P2P LOAN ACCEPTANCE AND DEFAULT PREDICTION WITH ARTIFICIAL INTELLIGENCE

A PREPRINT

Jeremy D. Turiel
Department of Computer Science
University College London
Gower St, Bloomsbury
London WC1E 6BT, United Kingdom
jeremy.turiel.18@ucl.ac.uk

Tomaso Aste *
Department of Computer Science
University College London &
Gower St, Bloomsbury
London WC1E 6BT, United Kingdom
t.aste@ucl.ac.uk

July 4, 2019

ABSTRACT

Logistic Regression and Support Vector Machine algorithms, together with Linear and Non-Linear Deep Neural Networks, are applied to lending data in order to replicate lender acceptance of loans and predict the likelihood of default of issued loans. A two phase model is proposed; the first phase predicts loan rejection, while the second one predicts default risk for approved loans. Logistic Regression was found to be the best performer for the first phase, with test set recall macro score of 77.4%. Deep Neural Networks were applied to the second phase only, were they achieved best performance, with validation set recall score of 72%, for defaults. This shows that AI can improve current credit risk models reducing the default risk of issued loans by as much as 70%. The models were also applied to loans taken for small businesses alone. The first phase of the model performs significantly better when trained on the whole dataset. Instead, the second phase performs significantly better when trained on the small business subset. This suggests a potential discrepancy between how these loans are screened and how they should be analysed in terms of default prediction.

***K*eywords** P2P lending · Artificial Intelligence · Big Data · Default risk · Financial automation

1 Introduction

Accurate prediction of default risk in lending has been a crucial theme for banks and other lenders for over a century. Modern days availability of large datasets and open source data, together with advances in computational and algorithmic data analytics techniques, have renewed interest in this risk prediction task. Furthermore, automation of the loan approval process opens new financing opportunities for small businesses and individuals. These previously suffered from more limited access to credit, due to the high cost of human processing. Ultimately, automation of this process carries the potential to reduce human bias and corruption, making access to credit fairer for all. Financial technologies are having a strong impact on this domain, which is rapidly changing [1]. The application of this model to P2P lending is just one example, with others being micro-financing in developing countries and loan-by-loan evaluation of loan portfolios for investment.

P2P lending has attracted the attention of industry, academics and the general public in recent years. This is also due to the large expansion of major P2P lending platforms like the Lending Club, which has now lent over $45bn to more than 3mln customers. Another reason for the increasing coverage and popularity of P2P lending is its fast expansion to less developed markets in Eastern Europe, South America and Africa. As the monetary and social relevance of the industry grows, the need for regulation arises. The FCA is among the regulars which have set rules for this industry [2, 3], indicating the importance of the trend in developed countries other than the United States.

*Head of the Financial Computing and Analytics Group http://www.cs.ucl.ac.uk/staff/tomaso_aste/. Director, UCL Centre for Blockchain Technologies http://blockchain.cs.ucl.ac.uk/tomaso-aste/.

Thanks to its easily accessible historical datasets, the Lending Club is the subject of multiple publications investigating the drivers of default in P2P lending [4, 5]. The growth of P2P lending in emerging countries has also attracted research interest, for instance [6] investigates lending in Mexico. This highlights the crucial role of P2P lending in providing access to credit for the population of emerging countries. Interdisciplinary scientific communities such as that of network science have also started to show interest in the socioeconomic dynamics of P2P lending [7]. More theoretical works have also inquired about the reason for the need and growth of P2P lending. This was often connected to the concept of credit rationing due to asymmetric information between lending counterparts [8]. A solution to the problem of credit rationing, focused towards allowing fair access to credit and reducing poverty, are micro-finance institutions. Chris Anderson, Editor in Chief of $Wired$ magazine, already identified the concept of "selling less of more", which is now making its way through to the lending market [9]. In order to reduce frictions and allow MFIs to have a self-sustainable business model Serrano-Cinca et al. already suggested that technology will allow to reduce costs and interest rates, leading to an e-commerce like revolution [10]. This work aims to contribute to this goal.

To the best of our knowledge, academic publications investigating the drivers of P2P lending [4, 5, 6] have applied simple regression models to this task. This work constitutes a significant step forward to applying Big Data and Artificial Intelligence techniques to P2P lending, combining two major disruptive emerging fields.

The rest of the paper is organised as follows: in Section 2 we describe the dataset used for the analysis and the methods applied, in Section 3 we present results and related discussion for the first (Section 3.1.1) and second phase (Section 3.1.2) of the model applied to the entire dataset, Section 3.2 then investigates similar methods applied in the context of "small business" loans, Section 4 draws conclusion from our work, followed by acknowledgments in Section 5.

