Credit and systemic risk in P2P lending driven by Big Data Analytics: Current and Future Research

Alessandro Spelta, UNIPV and Tomaso Aste, UCL

FINTECH - HO2020

January 22, 2019

P2P Platforms

- Among FinTech applications that rely on big data analytics, innovative ones are those based on **peer-to-peer** (P2P) financial transactions, such as peer to peer lending, crowdfunding and invoice trading.
- The concept peer-to-peer captures the **interaction** between units, which eliminates the need for a central intermediary
- **Advantages** of P2P Lending Platforms:
  - Improved financial inclusion
  - Higher rates of return compared to bank deposits
  - Lower fees
  - High speed of service
  - Customized user experience

Banks vs P2P Platforms: Risk concerns and data availability

- Both classic banks and P2P platforms rely on **credit scoring models** for estimating credit risk but the incentive for model accuracy is different:
  - Banks assumes the risk so they interested in having the most accurate possible model
  - In a P2P lending platform, the risk is fully borne by the lender
- P2P Platform often do not have access to borrowers' data usually employed by banks
- P2P Platforms operate as **social networks:**
  - Data from such activity can be leveraged for improving credit risk accuracy
  - If we lack P2P direct interaction data we can exploit similarity patters between borrower feature

Similarity patterns, Correlation network and credit scoring models (Giudici and Hadji-Misheva, 2018. Hadji-Misheva, Spelta, 2019)

Aim of Research

- Analyze the **predictive performance** of scoring models employed by P2P platforms.
- Built **similarity networks** from the available platform data.
- Extract **topological information** for describing the relationships between players' local interactions and the global network structure
- Investigate whether **pattern of similarities** between borrowers features can **improve loan default predictions**.

Similarity Network

- In the distance network **nodes** represent borrowing companies and the **edges** the similarities between adjacent nodes.
- There exist different metrics to build up **distances** between objects:
  - Correlation
  - Cosine
- There exist different algorithm to extract a **sparse representation** of the fully connected distance matrix revealing the strongest pattern of similarities
  - Minimum Spanning Tree
  - Maximum Planar Graph
  - F-test

Similarity Networks

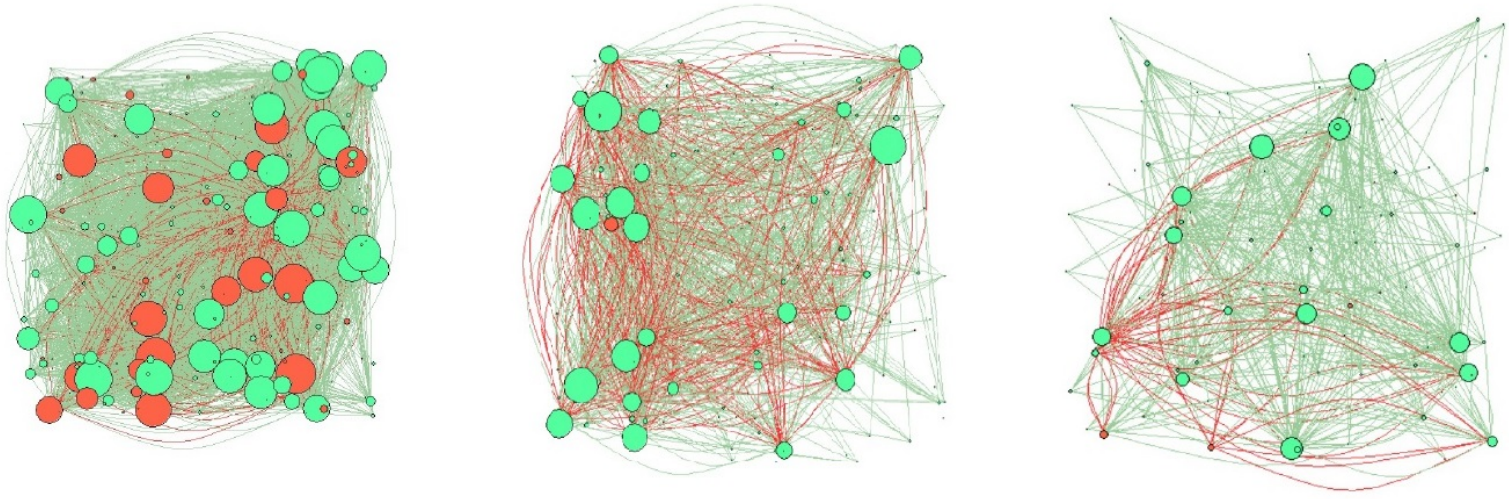
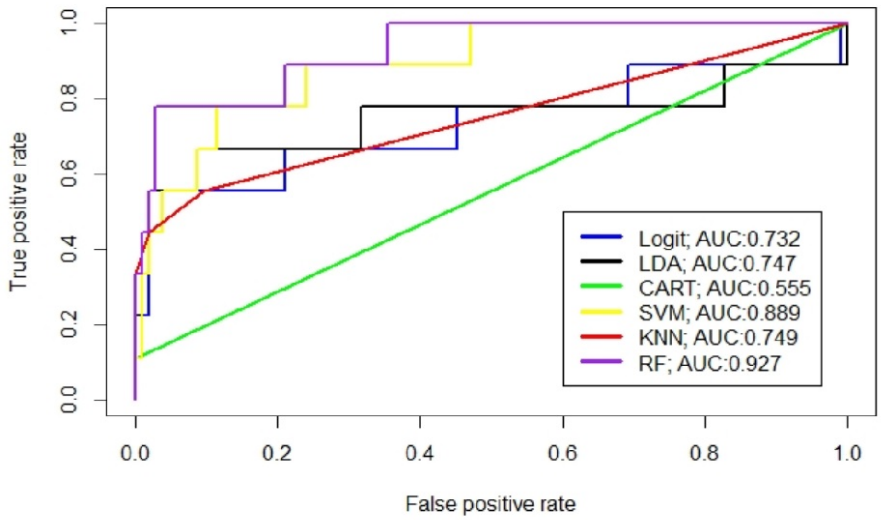


