

# A P2P Lending Credit Model for Chinese Overseas Students Loan Market

MAIMAITI CONTEST DEMO

09/25/2016

# TEAM BIO



**LYLA YANG**

Business Analyst

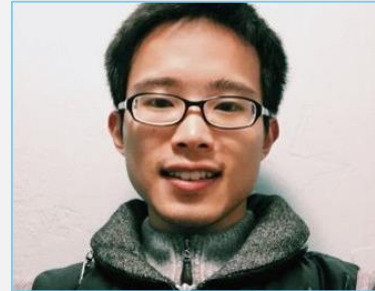
Origin Consulting, LLC



**SHEN LI, CMA**

Sr. Financial Analyst

Techstyle Fashion Group, Inc



**WENHUAN WEI**

Master Student

University of Washington



**ZHAOHU FAN**

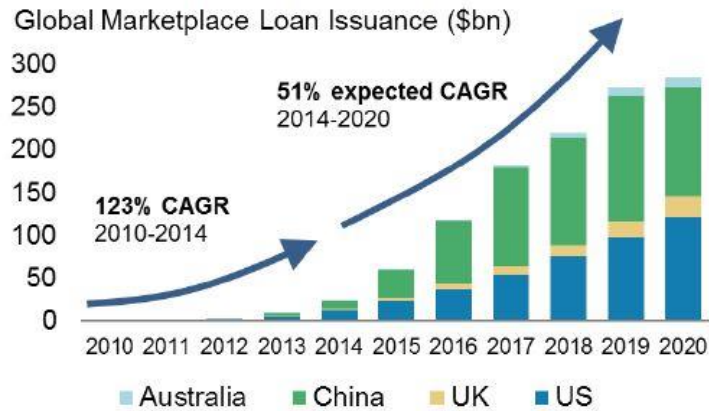
PHD Student

Pennsylvania State University

# LENDING MARKET SIZE GROWTH

Exhibit 4

**We estimate global marketplace lending can reach \$290 billion by 2020 (base case)**



Source: Company Data, Morgan Stanley Research

51% CAGR in lending market size  
20% CAGR in customer base

Assume..

