algorithm based on size of data set and characteristic type of features. Next step we have used the distance metric to handle mixed data types. Here we have used Gower distance metric and transformed our data. Then we decided the number of clusters it need to have by silhouette width internal validation metric. Then we visualized the clusters and tried to find insights from it.

**Characteristics and definitions of the features**

Data Folder consists of 2 data sets one is for training the model and other is for validation. Training model consists of 41188 records and 21 features. Validation data set consists of 4119 records and 21 features. From these 21 features we have 20 explanatory features and 1 response feature or predictor. The observation of predictor says whether the client is subscribed to term deposit.

**Definition of input variables:**

**age** - Age of the client- (numeric)

**job** - Client’s occupation - (categorical)  
(admin, bluecollar, entrepreneur, housemaid, management, retired, selfemployed, services, student, technician, unemployed, unknown)

**marital** - Client’s marital status - (categorical)  
(divorced, married, single, unknown, note: divorced means divorced or widowed)

**education** - Client’s education level - (categorical)  
(basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown)

**default** - Indicates if the client has credit in default - (categorical)  
(no, yes, unknown)

**housing** - Does the client as a housing loan? - (categorical)  
(no, yes, unknown)

**loan** - Does the client as a personal loan? - (categorical)  
(no, yes, unknown’)

**contact** - Type of communication contact - (categorical)  
(cellular, telephone)

**month** - Month of last contact with client - (categorical)  
(January - December)

**day\_of\_week** - Day of last contact with client - (categorical)  
(Monday - Friday)

**duration** - Duration of last contact with client, in seconds - (numeric)  
For benchmark purposes only, and not reliable for predictive modeling

**campaign** - Number of client contacts during this campaign - (numeric)  
(includes last contact)

**pdays** - Number of days from last contacted from a previous campaign - (numeric)  
(999 means client was not previously contacted)

**previous** - Number of client contacts performed before this campaign - (numeric)

**poutcome** - Previous marketing campaign outcome - (categorical)  
(failure, nonexistent , success)

**emp.var.rate** - Quarterly employment variation rate - (numeric)

**cons.price.idx** - Monthly consumer price index - (numeric)

**cons.conf.idx** - Monthly consumer confidence index - (numeric)

**euribor3m** - Daily euribor 3 month rate - (numeric)

**nr.employed** - Quarterly number of employees - (numeric)

**Definition of predictor:**

Output variable (desired target) - **Term Deposit** - subscription verified(binary: ‘yes’,‘no’)

**Flow chart of both the Models**

**A screenshot of a cell phone

Description automatically generated**

**Model-1**

* 1. **Data Pre-Processing:**

Data set pre-processing starts with check of duplicate records and found out 12 records has been duplicated in our set. We removed this rows by using distinct. Next step we checked whether there is any missing data. We found out that data is missing in these columns Housing, default, loan. The number of rows missing in housing and loan are 990 and in default it is 8597. We found out that all missing values in housing and loan fall in same rows. Upon that these are binary categorical variables we can’t even take the mean and impute it, and to even impute them with the mode the difference of occurrence of no and yes are hardly 1500, so we can’t even impute with its mode so we have removed those rows. Assuming from 41,000 removing 900 records doesn’t do much damage to our model. As we go the default feature the number of No’s is very large that is around 33000 and yes are 3. So we imputed the data ‘No’ for Default feature by taking mode into consideration.

* 1. **Exploratory Data Analysis:**

**Age distribution:**

**A picture containing text

Description automatically generated**

We can clearly observe that clients who are above legal age for subscription to a life expectancy age, the age has been distributed normally. By observing above data set we can conclude that our dataset has covered all types of ages which helps in our training our model which will get a larger exposure for it.

**Housing loan and personal loan vs subscription:**

**A screenshot of a social media post

Description automatically generated**

By observing we can clearly visualize that customers who has no personal loan and housing loan are high for saying yes to a subscription and rejection for a subscription is high where they have no personal loans and having housing loans.

**Subscription based on Number of Contact during the Campaign:**

**A screenshot of text

Description automatically generated**

We can clearly observe that in a campaign after 5 contacts the client saying yes to a subscription is null. So they can actually preserve these effort and can use them efficiently for other productivity.

