P2P Loan Repayment Prediction with Imbalanced Training Sets

Bernard Mizzi

ABSTRACT

Loan defaulting was one of the major causes leading to the Great Recession in 2008-2009. Having systems which correctly identify loan defaulters is essential to the financial markets to avoid major losses which might negatively impact the economy. Recent advancements in technology have introduced the creation of online platforms on which people can apply for loans. These platforms are known as online Peer-to-Peer lending platforms (P2P). Loans issued through these types of platforms are normally unsecure, and, thus, it is crucial to correctly identify loan defaulters so that lenders avoid losses. Using the data obtained from a P2P lending platform based in the USA, we apply machine learning techniques to attempt predicting defaulted loans in the P2P lending environment. The role of data preparation and training-testing selection techniques are also investigated to improve the predictive capability of a classifier. Due to having a disproportionate number of defaulted loans, such environments suffer from the class imbalance problem. Hence, sampling techniques to tackle class imbalance are also included. We discover that applying a dynamic approach which constantly updates a classifier is effective in such environments. This dynamic approach is combined with existing classifiers which outperform the traditional machine learning techniques. The hypothesis tests indicate that dynamic models outperform static ones.

1 INTRODUCTION

A loan is a business process in which a borrower borrows money from a lender. The borrower might fail to fully pay the loan, and, hence, the lender risks losing all the funds initially invested in the borrower. This is called loan defaulting, and this problem was one of the major issues which led to the 2008 financial crises [32].

Apart from the typical loans issued by banks, loans can also be issued on Peer-to-Peer (P2P) lending platforms. P2P lending platforms allow people to apply for loans which are funded by other individuals [48]. On such platforms, loans are normally unsecure [30], and due to lenders not being entirely knowledgeable of how to properly distinguish between a good and a bad loan, lenders are at higher risk of suffering losses. Hence, this increases the need to have systems which automatically predicts whether an applicant will most likely fail to fully pay a loan. This task is a binary classification problem called loan default prediction, and it can incorporate machine-learning concepts to achieve the required goal [12]. Additionally, the disproportionate number of defaulted loans within the data being investigated is a major drawback for loan default prediction, this problem is called class imbalance [22].

Furthermore, the literature has so far failed to fully address loan default prediction as a time-series problem [27]. Such data may contain hidden information that may change as time passes. This change is also known as concept drift [43], and although it might not be clearly evident that it resides within the data, it may still negatively impact the predictive capability of a classifier [45]. Therefore, research addressing systems which are capable of coping with concept drift is essential to this area of study.

1.1 Aims and Objectives

The aim of this study is to create a system which correctly identifies defaulted loans. Such a system will make use of machine learning aspects along with data preparation and training-testing selection techniques which might improve the predictive capability of a classifier. The aim is also to confirm whether a dynamic approach of classifying loan defaulting is worth investigating. Dynamic techniques are models which are repeatedly trained on new data to cope with the possible drift which may exist within the data being investigated.

Hence, the aim is on reaching following objectives in order to achieve the aforementioned aims of this work:

1. Identify a reference work and investigate the role of data pre-processing to check for any improvements in the results.
2. Identify the most effective classification models and combine them with any useful techniques identified in the first objective to improve the results of the reference work.
3. Investigate whether dynamic models perform better than static models in time-series P2P lending environments.

2 BACKGROUND & LITERATURE OVERVIEW

In this section we provide a brief overview of the background - an exposition of the main topics, ideas, concepts, and theory related to the problem of predicting loan defaulting.

2.1 Loans

A loan is a process where a borrower requests money from a lender. Should a borrower fail to pay his loan within an agreed period of time, also known as loan term length, the loan is defaulted [4, 33]. The continuous decline of performing loans and the increase of non-performing loans was the major factor behind the financial catastrophe in developed countries [2]. Loan defaults in the United States were one of the main issues which created financial instability [2]. In this regard, one may emphasize the need of a system which correctly detects defaulters.

2.2 P2P Lending

Recent advancements in technology have enabled the use of online portals in which borrowers can apply for a loan without the need of being physically present [11]. Zhao et al. [48] define P2P lending as individuals lending money to other individuals using online platforms which match the lenders and borrowers. Borrowers apply for the loan on an online platform and the lenders view and select the loans which they would like to fund. According to Klafft [24], there are several benefits incorporated in P2P lending:

- P2P lending reduces the costs for loan application as loans are applied for on an online platform and not through the use of a middle-man.
- There may be more information provided through the use of the online platform, since with banks, some information

that comes with the loan application may be lost through the banks' policies and standard processes.

- All loans are visible to the lenders, this creates a sense of fairness.
- Loans on these types of platforms are normally unsecure [30], thus the interest rates are higher, and so implying that lenders have higher returns.

Although these P2P lending platforms come with the aforementioned advantages, other disadvantages are also included in these platforms' business model. Unlike the banking sector, where the credit risk is analysed by experts, in P2P lending the credit risk is assumed by the lenders, who are normal individuals [39]. The P2P lending online platforms must make sure that there is enough information about the loan application so that the lenders are able to make good decisions. In our work, we concentrate on the data acquired from Lending Club, which is an online P2P lending platform based in the United States.[1]

2.3 Class Imbalance

Japkowicz [20] defines class imbalance as having one class having a disproportionate number of examples. This problem affects the performance of classification models which assume a balanced class distribution. In loan default prediction, this occurs when there is a small amount of defaulted loans examples [28]. In this section, we refer to data-level approaches, which address class imbalance by rebalincing the minority and majority classes.

2.3.1 Random Undersampling. Random Undersampling (RU) randomly undersamples the majority class to achieve balances between classes [1, 20]. As pointed out by Xin [1], one clear disadvantage of RU is that it may omit useful data when removing examples from the majority class.

