DePaul University

Final Report
Bank Direct Marketing

Group 17

Amy Aumpansub

Wei Wei

Yueze Xu

IS 467: Fundamentals of Data Science

Professor Hamed Qahri Saremi

June 5, 2018

Abstract

Direct marketing campaign based on phone calls is one of important marketing tools that bank performs to boost subscriptions for a term deposit. However, the increasing number of marketing campaigns over time has negative effect. Thus, bank should focus their marketing efforts on a specific group of clients and tailor their marketing campaigns to increase the chance of getting a respond from contacted clients. This paper is a study in the application of data mining to build classification model to identify the important attributes and to predict whether or not the client will subscribe a bank's term deposit after getting phone calls during a campaign period. Using classification analysis as a framework, client's characteristics, campaign's features, and economic indicators were considered in this study. The supervised learning algorithms including Decision Tree and K-Nearest Neighbors Classification were performed. All models have a high accuracy rate and exhibit predictive performance in predicting the outcome. The final model was selected based on strong validation statistics and used to predict the target variable and to identify the important attributes that have a significant effect on the outcome of the marketing campaign. Our study found that compared to client characteristics, the campaign and economic attributes have more influence on whether or not the client will subscribe for a term deposit after getting contacted during the campaign period.

Objective

This project is aimed to identify the important attributes of direct marketing campaigns and client's characteristics that contribute to a success of campaign, as well as to predict whether or not the customers who received phone calls during campaign period will respond the bank's campaign by subscribing a bank's term deposit.

Problem Description

The Portuguese bank performed a direct marketing campaign based on phone calls to stimulate clients to subscribe a term deposit. However, the study shows that the increasing number of marketing campaign over time has less influence on general public [1]. Therefore, bank needs to identify important attributes of successful campaigns and a specific group of clients who are likely to subscribe a bank's term deposit, so that they can improve the effectiveness of campaign and focus their marketing efforts on those specific clients. We will analyze data to address the following questions:

- Which types of customers are more likely to respond to a bank's direct marketing campaign based on phone calls and subscribe a bank term deposit?
- How effective is the bank's marketing campaign based on phone calls?
- Which campaign's attributes have a high influence on target variable?

Dataset

Dataset is related to Portuguese bank's direct marking campaign based on phone calls. It contains 41,188 records collected from May 2008 to November 2010. The dataset was retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. It consists of one binary response (dependent variable) and 20 predictors, which include 10 numeric variables and 10 categorical variables. The independent variables can be grouped into 2 categories including client's characteristics and campaign's attributes. The description of 20 input attributes are shown in the following table.

[1] Kadhim Al-Shayea, Qeethara. (2013). Evaluating Marketing Campaigns of Banking Using Neural Networks. Proceedings of the World Congress on Engineering 2013 Vol II.

Dependent Variable

y = Has the client subscribed a term deposit? (binary variable: yes, no)

Independent Variables

	Name	Attribute	Descriptions
1	default	binary	has credit in default?
2	housing	binary	has housing loan?
3	loan	binary	has personal loan?
4	job	categorical	type of job. e.g. management, entrepreneur, housemaid, student
5	marital	categorical	marital status
6	education	categorical	education level. e.g. unknown, secondary, primary, tertiary
7	contact	categorical	contact communication type. e.g. telephone, cellular
8	day	categorical	last contact day of the week
9	month	categorical	last contact month of year
10	poutcome	categorical	outcome of previous marketing campaign, e.g. success, failure, other
11	age	numeric	age of the customer
12	balance	numeric	average yearly balance in euros
13	duration	numeric	last contact duration in seconds
14	campaign	numeric	number of contacts performed during this campaign and for this client
15	pdays	numeric	number of days passed by after the client was contacted from a previous campaign
16	emp.var.rate	numeric	employment variation rate - quarterly
17	cons.price.idx	numeric	consumer price index - monthly
18	cons.conf.idx	numeric	consumer confidence index - monthly
19	euribor3m	numeric	euribor 3 month rate - daily
20	nr.employed	numeric	number of employees - quarterly

Methodology

Using a supervised learning analysis as a framework, the classification models from Decision Tree and K-Nearest Neighbors were developed and used to predict the target variable. We implemented the following steps as a methodology for building the classification model.

