Loan screening and default prediction with Machine Learning and Deep Neural Networks

Jeremy D. Turiel and Tomaso Aste

Abstract Logistic Regression and Support Vector Machine algorithms, together with Linear and Deep Neural Networks, are applied to lending data. A two layer model is formed, where the first layer predicts loan rejection, while the second one further screens the loans for default risk. Logistic Regression was found to be the best performer for the first layer, with test set recall macro score of 77.4%. Deep Neural Networks, applied to the second layer alone, were the best performer, with validation set recall score of 72%, for defaults. The models were also tested on the subcategory of the dataset comprising loans taken for small businesses. Benefits of training on the whole dataset to predict the specific category were tested. The first layer of the model was found to benefit strongly from the larger training dataset, while the second layer was found to perform significantly better when trained on small business data only. This suggests a potential discrepancy between how these loans are screened and how they should be analysed in terms of default prediction.

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 techniques, have renewed interest in risk prediction. Here, we present the analysis of two rich open source datasets reporting loans including credit card-related loans, wedding, house, small business and others. One dataset contains loans that have been rejected by credit analysts, whilst the other, which includes a significantly higher number

Department of Computer Science, University College London, Gower St, Bloomsbury, London WC1E 6BT, United Kingdom,
e-mail: jeremy.turiel.18@ucl.ac.uk,
e-mail: t.aste@ucl.ac.uk,
WWW home page: http:/www.cs.ucl.ac.uk/staff/tomaso_aste/

of features, represents loans which have been accepted and indicates their current status. Our analysis focuses on the loans which have been accepted or rejected as well as on loans which have been accepted and have subsequently been paid or have defaulted.

2 Method and Dataset

2.1 Dataset

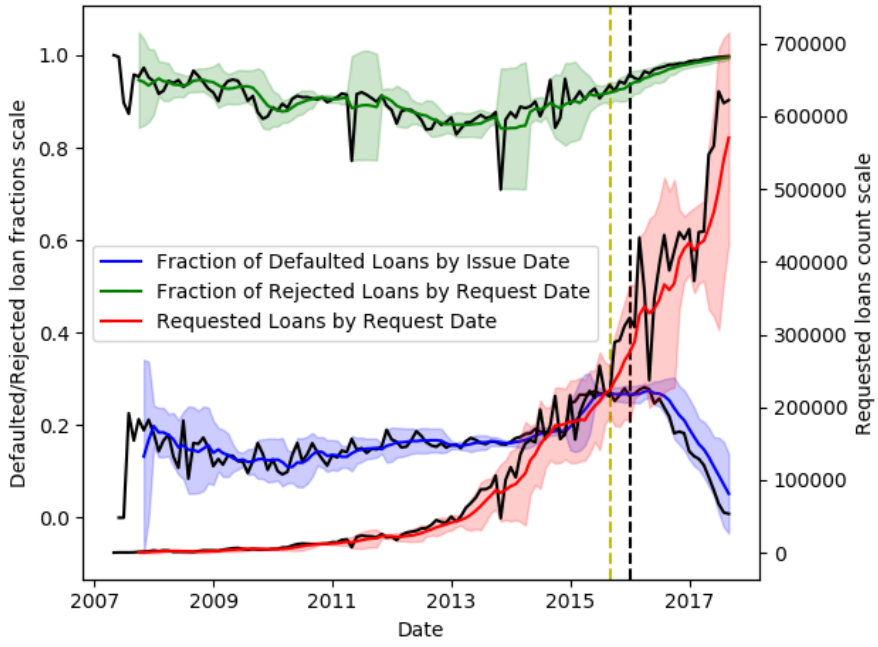

Fig. 1 Time series plots for the dataset. [1] Three plots are presented: the number of defaulted loans as a fraction of the total number of accepted loans (blue), the number of rejected loans as a fraction of the total number of loans requested (green) and the total number of requested loans (red). The 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 of the plots (black lines) and the shaded areas of the corresponding colours represent the standard deviation of the averaged data. The data on the right of the black dotted line was excluded due to the clear decrease in the fraction of defaulted loans, this was analysed 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 is collected from loans evaluated by Lending Club in the period between 2007 and 2017 (`https://www.lendingclub.com`). The data was downloaded from Kaggle (`www.kaggle.com`). We first analysed the dataset [1] 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 does not relevantly affect our analysis as the cumulative mean imputation is below 10% of the overall feature data. Furthermore, means or statistics are calculated here 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 presented in Figure 1.

Differently from other analyses of this dataset (or of earlier versions of it) [2] here only features which are known to the lending institution prior to evaluating the loan and issuing it are used, for analysis of default data. Features which were found to be very relevant [2] 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. The 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 the analysis.

