**Predicting If a Client Will Subscribe a Term Deposit**

1. **Introduction**

Portuguese banking institution ran a direct telemarketing campaign to promote its term deposit to its clients. We want to predict if a client will subscribe a term deposit or not after a call. To understand this question, we need to determine the most important factors that lead to a successful telemarketing call. The factors come from three aspects that determine ‘who’, ‘when’ and ‘how’: demographic information, economic index, and phone call behavior. The European banking institutions can use such a model to make better telemarketing strategies from 2008-2013.

First, we need to optimize the target of the telemarketing phone call. Certain customer segment that shares same characteristics, like age, occupation, financial situation and etc. may be more likely to subscribe a term deposit. Understanding customer data can help institutions to select a high-quality set of potential customers.

Second, general economic and social attributes can also impact the success of telemarketing phone call. People may feel decreased or increased need of saving when the economy is up or down. Although marketers cannot control the economic indexes, they can utilize the economic features to predict the outcome of a telemarketing campaign and set up a better marketing plan.

Third, how the telephone call is made can reach the best result. For example, the length of the call, the day of week that the call is made, and the frequency of a call to a same customer can be possible factors to determine if the call would be successful or not.

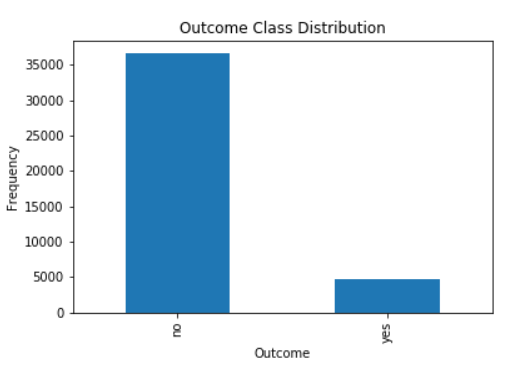
1. **Data Acquisition and Cleaning**

We acquire the Bank Marketing Data Set from UCI Machine Learning Repository. The dataset is collected from a Portuguese banking institution’s direct telemarketing campaigns implemented from May 2008 to November 2010. We downloaded bank-additional-full.csv with all examples. The dataset has some missing values named ‘unknown’ and ‘nonexistent’. We convert missing values to default ‘nan’.

The ‘pday’ column represents the number of days that passed by after the client was last contacted from a previous campaign has numeric values. The value ‘999’ means client was not previously contacted, instead of 999 days. We convert the ‘pday’ column from numeric variable to categorical variable for consistency.

1. **Data Exploration** 
   1. **Introduction to the cleaned data**

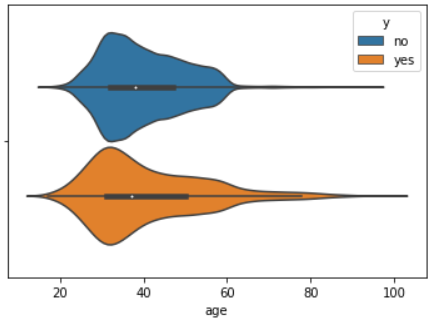
There are 41188 phone records in the dataset, ordered by date from May 2008 to November 2010. However, the date information is not included in the dataset. The dataset includes 20 columns as variables, and 1 column as outcome. We identify 10 numerical variables, and 10 categorical variables. And the outcome column contains categorical data ‘yes’ and ‘no’. There are more than 35,000 call failed, and only around 5000 calls are successful. The imbalanced dataset needs to be fixed when applying models to improve effectiveness.



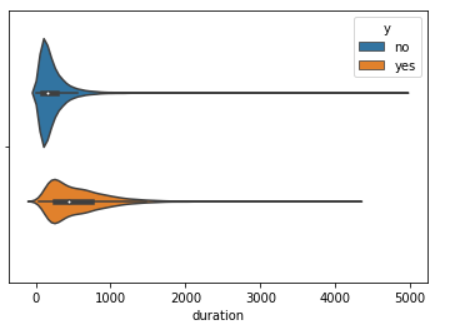
* 1. **Introduction to the numerical columns**

We will go through the numerical fields in the data set to explore their relationship with telemarketing call’s outcome.

