Bank Marketing Analysis: Term Deposit Cold Call Prediction

Zheyuan Hu, Hanming Li, Jiaqi Song
Georgetown University, Data Science and Analytics Program

GEORGETOWN UNIVERSITY

Introduction

Cold call has a long history in bank marketing because of its effectiveness. Bankers and salesman have been told their entire life that cold calling is how a business gains traction and develops a sizable customer base. But a constant limitation of the cold call is it is a rather low successful rate. Only about 28% cold calls result in a conversion and even 2% of cold calls are successful. That can be improved by data analysis.

The problems we want to solve:
-1. Can we predict if a client will subscribe to a term deposit?
-2. What strategies can we use to get a client subscription?
-3. How much of time should a cold call last?
-4. Can we build a scorecard for bank client based on their information?

Dataset

The data contains marketing information collected by direct phone calls to evaluate whether clients subscribe to a bank term deposit for a Portuguese banking institution.

Variable Introduction

Demographics:
- age: Customer's age (numeric)
- job: Type of job
- marital: Marital status
- education: Level of education

Past customer events:
- housing: Has housing loan?
- loan: Has personal loan?

Past direct marketing contacts:
- contact: Contact communication type
- month: Last contact month of year
- dayofweek: Last contact day of the week
- duration: Last contact duration, in seconds (numeric).

Campaign information:
- campaign: Number of contacts performed during this campaign and for this client
- pdays: Number of days that passed by after the client was last contacted from a previous campaign (numeric)
- previous: Number of contacts performed before this campaign and for this client (numeric)
- poutcome: Outcome of the previous marketing campaign

Socioeconomic factors:
- 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: Employment rate

Target variable: y - has the client subscribed a term deposit? (binary: "yes","no")

Data Visualization

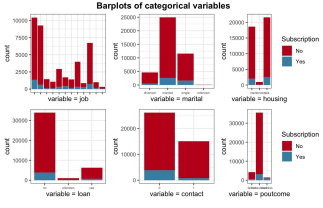


Figure 1: Distribution of categorical variables with subscription

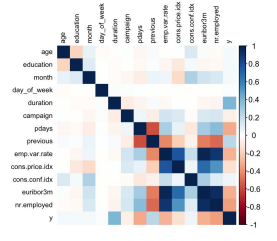
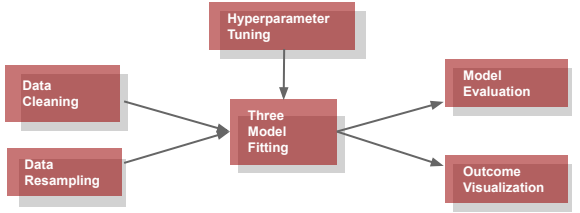
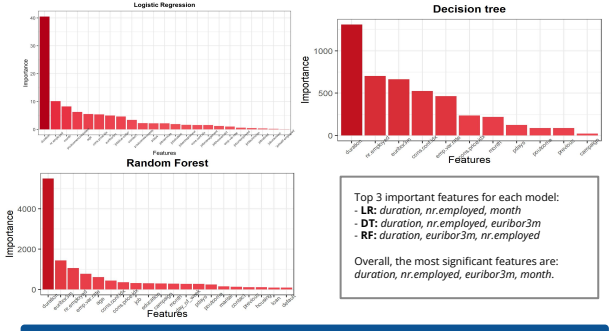


Figure 2: Correlation heatmap of numeric variables

Model Summary



Feature Importance



Top 3 important features for each model:
- **LR:** *duration, nr.employed, month*
- **DT:** *duration, nr.employed, euribor3m*
- **RF:** *duration, euribor3m, nr.employed*

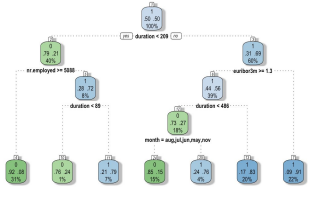
Overall, the most significant features are: *duration, nr.employed, euribor3m, month.*

Modeling Pipeline

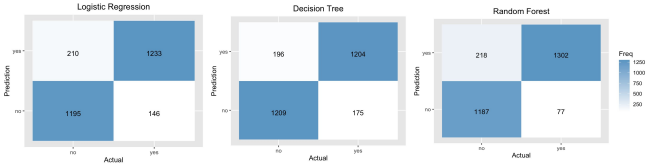
Logistic Regression:
1. Compare the Information Value and drop the features with low IV.
2. Apply WOE on some of the variables.
3. Fitting and comparing.

Decision Tree:
1. Find the best value of cp by cross-validation
2. Pruning the tree using the best value of cp

Random Forest:
1. Find optimal mtry and ntree via iteration
2. Apply weights to the training set based on data imbalance in *y*
3. Fit the model



Model Evaluation - Confusion Matrix

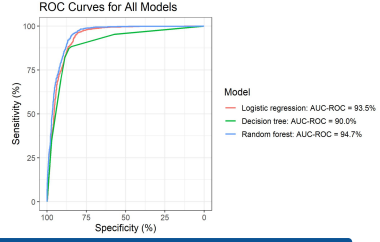


Confusion matrix for each model was plotted in order to show the difference in performance. The color closed to white represents a smaller frequency, while that closed to deep blue represents a larger frequency. Overall, all of three models produce convincing predictions, but further evaluation should be conducted for more accurate comparisons.

Model Evaluation - ROC & AUC

Table 1: Models summary

Model	Accuracy	Kappa	AUC
LR	0.872	0.744	0.935
DT	0.869	0.737	0.900
RF	**0.900**	**0.790**	**0.947**



The table above summarizes the evaluation results of the three models. Considering three essential criteria: accuracy, kappa and auc, random forest has the best performance in every field.

Model Evaluation - Analysis

To give a good prediction for clients' subscription, three supervised learning models are trained respectively using balanced undersampling data. For the test dataset, three criteria are applied to evaluate the predicting results.

To sum up, **Random Forest** is the best among the three models with the highest AUC-ROC score of 94.7%. The weighted model counters data imbalance well. **Logistic Regression** has the second higher accuracy and Kappa. Its AUC is slightly lower than Random Forest. This model still faces an imbalance problem in spite of resampling. **Decision Tree** is the worst one with the lowest accuracy and Kappa. Its AUC is much lower than the other models.

This summary shows that a single tree is not sufficient in producing effective result with this dataset. The application of ensemble learning, which combines the output of multiple trees in order to generate the final output, highly improves the performance of a single decision tree.

Conclusion

During modeling, it is discovered that *duration, nr.employed, euribor3m*, and *month* are the most important factors that correlate with whether the client will subscribe to the bank term deposit or not. In other words, it is a good strategy to prolong the duration. The longer the duration of your calling, the higher possibility that you have to succeed. Most of the successful cases have a duration around 8 minutes. Moreover, during the period that has a lower employment rate, the client will be more likely to subscribe a term deposit. Additionally, since the logistic regression produces a simple linear relationship, a scoreboard is also workable for building up a client base.