A Boosted Decision Tree Model for Predicting Loan Default in P2P Lending Communities

Semiu A, Akanmu Abdul Rehman Gilal

Abstract Loan Default Prediction For Social Lending Is An Emerging Area Of Research In Predictive Analytics. The Need For Large Amount Of Data And Few Available Studies In The Current Loan Default Prediction Models For Social Lending Suggest That Other Viable And Easily Implementable Models Should Be Investigated And Developed. In View Of This, This Study Developed A Data Mining Model For Predicting Loan Default Among Social Lending Patrons, Specifically The Small Business Owners, Using Boosted Decision Tree Model. The United States Small Business Administration (Usba) Publicly-Available Loan Administration Dataset Of 27 Features And 899164 Data Instances Was Used In 80:20 Ratios For The Training And Testing Of The Model. 16 Data Features Were Finally Used As Predictors After Data Cleaning And Feature Engineering. The Gradient Boosting Decision Tree Classifier Recorded 99% Accuracy Compared To The Basic Decision Tree Classifier Of 98%. The Model Is Further Evaluated With (A) Receiver Operating Characteristics (Roc) And Area Under Curve (Auc), (B) Cumulative Accuracy Profile (Cap), And (C) Cumulative Accuracy Profile (Cap) Under Auc. Each Of These Model Performance Evaluation Metrics, Especially Roc-Auc, Showed The Relationship Between The True Positives And False Positives That Implies The Model Is A Good Fit.

Keywords: loan default prediction, peer-to-peer lending, boosted decision tree, data mining

I. INTRODUCTION

This study aims to develop a data mining model for predicting loan default among social lending patrons, otherwise known as peer-to-peer (P2P) lending. Loan default prediction has been extensively studied and varieties of predictive models have been proposed. Notably, these past studies mainly focused on customers' loan among banks and other conventional financial institutions [1]–[4]. The rapid rise of P2P lending [3], [5], however, is necessitating development of decision support system, specifically loan default prediction, to help in making reliable lending decision.

Loan default prediction for social lending is just getting academic and practitioners' attentions as evident in the availability of few studies when compared to the conventional financial institutions' loans. The studies accessed and reviewed for this project used Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) [3] and Random Forest [5], among others. The need for large amount of data and few available studies in the current loan default prediction models for social lending suggest that other viable and easily implementable models should be investigated and developed.

Boosted Decision Tree is an improved version of the basic decision tree classification algorithm. It is achieved through an ensemble of decision trees using boosting [6]. Boosting implies that, while using more than one decision tree as the classification algorithm during the training phase, each tree is dependent on the prior trees [7], [8]. Bagging and Random Forests are other classic methods of creating ensemble models. These ensemble methods decrease the variance of a single estimate because they combine several estimates from different models. However, boosting generates a combined model with lower errors through optimization [8]. Random Forest is just a decision tree variant of bagging [9].

Boosted Decision Tree is a supervised learning method. Its dataset must be labeled with columns containing numerical values. The algorithm learns better through the fitting of the residual of the trees preceding it. Therefore, boosted decision tree ensemble often improves the accuracy of the algorithm with lesser risk of coverage by optimizing tree using arbitrary differentiable loss function [6]–[8].

There are three simple steps in Boosting. These are (a) initialization of the model, (b) new model fitting based the residual of the previous step, and (c) combination of the previous models for optimization.

An initial model F0 is defined to predict the target variable y. A residual y – F0 is therefore associated. Then, if h1 be the new model fitting, that is, the residual from the previous step, F0 and h1 would give F1 which is the boosted version of F0.

The mean squared error from F1 will be lower than that from F0:

$$F1(x) \leq F0(x) + h1(x) \ldots\ldots (i)$$

In improving the performance of $F1$, a new model $F2$ can be modelled using the residuals of $F1$:

$$F2(x) \leq F1(x) + h2(x) \ldots\ldots (ii)$$

In iterations m, till the residuals have been completely minimized as much as possible:

$$Fm(x) \leq Fm-1(x) + hm(x) \ldots\ldots (iii)$$

The addictive learners of the decision tree ensured information is impacted and the errors due to variance and bias are reduced significantly.