- Explored distributions and relationships between independent variables and dependent variable
- Cleaned data by filling the missing values, detecting outliers, and removing extreme outliers
- Transformed original data and normalized attributes using z-scores
- Selected features and extracted new features to reduce the dimensionality in the data
- Split data into a training set and a test set using holdout partition method
- Built classification models using Decision Tree and K-Nearest Neighbors algorithm
- Evaluated and compared different models based on validation statistics
- Selected final model and use it to analyze the results and provide recommendations

Data Exploration

Client Profile

The dataset contains 8 clients attributes including age, job, marital status, education level, default status, housing, bank balance, and amount of loans.

- Most clients are married, accounting for 60.5% of total number of clients
- 29.5% of clients hold a university degree, whereas 23.1% and 12.7% of clients hold a high school diploma and professional certificate, respectively
- 25.3% of clients are administrators, 22.5% are blue-collar workers, and 16.4% are technicians.
- The age of clients ranges from 17 to 98 years old, which median age is 40 years old.
- The majority of bank clients are between 28 and 37 years old.
- 82.4% of clients do not have loans and 79.1% of clients have no credit in default.
- 53.7% of clients are home owners, whereas 46.3% do not own the house.

Campaign Attributes

The dataset contains 12 campaign attributes including the campaign period, number of phone calls, the outcome of previous campaign, the contact durations, and etc.

- 63.5% of calls are made via cellular phones and 36.5% of calls are made via home telephones
- Bank contacted clients frequently in May, July, and August, implying the campaign is seasonal.
- Contact duration ranges from 0 to 4,918 seconds with mean of 258 secs. and mode of 85 secs.
- The number of contacts performed during the current campaign ranges from 1 to 56 contacts per client with the mean of 2.57 contacts and the mode of 1 contact per client.

Data Distribution

- The age, campaign, duration, and previous have a right-skewed distribution, pdays has a left-skewed distribution, and they all contain outliers. The distribution of nr.employed is left-skewed. Other attributes are normally distributed.
- The z-scores for 10 numeric attributes are calculated for detecting the outliers.
- The age, campaign, duration, pdays, and previous contains outliers which the absolute value of z-scores is more than 3. The outliers were handled in the next section.

Data Pre Processing

Data Cleaning

1. Missing Values

    - Only categorical attributes contain the missing values which account for 0.7% of the total records. This, the missing values were filled with the class label “Unknown.
    - The numeric attributes do not contain any missing values.

2. Outliers

    - The extreme outliers of age, campaign, and duration are removed before binning.
    - Pdays and Previous contains a large Z-score as 96.3% of clients have never got a contact before. While the number of employees contains only 10 values which are not reasonable, thus these three attributes were not included in analysis.

- nr.employed represents the number of employees quarterly indicator, but it does not explain clearly whether it is the total employee from the bank or the whole country. We decide to remove this predictor for our final model.

Data Transformation

1. Equal Frequency Binning

   - The age and duration variables were binned into 10 bins, in which each bin contains approximately 4,000 records or 10% of total records.
   - The campaign variable was binned into 5 bins which contains 42.8%, 25.7%, 13.0%, 10.3%, 8.2% of total records.

2. Smoothing by means

   - After binning variables, the age, duration, and campaign variables were smoothed by using the mean of each bin.
   - The new variables are age_means, campaign_means, and duration_means.

Data Reduction

1. Attribute Selection

   - The decision tree is performed with 10 categorical attributes. We ran the decision tree with the CART method and Gini index. The tree was set with the number of records in parent node and child node of 200 and 100, respectively. The maximum depth was set to 20.
   - The most tree important attributes are poutcome, month, and job (Figure 1).
   - The housing, default, loan and marital status were not included in our analysis as its importance is less than 0.001 (Figure 1).
   - The reduced set of categorical attributes contains 6 attributes including Poutcome, Month, Job, Contact, Day of Week, and Education.

2. Attribute Extraction

   - All 7 numeric variables were standardized using z-scores to remove scale differences.
   - The Principal Component Analysis (PCA) was performed for standardized variables.
   - The 3 new components were extracted, which explained 69.6% of total variance (Figure 2).