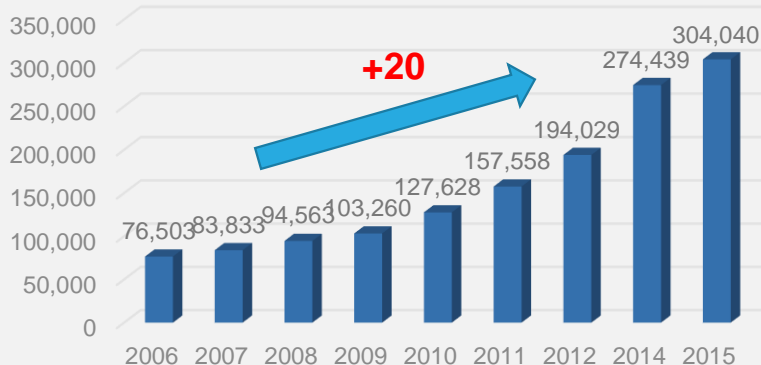
Total Customer acquisition Rate 10%

Average APR 5%

Average Loan per Customer \$5,000/Year

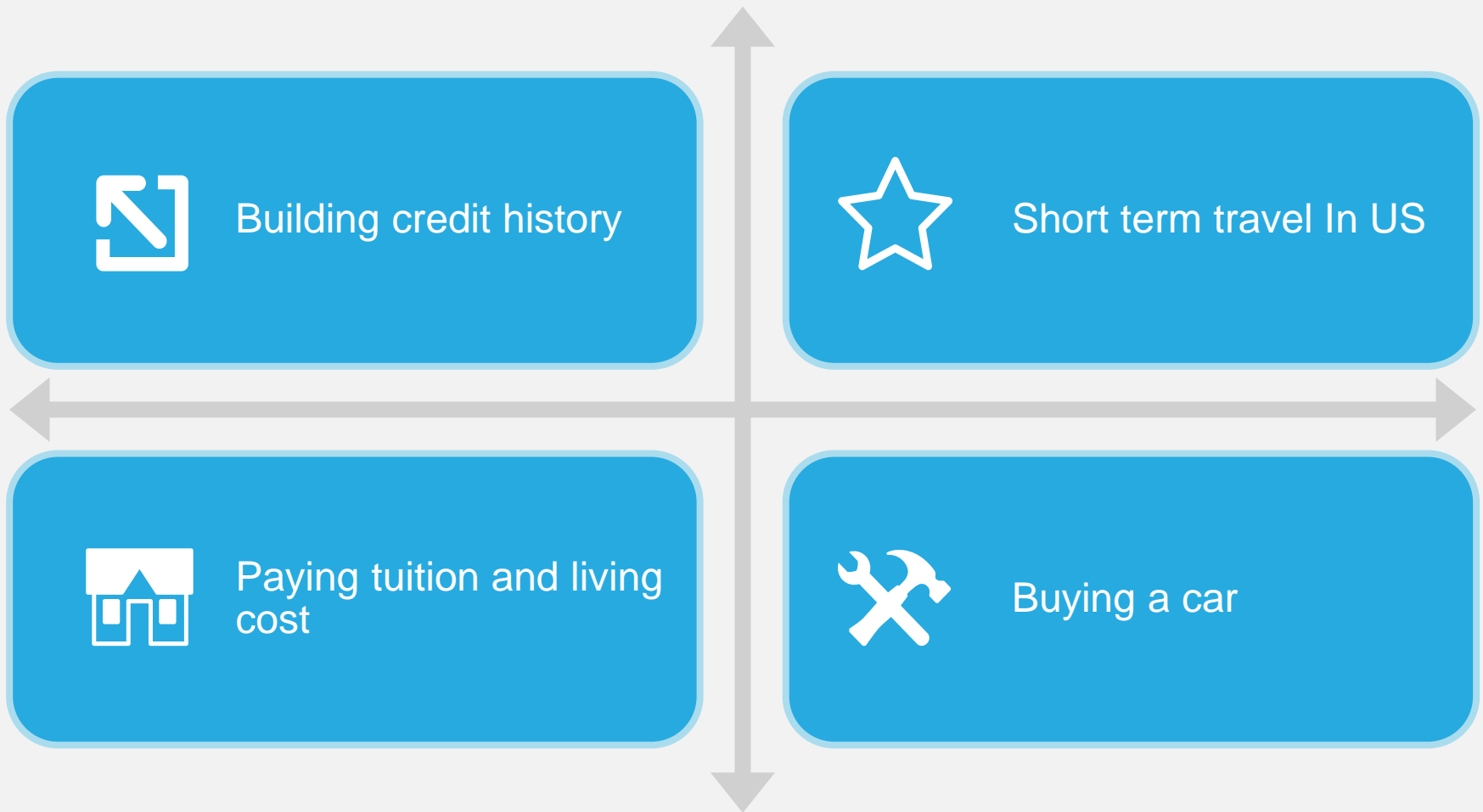
Bad Debts Rate 2%

## Number of Chinese Students in US

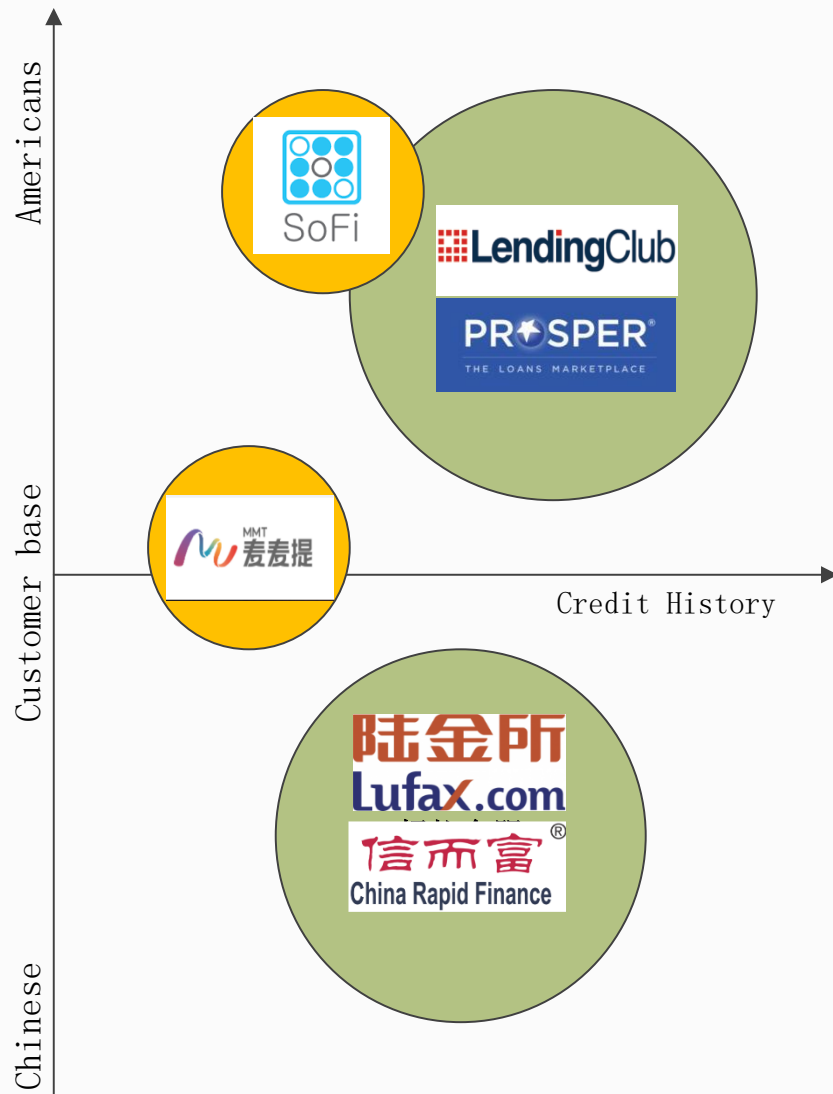


Annual Revenue Estimate: \$4.5M

# CUSTOMER NEEDS



# MARKET COMPETITION



Major P2P lending platform:

- Traditional credit model based on FICO
- Minimum credit score
- 3-5 years loan terms
- Common use is credit card debt

LendingClub

PROSPER<sup>®</sup>  
THE LOANS MARKETPLACE

Player in US college student loan niche market:

- Based on alumni-funded lending model
- Evaluate forward looking factors to determine borrower's future potential
- Use FICO as part of the credit rating process