This study developed Boosted Decision tree induction model [10], with data pre-processing steps and techniques that attend to the specifics of loan administration in social lending.

Revised Manuscript Received on October 15, 2019

Semiu A. Akanmu, Department of Computer Science North Dakota State University
Fargo, USA semiu.akanmu@ndsu.edu

Abdul Rehman Gilal, Department of Computer Science Sukkur IBA University Sindh, Pakistan
a-rehman@iba-suk.edu.pk

The model presents an accurate, yet simplified, loan default prediction opportunity with less data demand for social lending communities. It also contributes to studies on loan default prediction specifically, and application of artificial intelligence in finance generally.

The next section presents the summary of past related studies. Later, the processes involved in building this model, the data cleaning, feature selection, model testing, and their associated findings are presented. The final section describes the limitations of the study and suggests plan for future work.

II. RELATED STUDIES

Systematically selected twenty (20) past related studies are reviewed to attempt a scholarly footing for this study. The main inclusion criteria for articles in this systematic literature review (SLR) is: the study must have worked on loan default prediction, irrespective of the domain. The limit of 20 studies is set simply because of the limited time for the execution of this study.

Themes identified from the reviewed studies can be summarily categorized into two (2). These are: (i) application areas and the problems, and (ii) data features and the machine learning models.

a) Application areas and the Problem

Peer to peer lending [3], [5], [7], [11], [12], commercial banking [2], [4], [13]–[15], insurance [6], agriculture [16], mortgage [17], and small and medium enterprises (SMEs) [8], [12] are different application areas of loan default prediction studies. However, because of certain specific problems and different available dataset, the studies employed different machine learning models. Thus, the results are comparatively different.

The need to understand the associated risk to loan administration [2] and predict reliability of customers' loyalty in retailing banking [4] are examples of other application problems addressed. It is also shown that most studies on loan default prediction are recorded in the mainstream commercial banking, with few in agriculture loan, mortgage and SMEs, and P2P.

The P2P is used as the application area of interest in this study, not only because past studies have sparingly developed loan prediction models for the peculiarity of the area, but also for its growing adoption among informal lending organizations [3], [5]–[8]. The nature of the applicants' profiles and the attending data features in these application areas are also different. Therefore, due attention is required in characterizing machine learning models and their application research areas.

b) Data Features and the Machine Learning Models

The sizes of the datasets and the features present vary significantly in studies. This, as shown, suggests the type of applicable machine learning model, especially where the size of the available dataset determines the performance. The data features used in building the models reviewed range from eight (8) [2], seventeen (17) [4] to twenty-four (24) [13].

Studies with considerable large datasets are Bagherpour [18] of about 20 million loan observations between 2001 – 2006, and Rivet [6] which used 1 million loans data. Yang et al. [3], which employed a dataset of 10,000 anonymous users of a peer lending service, Sivasree and Sunny [4] of 4520 data instances for loan credibility prediction system, and Hassan and Abraham [13] with 1000 cases are others with small data sizes.

In these, K-Nearest Neighbors [18], Decision Tree [4], [6], [18], Random Forest [18], logistic regression [5], [6], Support Vector Machine [18], Artificial Neural Network [3], [13], [19], and Naïve Bayes [2], [11] are the recorded machine learning models. Asides from studies that did comparative study for loan default prediction performances [2], [5], [18], others employed single [4], [17] and hybrid/ensemble models [3], [6], [8], [9]. Hybrid and ensemble are mostly used. In these studies, random forest and boosted decision trees, which are bagging and boosting forms of decision trees respectively, recorded significant high performances [4], [8], [9], [14]. The boosted decision tree is used in this study because, as earlier described, it has the potential of performing better than random forest (bagging ensemble) and the basic decision tree classifier.

The next section describes how the boosted decision tree model is built. This includes data collection, data cleansing and features selection, cross-fold validation and testing of the model.

