Original Article

A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression Algorithm

Ch.Venkata Sandeep[1], T. Devi[2*]

Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602 105.
Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105

Abstract

Aim: To analyze the accuracy of Novel Random Forest (RF) and Linear Regression Algorithm (LR) algorithms used to approve bank loans. **Materials and Methods:** The existing model uses Linear Regression Algorithm (LR) and the proposed model employs a Novel Random Forest (RF). The Random Forest is a supervised learning model, it constructs solutions for different regression problems. It provides a high rate of accuracy by cross-validation. The 20 sample values are used to find out the Mean, Std. Deviation and Std. error means. The sample size was measured as 40 per group using G power (80%). **Results:** The resultant graph explains the comparison of the mean accuracy values of algorithms Novel Random Forest (RF) and Linear Regression (LR) where the mean accuracy of the Novel random forest is about 70.5% and the mean accuracy value of the Linear Regression is about 69.5%. The significance obtained is p=1.0 that is p>0.05, it shows insignificance between the groups based on independent sample T-Test. **Conclusion**: The mean accuracy rate of the Novel Random Forest algorithm has been improved to 70.5% compared to Linear Regression which is having around 69.5% mean accuracy.

Keywords: Machine Learning, Novel Random Forest, Linear Regression, Mean Accuracy, Credit Risk, Bank loan.

DOI: 10.47750/pnr.2022.13.S04.211

INTRODUCTION

Credit risk has tremendously increased in the past decade with an increasing number of transactions happening every day. Typically, credit risk assessment involves multiple parameter checks on various levels. Background verification primarily falls into one of the major buckets of evaluation. Background verification of an individual involves verifying the payment behaviors and patterns, repayment periods, credibility scores from various institutions, etc. It plays a significant role in identifying defaulters (Ayyadevara and Kishore Ayyadevara 2018) while approving the applications for loans (Ross Quinlan 1993). This research explains the significance of employing the best algorithm with the greatest accuracy for the Bank loan approval process by assessing credit risk. The analysis of either algorithm is carried out using real-time customer data aggregated from various sources (McCullagh et al. 2009). The mean accuracy of the algorithms is used to understand (Boehmke and Greenwell 2019) the best out of the two to use for efficient creditworthiness. Applications are building an effective predictive model in banking and computation of accuracy for credit card background verification.

There are about 5490 articles published since 2017 which are relevant to the topic on Google Scholar and about 10 articles on ScienceDirect. The existing algorithm, Linear Regression (LR), a popular statistical machine (Keith 2020) is a learning algorithm used to find the relation between the variable and the prediction. Given an

independent variable – x, Linear Regression essentially performs tasks to predict the dependent variable value – y. A best-fit line or rather a regression line is defined as the equation of dependent variable y, independent variable x and a slope a. The equation goes as y = ax + b where b is the intercept. The algorithm provides a model that (Elgawi 2010) gets the best regression line and best fit line by finding the best values of a and b. Minimal error is required to achieve the best value. Training and testing are done on the data files to finalize the prediction of the target variable. The variables are either (Taraba 2021) categorical, ordinal or numerical. The categorical features include employment status, credit histories such as creditworthiness, credit risk, and bank loan status. Numerical features include income status whereas ordinal features include education status and properties. The mean of the dependent and independent variables are only mapped and it is outlier sensitive. Real-time data isn't always linearly separable. This algorithm only works fine in the cases where the data is independent. Novel Random Forest, a supervised learning method, is introduced. It is built on Decision trees that begin by picking random characteristics, say k out of m. The best split criteria are then utilized to locate the root node among the randomly chosen k characteristics. The best splitting qualities are then determined using the same criteria, a function such as gin Index () or Infogain (). A tree is constructed using a root node and a leaf node (Han 2019) as the target. This method is repeated to generate several randomly generated forests. In this approach, the number of trees, as well as the random variables, are critical factors. Information Gain (G) is used to divide data in a certain tree into daughter nodes. This method decreases (Steyerberg, van der Ploeg, and Van Calster 2014) overall cost in addition to the processing time while achieving greater accuracy.