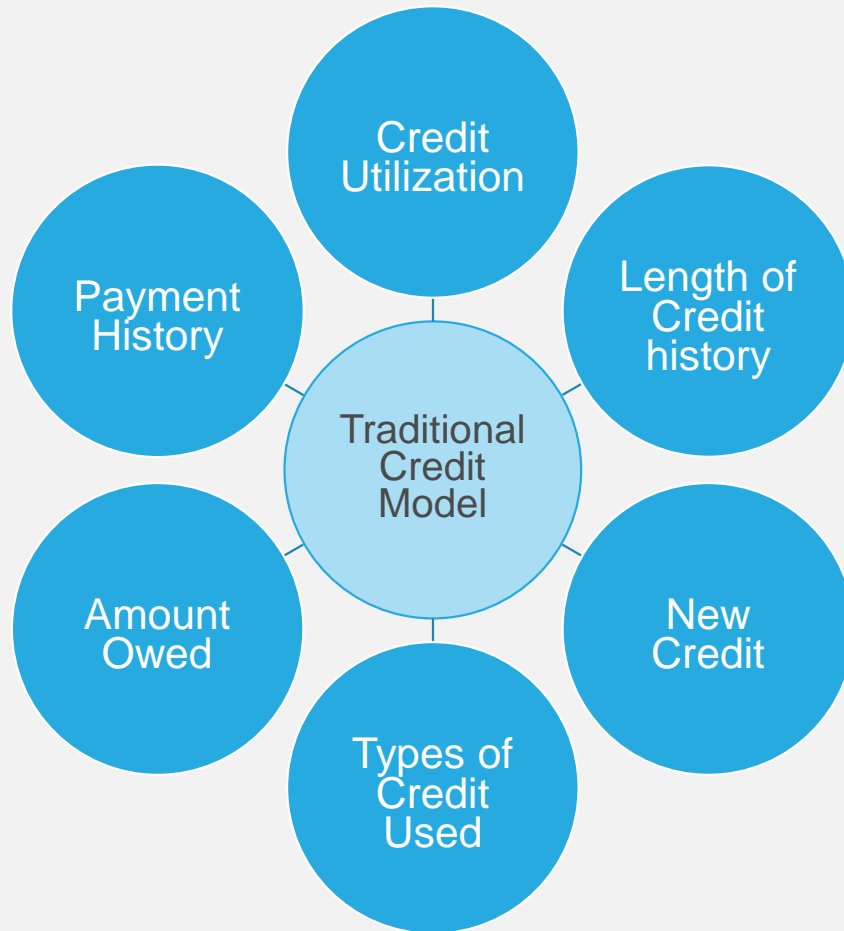


SoFi



CommonBond

# LIMITATION OF TRADITIONAL CREDIT MODEL



However...Our target customer

- Lack of Credit History
- No fixed salary

WE HAVE NO INFORMATION  
TO RATE THEM THEN WHAT  
WE DO?

# DO YOU FEEL LIKE IT IS SIMILAR TO...



## IS HE/SHE THE RIGHT GUY/GIRL?

After chat how do you like him/her?

What are the information you care about?

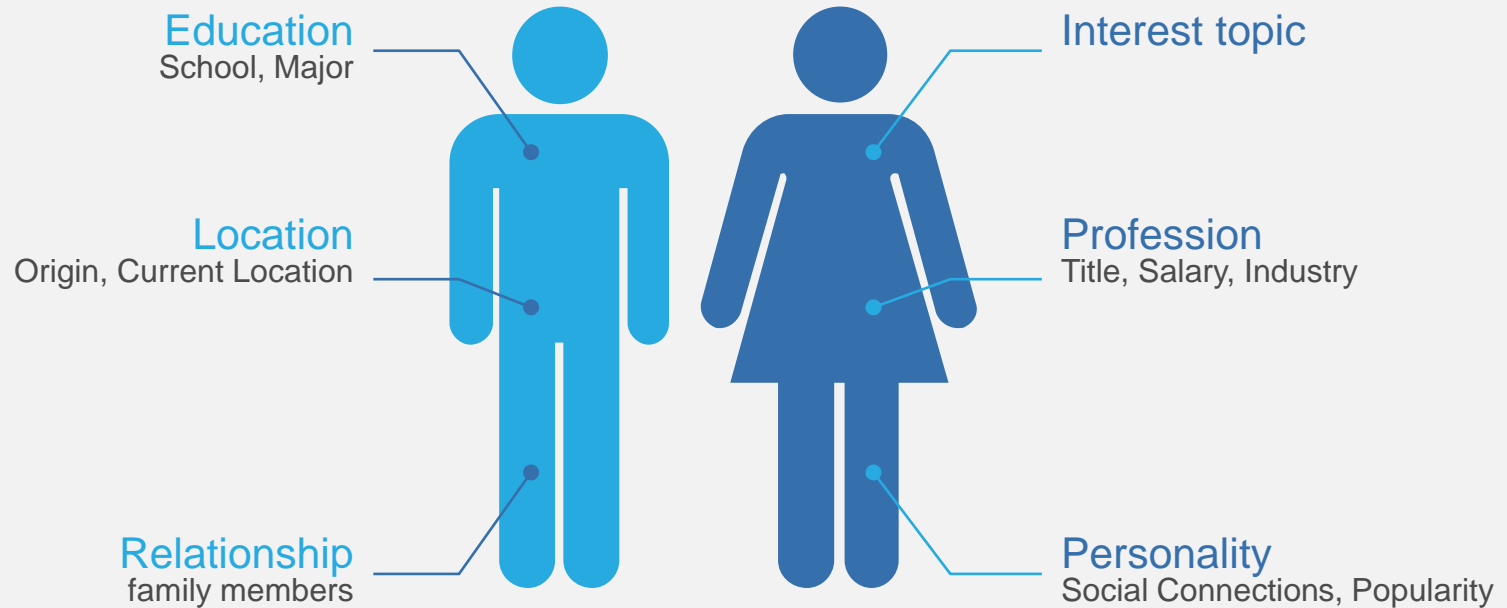
How to prove what he/she says?

