greatlearning

Capstone Presentation

Taiwan: Customer Defaults

Submitted By:

- *Suprasanna Pradhan*
  BABI Dec -2018

greatlearning

Business Problem Understanding

Problem Statement :

A Taiwan-based credit card issuer wants to better predict the possibility of default for its customers , trying to identify the key drivers that determine and minimize the risk . This would provide the appropriate information to offer credit card with adequate credit limit . The issuer also needs to understand their current and potential customers and their future strategy, including the planning of offering targeted credit products to their customers.

Constrain:

Carefully deciding what the dependent 0/1 variable is can be the most critical choice of a classification analysis. This decision typically depends on contextual knowledge and needs to be revisited multiple times throughout a data analytics project.

Scope:

The credit card issuer has gathered information on 30000 customers. The dataset contains information on 24 variables, including demographic factors, credit data, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default or not

Objectives:

From a Risk Management Perspective a Bank/Credit Card Company is more interested in minimizing their losses towards a particular customer.to compute the predictive accuracy of probability of default for a Taiwanese Credit Card Client. We have to predict the right value of probability of defaulters.

greatlearning

Modelling Approach Used & why

These models we have used to predict the right value:

Logistic Regression: Since the dependent variable is binary and we need check the multicollinearity of the variable before preparing the final model of logistic regression. We found six variables are carried out good coefficient and four of them are negative coefficient ,only two predictor show positive impact . Further we expected the other model may perform better

CART : The CART method is able determine the complex interactions among variables in the final tree, in contrast to identifying and defining the interactions in a multivariable logistic regression model. After pruned the AUC value found 76%, so we decide to check other models.

Random Forest: RF is tree-based strategies naturally it ranks by how well the model improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity)and It reduces the complexity of a model and makes it easier to interpret.

KNN Classification: KNN is a non-parametric model which means that it does not make any assumptions about the data set. It finds intense application in pattern recognition, data mining and intrusion detection

Naive Bayes -The Bayes theorem is used to calculate the conditional probability, it provide the equal opportunities to all variables hence we thought the model will perform better compare to other models

Bagging and Boosting :Bagging and Boosting decrease the variance of single estimate as they combine several estimates from different models. So the result may be a model with **higher stability**.

KNN Cross Folding :The k-fold cross-validation procedure is repeated n times, where importantly, the data sample is shuffled prior to each repetition, which results in a different split of the sample from the data set.

greatlearning

Insights from Analysis

- We found data set consisted 30000 observations with 25 variables.
- The categorical data value like Sex , marital status and education has changed to numeric value .
- We have realized that 22.1 % percent defaulter and 77.9% are not default cases
- Default category whereas male customer is 9.6% and female category shows 12.5% leaning to defaulted
- University level - graduate or PG is more into default side
- Married customers somehow lean towards defaulter
- Average age of 25 to 30 is the highest risk .
- We have also checked the multicollinearity problems is existed in the data set , pay status categorical variables are dependent on each other and impact of REPAY_ SEP to REPAY_APR variables to default. Payment DEFAULT is high.
- We have also created some dummy variables like ratio of the payment for each month SEP to APR and balance amount month wise from SEP to APR .
- We have added some new variables like payment ratio , timely payment and found in September22%, August 14%, July 14%, June 11%, May 09 and April it is 10% customer are paid on time
- Performed FA and created final data set with new variables (enclosed train data set below)
- Performed data split with balancing of the samples in test and train data set

Recommendations

Models	Accuracy
Logistic Regression	87%
CART	76%
Random Forest	76%
KNN	66%
NB	73%
Bagging	76%
Boosting	77%
K-flod	50%

- The best models are **Logistic Regression** , by which we predicted with **87 % accuracy**, whether a customer is likely to default next month. Whereas **XGBOOST method can 77% accuracy .**
- The strongest predictors of default are the **PAY_X (ie the repayment status in previous months), the LIMIT_BAL & the PAY_AMTX (amount paid in previous months)**.
- We have seen also that being Female, More educated, Single and between 30-40years old means a customer is more likely to make payments on time.