Capstone Project Proposal Report

Contents

Introduction

Project Name: Bank Marketing Prediction

Submitted By: Shashank Varshney

Domain Background

This project is related to the telemarketing. Telemarketing is a domain in which companies directly contact customers and try to sell the products which can help customer to achieve their goals.

In telemarketing, customer can be contacted by any means like e-mails, calls, etc. As part of telemarketing process, customer is contacted and provided information about a product and it can sometimes be required to contact customer more than one time.

The problem in telemarketing domain is company have to call all of the customers be it a potential customer or not. This become a lot of logistical work and lot of efforts are wasted on the customers which don’t fit in the potential buyer category for that product.

This problem requires solution so that company can target only the potential customers which have higher probability to buy that product instead of wasting efforts on the customers which have lower probability of buying that product.

So past data can be used for the data analysis and ultimately machine learning techniques can be applied to predict the potential buyers as per various attributes related to customers.

This kind of problem has already been researched and following is the link of the same.

http://media.salford-systems.com/video/tutorial/2015/targeted_marketing.pdf

Problem Statement

This dataset has been taken from the UCI Machine learning repository. This data is related to direct marketing campaigns (phone calls) of Portuguese banking institution. The goal is to predict if client will subscribe a term deposit or not based on some attributes of the client so that potential clients can be targeted instead of wasting efforts on the clients which have very low probability of buying the term deposit.

Datasets and Inputs

Dataset contains information related to direct marketing campaigns of Portuguese banking institution. The marketing campaigns were based on phone calls and often more than one contact to the same client was required, to access if the product would be yes or no for the subscription.

Dataset contains 41188 examples and 20 input variables. Following are the input variables.

1. **Age:** Age is the factor which can impact client interest in the term deposit.
2. **Job:** This is type of job client have.
3. **Marital:** This is marital status of the client.
4. **Education:** This gives educational background of the client.
5. **Default:** Whether client has credit in default.
6. **Housing:** Client has housing loan or not.
7. **Lona:** Client has personal loan or not.
8. **Contact:** contact communication type. Cellular or telephone.
9. **Month:** Last contact month of the year.
10. **Day_of_week:** Last contact day of the week.
11. **Duration:** last contact duration in seconds. This attribute highly affects the output target (if duration = 0 then y = "no"). This input will only be included for the benchmarking purposes and will be discarded for predictive modelling.

12. **Campaign:** Number of contacts performed during this campaign for this client including last contact.
13. **Pdays:** Number of days that passed by after the client was last contacted from a previous campaign. 999 means client was not previously contacted.
14. **Previous:** number of contacts performed before this campaign and for this client.
15. **Poutcome:** Outcome of the previous marketing campaign.
16. **Emp.var.rate:** Employment variation rate - quarterly indicator.
17. **Cons.price.idx:** Consumer price index - monthly indicator.
18. **Cons.conf.idx:** Consumer confidence index - monthly indicator.
19. **Euribor3m:** Euribor 3 month rate - daily indicator.
20. **Nr.employed:** Number of employees - quarterly indicator.
21. **Y:** Output variable. Has client subscribed to term deposit or not?

All the above mentioned will be used for analysis and the prediction model building except the "Duration".

Solution Statement

Solution of this problem is to create a model which can accurately predict whether client is going to subscribe to term deposit or not. Accurately predicting it can help the organization to put efforts towards the potential clients which are going to subscribe the term deposit instead of calling each client.

So, a predictive model should be created which can accurately predict whether client is going subscribe to term deposit or not.

Benchmark Model

Simple Bayes model will be used as the benchmark model which will consider that every client is going to subscribe the term deposit because this is how current process of the organization is working. Currently organization is calling every client.

Predictive model which will be used as a solution should have way higher accuracy than the benchmark simple Bayes model.

Evaluation Metrics

Counts of clients who said "yes" is 4640 and those who said "no" is 36548. So this is clearly an unbalanced distribution. If we consider all "yes" which is usually the case and call to every client then we will get 11.27% of accuracy.

F-beta score and accuracy shall be used as evaluation metrics for the predictive model and same will be compared with the benchmark model. Following is the formula of the F-beta score.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

Accuracy measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

Precision tells us what proportion of clients we classified as "yes" were actually "yes".. It is a ratio of true positives (clients which classified as "yes", and which are actually "yes") to all positives (all clients classified as "yes"), in other words it is the ratio of

[True Positives/(True Positives + False Positives)]

Recall(sensitivity) tells us what proportion of clients are "yes" were classified as "yes". It is a ratio of true positives (clients classified as "yes", and which are actually "yes") to all the words that were actually spam, in other words it is the ratio of

[True Positives/(True Positives + False Negatives)]

We don't want to miss any client which can say "yes" so we would like to focus more on the Recall. So I would like to keep value of beta as 0 so that recall can be focused.

Project Design

Following approach will be used to build the predictive model.

1. **Data Analysis:** First step is exploratory data analysis of the input variables to get the insight of the dataset.
2. **Data Cleaning and preprocessing:** Lot of missing values can be there so data cleaning and preprocessing steps needs to be performed to prepare data so that predictive modeling can be applied.
3. **Training, validation and test set creation:** As no test set has been provided so dataset will be broken into 3 parts : Training set (60%), validation set (20%) and test set (20%) for training, validating and testing purposes.
4. **Training the model:** Then model training procedure will be started starting from logistics regression, K-nearest neighbors, tree-classifier, random-forest, boosting algorithms, etc. and ultimately best classifier will be identified as per F-beta score and accuracy.
5. **Hyperparameters tuning:** Hyperparameters tuning will be performed for the selected model for better F-beta score.

6. **Predictive Model finalization:** Finally, predictive model will be finalized and trained, validated and tested.