#### Main Analysis

**Model assumptions**

1. binary logistic regression requires the response variable to be binary.
2. logistic regression requires the observations to be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
3. logistic regression requires there to be little or no multicollinearity among the independent variables.  This means that the independent variables should not be too highly correlated with each other.
4. logistic regression assumes linearity of independent variables and log odds.  although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

5. logistic regression typically requires a large sample size.

**Model selecting**

According to the assumptions of Logistic Regression, logistic regression requires there to be little or no multicollinearity among the independent variables. If we use all variables in the model, the multicolinearity will exist among those numerous variables. So, at first, we need to reduce the variables and select the key variables in those original variables. As the same as in simple linear regression, we can use AIC criterion for model comparison or in a stepwise model selection routine. However, the result of logistic regression has association with the unbalanced labels in data set and the bank marketing data set is unbalanced data. If we use unbalanced data set to select model, it may cause some disadvantages influence for our result. So, for using AIC criterion to select model, we will use the overall data which label is yes and the same number of the data which label is no and selected randomly. Those data construct a balanced data set for model selecting. There are 9 variables in final model. Its are age, marital, education, housing, duration, emp.var.rate, cons.price.idx, cons.conf.idx and euribor3m.

**Data splitting**

At this step, we need to split the overall data in to training set and testing set. We will use the training set for training the model and using the testing set for evaluating the model. Our strategy is that random choosing 80% of data which label is yes and choosing 80% of data which label is no. Those two data set construct the training set and the rest of data are included into testing set. There are 3713 data which label is yes and 29239 data which label is no for training. There are 927 data which label is yes for testing and 7309 data which label is no for testing.

**Model fitting**

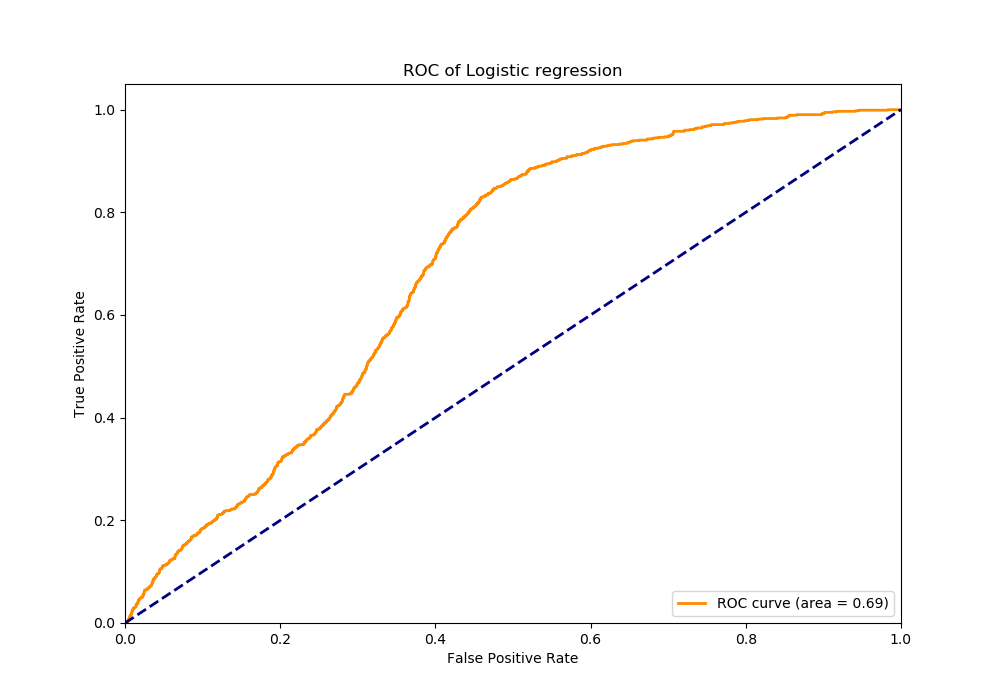
Because the training data set is an unbalanced data set. We need to fix this problem. Our strategy is that using class weight to balance this effect. The weight for data which label is no is 3713/29239 = 0.127 and the weight for data which label is yes is 1.. Because we do not want a overfitting modeling, we need to add L2 penalties for the coefficients of logistic regression model. L2 penalty is an method of Regularization which can prevent model overfitting by penalizing high-valued regression coefficients. In simple terms, it reduces parameters and [shrinks](https://www.statisticshowto.datasciencecentral.com/shrinkage-estimator/) (simplifies) the model. We also do Z-Score transformation on each numeric variables because we need to reduce the influence of outliers if this variable contains outliers. Z-Score transformation is an transformation which can center our data points and reduce the distance between two points.

**Model evaluation**

There are many evaluation index for evaluating the performance of one model like accuracy, F1 score Area Under ROC Curve(AUC) and confusion matrix. For an unbalanced testing data set, only looking the accuracy of the predict is not correct because the data which have more same label will effect the accuracy more powerful. So, we will select the F1 score, AUC and confusion matrix for evaluating the results.

The F1 score of this regression model is 0.257.

The ROC curve of this regression model is :



The AUC is 0.69.

The confusion matrix is :

|  |  |  |  |
| --- | --- | --- | --- |
|  | No (Predict) | Yes (predict) | Total |
| No (True) | 2293 | 5016 | 7309 |
| Yes (True) | 51 | 876 | 927 |
| Total | 2344 | 5892 | 8236 |

**Model diagnostic**

1. Obviously, the response variable of this regression model only contains double factors, yes and no. It is satisfy the first assumption.
2. The is the bank marketing data set. The observations(which indicate to one of custom) are not generated from from repeated measurements or matched data obviously.
3. For detecting the multicolinearity, we should calculate the VIF of each variables which are numeric. There are 6 numeric variables in this model. Its are age, duration,emp.var.rate, cons.price.idx, cons.conf.idx and euribor3m. The VIF of each variables are 1.0085, 1.0114, 2.6824, 2.9288, 1.2967 and 2.4287. Those VIF are all under 10. So, there is a less multicolinearity in those variables. It can roughly say that those variables are independent.
4. For detecting linearity of independent variables and log odds, we will use Pearson correlation to measure the linearity of log odds and those independent variables. Because Pearson correlation can measure the linear correlation between two variables. The Pearson correlation between log odds and those variables are -0.0036, 0.7884, -0.6632, -0.5324, -0.0684 and -0.6448. From those number we can know that the variables age and cons.conf.idx are not linear with the log odds but the other variables are linear with log odds. This may indicates that the variable age and cons.conf.idx should remove from our model.
5. Obviously, we have large training sample size.
6. At last, we want to know how well the model fits the data. We will use Hosmer-Lemeshow goodness-of-fit test to compare the observed and expected frequencies of events and non-events. The null hypothetical of this test is the model fits the data well. The alternate hypothetical is that the model fits the data not well. The q-value is 0. This means we need reject the null hypothetical. This indicates that logistical regression is not suitable for this data set and if possible, we can change another model to fit this data set.