Project Proposal

Tultul Majumder
1.12.2017

1. Motivation:

The problem is to classify borrower as defaulter or non-defaulter. It is commonly desired for loan giving agencies to classify borrower accurately so as to manage their loan risk better and increase business. Before giving a credit loan to borrowers, any agency decides who is bad (defaulter) or good (non defaulter) borrower. The prediction of borrower status i.e. in future borrower will be defaulter or non-defaulter is a challenging task for them.

However developing such a model is a very challenging due to growing demand for loans. Data size is very big and some question arises about relevancy of data. Some institutes use data warehouse technologies to support development and update the models. Mostly banks use the logistic regression to evaluate customer risk.

In this project, I will try to create a delinquency model which can predict in terms of a probabilistic label for each loan transaction whether the customer will be paying back within 5 days of issuance of loan.

This project will attempt to use some traditional classification models like Logistic regression, Random Forrest, Naïve Bayesian classifier and attempt to create model to better predict customer default.

2. Data description:

Pinnacle is a leading provider of innovative revenue uplift and customer value management products for Telecom operators globally. For this project, a data set with features created by the team is used. Data contains around 1M observations for past 3 months.

2.1. The list of variables and their definitions are as follows:

Variable	Definition
label	Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success, 0:failure}
msisdn	mobile number of user
aon	age on cellular network in days
daily_decr30	Daily amount spent from main account, averaged over last 30 days
daily_decr90	Daily amount spent from main account, averaged over last 90 days
rental30	Average main account balance over last 30 days

rental90	Average main account balance over last 90 days
last_rech_date_ma	Number of days till last recharge of main account
last_rech_date_da	Number of days till last recharge of data account
last_rech_amt_ma	Amount of last recharge of main account
cnt_ma_rech30	Number of times main account got recharged in last 30 days
fr_ma_rech30	Frequency of main account recharged in last 30 days
sumamnt_ma_rech30	Total amount of recharge in main account over last 30 days
cnt_ma_rech90	Number of times main account got recharged in last 90 days
fr_ma_rech90	Frequency of main account recharged in last 90 days
sumamnt_ma_rech90	Total amount of recharge in main account over last 90 days
cnt_da_rech30	Number of times data account got recharged in last 30 days
fr_da_rech30	Frequency of data account recharged in last 30 days
cnt_da_rech90	Number of times data account got recharged in last 90 days
fr_da_rech90	Frequency of data account recharged in last 90 days
cnt_loans30	Number of loans taken by user in last 30 days
amnt_loans30	Total amount of loans taken by user in last 30 days
maxamnt_loans30	maximum amount of loan taken by the user in last 30 days
medianamnt_loans30	Median of amounts of loan taken by the user in last 30 days
cnt_loans90	Number of loans taken by user in last 90 days
amnt_loans90	Total amount of loans taken by user in last 90 days
maxamnt_loans90	maximum amount of loan taken by the user in last 90 days
medianamnt_loans90	Median of amounts of loan taken by the user in last 90 days
payback30	Average payback time in days over last 30 days
payback90	Average payback time in days over last 90 days
pcircle	telecom circle
pdate	date

3. Approach:

Building different score model candidates using different classes of methods. Use only the training subset to estimate the model parameters. For each of the classes of models will try different parameter settings. My experience is that the parameter settings could have a great impact on the performance of the model. I plan to use an ensemble of models.

4. Statistical tools to be used:

Python : Scikit-Learn, Pandas, Numpy, Matplotlib, Seaborn

Visualization: Tableau