2.3.2 Random Oversampling. Random Oversampling (RO) randomly augments examples from the minority class to achieve balance between classes [1]. Due to having exact replicates of certain examples from the minority class, RO suffers from over-fitting [1]. Over-fitting is defined as having a learner which generalises poorly to new unseen data after having seen the observed data [47].

2.3.3 Synthetic Minority Oversampling Technique and its variations. Synthetic Minority Oversampling Technique (SMOTE) [10] oversamples the minority class by creating "synthetic" examples to achieve balance between the two classes. The new examples are generated from minority class examples by creating new instances along the lines which join any/all of the k neighbours from the minority examples. The k nearest neighbours are identified using a distance function (metric) such as, for example, Euclidean distance [3].

Classifiers tend to learn more the examples at the boundaries of each class to attempt achieving better results [17]. These examples tend to be more often mis-classified, nonetheless, they are still the most important instances for the purpose of classification. Thus, Han et al. [17] proposed borderline-SMOTE1 (SMOTE BL 1) and borderline-SMOTE2 (SMOTE BL 2) to over-sample the examples at the boundaries.

2.3.4 Dealing with Class Imbalance in P2P Loan-Default Prediction. Ferreira et al. [15] utilised four sampling techniques combined with machine learning models to tackle the class imbalance problem in the Lending Club data. The authors used SMOTE, SMOTE BL 1, SMOTE BL 2 and RU. Data-level sampling techniques were also used by Namvar et al. [30]. The results showed that the inclusion of sampling techniques improves the performance of a classifier.

Research also showed that hybrid models were found to be effective in tackling the class imbalance problem. Zieba and Härdle [49] combined an ensemble and a re-sampling diversification technique. The sampling used a beta binomial distribution for each class to generate indexes of examples that are used for the boosting iteration that follows. The proposed approach outperformed the AdaBoost (AB) and Balanced Bagging (BalBag). Niu et al. [31] also created a hybrid approach, which consisted of a sampling ensemble model, based on data distribution (REMDD). REMDD utilised undersampling based on majority class data distribution (UMCDD). Based on bagging, it first generates minority training subsets, followed by generating majority training subsets. Niu et al. incorporated their technique with a DT. The results showed that stand-alone classifiers are not suitable for unbalanced areas.

2.4 Concept Drift

Tsymbal [43] defined concept drift as the change in the hidden environment which can cause changes in the target environment. The hidden environment is hidden information which may negatively affect the capability of a classifier. Such information may cause the data to be dynamic and, thus, it causes deterioration of model performance [45]. In our study, the target environment is whether a loan is defaulted.

2.4.1 Concept Drift in P2P Loan Default Prediction. Krempl et al. [25] created a system called Temporal Density Extrapolation (TDX), which predicted the probability density of a univariate feature. TDX was applied on 10 years worth of Lending Club data from 2007 till 2017. Their system was capable of detecting drift in features `revol_util` and `int_rate`. Jelsma [21] also showed there exists drift in the Lending Club data between 2007 and 2017. Jelsma suggested that drift might have occurred due to changes in definitions or in business decision rules.

Liu et al. [27] also showed that classifying data which is far from the data used for training shows decay in performance. Better results were achieved when classifying data closer to the training data. Although not tested for by Liu et al., these results might have occurred due to having drift present in the data.

2.5 Machine Learning

Machine learning addresses the subject of building systems which learn and adapt without given any specific instructions, such an example is supervised learning. In this study, we focus on supervised learning, due to the fact that our classifiers train and adapt from past data, and then, they predict unseen data.

2.5.1 Artificial Neural Network. An ANN is a type of supervised learning technique consisting of connected nodes, whose functionality is based on the biological neuron [16]. In an ANN, the neurons are called nodes, and nodes in the ANN are interconnected

[1] https://www.lendingclub.com/ - Last Access Januaray 2020

through (multiple) layers. The input layer receives and forwards the input to the hidden layers, which use an activation function to send outputs to the following layers [18]. The output layer uses an activation function to produce the final output signal of the ANN. Feed-forward networks are the typical ANNs which receive input at the input layer, and forward the signals until they reach the output layers, an example of which is presented in Figure 1.

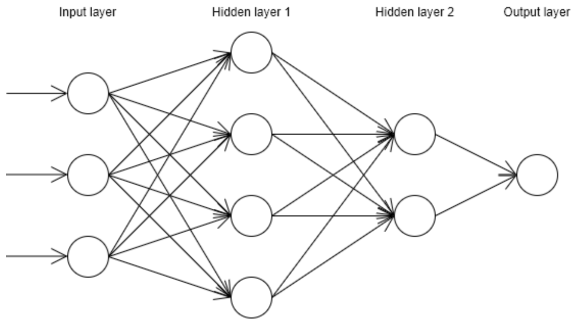

Figure 1: Adapted from [13], this figure shows an example of an ANN with two hidden layers.

Signals between nodes are computed by multiplying each input with the weight of the associated connections [26], using Equation 1:

$$f(x) = \sum_{i=1}^{N} w_i x_i + b \tag{1}$$

where $f(x)$ is the output, w_i is the weight of an incoming connection, x_i is the input of an incoming connection, b is the bias of the node, and N is the number of incoming connections. The weight of each node is updated using back-propagation [37]:

(1) Feed-forward computation
(2) Back-propagation to the output layer
(3) Back-propagation to the hidden layer
(4) Weight updates

The aforementioned steps are repeated until the error becomes significantly small. In simpler terms. the error is the difference between the predicted and the actual value.

2.5.2 Classification Models in P2P Loan Default Prediction. There exists an extensive list of literature which study the loan repayment prediction, and Neural Networks (NN) [14, 41, 44] show promising results.

