Huaqi Nie
Banner ID: B01694250

DATA1030 Final Project Report

Introduction

In this project, I'm going to make a classification model to predict if the client will subscribe a term deposit (variable y). It may help the bank to have a clear view about what factors would be important to affect the client's decision, and to improve its marketing strategy.

The data subset from the original dataset, there are 41189 data points with 20 features, and the target variable is the column 'y', which contains 2 classes, 'yes' or 'no'.

	Features	**Description**
Bank client data:	Age	Client's age
	Job	Client's type of job
	Marital	Client's marital status
	Education	Client's education level
	Default	Has credit in default?
	Housing	Has housing loan?
	Loan	Has personal loan?
Related with the campaign:	Contact	Contact communication type
	Month	Last contact month of year
	Day_of_week	Last contact day of the week
	Duration	Last contact duration, in seconds
	Campaign	Number of contacts performed during this campaign for this client
	Pdays	Number of days that passed by after the client was last contacted from a previous campaign
	Previous	Number of contacts performed before this campaign for this client
	Poutcome	Outcome of the previous marketing campaign
Social and economic context attributes	Emp.var.rate	Employment variation rate
	Cons.price.idx	Consumer price index
	Cons.conf.idx	Consumer confidence index
	Euribor3m	Euribor 3 month rate
	Nr.employed	Number of employees

This dataset has already been used to created several classification models on Kaggle. It is common that these models uses all 20 features and the best model has the highest score 0.914306, measured by the area under the ROC curve. My goal is to adjust and select features to build a better model to predict the clients' decision.

Exploratory Data Analysis

In this section, I try to discover the relationship between the features and the target variables. The following plots would be helpful:

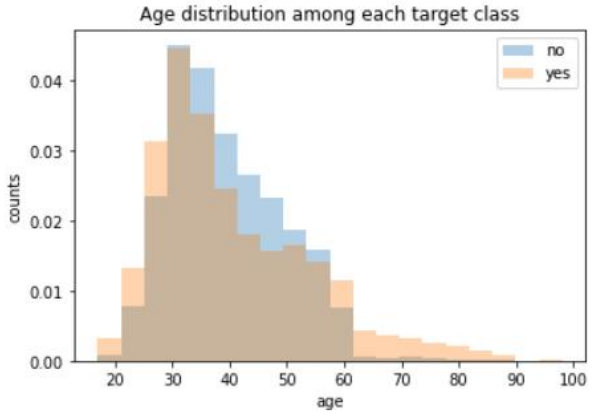


Figure 1: x-axis is age and y-axis is the count of each class of client's response.

From the plot, we could see that the younger people and the elder people tends to subscribe the campaign.

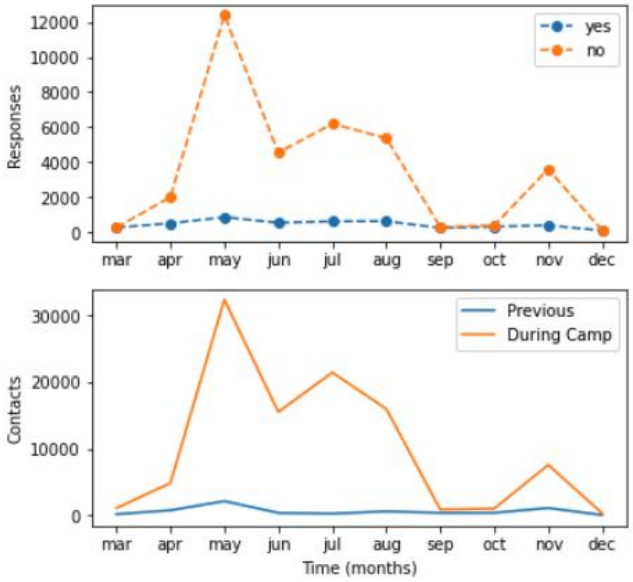
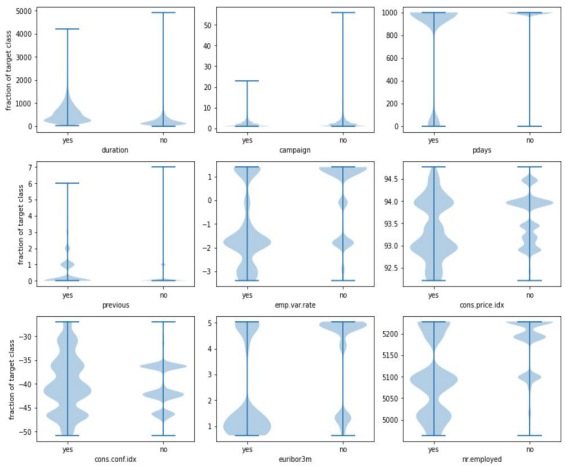


Figure 2: the x-axis is the month when the bank has the last contact to the client, the y-axis in the upper plot shows the count of each class of client's response, and the y-axis in the lower plot shows the count of contact number of contacts performed during this campaign and the contact number of contacts performed before this campaign for the client.