So this is the same concept we will use in risk control process:

- Prove what he says (he is telling the truth)
- Evaluate qualification (he qualifies)

# ATTRIBUTES OF THE SOCIAL NETWORK DATA

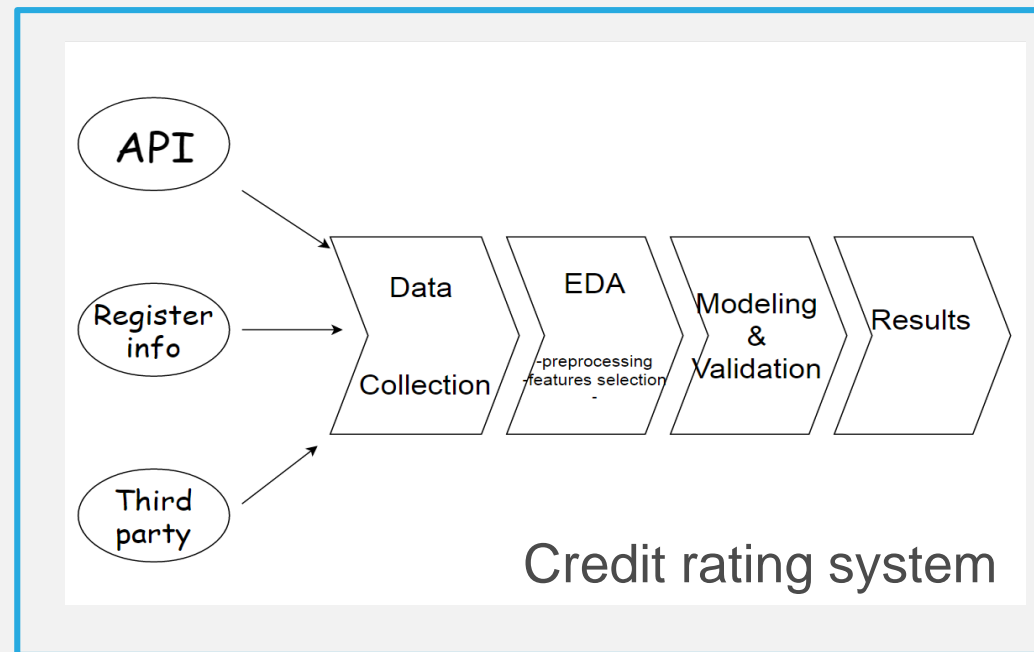
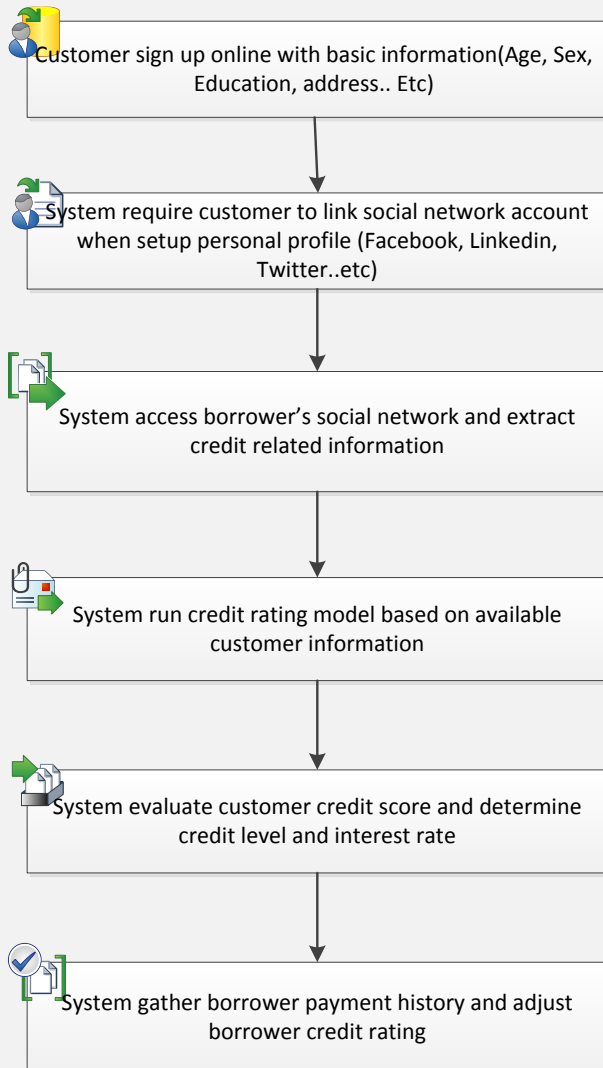
## FIRST IMPRESSION?



- Content – Profile categorization, topic modeling, sentiment analysis, interest mining, etc.
- Context – Location, temporal analysis, behavior trajectory, community, etc.
- Connection – Relationship mining, core network detection.



