XGB: Predicting Car Insurance Cold Calls Success

Mahendri Dwicahyo
Department of Computational Statistics
Politeknik Statistika STIS
Jakarta, Indonesia
15.8727@stis.ac.id

***Abstract*— Data have been widely used by insurance industry to provide better service or improve their process. Cold call is one of activities that can gain values by using analysis of the data. In this paper, Extreme Gradient Boosting (XGB) is used to develop a classification model that could help in predicting whether or not cold call can succeed. Performance of the model is compared to another algorithm such as Logistic Regression, Random Forest. The results show that XGB performs better than the other two, which has an overall accuracy of 75.84%.**

I. BACKGROUND

Car insurance is insurance for car's protection against physical damage or liability of traffic accidents when using the car. Some of them even include stolen car insurance. They always compete to get a new client in all services they had, for example car insurance. Industry often organize campaign to attract potential clients to buy their services, often spend money in the process. Randomly pick people can potentially waste money and time. In order to produce efficient results, it is necessary to know what trait will improve chance of client to buy the insurance.

Data are one ways to help solve the problem of calling the correct person. With a lot of records containing the details of potential client and previous history, one branch of statistics which involves prediction will help in making educated guess what type of person is should be considered in making calls. Prediction here is of classification type with two choices, success or not in persuading potential clients to buy car insurance.

In the last decades, computing power improve greatly and execution time of computer program become many folds faster. Area that explicitly affected by the recent progress is statistics. Algorithm that involve iteration and optimization seen as possible solution in predictive modeling thanks to the advancement. Logistic regression, random forest, and XGB[1] are several examples of algorithm used in prediction. They used iteration to minimize error and improve accuracy.

In this paper, the modeling process start from obtaining the data which available on Kaggle, data is from a bank in USA. Preprocessing is needed to address missing values and encoding variables. Visualize the data to know relationships among variables. Feature engineering, feature selection to improve accuracy and robustness of the model. Finally, test data on logistic regression, random forest, and XGB. The best model is the model which have higher accuracy, the chosen model then interpreted.

II. METHODS

Predicting cold calls success involves classification where the classification algorithm find the pattern from the data. There are many algorithm that able to accomplish the goal, Naive Bayes, K-Nearest Neighbors, Neural Networks. In this paper the algorithm used is Extreme Gradient Boosting (XGB) and compared to Random Forest, Logistic Regression. All of the processes involved in this paper is implemented in R programming.

XGB works by combining weak model, start from very high bias predictor then train on error produced and sum up the two predicting model. The process is iterated many more until the error is low. This is why XGB has potential to over-fitting the data, because the iteration fit on error of the model. Hyper-parameters in XGB is flexible to address over-fitting, this is why the process is not automatic. The one who implement XGB must balance between bias and variance trade-off. XGB is more general algorithm than Ada Boosting, where XGB can use several loss function. Loss function for success and fail (binary problems) is log loss function, while for continuous is root mean square error function.

Decision Tree also can be utilized when dealing with binary classification. Random Forest is an algorithm of Decision Tree. Random Forest elevate the bagging (bootstrap aggregating) where it solve the problems found in CART like over-fitting and correlation. Bagging works by building a lot of tree, where each tree is trained on sub-sample of the data by bootstrap. Each split on a tree use subset of variables, thus ensuring in-dependency and minimum correlation between tree. Random Forest test the performance by Out Of Bag (OOB), this resemble cross-validation process. Random Forest is quite robust to over-fitting.

Logistic Regression is classical way of dealing with binary classification. The prediction is the result of boundary line that constructed by combination of variables. Implementation of Logistic Regression is much more simple than previously mentioned algorithms.

A. Dataset

Dataset comes from Data Mining Cup (SS 2017) of TU Munchen where it describe data originated from a Bank in the US, data publicly available to download at Kaggle. Data consist of 4000 rows and 19 variables, included the label (buy or not buy car insurance). The variable description is following

- Id: Unique ID number. Predictions file should contain this feature **numeric**
- Age: Age of the client **numeric**
- Job: Job of the client. admin., blue-collar, etc.
- Marital: Marital status of the client. divorced, married, single
- Education: Education level of the client. primary, secondary, etc.
- Default: Has credit in default? yes, no
- Balance: Average yearly balance, in USD **numeric**
- HHInsurance: Is household insured? yes, no
- CarLoan: Has the client a car loan? yes, no
- Communication: Contact communication type. cellular, telephone, NA
- LastContactMonth: Month of the last contact. jan, feb, etc.
- LastContactDay: Day of the last contact **numeric**
- CallStart: Start time of the last call (HH:MM:SS). 12:43:15
- CallEnd: End time of the last call (HH:MM:SS). 12:43:15
- NoOfContacts: Number of contacts performed uring this campaign for this client **numeric**
- DaysPassed: Number of days that passed by after the client was last contacted from a previous campaign. **numeric**, -1 means client was not previously contacted
- PrevAttempts: Number of contacts performed before this campaign and for this client **numeric**
- Outcome: Outcome of the previous marketing campaign. failure, other, success, NA
- CarInsurance: Has the client subscribed a CarInsurance? yes, no

From 18 predictor variables only Age, NoOfContacts, DaysPassed, PrevAttempts is of numeric type, the others is considered to be factor.

B. Pre-processing

Performing analysis from the data require the data must clean from missing values and understanding the data. Fail to do that might produce anomalies or exception in calculation. Ensure that structure of the data and values represented are in format that desired. Missing values reduce usable records for building prediction model. In this process, minimize number of records that must be deleted by using knowledge that might help in imputation or use algorithm such as K-Nearest Neighbors. Data that will be used contain many missing values and in analysis section method used for imputation is described.