Turiel and Aste [44] utilised a Deep Neural Network (DNN) for loan default prediction using Lending Club data. The DNN was trained with an Adam optimiser [23], utilised the softmax cross-entropy loss function, consisted of two hidden layers with a linear activation function, and an output hidden layer having the *tanh* activation function. Turiel and Aste suggested that DNNs should be used for loan default prediction.

NNs also showed better results in Duan [14]. Lending Club data covering between 2007 and 2015 were used. Due to class imbalance, Turiel and Aste made use of SMOTE to increase the distribution of the minority class. In terms of accuracy, the created NN outperformed the other models, which were AB, NN with one hidden layer, Decision Tree (DT), Logistic Regression (LR) and Support Vector Machine (SVM).

Sun and Vasarhelyi [41] also created a DNN which managed to outperform a DT, LR and Gaussian Naive Bayes (GNB). The DNN had three hidden layers. The *sigmoid* activation function was used for the final layer, and the *Rectifer* [6] function was used for the three hidden layers. With their results, the authors showed that the DNN is powerful in the area of credit card fraud detection.

The majority of literature using the Lending Club data normally randomly sample the data for training, and then test it on out of sample data [27]. Liu et al. [27] applied neuroevelution to show that model performance decays if the model is applied when trained on old data. The authors also showed how their model performs when training and classifying window by window, and they achieved promising results. However, although the authors claim that the models should be trained on most recent data, they fail to compare the results to models which are trained on randomly sampled data.

3 METHODOLOGY

We now describe the methodology used to achieve the research aims and objectives of this study. We present reference work and the dataset that was used. We successfully replicated the data pre-processing and results of Ferreira et al. [15], and we also established a baseline. We then describe how we improve upon the work of Ferreira et al. by using much of the concepts, theory, and techniques presented in Section 2.

3.1 Reference Work

The identified work which we analyse in this study is that of Ferreira et al. [15]. It addresses the class imbalance problem in P2P loan default prediction, and uses a set of machine learning models to predict loan defaults. Ferreira et al. use the Lending Club data from 2007 till 2016. Their work covers an extensive range of data on which research can be conducted. Furthermore, the data is made publicly available by Lending Club, and there exists several work which investigate it. Ferreira et al. establish three baseline machine learning models: DT, LR and GNB. To address the class imbalance problem, they include four sampling techniques: RU, SMOTE, SMOTE BL 1 and SMOTE BL 2. They further include the Random Forest (RF), AdaBoost (AB) and Bootstrap Aggregating (BG), because, according to Ferreira et al., stand-alone ensemble classifiers can also cope with class imbalance. Ferreira et al. show that the baselines incorporated with the sampling techniques outperformed the baselines with the cost-sensitive approach and the ensembles.

3.2 Data Pre-Processing

The dataset (training and testing) was downloaded from Lending Club,[2] it covered from 2007 till 2016. The features filtered by Ferreira et al. [15] are features which are only available via API, which refers to data being available only at the time that the lenders are bidding for the loan. This avoids data leakage, as other features which were ignored are made available only after the loan is issued to the borrower. In our work, we also use the same features.

[2] https://www.lendingclub.com/public/about-us.action - Last Access Januaray 2020

Ferreira et al. [15] applied one-hot encoding to the categorical features. Ferreira et al. also set out to remove loans which were missing 50% of their features. It resulted that there were none, and, thus, there were 578,331 loans remaining in the data. Subsequently, features which had 50% of their data missing were removed. The authors then removed features which had variability less than 0.25. After this exercise, the total number of features was reduced to 134. Missing numeric values where imputed with the mean, whereas the missing categorical fields where imputed with the mode.

After the data was processed, a total of 578,331 loans were captured, 79.71% of which were fully paid loans, and 20.29% were defaulted loans. In these figures, class imbalance is evident. This information is depicted in Figure 2. The 79.71% are 461,007 fully paid loans, and the 20.29% are 117,324 defaulted loans.

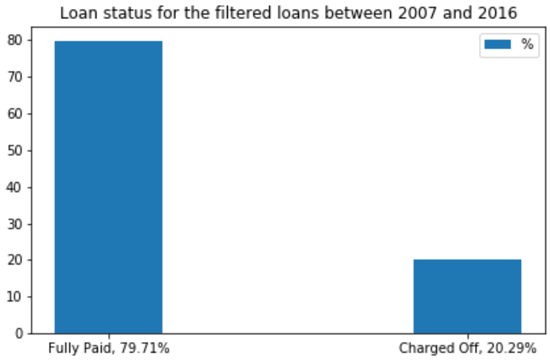

Figure 2: Produced with `matplotlib`,[3] this figure shows loan statuses for filtered loans from 2007 till 2016.

3.3 Experimentation

In Section 1.1, we defined the aims and objectives of our study. In this section, we will provide the details of how we will be addressing each objective. However, we must first establish the baseline. Taking into consideration the results presented in [21, 25], which showed that concept drift exists in the Lending Club data, and also the remarks made in [27], which mainly were that data should be should be sorted for the experimentation, we establish new baselines.

3.3.1 Baselines. The new baselines use the same models and sampling techniques used by Ferreira et al. [15], but the training and testing data are sorted chronologically. We use a 70%-30% training-testing set split. As done in [15], we also randomly sample 1% and 5% subsets from the 70% training set, to train and classify with the sampling techniques. Table 1 provides the results of the newly established baselines. It is clearly evident that the difference in improvement between the stand-alone models and the models combined with the sampling techniques pointed out by Ferreira et al. [15], in certain cases, is almost insignificant. The LR combined with the SMOTE BL 2 achieved the best result, however, by a very small margin. The applicability of ensemble techniques in addressing the class imbalance problem which was pointed out by Ferreira et al. [15] is evident, as the stand-alone RF managed to achieve the same AUROC as the best LR. However, it failed to learn the defaulted class, as both the TPs and Recall are very low. This implies that a considerable amount of losses are incurred, because bad loans which are mis-classified as good loans are still issued. With the exception of BG, AB and RF, the sampling techniques did well in assisting the classifier to learn the minority class, as the TPs are higher.

