**Bank Term Deposit Subscription Prediction**

**Based on Classification**

Zimeng Zhang

School of Information Studies

Syracuse University

Syracuse, NY, 13210

zzhang@syr.edu

**Abstract**

Using telemarketing dataset generated from a Portuguese banking institution, I used multiple machine learning methods (Logistic Regression, SVM, and Random Forest) to analyze consumers demographic influence on telemarking calls for selling bank long-term deposits. In this project, I performed and compared three machine learning models: Logistic Regression, Support Vector Machine and Random Forest . Model evaluation process will be conducted by measuring the Area Under the Curve (AUC), to testify the ability of a classifier to capture the degree of variance between classes. The entire project has demonstrated the validation of using machine learning models to anticipate and predict consumers’ loyalty for telemarketing campaign.

1. **Introduction**

Marketing is a crucial element for a commercial enterprise to define its targeting customers and enhance business. Telemarketing is one of which methods that can be used to directly figure out whether customers have the potential to be targeted as loyal consumers by contacting over telephone. Customers who show larger interest over telephone are more likely to subscribe a product from the bank. From the perspective of the Portuguese banking institution, more customers like that can help improve company revenue over years.

This telemarketing dataset includes consumer demographic records including age, marital and education status, etc. It is also having the records of if customers have subscribed to a banking term deposit. Details of the dataset will be included in future section. From the banking aspect, I want to use these telemarketing records to predict and evaluate whether a consumer would subscribe to a deposit. And through analyzing results, I plan to help the banking to find the best target segments of customers.

1. **Methodology**
   1. **Logistic Regression**

Logistic Regression is a predictive regression analysis that is used when the dependent variable is categorical, that is considered a linear classifier such that the classification decision is made based on a linear combination of predictors.

* 1. **Support Vector Machine**

Support vector machine (SVM) is a method used for data classification and regression analysis. Here, we use SVM to learn from the training data, and then build a model to predict whether the record would be a subscription based on client information. SVM is a hyperplane classifier, which works by determining which side of the hyperplane is on. SVM maximizes the margin around the separating hyperplane. The decision-making function is completely specified by a subset of the training samples. This subset of the vector is called the support vector. SVM training algorithm is a binary linear classifier with no probability. This method divides the data into categories separated by hyper lanes, and the marginal width of the hyper lanes becomes as wide as possible. The new data will be allocated to the classified space according to which side of the space it is located.

* 1. **Random Forest**

Random forest is a typical ensemble learning method.

When single decision is vulnerable to danger of high variance, we do a tree bagging to decrease variance that simulate training datasets by aggregating multiple bootstrap sample. Bootstrap sampling is the process of sampling with replacement. Every time an observation is drawn from a population, the observation is returned to the population. The process of sampling with replacement means that a bootstrap sample can have multiple instances of the same observation from the population. Since a sample can have multiple copies of an observation, sample variance is less than the population variance. And tree bagging method would fit a unique decision tree to each bootstrapped data sample and average the prediction of the models, therefore, based on numbers of trees n, variance decreases n times.

When utilizing the same training data over and over again, trees become correlated. Here comes the idea Random Forest that tries to remove correlation by randomly sampling training data columns at each split point. Therefore, each tree uses different feature subsets for the prediction. A typical number of columns to sample m is where p is the number of columns in the training data.

Random Forest has the following steps:

* Start with a training data set.
* Take n bootstrap samples where n is equal to the desired number of trees in the forest.
* For each bootstrap sample, grow a decision tree by sampling n columns of the bootstrap sample at each decision point.
* Prediction is determined by the majority vote.
  1. **Principal Components Analysis**

Principal components Analysis is an unsupervised learning method. PCA transforms the dataset into a lower dimension via linear transformation to implement dimension reduction and produce a lower dimensionality representation of the original data. PCA quantifies the relationship between attributes by finding a list of the most important axes in the data and using these axes to describe the data set. The principal axes are linear transformations of the original axes. The "principal axis" aims at decreasing the amount of variance in the data set with capturing majority variance of data, i.e., the first principal component picks up the most variance, PC2 is orthogonal to PC1 and picks up less variance. After transformation we can utilize certain number of principal components that capture enough variance as substitution of original data, so that dimension reduction is completed. Loading vector shows the relationship between principal components and original observations. Loading vector is the transformation vector for each recording. It is used to gain insight about the lower dimensionality representation, and coefficient in the vectors demonstrate the feature importance from original record.