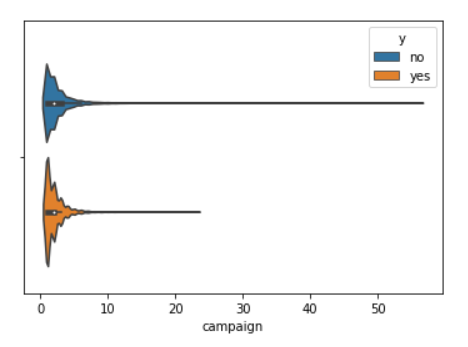
First, customer’s age. Most clients recorded in the data are between 20 and 60 years old. The largest group of people is around 30 years old.50% of clients are from 30-50 years old. People older than 60 years old are more likely to subscribe the term after phone call.



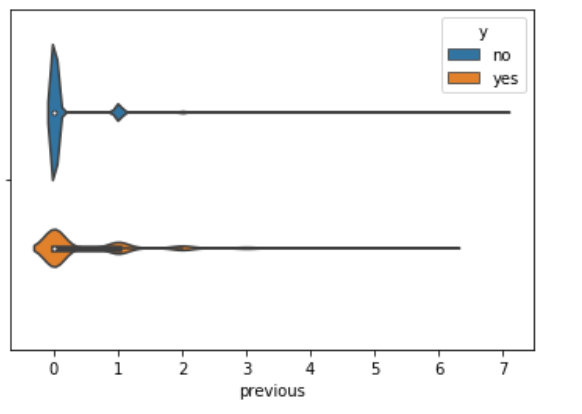
Second, last contact duration in seconds. We examine the relation between duration and outcome. Most phone calls are made shorter than 1700 seconds. Phone call that last longer are more likely to have a positive result. However, this attribute will be dropped when applying predictive models, because if duration is known, the result is known.



Third, campaign. Campaigns represents number of contacts performed during this campaign and for this client. We don’t see any big difference from the plot.

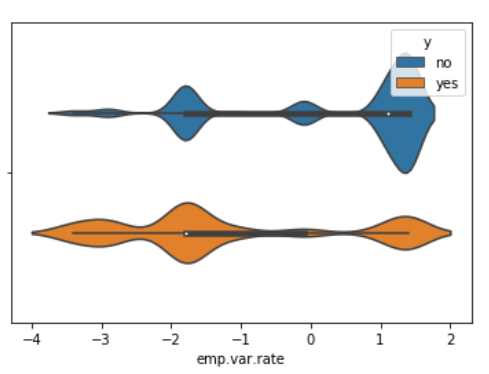


Fourth, previous. It represents number of contacts performed before this campaign and for this client. From the plot, we can see that 50% of subscribed clients have been contacted 0-1 time before this campaign. For those who have been contacted twice, they are more likely to subscribe.

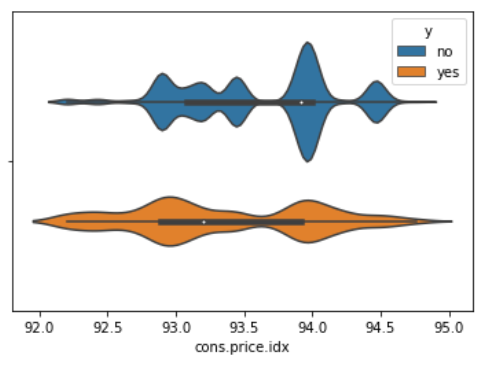


The following numerical fields will all be economic attributes. We will go through each one of them.

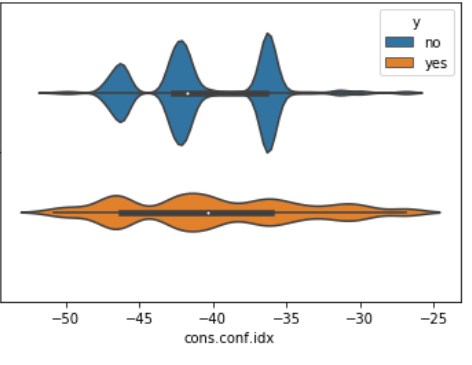
When the employment variation rate (with quarterly frequency) (emp.var.rate) is below -1, it is more likely to have a successful phone call. The median of employment variation rate for clients who subscribe the term after phone call is around -2, while the median for clients who don’t subscribe the term around 1. 50% of the employment variation rate range between -2 and 0 among successful phone calls, compared to unsuccessful phone calls with 50% of employment variation rate data from -2 to 1.5.