III. EXPERIMENTAL RESULTS AND FINDINGS

This section describes the process of data collection and selection, its preprocessing stage, feature selection and the model building. The testing of the boosted decision tree induction model, the comparison with basic decision tree classifier and other performance evaluation metrics are also explained.

a) Datasets Collection

Publicly-available dataset for social lending communities accessed are United States Small Business Administration (USBA) [1] and Imperial College London Kaggle competition Dataset [20]. The USBA dataset is unprocessed with all its data features still retaining its categorical descriptions and data instances in forms that are uncompliant with machine learning model building. On the other hand, the Imperial College's is processed, and thus has its features converted to numerical labels and data instances are model-compliant.

This study preferred the USBA dataset even though the Imperial College's has more features (771, as against the latter of 27), but the latter has more instances (899, 164) as against 211,000 of the former. The justifications for this are (a) having unprocessed data allows a comprehensive coverage of the model building process, where the data preprocessing and feature engineering are significant phases, and (b) knowing what constitutes each of the data features (which is impossible in the Imperial College's dataset) will support the implementation of the model built in a web application development as suggested as future work.

The USBA dataset is a real-life data which illustrates loan administration experience within a social circle of US small businesses.

Small businesses have been a primary source of job creation in the United States. It fosters small business formation and growth by creating job opportunities and reducing unemployment [1]. Table 1 presents the description of the features in the dataset.

Table 1. Description of the 27 features in the datasets [1].

Feature Name	Data type	Description
LoanNr_ChkDgt	Text	Identifier – Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American Industry Classification Code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in month
NoEmp	Number	Number of business employees
NewExist	Text	1= Existing business, 2 = New business
CreateJob	Number	Number of job created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, 1= Franchise, (00000 or 00001) = No franchise
UrbanRural	Text	1 = Urban, 2= Rural, 0 = Undefined
RevLineCr	Text	Revolving line of credit: Y= Yes, N= No.
LowDoc	Text	Low Doc Loan program: Y= Yes, N= No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged off amount
GrApprv	Currency	Gross amount of loan approved by bank
SBA_Apprv	Currency	SBA's guaranteed amount of approved loan

b) Data Preprocessing

In this stage, three data preprocessing steps are taken. These are (a) removing and imputing missing values from the dataset, (b) getting the categorical data into shape for machine learning through class mapping, and (c) selecting relevant features for model building. The data features with missing values and the number of rows affected are presented in Table 2.

Table 2. Data Features with missing values and the n the datasets [1].

Data Feature	Are Missing Values Present?	Number of missing values
LoanNr_ChkDgt	No	0
Name	Yes	14
City	Yes	30
State	Yes	14
Zip	No	0
Bank	Yes	1559
BankState	Yes	1566
NAICS	No	0
ApprovalDate	No	0
ApprovalFY	No	0
Term	No	0
NoEmp	No	0
NewExist	Yes	136
CreateJob	No	0
RetainedJob	No	0
FranchiseCode	No	0
UrbanRural	No	0
RevLineCr	Yes	4528
LowDoc	Yes	2582
ChgOffDate	Yes	736465
DisbursementDate	Yes	2368
DisbursementGross	No	0
BalanceGross	No	0
MIS_Status	Yes	1997
ChgOffPrinGr	No	0
GrApprv	No	0
SBA_Apprv	No	0

The features with missing values are both of numerical or date/time (e.g. ChgOffDate, etc.) and categorical (e.g. Name, Bank, etc.) data types. The option of dropping features or rows with missing values is highly disadvantageous because, asides from losing huge data instances, there is possibility of losing hypothetically important feature (e.g. NewExist, MIS_Status, etc.) that has missing values. Therefore, appropriate imputation techniques are required. However, certain categorical data, specifically RevLineCr, LowDoc and MIS_Status are replaced with correspondingly-mapped numerical values because of their potentials of influencing the machine learning model performance. Notably, the MIS_Status is the predicted class label. The missing data of other features, such as Name, Bank, BankState, etc., are replaced with the most frequent value in their respective columns.

c) Feature Engineering, Data Splitting and Boosted Decision Tree Model Building

Feature engineering consists of feature scaling and feature selection. Feature scaling is ensuring that features in the dataset are on the same scale using normalization or standardization techniques. However, since Decision tree is the machine learning model, feature scaling is not required. Feature selection, on the other hand, is one of the practical ways of avoiding model overfitting [21]. This study used the feature selection method.