2.2 Method and Discussion

The data introduced in Section 2.1 originates from two datasets. One contains all the loans which were rejected by analysts of the Lending Club and the other contains all loans which were accepted. The first dataset has a much lower number of features, as less information is gathered and recorded about rejected loans. The second dataset originally contained over 100 features, as issued loans are carefully recorded. The dataset of accepted loans indicates the status of each loan. Loans which had a status of fully paid or defaulted were selected for the analysis and this feature was used as target label for default prediction.

Two machine learning algorithms were applied to both datasets: logistic regression with underlying linear kernel and Support Vector Machines. [6, 7] The latter also aimed to clarify whether feature engineering in the form of polynomial features was to be beneficial. Neural Networks were also tested, but they were applied 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. [8] Neural networks were applied to default prediction with the aim to construct more complex features to improve the prediction of a non-trivial outcome such as loan default.

Machine learning is applied to the dataset discussed in Section 2.1 in order to form a two-layers model. The first layer tries to reproduce human decisions in accepting or rejecting loan applications, this combines the two datasets (one of accepted and the other of rejected loans). The second layer improves on human deci-

sions (or predicted human decisions). The dataset of accepted loans is analysed in order to predict whether the loan will default or will be fully paid. A higher number of features is available for this dataset alone, although certain features were excluded for the choice of field stated in Section 2.1.

Features for the first layer are reduced to those shared between the two datasets, geographical features (U.S. state and postcode) for the loan applicant were excluded, but should be evaluated carefully in further work. Features for the first layer are: Debt to Income ratio (of the applicant), employment length (of the applicant), loan amount (of the loan currently requested), 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 previous loans were used to train the model. This also simulates the human process of empirical 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 approximation, which is allowed as time sections are only introduced to refine model testing, does not apply to the second layer of the model where all dates correspond to the issue date.

Features considered for the second layer of the model are: loan amount (of the loan currently requested), term (of the loan currently requested), instalment (of the loan currently requested), employment length (of the applicant), home ownership (of the applicant, whether is owned, owned with a mortgage on the property or rented), verification status of the income or income source (of the applicant. Whether this was verified by the Lending Club), purpose for which the loan is taken, Debt to Income ratio (of the applicant), earliest credit line in the record (of the applicant), number of open credit lines (in applicant's credit file), number of derogatory public records (of the applicant), revolving line utilisation rate (the amount of credit the borrower is using relative to all available revolving credit), total number of credit lines (in applicant's credit file), number of mortgage credit lines (in applicant's credit file), number of bankruptcies (in the applicant's public record), logarithm of the applicant's annual income (the logarithm was taken for scaling purposes), FICO score (of the applicant), logarithm of total credit revolving balance (of the applicant).

For the second layer of the model, most recent loans were used as the validation set while earlier ones were used to train the model, as for the first layer. All numeric features for both layers were scaled by removing the mean and scaling to unit variance.

Hyperparameter tuning was performed for the model, mainly through grid searches. Regularisation techniques were applied to avoid overfitting, L2 regularisation was the most frequently applied, but the set of L1 and L2 regularisation techniques was, at times, included as a parameter in the grid search. [5] Hyperparameters were manually tuned, when specified in Section 3, either by shifting the parameter range in the grid search or by providing a specific value for the hyperparameter. This was mostly done when overfitting was clear 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.

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. [3] 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. [4]

Training and validation (or test) sets were used in the analysis. The dataset is 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 simulates future unseen data for the model to be tested on. Feature scaling is also trained on the training set alone and applied to the test set, in order to avoid information leakage. Grid searches for Logistic Regression and Support Vector Machines were trained to optimise the regularisation parameter and, occasionally, to choose between L1 and L2 regularisation. [5] The ranges for the regularisation parameter varied, but the widest range was $\alpha = [10^{-5}, 10^{5}]$. Despite the different ranges, values of alpha were all powers of 10 with integer exponents.

3 Results and Analysis

3.1 General two layers model for all purpose classes prediction

3.1.1 First Layer

Logistic regression was applied to the combined datasets to reproduce decisions made by the credit analysts in accepting or rejecting individual loans. The size of the combined and cleaned dataset is of $\approx 15 \cdot 10^6$ loans, including $\approx 8 \cdot 10^5$ accepted loans. Parameter tuning in the grid search was optimised to maximise the unweighted recall average of the two class labels (accepted and rejected). This was done as maximising AUC in the grid search led to overfitting the rejected class, which bares most of the weight in the dataset. The unweighted recall average is referred to as recall macro.

The grid search returned an optimal model with $\alpha \simeq 10^{-3}$, where α is the regularisation coefficient for L2 regularisation. The recall macro score for the training set was of $\simeq 79.8\%$. Test set predictions returned an AUC score of $\simeq 86.5\%$ and a recall macro score of $\simeq 77.4\%$. Test recall scores were of $\simeq 85.7\%$ for rejected loans and $\simeq 69.1\%$ for accepted loans.