2 Dataset and Methods

2.1 Dataset

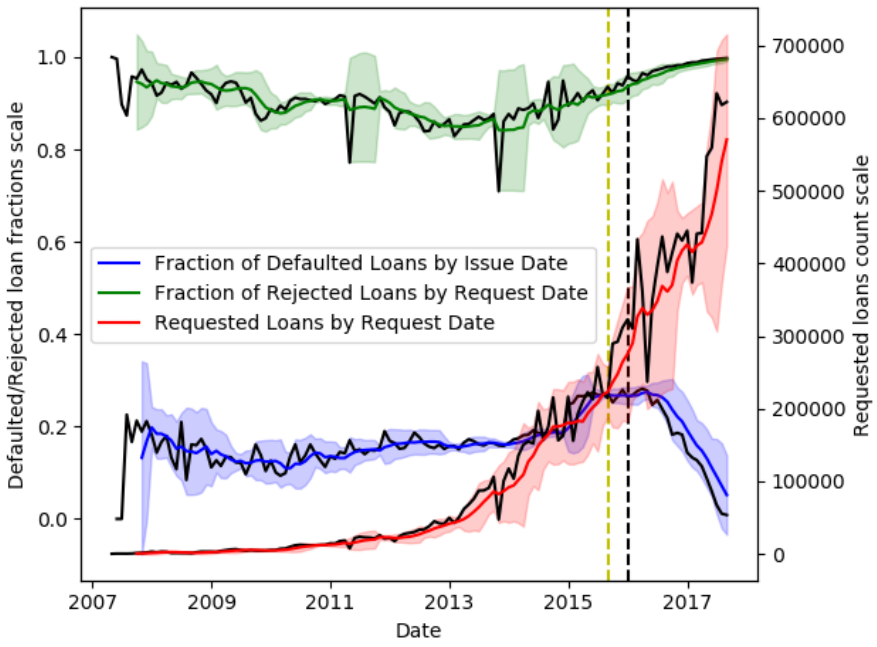

Figure 1: Time series plots of the dataset [11]. Three plots are presented: the number of defaulted loans as a fraction of the total number of accepted loans (blue), the number of rejected loans as a fraction of the total number of loans requested (green) and the total number of requested loans (red). The black lines represent the raw time series, with statistics (fractions and total number) computed per calendar month. The coloured lines represent six-month moving averages and the shaded areas of the corresponding colours represent the standard deviation of the averaged data. The data on the right of the vertical black dotted line was excluded due to the clear decrease in the fraction of defaulted loans, this was argued to be due to the fact that defaults are a stochastic cumulative process and that, with loans of 36-60 months term, most loans issued in that period did not have the time to default yet. A larger fraction of loans is, instead, repaid early. This would have constituted a biased test set.

The data was collected from loans evaluated by Lending Club in the period between 2007 and 2017 (`https://www.lendingclub.com`). The dataset was downloaded from Kaggle (`www.kaggle.com`).

In this paper, we present the analysis of two rich open source datasets [11] reporting loans including credit card-related loans, weddings, house-related loans, loans taken on behalf of small businesses and others. One dataset contains loans that have been rejected by credit analysts, whilst the other, which includes a significantly higher number of features, represents loans which have been accepted and indicates their current status. Our analysis concerns both. The first dataset comprises over 16 million rejected loans, but has only 9 features. The second dataset comprises over 1.6 million loans and it originally contained 150 features. We cleaned the datasets and combined them into a unique dataset containing ≈ 15 million loans, including $\approx 800,000$ accepted loans. Almost $800,000$ accepted loans labelled as "current" were removed from the dataset, since no default or payment outcome was available. The dataset of accepted loans indicates the status of each loan. Loans which had a status of fully paid (over $600,000$ loans) or defaulted (over $150,000$ loans) were selected for the analysis and this feature was used as target label for default prediction. The fraction of issued to rejected loans is $\simeq 10\%$, with the fraction of issued loans analysed constituting only $\approx 50\%$ of the overall issued loans. Defaulted loans represent $15 - 20\%$ of the issued loans analysed.

In the present work, features for the first phase were reduced to those shared between the two datasets. For instance, geographical features (U.S. state and postcode) for the loan applicant were excluded, even if they are likely to be informative. Features for the first phase are: 1) debt to Income ratio (of the applicant); 2) employment length (of the applicant); 3) loan amount (of the loan currently requested); 4) purpose for which the loan is taken. In order to simulate realistic results for the validation set, the data was sectioned according to the date associated with the loan. Most recent loans were used as validation set, while earlier loans were used to train the model. This simulates the human process of learning by experience. In order to obtain a common feature for the date of both accepted and rejected loans the issue date (for accepted loans) and the application date (for rejected loans) were assimilated into one date feature. This time-labelling approximation, which is allowed as time sections are only introduced to refine model testing, does not apply to the second phase of the model where all dates correspond to the issue date. All numeric features for both phases were scaled by removing the mean and scaling to unit variance. The scaler is trained on the training set alone and applied to both training and test sets, hence no information about the test set is contained in the scaler which could be leaked to the model.