Table 1: New baselines after we sorted the data.

Model	AUROC	Accuracy	Precision	Recall	TP	FP	TN	FN
GNB NA	0.67	0.65	0.35	0.58	23642	43720	88764	17374
GNB RU	0.67	0.50	0.29	0.79	32386	78134	54350	8630
GNB SM	0.65	0.51	0.29	0.76	30972	75121	57363	10044
GNB SM BL 2	0.64	0.51	0.29	0.75	30845	75507	56977	10171
DT NA	0.69	0.77	0.00	0.00	0	0	132484	41016
DT RU	0.69	0.68	0.38	0.57	23178	38542	93942	17838
DT SM	0.68	0.76	0.50	0.10	3924	3969	128515	37092
DT SM BL 2	0.68	0.76	0.13	0.00	1	7	132477	41015
LR NA	0.71	0.77	0.54	0.09	3840	3269	129215	37176
LR RU	0.71	0.67	0.38	0.61	25057	41586	90898	15959
LR SM	0.71	0.64	0.36	0.67	27677	48813	83671	13339
LR SM BL 2	**0.71**	0.63	0.35	0.71	29013	52872	79612	12003
BG NA	0.70	0.77	0.58	0.05	1936	1387	131097	39080
BG RU	0.70	0.65	0.36	0.64	26093	45511	86973	14923
BG SM	0.69	0.76	0.51	0.09	3740	3596	128888	37276
BG SM BL 2	0.69	0.77	0.53	0.06	3489	3151	129333	37527
AB NA	0.53	0.77	0.55	0.09	3582	2932	129552	37434
AB RU	0.65	0.66	0.37	0.61	25141	42404	90080	15875
AB SM	0.50	0.76	0.50	0.20	8175	8311	124173	32841
AB SM BL 2	0.50	0.77	0.51	0.15	6064	5797	126687	34952
RF NA	0.71	0.76	0.66	0.01	244	125	132359	40772
RF RU	0.71	0.68	0.38	0.61	24827	40240	92244	16189
RF SM	0.70	0.76	0.48	0.22	9029	9943	122541	31987
RF SM BL 2	0.70	0.76	0.49	0.16	6741	6954	125530	34275

3.3.2 Dynamic Window Model. The first objective of our study is to investigate the role of data pre-processing. Liu et al. [27] showed that most of the work sample and shuffle the data, and that testing a classifier on data which is far from the training data causes deterioration of model performance. This may occur due to having dynamic changing data caused by drift [21, 25]. Although such procedure may seem to work, applying it may still not be ideal. To tackle this, we introduce the dynamic model.

The dynamic model is continuously trained and tested on new data to cope with the drift. We separate the sorted test set into equally sized windows of 5,000 loans each, to have a considerable number of samples so that our hypothesis tests presented in Section 4 have more statistical power [29]. We create pairs of train and test sets. Each train set is the first 20,240 loans preceding the current test set. This construction of the dynamic sets is depicted in Figure 3. In total, there are 35 pairs of training and testing sets. Eventually, we end up with 35 sets of probability predictions. We group all the predictions together, and form one final set of predictions from which we compute the metrics.

3.3.3 Artificial Neural Network. Existing literature [14, 41, 44] clearly indicate that an ANN outperforms other classifiers in P2P loan default prediction. Hence, we the Keras package to implement an ANN which utilises the dynamic approach. [4] We use the same number of hidden layers used by Turiel and Aste [44], and we also use the Adam optimiser. We utilise non-linear activation functions for both the hidden and output layers. Furthermore, we incorporate dropout to avoid overfitting [5, 40]. We also include early stopping

[4] https://keras.io/ - Accessed October 2021

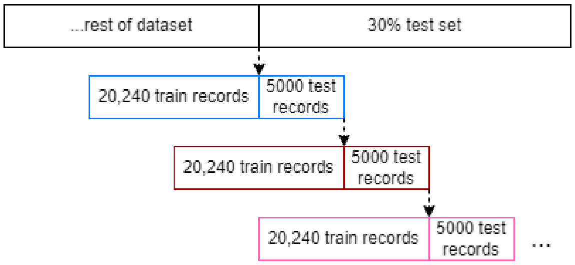

Figure 3: The dynamic model without validation data.

[9, 36]. Our early stopping mechanism uses the "patience" parameter, which means that if the ANN shows no further improvement in the validation loss after a certain number of epochs, the ANN stops training and it restores the state at which there was the lowest loss.

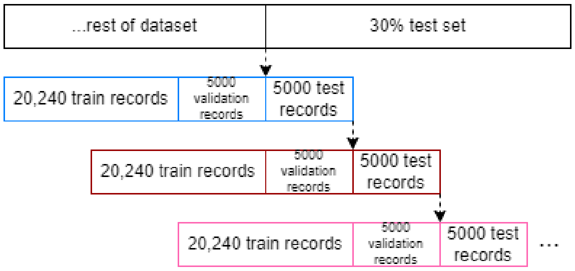

Figure 4: The dynamic model with validation data.

3.3.4 Comparing Dynamic and Static Models. Our last objective is to directly compare dynamic and static models to confirm whether dynamic models are better. Static models are models which are trained on a single set and used to classify data which, in terms of time, can either be close or far. This makes the classification task more challenging due to having drift within the data. The training data is a random subset of 20,240 loan records from the first 70% sorted data. The test data is the same test sets which are used for the dynamic approaches. The construction of the static model data is depicted in Figure 5. The classifier is trained on the single constructed training data, and is tested multiple times.

