

Spectral Clustering and Cost-Sensitive Deep Neural Network-Based Undersampling Approach for P2P Lending Data

Pankaj Kumar Jadwal, JK Lakshmi pat University, Jaipur, India

Sonal Jain, JK Lakshmi pat university, Jaipur, India

Basant Agarwal, Indian Institute of Information Technology, Kota, India

ABSTRACT

Peer-to-peer lending, also known as P2P lending, is the new generation loan disbursement process, where lenders and borrowers communicate through online services. Loans through P2P lending platforms are generally unsecured, due to the presence of borrowers with low credit scores. Lendingclub dataset, consisting of quantitative and qualitative information of borrowers from 2007 to 2011, is taken for the research. Machine learning models trained with such imbalanced dataset consists of biasing towards major class samples. The model performs significantly well on major class (safe borrowers) in terms of high precision but does not perform significantly well on minor class (defaulted) borrowers and provides low recall on minor class samples. To deal with the issue, a novel undersampling algorithm based on the combination of spectral clustering and cost sensitive deep neural network (SCSDNN) is proposed. Experimental results showcased the outstanding performance of the proposed technique, and it outperforms state of the art undersampling, oversampling and ensemble resampling techniques.

KEYWORDS

Class Imbalance, Clustering, Credit Risk, Deep Neural Network, Oversampling, P2P Lending, Resampling, Undersampling

1. INTRODUCTION

Peer to peer lending, also known as crowd lending, where lenders and borrowers can communicate directly without the involvement of financial bodies as the middleman (Ma and Wang 2016). Social lending market is spreading exponentially due to fast loan disbursement and less paperwork compared to traditional loan system (Guo et al., 2016). Borrowers with good credit can take the loan on lower interest rates, and lenders can also make more profits (Malekipirbazari and Aksakalli, 2015). Borrowers with average or lower credit score can communicate with lenders who are ready to provide them loans on high-interest rates (Serrano-Cinca and Gutiérrez-Nieto 2016).

There are several security concerns associated with P2P (Peer-to-Peer) lending market. Due to the absence of intermediate firm who is checking the authenticity of borrowers and lenders, it can be dangerous for lenders as well as for borrowers (Malekipirbazari and Aksakalli, 2015). Loan dispersed through P2P lending platforms contains risk of default and repayment delay. Borrowers and lenders communicates on common platform at various risk levels. High risk provides better returns but generates probable conditions of being defaulted. Therefore, precise prediction of the credibility of the borrowers is a crucial and significant issue in social lending. Statistical and machine learning

DOI: 10.4018/IJITWE.2020100103

Copyright © 2020, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

models can be utilized to predict the probable defaulters. Due to capability of generation of more optimal results, Machine learning models have already been outperformed statistical models. More accurate and specialized machine learning models are required to deal with such issues in P2P lending.

Several issues like long process time, limited lending money, not legal everywhere are associated with the P2P lending market. In the conventional process of loan disbursement, banks communicate directly with borrowers and assess the credibility of the borrowers, whereas, in P2P lending market, lenders and borrowers communicate directly with each other on the social lending platform. P2P lending datasets contains imbalance implicitly. Ratio of safe borrowers to defaulters is very high due to imbalance in the dataset. Machine learning models trained through such dataset provides good accuracy but accuracy is biased towards major class samples (Safe borrowers) and model predicts wrong class to most of the minor class samples. Therefore, traditional risk assessment machine learning models will not provide significant results to predict potential defaulters from borrowers.

In Imbalanced datasets, the ratio of the number of samples in major class to minor class, is significantly high. Machine learning models build on such datasets are biased towards majority class. Therefore, the precision of the major class is on the higher side and recall of minor class is on the lower side. Such kind of models will not be able to predict defaulters optimally (Lin et al., 2017; Xia & Liu 2017; Yijing et al. 2016). Although, Machine learning models trained through imbalanced dataset provides promising predictive accuracy, but due to less training of minor class samples compared to major class samples, ratio of predictive accuracy of minor class samples to major class samples is not close to one.

Therefore, to build machine learning models capable of providing high precision on major class samples and high recall on minor class samples, a novel undersampling approach SCCSDNN (Spectral Clustering and Cost Sensitive Deep Neural Network based Undersampling) is proposed. SCCSDNN is the combination of spectral clustering with cost sensitive deep neural network. Spectral clustering is applied to the major class samples and K clusters are obtained. Afterwards, K clusters are concatenated with minor class samples and K different datasets are obtained. Later, cost sensitive deep neural network model is built through k datasets, and dataset with the highest precision on major class and highest recall on minor class is chosen as the undersampled dataset.

The remaining of the paper is structured as follows. In Section 2, credit risk in social lending along with several resampling algorithms proposed to deal with the imbalance are discussed. In section 3, various quantitative and qualitative attributes of the lendingclub dataset are discussed. In section 4, the proposed undersampling algorithm (SCCSDNN) is discussed. In section 5, experimental results along with evaluation parameters are explored. The final section of the paper concludes the paper along the future directions.

2. RELATED WORK

2.1 Credit Risk Evaluation in Social Lending

P2P lending platform is providing vital services by providing loans to borrowers having wide range of credit scores and lenders are making more profits by targeting borrowers with low credit scores (Guo et al. 2016; Serrano-Cinca and Gutiérrez-Nieto 2016; Zhang and Chen 2017). In P2P lending, methods to evaluate the performance of credit risk models can be divided into two categories: statistical and machine learning methods. Credit risk evaluation involves a sophisticated mechanism to predict probable defaulters. Lenders in the P2P lending platform, do not have the specialized tools to assess the borrowers thoroughly. In terms of making maximum profits, sometimes lenders can opt wrong choice of borrowers (Byanjankar et al., 2015). Lenders need to balance the things between low to average risk and maximum profit (Guo et al., 2016). It can be done by introducing more specialized machine learning models. In P2P lending domain, various machine learning models like decision trees (Serrano-Cinca and Gutiérrez-Nieto, 2016), random forest (Malekipirbazari and Aksakalli, 2015),

and kernel methods (Guo et al. 2016) have been applied to predict the defaulters. Area under the curve(AUC), overall accuracy, precision on major class samples and recall on minor class samples are the various accuracy based measures used (Hand, 2009).

