DS-GA-1001 Term Project Report

Group Member:

Yihong Zhou (yz4552@nyu.edu) NetID - yz4552
Ziwei Wang (zw1365@nyu.edu) NetID- zw1365
Phu Mon Htut (pmh330@nyu.edu) NetID: pmh330
Yuqiong Li (yl5090@nyu.edu) NetID - yl5090

Team Name

99.9

Business Understanding

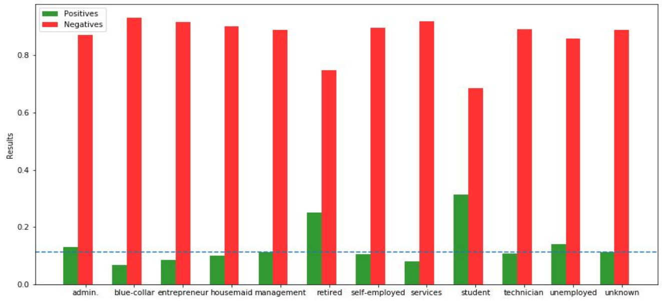
A Portuguese bank institute plans to run phone marketing campaigns to persuade customers to subscribe its term deposits. Success rate for such campaign is usually as low as 11%. Often, more than one contact to the same clients is made. Making phone calls not only requires manpower, money and time, but also has drawbacks including leading to negative customer attitudes and intrusion of customers' privacy.

The bank wants an accurate model for targeting customers so that it can gain the constant support from old customers and win term deposits market from new customers with minimized resources. The business problem is whether it is possible to predict the outcome of phone calls based on prospects' demographical information beforehand. It is important because then fewer numbers of contacts are needed to achieve the same number of subscriptions.

Data Understanding

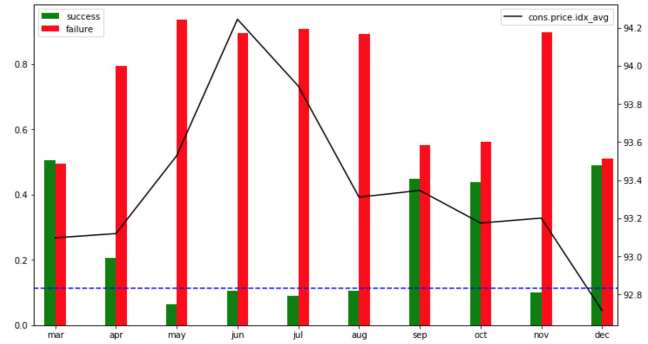
Data for this campaign comes from UCI machine learning repository. It can be formulated as a binary classification problem with target variable being positive (subscribe) and negative (not-subscribe). There are twenty feature variables as shown in the appendix. Basically, all the features could be categorized into four parts.

The first part contains information on clients' background, such as age and education and the second part contains information regarding the social and economic context, such as consumer price index and employment variation rate.



On the one hand, different customers with different backgrounds have different incentives to subscribe the term deposit. As the Figure (Figure 1) shows, students and retirees have much higher term-deposit rates than the rest of the groups.

Figure 1. Background vs Positive/Negatives



On the other hand, Figure (Figure 2) shows that term-deposit rates over different months are heavily correlated with the consumer price index. Thus, social and economic context also play important role in subscription outcome.

Figure 2. CPI vs Positive/Negatives

Therefore, this dataset is actually offering two different perspectives, one from customers' standing point and other from the social standing point.

Another two parts are about the campaign itself, one about the past campaigns and the other about the current campaign. Information on the past campaigns contains features like 'previous', a number of contacts performed during past campaign for one specific client, while information on the current campaign records the data, such as 'contact' and 'duration'.

It is important to note that 'duration' variable highly affects the output result because at the end of the call, the result is obviously known. The graph below (Figure 3) shows when duration increases, their success rate increases correspondingly. Especially, when duration goes beyond 260s(around 5mins), namely the phone call is conducted in a proper length, the success rate is weighted greater than the average success rate. It is obvious that human resources must play an important role in persuading customers. However, the duration is never known before a call is performed.