Features considered for the second phase of the model are: 1) loan amount (of the loan currently requested); 2) term (of the loan currently requested); 3) instalment (of the loan currently requested); 4) employment length (of the applicant); 5) home ownership (of the applicant. Rented, owned or owned with a mortgage on the property); 6) verification status of the income or income source (of the applicant. If this was verified by the Lending Club); 7) purpose for which the loan is taken; 8) Debt to Income ratio (of the applicant); 9) earliest credit line in the record (of the applicant); 10) number of open credit lines (in applicant's credit file); 11) number of derogatory public records (of the applicant); 12) revolving line utilisation rate (the amount of credit the borrower is using relative to all available revolving credit); 13) total number of credit lines (in applicant's credit file); 14) number of mortgage credit lines (in applicant's credit file); 15) number of bankruptcies (in the applicant's public record); 16) logarithm of the applicant's annual income (the logarithm was taken for scaling purposes); 17) FICO score (of the applicant); 18) logarithm of total credit revolving balance (of the applicant).

We first analysed the dataset [11] feature by feature to check for distributions and relevant data imbalances. Features providing information for a restricted part of the dataset (less than 70%) were excluded and the missing data was filled by mean imputation. This should not relevantly affect our analysis as the cumulative mean imputation is below 10% of the overall feature data. Furthermore, statistics were calculated for samples of at least 10,000 loans each, so the imputation should not bias the results. A time series representation of statistics on the dataset is shown in Figure 1.

Differently from other analyses of this dataset (or of earlier versions of it, such as [12]), here for the analysis of defaults we use only features which are known to the lending institution prior to evaluating the loan and issuing it. For instance, some features which were found to be very relevant in other works [12] were excluded for this choice of field. Amongst the most relevant features not being considered here are interest rate and the grade assigned by the analysts of the Lending Club. Indeed, our study aims at finding features which would be relevant in default prediction and loan rejection a priori, for lending institutions. The scoring provided by a credit analyst as well as the interest rate offered by the Lending Club would not, hence, be relevant parameters in our analysis.

2.2 Methods

Two machine learning algorithms were applied to both datasets presented in Section 2.1: logistic regression with underlying linear kernel and Support Vector Machines (see [13, 14] for general references on these methodologies).

Neural Networks were also applied, but to default prediction only. Neural Networks were applied in the form of a linear classifier (analogous, at least in principle, to logistic regression) and a deep (two hidden layers) neural network [15].

Regularisation techniques were applied to avoid overfitting, L2 regularisation was the most frequently applied, but also L1 regularisation was included in the grid search for LR and SVMs. These were included as mutually exclusive, hence not in the form of an elastic net [16, 17]. Initial hyperparameter tuning for the model was performed through extensive grid searches. The ranges for the regularisation parameter α varied, but the widest range was $\alpha = [10^{-5}, 10^{5}]$. Values of α were all powers of 10 with integer exponents. Hyperparameters were determined by the grid search and were manually tuned only in some cases specified in Section 3. This was done by shifting the parameter range in the grid search or by setting a specific value for the hyperparameter. This was mostly done when there was evidence of overfitting from training and validation set results from the grid search. Class imbalance was mitigated through regularisation as well as by balancing the weights at the time of training of the model itself. Manual hyperparameter tuning was applied as a consequence of empirical evaluations of the model. Indeed, model evaluations through different measures often suggest that a higher or lower level of regularisation may be optimal, this was then manually incorporated by fixing regularisation parameters or reducing the grid search range. Intuition of the authors about the optimisation task was also applied to prioritise maximisation of a performance measure or balance between different performance measures. Training and validation (or test) sets were used in the analysis. The dataset was split at the beginning in order to prevent information leakage, which might provide the model with information about the test set. The test set then contains future unseen data.

Two metrics were used for result validation, namely recall and AUC. AUC can be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [18]. This is very relevant to the analysis as credit risk and credit ranking are assessed in relation to other loans as well. The metric extrapolates whether defaulting loans are assigned a higher risk than fully paid loans, on average. Recall is the fraction of loans of a class (such as defaulted or fully paid loans) which are correctly predicted. The standard threshold of 50% probability, for rounding up or down to one of the binary classes, was applied. This is relevant as it does not test the relative risk assigned to the loans, but the overall risk and the model's confidence in the prediction [19].

3 Results and Discussion

3.1 General two phases model for all purpose classes prediction

3.1.1 First Phase

Logistic regression was applied to the combined datasets. The grid search over hyperparameter values was optimised to maximise the unweighted recall average. The unweighted recall average is referred to as recall macro and is calculated as the average of the recall scores of all classes in the target label. The average is not weighted by the number of counts corresponding to different classes in the target label. We maximise recall macro in the grid search as maximising AUC led to overfitting the rejected class, which bares most of the weight in the dataset. This is due to AUC weighting accuracy as an average over predictions. This gives more weight to classes which are overrepresented in the training set, a bias that can lead to overfitting.

