Prediction of Loan Approval via the Different Machine Learning Models

Bibarys Mussagaliyev[1], Nurken Kadirov[2] and Amin Zollanvari[3]

***Abstract*— Loan prediction that helps in estimating the ability of customers repayment is considered as one of the most significant issues of credit institutions, as the number of loan applicants is growing dramatically. For the recent years the concepts and techniques of machine learning have been significantly developed in modern computer science field. Thus, there have been presented and implemented different methods of classification, which can play a significant role during the implementation of models for evaluating the applicants' repayment capability. In this project, the following classifiers such as Logistic Regression, Random Forest, XG Boost and k-Nearest Neighbors (kNN) were considered for the development of the loan status predicting model. The listed classifiers were compared with each other according to their accuracies, AUC ROC for the defining of the most appropriate classifier. The reason of considering these classifiers was the fact that the main issue of the project is binary classification problem. And it was found that these classifiers are best suited for such classification issues. In addition, the Wrapper method (Sequential Forward Search) was used to highlight five features that have the greatest impact on the classification. This study has shown that the Random Forest classifier with features selected dominates over the rest classifiers due to its cross validation (accuracy) and score on overfitting equal to 80.64% and 81.47% respectively.**

***Index Terms*— Loan prediction; Machine learning; feature selection; XG Boost; SVM; Random Forest; Logistic Regression;**

I. INTRODUCTION

According to World Bank forecasts the number of private sector loan application has grown from 125.96% in 2015 to 129.24% in 2019 [17]. Such tendency of loan growth rate shows that the control of loan approval may become one of the most complicated issues in the financial institutions including banking and other market directions. Thus, the demand on the loan prediction models is high. In this work a model that is intended for predicting of loan approval status based on the data about the applicants was developed. Prior work was done in [21], author exploited different models for prediction of loan approval. As a novelty, we proposed new models to be tested as well as feature selection algorithm which can enhance the overall performance of the model. Also, different evaluation metrics of classifiers are presented in the paper. The target issue is considered as the binary classification problem, as the output of the model is "yes" or "no" ("1" or "0") of loan status. For this project the training data with 13 features and 614 samples was used. During the development of the model this data was divided into two parts with ratio of 20% to 80% indicating validation and training data respectively. Training data was used to train the model with an appropriate classifier and the rest data for evaluating the accuracy and performance of the designed model. The following four classifiers such as Logistic Regression, Random Forest, XG Boost and k-Nearest Neighbour have been considered. The most appropriate classifier for the developed model was chosen by evaluating the performance of each classifier in terms of accuracies, AUC ROC. Also, the feature selection process has been conducted via the Wrapper Method (Sequential Forward Search) where the most five influential features selected as well as the rest were eliminated. The given data was analyzed by preprocessing procedure and prepared for further study. This was done in order to avoid some missing and inconsistencies in the training data that is undesired for proper loan prediction and evaluation of the classifiers. The list of related papers and works have been reviewed during the implementation of the project for studying the concepts of listed classifiers and corresponding background information. These classifiers will be briefly described in the following section.

[1]B. Mussagaliyev - student at the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan *bibarys.mussagaliyev@nu.edu.kz*

[2]N. Kadirov - student at the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan *nurken.kadirov@nu.edu.kz*

[3]A. Zollanvari - Associate Professor at the Department of Electrical and Computer Engineering, Nazarbayev University, Astana, Kazakhstan *amin.zollanvari@nu.edu.kz*

II. BACKGROUND THEORY

A. Logistic Regression

Logistic regression is the method of binary classification that operates linearly [7]. In general, the predictive classification models operate as following. For example, the set of input training data variables is required in prediction of labels of class such as small, medium, big, etc. However, the binary classification models are applied in cases, where there are two labels of class that can be Yes/No, 1/0 and True/False. Thus, it was used in this project, as there are two class labels of the loan status, which are Yes or No.
From [14] it was found that there are many similarities of logistic and linear regression classifiers. However, linear regression is intended for predicting the numerical values, while logistic regression is used in problems related to classification. The line or hyperplane is used for modelling the class label by both of these techniques. Then, it is fitted to the data for following purposes depending on the type of regression. Linear regression performs this in order to predict new variable and logistic regression aims to separate two classes according to how well the line fits the data.