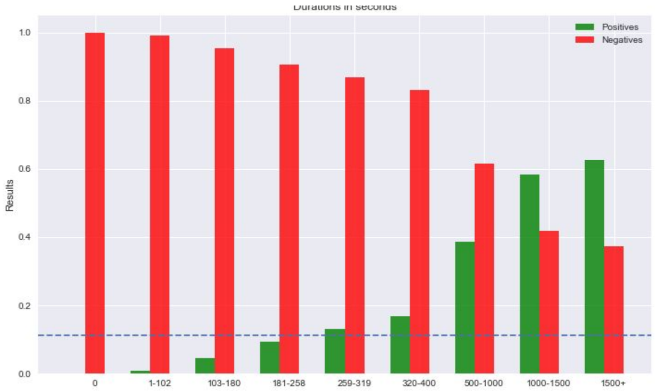


Figure 3. Duration vs Positive/Negatives

Data Preparation

This dataset contains two different categories of missing values. When the features are categorical, they are encoded as 'unknown'. For example, 'job', 'default' and 'education' variables all have 'unknown' values. Besides these, only one numerical feature in this dataset contains the missing value, which is 'pdays', defined as the number of days that passed by after the client was last contacted from the previous campaign. When 'pdays' is 999, it means that client was not previously contacted.

When the data contains the missing value, the first task is to check missing at random by computing the proportion of unknown values for each categorical variables over all observations.

variable	Y = 1 when variable unknown	unknown_portion
job	0.112121	0.008012
marital	0.150000	0.001942
education	0.145003	0.042027
default	0.051530	0.208726
housing	0.108081	0.024036

loan	0.108081	0.024036

Table 1. Unknown variables' relation with Response Y

As the table shows (table 1), the proportions of unknown values for 'loan' and 'housing' are exactly same. More importantly, the proportions of positive observations given that housing and loan are unknown, are also same. These two facts highly suggest that the dataset is not missing at random.

#[loan = unknown, housing = unknown]	990
#[loan = unknown, housing = unknown, pdays = 999(i.e. Never contacted before)]	956

Table 2. Unknown Comparison between related features

The reason why it is "not missing at random" is that they are the customers who were contacted for the very first time during the entire campaigns. 956 out of 999 observations whose housing and loan are both unknown, are never previously contacted by the bank. Because the bank had never talked to them before, the bank did not know them very well. With this regard, binning the categorical variables into dummy variables to indicate the missing should be a properer data preprocessing.

For another variable, 'pdays', which also contains missing values 999, the strategy is to remove it. The importance of 'pdays' is that it can indicate whether the customers are firstly reached and thereafter give sufficient sights on randomness for unknown values. However, When 'previous' variable(number of contacts performed before this campaign for this client) is zero, or 'poutcome'(the outcome of the previous marketing campaign) is 'nonexistent', it also suggests that the client was entirely new to the bank. Thus, it is sufficient to remove the 'pdays' variable since the other two variables are giving duplicate information.

The final step is to check the multicollinearity between numerical variables and normalize them.

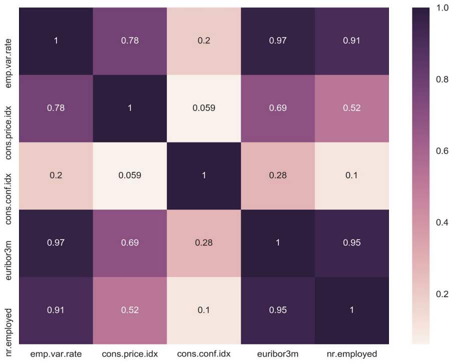


Figure 4. Multicollinearity between numerical variables

Based on the heatmap (Figure 4), 'euribor3m', 'emp.var.rate' and 'nr.employed' are heavily correlated, with 0.97, 0.95 and 0.91 correlation, so one or two correlated variables might be dropped to see whether the models will be improved in the model evaluation part.

The numerical variables, except for ‘age’, ‘campaign’ and ‘previous’, are normalized for model analysis. Finally, the proposed data ended up with 60 variables due to binning transformation.

Modeling Understanding

The following eight machine learning algorithms are used:

Naive Bayes:
It is a simple, computationally inexpensive model. However, Naive Bayes assumes conditional independence of classes, which is violated in our data as the social economic features are correlated with each other. Thus, the performance of this model might not be as competitive as other models.