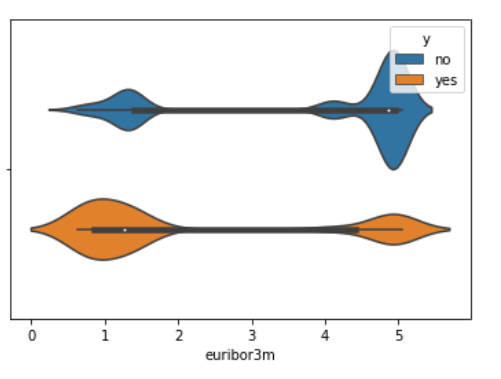
When the consumer price index (monthly) (cons.price.idx) is less than below 93, clients have a propensity for subscribing the term. The median of consumer price index among successful calls is about 93.25, compared to 93.85 among failed calls, with total range from 92 to 95 for the entire dataset.



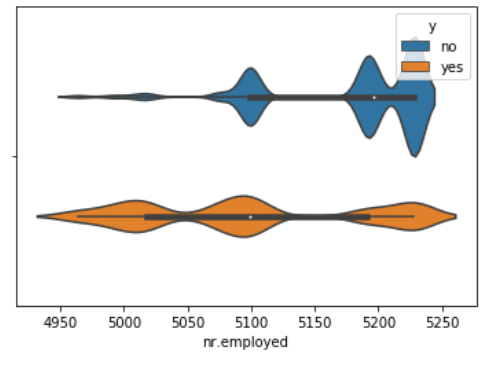
The total range of consumer confidence index (cons.conf.idx) in the dataset ranges from -54 to -26.There is wider variations of consumer confidence index (monthly) among positive outcome. 50% of positive outcome has consumer confidence index ranging from -46 to -36, with medium of -41. 50% of negative outcome has consumer confidence index ranging from -43 to –37. Telemarketing calls are more likely to fail at three points, when the consumer confidence index is –47, -43 and –36.



The median of Euribor three month rate (euribor3m) for negative outcome is 5, while the median rate for positive outcome is only 1.3. When the Euribor three month rate is low, clients have much more propensity for subscribing the term.



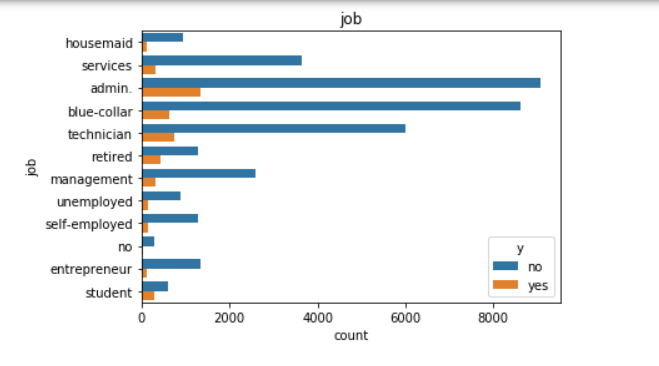
The total range of number of employees (quarterly) (nr.employed) in the dataset is from 4925 to 5275. 50% of negative outcomes have number of employees form 5100 to 5230, with median of 5200. The range of the same attribute for 50% of positive outcomes lies between 5025 and 5270, with median of 5100. Both negative outcomes and positive outcomes reach three peaks. They reach a peak at a same point when number of employees is 5230. And they reach another peak at a close point when number of employees are 5090 for positive outcome, and 5100 for negative outcome. The difference is that the negative outcomes reach a peak when the number of employees is 5180, and the positive outcomes reach a peak when the number of employees is 5010.



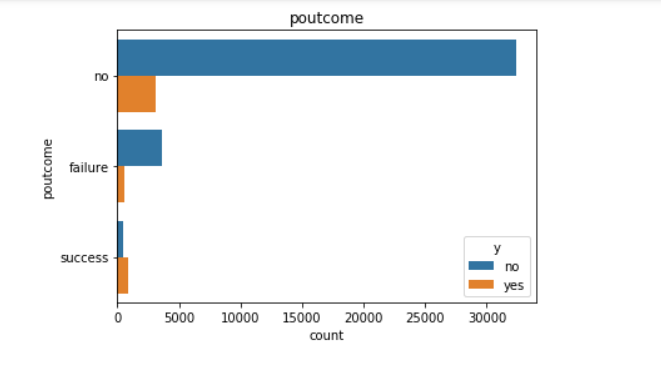
* 1. **Introduction to the categorical columns**

Next, we will go through all categorical columns in the dataset, and explore their relations to the outcome.