**1.3 Cross tabulation of bank marketing data set:**

Here main objective is to determine the categories that has greater effect on Term Deposit subscription on the basis of frequency of occurrences. This has been analyzed by using summary function in R. After analyzing we found out that there is high variance of data in some categorical features such as personal loan ,default, month and our predictor also. As we clearly observe that our data set is of high variance and low biased.

* 1. **Correlation of Bank Marketing data set:**

**A screenshot of a cell phone

Description automatically generated**

As we can see only 3 features are strongly correlated they are emp.var.rate, euribor3m, nr.employed features and strong weakly correlated are emp.var.rate and previous feature. Remaining are not strong weakly or strong positively correlated.

* 1. **Feature Selection:**

By observing the above correlation matrix we found out 3 strongly correlated features, keeping 1 of the among features and removed the other 2 features. Because strong correlation implies that those features are linearly dependent value increase in one feature leads to increase of other 2 features even. So the variations of one of the features affects will be same as keeping all of them. Keeping one of them or 3 of them doesn’t affect much on our hypothesis error rate. But due to more features may cause overfitting of the model which may effect in predicting of our predictor feature.

* 1. **Selection of Classification model :**

By analyzing the above exploratory analysis and cross tabulation of metadata categories we can say that our data has less variety of assumptions about the target data . If we see the response feature has 35654 No’s and 4532 Yes from over 40,000 records this clearly shows us there is a very less variety of Yes which clearly indicate that our data is low bias. By observing the above correlation matrix and exploratory analysis we can say that there is a high variance of data. In such cases Linear regression, Naïve Bayes, Logistic regression gets eliminated automatically because they are good for high bias / low variance dataset. Decision Trees were rejected because of inability to learn after initial processing, and possible overfitting of the data, therefore non-adaptable to new data. Random forest algorithm was chosen as it is probability based and have the ability of handling numerical/categorical variables.

* 1. **Feature Engineering :**

Random forest is not good with categorical variables, as it favors with the category variables having many levels. So, here starting we have transformed our whole categorical data into numerical data . Then the class of numbers are further changed to factor level representation.

* 1. **Training the model:**

We train our model by using random forest function and having default values of mtry and ntrees. Data passed to our model will be our cleaned dataset and selected features after the correlation. After first training of our model, the fitted model has OOB estimate error of 8.56% and from confusion matrix we can say that no class error rate is 3.6% where as for yes it is 46%. Then we need to tune the model to decrease the error rate .Before that we have selected the features by using varimplplot function and plotting the feature importance.

A screenshot of a cell phone

Description automatically generated From the above plot our predictor is highly dependent on duration and least dependent on the mode of the contact and whether they have personal loans or not. So we are taking top 15 features from the importance list and we again train the model. We found out the OOB rate is 8.66% and No’s classification error is 3.5% and Yes’s classification is 46% which clearly mentions us that the removed features are very least dependent that it has no effect on our model. Now these features selected are our final features of our model.

* 1. **Tuning the model:**

Once the final features has been selected we want to tune the model by changing the mtry that is number of variables tried at each split. So we basically ran our model keeping in loop running for different ‘mtry’ i.e from 1 to 8 and store the accuracies in a vector and plot them. So with a higher accuracy one we can take as our selected model.

A screenshot of a cell phone

Description automatically generated

We can see that classification accuracy on training set for different mtry’s and it has been observed that after 5 splits it constant. So we decide our final tuned model with mtry’s = 5. Then also we have the OOB error rate around 8%. After carefully analyzing the data we can see that there is an imbalance of data where we can observe that Yes are around 4000 and No are around 36000 so we have performed the **over-sampling** using the Rose package so that our Yes are also duplicated of certain situations around to 33000. Then we again trained our model with OOB rate is 3.17% and no classification error of No around 5% and for Yes is 99.99%. So finally we got our model with mtry’s 5 and OOBrate is around 3%.