The system model of logistic regression can be described as following. For example, if the given input data is defined as X and the output as y. Then, the prediction of such model that is $\hat{y}$ can be described as $\hat{y} = model(X)$. This model is determined by coefficient parameters β. There is defined one β per each input. The input data is predicted by using the weighted sum of inputs.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \cdot \beta_m x_m \quad (1)$$

The output of such model will be $y = X \cdot \beta$, where X - input data matrix, and y - output vector.

The most confusing aspect of this classifier is to determine of what exactly is going to be predicted or calculated by logistic regression. With the linear weighted sum of inputs, the log-odds of successful event can be determined – the log-odds of that the sample is possessed to class 1.

$$log\text{ - }odds = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \cdot \beta_m x_m \quad (2)$$

Odds of success – is the converted variable of the ratio of success probability p, which is predicted by logistic regression to probability of failure *(1-p)*.

$$odds = \frac{p}{1-p} \quad (3)$$

$$log\text{ - }odds = log(\frac{p}{1-p}) \quad (4)$$

The conversion of the log-odds back to odds can be done as following:

$$odds = e^{log-odds} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\cdot\beta_m x_m} \quad (5)$$

$$p = \frac{odds}{1+odds} = \frac{1}{1+\frac{1}{odds}} = \frac{1}{1+\frac{1}{e^{log-odds}}} \quad (6)$$

, the obtained expression of probability p is is the closed form of the logistic regression model and it is designed to predict the probability of whether the sample belongs to default class.

In addition, it is necessary to estimate the parameters β properly, otherwise this may cause an issue in further data analysis. They should be estimated by the sample of conducted observations. The following two techniques can be used for overcoming this issue. These are least squares optimization and maximum likelihood estimation.

B. Random Forrest

Random Forest – is considered as a supervised learning algorithm, which is consisted from a great amount of individual decision trees [11]. This type of classifier can be intended for both classification and regression depending on the application. In this project it was used for classification of the input data. Moreover, it is one of the most user-friendly algorithms, as it is very flexible and easy to use. The more the number of trees, from which it is contained, the more stable the forest will be. The decision trees are created by the random forests based on data samples, which are selected randomly. Then the prediction from each designed decision tree is obtained and the best one is chosen. The selection of the most appropriate solution is determined by voting means.

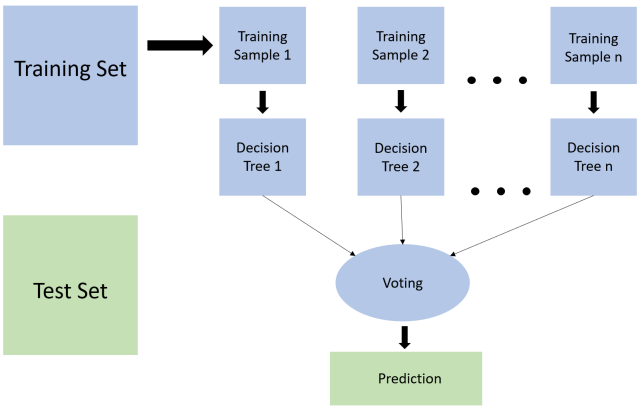


Fig. 1: Random Forrest classifier scheme

One of the main advantages of this classifier is that it can indicate the importance of specific features. Thus, this can also be used during the feature selection process for models that contain trees. However, in this project it has not been used for feature selection, as its feature selection algorithm did not fit to the loan prediction and other classifiers that were used. Its working principle is demonstrated in Fig. 1.

The Random Forest algorithm can be described by following steps:

1) Random samples are selected from input data.
2) A decision tree is constructed for each input sample and a prediction from each tree is obtained.
3) A vote is performed for each predicted value.
4) The final prediction is selected among the obtained ones according to the greatest number of votes.