In order to obtain a more complete and representative validation set, the split between training and validation sets was $75\%/25\%$ for the first phase of the model (differently from the $90\%/10\%$ split applied in Section 3.1.2). This provides 25% of the data for testing, corresponding to approximately two years of data. This indeed constitutes a more complete sample for testing and was observed to yield and more stable reliable results.

The grid search returned an optimal model with $\alpha \simeq 10^{-3}$. The recall macro score for the training set was $\simeq 79.8\%$. Test set predictions instead returned a recall macro score $\simeq 77.4\%$ and an AUC score $\simeq 86.5\%$. Test recall scores were $\simeq 85.7\%$ for rejected loans and $\simeq 69.1\%$ for accepted loans.

The same dataset and target label were analysed with Support Vector Machines. Analogously to the grid search for logistic regression, recall macro was maximised. A grid search was applied to tune α. Training recall macro was $\simeq 77.5\%$ while test recall macro was $\simeq 75.2\%$. Individual test recall scores were $\simeq 84.0\%$ for rejected loans and $\simeq 66.5\%$ for accepted ones. Test scores did not vary much, for the feasible range of $\alpha = [10^{-3}, 10^{-5}]$.

In both regressions, recall scores for accepted loans are lower by $\approx 15\%$, this is probably due to class imbalance (there is more data for rejected loans). This suggests that more training data would improve this score. From the above results, we observe that a class imbalance of almost 20x affects the model's performance on the underrepresented class. This phenomenon is not particularly worrying in our analysis, though, as the cost of lending to an unworthy borrower is much higher than that of not lending to a worthy one. Still, about 70% of borrowers classified by the Lending Club as worthy, obtain their loans.

The results for SVMs suggest that polynomial feature engineering would not improve results in this particular analysis. The surprisingly accurate results for logistic regression suggest that credit analysts might be evaluating the data in the features with a linear-like function. This would explain the improvements shown by the second phase, when just a simple model was used for credit screening.

3.1.2 Second Phase

Logistic Regression, Support Vector Machines and Neural Networks were applied to the dataset of accepted loans in order to predict defaults. This is, at least in principle, a much more complex prediction task as more features are involved and the intrinsic nature of the event (default or not) is both probabilistic and stochastic.

Categorical features are also present in this analysis. These were "hot encoded" for the first two models, but were excluded from the neural network in this work as the number of columns resulting from the encoding greatly increased training time for the model. We shall investigate neural network models with these categorical features included, in future works.

For the second phase, the periods highlighted in Figure 1 were used to split the dataset into training and validation sets (with the last period excluded as per the figure caption). The split for the second phase was of $90\%/10\%$, as more data improves stability of complex models and balanced classes had to be obtained through downsampling for the training set (downsampling was applied as oversampling was observed to cause the model to overfit).

In this phase, the overrepresented class in the dataset (fully paid loans) benefitted from the higher quantity of training data, at least in terms of recall score. In this case the overrepresented class is that of fully paid loans while, as discussed in Section 3.1.1, we are more concerned with predicting defaulting loans well rather than with misclassifying a fully paid loan.

3.1.3 Second Phase - Logistic Regression

The grid search for logistic regression returned an optimal model with a value of $\alpha \simeq 10^{-2}$. The grid was set to maximise recall macro, as for the models in Section 3.1.1. Training recall macro score was $\simeq 64.3\%$ and test AUC and recall macro scores were 69.0% and 63.7%, respectively. Individual test recall scores were 63.8% for defaults and 63.6% for fully paid loans, see Table 1. Maximising recall macro indeed yields surprisingly balanced recall scores for the two classes. Maximising AUC did not lead to strong overfitting, differently from what is discussed in Section 3.1.1. Test scores were lower, both in terms of AUC and recall macro.

3.1.4 Second Phase - Support Vector Machine

Support Vector Machines were also applied to the dataset. The optimal value of α returned by the grid search was $\alpha = 10^{-2}$, the same as for logistic regression in Section 3.1.3. Scores for the model were, though, worse than those returned by logistic regression. Test AUC was $\simeq 64.3\%$ and individual test recall scores were 58.7% for defaulted loans and 65.6% for fully paid loans, see Table 1. It can be inferred that the analysis of this dataset does not benefit from SVM kernel's non-linearities in its test set performance. Furthermore, recall scores are improved for the overrepresented class in the dataset. This is the opposite of what is aimed for in this analysis, where we prioritise high recall on the default class which has a higher impact on the borrower's balance sheet. Such a strong score imbalance is also not ideal in terms of quality of the predictor. It should be noted that the label class imbalance (defaulted and fully paid loans) is much weaker than that described in Section 3.1.1, with defaulted loans representing $15 - 20\%$ of the dataset.

Table 1: Table with main results from LR and SVM tested for the second phase of the model.