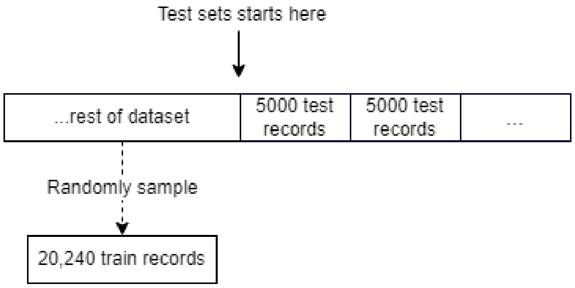

Figure 5: Construction of a static model.

4 RESULTS AND EVALUATION

We now present the experiments which address the aims and objectives established in Section 1.1, after we described their methodology in Section 3.

4.1 Dynamic Window Model

Following the construction presented in Figure 3, the 173,500 test loans created in Section 3.3.1 are split into 35 pairs of train and test sets. Each train set consists 20,240 loan records, whereas each test set contains 5,000 loan records. We use a grid search 5-fold cross-validation [35] to test the identical hyperparameters tested for in Ferreira et al. [15]. These are presented in Table 2. The selected hyperparameters are not indicated, as there were multiple training sets for each model. The results of these experiments are presented in Table 3.

We manage to outperform the baselines presented in Table 1, as the stand-alone LR obtained an AUROC of 0.73. Nonetheless, it failed to correctly identify the majority of the defaulted loans, as it achieved the second lowest amount of TPs, and a Recall of 0.18. The stand-alone RF experienced the same behaviour. The same results are summarized in Figure 6. The RF combined with sampling techniques did not do well in generalising over the defaulted class. Such situations require further analysis from the responsible parties who make use of such classification systems. The debate would be whether it would be ideal to issue more loans due to misclassifying defaulted loans as being fully paid. Issuing more loans would mean that the amount of profit made from interest increases. However, it also means that there will be more bad loans which are defaulted, thus, more losses are incurred. The other models combined with a sampling technique obtained an AUROC of 0.72, which also means that they outperform the baselines. The LR combined with the sampling techniques did better in identifying loan defaults, as the TPs are higher.

We applied the dynamic model data-processing technique and we successfully achieved the first objective. Going forward, we combine the dynamic approach with an ANN, to improve the results in Table 3.

Table 2: Hyperparameters tested for the dynamic model.

Classifier/ Samp. technique	Hyperparameter	Value
LR	c	0.001, 0.01, 0.1, 1, 10, 100, 1,000
	solver	newton-cg, lbfgs, liblinear, sag
RF	n_estimators	10, 50, 100
	max_features	auto, sqrt, log2, None
	max_depth	5, 10, 15
RU	ratio	0.8, 0.9, 1.0
SM	ratio	0.8, 0.9, 1.0
	k_neighbors	1, 3, 5, 7
SM BL 1	ratio	0.8, 0.9, 1.0
	k_neighbors	1, 3, 5, 7
SM BL 2	ratio	0.8, 0.9, 1.0
	k_neighbors	1, 3, 5, 7
RO	ratio	0.8, 0.9, 1.0

Table 3: The dynamic model results using the LR and RF.

Model	AUROC	Accuracy	Precision	Recall	TP	FP	TN	FN
LR NA	**0.73**	0.77	0.56	0.19	7674	5965	126519	33342
LR RU	0.72	0.69	0.39	0.60	24624	38083	94401	16392
LR SM	0.72	0.68	0.38	0.62	25228	40319	92165	15788
LR SM BL 1	0.72	0.68	0.39	0.61	25110	39521	92963	15906
LR SM BL 2	0.72	0.67	0.38	0.64	26164	42164	90320	14852
LR RO	0.72	0.68	0.39	0.62	25589	40440	92044	15427
RF NA	**0.73**	0.77	0.61	0.07	3005	1884	130600	38011
RF RU	0.72	0.67	0.38	0.65	26629	43308	89176	14387
RF SM	0.72	0.76	0.50	0.28	11365	11197	121287	29651
RF SM BL 1	0.72	0.76	0.50	0.27	11135	10971	121513	29881
RF SM BL 2	0.72	0.76	0.50	0.27	11050	11082	121402	29966
RF RO	0.72	0.70	0.41	0.56	22788	33125	99359	18228

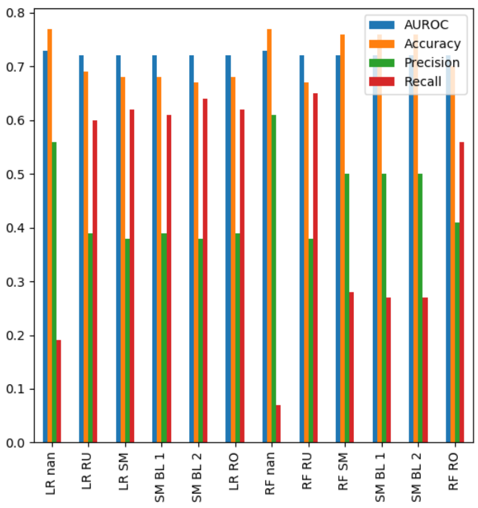

Figure 6: A depiction of the dynamic model results using the LR and RF.