Benefits of the classifier:

- Due to the amount of decision trees that are included during the process, the random forest can be considered as a stable model that has a high accuracy.
- There cannot be faced any overfitting issues, as the average of all predictions is considered by the classifier, which leads the biases to be eliminated.
- They can fix the missing values. This can be achieved by following two methods. First, by replacing the continuous variables by median values. Second, by calculating the proximity-weighted average of values that are missing.

Limitations of the classifier:

- The generation of random forests predictions is quite slow, as there are multiple decision trees. Thus, the overall process is time consuming.
- The architecture of the model is more difficult than the decision tree.

The last advantage of the Random Forests was very useful during the development of this project. Because there was significant number of values missing in features of the given input dataset. For example, the features such as gender, self-employed and loan amount term were absent. This classifier was a good addition to the conducted data preprocessing, where all of these missing values were filled respectively.

C. XG Boost

XG boost is the "state-of-the-art" machine learning algorithm, which is built by the implementation of decision tress that are gradient boosted constructed for enhancing of the capacity and the speed of performance [5]. It is intended for the operation with structured data. Thus, it was chosen for this project as the data, tested during the implementation was contained in relational databases.

Moreover, it is faster compared with other classifiers due to the fact it was originally designed and written by using C++ language. There are wide range of model parameters including internal parameters that can be used for cross-validation [15]. Also, there are objective functions, which are defined by user, missing values and parameters of tree. This classifier is possessed by the family of boosting algorithms and the gradient boosting framework is used as core. However, there may still be some confusions regarding to the term boosting.

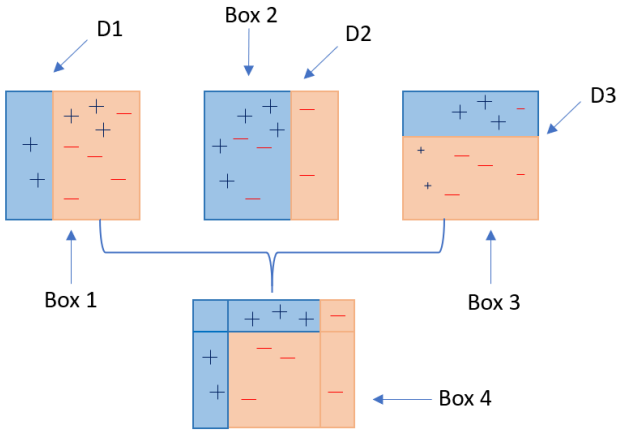


Fig. 2: XG Boost classifier scheme

Boosting is a technique that sequentially operates based on the ensemble principles, where a set of weak learners are combined, and the enhanced accuracy of prediction is delivered. Weak learners referred to ones that are slightly better compared with the random guessing. As an instance, there can be assumed a decision tree, the predictions of which are narrowly more than 50%. The outcomes of the model can be defined at any instant *t* according to the outputs of previous *t-1* instant. A lower weight is assigned to the outcomes that were correctly predicted, while the ones that are misclassified are given by higher weights. There are four classifiers illustrated in the figure below. They are tending to classify the shown "+" and "-" classes homogeneously.