Our institution is passionate about high quality evidence based research and has excelled in various fields (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020).The existing system has issues and major parts (Zhu et al. 2021).It is quite apparent when it is looked at the rising need for new algorithms to solve the drawbacks of existing traditional models. Since the Novel Random forest algorithm has a growing edge over the linear regression algorithm which confines the assumption that the dependent and independent variables are always related. The study aims to determine the better performing algorithm out of Novel Random forest and Linear Regression Models for accurately analyzing real-time numerical data of customers for bank loans by avoiding credit risks.

MATERIALS AND METHODS

The setup of the research has been performed in the Data Analytics Laboratory, Department of CSE in Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The study uses a credit card dataset downloaded from Kaggle. The sample size was measured as 20 per group using G power (80%) (Kuhn and Johnson 2019) with an alpha value of 0.05 and a beta value is 0.95.

The group 1 existing model uses a linear regression algorithm (LR). The 20 sample values are used to find out the Mean, Standard deviation, and Standard error mean. Logistic Regression uses an equation for the representation similar to linear regression. It is a supervised classification algorithm. The sigmoid function is calculated using Levene's Test for Equality of Variances both assumed and non-assumed. The Sigmoid 2-tailed function is calculated using the T-test for equality of means.

The group 2 is proposed Novel Random forest algorithm uses 20 sample values where various statistical metrics are evaluated to get a mean accuracy. This value is used to find the comparison between the existing and proposed models. Random forest uses bagging and features randomness when building each tree to create an uncorrelated forest of trees.

This study was implemented using Jupyter lab, and the hardware configuration used Intel i3 processor, 50 GB HDD, 4GB RAM, and the software configuration required is Windows OS.

The dataset used for the existing model has been imported from Kaggle by downloading the dataset. The 20 sample values and different attributes related to data output were collected. For comparison of two algorithms one being Novel Random Forest (RF) and the other being Linear Regression (LR), where a data set named DataSet2 of customers containing 20 rows is used to find the mean accuracy of both algorithms.

Statistical Analysis

The factual programming which is utilized for examining IBM SPSS rendition 22(64 cycle) is examination programming, which is finished by transferring a dataset (Keith 2020) to the product which gives the result as free factors N, implies, Std. deviation, Std. mistake implies with the mean precision as the result for the given models

Novel Random forest and Linear Regression. The dependent variables are output accuracy and cross-validation. The independent variables are the time period of experience values (S.Vijayarani et al. 2011).

RESULTS

Table 1 describes the pseudocode for supporting Random Forest by updating every value which is needed to operate. All the data is stored in the specified memory location.

Table 2 describes Pseudocode for Linear Regression which is an optimization algorithm for finding the local minimum of a differentiable function. Gradient descent is used to find the values of a function.

Table 3 describes a Comparison between the Novel Random forest and Linear regression algorithms mean accuracy is shown as 70.5% and 69.5% respectively for the sample size N=20.

Table 4 explains the independent variables which define the Equality of the Variances using independent sample T-Test, the significance obtained p=1.0.
Figure 1 is a bar graph created by density with the current amount. There are drastic changes between the intervals of time CGraph table displays the Input parameters such as the name of the active data set, filters, weight, split files, and the number of rows in the working data file, Missing Value Handling, Syntax and Resources are described further.

Figure 2 represents the comparison of the mean accuracy value with algorithms Novel Random Forest and Linear Regression where the mean accuracy of random forest is greater than Linear Regression.

DISCUSSION

Based on the results obtained by independent T-test analysis, the significance value is determined. The significance value obtained has a significant difference of 1.0 ($p>0.05$) between the two groups for the selected dataset and the accuracy of RF is 70.5% which is higher than LR is 69.5%. This shows there does not have any significance due to the inconsistency dataset of bank loans.
The analysis of both the algorithms has been done in Table 1 representing the group statistics and Table 4 representing the independent variables and bar graph which represents the comparison of the (Siddig, Ibrahim, and Elkatatny 2021) two algorithms with the accuracy percentages of 70.5% for Novel Random Forest (Elgawi 2010) and Linear Regression with an accuracy of 69.5%. With the accelerated growth of the economy and the transaction amounts in businesses, the risk factors have also increased. To overcome all the shortcomings of the existing algorithms, we compare and contrast the measures of accuracy of the Novel Random Forest (RF) over the traditional Linear Regression algorithm. There are many studies which are related to the similar study of proposed research where the findings are, “Credit card fraud detection using (Ross Quinlan 1993) AdaBoost (Steyerberg, van der Ploeg, and Van Calster 2014) and majority voting”, “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy", “Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism", “Prediction of BankLoan Status in Commercial Bank using (Arutjothi and Senthamarai 2017) Machine Learning Classifier".
Factors affecting the research work are the predictive models that specify the comparison of two models with the best performance and accuracy. Although the results of the study are better in both experimental and statistical analysis, there are certain limitations in the work. The evaluation of accuracy cannot provide a better outcome on larger data sets. However, the work can be enhanced by applying optimization algorithm techniques, to achieve better accuracy. Feature selection algorithms can be used before classification to improve the accuracy of classifiers. The future scope of the study explains how it will be useful in future for many applications with improved accuracy than other algorithms that don't take into account the necessary number of variables by carefully observing the credit risk while evaluating the credit score or to be precise, the approval score.