Instead of extensive literature available for credit risk evaluation of the traditional banking system, minimal research has been carried out in P2P lending domain. Emekter et al. (2015) proposed a credit risk evaluation model, derived from P2P lending dataset based on survival analysis and Logistic Regression. In the proposed framework, FICO (Fair Isaac Corporation) scores, Debt to income ratio (DTI), and revolving line utilization are concluded as most prominent attributes for credit risk evaluation. Guo et al. (2016) proposed an instance-based credit risk model using five features, namely, inquiries in the last six months, FICO (Fair Isaac Corporation) score, homeownership, funding amount and DTI, and linear programming is used for the optimal portfolio construction. Malekipirbazari and Aksakalli (2015) applied Random forest to P2P lending dataset and predictive outcomes outperformed the FICO scores and LC grades. Lending club assigns various grades, LC grades from A to H, based on borrower's quantitative and qualitative attributes. Although, the performance of the proposed model decreased under a high acceptance level. Malekipirbazari and Aksakalli (2015) applied following machine learning models: support vector machine, logistic regression, random forest and K nearest neighbor to draw optimal decision boundary between potential borrowers and defaulters and concluded the significance of the proposed model compared to LC grades and FICO grades. Serrano-Cinca and Gutiérrez-Nieto (2016) proposed profit scoring model instead of the credit scoring model, which concentrates more on the predictive profits by investment in P2P lending platform. Exhaustive chi-square automatic interaction detection (CHAID) Multivariate linear regression was applied to build the model. Therefore, further extensive research is still required to generate more specialized models that able to predict the defaulters in an optimal way. Table 1 represents synthesize related work through a comparison table.

Table 1. Studies of default prediction in P2P lending

Year	Author	Assigned Task	Algorithm	Dataset	Dataset properties
2020	(Liang & He, 2020)	Default prediction	Adaboost ensemble learning	P2P transactions dataset	1,20,000 borrowers
2019	(Chi, Ding, & Peng, 2019)	Prediction of expected return for lenders	Data driven robust model	Prosper dataset	17,000 borrowers
2018	(Wang, Jiang, Ding, Lyu, & Liu, 2018)	Behavioral scoring	Ensemble random forest	Time series	46,494
2018	(X. Ma et al., 2018)	Prediction of Defaulters	XGboost and LightGBM	P2P	569,339 transaction
2017	(Z. Li, Tian, Li, Zhou, & Yang, 2017)	Prediction of Defaulters	LSTM neural network	Time Series	600
2017	(Fu, 2017)	Prediction of Defaulters	random forest	P2P transaction	1,320,000
2017	(Y. Zhang, Yu, & Wang, 2017)	Prediction of Defaulters	Random forest, Support vector machine	P2P transaction	39,538
2017	(Chen, 2017)	Prediction of Defaulters	Logistic Regression	Time Series	3650
2017	(Huo, Chen, Chen, & others, 2017)	Prediction of Defaulters	Neural Networks, Logistic Regression	P2P transactions	4518
2017	(Ge, Feng, Gu, & Zhang, 2017)	Prediction of Defaulters	Logistic Regression	Social Media information	35457
2016	(Polena & Regner, 2018)	Prediction of number of defaulters	Nonlinear regression	P2P transactions	70,673
2016	(J. Li, Hsu, Chen, & Chen, 2016)	Prediction of defaulters	Logistic regression	P2P transactions	48,784
2015	(Byanjankar, Heikkil, & Mezei, 2015b)	Credit Scoring for default prediction	Neural Networks	P2P transactions	16,037

2.2 Class Imbalance Problem

Most of the real time classification problems contains class imbalance implicitly. In class imbalanced dataset, samples belongs to one class are significantly higher than those samples belongs to another class. Model trained on such dataset produces biased results (Chawla et al., 2004). Therefore, precision of major class samples is comparatively much higher than recall of minor class samples. Probability of classifying minor class samples into the desired class is much lower due to the imbalance between classes. Several methods (Cao et al., 2019; Wei and Ma, 2019) have been proposed by various researchers to deal with the issue: modify the loss function, modify the dataset (resampling) and ensemble methods.

2.3 Modify the Loss Function

In traditional machine learning models, both classes are allocated equal weights. To deal with the class imbalance, loss function can be modified. Modification of loss function for balancing the dataset is known as cost sensitive learning. Cost-sensitive learning tries to modify the cost function, and utilizes a function $C(p, t)$ (usually represented as a matrix) that denotes the overall cost of wrongly classified classes (Xia, Liu, and Liu 2017). This allows us to provide more weight to minor samples than major class samples, in hopes that this increases the true positive rate.

2.4 Modify the Dataset (Resampling)

Resampling methods can be partitioned into three broad areas: oversampling, undersampling and combination of undersampling and oversampling (Faris et al., 2020; Raghuwanshi and Shukla, 2020). Oversampling refers to integrate synthetic instances to the minor class samples to balance the dataset. State of the art oversampling techniques are random oversampling (Zhu et al., 2018), SMOTE (Chawla et al., 2002) and adaptive synthetic sampling (He et al., 2008). Random oversampling generates synthetic instances randomly and integrates them with the minor dataset to obtain the balanced dataset. SMOTE generates new synthetic samples based on K Nearest neighbors. Adaptive synthetic sampling generates synthetic samples to the minor class using density distribution. These oversampling algorithms have been applied to credit datasets to achieve balanced datasets, and further, several machine learning models have been applied to obtain better outcomes in terms of precision, recall and F-score (Haixiang et al., 2017). Undersampling methods remove samples from major class to achieve balanced dataset (Gong & Kim, 2017). Random undersampling and instance hardness threshold are most significant and widely used undersampling techniques. Random undersampling randomly picks samples from major class and removes them to generate balanced dataset (Zhu et al., 2018). The instance threshold algorithm discards the samples with a high probability of being classified into the wrong class (Smith et al., 2014).

