Yadi Yang
Banner ID: B01189775

DATA1030 Midterm Report

Introduction

The data for the project originated from a direct marketing campaign of a Portuguese banking institution. Phone communication was the main source of the campaign and often times than not, more than one contact per client is needed in order to conjure a response on whether the client would like to subscribe to the product of the bank term deposit. The intention of the entire project is classification-based in order to predict if the client will subscribe to a term deposit, which would assist in generating a larger cash flow for the bank.

Data

The target variable is whether or not the client will subscribe to a term deposit, which is the binary column of 'y' that has the classes of 'yes' or 'no'. The data that I am working with has 41188 entries with 20 features.

	Features	Descriptions	Data Type	Units
Client Data	Age	Age of the client	Numeric	0-100
	Job	Type of job	Categorical	'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown'
	Martial	Marital Status	Categorical	'divorced', 'married', 'single', 'unknown' ; note: 'divorced' means divorced or widowed'
	Education	Education Level	Categorical	'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
	Default	Has credit in default	Categorical	'no', 'yes', 'unknown'
	Housing	Has housing loan	Categorical	'no', 'yes', 'unknown'
	Loan	Has personal loan	Categorical	'no', 'yes', 'unknown'
Campaign	Contact	Communication Type	Categorical	'cellular','telephone'

Yadi Yang
Banner ID: B01189775

Information	Months	Last contact month	Categorical	'jan', 'feb', 'mar', ..., 'nov', 'dec'
	Days_of_week	Last contact day of week	Categorical	'mon','tue','wed','thu','fri'
	Duration	Last contact duration in seconds	Numeric	If duration=0 then y='no'
	Campaign	Number of contacts performed	Numeric	Numeric
	Pdays	Numbers of days that passed since the client was contacted	Numeric	999 means the client was not previously contacted
	Previous	Number of contacts performed before this campaign	Numeric	Numeric
	Poutcome	The outcome of the previous campaign	Categorical	'failure','nonexistent','success'
Social and Economic Context Attributes	Emp.var.rate	Employment variation rate - quarterly indicator	Numeric	Numeric
	Cons.price.idx	Consumer price index - monthly indicator	Numeric	Numeric
	Cons.conf.idx	Consumer confidence index - monthly indicator	Numeric	Numeric
	Euribor3m	Euribor 3 month rate - daily indicator	Numeric	Numeric
	Nr.employed	Number of employees - quarterly indicator	Numeric	Numeric

The dataset is well known on Kaggle and has been used in many previous projects for machine learning such as logistic regressions, KNN, SVM, Decision Tree, Random Forest, Naive Bayes, and Deep Learning. This dataset has also been used for deep learning techniques. I reviewed through all the models that were entered into the competition and the highest scoring one is a Gradient Boosting Model that scored 0.914306 as measured by the area under the ROC curve.

Yadi Yang
Banner ID: B01189775

Exploratory Data Analysis

I performed a thorough exploratory analysison each column of my dataset and graphed the following graphs to demonstrate some variable relationships and significance.

Figure 1

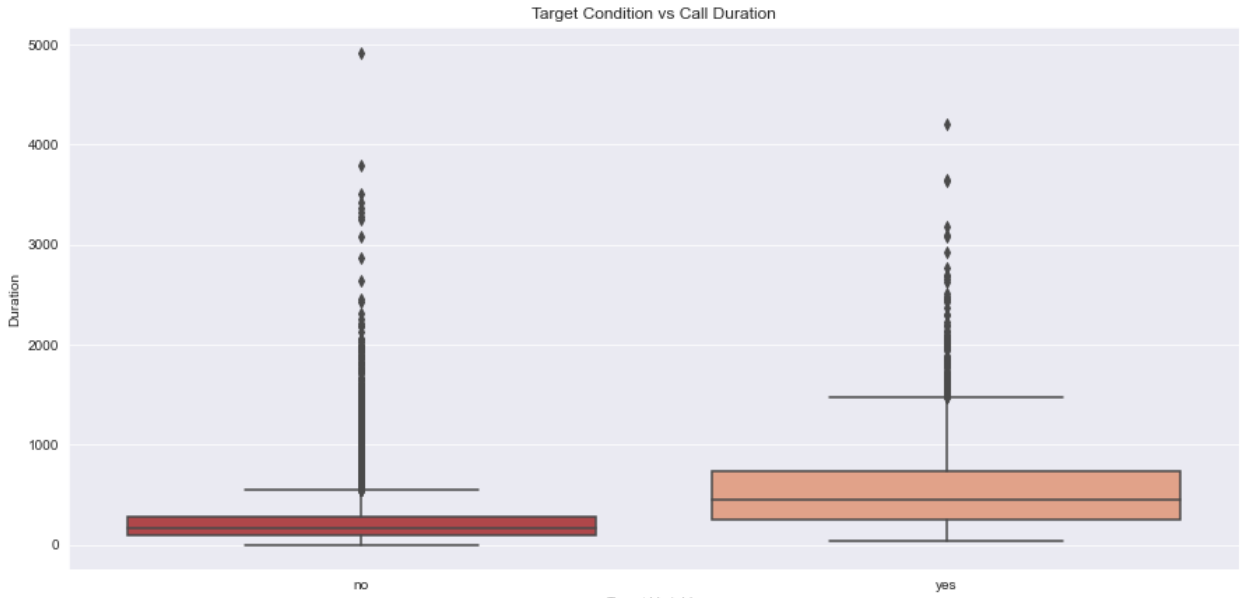

Fig.1 y is duration of the call campaign and x is target variable of yes or no. From the plot, it seems that there is the clients who respond yes have longer durations while the extremely long calls are outliers for the refusal to set up term deposit