CONCLUSION

The loan assessment involves multiple parameter checks on various levels. It is quite apparent when you look at the rising need for new algorithms to solve the drawbacks of existing traditional models. Background verification primarily falls into one of the major buckets of evaluation. The mean accuracy rate of the Novel Random Forest algorithm has been improved to 70.5% compared to Linear Regression which is having around 69.5% mean accuracy. This suggests the proposed system provides an accurate improvement for bank loan approval.

DECLARATIONS

Conflicts of interest
No conflicts of interest in the manuscript.

Authors Contribution
Author CHVS was involved in data collection, data analysis, and manuscript writing. Author TD was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements
The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding
We thank the following organizations for providing financial support that enabled us to complete the study.
1. Sree Vidya High School, Nandigama.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Arutjothi, G., and C. Senthamarai. 2017. "Prediction of Loan Status in Commercial Bank Using Machine Learning Classifier." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iss1.2017.8389442.
2. Ayyadevara, V. Kishore, and V. Kishore Ayyadevara. 2018. "Linear Regression." *Pro Machine Learning Algorithms*. https://doi.org/10.1007/978-1-4842-3564-5_2.
3. Boehmke, Brad, and Brandon Greenwell. 2019. "Linear Regression." *Hands-On Machine Learning with R*. https://doi.org/10.1201/9780367816377-4.
4. Conway, Drew, and John Myles White. 2012. *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started*. "O'Reilly Media, Inc."
5. Elgawi, Hassab. 2010. "Random Forest-LNS Architecture and Vision." *New Advances in Machine Learning*. https://doi.org/10.5772/9386.
6. Han, Chang. 2019. *Loan Repayment Prediction Using Machine Learning Algorithms*.
7. IEEE Staff. 2019. *2019 International Conference on Intelligent Sustainable Systems (ICISS)*.
8. Keith, Michael. 2020. "Random Forest." *Machine Learning with Regression in Python*. https://doi.org/10.1007/978-1-4842-6583-3_5.
9. Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
10. Kleinbaum, David G. 2013. *Logistic Regression: A Self-Learning Text*. Springer Science & Business Media.
11. Koning, Mark, and Chris Smith. 2017. *Decision Trees and Random Forests: A Visual Introduction for Beginners*. Independently Published.
12. Kumar, Manish. 2016. "Superiority of Rotation Forest Machine Learning Algorithm in Prediction of Students' Performance." *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2016908712.
13. Pavlov, Yu L. 2019. *Random Forests*. Walter de Gruyter GmbH & Co KG.
14. Singh, Pradeep Kumar, Zdzisław Pólkowski, Sudeep Tanwar, Sunil Kumar Pandey, Gheorghe Matei, and Daniela Pirvu. 2021. *Innovations in Information and Communication Technologies (IICT-2020): Proceedings of International Conference on CRIME - 2020, Delhi, India : IICT-2020*. Springer Nature.
15. McCullagh, Peter, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. 2009. "Conditional Prediction Intervals for Linear Regression." *2009 International Conference on Machine Learning and Applications*. https://doi.org/10.1109/icmla.2009.115.
16. Molnar, Christoph. 2019. *Interpretable Machine Learning*. Lulu.com.
17. Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
18. Steyerberg, Ewout W., Tjeerd van der Ploeg, and Ben Van Calster. 2014. "Risk Prediction with Machine Learning and Regression Methods." *Biometrical Journal*. https://doi.org/10.1002/bimj.201300297.
19. Arutjothi, G., and C. Senthamarai. 2017. "Prediction of Loan Status in Commercial Bank Using Machine Learning Classifier." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iss1.2017.8389442.
20. Ayyadevara, V. Kishore, and V. Kishore Ayyadevara. 2018. "Linear Regression." *Pro Machine Learning Algorithms*. https://doi.org/10.1007/978-1-4842-3564-5_2.
21. Boehmke, Brad, and Brandon Greenwell. 2019. "Linear Regression." *Hands-On Machine Learning with R*. https://doi.org/10.1201/9780367816377-4.
22. Conway, Drew, and John Myles White. 2012. *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started*. "O'Reilly Media, Inc."
23. Elgawi, Hassab. 2010. "Random Forest-LNS Architecture and Vision." *New Advances in Machine Learning*. https://doi.org/10.5772/9386.
24. Han, Chang. 2019. *Loan Repayment Prediction Using Machine Learning Algorithms*.
25. Keith, Michael. 2020. "Random Forest." *Machine Learning with Regression in Python*. https://doi.org/10.1007/978-1-4842-6583-3_5.
26. Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