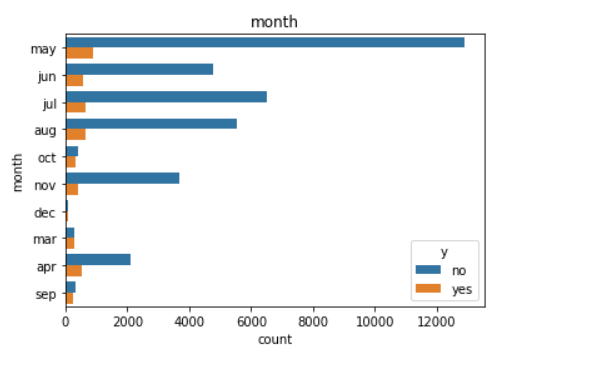
First, job. most clients have jobs as administrator, blue-collar, technician and service person. Among people who subscribed the term, the largest number is from administrative job. The least number is from entrepreneur and housemaid. Students are more likely to subscribe than other occupations. Although the groups of clients being students and are retired are small, the percentage of subscriber in the two groups are comparably higher than other occupations.



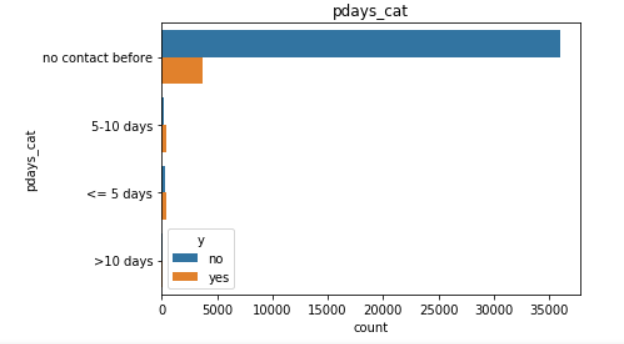
Second, previous outcome. Very few clients have a positive previous outcome. However, those who have a successful previous outcome are very possible to subscribe again.



Third, month. Calling on October, March, April and September are more likely to achieve positive outcomes than calling on other months.



Fourth, number of days that passed by after the client was last contacted from a previous campaign. 5-10 days and less than 5 days generate better result than no contact before.

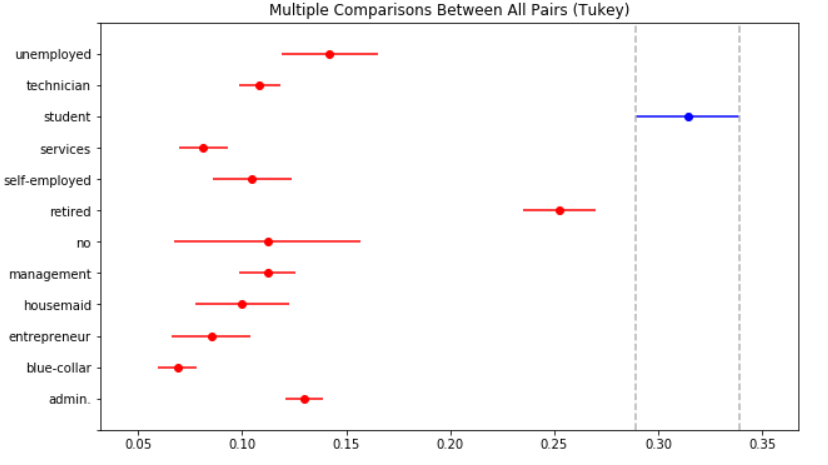


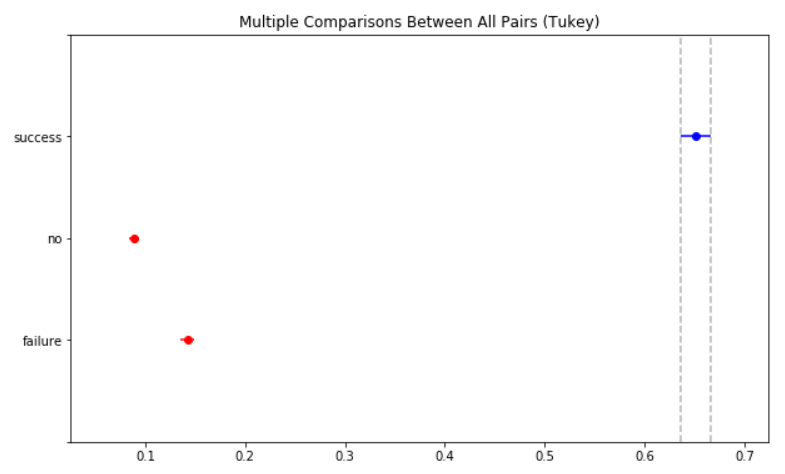
Fifth, housing, marital, contact method, education, day\_of\_week, default, having a loan. We cannot tell from the graph that these variables would have any impact on the outcome.