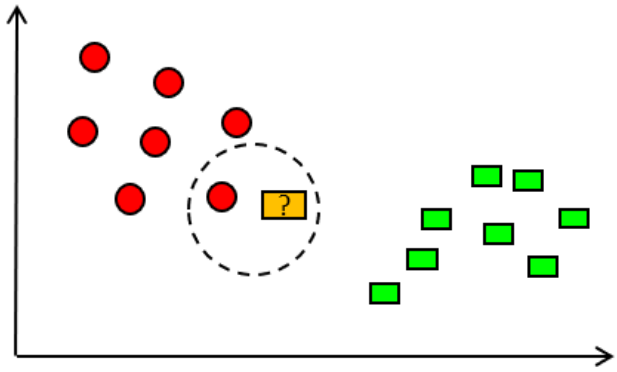


Fig. 3: Classification of new example with kNN algorithm

1) Box 1 – a vertical line at D1 is created by the first classifier, which states that everything that are located at the right side of D1 is "-" and at the left side "+". Nevertheless, it can be seen that three "+" class variables are misclassified. This classifier is considered as the Decision Stump, which conducts the split only at one level. Therefore, its prediction is done according to one feature.
2) Box 2 – three misclassified variables of class "+" are assigned by more weight by the second classifier in this box. And a vertical line is created at D2, which decides that everything at the left of D2 is "+" and at the right side is "-". But there are still three misclassified variables.
3) Box 3 – third classifier assign to the three misclassified variables more weight as it was done previously. And another line at D3 is created, which still tends to fail in classification of the variables.
4) Box 4 – this box is created by the weighted combination of weak classifiers described above (1, 2 and 3 boxes). It can be noticed that it handles well with the classification of variables, as there are no any misclassified points.

The main idea of this type of classifier is the same as the principle described in the above example. This is constructing the new stronger model based on the conclusions made on the importance of features and parameters. And try to decrease the misclassification error of previous model.

D. k-Nearest Neighbour

According to [15] the k-Nearest Neighbour (kNN) classifier is a simple supervised machine learning algorithm, which is applied in solution of classification and regression issues. The main idea of the kNN is that it is assumed by classifier that there are identical points exist close to each other [5]. It is widely used due to the ease of interpretation and fast calculating time.

It is important to mention that the parameter K - defines the number of nearest neighbors. The condition for parameter K is that it should be odd, rather than even. If it is very small there may occur an overfitting, similarly if

K is quite great there will be oversmoothing. Thus, it is required to choose an optimal value for K.

Suppose, there is a case shown in Fig. 3 where the new example or point should be determined for what label it will be predicted. Firstly, the k that is closest to this point is determined after which points are classified via the majority vote of k neighbors of this point. The class that has the greatest amount of votes is considered as a prediction. It is necessary to find the distance between points according to the Euclidean distance. It can be found by following expression:

Euclidean distance *d(a,b)*, where a and b - two points in Euclidean m - space.

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + \cdots + (a_m - b_m)^2} \qquad (7)$$

III. MODEL DEVELOPMENT

This section will provide readers with the fundamental aspects of the loan-predicting model. At first, the statistical analysis of the data sets is performed. This straightforward analysis is very fundamental, as it briefly reveals some important portion of information regarding the data without complicated analysis [18]. Then, the background knowledge regarding the so-called data preprocessing techniques will be provided. This stage of the project is essential, since the original data had some missing values, anomalies, huge spreads and categorical data, which should be encoded to become numerical [20]. As the data sets are fixed and preprocessed, the declaration of classifiers (*Logistic Regression*, *Random Forest*, *XG Boost*, *kNN*) and their fitting to the training data will take place in the project development. Classifiers' evaluation will take place here according to their accuracy, AUC ROC etc. Then, the major part of the model begins, feature selection based on Wrapper method (Sequential Forward Search, SFS). After feature selection algorithm has been applied, fitting of the new versions of classifiers starts, but with several features discarded from training data. Again, evaluation of updated classifiers is done the sane way as it was done previously. By the end, it will be clear, which classifier dominates others in terms of performance metrics. Analyzing the performance evaluation of each classifier, then leads to the choosing the best classifier.

A. Statistical Analysis

As it was said before, this subsection is about statistical analysis related to the data set. By carefully analyzing the train data several important observations were made. Indeed, functions and methods provided in *sklearn* libraries have contributed to the analysis of the data.

- Categorical data counts for whole sample size
- Histogram plot for *ApplicantIncome*, *CoApplicantIncome*, *LoanAmount*, *LoanAmountTerm* features
- Loan status based on credit history bar chart constructed

Categorical data such as *Married*, *Gender*, etc. are counted and output demonstrates, for instance, how many male or female in the sample space or how many applicants are graduated or not graduated.