* 1. **Data Set**

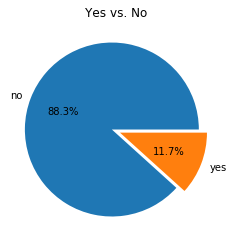
The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

This marketing dataset includes in total 45,211 records along with 18 attributes. Among these 18 attributes, there is metadata such as job, whether consumer credit is default (no/yes), and the duration of last contact etc. That being said, this dataset includes both categorical and numeric data type. The target variable would be a column named ‘y’, in which ‘no’ stands for not subscribe and ‘yes’ indicates subscribed.

* 1. **Experimental Procedures**
     1. **Exploratory Data Analysis**

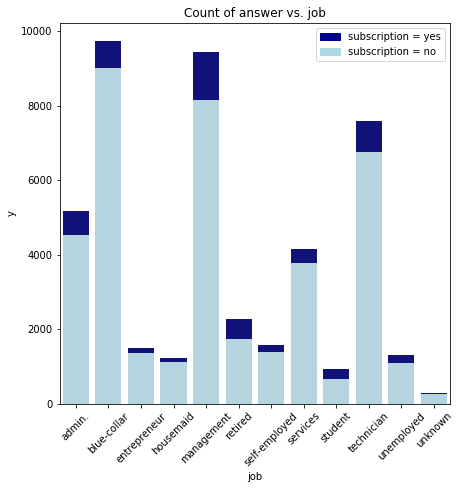
Our goal is to predict whether a customer would subscribe the product. First, we look at the subscription condition, ‘yes’ demonstrate that a customer successfully subscribe the product and ‘no’ demonstrates a customer would not buy the product. The pie chart shows that the data set is extremely imbalanced, having positive samples far more than negative samples, which means data resample is needed. And the campaign does not have a good result according to the subscription rate.

**Figure 1** TotalSubscription condition

****

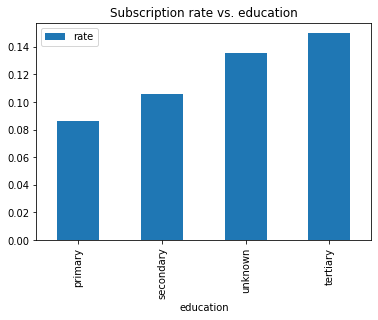
Next, to furtherly analyze subscription conditions, we make a stacked bar plot to show counts of subscription grouped by each customer job. There is a clear comparison between subscribed and not subscribed for each job. According to the chart, people working in management are retired or students are relatively inclined to subscribe to bank term deposits. Blue-collar, management and technician are the main audience industry of telemarketing.

**Figure 2** Subscription counts by jobs.

****

Education level is also a crucial factor for marketing campaigns. Through this chart, we can tell that people who have higher education are more likely to accept telephone marketing promotions.

**Figure 3** Subscription rate vs. Education

****

* + 1. **Data Preprocessing**
       1. **Data Resample**

The dataset is unbalanced with 90% negative samples and 10% positive samples. Coming to modelling procedure, many machine learning models suffer from frequency bias, and they place more emphasis on learning from more frequently occurring data observations. However, when class imbalance is introduced, the learning challenges of datasets that are inherently more difficult to learn are magnified. Therefore, we construct a balanced dataset by oversampling from the minority.

* + - 1. **Data Transformation**

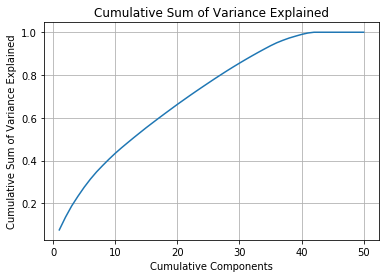
Notice that the original data set has a mixture of categorical and numerical data which could not possibly be processed. In order to be able to apply the classification model, here we transform all the category data into numeric type data by one-hot encoder. The transformed dataset has a total of 50 features. To speed up the learning process, a normalization process is also performed.

* + 1. **Dimension Reduction**

Due to the large sample size and the number of features, the linear kernel SVM is far overloaded, resulting in kernel that cannot converge even with a high number of iterations. So, we need to downsize the number of features. To reduce the sample dimensionality, we applied the PCA algorithm to transform the original samples and selected the principal component with a cumulative contribution of 85% as the new input to the SVM.