Figure 1: Left: Correlation network based on the activity indicator. Number of nodes= 386; Middle: Correlation network based on the solvency indicator. Number of nodes= 288; Right: Correlation network based on the return on equity ratio. Number of nodes= 226

Topological Coefficients

- We extract various type of information from such networks
- **Nodes Importance**
  - How many partners a Company has (degree)?
  - How strong are the weights embedded in such connections (strength)?
  - How crucial a node is in letting information to spreads over the network (betweennes)
- **Community Structure**
  - Dense sub-graphs sharing some common characteristics
  - For each node we have the identifier of the community the node belongs
- Such information are used to complement balance sheet information of each Company.
- We compare different credit scoring models (logistic, knn, svm, random forest...) using non parametric measures (roc, accuracy, precision).

ROC comparison



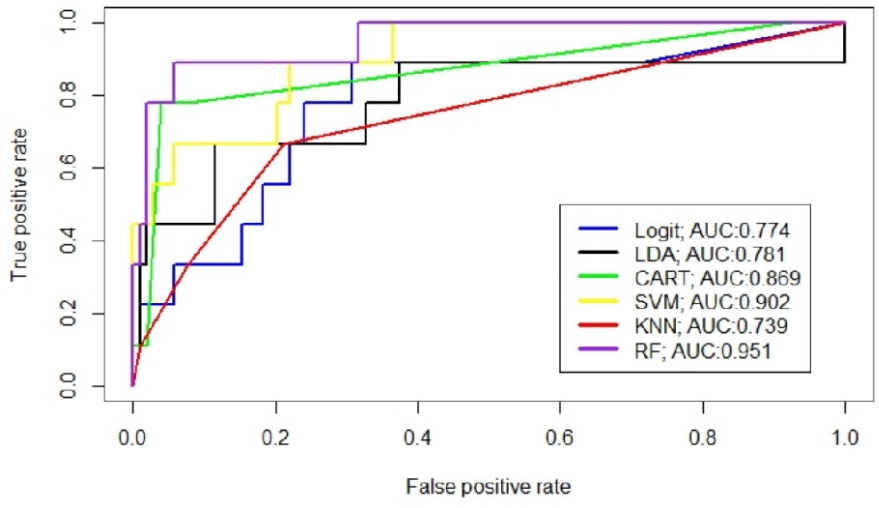


Figure 2: Comparison predictive utility of models with and without network information

Predictive Modeling

Company	PD from BLR	PD from TNBS	PD from WNBS	Status
Company A	0.325995937	0.203178979	0.102347865	Active
Company B	0.207411016	0.220493154	0.121339165	Active
Company C	0.198157788	0.101808476	0.047102588	Active
Company D	0.107315436	0.103854312	0.081768604	Active
Company E	0.006017395	0.000907596	0.000364361	Active
Company F	0.127879968	0.248898102	0.373049367	Default
Company G	0.002658514	0.125033683	0.149287131	Default
Company H	0.045663684	0.177499378	0.510471885	Default
Company I	0.000419074	0.040453597	0.06446989	Default
Company J	0.016839456	0.07174686	0.091508018	Default

Table 1: Comparison of PD Estimates across different models. BLR indicated the baseline regression model; TNBS the network based model, with all types; WNBS the network based model, with only Type C edges.

Deep Learning based credit scoring models (Aste and Turiel, 2018)

Aim of Research

- Automation of loan screening and acceptance through Machine Learning.
- Accurate prediction of default risk through Big Data analytics and Machine Learning techniques.
- P2P lending data investigated to understand limitations of default predictability.