K Nearest Neighbors:
The advantages of KNN models are that they are robust non-parametric models that work well with non-linear features, and relatively simple to implement and analyze. However, they require high memory, longer computation time, and do not work well with a large number of features.

Logistic Regression:
Unlike Naive Bayes, Logistic Regression accounts for interactions between features. In addition, it can be tuned to generalize better as there are a few ways to regularize it. It also provides the probabilistic interpretation of parameters, unlike decision trees.
However, it does not work well with non-linear features. Moreover, it’s convenient to learn about new data using stochastic gradient descent.

Neural Networks:
Neural Networks are powerful function approximators that need little feature selection and normalization. However, they are computationally expensive and can overfit the data if the data set is small. As it is a large dataset, Neural Networks should work very well.

Decision Tree:
Decision Tree is easy to interpret, fast and scalable. They are also robust as they are the non-parametric model and can be applied to nonlinear features. However, they can easily overfit to training data if the number of parameters is not restricted. Additionally, they won’t support online learning, so have to rebuild the tree when new examples come on.

Random Forest:
The drawback of Decision Trees can be adjusted by using ensemble methods like Random Forest. The pros of random forest are that it is a bagging method that improves the performance of basic Decision Tree models. However, it is computationally expensive and hard to interpret.

XGBoost:
XGBoost is a decision tree gradient boosting method that is designed to be computationally scalable and regularizes better. Thus, the pros are that it’s more computationally efficient, generalizes well and more accurate. However, it’s also harder to analyze and interpret.

Support Vector Machine :

The Pros of SVM are that they have High accuracy, are well-suited for smaller datasets, nice theoretical guarantees regarding overfitting, and they work well for nonlinear feature space with an appropriate kernel and can handle high dimension of data.
The cons are that SVM is memory-intensive, hard to interpret, and not suited for a large dataset because the training time can be very high.

Modeling and Evaluation:

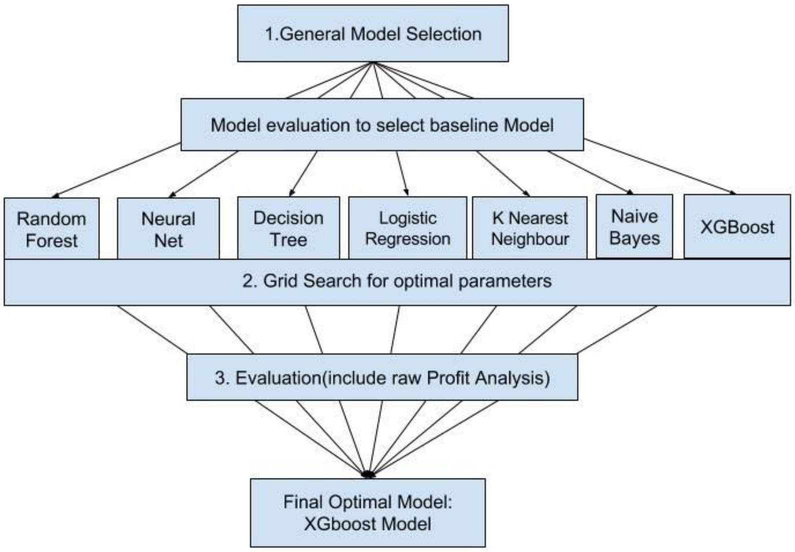

1.General Model Selection
Model evaluation to select baseline Model
Random Forest
Neural Net
Decision Tree
Logistic Regression
K Nearest Neighbour
Naive Bayes
XGBoost
2. Grid Search for optimal parameters
3. Evaluation(include raw Profit Analysis)
Final Optimal Model: XGboost Model

Figure 5. Overall Model Analysis Procedure

Model analysis and selection is a two-stage process. The optimal hyperparameters are selected from each model family by using grid search method. AUC performs as an evaluation metric for this stage. In the second stage, with tuned parameters models goes through a general evaluation framework including ROC curve, AUC, raw profit matrix, MSE, F1 score, Precision to compare model performances.