# CREDIT RATING BUSINESS PROCESS



# CREDIT RATING MODEL DESIGN PROCESS



Lending Club



LASSO

80 Variables  $\gg$  10 Variables

**Machine Learning**  
**Algorithms**



( Python and R Codes )

# SOURCE OF DATA COLLECTED

## Social data sets:

Current Location,  
Education Experience,  
Degrees,  
Majors,  
Grades,  
Cell phone type,  
Place of living,  
Graduation Year,  
Position Title,  
Internships  
Work Experience,  
Traveling History,  
Active social connections  
Active social events  
Etc

## Lending Club:

Variable	Description
LOAN AMOUNT	The listed amount of the loan applied for by the borrower.
TOTAL FUNDING	The total amount committed to that loan at that point in time.
INVESTOR FUNDING	The total amount committed by investors for that loan at that point in time.
TERM	The number of payments on the loan. Values are in months and can be either 36 or 60.
INTEREST RATE	Interest rate on the loan.
INSTALLMENT	The monthly payment owed by the borrower if the loan originates.
INCOME	The annual income provided by the borrower during registration, measured in \$1,000s.
DTI	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
FICO	The average of the last upper and lower boundaries of the range of the borrowers FICO.
CHARGED OFF	An indicator variable equal to 1 if the borrower defaults on the loan and 0 otherwise.

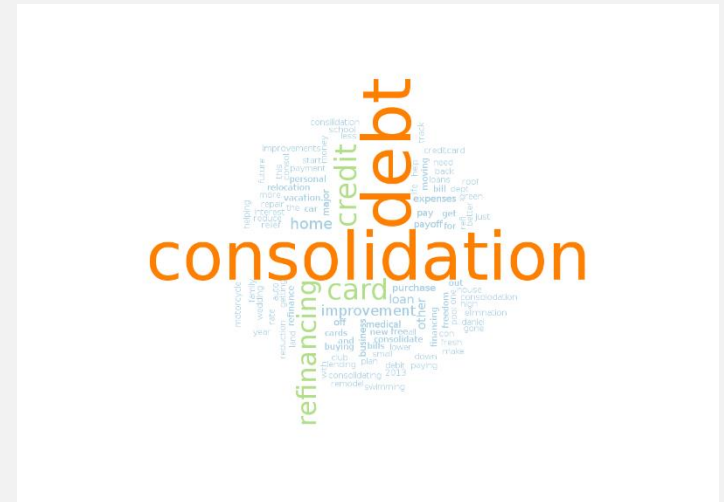
1000 single cases are collected



Used Python BeautifulSoup4 and Selenium library and assign the collecting jobs in distributed mode to collect social network leads data

## CATEGORIES OF INFLUENTIAL PREDICTORS

- Credit history(if available)
  - o FICO Score
  - o Credit length
- Education essentials
  - o School tier
  - o Major(Stem, non Stem)
  - o Degree level
- Ability to pay
  - o Own car?
  - o Account balance
  - o Position title
  - o Place of living
  - o Past purchase transactions on Amazon
- Social interaction
  - o # of active connection
  - o # of Places visited
  - o Frequency of login and communication