4.2 Artificial Neural Network

We now combine the dynamic approach with the ANN. Similar to the experiments described in subsection 4.1, we follow the construction presented in Figure 4, and we train, validate and test for each 35 window which we have available. The first 5,000 loans preceding the current test set are used for validation, and then the first 20,240 loans preceding the validation set are the train set. As was done in Turiel and Aste [44], we remove the mean in the training, validation and test sets, and we scale to unit variance. [5]

All ANNs which we test consist of two hidden layers. Each ANN has one neuron in the output layer, which uses the *sigmoid* activation function to output a probability between 0 and 1 of whether a borrower is likely to default. In our experiments, a probability greater than or equal to 0.5 means that the loan is predicted as default, whereas a probability less than 0.5 means that the loan is predicted as fully paid. The number of neurons in the input layer for each model is 133. We varied the number of neurons in the two hidden layers between the values 120 and 60, 80 and 40, 60 and 30, and finally 40 and 20. We set the *LeakyReLU* as the activation function of both hidden layers. The *LeakyReLU* function was proposed to deal with the *dying* neuron problem of the ReLU activation function [34]. Due to its recent increase in usage over the past years, and also due to being one of the most commonly used ANN optimization techniques [7, 8, 44, 46], we choose the Adam optimizer [23] to train the ANNs. Initially, we set the learning rate to a small value of 0.0005, and we also train the ANNs using a fixed number of 25 epochs. However, we noticed that in certain cases, the ANN over-fitted, and in other cases, training was stopped early. Overfitting is shown in Figure 7, the validation loss showed no further improvement, and it started overfitting right after 15 epochs. On the other hand, Figure 8 shows an ANN which might have been stopped earlier than expected. Hence, we increase the number of training epochs to 300, and we include early stopping. Our early stopping configuration stops the training of an ANN and stores the weights before there have been 20 epochs without any improvement in the validation loss. We also increased the learning rate from 0.0005 to 0.001.

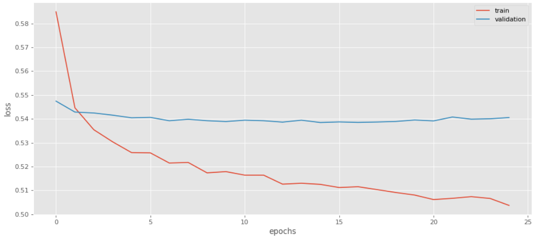

Figure 7: An ANN which over-fitted after 15 epochs.

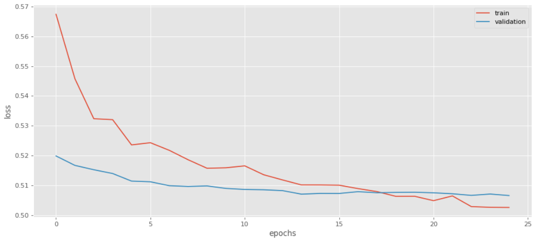

Figure 8: An ANN which was unnecessarily stopped early.

Due to time constraints, we did not include experiments involving sampling techniques. We obtain and AUROC of 0.74, and we outperform the LR and RF combined with dynamic approach in Table 3. Nonetheless, the number of TPs is low, and the number of FNs is high. This occurred due to not having involved sampling techniques.

4.3 Dynamic vs Static Approaches

We address the last objective by checking whether dynamic models perform better than static ones in loan default prediction. The

[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html - Last Accessed November 2021

Table 4: The dynamic model results using the ANN.

AUROC	Accuracy	Precision	Recall	TP	FP	TN	FN
0.73	0.77	0.58	0.16	6711	4866	127618	34305
0.73	0.77	0.56	0.17	7047	5427	127057	33969
0.73	0.77	0.58	0.16	6419	4675	127809	34597
0.73	0.77	0.59	0.14	5762	4056	128428	35254
0.73	0.77	0.57	0.16	6551	4914	127570	34465
0.74	0.77	0.57	0.17	6965	5245	127239	34051
0.74	0.77	0.57	0.17	7012	5274	127210	34004
0.74	0.77	0.57	0.17	6804	5161	127323	34212
0.74	0.77	0.57	0.18	7248	5490	126994	33768
0.74	0.77	0.57	0.16	6724	5073	127411	34292
0.73	0.77	0.57	0.17	6950	5186	127298	34066
0.73	0.77	0.58	0.15	6341	4543	127941	34675
0.73	0.77	0.57	0.16	6551	4914	127570	34465
0.74	0.77	0.58	0.17	6778	4960	127524	34238
0.74	0.77	0.57	0.17	7082	5389	127095	33934
0.74	0.77	0.57	0.17	6800	5091	127393	34216
0.74	0.77	0.57	0.16	6682	4952	127532	34334
0.74	0.77	0.57	0.17	7003	5193	127291	34013
0.73	0.77	0.57	0.16	6514	4822	127662	34502

static model uses the same test sets of the dynamic model, however, training is performed on a single set. The train and test set construction for the static model is depicted in Figure 5. Here, we compare the results of the dynamic and static approaches using the stand-alone LR and RF, and combined with RU, SMOTE, SMOTE BL 1, SMOTE BL 2 and RO. The tested hyperparameters for the static models are identical to those tested for the dynamic approach, and are presented in Table 2.

We conduct the hypothesis tests with 0.05 level of significance to compare the results between the dynamic and static models. We first start with the Paired-Sample T-Test to compare the mean AUROC of the dynamic and static approaches. The results are presented in Table 5. The columns refer to the six different types of models which are tested, one were there is no sampling technique, and another five which refer to the models having a sampling technique. All the tests return a value smaller than 0.05, thus meaning that the difference in the means of the dynamic and static models is not 0. Hence, there is a significant difference between the two models.

Table 5: The significance results obtained from the Paired-Sample T-Test.

Model	NA	RU	SMOTE	SMOTE BL 1	SMOTE BL 2	RO
LR	0.00	0.00	0.00	0.00	0.00	0.00
RF	0.00	0.00	0.00	0.01	0.00	0.00

Due to the fact that the Paired-Sample T-Test assumes that the data is normally distributed [38], we utilise the Wilcoxon Signed-Rank Test. This test compares the median AUROC of the dynamic and static approaches. The results of the Wilcoxon Signed-Rank Test are presented in Table 6. All the tests return a value smaller than 0.05. This also implies that there is a significant difference between the static and dynamic models.