2.5 Ensemble Methods

In this category, algorithms of oversampling and undersampling are merged to produce hybrid models which can take advantages from both techniques. SMOTE-Tomek and SMOTE-ENN are the most popular choices for ensemble methods (Zeng et al. 2016). In SMOTE-Tomek, oversampling is done by SMOTE and cleaning of data is done by removing Tomek links. Tomek links are the pair of examples which belongs to different classes and are K nearest neighbors of each other. SMOTE-ENN uses edited nearest neighbors for the balance in the dataset. Apart from that, Brown and Mues (Brown and Mues 2012), Abeysinghe et al. (Abeysinghe, Li, and He 2016) and Sanz et al. (Sanz et al. 2015) proposed various modified variants of ensemble methods to deal with the class imbalance problem.

2.6 Clustering for Undersampling

Clustering algorithms are the unsupervised methods where clusters are generated based on similarity among data samples. Generally, clustering is not applied on the dataset contains predictors and response variable, but clustering can be utilized in the preprocessing phase of the dataset where clustering may help in obtaining the balanced distribution of data samples. Clustering algorithms can play a significant role in the preprocessing phase of the dataset (Jadwal et al., 2017, Jadwal et al., 2019). Oversampling and undersampling both can be achieved with various clustering algorithms. Recently, various clustering-based methods have been proposed to deal with within class imbalance problem (Bunkhumpornpat et al., 2012; Cieslak and Chawla, 2008; Cieslak et al., 2006; Jo & Japkowicz 2004). Drummond et al. (2003) concluded through experimental results that undersampling on major class samples is far better than oversampling on minor class samples. Undersampling suffers from information loss. If the information loss issue can be solved, then the performance of the classifier can be improved in a significant way. In order to deal with the mentioned issue, Yen and Lee (2009)

proposed clustering based undersampling. In the proposed approach, initially, K Means clustering is applied to the major class sample, and k clusters are obtained. Later, K clusters are merged with minor class sample and K independent datasets are retrieved. Now, if a cluster contains more major class samples and very less minor class samples than cluster represents the characteristics of minor class and considered as major class. Finally, a suitable number of major class samples are picked and merged with a minor class sample to keep the cluster balanced. Rahman and Davis (2013) proposed another variant of clustering based undersampling where K Means clustering is applied on major class samples and major class samples are segmented into k clusters. Now, K clusters are merged with minor class samples and K independent datasets are generated. Later, FURIA machine learning model is built based on K datasets and dataset with the highest accuracy is chosen as the undersampled dataset.

Various methods (resampling, changing the cost function and ensemble methods) and their extensions have been proposed to deal with class imbalance problem (Cao et al., 2019; Faris et al., 2020; Raghuwanshi and Shukla, 2020; Wei and Ma, 2019). However, combination of such methods with clustering algorithms are also producing balanced distribution of the dataset. K Means clustering has already been applied to several imbalanced datasets and balanced distribution has been compared with various state of the art resampling algorithms (Yeh and Lien, 2009). K Means clustering can not be applied on the dataset which contains non globular clusters implicitly. P2P dataset is multidimensional large dataset, where K Means clustering can discover only globular clusters while spectral clustering can discover convex shape clusters. Clusters which cannot be identified with K Means clustering, spectral clustering is the significant choice in such scenario. Very limited research work is carried out for the combination of spectral clustering and deep feed forward neural networks to deal with class imbalance in lendingclub dataset. Cost sensitive deep neural network model is trained through K datasets obtained from the merge of K clusters to minor class samples. Spectral clustering is applied on major class dataset to obtain K clusters. To the extent of our knowledge, very limited literature is available, which addresses the imbalanced issue in the social lending context and provides solutions based on clustering algorithms.

3. DATASET

Due to the popularity and dominance of P2P lending platforms, the large amount of dataset has been generated, which contains quantitative and qualitative properties of the borrowers. Lendingclub is one of the most popular P2P lending platforms. Dataset of approved loans from 2007 to 2011 has been taken from lendingclub. Table 2 represents the description of the attributes used in Lending club dataset.

Columns like desc, url tends to be unnecessary for analysis, are removed from the dataset. Attributes having more than 50 percent of missing values are removed from the dataset. Features like id, member_id, funded_amnt, funded_amnt_inv, int_rate, sub_grade, emp_title, out_prncp, out_prncp_inv, total_pymt, total_pymt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt and issue_d are identified as leaky features and removed from the dataset. Average of Fico_range_low and FICO_range_high is calculated and Fico_range_average replaced the both attributes. Last_Fico_range_low and Last_FICO_range_high are identified as leaky attributes and removed from the dataset. Loan status is the response variable and contains following categories: Fully Paid, Charged Off, Does not meet the credit policy and Fully Paid, Does not meet the credit policy and Charged Off, Current, In Grace Period, Late (31-120 days), Late (16-30 days) and default. From the above categories, only the Fully Paid and Charged Off values describe the outcome of a loan. Fully_paid is encoded as '1' and charged_off is considered as defaults and encoded as '0'. Pymnt_plan is a binary attribute which denotes whether, payment plan has been put in place for the loan or not. Pymnt_plan contains 39238 samples as 'n' and only 1 sample as 'y'. Due to almost same value for all observations, pymnt_plan is not contributing towards machine learning model. Therefore, pymnt_plan is discarded from the dataset. Revol_util contains