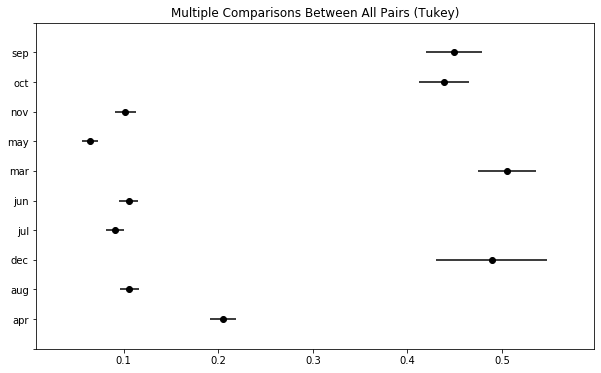
1. **ANOVA Analysis**

We initiate a null hypothesis for each numerical variable that there is no statistically significant difference in the mean between the groups of positive outcome and negative outcome of each variable. The F test generates high F score and low p value. We reject all null hypothesis for all numerical variables.

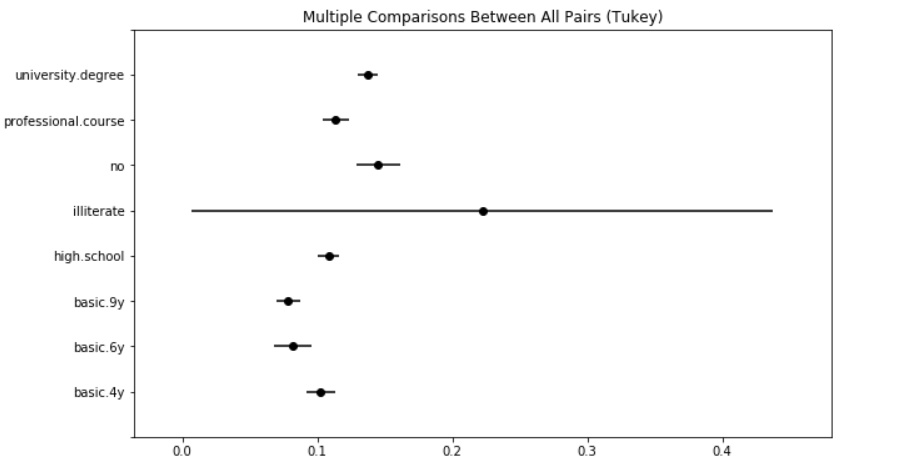
For categorical variables, we select a few variables that may have impact on the outcome based on the insights we get from the exploratory analysis. Same as the observations from exploratory analysis, the group of student and retired people are more likely to subscribe than other groups. Clients are more likely to subscribe if previous outcome is successful. The highest subscribe rates occur in March, August, September, October.



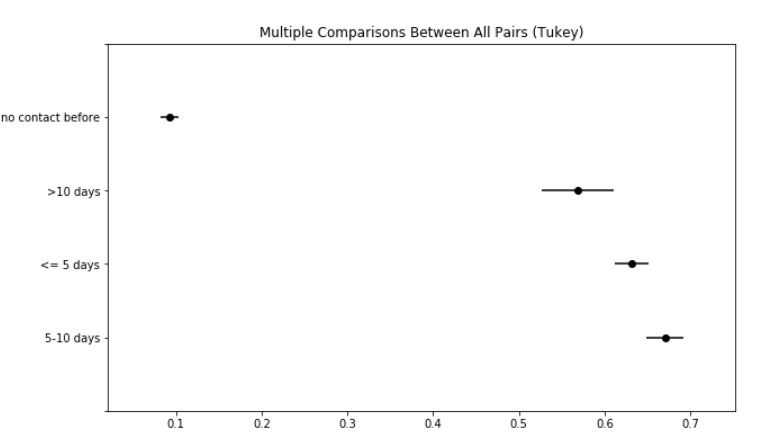




Although it is hard to tell any difference between group of varied education, the ANOVA test shows People who get advanced education (high school, professional course, and university degree) are more likely to subscribe. From the below simultaneous plot, we see that although the mean of illiterate group is high, it doesn't mean that the illiterate people are more likely to subscribe. The reason is that the confidence interval for illiterate group is too wide. We can also ignore 'no' group, that contains unknown values.