1.Baseline Selection

After polishing the dataset, a general model selection is tried to produce the potential candidate of the optimal models. The cleaned and preprocessed train data directly feed into the models pool containing 8 models with default parameters: Support Vector Machine, Naive Bayes, Neural Net, Decision Tree, Random Forest, Logistic Regression, K Nearest Neighbour and XGBoost classifier. Using AUC in baseline selection is a good alternative to the standard empirical risk (classification error) and it gives an intuitive result of how good default models are trained.

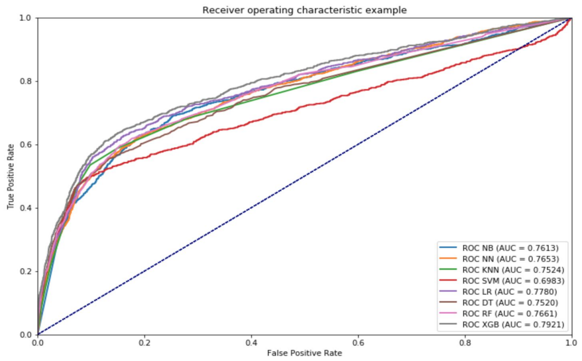


Figure 6. First Round AUC for 8 models: Support Vector Machine, Naive Bayes, Neural Net, Decision Tree, Random Forest, Logistic Regression, K Nearest Neighbour, XGBoost classifier

Except for the SVM, the other 7 models have pretty close AUC results(Figure 6). While the XGBoost has highest AUC 0.7921, the SVM runs under 0.7. Thus, the SVM becomes the baseline model for all the other models.

2.Optimal Hyperparameter Selection

We used K-fold function to cross-validate 5-fold on train data and GRID Search on top 7 models to find the best hyperparameters for each of the models. The criteria for choosing hyperparameter is using AUC metric. In this case, AUC produces clearer comparisons between the performance of different permutations of hyperparameters and chooses the best combination for the corresponding models. This will improve the performance of models besides the baseline.

3. Evaluating top models

3.1 Evaluation based on model performance.

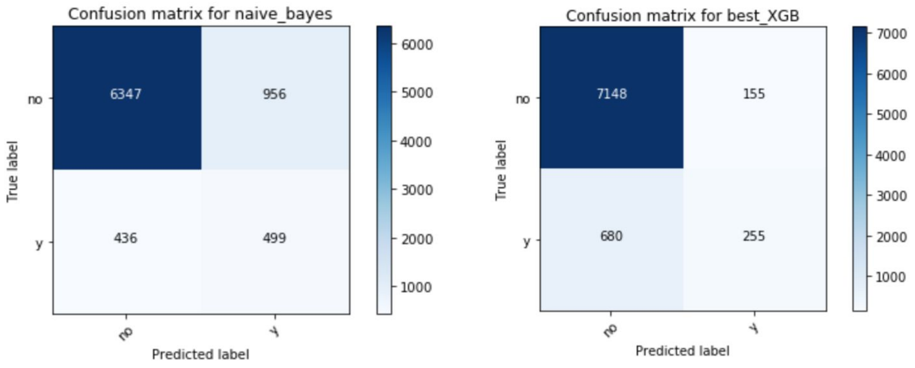


Figure 7.Confusion matrices for Naive Bayes and XGBoost Classifier

Plotted in Figure 8, the performance of all models are compared. From pure evaluation aspects, the XGBoost Model results in the best to fit the test data and it has the highest precision (Figure 15). However, Naive Bayes gets the most of true prediction right in confusion matrix, that is the model predicted subscribe and the customer indeed subscribed, and highest F1 Score, shown in Figure 7. The suspicion of the neural net not doing well is it works too well for training data so that the model overfits.

Logistic regression, Decision Tree, Random Forest and XGBoost perform robustly and rank at the top from error, variance and accuracy perspectives among all models.