### Plot of world cloud - Loan Title

# SAMPLE SOCIAL NETWORK DATA COLLECTED

Titles	Skills ID	Connections	Grades	Honors	Degrees	Schools	NumExperience	AllLocations
['Teaching Assistant', 'Software Developer Internship']	26 en	500+	['3.67']	4	['Masters degree ', 'Bachelor of Science BS ']	['Carnegie Mellon University', 'Nanjing University']	2	['Mountain View, California', 'Greater Seattle Area', 'Education']
['Quantitative Investment Analyst', 'Quantitative Quality Eng	13 en	287	['3.98', 'top 1%']	2	['Master of Science ', 'Bachelor of Science BS ', 'Bachelor of	['Carnegie Mellon University', 'University of Michigan', 'Shanghai Jiao Tong University']	2	['New York, New York', 'Shanghai City, China', 'Shanghai Suburb, China']
['Finance Accounting Intern', 'Research Intern']	6 jiaqi-he-078	219	['3.75']	0	['Master of Science ', 'Bachelor of Science ']	['Carnegie Mellon University', 'Tongji University']	2	['Pittsburgh, Pennsylvania', 'Shanghai City, China', 'Shanghai City, China', 'Social S
['Software Engineer Intern']	15 shaokun-zo	500+	['3.97', '3.77']	0	['Master of Science MS ', 'Bachelor of Science BSc ']	['Carnegie Mellon University', 'City University of Hong Kong']	1	['Greater Pittsburgh Area', 'Hong Kong']
['Department leader of The Student Sports Association']	28 wenhuan-w	384		2	['Masters Degree ', 'Bachelors Degree ']	['University of Washington', 'Shanghai Jiao Tong University']	1	['Seattle, Washington']
['Web Developer']	32 zhitian-zhan	235	['3.74']	0	['Masters Degree ', 'Bachelors Degree ']	['University of Southern California', 'University of Washington']	1	['Greater Los Angeles Area', '11100 Valley Blvd., Suite #200, El Monte, CA 91731']
['Hadoop Engineer', 'Software Engineering Intern', 'Intern',	22 qhdai	500+		2	['Masters degree ', 'Bachelors degree ']	['The University of Texas at Dallas', 'Capital Normal University']	5	['San Francisco Bay Area', 'San Francisco Bay Area', 'Raleigh-Durham, North Caro
['Software Engineer intern', 'Data Analyst intern', 'Intern of	17 tong-zhang	331	['3.80', '3.80']	0	['Master of Science MS ', 'Bachelor of Engineering BE ', 'Exc	['University at Buffalo', 'Xi'an Jiaotong University', 'University of California, Riverside']	3	['San Francisco Bay Area', 'Beijing', ' ', 'Xi'an, Shaanxi, China', 'Children']
['Intern Robotics', 'Software Engineering Intern']	20 en	500+	['90']	0	['Masters degree ', 'Bachelor of Engineering BE ']	['Carnegie Mellon University', 'University of Electronic Science and Technology']	2	['Pittsburgh, Pennsylvania', 'Beijing City, China', 'Social Services', 'Social Services']
['Airline Operations Analyst Intern', 'Grader of Discrete Mat	19 xinpeiz	384	['3.86']	3	['Master of Science in Engineering MSE ', 'Bachelor of Scien	['University of Michigan', 'Shanghai Jiao Tong University']	5	['Ann Arbor, Michigan', 'Gothenburg, Sweden', 'Ann Arbor', 'Shanghai City, China
['Software Engineer', 'Software Engineer Intern', 'Software	27 haishanye	500+	['3.8', '4.63']	3	['Masters Degree ', 'Bachelor of Engineering BE ']	['University of Southern California', 'Tongji University']	3	['Los Angeles, California', 'Cupertino, CA', 'San Francisco Bay Area']
['Student']	19 keyi-kelsey-	500+		0	['Bachelors degree ', 'Bachelors degree ']	['University of Wisconsin-Madison']	1	['Madison, Wisconsin Area']
['Intern']	15 mengqing-q	93		0	['Master of Science MS ', 'Bachelor of Network engineering	['State University of New York at Binghamton']	1	['Binghamton, New York']
['Research Scientist', 'Graduate Teaching Assistant For Data	26 yuedeng05-	500+	['3.73']	0	['Master of Science MS ', 'Bachelor of Engineering BE ']	['University of Maryland College Park', 'Beijing University of Posts and Telecommunicatio	8	['San Francisco, California', 'Hopkinton, Massachusetts', 'Hopkinton, MA', 'Bowie,
['Software Engineer', 'Research Assistant']	19 shaoyi-hu-8	500+	['3.71', '3.68']	0	['Masters degree ', 'Bachelors degree ']	['Carnegie Mellon University', 'Harbin Engineering University']	2	['San Jose, California', 'San Francisco Bay Area', 'Pittsburgh']
['M.S Graduate Researcher', 'Research Intern in Clinical Info	26 en	500+		3	['Master of Science MS ', 'Bachelor of Science BS ', 'High Sc	['Carnegie Mellon University', 'China Pharmaceutical University']	4	['Pittsburgh, Pennsylvania', 'Greater Pittsburgh Area', 'Tarrytown, NY', 'Chengdu,
['Software Engineer', 'Technical Yahoo', 'Software Engineer'	7 xing-jin-a37	219		0		['The University of Texas at Austin']	3	['San Francisco Bay Area']
['EC2 networking software develop engineer', 'co-founder&	11 shuo-chen-	388		0	['Bachelors degree ', 'Bachelors degree ']	[' ', '']	3	['Greater Seattle Area', 'Seattle', 'Shang hai', 'Seattle']
['Internship', 'Student']	15 zijian-hu-68	500+		0	['Masters degree ', 'Masters degree ', 'Bachelor of Science E	['The Ohio State University', 'Carnegie Mellon University', 'Central South University']	2	['Columbus, Ohio', 'Hangzhou, Zhejiang, China', 'Pittsburgh']
['Graduate Teaching Assistant']	27 en	500+		0	['Master of Science MS ', 'Bachelor of Engineering BE ']	['Stevens Institute of Technology', 'Dalian Nationalities University']	1	['Greater New York City Area', 'Hoboken, NJ', 'Science and Technology', 'Educatio
[]	23 jingjingshac	93	['Master', '3.7']	0	['Masters Degree ', 'Bachelors Degree ', 'Exchange Student	['The Ohio State University', 'Beijing Institute of Technology', 'The Hong Kong Polytechnic	0	['Columbus, Ohio']
['Human Resources Trainee', 'Customer Service']	13 omar-wong	18	['3.3']	0	['Bachelors degree ', 'High school diploma ', 'High School ']	['University of Washington Tacoma', 'Green River Community College']	2	['Greater Seattle Area', 'Hong Kong', 'Hong Kong', 'Social Services']
['Software Engineer', 'Computer Science Student']	14 xinpeng-zha	500+	['3.76', '3.67']	1	['Master of Engineering MEng ', 'Bachelor of Engineering BE	['University of California, Berkeley', 'University of Michigan', 'Shanghai Jiao Tong Univers	2	['San Francisco Bay Area', 'Santa Clara', 'San Francisco Bay Area']