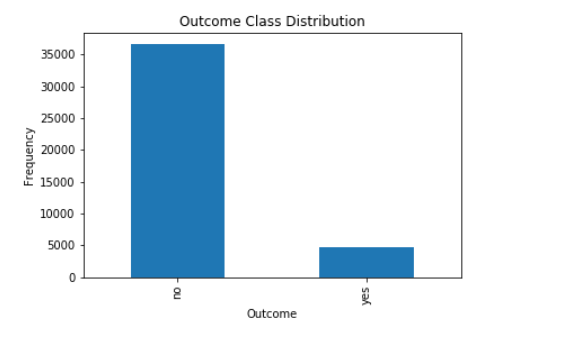


The simultaneous plot also shows more clearly than the graph in exploratory analysis for the previous contact variable. The group of 5-10 days that passed by after the client was last contacted from a previous campaign have the highest mean. The group of no contact before has the lowest mean. And clients who have previous contact in less than 10 days are more likely to subscribe than clients who have no contact before.



1. **Modeling**

Knowing the outcomes of phone calls are labeled, which are yes (1), and no(0), we use supervised machine learning algorithms. Furthermore, there are only two outcomes of the phone call, yes or no. We use binary classification algorithms to predict the outcome. We build three models: random forest model, logistic regressions model, and gradient boosting classifier. The models are trained with 80% of data, and the remaining 20% data is used as test data. In exploratory data analysis, we plot the outcome. 35,000 out of 41,188 phone calls fails, and only less than 5000 phone calls have outcomes of ‘yes’. That means we have very imbalanced data. Too few examples of successful calls will cause the model to learn ineffectively, because learning is biased towards the majority classes. To address this issue, we apply SMOTE to resample training data. Next, we will go through three predictive models we use to predict the phone call outcome.

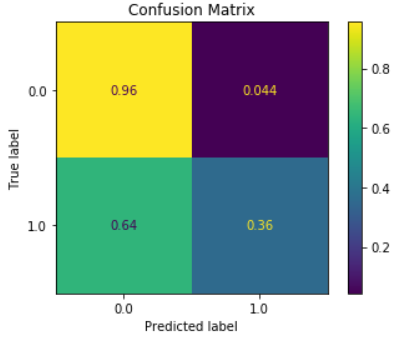
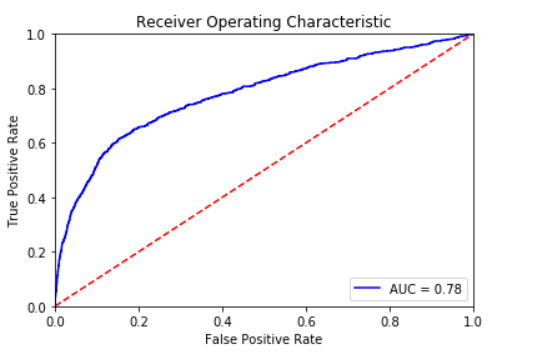


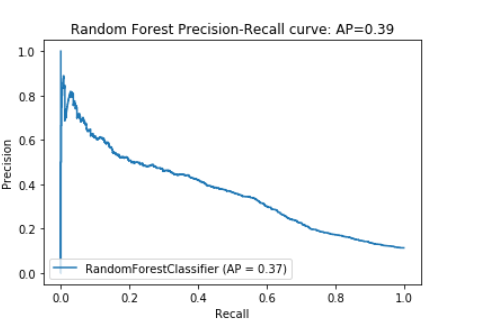
* 1. **Data Pre-Processing**

1. One-hot encoding: We split the dataset into numerical variable group and categorical group. We apply one-hot encoding to categorical variables. We use the label encoded data to build models. The column names are changed after one-hot encoding. The outcome column y is split into two columns: y\_no and y\_yes.
2. Data splitting: We drop the y\_no column and make y\_yes column only target variable. We split data into train and test dataset by 80% to 20% ratio.
3. Scaling: We scale the values of all features by standardization, mean removal and variance scaling, so the mean of all features is 0, and variance would be 1.
4. Resampling: We use SMOTE to resample unbalanced training data, so the resampled data can be trained more effectively in the machine learning models.
   1. **Random Forest Model**