It seems the the different type of contact number could be a good indicator to predict clients response.



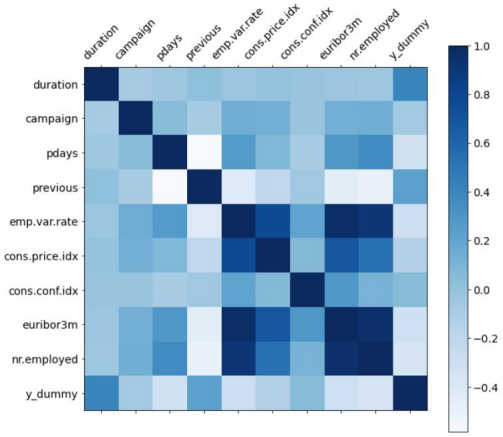


Figure 3: the violin plot shows the distributions of each continuous features among two classes in clients response(target).

Figure 4: the heatmap shows correlation matrix among each continuous features and the clients response(target).

From these two figures, we could see that the violin plots represents that every features have different distributions among two classes, which indicates there should be some relation between these features and the target variable. However, the heatmap shows the correlations of the continuous features and the target variable are very week. This contradiction between the two plots indicates that the target variable cannot be determined by only one features, and that's why machine learning is necessary to solve this classification problem.

Methods

In this section, I would present how I split and preprocess my data, and what machine learning algorithms I have used to build the classification model.

Data Splitting

I would like to split my data into 90% train and 10% test set, so that I could apply cross validation in my following experiments. Also, it is a stratified split, because the dataset is imbalanced and I want to make sure that every class in the target variable could keep certain proportion in each subset.

Data Preprocessing

Encoders	**Features**
Ordinal Encoder	Education
One-hot Encoder	Job, Marital, Default, Housing, Loan, Contact, Month, Dayofweek, Poutcome
Mean-Max Scaler	Age
Standard Scaler	Duration, Campaign, Pdays, Previous, Emp.var.rate, Cons.price.idx, Cons.conf.idx, Euribor3m, Nr.employed.
Label Encoder	y (target variable)

After preprocessing, the feature number increases from 20 to 56.

Machine learning algorithms

In this section, I would show you several machine learning algorithms that I choose to build the classification model and how tune them to get the best parameters:

ML algorithms	**Parameters**
Logistic Regression	'penalty':['l1','l2'], 'C': [0.01,0.1,1,10,100,1000]
Random Forest Classifier	n_'n_estimators': [5, 10, 30,100,200] 'max_depth': [5,10,20,30]
Support Vector Machine -- linear kernel	'C': [0.01,0.1,1,10,100],
Support Vector Machine -- rbf kernel	'C': [0.01,0.1,1,10,100], 'gamma': ['scale',1,0.1,0.001]
XGB Classifier	"colsample_bytree": [0.6, 0.9], "max_depth": [1,3,10,30,100,200], "subsample": [0.66,0.8]

Evaluation metric

To choose the evaluation metrics, I consider the imbalance of the dataset, which means true-negative would be large. The f_beta score is a better choice because it is the weighted harmonic mean of P(precision) and R(recall), which do not include the count of true-negative. Also, because the main way of Bank marketing is telemarketing, which is cheap, we could like to capture more clients who would response 'yes' to the campaign. Thus, I choose f1.5 to be my evaluation metric which would capture the largest fraction of the condition positive samples even if false-positives will be large as a result.

Uncertainty

For each split in the cross validation, different random state would be assigned. For random forest, each time it could split from different points. These could cause uncertainty so I would use the mean of the test scores to evaluate the models.

Re-sample

After I implement these model to my dataset, all the models are not performing well. It could be due to the imbalance. Thus, I resample my training set, grabbing all the data points with label 1 and randomly pick same number data points with label 0, to make it 50-50 balanced.

Results

My baseline score is 0.48 if I predict all the response to class 0. To improve the score, I have tried 5 models. The following table could summarize my results.