Loan Default Prediction Results				
Model	α	Recall Train	AUC Test	Recall Test (Macro/Default/Paid)
LR	10^{-2}	64.3%	69%	$63.7\%/63.8\%/63.6\%$
SVM	10^{-2}	–	64.3%	$62.2\%/58.7\%/65.6\%$

3.1.5 Second Phase - Neural Network

Linear Neural Network classifiers as well as Deep (two hidden layers) Neural Networks were also trained on the dataset for the second phase of the model. Linear Neural Network classifiers were trained on numerical features alone as well as on both numerical and categorical features. L2 regularisation was then applied. Numerical features-only test scores

returned an AUC of 67.8% and a recall of 60.0% (for defaulted loans). The model yielded improved results when trained on categorical features too. Test scores returned an AUC of 68.7% and recall of 62.7% (for defaulted loans). These scores are slightly worse than those for logistic regression, but they do not implement regularisation yet. Once L2 regularisation ($\alpha = 10$) was manually set and applied, test AUC improved to 69% and recall improved to 65% (for defaulted loans).

A Deep Neural Network (with an arbitrary two hidden layers node structure - DNN^a in Table 2) was initially applied to numerical data alone. In comparison with the Linear Classifier, test AUC and recall (for defaulted loans) scores improved to 68% and 67%, respectively. This indeed shows how more advanced feature combinations improve the predictive capabilities of the model. The improvement was expected, as the complexity of the phenomenon described by the target label surely implies more elaborated features and feature combinations than those originally provided to the model.

The DNN was then refined with a grid search on node numbers n_1, n_2 for the two hidden layers. The grid search was run over all combinations of values from the sets $n_1 \in \{5, 10, 15, 20, 30\}$, $n_2 \in \{1, 3, 5, 10\}$ and by applying a high level of dropout regularisation (20%). The level of dropout regularisation was empirically chosen from a $[0\%, 30\%]$ range, this is a reasonable range for this type of models often found in the literature. The strong regularisation aimed to reduce the DNN's intrinsic tendency to overfit, leading to a more robust and general model infrastructure. Results on the test set were indeed verified to be largely in line throughout the grid search, suggesting a model which is robust in the context of hyperparameter tuning.

Results for two network structures selected from the grid search (together with DNN^a - arbitrary two hidden layers node structure) are described in Table 2. These network structures are selected, as their results display the desirable properties of stable AUC and high recall on defaults.

Table 2: Table with main results from DNN architectures tested for the second phase of the model.

Loan Default Prediction Results					
Model	Dropout	Recall Train	AUC Test	Recall Test	Default
DNN [a]	20%	-	68%	67%	
DNN [b]	20%	71%	66%	75%	
DNN [c]	20%	68%	69%	72%	

[a] DNN with arbitrary node numbers $[n_1 = 20, n_2 = 5]$
[b] DNN with node numbers fine-tuned to $[n_1 = 30, n_2 = 1]$
[c] DNN with node numbers fine-tuned to $[n_1 = 5, n_2 = 3]$

Figure 2 is a representation of one of the weight instances of the fully trained DNN^c network in Table 2. The network representation in Figure 2 encodes the weight of each link in the fully connected layer as line thickness. Node size and colour are indicative of the normalised sum of outgoing weights from the node. This representation clearly constitutes an approximation, as the nodes contain non-linearities, but it still provides a useful visual interpretation and stability check tool.

3.2 Two phases analysis for "small business" category

The "purpose" feature described in Section 2.2 provides information about the purpose for which the loan was requested. The small business class of this feature is of particular interest here. This loan category was observed to have the highest fraction of defaulted loans amongst all categories and the least likelihood to survive throughout the lending term period [12]. Furthermore, this purpose is arguably different from the others and is more business-focused, rather than just a personal loan.

We therefore decided to look at this category in isolation, although it was included in the entire dataset used for the analyses described in the previous sections.

3.2.1 First Phase - Small Business Training Data Only

Logistic Regression and Support Vector Machines were trained and tested on "small business" loans alone. Two grid searches were trained for Logistic Regression, one maximises AUC whilst the other maximises recall macro. The former returns an optimal model with $\alpha = 0.1$, training AUC score $\simeq 88.9\%$ and test AUC score $\simeq 65.7\%$. Individual recall scores are $\simeq 48.0\%$ for rejected loans and 62.9% for accepted loans. The discrepancy between the training and