* 1. **Testing the data:**

Once our model is ready we are ready to test our model with other bank addituional file which consists of 4119 records . This data will be pre-processed and transformed and then it goes out for testing with our model. After testing we got highest accuracy for the model .The. confusion matrix for our testing is shown below.

|  |  |  |
| --- | --- | --- |
|  | **no** | **yes** |
| **no** | 3570 | 2 |
| **yes** | 0 | 442 |

This shows our model has an 99% of accuracy from our testing data set.

**Model-2**

The second part of our model is basically customer segmentation by using clustering techniques.As our data consists of bot categorical and numerical data types it is difficult to cluster normally so we had to use some distance metric and transform the data .

**2.1 Feature Engineering:**

As our data set is of mixed data types we are using an distance metric called Gower Distance metric. In Gower Distance metric the formulation of data is done by for each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1. Then, a linear combination using user-specified weights (most simply an average) is calculated to create the final distance matrix. The metrics used for each data type are described below:

* quantitative (interval): range-normalized Manhattan Distance.
* ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
* nominal: variables of *k* categories are first converted into *k* binary columns and then the Dice coefficient is used.

As we have a large dataset it is very difficult to transform into distance metric which needs a good machine configuration and its also takes a large computational time. Due to these reasons we have selected a test data set which has low number of records of 4119 .

**2.2 Choosing a Clustering Algorithm:**

We are selecting here Clustering of PAM model because as we have imbalanced data we need to even handle the outliers and noisy and also more robust than K-means. It also have an added benefit of having an observation serve as the exemplar for each cluster. Even our dataset is small so we have choosen the PAM model.

**2.3 Selecting the number of Clusters:**

The number of clusters are decided by using silhouette width which is an internal validation metric which is an aggregated measure of how similar an observation is to its own cluster compared its closest neighboring cluster. The metric can range from -1 to 1, where higher values are better.

We found out the silhouette width for different clusters by running it in a loop and plotted them .

Where we found out that model having 7 clusters has large silhouette value than other. Deciding with 7 Clusters we ran our PAM model and brought out the clusters.

A close up of a mans face

Description automatically generated

**2.4 Insights from the Clusters:**

By observing the above clusters we can derive some patterns in every cluster which will help us to know what groups of patterns of customers are available and the organization can plan campaigns according to the groups. We are showing some of patterns in each cluster:

A screenshot of a cell phone

Description automatically generated

By using these patterns we can find out what is the average duration of time spent on each customer group. And how many times they have been contacted on each customer group and so on.The organization use this descriptive statistics and plans their campaign to tackle them different groups of customers.

**Conclusion**

By using Bank marketing data set to predict the subscription of term deposit by using a random forest classification was highly successful with 99% accuracy. But in our other goal we are not able to get valuable insights from the clusters even though having a certain patterns of it.

**Further Study**

Mostly need to focus on clustering techniques which can handle large mixed data set types. So, that we can get valuable insights of data and even having an imbalanced data is one of the main disadvantage. By having some domain knowledge on banking it will help us to find us some relations. Finding some distance metric which can have less computational time in handling large datasets.

**References**

# Works Cited

abhinav. (n.d.). *data is linear or not*. Retrieved from medium: https://medium.com/@abhinav.mahapatra10/ml-basics-regression-how-to-tell-if-a-dataset-is-linear-or-not-594a4f1e8aaf

analyst, p. (n.d.). *random forest classifier*. Retrieved from r bloggers: https://www.r-bloggers.com/how-to-implement-random-forests-in-r/

*bussiness science.io*. (n.d.). Retrieved from Clustering: https://www.business-science.io/business/2016/08/07/CustomerSegmentationPt1.html

*Clustering mixed data types*. (n.d.). Retrieved from wicked good data: https://dpmartin42.github.io/posts/r/cluster-mixed-types

igras, K. (n.d.). *Customer Segmentation*. Retrieved from Knuggets: https://www.kdnuggets.com/2019/09/customer-segmentation-r-users.html

maklin, c. (n.d.). Retrieved from medium: https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

*Rose package*. (n.d.). Retrieved from Cran-r.org: https://cran.r-project.org/web/packages/ROSE/ROSE.pdf

**DATA SOURCE**

<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>