Big Data Programming
(CBG1C04)

ASSIGNMENT PRESENTATION

22-FEB-2018

Presented By:

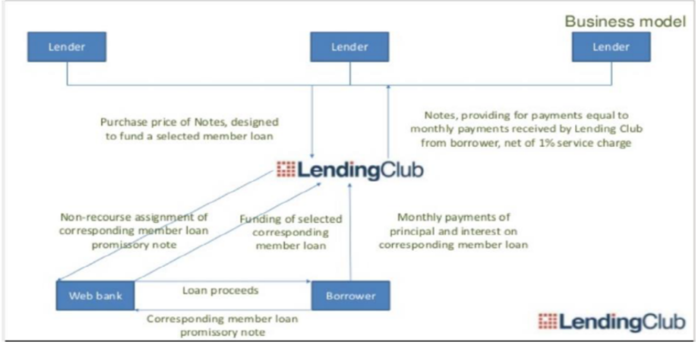
Lim Yuan Her (1780113E)

Agenda

- **Application Domain**
- **Research Goal**
- **Data Sources/ Schema**
- **Data Cleaning/ Transformation/ Exploration**

Lending Club - Application Domain

- ❑ Online crowdfunding platform for peer to peer lending, facilitating personal loans, business loans, and financing.
- ❑ Borrowers access loans through online or mobile interface and investors provide capital in exchange for earning interest (peer-to-peer (P2P) lending).
- ❑ Online-only operation results in cheaper operating costs and overheads, thus offering lenders higher returns compared to traditional bank products, and borrowers can borrow money at lower interest rates, even after accounting for platform and credit checking fees.
- ❑ Interest rates set by lenders who compete for lowest rate based on reverse auction model or fixed based on borrower's credit profile.


Business model
Lender
Lender
Lender
Purchase price of Notes, designed to fund a selected member loan
Notes, providing for payments equal to monthly payments received by Lending Club from borrower, net of 1% service charge
LendingClub
Non-recourse assignment of corresponding member loan promissory note
Funding of selected corresponding member loan
Monthly payments of principal and interest on corresponding member loan
Web bank
Loan proceeds
Borrower
Corresponding member loan promissory note
LendingClub

Research Goal

- ❑ Attempt to predict the risk of the loan being default based on past loan data

- ❑ Data from LendingClub's website (https://www.lendingclub.com/info/download-data.action).

- ❑ Use partial loan data from period 2017 (Q1 to Q3) as training and cross-validation set and rest as testing set.

- ❑ Logistic Regression/ Decision Tree used with goal of maximizing areaUnderROC

Data Sources/ Schema

- ❑ Dataset from 2 input files:
  - ❖ LoanStats.csv - loan data for all loans issued including current loan status and payment information
  - ❖ LCDataDictionary.xlsx - Data Dictionary containing definitions for all data attributes

- ❑ Each loan record (324,939 records in loan data file) has 145 attributes (columns) and Data Dictionary has 2 attributes (attribute name, description)

- ❑ 7 loan statuses in “loan_status” column : Charged Off, Current, Default, Fully Paid, In Grace Period, Late (16-30 days), Late (31-120 days)
  - ▪ “Charged Off” (defaulted loan) and “Fully Paid” (non-defaulted loan) statuses are considered in analysis

Data Cleaning/ Transformation/ Exploration

- ❑ Data Cleaning:
  - Drop columns with >50% missing values
  - Drop columns with strong correlations with each other
  - Remove columns of date datatype
  - Remove duplicate records
  - Remove records having missing values in any of the columns
  - Check Outlier Data
  - Check for distinct values in columns

- ❑ Data Transformation
  - Encode categorical features into numerical representation

- ❑ Data Exploration
  - Descriptive Statistics
  - Correlation/ Skewess for numerical features
  - Histograms (distribution profile of numerical features)
  - Box/ Density Plots (correlation of loan status with numerical features)
  - Mosaic Plots (correlation between categorical features)
  - Heat Maps (correlation of categorical feature groups with numerical features)
  - Pareto plots (distribution profile for categorical features)

Q&A