The same dataset and target label were analysed with Support Vector Machines. A grid search was applied to tune α, the regularisation coefficient for L2. Training and test set results were in line with those for logistic regression. Analogously to the grid search for logistic regression, recall macro was maximised for SVM. 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 $\alpha = [10^{-3}, 10^{-5}]$.

Class imbalance for the target feature in the training set is observed to affect the recall scores for the two classes. Recall scores for accepted loans are lower by $\approx 15\%$, this indeed suggests that more training data would improve this score. A class imbalance of almost 20x does indeed impact 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 worthy borrowers do 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 layer, when just a simple model was used for credit screening.

3.1.2 Second Layer

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 as more features are involved and the intrinsic nature of the event (default or not) is probabilistic and stochastic.

Categorical features are also present in this analysis. These were "hot encoded" for the first two models, but they were were just provided as categorical features to the neural network (as the algorithm for this is able to handle categorical variables).

In order to obtain a more complete and representative validation set, the split between training and validation sets was modified to be $75\%/25\%$ for the first phase of the model. This provides 25% of the data for training, corresponding to approximately two years of data. This indeed constitutes a more complete sample for testing and was observed to yield better and more stable results. For the second phase, the periods highlighted in Figure 1 were used to split the dataset into training and test 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)

3.1.3 Second Layer - 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. 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 though lower, both in AUC and recall macro.

3.1.4 Second Layer - Support Vector Machine

In this layer, 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.

Support Vector Machines were also applied to the dataset. The optimal value of α returned by the grid search was the same as for logistic regression is 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. It can be inferred the analysis of this dataset does not benefit from higher dimensions. Furthermore, recall scores are improved for the overrepresented class in the dataset, this is the opposite of what is aimed for in this analysis. 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.

3.1.5 Second Layer - Neural Network

Linear Neural Network classifiers as well as Deep (two hidden layers) Neural Networks were also trained on the dataset for the second layer 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 in order to show the benefits of this on test set scores. 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 applied, test AUC improved to 69% and recall improved to 65% (for defaulted loans). This indeed shows how L2 regu-

larisation allows the model to better generalise and apply its predictive capabilities to future data.

A Deep Neural Network (with an arbitrary two hidden layers node structure) 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 $n_1 \in \{5, 10, 15, 20, 30\}$, $n_1 \in \{1, 3, 5, 10\}$ and by applying a high level of dropout regularisation (20%). 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 in the grid search are described in Table 1, as their results display desirable properties.

Table 1 Table with main results from DNN architectures tested for the second phase of the model.

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

[a] DNN with arbitrary node numbers [20, 5]

[b] DNN with node numbers fine-tuned to [30, 1]

[c] DNN with node numbers fine-tuned to [5, 3]

The Deep Neural Network and the Logistic Regression models provide substantial improvements on the first credit analyst screening. Furthermore, they include only features which can be generalised to any lender and not only to P2P lending. A recall significantly and robustly above 70%, with AUC scores of $\simeq 70\%$ for the Deep Neural Network improves even on the logistic regression where features such as assigned loan grade and interest rate are provided to the model. [2] These features were found to be the most relevant for predicting loan default. The current model tries to predict default without biased data from credit analysts' grade and assigned interest rate. This could also be further improved in order to predict loan risk default without the need of human credit screening.

For interpretability purposes, taking advantage of the careful parsimonious modelling applied, we present in Figure 2 a visual representation of the Neural Network with results reported in Table 1.