International Journal of Information Technology and Web Engineering

Volume 15 • Issue 4 • October-December 2020

Table 2. Lending Club Dataset

Attributes	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
alt_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrowers self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are: 'W', 'F'
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issued	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts
member_id	A unique LC assigned id for the borrower member.
mo_sin_old_il_acct	Months since oldest bank installment account opened
mo_sin_old_rev_tl_op	Months since oldest revolving account opened
mo_sin_recent_rev_tl_op	Months since most recent revolving account opened
mo_sin_recent_tl	Months since most recent account opened
mort_acc	Number of mortgage accounts.
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
mths_since_recent_il	Months since most recent installment accounts opened
mths_since_recent_bc	Months since most recent bankcard account opened.
mths_since_recent_bc_dlq	Months since most recent bankcard delinquency
mths_since_recent_inq	Months since most recent inquiry.
mths_since_recent_rev_delinq	Months since most recent revolving delinquency.
next_pymnt_d	Next scheduled payment date
num_accts_ever_120_pd	Number of accounts ever 120 or more days past due
num_actv_bc_tl	Number of currently active bankcard accounts
num_actv_rev_tl	Number of currently active revolving trades
num_bc_sats	Number of satisfactory bankcard accounts
num_bc_tl	Number of bankcard accounts
num_il_tl	Number of installment accounts
num_op_rev_tl	Number of open revolving accounts
num_rev_accts	Number of revolving accounts
num_rev_tl_bal_gt_0	Number of revolving trades with balance >0
num_sats	Number of satisfactory accounts
num_tl_120dpd_2m	Number of accounts currently 120 days past due (updated in past 2 months)
num_tl_30dpd	Number of accounts currently 30 days past due (updated in past 2 months)
num_tl_90d_dpd_24m	Number of accounts 90 or more days past due in last 24 months
num_tl_op_past_12m	Number of accounts opened in past 12 months
open_acc	The number of open credit lines in the borrower's credit file.
open_acc_6m	Number of open trades in last 6 months
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
open_il_6m	Number of currently active installment trades
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
out_prncp	Remaining outstanding principal for total amount funded

continued on following page

Table 2. Continued

out_pncp_inv	Remaining outstanding principal for portion of total amount funded by investors
pct_tl_nvr_dlt	Percent of trades never delinquent
percent_bc_gt_75	Percentage of all bankcard accounts > 75% of limit.
policy_code	publicly available policy_code=1 new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
tax_liens	Number of tax liens
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all accounts
tot_hi_cred_lim	Total high credit/credit limit
total_acc	The total number of credit lines currently in the borrower's credit file
total_bal_ex_mort	Total credit balance excluding mortgage
total_bal_il	Total current balance of all installment accounts
total_bc_limit	Total bankcard high credit/credit limit
total_cu_tl	Number of finance trades
total_il_high_credit_limit	Total installment high credit/credit limit
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_pncp	Principal received to date
total_rev_hi_lim_A	Total revolving high credit/credit limit
url	URL for the LC page with listing data.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
verified_status_joint	Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.

50 missing values, and pub_rec_bankruptcies contains 697 missing values. Due to very low missing values in the preciously discussed attributes, these missing values are replaced by the mean of the corresponding column. Following variables home_ownership, verification status, em_length, term, addr_state, grade, purpose and title contains categorical values. Last_credit_pull_d, addr_state, title, earliest_cr_line are discarded from the dataset due to having DATE data type. Finally, labels of remaining categorical features are encoded, and one hot encoding is applied to all categorical features. Tables 3 and 4 shows the final picture of the preprocessed dataset.

4. PROPOSED UNDERSAMPLING APPROACH (COMBINATION OF COST SENSITIVE DEEP NEURAL NETWORK WITH SPECTRAL CLUSTERING FOR UNDER-SAMPLING (SCCSDNN))

Input: Class imbalanced dataset.

Step 1: Data preprocessing (Normalization, removal of outliers, missing value imputation, removal of leaky features).

Step 2: Spectral Clustering is applied on the Major class samples of the preprocessed dataset, and K clusters are obtained from the preprocessed dataset.

Step 3: Obtained clusters are concatenated with minor class samples, and K datasets are generated.

Step 4: Feed forward artificial neural network is applied on the K datasets (While doing the same, loss function is modified and minor class samples are provided x times for weight, where x= ratio of major class samples to minor class samples.

Table 3. Predictors of final preprocessed Dataset

Feature	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
at	A ratio calculated using the borrowers total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
revol_bal	Total credit revolving balance.
installment	The monthly payment owed by the borrower if the loan originates.
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file.
loan_amnt	The total amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
fico_average	The average boundary range the borrower's FICO at loan origination belongs to.
open_acc	The number of open credit lines in the borrower's credit file.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
grade	LC assigned loan grade.
inq_last_6months	The number of inquiries in past 6 months (excluding auto and mortgage inquiries).
term_60months	The number of payments on the loan. Values are in months and can be either 36 or 60.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.
purpose_debt_consolidation	A category provided by the borrower for the loan request.
verification_status_Not Verified	Indicates if income was not verified by LC.
home_ownership_MORTGAGE	The home ownership status is provided by the borrower during registration. Our value is MORTGAGE.
home_ownership_RENT	The home ownership status is provided by the borrower during registration. Our value is RENT.
term_36months	The number of payments on the loan. Value is 36.
verification_status_Source Verified	Shows borrowers having Source verified as verification status.
verification_status_Verified	Shows borrowers having verified as verification status.
purpose_small_business	Shows purpose of taking loan: Small business.
purpose_other	Shows purpose of taking loan: Other.
purpose_credit_card	Shows purpose of taking loan: Credit Card.
home_ownership_OWN	The home ownership status provided by the borrower during registration. Our value is OWN.
pub_rec_bankruptcies	Number of public record bankruptcies.
pub_rec	Number of derogatory public records.
purpose_home_improvement	The home ownership status provided by the borrower during registration. Our value is home improvement.
purpose_major_purchase	The home ownership status provided by the borrower during registration. Our value is major purchase.
purpose_car	The home ownership status provided by the borrower during registration. Our value is car.
purpose_medical	The home ownership status provided by the borrower during registration. Our value is medical.
purpose_wedding	The home ownership status provided by the borrower during registration. Our value is wedding.
purpose_moving	The home ownership status provided by the borrower during registration. Our value is moving.
purpose_educational	The home ownership status provided by the borrower during registration. Our value is educational.
purpose_house	The home ownership status provided by the borrower during registration. Our value is house.
purpose_vacation	The home ownership status provided by the borrower during registration. Our value is vacation.
purpose_renewable_energy	The home ownership status provided by the borrower during registration. Our value is renewable energy.
home_ownership_OTHER	The home ownership status provided by the borrower during registration. Our value is OWN.
home_ownership_NONE	The home ownership status provided by the borrower during registration. Our value is OWN.

Table 4. Final Dataset

N	Attributes	Fully Paid	Default	Imbalance Ratio
39226	40	83.18%	16.82%	4.94

Step 5: Dataset, providing optimal results (high precision on major class samples and high recall on minor class samples), is chosen as final dataset.