According to the figure below, we select the first 30 principal components so that the SVM linear kernel completes convergence with a relatively small number of iterations.

**Figure 4** Cumulative Sum of Variance Explained



* + 1. **Hyperparameter Tuning**

For random forest model, there are several important hyperparameters that are able to be tuned and have a great influence on model performance.

Parameter ‘criterion’ Decision tree classifier predict the majority of a region and evaluate on distribution of predictions. Gini index and entropy are two criteria of it, respectively represent a measure of node purity and randomness or information. Parameter ‘max\_features’: Each step random forest classifier randomly sample columns from original dataset, max\_features depict the numbers of columns it chooses. After binning, there are 50 features in data set, so the best choice is around 7 each step. Parameter ‘max\_depth’: Maximum depth of each tree in the forest. Parameter ‘n\_estimators’: The number of trees in the forest.

To achieve the model with best performance, we conduct grid search via sklearn package, set up parameter grids and evaluate the model performance by cross validation. The best parameters setting is shown in the table.

**Table 1** Radom Forest Parameters Setting

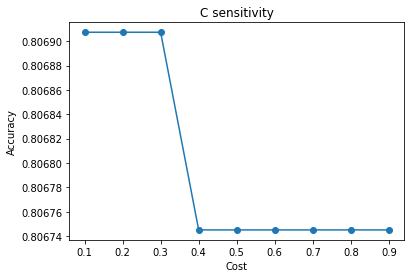
|  |  |
| --- | --- |
| Hyperparameter | Value |
| criterion | entropy |
| max\_features | 7 |
| max\_depth’ | 50 |
| n\_estimators | 100 |

For SVM model, the most important parameter to be tuned is the cost, C acts on the loss term and is the opposite of the regularization parameter.

The larger the C, the greater the effect of increasing the number of eligible relaxation variables and the model tends to be complex. the smaller the C, the smaller the effect of increasing the number of variables that violate the constraints and the model tends to be simple. Thus.

When low deviation and high variance, overfitting is encountered, the model complexity should be reduced, and C value should be decreased. If the deviation is high and the variance is low, the model complexity should be increased, and the C value should be increased when an underfitting situation is encountered. In this project, we arrange different value to C and draw a line plot depicting model accuracy vs. cost and find that the best cost = 0.1.

**Figure 5** C sensitivity



1. Results
   1. **Classification Report**

For classification evaluation, we use accuracy and AUC score for accuracy of model and ability of generalization. The result is as follows, it can be shown that because of the data loss through dimension reduction, SVM has relatively bad performance compared to the other two models. Logistic regression might have a tight decision boundary; thus, the AUC score is lower than expected while the accuracy is good. Random forest has the best performance with predicting 92 % data correctly.

**Table 2** Classification report

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | LR | SVM | RF |
| AUC | 0.51 | 0.88 | 0.98 |
| Accuracy | 0.83 | 0.81 | 0.92 |

* + 1. **Logistic Regression**

Logistic Regression model is trained on our training data set and after which we test the model with test data set and compiled a classification report to observe the classification results. For these binary classification problems, the classification ability of the logistic regression classifier for the two categories is almost the same. The average accuracy of the classifier is 83%, which is 33% higher than random guess.

**Table 5** Logistic Regression classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Logistic Regression | precision | recall | f1-score | support |
| no | 0.83 | 0.84 | 0.83 | 3023 |
| yes | 0.84 | 0.83 | 0.84 | 3145 |
| accuracy |  |  | 0.83 | 6168 |
| macro avg | 0.83 | 0.83 | 0.83 | 6168 |
| weighted avg | 0.83 | 0.83 | 0.83 | 6168 |

* + 1. **Support Vector Machine**

After tuning the hyperparameter of the SVM model, we set c=0.1 for the best model performance accuracy. We compiled a classification report to observe the classification results. For these binary classification problems, the classification ability of the SVM classifier for the two categories is almost the same. The average accuracy of the classifier is 81%, which is 31% higher than random guess.