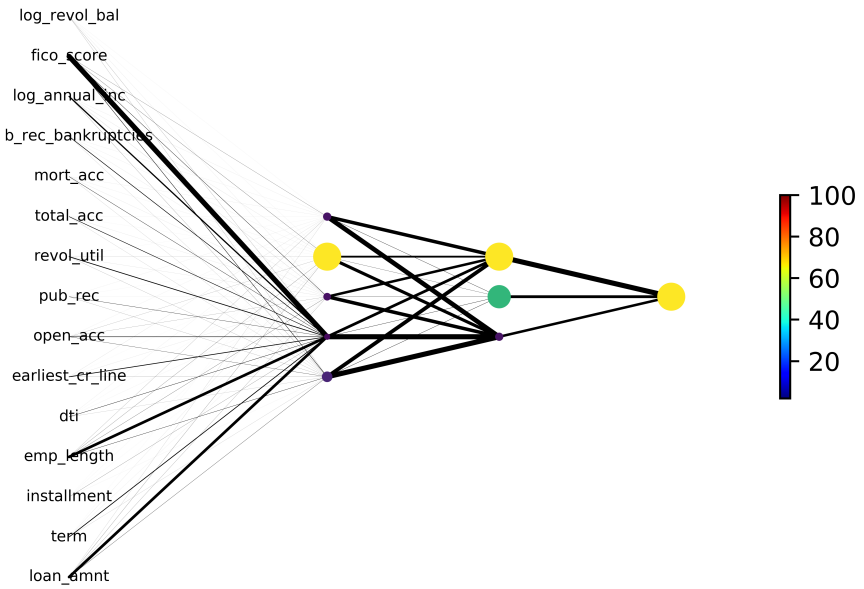

Figure 2: Neural network representation with node size and colour representing total outgoing weight and edge width proportional to the weight. The DNN represented is with node numbers fine-tuned to [5, 3] and $tanh$ non-linearities.

Table 3: Small business loan acceptance results and parameters for SVM and LR grids trained and tested on the data's "small business" subset.

Model	Grid metric	α	Training Score	AUC Test	Recall Rejected	Recall Accepted
LR	AUC	0.1	88.9%	65.7%	48.5%	62.9%
LR	recall macro	0.1	78.5%	65.5%	48.6%	57.0%
SVM	recall macro	0.01	-	89.3%	47.8%	62.9%
SVM	AUC	10	-	83.6%	46.4%	76.1%

test AUC scores indicates overfitting to the data or the inability of the model to generalise to new data for this subset. The latter grid search returns results which somewhat resemble the former one. Training recall macro is $\simeq 78.5\%$ whilst test recall macro is $\simeq 52.8\%$. AUC test score is 65.5% and individual test recall scores are 48.6% for rejected loans and 57.0% for accepted loans. This grid's results again show overfitting and the inability of the model to generalise. Both grids show a counterintuitively higher recall score for the underrepresented class in the dataset (accepted loans) whilst rejected loans are predicted with recall lower than 50%, worse than random guessing. This might simply suggest that the model is unable to predict for this dataset or that the dataset does not present a clear enough pattern or signal.

Support Vector Machines perform poorly on the dataset in a similar fashion to Logistic Regression. Two grid optimisations are performed here too, in order to maximise AUC and recall macro, respectively. The former returns a test AUC score of 89.3% and individual recall scores of 47.8% for rejected loans and 62.9% for accepted loans. The latter grid returns a test AUC score of 83.6% with individual recall scores of 46.4% for rejected loans and 76.1% for accepted loans (this grid actually selected an optimal model with weak L1 regularisation). A final model was fitted, where the regularisation type (L2 regularisation) was fixed by the user and the range of the regularisation parameter was shifted to lower values in order to reduce underfitting of the model. The grid was set to maximise recall macro. This yielded an almost unaltered AUC test value of $\simeq 82.2\%$ and individual recall values of 47.3% for rejected loans and 70.9% for accepted loans. These are slightly more balanced recall values. However, the model is still clearly unable to classify the data well, this suggests that other means of evaluation or features could have been used by the credit analysts to evaluate the loans. The hypothesis is reinforced by the discrepancy of these results with those described in Section 3.1.1 for the whole dataset. It should be noticed, though, that the data for small business loans includes a much lower number of samples than that described in Section 3.1.1, with less than $3 \cdot 10^5$ loans and just $\approx 10^4$ accepted loans.

3.2.2 First Phase - All Training Data

Given the poor performance of the models trained on the small business dataset and in order to leverage the large amount of data in the main dataset and its potential to generalise to new data and to subsets of its data, Logistic Regression and Support Vector Machines were trained on the whole dataset and tested on a subset of the small business dataset (the most recent loans, as by the methodology described in Section 2.2). This analysis yields significantly better results, when compared to those discussed in Section 3.2.1. Results are presented in Table 4.

Table 4: Small business loan acceptance results and parameters for SVM and LR grids trained on the entire dataset and tested on its "small business" subset.