# EXPLORATORY DATA ANALYSIS

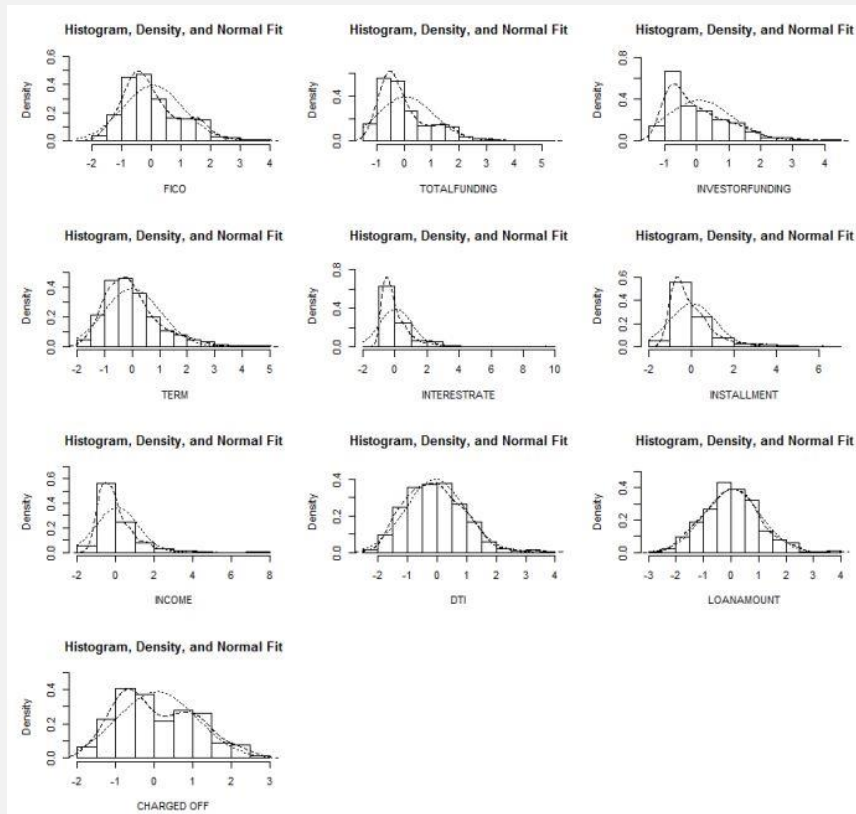


Figure 1: Plot of empirical distribution

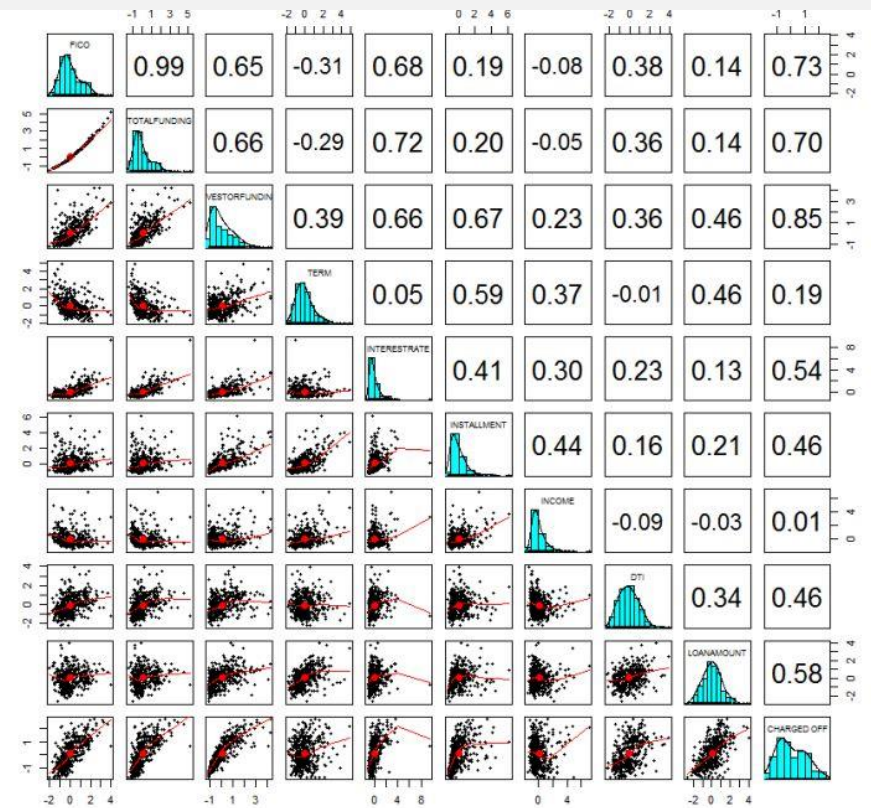


Figure 2: Box plot of selected variables

# PERFORMANCE COMPARISON

Machine Learning Methods	Mean Square Error	Speed
SVMs	-----	Very Slow
RandomForestRegressor(max_depth=5)	0.032	Slow
xgb.XGBRegressor	0.024	Fast
<b>MLR</b>	<b>0.023</b>	
GradientBoostingRegressor	0.026	Fast

So MLR-Modified logistic regression model will be adopted.

# FINAL MODEL

```
> model61=multinom(Grade~FICO+DTI+TOTALFUND+INCOME, data=LendingClub)
> summary(model61)
```

Call:  
`multinom(formula = Grade ~ FICO+DTI+TOTALFUND+INCOME, data=LendingClub)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.2542	-0.3112	-0.0912	0.1160	2.8321

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.02894	0.23822	0.121	0.9033
FICO	2.89578	0.46425	6.238	4.44e-10 ***
DTI	2.05996	0.34310	6.004	1.93e-09***
TOTALFUND	1.22020	0.57564	2.120	0.0340 *
INCOME	-0.61886	0.24286	-2.548	0.0108 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 542.71 on 396 degrees of freedom  
Residual deviance: 172.27 on 394 degrees of freedom  
AIC: 178.27