Time Series Features

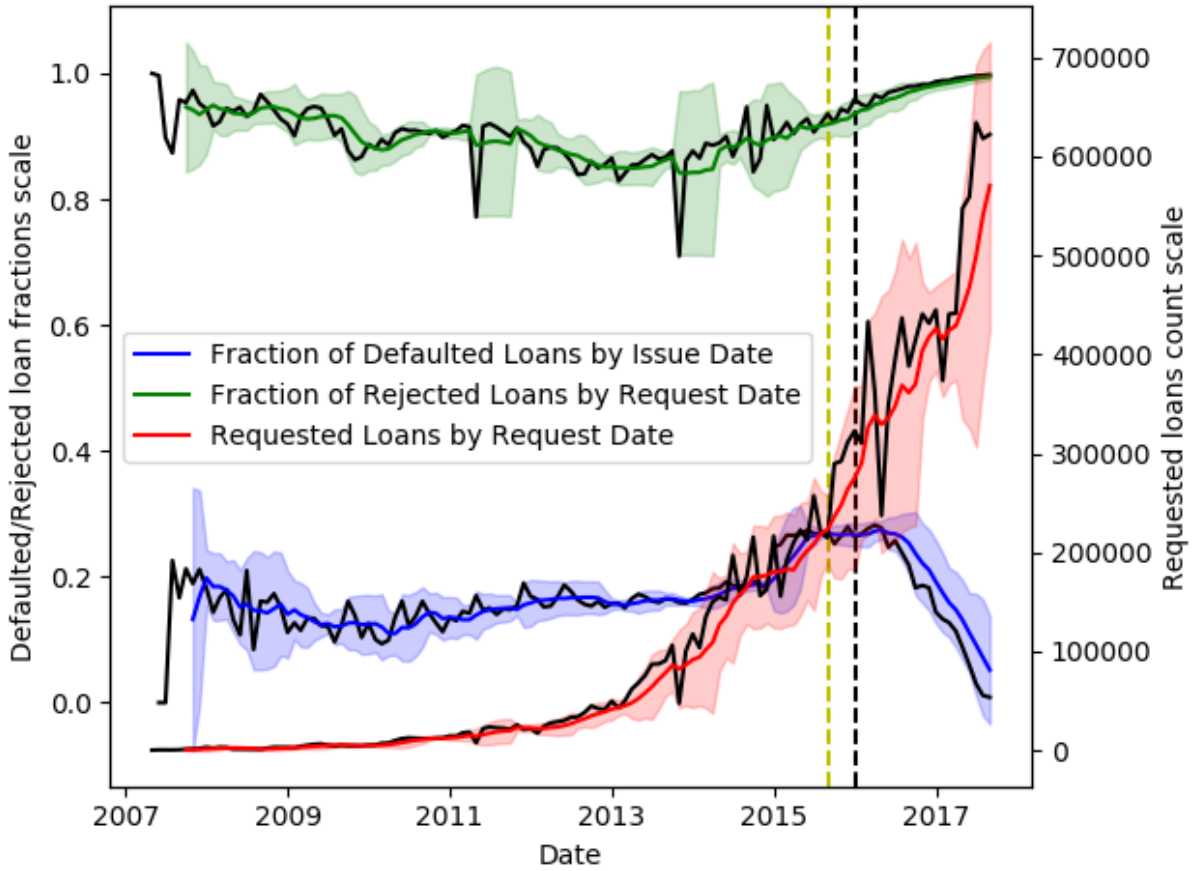


Figure 3: Time series plots for the dataset. Three plots are presented: the number of defaulted loans as a fraction of the total number of accepted loans (blue), the number of rejected loans as a fraction of the total number of loans requested (green) and the total number of requested loans (red).