After the preprocessing of the dataset, the target variable, MIS_Status, is assigned to a separate data frame. Then, intuitively insignificant variables, such as Name, City, State, Bank, Bank State, NAICS, are deleted, and the predicting data variables are assigned to a separate data frame.

This illustrates the feature selection strategy of this study since the remaining predicting features (16 in number) is not too many for feature embedding, among others.

The data frames for the target and predicting variables are then split into the train and test set using the *train_test_split* function from scikit-learn's *cross_validation* submodule in 80:20 percent ratio. The choice of the train to test set ratio, as suggested by Sebastian [21] fits well for the type of dataset size used in this study.

d) Model Testing and Evaluation

The basic decision tree classifier and boosted decision tree models are developed, tested and evaluated [22], [23]. The accuracies for the decision tree classifier and boosted decision tree model, when tested, are 98% and 99% respectively. Asides from the great performances of both models, the better performance of the boosted decision tree supported the theoretical assertion that boosting improves weaker learning models. Figure 1 presents accuracy values of the decision tree classifier and boosted decision tree models.

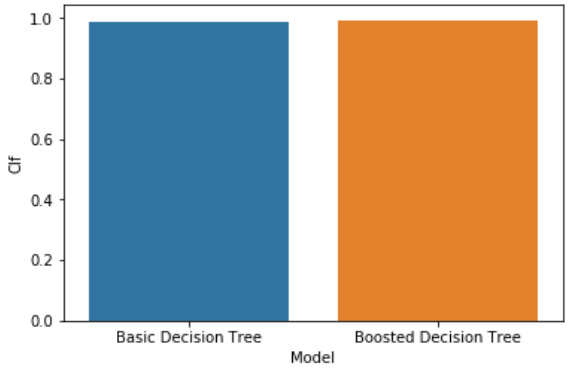

Figure 1: Accuracy values for the Basic Decision Tree and Boosted Decision Tree

The boosted decision tree model is evaluated using (a) Receiver Operating Characteristics and its Area Under Curve (AUC), (b) Cumulative Accuracy Profile (CAP) Curve and (c) Cumulative Accuracy Profile (CAP) Analysis Using AUC.

Receiver Operating Characteristic (ROC) Curve

The Receiver Operating Characteristic (ROC) Curve is used to evaluate the performance of the boosted decision tree model, with the True Positive Rate being plotted against the False Positive Rate. The model's Area Under Curve (AUC) is given as 100%. ROC illustrates the diagnostic ability of the binary classifier system when its discrimination threshold is varied. In this study, the class labels are 1 = Pay-In-Full, that is for those that paid in full (did not default), and 0 = CHGOFF for those that defaulted. It implies that the model is perfectly fit for the classification of the two class labels.

Cumulative Accuracy Profile (CAP) Curve

The Cumulative Accuracy Profile (CAP)Curve analyzes the effectiveness of identifying all data instances of a class based on the minimum number of trials. This is used by evaluating how effective can the Boosted Decision tree identifies the class labels of those that did not default. Two models, random and perfect, were plotted. The random model illustrates the fact that the class 1.0 will be detected linearly. The perfect model, on the other hand, detects all the class 1.0 data points in the same number of tries. The ROC-AUC and CAP (with random and perfect model), and the CAP (with random, perfect and boosted decision tree models) are presented in figures 2, 3 and 4 respectively.