	Logistic regression	K Nearest Neighbour	Decision Tree	Random Forest	Neural Net	Naive Bayes	XGBoost
MSE	0.1034	0.1092	0.1066	0.1021	0.1135	0.169	0.1014
F1 Score	0.2934	0.2695	0.3860	0.3362	0.1226	0.4176	0.3792
Precision	0.6604	0.5589	0.5576	0.6289	0.5872	0.3430	0.62
Sensitivity	0.1869	0.1775	0.295	0.2139	0.684	0.5336	0.2727

Table 3. Prime evaluation matrix and its result for 7 models

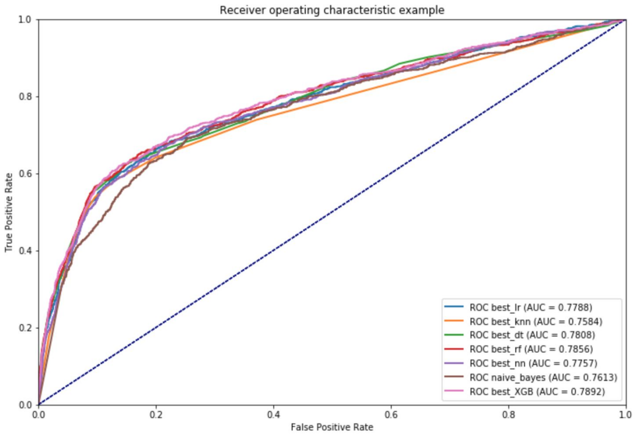


Figure 8 Second Round(with parameter tuning) AUC for 7 models:Naive Bayes, Neural Net, Decision Tree, Random Forest, Logistic Regression, K Nearest Neighbour, XGBoost classifier

Overall from Figure 12 to Figure 15, and confusion matrix in Appendix, the XGBoost is the optimal choice for modelling because it has a relatively low Mean Square Error, high F1 score, and the best learning ability. It outperforms not only other models, but also the baseline of Naive Bayes very much about 13% in AUC after adjusting the parameters.

However, the model scores do not directly reflect their appropriateness for the business situation. The realistic business situation will be considered in the next evaluation.

3.2 Evaluation in business sense.

The cost matrix for the Portuguese banking institution business is defined in this section. The model should be heavily rewarded for true positive predictions; for every false positive prediction, the company will lose -1 per call for calling the customer who will not subscribe. When model predicts no subscription for the customer, the bank is assumed to not call. Thus, there is no any spends or loss for the bank.

Raw Profit Matrix:

	True Label: Yes	True Label: No
Predict: Yes	99	-1
Predict: No	0	0

Table 4.Raw Profit Matrix

The profit-cost table will be as below when test data of size = 8238 is put into models:

	Logistic Regression	K Nearest Neighbour	Decision Tree	Random Forest	Neural Network	Naive Bayes	XGBoost
Total Calls Made	269	297	495	318	109	1455	410
Raw Profit	17331	16303	27105	19681	6291	48445	25090
Success Rate (Precision)	66.04%	55.89%	55.76%	62.89%	58.72%	34.30%	62%
Profit per phone call	64.42	54.89	54.75	61.89	57.71	33.3	61.20

Table 5. Raw Profit-cost table

The sampling process for each customer call follows Bernoulli trials with around 11% success rate of converting the prospects into real customers. Thus, the expected return of probabilistic model follows the following formula:

$$Expected\ Profit\ per\ call\ = Gain\ per\ subscription\ \times\ subscribe\ rate\ +\ Loss\ per\ subscription\ \times\ not\ subscribe\ rate$$

In this case, the expected return per random call is $\$99 \times 11\% + (-1\$) \times (1 - 11\%) = \$10$
Namely, in the past, every random call can give the bank $10 return. However, the expected profit per call for all predictive models dramatically exceeds $10, as the range of expected profit for all predictive model is from 33.3 to 64.42.

In terms of raw profit, Naive Bayes performs the best. However, it achieved the highest raw profit by making large numbers of outgoing calls. Profit per phone call of Naive Bayes is only about 33.3, much lower than those produced by other models, so it is not a very efficient method.