Model	Grid metric	α	Training Score	AUC Test	Recall Rejected	Recall Accepted
LR	AUC	1	89.0%	71.9%	53.5%	60.2%
LR	recall macro	0.1	77.9%	71.7%	54.0%	59.9%
LR	fixed	0.001	80.0%	71.1%	55.2%	65.2%
LR	fixed	0.0001	80.1%	71.0%	55.9%	62.9%
SVM	recall macro	0.01	-	77.5%	52.6%	68.4%
SVM	AUC	10	-	89.0%	97.3%	43.3%

The results presented in Table 4 for Logistic Regression still present consistently higher recall for accepted loans. There is an apparent credit analyst decision bias towards rejecting small business loans. This could, though, be explained as small business loans have a higher likelihood of default, hence they are considered more risky and the model, trained on all the data, does not have this information. Information on loan defaults is present as a label only in default analysis, as no data is present for rejected loans. Future works might input the percentage of defaulted loans corresponding to the loan purpose as a new feature and verify whether this improves the model.

Results for Support Vector Machines are in line with those for logistic regression. The grid trained to maximise AUC is clearly overfitting the rejected class to maximise AUC and should be discarded. Results for the grid maximising recall macro follow the same trend of those from Logistic Regression. Recall scores are slightly more unbalanced. This confirms the better performance of Logistic Regression for the prediction task, as discussed in Section 3.1.1.

3.2.3 Second Phase

Logistic regression and Support Vector Machines were trained on accepted loan data in order to predict defaults of loans with "small business" purpose. Analogously to the analysis discussed in Section 3.2.1, the models were trained and tested on small business data alone. Results for models trained on small business data alone are presented in Table 5. Results for Logistic Regression are slightly worse and more unbalanced in individual recall scores than those presented in Section 3.1.2, this can be explained by the smaller training dataset (although more specific, hence with less noise). Surprisingly, again, the underrepresented class of defaulted loans is better predicted. This could be due to the significant decay of loan survival with time for small business loans, this data is obviously not provided to the model, hence the model might classify as defaulting, loans which might have defaulted with a longer term. Alternatively, most defaulting loans could be at high risk, while not all risky loans necessarily default, hence giving the score imbalance. Maximising AUC in the grid search yields best and most balanced results for Logistic Regression in this case. Analogously to the analysis in Section 3.2.1 class imbalance is strong here, defaulted loans are $\approx 3\%$ of the dataset. The better predictive capability on the underrepresented class might be due to loan survival with time and should be investigated in further works. Three threshold bands might improve results, where stronger predictions only are evaluated.

Support Vector Machines provide more balanced results, although worse overall, for this task. In both SVMs and LR we observe how stronger regularisation, corresponding to higher values of α, improves recall results on the test set for the overrepresented class. AUC test scores improve as well, suggesting an improvement in the model's ability to generalise.

Table 5: Small business loan default results and parameters for SVM and LR grids trained and tested on the data's "small business" subset.

Model	Grid metric	α	Training Score	AUC Test	Recall Defaulted	Recall Paid
LR	AUC	0.1	64.8%	66.4%	65.2%	57.4%
LR	recall macro	0.01	60.4%	65.3%	64.6%	53.3%
SVM	recall macro	0.01	-	59.9%	59.8%	58.8%
SVM	AUC	0.1	-	64.2%	50.8%	65.8%

Analogously to the analysis presented in Section 3.2.2, Logistic Regression and Support Vector Machines were also trained on all the data and tested on small business data only, in order to leverage the larger datasets, which might share signals with its "small business" subset. Results in this case, differ from those in Section 3.2.2, where an improvement was observed. Results are presented in Table 6. The model poorly predicts fully paid loans, with a recall score even below 50%. This might suggest that the way these loans are screened is similar to that of other categories, but their intrinsic default risk is very different indeed. This is also observed in the discrepancy in loan survival between these loans and all other loan categories. [12] The optimal parameters returned by the grid suggest weaker regularisation than that for results in Table 5. For predicting a subset of its data, stronger regularisation might improve results, this could be verified in future works. It should be considered, though, that regularisation might reduce the importance of a small subset of the data, such as that of small business loans. The fraction of the small business subset with respect to the complete dataset is roughly the same for loan acceptance ($\simeq 1.3\%$) and loan default prediction ($\simeq 1.25\%$). This indeed suggests a difference in the underlying risk of the loan and its factors.

As the conclusions about model generalisation described in Section 4 can be drawn already by comparing LR and SVM models, DNNs are not considered for to the small business dataset analysis in Section 3.2. DNNs are considered only for the purpose of improving model performance through more complex models and feature combinations, which is the theme of Section 3.1.

Table 6: Small business loan default results and parameters for SVM and LR grids trained on the entire dataset and tested on its "small business" subset.

Model	Grid metric	α	Training Score	AUC Test	Recall Defaulted	Recall Paid
LR	AUC	0.001 (L1)	69.8%	68.9%	81.0%	43.3%
LR	AUC	0.001	69.7%	69.2%	86.4%	35.0%
LR	recall macro	0.001	64.2%	69.2%	86.4%	35.0%
SVM	recall macro	0.001	-	64.1%	77.7%	48.3%
SVM	AUC	0.001	-	69.7%	77.7%	48.3%

4 Conclusions

In this paper we demonstrate that P2P loan acceptance and default can be predicted in an automated way with results above $\simeq 85\%$ (rejection recall) for loan acceptance and above $\simeq 75\%$ (default recall) for loan default. Given that the present loan screening has a resulting fraction of default around 20% (see Figure 1) we can infer that potentially the methodology presented in this paper could reduce the defaulting loans to 10% with positive consequences for the efficiency of this market. The best performing tools were Logistic Regression for loan acceptance and Deep Neural Networks for loan default. The high recall obtained with linear models on replicating traditional loan screening suggests that there is significant room for improvement in this phase as well.