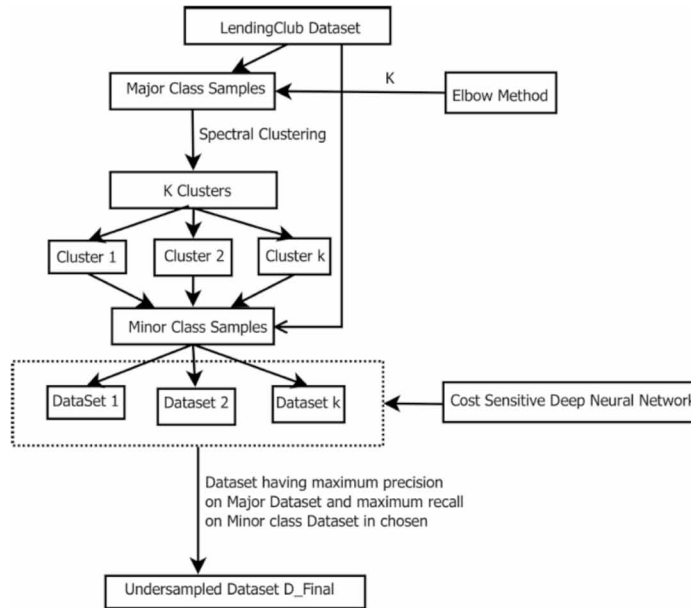
Output: Class balanced dataset.

4.1 Study Case

In the proposed approach, initially lendingclub dataset has been taken and preprocessed in terms of normalization, removal of outliers and leaky attributes etc. Now, dataset is segregated into major and minor class samples. Spectral clustering is applied on the major class samples and K number of clusters are obtained. Here K is confirmed to be 4 by elbow method. Later, K clusters are concatenated with minor class samples and k new datasets are generated. After generation of k datasets, deep neural network model (feed forward) is applied on the K datasets where, class weights are balanced for both classes. Precision from major class samples and recall from minor class samples is calculated. Dataset with optimal values for these parameters, is chosen as undersampled dataset.

To deal with the imbalance in the lendingclub dataset, a novel undersampling algorithm based on the combination of cost sensitive deep neural network with spectral clustering (SCCSDNN) is

Figure 1. SCCSDNN Undersampling Algorithm



proposed. Steps involved in the proposed approach of undersampling (SCCSDNN) is shown in Figure 1 and various stages are described below.

4.1.1 Preprocessing of the Dataset

In the preprocessing phase of the dataset, outliers are identified for each attribute. Lower threshold and the upper threshold is calculated for each attribute and values outside the lower and upper threshold are removed from the dataset. Lower and upper threshold are calculated using equation 2 and 3. Equation 1 represents the formula for normalizing the range of predictors. In equation 1, $\text{Min}(x)$ represents the minimum value of the corresponding feature x , $\text{Max}(x)$ represents the maximum value if the feature x , New_Max is taken as 1, and New_Min is taken as 0. In the equation 2, Q_{t1} represents the quarter 1 and Q_{t3} represents the quarter 3. Values, for any input feature x , in the range of Lower threshold and Upper threshold are accepted from the dataset:

$$\text{Norm}v' = \frac{v - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)} * \text{New_Max}(x) - \text{New_Min}(x) + \text{New_Min}(x) \quad (1)$$

$$\text{Lowerthreshold} = \left[Q_{t1} - 1.5 * (Q_{t3} - Q_{t1}) \right] \quad (2)$$

$$\text{Upperthreshold} = \left[Q_{t3} - 1.5 * (Q_{t3} - Q_{t1}) \right] \quad (3)$$

4.1.2 Spectral Clustering on the Preprocessed Dataset

- Initially, Major class samples of the preprocessed dataset is projected into R^n (where, R is a n dimensional vector);
- Affinity matrix B is defined, using a Gaussian Kernel K or say just an Adjacency matrix (i.e. $B_{ij} = \beta_{ij}$);
- Laplacian matrix B is constructed by subtracting adjacency matrix B from degree matrix D ;
- Eigenvalue problem is solved, like $AX = \alpha X$, where X is the eigenvector and α is the eigenvalue;
- K eigenvectors ($e_i, i=1$ to k) corresponding to K highest or lowest eigenvalues ($\alpha_i, i=1$ to k) are selected to define a K -dimensional subspace P^TLP ;
- Finally, Spectral clustering is applied to form K clusters in the subspace.

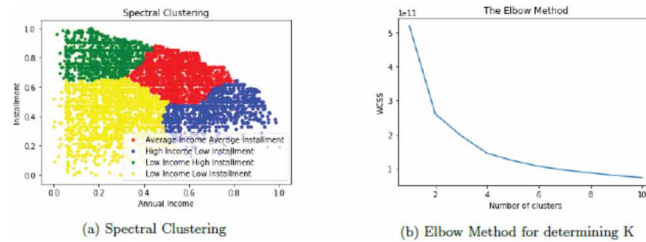
Preprocessed dataset is divided into major and minor class samples. Elbow method is applied on major class samples to retrieve the optimal number of clusters K . Elbow method is the most popular method for finding optimal value of K . In elbow method, two dimensional plot is generated between number of clusters and Cost function. Elbow point on the curve is picked and corresponding value of K is taken as number of clusters. As per elbow method, K is confirmed to be 4. Elbow method draws the plot between number of clusters and Cost function. Cost function is an objective function that showcases the relation between inter class similarity and intra class similarity. It is also denoted as within cluster sum of squared error. As number of clusters increases, cost function also decreases. We need to identify the elbow inside the plot and number of clusters corresponding to elbow represents the total number of clusters. While applying elbow method on the preprocessed dataset, value of k is retrieved as four. A plot between the within-cluster sum of squared error (Cost function) and the number of clusters is shown in figure 2(b) and Outcome of spectral clustering is showcased in the figure 2(a). Figure 2(a) shows four kinds of clusters.