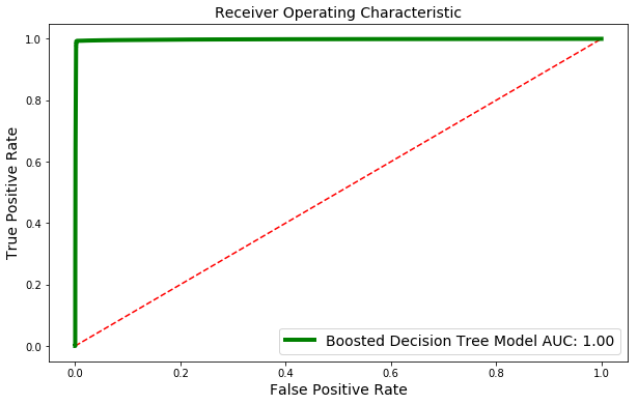

Figure 2: ROC-AUC for the Boosted Decision Tree Model

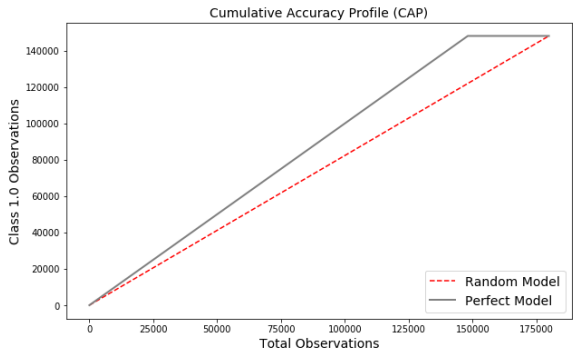

Figure 3: Random and Perfect Models for the CAP curve

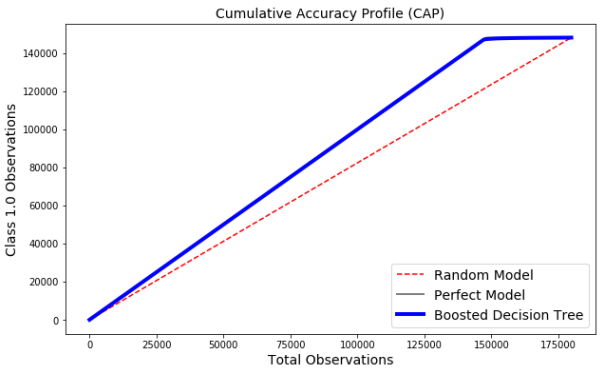

Figure 4: Random, Perfect, and Boosted Decision Tree Models for the CAP curve

IV. CONCLUSION

This study developed a boosted decision tree model for predicting loan default in P2P lending communities. It aimed at improving lending decision making, specifically in social lending which had not received adequate research attention compared to the conventional banking system. The USBA dataset of 27 features and 899164 is used.

After feature selection, a total of 16 features was used in the model development. The decision tree classifier and boosted decision tree accuracies recorded were 98% and 99% which suggested a strong model. The model was finally evaluated using a) Receiver Operating Characteristics (ROC) and Area Under Curve (AUC), (b) Cumulative Accuracy Profile (CAP), and (c) Cumulative Accuracy Profile (CAP) under AUC. Each of these model performance evaluation metrics showed that the model is a good fit.

However, it is noteworthy that the model evaluation is not exhaustive. But due to limited time, this study is unable to explore plot analysis of the CAP curve. To this end, the effect of class imbalance would be checked and the model fine tuned. It is also important to experiment the performance of this model, especially with the specifics of its training data, in comparison with other classification models like Support Vector Machine, Naïve Bayes, and Random Forest. Future work should also work on the implementation of the model built in a web application development for the use of the loan administrators.