The loan grade and interest rate features were found to be the most relevant for predicting loan default in [12]. The current model tries to predict default without biased data from credit analysts' grade and assigned interest rate, hence these features are excluded. The Deep Neural Network and Logistic Regression models provide substantial improvements on traditional credit screening. A recall score significantly and robustly above 70%, with AUC scores $\simeq 70\%$ for the Deep Neural Network, improves even on the logistic regression in [12]. The features provided to the model in our study generalise to any lending activity and institution, beyond P2P lending. The present work could therefore be augmented in order to predict loan default risk without the need for human credit screening.

The two phases model for all loan purposes described in Section 3.1 showed better performance overall, with well-balanced individual test recall scores for the second phase of 75% for defaulted loans. This shows the ability to predict well above 50% of defaults on loans screened and accepted by credit analysts, while not penalising excessively the acceptance of well performing loans. Training on the whole dataset for the first phase resulted in higher scores when applied to small business loans than when trained on small business loans alone. The opposite was true for the second phase, where default prediction was significantly better overall, when trained on small business loans alone. This suggests a discrepancy between how credit analysts treat these loans and how they might be treated more efficiently, in terms of their default risk and characteristics. Neural Networks were shown to significantly outperform the other models, suggesting that they might be used for default prediction, further to credit analyst screening. Neural Networks could also be combined with Logistic Regression in a conservative model, in order to mitigate their complex and not well-predictable nature. This and further data preprocessing and augmentation should be the subject of further work. We shall further extend our work to areas such as micro-financing in developing countries and loan-by-loan evaluation of loan portfolios for investment as well as to traditional lending. The integration of the present model with predictive

modelling based on information filtering network techniques [20], [21], [22], [23] will also be the subject of future research.

5 Acknowledgments

The authors acknowledge the EC Horizon 2020 FIN-Tech project for partial support and useful opportunities for discussion. JT acknowledges support from EPSRC (EP/L015129/1). JT acknowledges Dr. Guido Germano for useful feedback and discussions. TA acknowledges support from ESRC (ES/K002309/1), EPSRC (EP/P031730/1) and EC (H2020-ICT-2018-2 825215).

References

[1] Deloitte Reports. Beyond fintech: Disruptive innovation in lending. 2017.

[2] Kate Beioley. Fca sets out crackdown on peer-to-peer lending. *Financial Times*, 2018.

[3] Financial Conduct Authority. Cp18/20: Loan-based ('peer-to-peer') and investment-based crowdfunding platforms: Feedback on our post-implementation review and proposed changes to the regulatory framework. 2018.

[4] Nilas Möllenkamp. Determinants of loan performance in p2p lending. B.S. thesis, University of Twente, 2017.

[5] Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1):54–70, 2015.

[6] Carlos Eduardo Canfield. Determinants of default in p2p lending: the mexican case. *Independent Journal of Management & Production*, 9(1):1–24, 2018.

[7] Mingfeng Lin, Nagpurnanand R Prabhala, and Siva Viswanathan. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1):17–35, 2013.

[8] Joseph E Stiglitz and Andrew Weiss. Credit rationing in markets with imperfect information. *The American economic review*, 71(3):393–410, 1981.

[9] Chris Anderson. *The long tail: Why the future of business is selling less of more*. Hachette Books, 2006.

[10] Carlos Serrano-Cinca and Begoña Gutiérrez-Nieto. Microfinance, the long tail and mission drift. *International Business Review*, 23(1):181–194, 2014.

[11] Nathan George. All Lending Club loan data version 6, february 2018. `https://www.kaggle.com/wordsforthewise/lending-club`. Accessed: 2018-10-1.

[12] Carlos Serrano-Cinca, Begoña Gutiérrez-Nieto, and Luz López-Palacios. Determinants of default in p2p lending. *PloS one*, 10(10):e0139427, 2015.

[13] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[14] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[15] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015.

[16] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[17] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

[18] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[19] David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2011.

[20] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna. A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421–10426, 2005.

[21] Riccardo Marcaccioli and Giacomo Livan. A pólya urn approach to information filtering in complex networks. *Nature communications*, 10(1):745, 2019.

[22] Guido Previde Massara, Tiziana Di Matteo, and Tomaso Aste. Network filtering for big data: Triangulated maximally filtered graph. *Journal of complex Networks*, 5(2):161–178, 2016.

[23] Rosario N Mantegna. Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197, 1999.