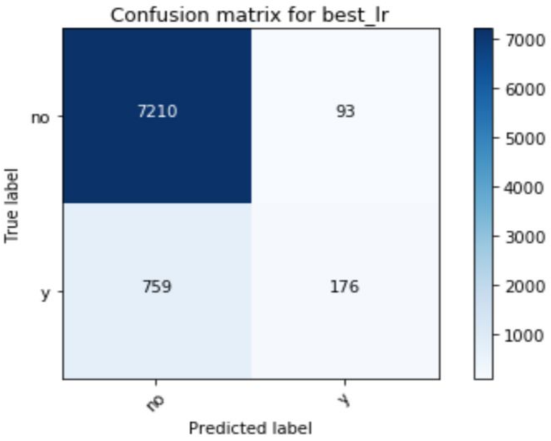
The Portuguese banking wants to attract more term deposit while spending less time and money on the campaign. In this regard, the bank should rise up success call rate and increase the profit per phone call. Thus, the bank is highly recommended to use the logistic regression to maximize the profit in the market campaign.

Deployment

Data mining can significantly help Portuguese banking target the most potential customers while making a fewer number of contacts. The subscription rate of campaign outcome is around 11%. However, with the help of the predictive models, the bank can eventually achieve around 34~66% of success rate.

Opportunity cost is an important issue that the bank should realize when deploying these predictive models. Economically, the opportunity cost is defined as the loss of potential gain from other alternatives when one alternative is chosen.

Figure 9 Confusion matrix for best logistic regression



Take the logistic model with the optimal parameters as an example. It has highest precision rate and profit per phone call among all the models. However, at the same time, it has a pretty high false negative rate as well. That's to say, this model is too prudent to categorize an observation as a potential client. When this model is deployed, it causes a large number of real customer loss.

There are two kinds of costs the bank needs to evaluate. Phone calls are made when the customers are predicted to be 'positive'. Firstly, when phone calls are made to the people who will not subscribe, the cost is the investment of human labours, who takes time to call and persuade. Secondly, when models misclassified customers who will subscribe as 'negative', the cost is the loss of potential gain.

To mitigate the opportunity cost due to False Negative, the model will instead also put a cost of 10 to represent the influences caused by mislabelling the 'true customers'.

New Profit Matrix:

	True Label: Yes	True Label: No
Predict: Yes	99	-1
Predict: No	-10	0

Table 6. New Profit Matrix with opportunity cost

The profit considering opportunities cost table will be as below when test data of size = 8238 is put into models:

	Logistic regression	K Nearest Neighbour	Decision Tree	Random Forest	Neural Network	Naive Bayes	XGBoost
Total Calls Made	269	297	495	318	109	1455	410
New Profit	9741	8613	20515	12332	-2419	44085	18290
Sensitivity	18.69%	17.75%	29.5%	21.39%	6.84%	53.36%	27.27%
New Profit per phone call	36.211	29.0	41.19	38.78	-22.19	30.30	44.609

Table 7. Opportunities Cost considered Profit-cost table

Notice that Neural network even has negative profit in this case because it misclassified nearly 93.16% of real customers. Thus, its opportunity cost is extremely high. Even though a good precision model can help the bank seize the clients efficiently, the bank does not want to lose the customers, who will actually subscribe term deposits but are missed by the predictive models either. Therefore, based on this trade-off, XGBoost models actually work best with both high precision and sensitivity scores.

More importantly, the data mining shall enlighten the bank on conducting marketing strategies. Treasure the old customers who are 6 times more likely to subscribe compared to new customers. Promotion to prevent current customers from churning might be necessary. Do not bother same clients another time if they have already refused to subscribe term deposit for the first 3 or 4 times. Basically, making phone calls to the clients like that is a waste of resource, and more importantly, is not good for bank's public relationship.