C. Feature Engineering & Selection

Feature engineering deals with issues related to gain more information from the data by combining several variables to create a new feature or vice versa. The decision for feature engineering mainly based on domain knowledge or exploration through Exploratory Data Analysis (EDA).

Feature selection aim is to remove some of features that is used in feature engineering, ensuring that variables only suffer from small multicollinearity. Features that do not group target variable efficiently will be removed. For example where one group from another is only differ slightly. The reason for this is to prevent over-fitting and build features that only contain high information (split of tree algorithm will be more decisive).

III. Analysis

A. Pre-processing results

All of the process below are done by using R. Default, HHInsurance, CarLoan, LastContactDay, CarInsurance are numeric, so need to convert it into factor. Another case is with LastContactMonth. Looking at the str of the data, the order of month is messed up so need to manually re-order it. CallStart and CallEnd is treated as factor, change it as POSIXct.

Missing values is in Job, Education, Communication, and Outcome.

- Job: have values like retired and unemployed, so assume that NA is really a missing value and need to remove it
- Education: contain three levels of education and 169 missing values. But rather remove all missing values, prefer filling to some of it. Student with age more than 18 treat it as having tertiary education. As in the US, average age of freshman is 18-22 years old. For management, services, admin. assume that minimum require diploma or bachelors degree (according to US Bureau of Labor), so fill three jobs above with tertiary education. The rest of missing values need to be removed
- Communication: assume NA here mean that bank do not communicate with client using phone, so code it to no-comm
- Outcome: The number of missing values is high, approximately only a quarter have values, but do not remove it and address it in feature engineering

After check on each feature, remove row that still contain NA in Job and Education, but exclude Outcome as it will be addressed later. In the end from the data, 3894 row can be used. That number is much better rather than deleting all NA in communication and education.

B. Exploratory Data Analysis

Skip the correlation plot, as the data donimated by categorical features. The approach used is to use bar chart between features and label to see how each of features affect label and here are the results

- Student, retired, and unemployed have chance to buy insurance more than 0.5, it can be an indication that people that do not work will have higher chance of buying. While the others have chance less than 0.5
- In marital status, married person has less percentage in buying insurance so it is save while divorced or single having higher percentage. It could be that not married person will more likely to buy insurance
- People with tertiary education are more likely to buy insurance. Primary to seconday level saw negligible

increase in chance. Based on this, feature can be make more simple

- People that have credit in default less likely to buy insurance. This could be interpreted that people have financial difficulty
- Having household insurance make people less likely to buy insurance. The reason may people with HH insurance bought car insurance too when buying HH insurance
- As to be expected, people that have car loan will less likely to buy car insurance. This might be to loan a car needed to buy insurance too for the loaned car
- Communicate with people that have cellular or telephone will have greater chance to success. Either cellular or telephone is fine, there is no significant difference. This might be used to simplify features
- No clear pattern in LastContactDay, it might be because difference in month or year make effect differ randomly
- March, Sept, Oct, Dec are peak months where people will consider buying car insurance. This might have something to do with people searching another insurance option or from the effect of buying a new car (when in December)
- Outcome of previous campaign contribute largely to people decision to buy insurance. This result means that people satisfied with bank services so client decided to buy another services, which in this case is a car insurance
- Days passed in previous campaign has influence in decision of buying insurance or not. After more than 390 days not called, client more likely to buy insurance
- People that not called in previoys will more likely to reject offer from insurance call

C. Feature Engineering Results

This section deals with action that inspired by results in EDA section, mainly focus on creating or recode features

- Outcome is first that should be handled. After seeing outcome plot, it is reasonable then to group outcome into success or not success. Included in not success is other, failure, and NA
- Move on to Job, as described on EDA section, it might good to try create a new feature based on working or not working condition
- Marital status can also be used as basis to create a new feature, indicator of married or not. Divorced or single will become one group
- Education will be simplified into tertiary or not tertiary
- Next is Communication, plot of communication shows that grouping cellular and telephone as one might reasonable as difference between them is small
- Last come from DaysPassed, where it will become basis for two new features. First is indicator whether 390 days has passed since last contact from a previous campaign. Second is indicator whether client was contacted in a previous campaign

D. Feature Removal

Before the data can be used for modeling, several variables are needed to be deleted to free the predictor from autocorrelation or over-fitting. Variables that already coded into new features in previous section are removed, in addition to ID where it does not mean anything for prediction

E. Build Model

Logistic Regression use generalized linear model with five fold cross validation. Random Forest use 5 variables at split decision and five fold cross validation. XGB use 1000 maximum iteration, maximum depth of a tree of four, learning rate by 0.008, gamma equal one, subsample of records by 0.8 ratio per iteration, 0.8 ratio subsample of variables per iteration, minimum sum of instance weight of 0.8.

IV. Results

The model then runned using control and hyperparameters above. The results are following

TABLE I
LOGISTIC REGRESSION

Accuracy	0.7365
Kappa	0.4138
Sensitivity	0.9140
Specifity	0.4728

TABLE II
RANDOM FOREST

Accuracy	0.7468
Kappa	0.4354
Sensitivity	0.9269
Specifity	0.4792

TABLE III
XGB

Accuracy	0.7584
Kappa	0.4627
Sensitivity	0.9312
Specifity	0.5016

V. Discussion

According to the results on three algorithms, XGB perform better than logistic regression and random forest. Although for XGB, tuning for balance between overfitting and underfitting is required to get the best results. There is some degree of agreement between model, PrevSuccess, CommPhone, HHInsurance are features that placed in top 10 in all three models. Previous success in campaign will help call turn to a success. With communication over phone bring success to persuade client. Household insurance effect is reciprocal to call success, someone with household insurance less likely to buy car insurance. Without current campaigns condition variables, accuracy obtained is more than 75%.

REFERENCES

[1] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.