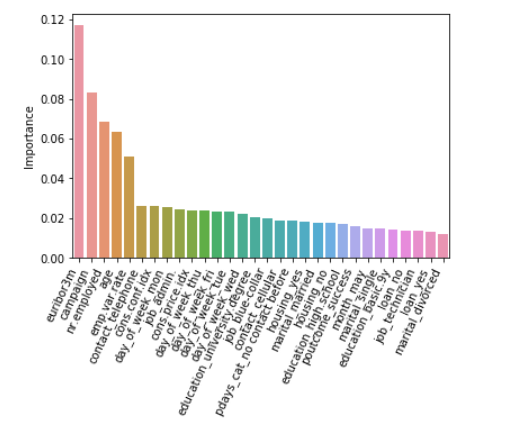
We first build a base random forest model using pro-processed data to train and fit with default parameters. We get an accuracy score of 0.883. We then apply random search for hyper-parameter tuning with 5 fold cross validation. We train the data with the best parameter grid obtained from random search. The accuracy score improves by 0.43%.

We then apply grid search with 5 fold cross validation. The accuracy score improves to 0.888, which performs better than random search. We plot the confusion matrix and ROC curve as below, as well as the precision-recall curve.

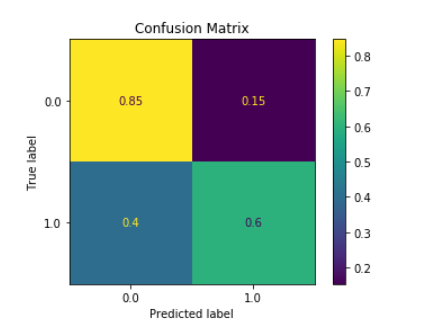
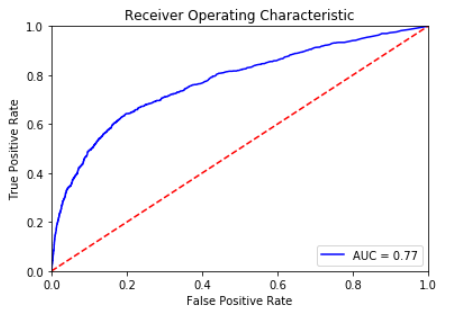


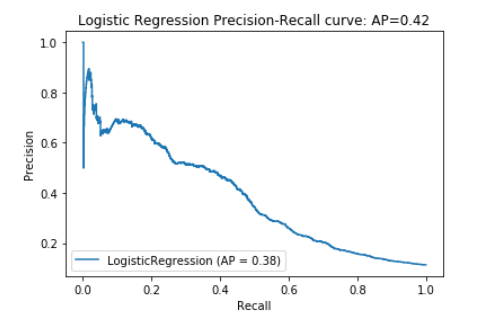
We use random forest model with optimized parameters from grid search to check the top 30 feature importance, which is the by-product of random forest model. From below feature importance table and chart, we can see that Europe's Libor rate ('euribor3m'), number of contacts performed during this campaign and for a client ('campaign'),number of employees ('nr.employed'), client's age ('age'), employment variation rate ('emp.var.rate'), contact by telephone ('contact telephone'),consumer confidential index (cons.conf.idx),contact on Monday (day\_of\_week\_mon), job as admin ('job\_admin'), consumer price index ('cons\_price.idx'), call on Thursday, Firday, Tuesday and Wednsday ('day\_of\_week\_thu', 'day\_of\_week\_fri', 'day\_of\_week\_tue', 'day\_of\_week\_wed'), client who has university degree (education\_university\_degree),job as a blue\_collar ('job\_blue-collar'), no contact before ('pdays\_cat\_no contact before'), client who doesn't own a house (housing\_no), client who owns a house (housing\_yes) are the top important features for outcome prediction.



