2016 International Conference on Computational Social Science
June 23-26, 2016 IC^2S^2

Forecasting Financial Well-Being for Small and Medium Enterprises Using Network-Based Signals.

Julius Adebayo, Yoshihiko Suhara, Vivek Singh, Burcin Bozkaya, and Alex (Sandy) Pentland
{juliuad, suhara, singhv, pentland}@mit.edu,{bbozkaya}@sabanciuniv.edu

Keywords: networks, business, credit-worthiness, prediction, financial

Abstract: We leverage machine learning methods to forecast the short-term financial well-being of Small and Medium Enterprises (SMEs) using millions of credit card transactions. Our approach incorporates features derived from a city-level network of all merchants. We create a network of merchants from customer co-purchases derived from aggregated credit card transactions. We combine network-based features with features derived from the socio-demographic characteristics of an SME's customer base, and financial indicators into a comprehensive predictive model. For 1200 retail stores in a large metropolitan region, our comprehensive model shows a **36%** improvement over a baseline model consisting of traditional financial metrics. To the best of our knowledge, our study is the first to leverage a network-based approach in characterizing SME financial well-being.

Introduction and Problem Overview

SMEs account for a significant portion of the economic activity in most countries. As of 2012, there are about 20 million SMEs in Europe responsible for about 67 percent of jobs [1]. From a bank or lending agency's perspective, an accurate model for forecasting future SME 'financial well-being' is crucial for effective capital allocation. Models that can accurately identify top performing SMEs will improve decision-making, and help reduce financial loss.

Currently, traditional approaches for predicting an SME's well-being, defined as the likelihood of defaulting on a loan, rely on financial metrics like earnings per asset and equity per asset. However, these metrics are often difficult to obtain, and tend to have limited predictive performance [2]. We address these difficulties in two ways. First, we present a new definition of 'financial well-being': the change in future revenue of an SME. Second, in deriving our complete feature set from financial transactions data, we ensure that banks can easily incorporate the approach presented here into their models.

The use of network-based signals here derives from the fact that networks, in various forms, have been shown to form the backbone of economic life [3]. The complexity of a nation's product space embeds information about the potential for improvement in GDP [4]. The diversity of a community's social network is closely correlated with its economic wealth [3]. Recently, the centrality of authors in a citation network has been shown to predict scientific success [4]. In these examples, one observes that nodal structural properties seem to encode the capacity for improvement. Given this insight, we consider how a merchant's structural characteristics in a city-level network might predict its future 'financial well-being.'

Methodology & Approach

Our dataset consists of anonymized credit card transactions. The data spans a 3-month period in 2013 with approximately 15.7 million transactions. We consider 1200 retail stores in the country's largest metropolitan area. We derive signals from the first 2 months to predict change

in merchant revenue in the third month. For each merchant, we calculate the change in revenue in the third month given the average revenue over the first two months. Now, we define the target variable as a binary variable as follows: 1 for merchants with above median values and 0 otherwise. From a lender's perspective, the target variable specified drives decision-making since it provides insight into the future growth in revenue of an SME.

Our setup corresponds to a supervised learning problem, particularly, classification. For all the analysis presented, we leverage a Random Forest classifier, and report 10-fold nested cross validation scores per model. Our measure of model performance is the AUROC; area under the receiver-operating curve, which quantifies the overall performance of a binary classifier.

Results & Discussion

The baseline financial model includes features derived from three signals: transaction volume, customer visit volume, and previous revenue over the first two months. A 10-fold cross-validated classifier built on these features resulted in an AUROC of 0.58. This suggests that the baseline features possess minimal signal with which a model can distinguish between merchants with a high likelihood of future growth and those without.

For customers in the dataset, we have information regarding age, income, asset, gender, education level, and marital status. We aggregate these features for each merchant into a 47-feature model consisting of the minimum, maximum, median, and standard deviation of each characteristic. A classifier built on these features resulted in an AUROC of 0.53.

Using the database of financial transactions, we create a network of merchants representing the flow of individuals across merchants. Each merchant corresponds to a node in the network. We draw an edge between two nodes if a customer makes a purchase from these two merchants during the first two months. Consequently, we obtain a weighted undirected network among merchants where the weights correspond to the total number of individuals that patronize two merchants in the two months considered. The network obtained captures city-level flow of economic activity between these merchants.