Cluster with yellow color shows the segmented borrowers with low income and low instalment. Blue colored cluster shows the borrowers with high annual income and low instalment. Segmented borrowers having average income and average instalment are showcased with red colored cluster and green colored clusters show the aggressive borrowers. Finally, 4 clusters are merged with minor class samples and four new datasets are generated. Figure 2(a) represents four kind of clusters build from major class samples of the preprocessed dataset. For plotting, two most prominent features are used. Feature importance is calculated based on random forest. In random forest, several decision trees are build, and based on several decision trees, feature importance is decided. Installment and income are the two most prominent attributes. Major class samples represents the borrowers, which are not defaulted. We have blue clusters, in which, borrowers having high income and low installment are plotted. Borrowers with high income and low installment (blue cluster) have the minimum chances of being defaulted. That's why, it is denoted as safe group. If all clusters are compared with each other to obtain the safest group of clusters, the blue colored cluster will be on the winning side. After concatenation (Merging major class samples with minor class samples) of four clusters with minor class samples, four random datasets are generated. Now, separate feed forward artificial neural network model is applied on the each dataset and dataset produces highest sensitivity and highest specificity is chosen as the final balanced dataset.

4.1.3 Cost Sensitive Deep Neural Networks

Spectral clustering phase provides K number of clusters and K clusters are concatenated with minor class dataset. Therefore, K new datasets are generated. Further, the K datasets are splitted into training and testing set. The testing dataset comprises of 20 percent of the main dataset and training dataset consists 80 percent of the samples. Now, the feed forward deep neural network is trained from each of the training dataset and performance of the model is evaluated by calculating precision on Major class, recall on minor class and overall accuracy. Precision represents the percentage of positive

Figure 2.



samples which are correctly classified. Recall attempts to answer the question “What ratio of positive observations was classified correctly?”. Since, number of hidden layers are more than 2, proposed feed forward artificial neural network models is treated as feed forward deep neural network model. Hidden layer are intermediate layers between input and output layer. In the first hidden layer, number of nodes are taken ranging from 15 to 25 and optimal results are retrieved in 25. Therefore, 25 nodes are finalized for the first hidden layer. Rectifier activation function is used at the first hidden layer. Rectifier activation function is the commonly used activation function for hidden layers. The function returns zero if it receives negative input, otherwise, it returns the same value back for positive input. It is widely used activation function for the hidden layer due to easier to train and allows the model to account for interactions and non-linearity. Dropout is applied between the two hidden layers and between the last hidden layer and the output layer. Dropout represents the percentage of units being dropped from the hidden layers. Dropout rate of 25% is used. Second hidden layer contains same parameters as of the first hidden layer. In the output layer, sigmoid activation function is used. Sigmoid activation function is a powerful activation function, used for binary classification. Output of sigmoid function exists between 0 and 1. It is also called logistic function. Adaptive learning rate optimization algorithm (adam) is used as a optimizer, and it computes individual learning rates for various parameters. During the fitting of the model to the training and testing dataset, hyper parameters are assigned the following values. Number of epochs are 100, batch size is 32 and class_weight is assigned the ‘balance’ value which alters the weight that each training example carries when computing the loss. The number of samples in the major class is 83.18% of the original dataset, and the number of samples in minor class are 16.82%. Major class is assigned unit weight while minor class is assigned 4.94 times more weight than major class. Tenfold cross-validation is used to deal with bias vs. variance tradeoff. Dataset with the highest precision on the major class sample and maximum recall on the minor class sample is chosen as the undersampled dataset. Table 5 shows the hyper parameters along with values used in the model.

5. CLASSIFICATION RESULTS

Lendingclub dataset comprises quantitative and qualitative information of borrowers. The number of borrowers being defaulted is approximately five times lesser than the safe borrowers. In simple words, the dataset is highly imbalanced. Therefore, accuracy cannot be the only parameter to judge the effectiveness of the feed forward deep neural network model. Performance of the model will be significant if the precision of the major class is high and recall of minor class is also on the higher side. F-score is the harmonic mean of precision and recall. F-score provides a balance between classification of major and minor class samples. High precision for major class samples and high recall for minor class samples is required for getting F-score value closer to 1. Range of F-score is between 0 to 1. F-score is also an important parameter for the imbalanced dataset. Formula of the F-score is depicted below:

Table 5. Hyper parameters in deep Neural Network Model

Hyper Parameter	Value
Learning rate	.02
Batch Size	32
Number of epochs	100
Number of hidden layers	2
Number of hidden units per hidden layer	25
Dropout Regularization	.25

$$F\text{-score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

Precision, also known as, positive predictive value, is the proportion of predicted positive observations that truly are positive. Recall, also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate. The F score reaches the best value, meaning perfect precision and recall, at a value of 1. The worst F score, which means lowest precision and lowest recall, would be a value of 0.

Results of proposed algorithm (CSNNNSC) are compared with several state of the art underdamping, oversampling and combination of under and oversampling algorithm. Results are showcased in the Table 6. Proposed algorithm SCCSDNN is applied on four datasets (dataset 1, dataset 2, dataset 3 and dataset 4) generated through concatenation of minor class samples with four clusters: Cluster 0 (Low income and Low installment), Cluster 1 (High income and High installment), Cluster 2 (Low income and high installment) and Cluster 3 (Average income and Average installment) with minor class samples. Results obtained from SCCSDNN are compared with Random Oversampling (ROS), SMOTE (Synthetic Minority Oversampling technique), ADASYN (Adaptive synthetic Oversampling), Random Undersampling (RUS), Near-Miss1, Near-Miss2, Near-Miss3, SMOTE-ENN, and SMOTE-TOMEK. TP (True positive), TN (True Negative), FP (False Positive), FN (False Negative), Precision, recall, Accuracy and F-score are the evaluation parameters for the resampling algorithms. Proposed algorithm outperforms state of the art resampling techniques based on precision (.90), recall (.85), accuracy (.88) and F-score (.87).

6. CONCLUSION AND FUTURE WORK

Accurate evaluation of credit risk associated with borrowers, is a crucial and challenging issue on the P2P lending platforms. Due to the absence of middleman between lenders and borrowers, it can be more risky for borrowers and lenders. P2P lending datasets contains imbalance implicitly. Traditional machine learning models built through the P2P lending datasets, may not be able to classify the minor class borrowers in the desired class. Biasing towards major class makes it difficult to achieve a high recall of minor class samples.