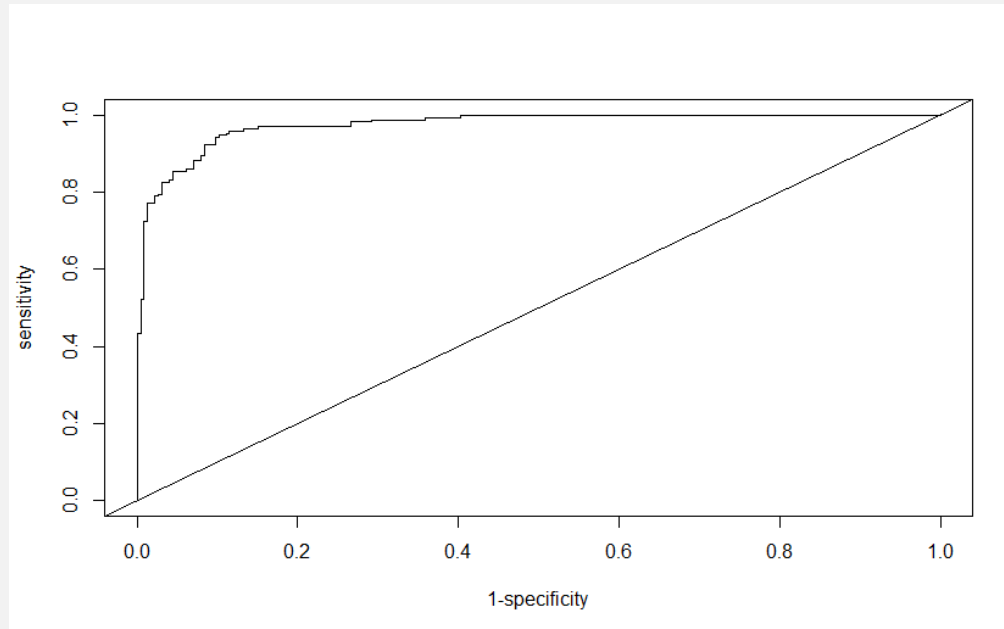
Final proposed model is given as follows(Logistic regression):

$$\log \frac{\pi_i}{1 - \pi_i} = 0.02894 + 2.89578X_{1i} + 2.05996X_{2i} + 1.22020X_{3i} - 0.61886X_{4i}$$



# MODEL VALIDATION

The ROC curve is shown below. From the above output, the A value (i.e., Area Under the Curve) is 0.9745, indicating that the model has good predictive power.



Plot of ROC curve(receiver operating characteristic)

# FUTURE MODEL DEVELOPMENT

- To reduce prediction error, we will split the data set into a training set (80%) and test set (20%). We build our model on the training set while comparing the predictive performance on the test set. We measure performance in terms of the root mean squared error, RMSE.
- When there have been enough successful loan records of many students, we can generate more features of credit of a student from the real loan history of his own, we can modify this social media model to strengthen the scalability and enhance the reliability of our whole risk management system.
- Collect additional data from online purchase history and utilities payment history

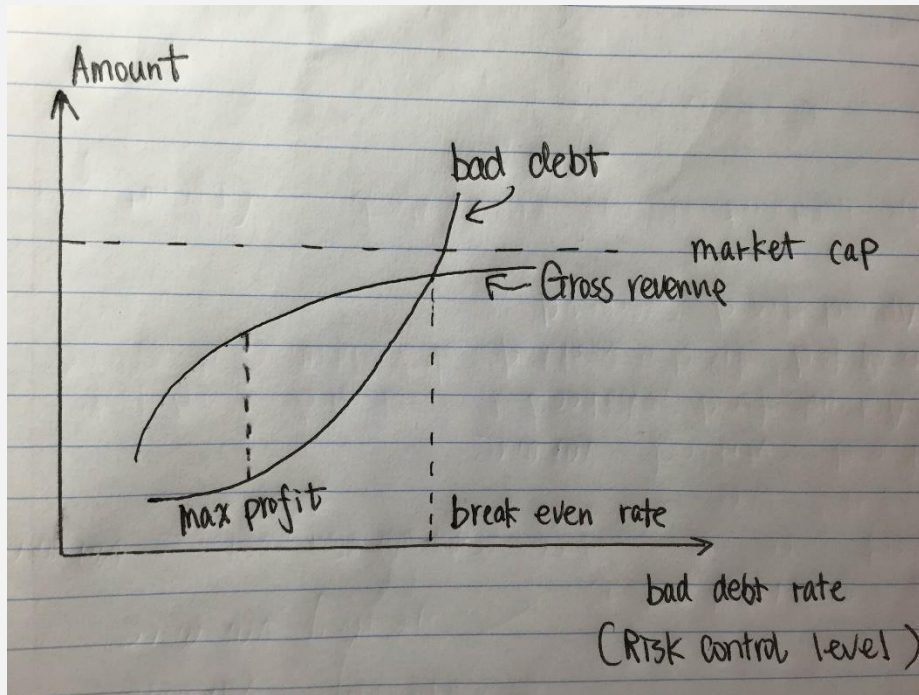
## OTHER TAKEAWAYS

- The customer acquisition should focus on a group of selected attributes(Education, Major, Grades, etc)
- Building Financial KPIs for credit model ( Average loan amount, Average APR, average credit score, late payment, bad debt rate, etc )
- Model need to be adjusted when sufficient loan data get collected

## ADDITIONAL PLAN ON RISK CONTROL

- Collect borrower's family information and Chinese personal identification
- Alumni or Friends who has a credit score referral
- Use social network API to track user's real time update, major events, identify cross region fraud

# RELATIONSHIP BETWEEN RISK LEVEL AND PROFIT



- Max Profit point
- Breakeven point
- Build income projections based on credit model output

**THANK YOU**