Tournament

The Pheasant

Problem Statement

- People in the informal sector such as farmers and freelance workers usually have no access to formal credit due to their lack of formal credit history and document such as pay slips and employment records.
- Some research (Björkegren and Grissen 2015) suggests that mobile phone records can provide an equivalent information to formal credit records in classifying good and bad credit.
- As a Fin Tech analyst, you are assigned to build a machine learning model to classify credit customers from their pre-paid phone records.

Model

Predictors

Customer Profile
- Age
- Gender
- PayType
- Province

Mobility
- Count of call numbers
- Count of call locations

Call Pattern
- Frequency
- Mean call duration
- SD call duration

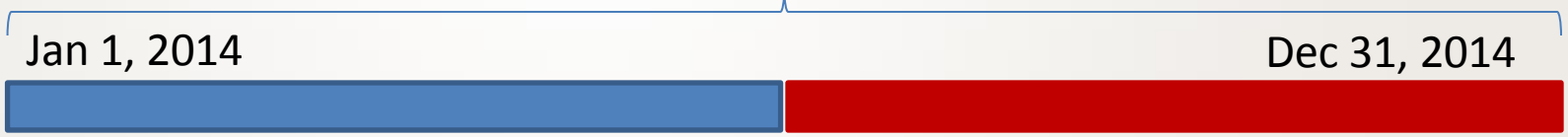
Payment stability
- Payment pattern

Response

Delinquency
- Default (bad=1)
- Non-default (good=0)

Data

- Prepaid phone data:
  - Consist of 3 tables: customer profiles, call detail records, payment records.
  - Collect from Jan 01, 2014 to Dec 31, 2014.



- Credit data:
  - Consist of 1 table showing customers' delinquency.
  - Collect only customers applying for credit and getting approved from July, 2014 to Dec 31, 2014.
  - Follow the delinquency 1 year after credit approved.

Data Structure

Predictor Tables

Profile		
Column name	Data Type	Note
id	Character	
phonenum	Character	
birthdate	Date	
gender	Factor {M,F}	
paytype	Character	Prepaid
province	Character	

CDR (Call Detail Records)		
Column name	Data Type	Note
timestamp	DateTime	
callingnum	Character	
callednum	Character	
duration	Numeric	Minutes
location	Character	

Payment Records		
Column name	Data Type	Note
timestamp	DateTime	
callingnum	Character	
topup	Numeric	Bahts
spending	Numeric	Bahts
balance	Numeric	Bahts
status	Factor {ACTIVE,INACTIVE}	

Data Structure

Response Table

Response		
Column name	Data Type	Note
id	Character	
status	Factor {0,1}	

Rules

- Individual effort.
- Sample data will be given on Tue March 8.
- Competition date: Sat March 12 from 9:00 to 16:00.
- Allow discussion before the competition day.
- Discussion is not allowed on the competition day.
- Code can be prepared prior to the competition day, but required to be written individually.
- Encourage applicants to use their own personal computer.
- Encourage R and Python.

Format

- Full set of training data (both predictors and responses) will be given on **Tue March 8**
- Full set of test data (only predictors) will be given on Sat March 12 at **11:00**.
- Each applicant will submit the predicted responses of the test data and the AUC will be shown on the leader board
- Each applicant may submit a maximum of 3 entries before **16:00**.
- Evaluation will be based solely on **AUC**

Submission

- A csv file with 2 columns
  - profile_id
  - probability of being in the default state P(response = 1)
- The file must contain 1000 profile_id from the given test set

Important Dates

- March 8, 2016: Upload sample data and training data
- March 12 2016: Competition day
- March 21, 2016: Winners announcement