Given the merchant network, we then compute structural properties of the nodes to use as features in a classifier. For example, one notion of importance, eigenvector centrality, in this network roughly encodes the probability that a random customer will visit a particular merchant. Other notions of centrality on this network capture how far away a merchant is, on average, from every other merchant. For the metropolitan city considered, we obtained a network of approximately 50 thousand nodes and 900 thousand edges.

We also capture features characterizing a merchant's ego network, and representative cluster, derived as in [6]. For example, one important feature in our analysis was the diversity, in category, of a merchant's ego network. Our measure of diversity quantifies how homogeneous a merchant's ego network is, derived from a similar measure proposed in [7]. Table A shows the top 5 ranked features in a model built on network-based features. A 10-fold cross-validated classifier built on these features resulted in a performance of 0.71 AUROC, which is a **22% improvement** on the baseline model. A model combining the entire top ranked features results in a 0.79 AUC, which represents a **36% improvement** over the baseline model.

A

Feature Rank	Baseline Financial Metrics	Socio-Demographic Features	Network Features
1	Average Revenue	Age	Eigenvector Centrality
2	Change in Revenue	Education Level	Ego Diversity
3	Average Number of Customers	Income and Asset	Network Cluster Identity
4	Change in number of Transactions	Gender	Degree Centrality
5	Total Number of Transactions	Marital Status	Weighted Neighbor Eigenvector Centrality

B

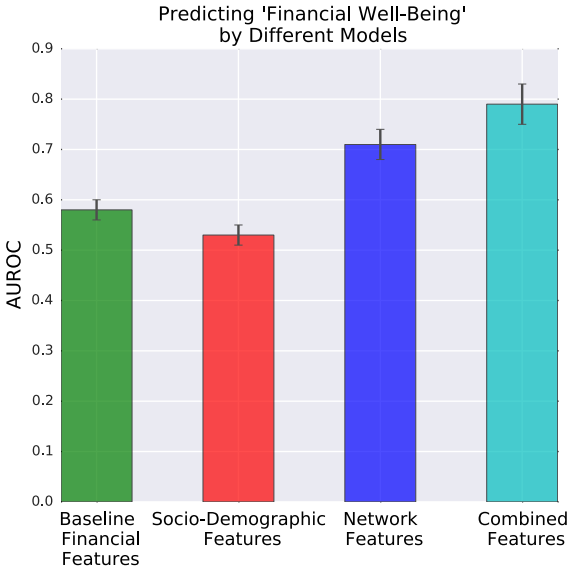

Table A shows a list of the top ranked features for each category. **Figure B:** shows AUROC for models built using different feature groups including a combined model with all top ranked features. The combined model provides a 36% improvement (2-sided test, t(18) = 28.5, p=4.95e-16), while the network model provides a 22% improvement (2-sided test, t(18) = 19.5, p=1.45e-13) over the baseline model.

Conclusion & Future Work

Here, we have shown that features derived from a city-level merchant network, that captures the flow of consumers, improves the ability to forecast future financial well-being for these merchants. Our results suggest that incorporating network-based signals, and socio-demographic information into current models will improve lending decisions. Going forward, work is underway to expand the model presented here to a larger set of merchants across several categories. Further, we expect to verify our results for different definitions of financial well-being to understand the robustness of the results presented. Beyond this, we also expect to undertake careful regression analysis, and significance testing of all features used in our models to ascertain the strength of each signal for financial well-being prediction.

Citations

1) Gagliardi, D., "A recovery on The horizon." Annual report on European SMEs 2013 (2012).
2) Fantazzini, Dean et. al. "Random survival forests models for SME credit risk measurement." Methodology and Computing in Applied Probability 11.1 (2009): 29-45.
3) Eagle, Nathan et. al. "Network diversity and economic development." Science 328.5981 (2010): 1029-1031.
4) Hidalgo, César. "The dynamics of economic complexity and the product space over a 42 year period." CID Working Paper189 (2009).
5) Sarigöl, Emre,"Predicting scientific success based on coauthorship networks." EPJ Data Science 3.1 (2014): 1-16.
6) Adebayo, Julius, et al. "An exploration of social identity: The structure of the BBC news-sharing community on Twitter." Complexity 19.5 (2014): 55-63.
7) Singh, Vivek Kumar, Burcin Bozkaya, and Alex Pentland. "Money Walks: Implicit Mobility Behavior and Financial Well-Being." PloS one 10.8 (2015): e0136628.