27. McCullagh, Peter, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. 2009. "Conditional Prediction Intervals for Linear Regression." *2009 International Conference on Machine Learning and Applications*. https://doi.org/10.1109/icmla.2009.115.
28. Molnar, Christoph. 2019. *Interpretable Machine Learning*. Lulu.com.
29. Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
30. Siddig, Osama, Ahmed Farid Ibrahim, and Salaheldin Elkatatny. 2021. "Application of Various Machine Learning Techniques in Predicting Total Organic Carbon from Well Logs." *Computational Intelligence and Neuroscience* 2021 (August): 7390055.
31. Steyerberg, Ewout W., Tjeerd van der Ploeg, and Ben Van Calster. 2014. "Risk Prediction with Machine Learning and Regression Methods." *Biometrical Journal*. https://doi.org/10.1002/bimj.201300297.
32. Taraba, Peter. 2021. "Linear Regression on a Set of Selected Templates from a Pool of Randomly Generated Templates." *Machine Learning with Applications*. https://doi.org/10.1016/j.mlwa.2021.100126.
33. Arutjothi, G., and C. Senthamarai. 2017. "Prediction of Loan Status in Commercial Bank Using Machine Learning Classifier." *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. https://doi.org/10.1109/iss1.2017.8389442.
34. Ayyadevara, V. Kishore, and V. Kishore Ayyadevara. 2018. "Linear Regression." *Pro Machine Learning Algorithms*. https://doi.org/10.1007/978-1-4842-3564-5_2.
35. Boehmke, Brad, and Brandon Greenwell. 2019. "Linear Regression." *Hands-On Machine Learning with R*. https://doi.org/10.1201/9780367816377-4.
36. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
37. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
38. Elgawi, Hassab. 2010. "Random Forest-LNS Architecture and Vision." *New Advances in Machine Learning*. https://doi.org/10.5772/9386.
39. Han, Chang. 2019. *Loan Repayment Prediction Using Machine Learning Algorithms*.
40. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
41. Keith, Michael. 2020. "Random Forest." *Machine Learning with Regression in Python*. https://doi.org/10.1007/978-1-4842-6583-3_5.
42. Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
43. McCullagh, Peter, Vladimir Vovk, Ilia Nouretdinov, Dmitry Devetyarov, and Alex Gammerman. 2009. "Conditional Prediction Intervals for Linear Regression." *2009 International Conference on Machine Learning and Applications*. https://doi.org/10.1109/icmla.2009.115.
44. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
45. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
46. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
47. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
48. Ross Quinlan, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
49. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
50. Siddig, Osama, Ahmed Farid Ibrahim, and Salaheldin Elkatatny. 2021. "Application of Various Machine Learning Techniques in Predicting Total Organic Carbon from Well Logs." *Computational Intelligence and Neuroscience* 2021 (August): 7390055.
51. Steyerberg, Ewout W., Tjeerd van der Ploeg, and Ben Van Calster. 2014. "Risk Prediction with Machine Learning and Regression Methods." *Biometrical Journal*. https://doi.org/10.1002/bimj.201300297.
52. S.Vijayarani, Dr Dr, S. Vijayarani Dr Vijayarani, Assistant Professor, School of Computer Science and Engineering, Bharathiar University, Coimbatore, and M. Sangeetha M. Sangeetha. 2011. "A Novel Privacy Preserving Approach for Decision Tree Learning." *Indian Journal of Applied Research*. https://doi.org/10.15373/2249555x/mar2014/40.
53. Taraba, Peter. 2021. "Linear Regression on a Set of Selected Templates from a Pool of Randomly Generated Templates." *Machine Learning with Applications*. https://doi.org/10.1016/j.mlwa.2021.100126.
54. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.
55. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
56. Zhu, Siyao, Cassandra Mitsinikos, Lisa Poirier, Takeru Igusa, and Joel Gittelsohn. 2021. "Development of a System Dynamics Model to Guide Retail Food Store Policies in Baltimore City." *Nutrients* 13 (9). https://doi.org/10.3390/nu13093055.