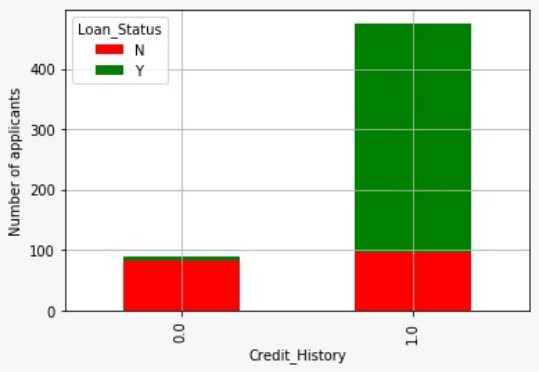


Fig. 4: Loan status based on credit history bar-chart

Histogram plot of numerical data plays an important role of understanding of how do the train data's numerical values look in the full scale. Analyzing such histogram plots, particular numerical feature's spread can be noticed.

From the Figure 4, it can be noticed that the feature *CreditHistory* can provide very fruitful information, it can be seen that majority of those applicants who has credit history have much more chances to have their loan approved.

B. Data Preprocessing

Data preprocessing is very important in any machine learning problem. In majority of cases, the data in hands is not perfect and has several issues. The issues mainly are: missing values, outliers, skewed values, range issues etc. [20]. Using the prior statistical analysis, following data preprocessing steps were done:

- Combining both train and test data
- Deleting unnecessary symbols, namely "+" symbol in *Dependents* feature
- Categorical imputation (fill missing values with majority category)
- Numerical imputation (fill missing values with median or mean value)
- Log transform (distribution of data become approximately normal)
- Categorical data is encoded to achieve analogous numerical data (one-hot encoding)

For instance, the histogram of loan amount term is depicted in Figure 5. It can be seen that data is very spread and has a big deviation. After performing of log transform, the data became normalized, the deviation in range is decreased and it is shown in Figure 6.

C. Classifiers Declaration

Classifiers are defined by set of different parameters, each of these parameters individually affect classifier and the overall performance of model. Therefore, the declaration of classifier can be considered as quite significant step in model development. According to the aims of the project, following parameters for classifiers were chosen for the models.

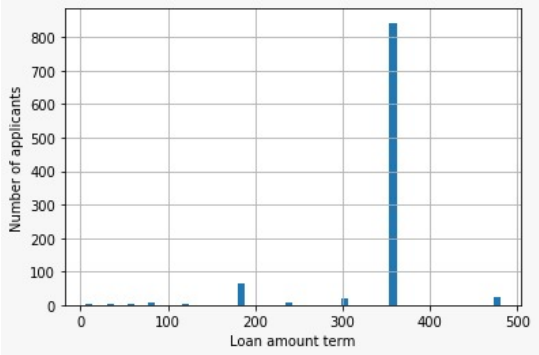


Fig. 5: Histogram of loan amount term

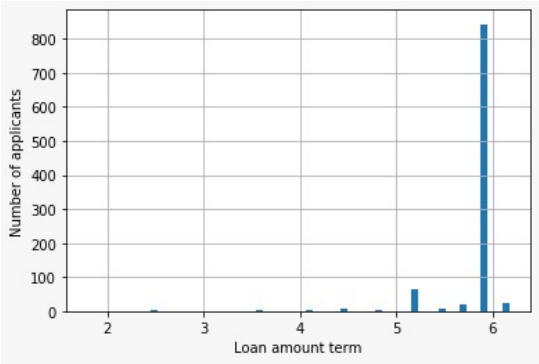


Fig. 6: Histogram of loan amount term (log)

- *Logistic regression*: standard logistic regression model with Limited-memory BFGS solver (*lbfgs*)
- *Random forest*: number of estimators-100, max depth-5, random state-0
- *XG Boost*: number of estimators-20, learning rate-1, max features-0.1, max depth-2, random state-0
- *kNN*: number of neighbors-5, metric-"minkowski", p-2

As the classifiers are declared, it is crucial to train them and validate their performance. To get credible accuracy results and avoid bias estimation, data is split into train data and validation data sets. Former is for training and latter is for validation. Split is achieved with ratio of 80% of data for training set and 20% for validation set.

D. Feature Selection

The approach for selection of features in samples was chosen wrapper method, namely *sequential forward search (SFS)*. The SFS evaluates each individual feature, analyzes it and selects the one that corresponds to best performing model [19]. Metrics of evaluation are defined as accuracy, cross-validation, AUC ROC etc. Then, SFS algorithm was used to select five best features, features that have highest importance out of entire feature set. Using these best five features, new classifiers are constructed and trained on the updated data (samples with five features).