* 1. **Logistic Regression**

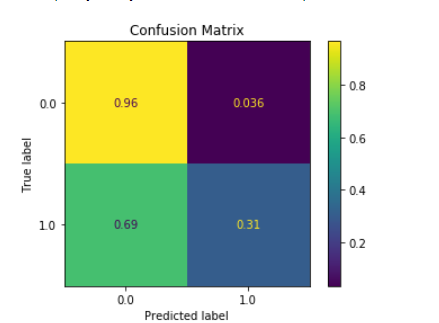
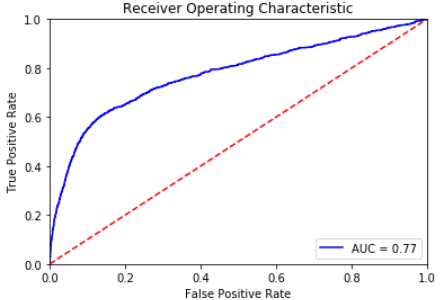
We also apply logistic regression model for prediction. We build a base logistic regression model first. We then use both random search and grid search with 5 fold cross validation for hyperparameter optimization. We get the best test accuracy score of 0.81. Below is the confusion matrix, roc curve and precision-recall curve we generated form Logistic Regression model.

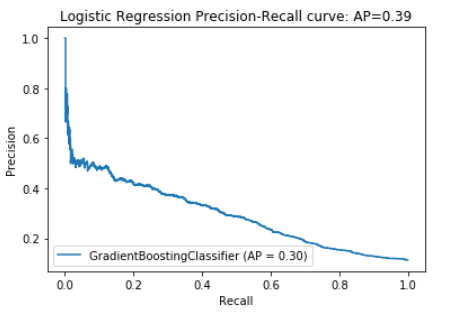
 



* 1. **Gradient Boosting Classifier**

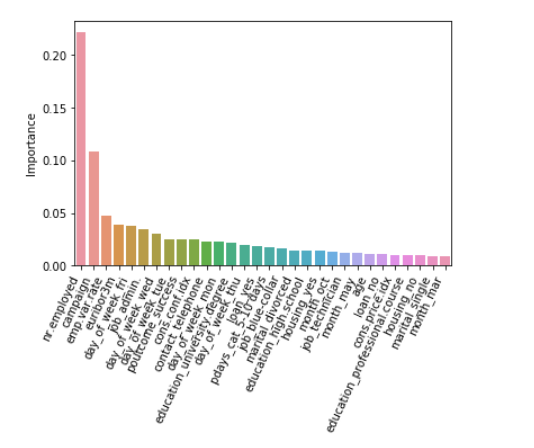
Our third model is gradient boosting classifier. Like the other two models, we built a base model with accuracy score of 0.88. After hyperparameter tuning using random search and grid search with 5 folds of cross validation, we improve the test accuracy score to 0.89 by 1.19%. The grid search performs better then random search and improves the accuracy score to 0.89 by 1.28%. Below are the confusion matrix, ROC curve and precision-recall curve generated from Gradient Boosting Classifier model.



Gradient Boosting model also has feature importance as a by-product.

number of employees ('nr.employed'), number of contacts performed during this campaign and for a client ('campaign'), employment variation rate ('emp.var.rate'), Europe's Libor rate ('euribor3m'), are the top 4 most important features. Compared to the feature importance plot we get from Random Forest, client’s age is not considered as a very top important feature in gradient boosting classifies model.



By comparing three models, we got highest accuracy score from Gradient Boosting, highest AUC score from Random Forest, and highest AP score from Logistic Regression.

1. **Recommendation**

From the result of the Anova analysis, and feature importance plots, we can conclude that the economic indexes, including number of employees ('nr.employed'), employment variation rate ('emp.var.rate'), and Europe's Libor rate ('euribor3m') are very important features that can impact on the marketing call result. Also, number of contacts performed during this campaign and for a client (‘campaign’) can also have significant impact on the result. In addition, the group of clients’ job as ‘admin’, can also influence final the result. What’s more, contact made on which day of a week and client’s education degree are the factors that cannot be ignored. The marketing department in the bank should take the above important features into account, so they can make better decision on when is the best time to carry out the campaign, which day to call, and who should be called.

Generally, when the economic is down, and the employment rate is low, people are more likely to subscribe the term. The bank should invest more on this direct phone call campaign during this time. When considering segmenting clients, clients with higher education degree are more likely to subscribe. As for the calling strategy, making multiple phone calls to a same client can help to increase the possibility of subscription.