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
poutcome	.020	100.0%
month	.012	59.4%
job	.002	11.3%
contact	.002	8.5%
day_of_week	.001	5.1%
education	.001	4.0%
marital	.000	0.7%
housing	3.742E-5	0.2%
default	2.459E-6	0.0%

Growing Method: CRT
Dependent Variable: y

Figure 1: Attribute Selection

Independent Variable Importance from Decision Tree

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.721	38.871	38.871	2.721	38.871	38.871
2	1.107	15.815	54.686	1.107	15.815	54.686
3	1.044	14.920	69.606	1.044	14.920	69.606
4	.938	13.405	83.011			
5	.853	12.181	95.192			
6	.319	4.564	99.756			
7	.017	.244	100.000			

Extraction Method: Principal Component Analysis.

Figure 2: Attribute Extraction

Total Variance Explained (PCA)

Model Fitting I

Decision Tree

Total observations were spitted into a training set and a test set using holdout partition method. A training set contains 28,836 records or 70% of the total records. A training set is used to build models. A test set contains 12,353 records or 30% of total records, used for model evaluation.

- The decision tree was built using y as a dependent variable (Has client subscribed for a term deposit?). The independent variables include 6 categorical variables and 7 numeric variables, which are Poutcome, Month, Job, Contact, Day of Week, Education, age_mean, campaign_mean, duration_mean, emp.var.rate, cons.price.idk, cons.conf.idk, and euribor3m.
- The CART method was selected with GINI index of 0.001 impurity threshold. We applied both pre-pruning and post-pruning.
- The four different decision tree models were developed by changing the number of records in parent and child nodes. The maximum level of depth was set to 20. The performance of each model is presented in model evaluation metrics below (figure 3).
- All four models have a high accuracy rate and specificity rate, indicating that the models can accurately recognize the negative ("no") tuples. However, the lower sensitivity rate of four models indicates that the models have a poorer performance on recognizing positive ("yes") tuples. This is caused by the problem of imbalanced in class of target variable, in which the dataset contains a large number of "no" tuples. We will use ROC curve to address this problem.
- The increase in precision rate indicates the higher quality of exactness. From the model evaluation metrics below (figure 3). The precision rate of models increases as we increase the number of tuples in the parent and child nodes, but it has an inverse relationship with sensitivity rate (recall). The F-measure which is harmonic mean of precision and recall also decreases.
- We selected the last model as our final model because it has a high accuracy rate for both test and trainings and has the highest specificity and precision rate.

Model (Parent, Child)	Partition	Accuracy	Error	Sensitivity	Specificity	Precision	F Measure
(100, 50)	Train	91.4%	8.5%	46.1%	97.2%	67.5%	54.8%
	Test	91.1%	8.7%	44.7%	96.9%	63.9%	52.6%
(200, 100)	Train	91.1%	8.7%	42.4%	97.3%	66.8%	51.9%
	Test	91.0%	8.8%	41.4%	97.1%	63.8%	50.2%
(500, 250)	Train	90.9%	9.1%	37.2%	97.7%	67.9%	48.1%
	Test	90.9%	9.0%	36.4%	97.6%	65.4%	46.8%
(1000, 500)	Train	90.5%	9.3%	28.3%	98.5%	71.1%	40.5%
	Test	90.6%	9.2%	28.4%	98.4%	68.1%	40.1%

Figure 3: Model Evaluation Metrics

Decision Tree

ROC Curve

- The ROC curve is plotted in function of sensitivity (y-coordinate) and 1-specficity (x-coordinate) for different thresholds (figure 4).
- From the ROC curve, the area under the curve is .795 with 95% confidence interval.
- Also, the area under the curve is significantly different from 0.5, which means that the decision tree classifies the group significantly better than by chance