Figure 2

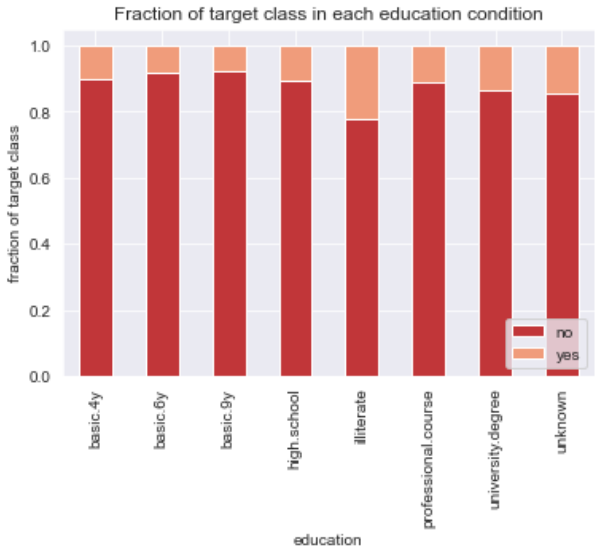

Fig.3 The X axis is education condition and y is fraction of target class.
From the plot, it seems that people who are illiterate tent to subscribe more to the term deposit

Yadi Yang
Banner ID: B01189775

Figure 3

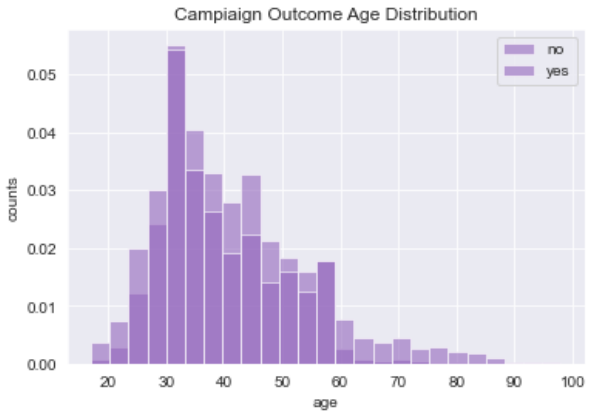

Fig.2 Since the colors are close in range, the darker ones are the ones that have both x and y. The X axis is age and y is the count of each class in target varibles. From the plot, we could see that the younger people and the elder people tends to subscribe the campaign.

Data preprocessing

My dataset is split into 90% train and 10 % test because we have more than 40,000 data points therefore 10% test is definitely enough. Usually, the more training data the better because it could assist in increasing accuracy as it is more sampling points for the model. The dataset is not independent and identically distributed because each campaign contact will affect the following contact result. In addition, the dataset utilizes a stratified split because from the EDA above, we recognize the imbalance. A stratified sampling would allow us to not oversample the minority class. The data could be clustered into age groups or groups according to other categorical variables however it is not a time series dataset. Given my machine learning question of predicting whether or not a client would subscribe to the term deposit, I should split the data like the one I have now with the 90/10 split to provide more data points for accuracy training. I also have separated the continuous features from the categorical features and applied the necessary encoder. In addition, after preprocessing, features increased from 20 to 55 from the utilization of the encoders.

Features	Encoder	Reasoning
Age	Min-Max Scaler	Age is a continuous variables and its values are reasonably bounded therefore this scaler is a good way to process this feature.
Job	One-Hot Encoder	Job is not a category that can be ranked so we use this encoder to converts categorical features into dummy arrays
Marital	One-Hot Encoder	Marital status is not a category that can be ranked so we use this encoder to converts categorical features into dummy arrays
Education	Oridinal Encoder	Since education can be ranked, we use ordinal encoder on categorical features like such as it converts categorical features into

		an integer array
Default	One-Hot Encoder	Loan default status is not a category that can be ranked so we use this encoder to convert categorical features into dummy arrays
Housing	One-Hot Encoder	Housing loan status is not a category that can be ranked so we use this encoder to convert categorical features into dummy arrays
Loan	One-Hot Encoder	Personal loan status is not a category that can be ranked so we use this encoder to convert categorical features into dummy arrays
Contact	One-Hot Encoder	Whether a client has been contacted is not a category that can be ranked so we use this encoder to convert categorical features into dummy arrays
Month	One-Hot Encoder	Months is not a category that can be ranked so we use this encoder to convert categorical features into dummy arrays
Day of Week	One-Hot Encoder	The day of th week is a categorical feature that cannot be ranked so we use this encoder to convert the categorical features into dummy arrays.
Duration	Standard Scaler	standardizes continuous features by removing the mean and scaling to unit variance
Campaign	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
Pdays	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
Previous	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
poutcome	One-Hot Encoder	Poutcome is not a category that can be ranked so we use this encoder to converts categorical features into dummy arrays.
emp.var.rate	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
cons.price.idx	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
cons.conf.idx	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
euribor3m	Standard Scaler	This feature is not reasonably bounded and it follows a tailed distribution therefore a standard scaler is better to be used
nr.employed	Standard Scaler	This feature is not reasonably bounded and it follows a tailed

		distribution therefore a standard scaler is better to be used
y	Label encoder	This transformer should be used to encode target values

References

Kaggle Data Website - https://www.kaggle.com/henriqueyamahata/bank-marketing
Scatter Matrix - https://www.marsja.se/pandas-scatter-matrix-pair-plot/
Matplotlib - https://matplotlib.org/stable/tutorials/colors/colors.html

Github Link

https://github.com/skyyaya28/1030Marketing_Project.git