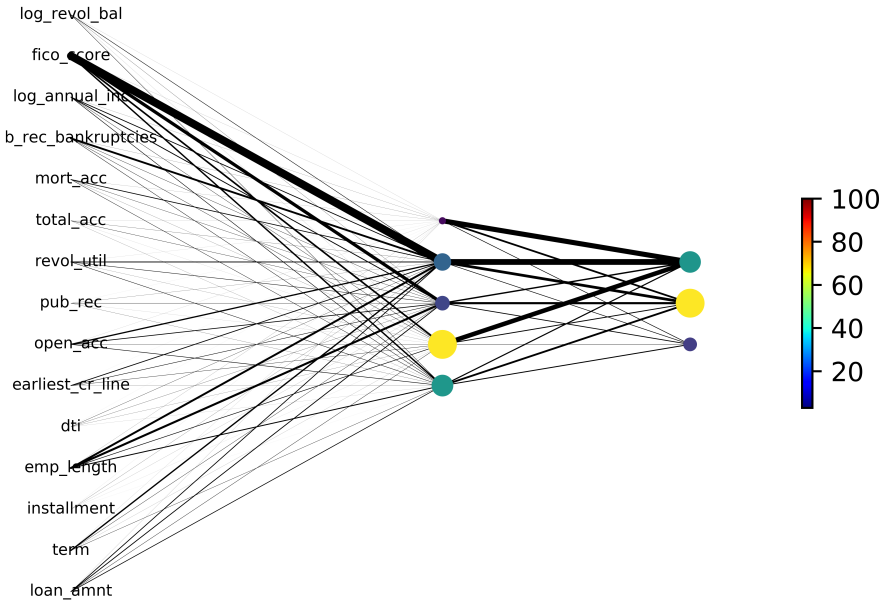

Fig. 2 Neural network representation with node size and colour representing total outgoing weight and edge width proportional to the weight. The DNN represented is with node numbers fine-tuned to [5, 3] and *tanh* non-linearities.

3.2 Two layers 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. [2] Furthermore, this purpose is arguably different from the others and is more business-focused, rather than just a personal loan.

3.2.1 Prediction of small business loan acceptance - Logistic Regression

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 of $\simeq 88.9\%$ and test AUC score of $\simeq 65.7\%$. Individual recall scores are $\simeq 48.0\%$ for rejected loans and 62.9% for accepted loans. The discrepancy between the training and 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.

3.2.2 Prediction of small business loan acceptance - Support Vector Machine

Support Vector Machines perform poorly on the dataset in a similar fashion to Logistic Regression. Two grids are fit 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. 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 loan. 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.

3.2.3 Prediction of small business loan acceptance - all training data

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 Sections 3.2.1 and 3.2.2. Results are presented in Table 2.

The results presented in Table 2 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

Table 2 Training and test set 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%

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. Further work 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 imbalanced. This confirms the better performance of Logistic Regression for this prediction task, as discussed in Section 3.1.1.

3.2.4 Prediction of small business loan default

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 Sections 3.2.1, 3.2.2 and 3.2.3, the models were trained on small business data alone as well as on all data, both trained models were then tested on small business data. Results for models trained on small business data alone are presented in Table 3. Results for Logistic Regression are slightly worse and more imbalanced 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 3 Training and test set 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 3.2.3, Logistic Regression and Support Vector Machines are now 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 3.2.3, where an improvement was observed. Results are presented in Table 4. 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. [2] The optimal parameters returned by the grid suggest weaker regularisation than that for results in Table 3. For predicting a subset of its data, stronger regularisation might improve results, this could be verified in further 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 ruffly 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.

4 Conclusion

Results for the method and dataset, described in Section 2, were presented in Section 3. The general two layers model for all loan purposes described in Section 3.1 showed better performance overall, with well-balanced individual test recall scores for the second layer of 63.8% and 63.6% for defaults and fully paid loans, respectively. This shows the ability to predict well above 50% of the defaults, further to

Table 4 Training and test set 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%

credit screening, while not penalising excessively the acceptance of well performing loans. Training on the whole dataset for the first layer resulted in higher scores when applied to small business loans, rather than when trained on small business loans alone. The opposite was true for the second layer, 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, this suggests that they might be used for default prediction, further to credit analyst screenings. Neural Networks could also be combined, due to their complex and not well-predictable nature, in a conservative model with Logistic Regression. This should be the subject of further work.

5 Data Availability

Data is freely available from Kaggle: `https://www.kaggle.com/wordsforthewise/lending-club`. All interested parties will be able to obtain the data in the same manner the authors did.

References

1. Nathan George.: *All Lending Club loan data. Version 6, February 2018. Kaggle. Available from https://www.kaggle.com/wordsforthewise/lending-club.*
2. Serrano-Cinca C, Gutiérrez-Nieto B, López-Palacios L (2015) Determinants of Default in P2P Lending. PLoS ONE 10(10): e0139427. https://doi.org/10.1371/journal.pone.0139427
3. Andrew P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, Volume 30, Issue 7, 1997, Pages 1145-1159, ISSN 0031-3203, https://doi.org/10.1016/S0031-3203(96)00142-2 (http://www.sciencedirect.com/science/article/pii/S0031320396001422)
4. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies.

2 (1): 37-63. http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

5. Ng, A.Y. (2004). Feature selection, L1 vs. L2 regularization, and rotational invariance. ICML '04.
6. David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant. Applied Logistic Regression. Wiley Series in Probability and Statistics, Volume 398. John Wiley & Sons (2013).
7. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998. doi:10.1109/5254.708428
8. Jürgen Schmidhuber, Deep learning in neural networks: An overview, Neural Networks, Volume 61, 2015, Pages 85-117, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2014.09.003. (http://www.sciencedirect.com/science/article/pii/S0893608014002135)