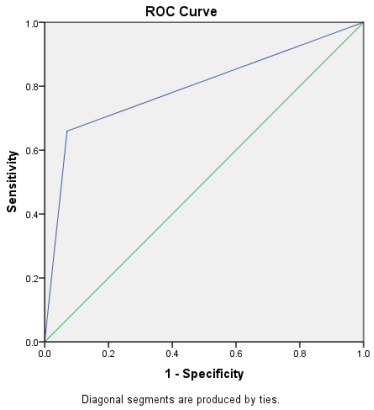


Figure 4: ROC Curve

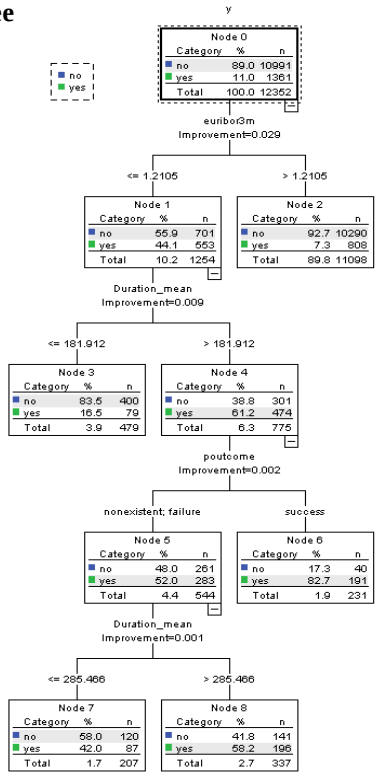
Final Model (Decision Tree)

- The final model was built with the number of records in parent and child nodes of 1,000 and 500 with the maximum level of depth of 20.
- The dependent variable is a binary variable, y (Has client subscribed for a term deposit?). The independent variables include 6 categorical variables and 7 numeric variables, which includes Poutcome, Month, Job, Contact, Day of Week, Education, age_mean, campaign_mean, duration_mean, emp.var.rate, cons.price.idk, cons.conf.idk, and euribor3m.
- The final model contains 9 nodes, which 5 nodes are terminal nodes which generate 5 classification rules (figure 5).

Figure 5: Final Decision Tree

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	y
	Independent Variables	poutcome, job, education, contact, month, day_of_week, Duration_mean, age_means, campaign_means, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	1000
	Minimum Cases in Child Node	500
Results	Independent Variables Included	euribor3m, cons.conf.idx, cons. price.idx, emp.var.rate, month, poutcome, Duration_mean, job, education, contact
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	4



Final Model (Decision Tree)

- The most three important attributes are euribor3m., Poutcome, cons.conf.idk (figure 6).
- The least important attributes are job and education level of clients (figure 6).
- The classification matrix is shown in figure 7. The final model has a high accuracy rate of 90.6 % (test set) and a high specificity rate of 98.4%, indicating that it can accurately recognize the negative ("no") tuples.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
euribor3m	.029	100.0%
poutcome	.023	79.1%
cons.conf.idx	.014	49.5%
emp.var.rate	.012	42.6%
Duration_mean	.010	35.3%
cons.price.idx	.010	34.3%
month	.009	30.8%
contact	.001	3.0%
job	1.522E-5	0.1%
education	8.610E-6	0.0%

Growing Method: CRT
Dependent Variable: y

Figure 6: Variable Importance Table

Classification

Sample	Observed	Predicted no	Predicted yes	Percent Correct
Training	no	25180	377	98.5%
	yes	2350	929	28.3%
	Overall Percentage	95.5%	4.5%	90.5%
Test	no	10810	181	98.4%
	yes	974	387	28.4%
	Overall Percentage	95.4%	4.6%	90.6%

Growing Method: CRT
Dependent Variable: y

Figure 7: Classification Table

Model Fitting II

K-Nearest Neighbors

Total observations were spitted into a training set and a test set using holdout partition method. A training set contains 28,836 records or 70% of the total records. A training set is used to build models. A test set contains 12,353 records or 30% of total records, used for model evaluation.

- The two KNN models was built using y as a dependent variable. The independent variables include both categorical variables and numeric variables which were normalized to remove an effect of the differences in scale and prevent distance measures from being dominated by the attributes with larger value of range.
- The two KNN models were built with selected range of k from 2 to 8 where k is the number of the nearest neighbors used to classify the new records.
- The first KNN model was developed using 6 categorical variables and 7 numerical variables.
- The second KNN model was built using a reduced set of attributes containing 6 categorical variables and 3 principal components
- We chose the smallest value of k which has the lowest error rate.
- The Euclidean metric is selected to measure the distance of records as a straight-line distance based on all dependent variables.