Table 6: The significance results obtained from the Wilcoxon Signed-Rank Test.

Model	NA	RU	SMOTE	SMOTE BL 1	SMOTE BL 2	RO
LR	0.00	0.00	0.00	0.00	0.00	0.00
RF	0.00	0.00	0.00	0.00	0.00	0.00

Due to the fact that the Wilcoxon Signed-Rank Test assumes that the data is symmetric [42], we utilise the Paired-Sample Sign Test. Similar to what was obtained in the previous two tests, here, also, we see that the results of the tests indicate that there is a significant difference between the two models, as the values are all less than 0.05. Thus, using the results obtained from the three tests, we demonstrate that one should use the dynamic approach when performing research in the area of loan default prediction which involves time-series data.

Table 7: The significance results obtained from the Paired-Sample Sign Test.

Model	NA	RU	SMOTE	SMOTE BL 1	SMOTE BL 2	RO
LR	0.00	0.00	0.00	0.00	0.00	0.00
RF	0.00	0.00	0.00	0.00	0.00	0.00

5 CONCLUSIONS AND FUTURE WORK

We created a system which detects borrowers who are likely to default on their loan. We investigate the effect that different data preparation and training-testing selection techniques have on classifier performance. We perform a direct comparison between dynamic and static models to demonstrate which performs best. We thoroughly reviewed the work presented by Ferreira et al. [15], and we managed to fully replicate the data pre-processing and results. We established a baseline and we proceeded with implementing methods which involve data pre-processing and training-testing set selection. We achieved the first objective by introducing the dynamic approach.

We achieved better results by combining the ANN with the dynamic approach. However, there is no clear winner between the tested classifiers. It was evident that the ANN failed to correctly identify the majority of defaulted loans due do not having a sampling technique combined with it. It is clear that such a task is not easy and straight-forward. The parties involved in implementing such a system have to cater for the business requirements and needs, because although our task was to correctly identify defaulted loans, care must be taken to not loose potential profit from goods loans which are incorrectly classified as bad.

The third objective was achieved by performing tests at 0.05 level of significance to check whether the improvements introduced by the dynamic model were significant. The test results showed that they were. Hence, we suggest that dynamic models should be used in the area of P2P loan default prediction instead of static ones.

5.1 Future Work

Although we limited ourselves to an ANN, further research involving other machine learning models and sampling techniques can be carried out. One may also look at the applied data pre-processing.

Imtiaz and J [19] introduced techniques which can be used instead of mean and mode imputation. One may also decide to use the original categorical features, because they might better aid the DT to establish better splitting criteria.

Finally, research can also be done to properly check the effects of varying the decision boundary with which a classifier classifies a defaulted loan. Our work shows that this decision has an effect on the amount of potential profits which lenders lose in cases where classifiers are biased towards the default class. Proper financial analysis can be carried out in collaboration with a loans expert so that each decision boundary can be properly analysed and evaluated.

REFERENCES

[1] 2008. On the Class Imbalance Problem. In *2008 Fourth International Conference on Natural Computation*, Vol. 4. IEEE, 192–201.

[2] 2013. Micro and Macro Determinants of Non-performing Loans. *International journal of economics and financial issues* 3, 4 (2013), 852–860.

[3] Haneen Arafat Abu Alfeilat, Ahmad B. A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh Eyal Salman, and V. B. Surya Prasath. 2019. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data* 7, 4 (2019), 221–248.

[4] Uzair Aslam, Hafiz Ilyas Tariq Aziz, Asim Sohail, and Nowshath Kadhar Batcha. 2019. An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience* 16, 8 (2019), 3483–3488.

[5] Pierre Baldi and Peter J. Sadowski. 2013. Understanding Dropout. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* 2814–2822.

[6] Y Bengio, P Simard, and P Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.

[7] Sebastian Bock, Josef Goppold, and Martin Georg Weiß. 2018. An improvement of the convergence proof of the ADAM-Optimizer. *CoRR* abs/1804.10587 (2018).

[8] Sebastian Bock and Martin Georg Weiß. 2019. A Proof of Local Convergence for the Adam Optimizer. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019.* IEEE, 1–8.

[9] Rich Caruana, Steve Lawrence, and C. Lee Giles. 2000. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA.* MIT Press, 402–408.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *The Journal of artificial intelligence research* 16 (2002), 321–357.

[11] Dongyu Chen and Chaodong Han. 2012. A Comparative Study of online P2P Lending in the USA and China. *Journal of internet banking and commerce : JIBC* 17, 2 (2012), 1.

[12] Ya-Qi Chen, Jianjun Zhang, and Wing W. Y. Ng. 2018. Loan Default Prediction Using Diversified Sensitivity Undersampling. In *2018 International Conference on Machine Learning and Cybernetics, ICMLC 2018, Chengdu, China, July 15-18, 2018.* IEEE, 240–245.

[13] Vlastimil Dohnal, Kamil Kuca, and Daniel Jun. 2005. What are artificial neural networks and what they can do? *BIOMEDICAL PAPERS-PALACKY UNIVERSITY IN OLOMOUC* 149, 2 (2005), 221.

[14] Jing Duan. 2019. Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute* 356, 8 (2019), 4716–4731.

[15] Luis Eduardo Boiko Ferreira, Jean Paul Barddal, Heitor Murilo Gomes, and Fabrício Enembreck. 2017. Improving Credit Risk Prediction in Online Peer-to-Peer (P2P) Lending Using Imbalanced Learning Techniques. In *29th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2017, Boston, MA, USA, November 6-8, 2017.* IEEE Computer Society, 175–181.