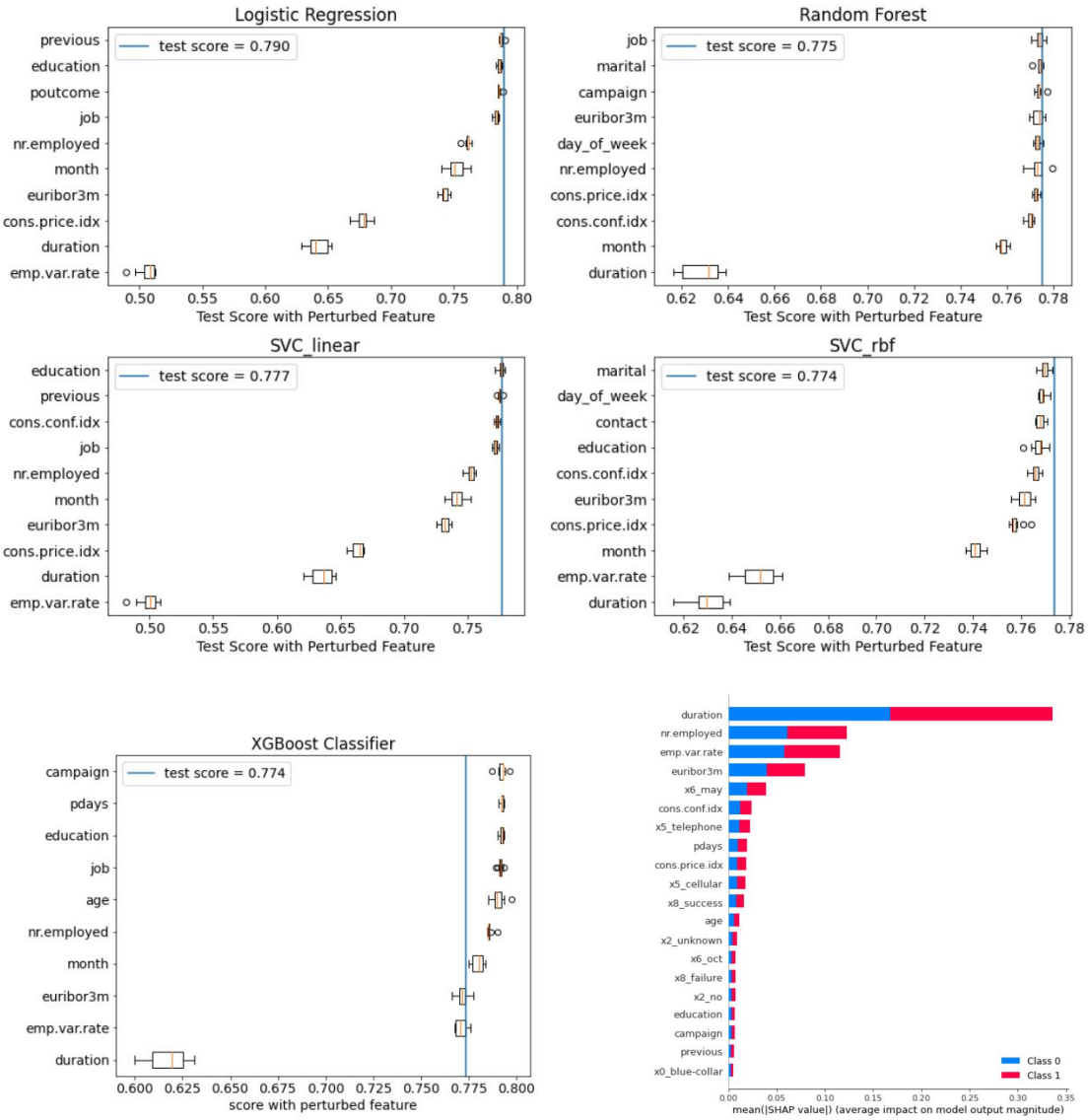


Figure 5: The global importance for each model and local importance for logistic model.

Models	**Test Score**	**Std above baseline**	**Top 3 important features**
Logistic Reg	0.790	690.31	Emp.var.rate, duration, cons.price.idx.
Random Forest	0.775	154.52	Duration, month, cons.conf.idx
SVC_linear	0.777	276.27	Emp.var.rate, duration, cons.price.idx
SVC_rbf	0.774	155.65	Duration, emp.var.rate, month
XGBoost	0.774	260.99	Duration, emp.var.rate, euribor3m

By comparing all the models, the logistic regression model gives the best test score a nd the highest number of standard deviations above the baseline. By calculating the global and local feature importance, duration and emp.var.rate could be the two most important features for the classification models, which make sense because people could tend to engage in the campaign would like to know more about it and lead to long duration, and when employee

rate is high people are more likely to subscribe their deposits. Features like education and marital status would be less significant.

It is interesting when I shuffle the features and I find for some features, shuffling them could even cause the increase of the test score, which means the machine could not learn from them very well.

Overlook

The first week spot is the point I mentioned in the end of previous section. When I use permutation to find the feature importance, I find that the models do not learn from some of the features very well. Thus, continuing to tuning the models could be a good way to improve the model performance. Also, I only shuffle single feature at each time so that the permutation importance could only show the importance of individual feature. However, one feature might appear unimportant but combined with another feature could be important. More experiment with permutation of two and more features could be tried.

It is also worth to try different techniques to improve the model, like deep learning. In the perspective of data source, collecting more relative data about the campaign might be helpful.

Reference

Bank Marketing Dataset on Kaggle:
https://www.kaggle.com/henriqueyamahata/bank-marketing

Github Link

https://github.com/Huaqi010/bank_marking_project