KNN model 1

- The target variable is a binary variable (Has the client subscribed for a term deposit?)
- The dependent variables contain 6 categorical variables (job, education, poutcome, contact, day, month) and 7 numerical variables (age_mean, campaign_mean, duration_mean, emp.var.rate, cons.price.idk, cons.conf.idk, euribor3m)
- Total records were spitted into a training set and a test set using a holdout partition method. The model was built using a training set (70% of total records) and evaluated with a test set (30% of total records)
- The best value of k is selected using 10 folds cross validation, in which the best k for model 1 is equal to 8 which is the smallest number of k that has the smallest error rate.
- As shown in a peer chart (figure 10), the 8 nearest neighbors are used to classify the new records.
- The KNN model 1 has an accuracy rate of 90.4% (test set), the error rate of 9.6%, the specificity of 98.1%, and the sensitivity of 26.6%. The classification table (figure 12) indicates that the KNN model 1 can accurately recognize "no" tuples but has a less ability to detect "yes" tuples.

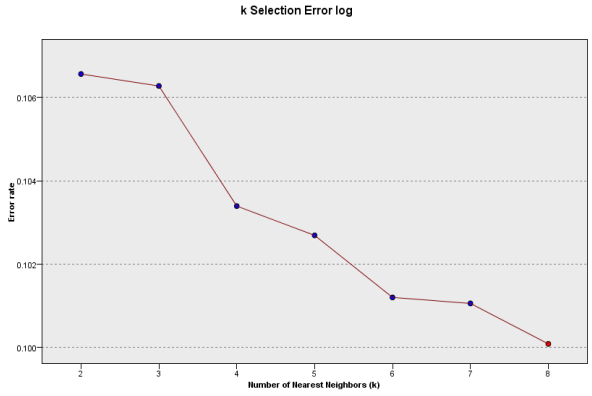


Figure 8: Error Rate of Different K

Classification Table

Partition	Observed	Predicted		
		no	yes	Percent Correct
Training	no	23934	457	98.1%
	yes	2286	815	26.3%
	Overall Percent	95.4%	4.6%	90.0%
Holdout	no	10301	197	98.1%
	yes	931	337	26.6%
	Missing	0	0	
	Overall Percent	95.5%	4.5%	90.4%

Figure 9: Classification Table

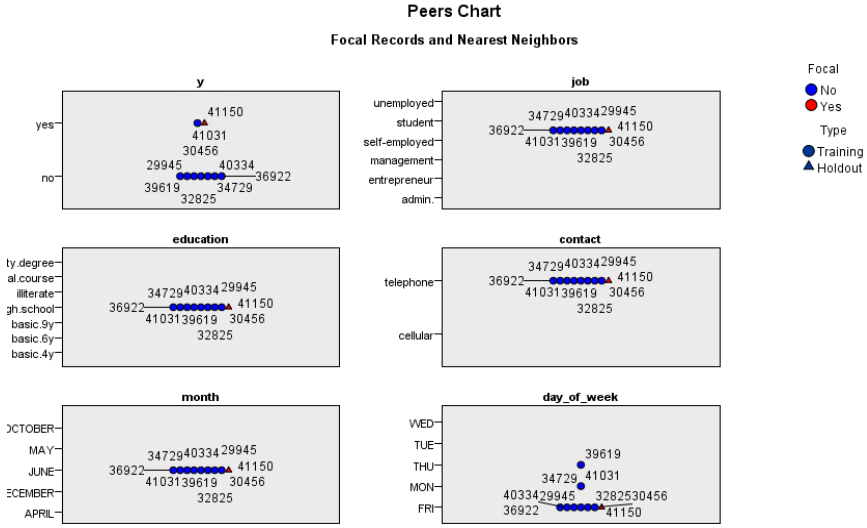


Figure 10: Peer Chart for k = 8

KNN model 2

- The target variable is a binary variable (Has the client subscribed for a term deposit?)
- The dependent variables contain 6 categorical variables (job, education, poutcome, contact, day, month) and 3 principal components. Categorical variables were normalized by selecting normalization function in classify (nearest neighbors) function in SPSS.
- Total records were spitted into a training set and a test set using a holdout partition method.
- The model was built using a training set (70% of total records) and evaluated with a test set (30% of total records)
- The best value of k is selected using 10 folds cross validation, in which the best k for model 2 is equal to 8 because it has the smallest error rate (figure 11)
- The 8 nearest neighbors are used to classify the new records.
- The KNN model 2 has an accuracy rate of 90.2% (test set), the error rate of 9.8%, the specificity of 98.6%, and the sensitivity of 20.7%. The classification table (figure 12) indicates that the KNN model 2 has a better performance on predicting "no" tuples than predicting "yes" tuples.