IV. RESULTS

In general, there are total of eight classifiers built. Four of them trained without feature selection and other four trained with feature selection. It should be noted that each of classifiers when selecting best five features, selects them differently due to the different model parameters of each classifier. Figure 7 shows the accuracy, cross-validation and score on overfitting performances for eight classifiers with and without feature selection. It can be seen that, the feature selection positively affects the performance of classifiers, meaning that cancellation of trivial features leads to the better performance of classifier.

	Without Feature Selection				With Feature Selection			
Classifier	LogReg	RF	XGBoost	KNN	LogReg	RF	XGBoost	KNN
Accuracy	83.74 %	82.11 %	82.11 %	82.11 %	83.74 %	82.11 %	82.93 %	82.93 %
Cross-Validation	80.45 %	80.44 %	77.38 %	76.37 %	80.85 %	80.64 %	80.85 %	80.44 %
Overfitting	80.65 %	82.69 %	87.17 %	82.28 %	80.86 %	81.47 %	80.86 %	80.45 %

Fig. 7: Classifiers' performance with and without feature selection

By analyzing Figure 7 and AUC ROC graphs (see Appendix), it can be stated that Random Forest with feature selection classifier is considered as the best candidate to choose. Since, it has highest accuracy and AUC ROC metrics. Therefore, test data (data with no label) can be exploited in the Random Forest classifier. The classifier outputs resultant columns of labels (Loan Status), either applicant gets loan approval or not. The output data is stored as the "csv" file.

V. CONCLUSION

In this paper, we designed a model for predicting the loan status with binary output indicating whether "yes" or "1" which means the loan for candidate is approved and conversely "no" or "0". For the construction of such model the list of following classifiers that include Logistic Regression, kNN, Random Forest and XG Boost was used. Then the performance of each classifier was tested separately in terms of their accuracy, AUC ROC and choose the classifier that best fits for our developing model. Also, five best features which mostly impact on the classification have chosen by using the Wrapper method feature selection, to be price Sequential Forward Search. Initial training data was divided into two parts in following ratio 20% to 80%. First part is validating data that was used for evaluation listed classifier characteristics, while the second part remained as training data and was used for training the model with four classifiers. In addition, preprocessing of data has been done in order to ensure that there will not be any missing and invalid characters in the training data set, which may be undesired for this process. The target binary classification problem was solved. We found that the Random Forest classifier can be considered as the most suitable classifier for this model. As it outperforms the rest classifiers with accuracy of 80.64% and score on overfitting equal to 81.47% respectively.

REFERENCES

[1] A. S, A. M R and S. Ananda Krishnan G, "Comparative Analysis of Heat Maps over Voronoi Diagram in Eye Gaze Data Visualization", International Conference on Intelligent Computing and Control, vol. 9, no. 31, 2017.

[2] A. U.R and S. Paul, "Feature Selection and Extraction in Data mining", IEEE, vol. 316, no. 978-1-5090-4556, 2016.

[3] C. Fahy and S. Yang, "Dynamic Feature Selection for Clustering High Dimensional Data Streams", IEEE Access, vol. 7, no. 127128, 2019. Available: http://creativecommons.org/licenses/by/4.0/.

[4] D. Cheng, Z. Niu, Y. Tu and L. Zhang, "Prediction Defaults for Networked-guarantee Loans", IEEE, vol. 978-1-5386-3788, no. 318, 2018.

[5] G. Arutjothi and C. Senthamarai, "Prediction of Loan Status in Commercial Bank using Machine Learning Classifier", IEEE Xplore Compliant, vol. 978-1-5386-1959-9, 2017.

[6] G. Attigeri, M. Pai and R. Pai, "Analysis of Feature Selection and Extraction Algorithm for Loan Data: A Big Data Approach", IEEE, vol. 978-1-5090-6367, 2017.

[7] H. Sutrisno and S. Halim, "Credit Scoring Refinement Using Optimized Logistic Regression", IEEE, vol. 978-0-7695-6163-9, 2017.