**Table 6** SVM classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SVM | precision | recall | f1-score | support |
| no | 0.80 | 0.81 | 0.80 | 3023 |
| yes | 0.82 | 0.80 | 0.81 | 3145 |
| accuracy |  |  | 0.81 | 6168 |
| macro avg | 0.81 | 0.81 | 0.81 | 6168 |
| weighted avg | 0.81 | 0.81 | 0.81 | 6168 |

* + 1. **Random Forest**

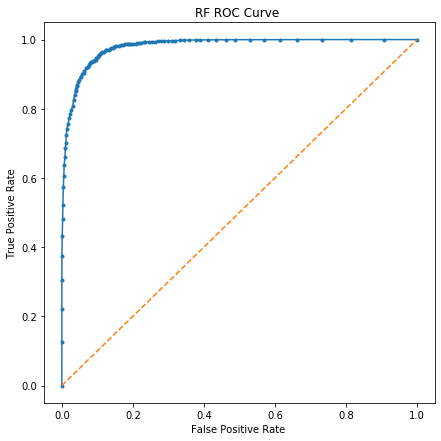
Random forest algorithm performed the best solving this problem, at the same time, his running time is slightly longer than the other two models. Rf model has better ability to predict among the audience of telemarketing, which kind of people are more likely to reject subscription rather than conducting acceptance. The precision of predicting ‘no’ class is 95%, 6% higher than that of predicting ‘yes’. Total performance is pretty good with 92% accuracy after tuning hyperparameters.

**Table 7** RF classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Random Forest | precision | recall | f1-score | support |
| no | 0.95 | 0.88 | 0.92 | 3023 |
| yes | 0.89 | 0.96 | 0.93 | 3145 |
| accuracy |  |  | 0.92 | 6168 |
| macro avg | 0.92 | 0.92 | 0.92 | 6168 |
| weighted avg | 0.92 | 0.92 | 0.92 | 6168 |

According to the former result, random forest has an auc score of 0.98. It can tell from the curve that, If we focus on which users will subscribe to bank term deposits, we can choose the upper left point of the curve and use its threshold, the prediction in this case will have a higher true positive rate and a lower false positive rate, and the model has a high generalization ability.

**Figure 6** ROC curve of random forest

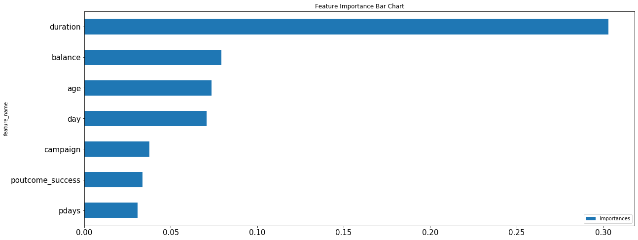


* 1. **RF Model Interpretation**

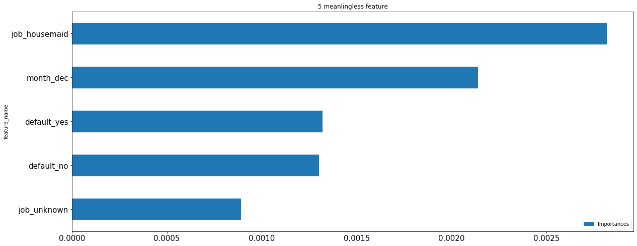
In order to better explain the model, we extracted the most meaningful features from the Random Forest model. According to the models, these features have great decision power influencing the prediction outcome.

The bar chart could tell that feature duration, balance and age are most correlated to a customer’s subscription, especially the call duration. Features like a people’s job is housemaid or the call is on Dec means little to the subscription condition.

**Fig 7** Meaningful Features



**Fig 8** Meaningless Features



1. **Discussion**

Logistic regression has a good accuracy score while a bad auc score, which means logistic regression has good performance in precision, but its generalization ability is not good, probably because it calculates the probability of samples belonging to different categories by sigmoid function, which has a stricter classification boundary with a shape of S-curve.

With 50 more features in this data set, linear kernel of SVM is hard to converge. To solve that problem, I took use of PCA to reduce the input dimension. At the same time, transformed data set can not capture all the information as original data set did. I am thinking about is it the caution of relatively lower classification power of SVM.

According to results of Random Forest algorithm, duration is the most impactive feature to subscription. However, the casual results could not be ignored that if customer were preparing to subscribe bank term deposit, they certainly would like to learn more about the product, so maybe balance and age are more insightful when choosing your audience.

1. **Conclusion**

Telemarketing does not have a high success rate, the ratio of success to failure is about 9:1. People with higher education are more inclined to accept bank deposit term. People in management are more inclined to accept bank deposit term.

Random Forest algorithm has the best predictive power with accuracy of 92%. The most informative features are call duration, account balance and customer’s age.

**Reference**

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS