The Value of Big Data for Credit Scoring: Enhancing Financial Inclusion using Mobile Phone Data and Social Network Analytics

María Óskarsdóttir, Cristián Bravo, Carlos Sarraute, Jan Vanthienen, Bart Baesens

Applied Soft Computing, 2017

History of Credit Scoring

Before

- limited data – only a few hundred observations
- simple classification techniques

Now

- models still tend to be simple
- newer classification techniques can only offer marginal performance gains
- new data sources

New Domain: Call Detail Records

Call Start Date	Call Start Time	Call Duration (sec)	From Number	To Number
01MAY2017	14:51:14	715	(202) 555-0116	(701) 555-0191
02MAY2017	14:34:37	29	(803) 555-0129	(202) 555-0116
01MAY2017	20:34:14	9	(803) 555-0117	(406) 555-0137
02MAY2017	20:03:38	89	(701) 555-0148	(803) 555-0129

1. What is the added value of including call data for credit scoring?
2. Can call data replace traditional data used for credit scoring?
3. How does default behavior propagate in the call network?

Graphs representation:

- directed and undirected (direction of calls)
- weighted and unweighted (number of calls)

Default propagation techniques:

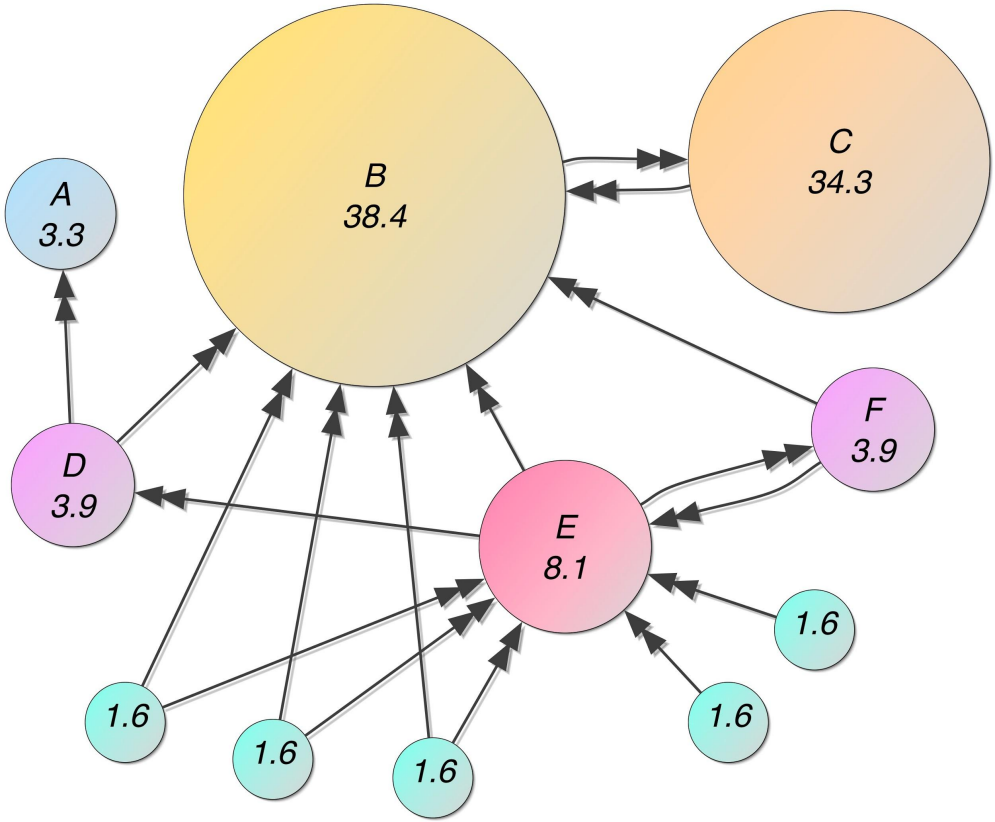
- Personalized PageRank
- Spreading Activation

Link-Based Features:

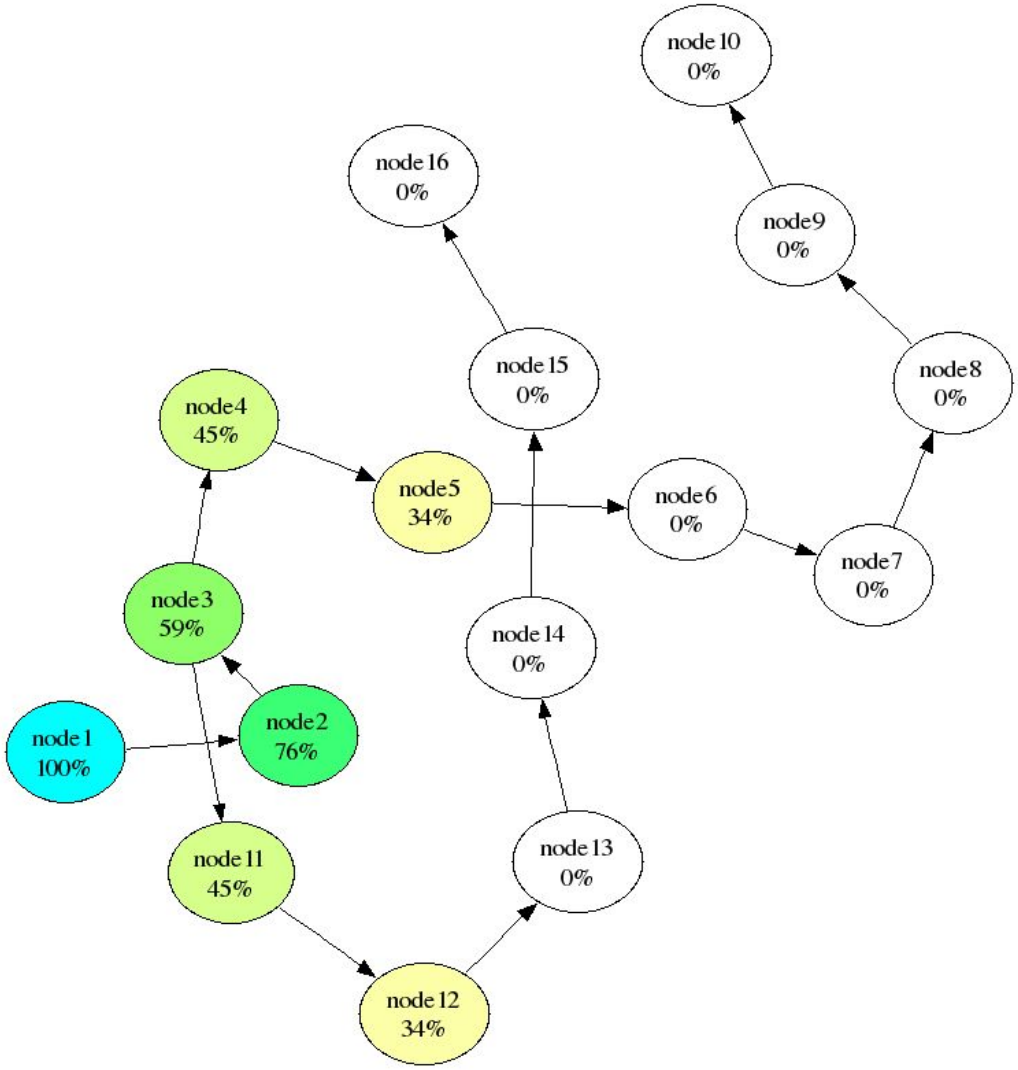
- count-link (the frequency of classes in the neighborhood)
- mode-link (mode of classes in the neighborhood)
- binary-link (binary indicator for each class)

Personalized PageRank

“PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites”



Spreading Activation



Expected Maximum Profit Measure

$$EMP = \int_{b_0} \int_{c_1} P(T(\Theta); b_0, c_1, c^*) \cdot h(b_0, c_1) dc_1 db_0$$

b_0 - benefit of correctly identifying a defaulter
c_1 - cost of incorrectly classifying a non-defaulter as a defaulter
c^* - cost of the action

$\Theta = \frac{c_1 + c^*}{b_0 - c^*}$ - cost-benefit ratio

(fraction of applications that should be rejected to receive maximum profit)

$$\bar{\eta}_{EMP} = \int_{b_0} \int_{c_1} [\pi_0 F_0(T(\Theta)) + \pi_1 F_1(T(\Theta))] \cdot h(b_0, c_1) dc_1 db_0$$

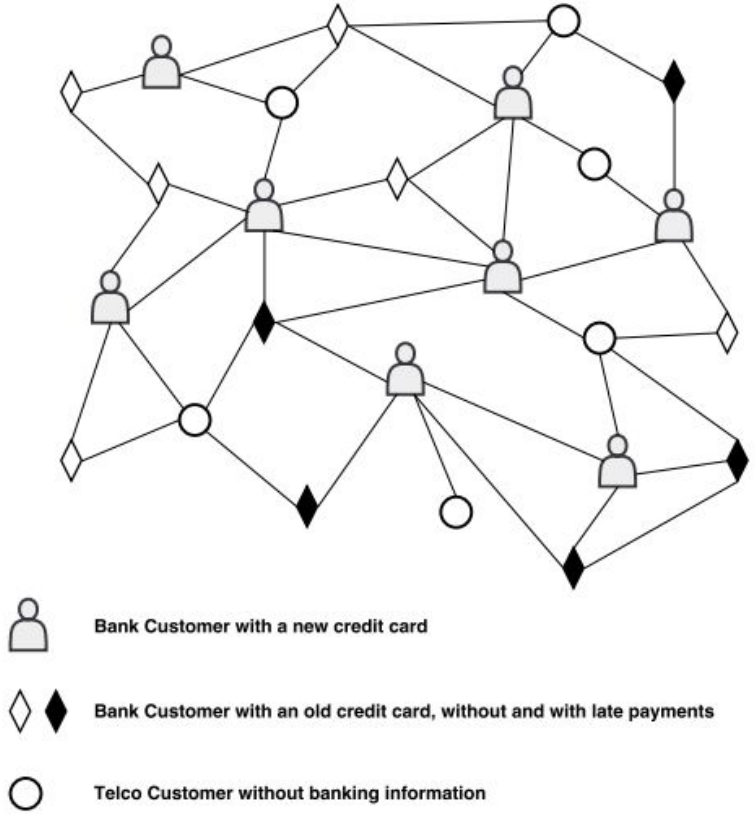
π_0 (π_1) - average classification profit per borrower given the prior probabilities of being a defaulter (nondefaulter)
$F_0(s)$ ($F_1(s)$) - cumulative density functions of defaulters (non-defaulters)

Feature Importance in Terms of Profit

1. Apply the random forest model to the test set and extract class predictions for each tree
2. For each tree compute the profit
3. For each feature in the test set, compute the mean decrease in profit
4. Sort the features with the highest values are those with the greatest mean decrease in profit

Data Description

- telecommunications operator + commercial bank
- 5 consecutive months of CDR data
- 90 million unique cell phone numbers
- bank data about ~2 000 000 customers
- sociodemographic info and debit & credit account activity for the last 3 months



Features: Sociodemographic (SD)

- current age of the customer
- total amount spent in the month before receiving the credit card
- average amount spent per day during the month before receiving the credit card
- diversity of value spent over non-empty bins during the month prior to receiving the credit card
- diversity of number of purchases over all seven bins during the month prior to receiving the credit card

Features: Calling Behavior (CB)

- total number of phone calls received during the three months of the social network
- aggregated duration of all phone calls made on weekends during the three months of the social network
- aggregated duration of all phone calls made and received on Tuesdays during the three months of the social network

Features: Link-based (LB)

- binary indicator of having neighbors with no late payments, in a network with incoming edges
- binary indicator of having neighbors with one late payment, in a network with outgoing edges
- binary indicator of having neighbors with two late payments, in a network with undirected edges
- binary indicator of having neighbors with three late payments, in a network with undirected edges
- also numerical

Features: Personalized PageRank (PR)

- binary indicator of having neighbors with high exposure scores after applying PR on a network with incoming edges and delinquent customers with one or more late payments
- binary indicator of having neighbors with high exposure scores after applying PR on a network with outgoing edges and delinquent customers with two or more late payments

Features: Spreading Activation (SPA)

- exposure score after applying SPA on a network with incoming edges and delinquent customers with one or more late payments
- binary indicator of having neighbors with high exposure scores after applying SPA on a network with incoming edges and delinquent customers with one or more late payments

Statistical Model Performance

Model		Classifier		
Model ID	Feature Groups	Logistic Regression	Decision Trees	Random Forest
A	SD	0.5869	0.7004	0.8993
B	CB	0.5351	0.7043	0.8700
C	LB	0.5485	0.7429	0.7697
D	PR	0.5163	0.7611	0.8339
E	SPA	0.5281	0.7188	0.8063
F	SD,CB	0.6115	0.7127	0.9227
G	CB,LB,PR,SPA	0.5182	0.7307	0.9154
H	SD,CB,LB,PR,SPA	0.6121	0.7263	0.9224

Mean Decrease in Accuracy
0.010
0.005
0.000
Weekday Duration UD
Night Count UD
Count UD
Tuesday Duration UD
Night Duration UD
Tuesday Duration IN
Weekday Duration IN
Night Duration IN
Duration UD
Count Low Exposure (3) UD
Duration IN
Wednesday Duration UD
Count Low Exposure (2) UD
Count (0) UD
Night Count OUT
Count Low Exposure (1) UD
Wednesday Duration IN
Monday Duration OUT
Exposure (2) UD
Count Low Exposure (1) UD
Feature
Group
CB
LB
PR
SPA

Mean Decrease in Profit
0
1
2
3
Night Duration IN
Diverstiy: Number in Non−Empty Bins
Age
Sunday Duration OUT
Exposure (3) UD
Exposure (1) UD
Average Amount Spent per Day
Monday Count IN
Sunday Count UD
Diversity: Value in Non−Empty Bins
Weekday Count IN
Sunday Count IN
Count Link (2) OUT
Sunday Duration UD
Weekend Duration UD
Friday Count IN
Sunday Count OUT
Diversity: Value in All Bins
Duration OUT
Thursday Duration OUT
Feature
Group
CB
LB
PR
SD

Results

1. Homophily amongst defaulters is proved
2. Social network features increase both statistical and profit scores and are amongst the most important features