[8] H. Zhong, X. Song and L. Yang, "Vessel Classification from Space-based AIS Data Using Random Forest", IEEE, vol. 319, no. 978-1-7281-3933, 2019.

[9] L. Aymon et al., "Leak detection using Random Forest and pressure simulation", IEEE, vol. 419, no. 978-1-7281-3105, 2019. Available: DOI 10.1109/SDS.2019.00008

[10] M. Ohsaki, P. Wang, S. Katagiri, H. Watanabe and A. Ralescu, "Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 29, no. 9, 2017.

[11] M. Zaffar, M. Ahmed Hashmani and K. Savita, "Performance Analysis of Feature Selection Algorithm for Educational Data Mining", IEEE, vol. 917, no. 978-1-5386-0790, 2017.

[12] S. Chen, Q. Wang and S. Liu, "Credit Risk Prediction in Peer-to-Peer Lending with Ensemble Learning Framework", IEEE, vol. 978-1-7281-0106, 2019.

[13] S. N K, A. Nikhil and H. P, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification", IEEE, vol. 978-1-5386-9471, 2019.

[14] V. Ashlesha, "Predictive and probabilistic approach using logistic regression: application to prediction of loan approval", IEEE, vol. 40222, 2017.

[15] Y. Li, "Credit Risk Prediction Based on Machine Learning Methods", IEEE, vol. 978-1-7281-1846, 2019.

[16] Y. Sayjadah, I. Hashem and F. Alotaibi, "Credit Card Default Prediction using Machine Learning Techniques", IEEE, vol. 978-1-5386-7167, 2018.

[17] "Domestic credit to private sector (% of GDP) | Data", Data.worldbank.org, 2019. [Online]. Available: https://data.worldbank.org/indicator/FS.AST.PRVT.GD.ZS. [Accessed: 28- Nov- 2019].

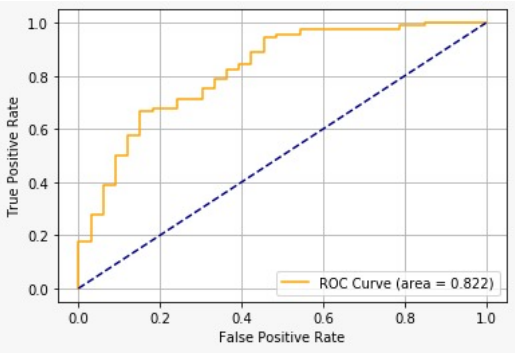
[18] "Fundamental Techniques of Feature Engineering for Machine Learning", Towards Data Science, 2019. [Online]. Available: https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114 [Accessed: 28- Nov- 2019].

[19] "Feature Selection Techniques in Machine Learning with Python", Towards Data Science, 2019. [Online]. Available: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e [Accessed: 28- Nov- 2019].

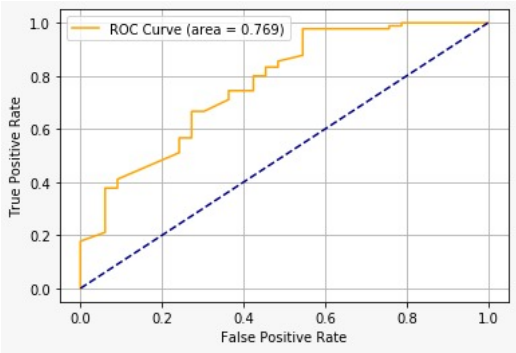
[20] "Bank Loan Default Prediction with Machine Learning", Medium, 2019. [Online]. Available: https://medium.com/henry-jia/bank-loan-default-prediction-with-machine-learning-e9336d19dffa [Accessed: 28- Nov- 2019].

[21] "Predicting Loan Repayment", Medium, 2019. [Online]. Available: https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92 [Accessed: 28- Nov- 2019].