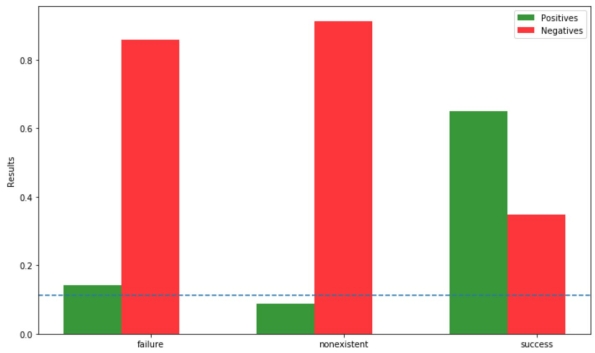


Figure 10. Last campaign results vs Positive/Negatives

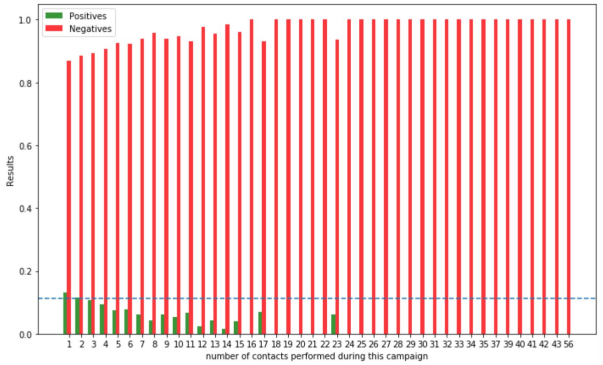


Figure 11. Number of Calls vs Positive/Negatives

Furthermore, one ethical problem arises in considering of the number of calls to the same clients. As the graph above shows(Figure 11), some customers were constantly reached more than 20 times. Cold Calling customers is a good marketing strategy. However, when the number of call to same clients goes beyond a certain limit, the marketing campaign is actually causing annoyance to the people who are constantly contacted no matter how many times they have already refused. Thus, the bank shall enforce a call limit. The campaign will never call the same client again once the number of rejections has reached that limit, such as 10 times. Or perhaps, at the end of every call, it should give customers an option to be not contacted any longer.

To sum up, the predictive models indeed help the bank achieve the highest profit return by using minimized resources. Within various business scenarios, different predictive models should be applied. If the bank cares more about the current campaign subscription benefit, the logistic regression model is recommended as it has maximum profit per call, even without thinking about loss risk in the future. On the other hand, if bank counts the opportunity cost and wants to seize as many as real customers it can, XGBoost is a favoured choice in this situation. From a practical business view, XGBoost is the best choice for Portuguese bank institute.

Appendix

Input variables:
# bank client data:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes
16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)
17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric)
20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

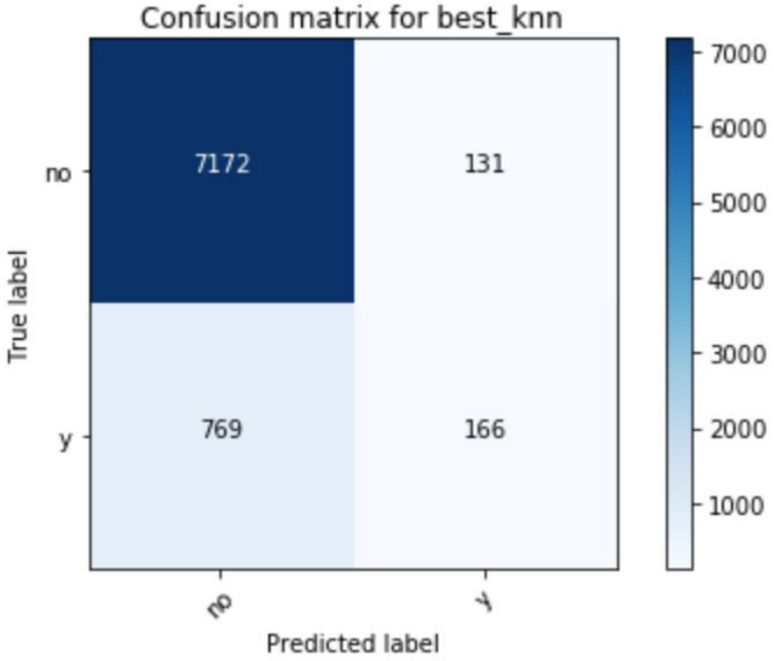


Figure 12: Confusion matrix for KNN model (above).