REFERENCES

1. M. Li, A. Mickel, and S. Taylor, "Should This Loan be Approved or Denied?: A Large Dataset with Class Assignment Guidelines," J. Stat. Educ., vol. 26, no. 1, pp. 55–66, 2018.
2. J. H. Aboobyda and M. A. Tarig, "Developing Prediction Model Of Loan Risk In Banks Using Data Mining Machine Learning and Applications," An Int. J., vol. 3, no. 1, pp. 1–9, 2016.
3. Z. Yang, Y. Zhang, B. Guo, B. Y. Zhao, and Y. Dai, "DeepCredit: Exploiting User Cickstream for Loan Risk Prediction in P2P Lending," in Twelfth International AAAI Conference on Web and Social Media, 2018.
4. M. S. Sivasree, "Loan Credibility Prediction System Based on Decision Tree Algorithm," Int. J. Eng. Res. Technol., 2015.
5. A. Namvar, M. Siami, F. Rabhi, and M. Naderpour, "Credit risk prediction in an imbalanced social lending environment," arXiv Prepr. arXiv1805.00801, 2018.
6. M. T. Maxime Rivet, Marc Thibault, "Dynamic Loan default prediction," 2017.
7. Z. Alomari and D. Fingerman, "Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications."
8. A. Pandit, "DATA MINING ON LOAN APPROVED DATSET FOR PREDICTING DEFAULTERS," Rochester Institute of Technology, 2016.
9. K. R. Rawate and P. P. A. Tijare, "REVIEW ON PREDICTION SYSTEM FOR BANK LOAN CREDIBILITY," Int. J. Adv. Eng. Res. Dev., vol. 4, no. 12, pp. 860–867, 2017.
10. P.-N. Tan, Introduction to data mining. Pearson Education India, 2018.
11. C. Jiang, Z. Wang, R. Wang, and Y. Ding, "Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending," Ann. Oper. Res., vol. 266, no. 1–2, pp. 511–529, 2018.
12. A. Bhimani, M. A. Gulamhussen, and S. da R. Lopes, "The role of financial, macroeconomic, and non-financial information in bank loan default timing prediction," Eur. Account. Rev., vol. 22, no. 4, pp. 739–763, 2013.
13. A. K. I. Hassan and A. Abraham, "Modeling consumer loan default prediction using ensemble neural networks," in 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE), 2013, pp. 719–724.
14. A. Goyal and R. Kaur, "Loan Prediction Using Ensemble Technique.," Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 3, pp. 523–526, 2016.
15. G. Sudhamathy, "Credit risk analysis and prediction modelling of bank loans using R," Int. J. Eng. Technol, vol. 8, pp. 1954–1966, 2016.
16. O. O. Odeh, A. M. Featherstone, and D. Sanjoy, "Predicting Credit Default in an Agricultural Bank: Methods and Issues," 2006.
17. Y. Ghulam, K. Dhruva, S. Naseem, and S. Hill, "The interaction of borrower and loan characteristics in predicting risks of subprime automobile loans," Risks, vol. 6, no. 3, p. 101, 2018.
18. A. Bagherpour, "Predicting mortgage loan default with machine learning methods," Univ. California/Riverside, 2017.
19. D. Goriunov, K. Venzhyk, and others, "Loan default prediction in Ukrainian retail banking," 2007.
20. "Imperial College London Kaggle competition Dataset." [Online]. Available: https://www.kaggle.com/c/loan-default- prediction/data. [Accessed: 24-Aug-2018].
21. S. Raschka and V. Mirjalili, Python machine learning. Packt Publishing Ltd, 2017.
22. A. R. Gilal, J. Jaafar, L. F. Capretz, M. Omar, S. Basri, and I. A. Aziz, "Finding an effective classification technique to develop a software team composition model," J. Softw. Evol. Process, vol. 30, no. 1, 2018.
23. J. Abdul Rehman Gilal, Mazni Omar, Ruqaya Gilal, Ahmed Waqas, Sharjeel Afridi and Jaafar, "A Decision Tree Model for Software Development Teams," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 5S, pp. 241–245, 2019.

AUTHORS FRFILE

Semiu A., Akanmu is currently a software research engineer with Dickinson Research Extension center, North Dakota State University, US. He is also working on a doctoral research with interest in ontology modelling for software security. He had his Bachelor of Science (BSc) in computer science from Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria, in 2008, Master of Science (MSc) and a Doctor of Philosophy (Ph.D.) in Information Technology (IT) from Universiti Utara Malaysia, Sintok, Malaysia, in 2013 and 2016, respectively.

Abdul Rehman Gilal, is a faculty member of Computer Science department at Sukkur IBA University, Pakistan. He has earned Doctor of Philosophy (Ph.D.) in Information Technology from Universiti Teknologi Petronas (UTP), Malaysia. He has been mainly researching in the field of software project management for finding the effective methods of composing software development teams. Based on his research publication track record, he has contributed in the areas of human factor in software development, complex networks, databases and data mining, programming and cloud computing.