VI. APPENDIX

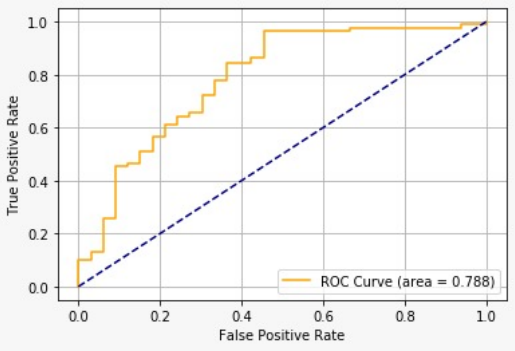


(a) Logistic Regression without FS

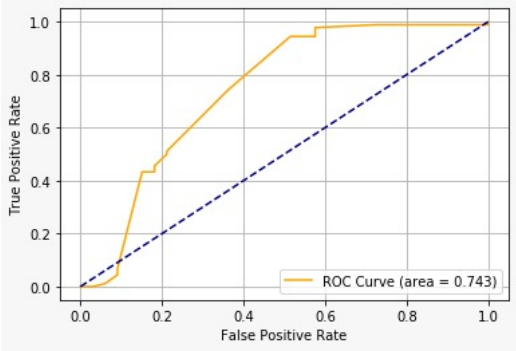


(b) Logistic Regression with FS

Fig. 8: AUC ROC Curve for Logistic Regression

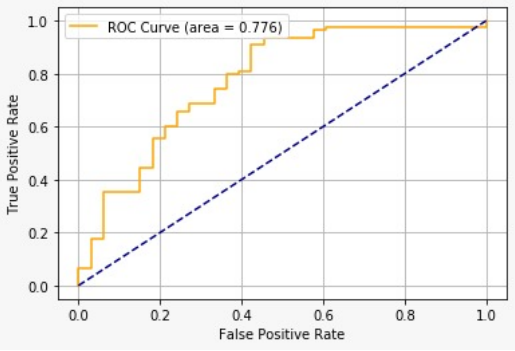


(a) Random Forest without FS

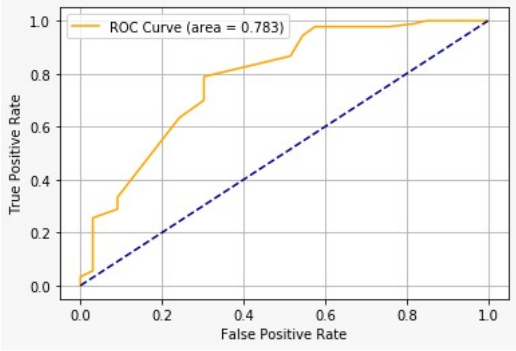


(b) Random Forest with FS

Fig. 9: AUC ROC Curve for Random Forest

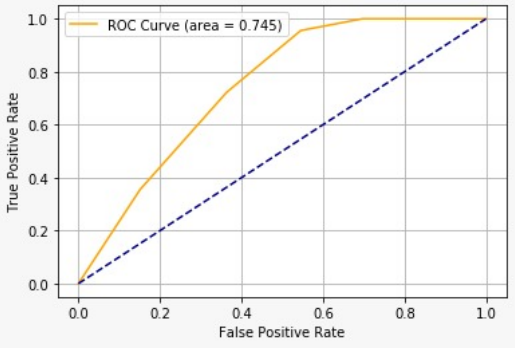


(a) XG Boost without FS

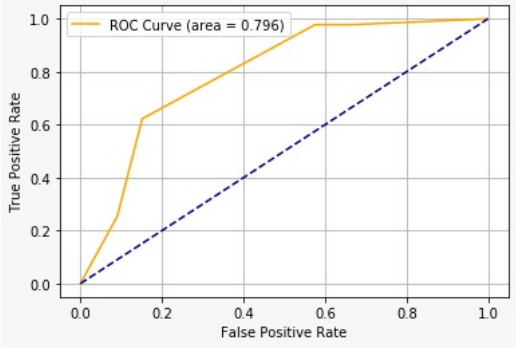


(b) XG Boost with FS

Fig. 10: AUC ROC Curve for XG Boost



(a) kNN without FS



(b) kNN with FS

Fig. 11: AUC ROC Curve for kNN