[16] Kevin Gurney. 2018. *An introduction to neural networks.* CRC press.

[17] Hui Han, Wenyuan Wang, and Binghuan Mao. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I (Lecture Notes in Computer Science)*, De-Shuang Huang, Xiao-Ping (Steven) Zhang, and Guang-Bin Huang (Eds.), Vol. 3644. Springer, 878–887.

[18] Robert Hecht-Nielsen. 1988. Theory of the backpropagation neural network. *Neural Networks* 1, Supplement-1 (1988), 445–448.

[19] Sharjeel Imtiaz and Allan J. 2017. A Better Comparison Summary of Credit Scoring Classification. *International journal of advanced computer science & applications* 8, 7 (2017).

[20] Nathalie Japkowicz. 2000. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, Vol. 56. Citeseer.

[21] Tineke Jelsma. 2021. *Detecting Prediction Influence Drift In Data Streams.* Master's thesis.

[22] Sihem Khemakhem and Younes Boujelbene. 2018. Predicting credit risk on the basis of financial and non-financial variables and data mining. *Review of accounting and finance* 17, 3 (2018), 316–340.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

[24] Michael Klafft. 2008. Online peer-to-peer lending: a lenders' perspective. In *Proceedings of the international conference on E-learning, E-business, enterprise information systems, and E-government, EEE.* 371–375.

[25] G Krempl, D Lang, and V Hofer. 2019. Temporal density extrapolation using a dynamic basis approach. *Data mining and knowledge discovery* 33, 5 (2019), 1323–1356.

[26] Ben Krøse and Patrick Van Der Smagt. 1996. Patrick van der Smagt. (1996).

[27] Yue Liu, Adam Ghandar, and Georgios Theodoropoulos. 2020. Online NEAT for Credit Evaluation - a Dynamic Problem with Sequential Data. *CoRR* abs/2007.02821 (2020).

[28] Ana Isabel Marqués, Vicente García, and José Salvador Sánchez. 2013. On the suitability of resampling techniques for the class imbalance problem in credit scoring. *J. Oper. Res. Soc.* 64, 7 (2013), 1060–1070.

[29] Charity J Morgan. 2017. Use of proper statistical techniques for research studies with small samples. *American journal of physiology. Lung cellular and molecular physiology* 313, 5 (2017), L873–L877.

[30] Anahita Namvar, Mohammad Siami, Fethi A. Rabhi, and Mohsen Naderpour. 2018. Credit risk prediction in an imbalanced social lending environment. *CoRR* abs/1805.00801 (2018).

[31] Kun Niu, Zaimei Zhang, Yan Liu, and Renfa Li. 2020. Resampling ensemble model based on data distribution for imbalanced credit risk evaluation in P2P lending. *Information sciences* 536 (2020), 120–134.

[32] Mwanza Nkusu and Martin Mühleisen. 2011. Nonperforming Loans and Macrofinancial Vulnerabilities in Advanced Economies. *IMF Working Papers* 2011, 161 (2011).

[33] Rising Odegua. 2020. Predicting Bank Loan Default with Extreme Gradient Boosting. *CoRR* abs/2002.02011 (2020).

[34] Rahul Parhi and Robert D. Nowak. 2020. The Role of Neural Network Activation Functions. *IEEE Signal Process. Lett.* 27 (2020), 1779–1783.

[35] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[36] Lutz Prechelt. 2012. Early Stopping - But When? In *Neural Networks: Tricks of the Trade - Second Edition.* Lecture Notes in Computer Science, Vol. 7700. Springer, 53–67.

[37] Raul Rojas. 1996. *Neural Networks A Systematic Introduction* (1st ed. 1996.. ed.).

[38] Amanda Ross and Victor L Willson. 2017. Paired samples T-test. In *Basic and advanced statistical tests.* Brill Sense, 17–19.

[39] Carlos Serrano-Cinca, Begoña Gutiérrez-Nieto, and Luz López-Palacios. 2015. Determinants of Default in P2P Lending. *PloS one* 10, 10 (2015), e0139427–e0139427.

[40] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1 (2014), 1929–1958.

[41] Ting Sun and Miklos A Vasarhelyi. 2018. Predicting credit card delinquencies: An application of deep neural networks. *Intelligent systems in accounting, finance & management* 25, 4 (2018), 174–189.

[42] S M Taheri and G Hesamian. 2013. A generalization of the Wilcoxon signed-rank test and its applications. *Statistical papers (Berlin, Germany)* 54, 2 (2013), 457–470.

[43] Alexey Tsymbal. 2004. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin* 106, 2 (2004), 58.

[44] J. D Turiel and T Aste. 2020. Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society open science* 7, 6 (2020), 191649–191649.

[45] Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. 2016. Characterizing concept drift. *Data mining and knowledge discovery* 30, 4 (2016), 964–994.

[46] David Wierichs, Christian Gogolin, and Michael Kastoryano. 2020. Avoiding local minima in variational quantum eigensolvers with the natural gradient optimizer. *Physical Review Research* 2, 4 (2020), 043246.

[47] Xue Ying. 2019. An Overview of Overfitting and its Solutions. *Journal of physics. Conference series* 1168, 2 (2019), 22022.

[48] Hongke Zhao, Yong Ge, Qi Liu, Guifeng Wang, Enhong Chen, and Hefu Zhang. 2017. P2P Lending Survey: Platforms, Recent Advances and Prospects. *ACM Trans. Intell. Syst. Technol.* 8, 6 (2017), 72:1–72:28.

[49] Maciej Zieba and Wolfgang Karl Härdle. 2018. Beta-boosted ensemble for big credit scoring data. In *Handbook of Big Data Analytics.* Springer, 523–538.