Method: Two-Layer Model

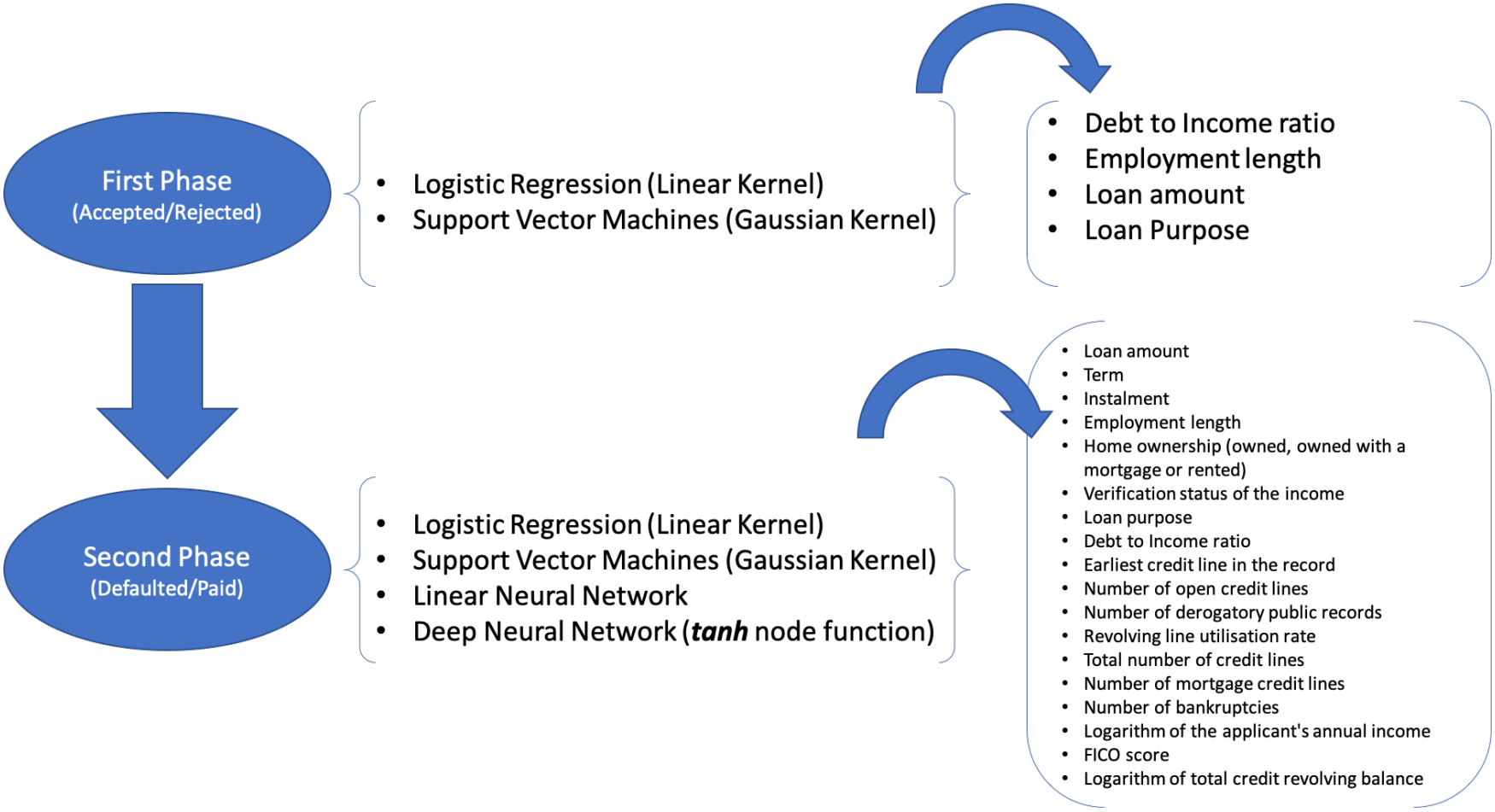


Figure 4: Representative diagram outlining the two phases of the model with machine learning methods applied and features considered for each phase

Loan Selection

Loan Selection Results					
Model	Recall Train	AUC Test	Recall Macro Test	Recall Accepted Test	Recall Rejected Test
LR	79.8%	86.5%	77.4%	69.1%	85.7%
SVM	77.5%	-	75.2%	66.5%	84.0%

Table 2: Results for the ML algorithms applied to the 1^{st} model phase.

- Simple Logistic Regression model replicates analyst rejections with recall above 85%.
- Target feature class imbalance in training set affects class scores. Would benefit from more training data.
- Replicability of screening leads to more complex models applied to default prediction.

Loan Default Prediction

Loan Default Prediction Results					
Model	Recall Train	AUC Test	Recall Macro Test	Recall Default Test	Recall Paid Test
LR	64.3%	69.0%	63.7%	63.8%	63.6%
SVM	-	64.3%	62.15%	58.7%	65.6%
LNN [a]	-	67.8%	-	60.0%	-
LNN [b]	-	67.8%	-	60.0%	-
LNN [c]	-	69%	-	65%	-
DNN [d]	-	68%	-	67%	-
DNN [e]	71%	66%	-	75%	-
DNN [f]	68%	69%	-	72%	-

[a] LNN with numerical features only
[b] LNN with numerical and categorical features
[c] LNN with numerical and categorical features, L2 regularised
[d] DNN with arbitrary node numbers [20, 5]
[e] DNN with node numbers fine-tuned to [30, 1]
[f] DNN with node numbers fine-tuned to [5, 3]

Table 3: Results for the ML algorithms applied to the 2^{nd} model phase.

- Increasing complexity of the model captures complex phenomenon of default, with NN outperforming in recall.

Future Discussion

Future Discussions

- Leveraging transactional data
- Levering social behavior patterns
- Quality of Data