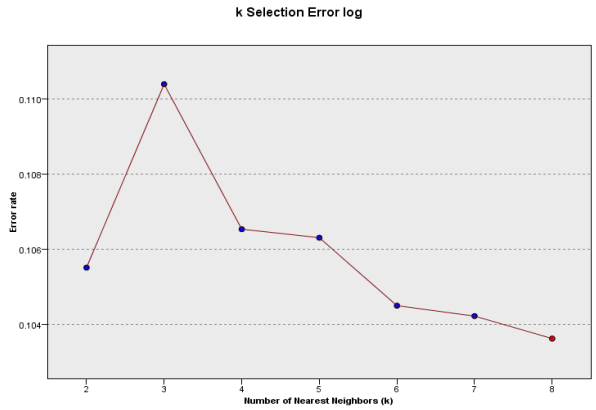


Figure 11: Error Rate of Different K

Classification Table

Partition	Observed	Predicted		
		no	yes	Percent Correct
Training	no	24041	350	98.6%
	yes	2481	620	20.0%
	Overall Percent	96.5%	3.5%	89.7%
Holdout	no	10351	147	98.6%
	yes	1005	263	20.7%
	Missing	0	0	
	Overall Percent	96.5%	3.5%	90.2%

Figure 12: Classification Table

KNN Model Comparison

Model	Partition	Accuracy	Error	Sensitivity	Specificity	Precision	F Measure
KNN 1	Train	90.0%	10.0%	26.3%	98.1%	64.1%	37.3%
	Test	90.4%	9.6%	26.6%	98.1%	63.1%	37.4%
KNN 2	Train	89.7%	10.3%	20.0%	98.6%	63.9%	30.5%
	Test	90.2%	9.8%	20.7%	98.6%	64.1%	31.3%

KNN model 1 performs slightly better than model 2 as it has a higher accuracy, sensitivity, and F-measures. The specificity and precision rate of model KNN1 are close to those of model KNN 2. So we selected model 1 from KNN to compare with the final model from decision in the following section.

Final Model Comparison

Model	Partition	Accuracy	Error	Sensitivity	Specificity	Precision	F Measure
Decision Tree (1000, 500)	Train	90.5%	9.3%	28.3%	98.5%	71.1%	40.5%
	Test	90.6%	9.2%	28.4%	98.4%	68.1%	40.1%
KNN 1	Train	90.0%	10.0%	26.3%	98.1%	64.1%	37.3%
	Test	90.4%	9.6%	26.6%	98.1%	63.1%	37.4%

The model from decision tree performs slightly better than model from KNN as it has a higher accuracy, sensitivity, specificity, precision, and F-measures. The precision rate of decision tree models is significantly higher than that of model KNN 1, So we selected model from decision tree as our final model to be used in our analysis and to predict the target variable for new records.

Conclusion

Data mining can provide help in diversity marketing strategies, and its applications can be influential when having complex data and large procedures. It also has the ability to reduce the number of negative decisions. In this paper, we find out bank performs a direct marketing campaign based on phone calls to stimulate clients to subscribe a term deposit. The target variable is a binary variable which indicates whether or not the clients will subscribe for a term deposit after getting phone calls during the campaign period. The classification models were built using both Decision Tree and K-Nearest neighbors method. All models were evaluated based on their performance on the test set. Our team selected the final model from decision tree with contains 5 terminal nodes. The final model exhibits a high accuracy and precision. We use the final model to classify the new records and identify important attributes.

- The most important attributes are euribor3m. (averaged interest rate that eurozone bank offers an unsecured funds to other banks), Poutcome (the fail or success outcome of previous campaign), cons.conf.idk (consumer confidential index reported monthly which is an economic indicators)
- The least important attributes are job and education level of clients.
- The campaign and economic attributes have more influence on whether or not the client will subscribe for a term deposit after getting contacted (target variable)

Recommendation

From our analysis, the campaign and economic attributes have more effects on target variable than the client's characteristics. Thus, bank should consider the economic circumstances and tailor their campaign to increase the effectiveness of their marketing campaign which will increase the chance of getting more subscriptions for a term deposit.

Future Studies

The target variable has a problem of imbalanced class which causes our model to have less ability to recognizing the "yes" records. For the future studies, the Under-sampling is needed to randomly eliminate tuples from negative ("no") class. This will help to solve the class imbalanced problem, which will increase the ability of our model to recognize the "yes" records.