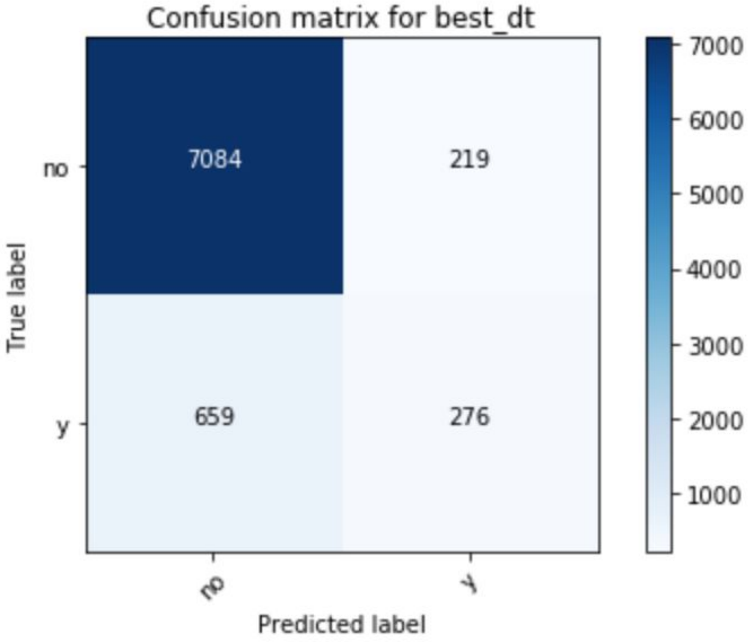


Figure 13: Confusion matrix for Decision Tree model (above).

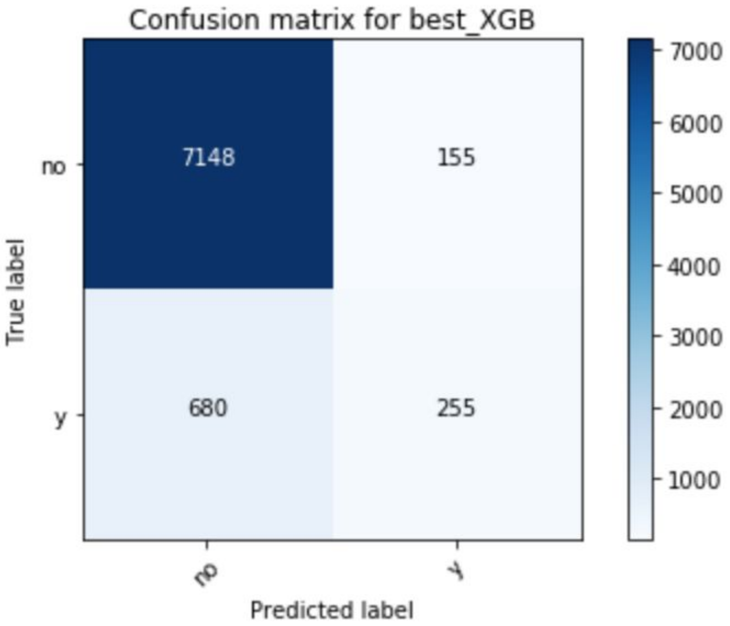


Figure 14: Confusion matrix for XGboost

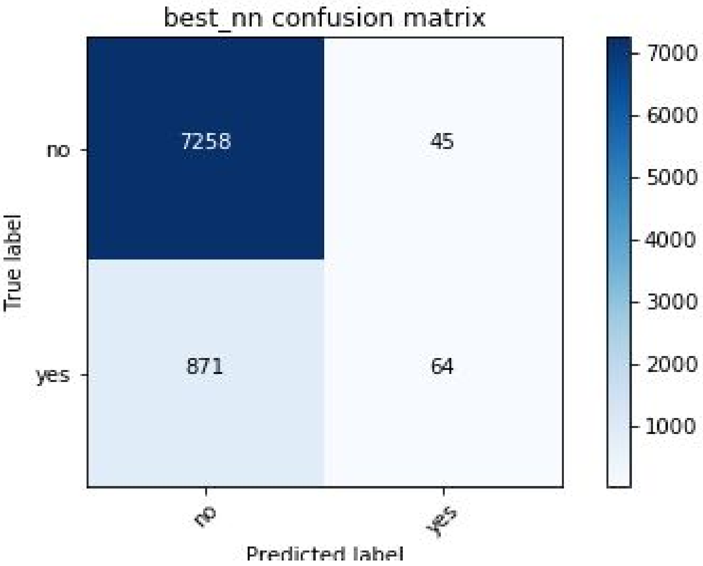


Figure 15: Confusion matrix for Neural Network