SCCSDNN algorithms provides balanced distribution of the dataset using spectral clustering and cost sensitive feed forward deep neural network. Spectral clustering provides non globular clusters generated from Lending club dataset. Further, obtained K clusters are merged with minor class samples in order to obtain K datasets and Feed forward deep neural network is applied on K datasets individually. Before applying, cost sensitive reweighting assigns weights to minor and major class samples proportionally to inverse to their frequency. Classification results are compared with several state of the art resampling techniques like ROS (Random undersampling), SMOTE (Synthetic minority oversampling), ADASYN, RUS (Random undersampling), NearMiss-1, NearMiss-2, NearMiss-3,

Table 6. Classification results

Category	Resampling Algorithm	TP	FP	FN	TN	Precision	recall	Accuracy	F-score
OverSampling	ROS+ Cost Sensitive Deep Neural Network	6743	2109	3289	5474	0.72	0.76	0.69	0.78
	SMOTE+Cost Sensitive Deep Neural Network	7050	1802	205	8231	0.82	0.80	0.88	0.77
	ADASYN+Cost Sensitive Deep Neural Network	7431	1656	478	8263	0.83	0.82	0.88	0.80
Undersampling	RUS+Cost Sensitive Deep Neural Network	979	493	624	835	0.63	0.67	0.62	0.67
	NearMiss1+Cost Sensitive Deep Neural Network	972	500	263	1196	0.71	0.66	0.74	0.64
	NearMiss2+Cost Sensitive Deep Neural Network	1255	304	91	1368	0.82	0.81	0.87	0.78
	NearMiss3+Cost Sensitive Deep Neural Network	846	626	306	1153	0.65	0.57	0.68	0.56
Combined	SMOTEENN	4784	1017	273	4297	0.81	0.82	0.88	0.79
	SMOTETOMEK	6602	1698	235	8010	0.83	0.80	0.88	0.77
	SCCSDNN applied on dataset 1(cluster 0+Minor class samples)	917	554	135	2408	0.81	0.62	0.83	0.62
Proposed Approach	SCCSDNN applied on dataset 2(cluster 1+Minor class samples)	1305	224	249	2103	0.90	0.85	0.88	0.87
	SCCSDNN applied on dataset 3(cluster 2+Minor class samples)	1184	301	44	2162	0.88	0.80	0.91	0.78
	SCCSDNN applied on dataset 4(cluster 3+Minor class samples)	1116	363	57	1583	0.81	0.75	0.87	0.73
	ROS->Random Oversampling								
Abbreviation	SMOTE->Synthetic Minority Oversampling Technique								
	ADASYN->Adaptive Synthetic								
	ENN->Edited Nearest Neighbours								

SMOTE-ENN and SMOTE-TOMEK. Dataset which provides optimal results based on evaluation parameters (precision, recall, F-score), is taken as final balanced dataset. The proposed algorithm outperforms another state of the art resampling algorithm in precision, recall and F-score. SCCSDNN provides best results on Dataset 2, which is obtained through merging cluster 1 with minor class samples. Precision, recall and F-score are obtained as .90, .85, and .87 respectively.

Other powerful deep neural networks like CNN (Convolutional Neural Networks), RNN (Recurrent Neural Network) and LSTM (Long Short Term Memory Neural Network) may be applied as future work. Other popular clustering algorithms like density-based clustering can be applied to segment borrowers into optimal clusters.

REFERENCES

- Abeyasinghe, C., Li, J., & He, J. (2016). A classifier hub for imbalanced financial data. *Australasian Database Conference*, 476–479. doi:10.1007/978-3-319-46922-5_43
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3), 3446–3453. doi:10.1016/j.eswa.2011.09.033
- Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-based synthetic minority over-sampling technique. *Applied Intelligence*, 36(3), 664–684. doi:10.1007/s10489-011-0287-y
- Byanjankar, A., Heikkilä, M., & Mezei, J. (2015a). Predicting credit risk in peer-to-peer lending: A neural network approach. *Computational Intelligence, 2015 IEEE Symposium Series On*, 719–725.
- Byanjankar, A., Heikkilä, M., & Mezei, J. (2015b). Predicting credit risk in peer-to-peer lending: A neural network approach. *2015 IEEE Symposium Series on Computational Intelligence*, 719–725. doi:10.1109/SSCI.2015.109
- Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 1565–1576.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. doi:10.1613/jair.953
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1), 1–6. doi:10.1145/1007730.1007733
- Chen, Y. (2017). Research on the credit risk assessment of chinese online peer-to-peer lending borrower on logistic regression model. *DEStech Transactions on Environment, Energy and Earth Sciences*.
- Chi, G., Ding, S., & Peng, X. (2019). Data-driven robust credit portfolio optimization for investment decisions in P2P lending. *Mathematical Problems in Engineering*, 2019, 2019. doi:10.1155/2019/1902970
- Cieslak, D. A., & Chawla, N. V. (2008). Start globally, optimize locally, predict globally: Improving performance on imbalanced data. *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference On*, 143–152.
- Cieslak, D. A., Chawla, N. V., & Striegel, A. (2006). Combating imbalance in network intrusion datasets. *GrC*, 732–737.
- Drummond, C., & Holte, R. C. et al. (2003). C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. *Workshop on Learning from Imbalanced Datasets II*, 11, 1–8.
- Emekter, R., Tu, Y., Jirasakuldech, B., & Lu, M. (2015). Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. *Applied Economics*, 47(1), 54–70. doi:10.1080/00036846.2014.962222
- Faris, H., Abukhurma, R., Almanaseer, W., Saadeh, M., Mora, A. M., Castillo, P. A., & Aljarah, I. (2020). Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: A case from the Spanish market. *Progress in Artificial Intelligence*, 9(1), 31–53. doi:10.1007/s13748-019-00197-9
- Fu, Y. (2017). Combination of random forests and neural networks in social lending. *Journal of Financial Risk Management*, 6(4), 418–426. doi:10.4236/jfrm.2017.64030
- Ge, R., Feng, J., Gu, B., & Zhang, P. (2017). Predicting and deterring default with social media information in peer-to-peer lending. *Journal of Management Information Systems*, 34(2), 401–424. doi:10.1080/07421222.2017.1334472
- Gong, J., & Kim, H. (2017). RHSBoost: Improving classification performance in imbalance data. *Computational Statistics & Data Analysis*, 111, 1–13. doi:10.1016/j.csda.2017.01.005
- Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, 249(2), 417–426. doi:10.1016/j.ejor.2015.05.050

- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73, 220–239. doi:10.1016/j.eswa.2016.12.035
- Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103–123. doi:10.1007/s10994-009-5119-5
- He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Neural Networks, 2008. IJCNN 2008 (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference On*, 1322–1328.
- Huo, Y., Chen, H., & Chen, J. (2017). Research on personal credit assessment based on neural network-logistic regression combination model. *Open Journal of Business and Management*, 5(02), 244–252. doi:10.4236/ojbm.2017.52022
- Jadwal, P. K., Jain, S., Gupta, U., & Khanna, P. (2017). K-Means clustering with neural networks for ATM cash repository prediction. *International Conference on Information and Communication Technology for Intelligent Systems*, 588–596.
- Jadwal, P. K., Jain, S., Gupta, U., & Khanna, P. (2019). Clustered Support Vector Machine for ATM Cash Repository Prediction. In *Progress in Advanced Computing and Intelligent Engineering* (pp. 189–201). Springer. doi:10.1007/978-981-13-1708-8_18
- Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *SIGKDD Explorations*, 6(1), 40–49. doi:10.1145/1007730.1007737
- Li, J., Hsu, S., Chen, Z., & Chen, Y. (2016). Risks of p2p lending platforms in china: Modeling failure using a cox hazard model. *Chinese Economy*, 49(3), 161–172. doi:10.1080/10971475.2016.1159904
- Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, 74, 105–114. doi:10.1016/j.eswa.2017.01.011
- Liang, K., & He, J. (2020). Analyzing credit risk among Chinese P2P-lending businesses by integrating text-related soft information. *Electronic Commerce Research and Applications*, 40, 100947. doi:10.1016/j.elerap.2020.100947
- Lin, W.-C., Tsai, C.-F., Hu, Y.-H., & Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409, 17–26. doi:10.1016/j.ins.2017.05.008
- Ma, H.-Z., & Wang, X.-R. (2016). Influencing factor analysis of credit risk in P2P lending based on interpretative structural modeling. *Journal of Discrete Mathematical Sciences and Cryptography*, 19(3), 777–786. doi:10.1080/09720529.2016.1178935
- Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31, 24–39. doi:10.1016/j.elerap.2018.08.002
- Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42(10), 4621–4631. doi:10.1016/j.eswa.2015.02.001
- Polena, M., & Regner, T. (2018). Determinants of borrowers default in P2P lending under consideration of the loan risk class. *Games*, 9(4), 82. doi:10.3390/g9040082
- Raghuwanshi, B. S., & Shukla, S. (2020). SMOTE based class-specific extreme learning machine for imbalanced learning. *Knowledge-Based Systems*, 187, 104814. doi:10.1016/j.knosys.2019.06.022
- Rahman, M. M., & Davis, D. (2013). Cluster based under-sampling for unbalanced cardiovascular data. *Proceedings of the World Congress on Engineering*, 3, 3–5.
- Sanz, J. A., Bernardo, D., Herrera, F., Bustince, H., & Hagrass, H. (2015). A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial applications with imbalanced data. *IEEE Transactions on Fuzzy Systems*, 23(4), 973–990. doi:10.1109/TFUZZ.2014.2336263
- Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89, 113–122. doi:10.1016/j.dss.2016.06.014

Smith, M. R., Martinez, T., & Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine Learning*, 95(2), 225–256. doi:10.1007/s10994-013-5422-z

Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, 27, 74–82. doi:10.1016/j.elerap.2017.12.006

Wei, C., & Ma, T. (2019). *Improved sample complexities for deep networks and robust classification via an all-layer margin*. ArXiv Preprint ArXiv:1910.04284

Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications*, 24, 30–49. doi:10.1016/j.elerap.2017.06.004

Yen, S.-J., & Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3), 5718–5727. doi:10.1016/j.eswa.2008.06.108

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94, 88–104. doi:10.1016/j.knosys.2015.11.013

Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. (2016). Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, 225–228. doi:10.1109/ICOACS.2016.7563084

Zhang, K., & Chen, X. (2017). Herding in a P2P lending market: Rational inference OR irrational trust? *Electronic Commerce Research and Applications*, 23, 45–53. doi:10.1016/j.elerap.2017.04.001

Zhang, Y., Yu, C., & Wang, L. (2017). Temperature exposure during pregnancy and birth outcomes: An updated systematic review of epidemiological evidence. *Environmental Pollution*, 225, 700–712. doi:10.1016/j.envpol.2017.02.066 PMID:28284544

Zhu, B., Baesens, B., Backiel, A., & Vanden Broucke, S. K. L. M. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. *The Journal of the Operational Research Society*, 69(1), 49–65. doi:10.1057/s41274-016-0176-1

Pankaj Kumar Jadwal has over ten years of academic experience and is currently associated with Swami Keshvanand Institute of Technology Management & Gramothan, Jaipur. Before joining SKIT, he has worked with Poornima University. Pankaj Kumar Jadwal has various publications to her credit and has presented research papers at National and International conferences. Pankaj Kumar Jadwal is life member Computer Society of India. His area of interest for Research is applications of machine learning, data mining and analysis of algorithms.

Sonal Jain has over 19 years of academic experience and is currently associated with JK Lakshmipat University, Jaipur. She has several publications to her credit. She has co-authored several books including Personal Computer Software (Wiley India), Simplifying C (Dreamtech Publications) and six textbooks for Gujarat State Education Board. She has also been visiting faculty at the University of J. Seyle, Slovenia, Hungary and also visited Amsterdam University of Applied Sciences under faculty exchange program at JKL. She is also the recipient of Young Scientist Award by Venus International Foundation, Chennai, and received Vaigyanik Drashtikon Award at 4th Rajasthan Science Congress.