TABLES AND FIGURES

Table 1. Pseudocode for supporting Random Forest over linear regression Algorithm based on the factors of initialize,compute,update, sum of vectors and compare the values.

Input: Dataset frame $x_{i,}$ labels y_i
Output: Sum of vectors ,a array, b and DT
Procedure: 1: Initialize:$a_{i,}$=0,f_i = -y 2:Compute:rdf_loan = RandomForestClassifier(max_depth=10).fit(xtrain_loan, ytrain) 3:Update:Loan status 4:Compute:find dataframe = pd.concat([scaled_dataframe, df_dummies], axis= 1) 5: Until :target = target.dataset('Fully Paid', 1) 6:Update the threshold b 7:Store the status value 8:Update the data entry 9: Determine the datasets of credentials.

Table 2. Pseudocode for Linear Regression which is an optimization algorithm for finding the local minimum of a differentiable function. Gradient descent is used to find values of a function.

Input: R: Dataset frameworks T: Unique terms in all documents
Output: Accuracy
Procedure: for dataframe = dataframe.loc[dataframe['Current Loan Amount']!=99999999] for dataframe['Maximum Open Credit'].describe().astype('int') wij= accuracy of approval Often t_i in document d_j End for document End for of term

Table 3. Comparison of mean accuracy for RF and LR for the sample size N=20, has the standard deviation as 3.02.

	Algorithm	N	Mean	std.dev	Std.error mean
Accuracy	RF	20	70.5000	3.02765	.95743
	LR	20	69.5000	3.02765	.95743

Table 4. Independent sample test for significance and standard error determination. The significance value obtained has an insignificant difference of 1.0 between the two groups for the selected dataset. The p-value = 1.0, Mean Difference = 18 and Confidence interval = (-1.84 - 3.84).

		Levene's test for equality variance		**T-test for Equality of Means**		**T-test for Equality of Means**				
						Sig (2-tailed)	**Mean Difference**	**std**	**95.5% confidence interval of the**	
		F	**sig**	**t**	**df**				**Lower**	**Upper**
Accuracy	Equal variance assumed	0.00	1.0	0.73	18	0.047	1.00000	1.35	-1.844	3.8446
	Equal variance not assumed			0.73	18.0	0.047	1.00000	1.35	-1.844	3.8446

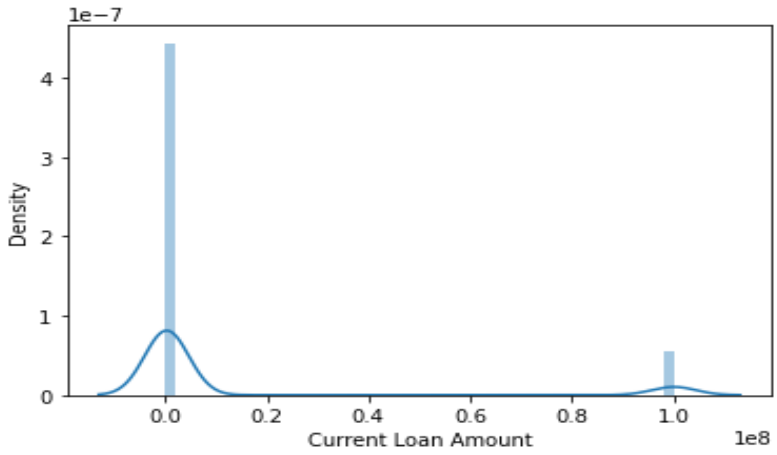

Fig. 1. Represents the density of the current loan amount after comparing all the credit scores of the person. The loan amount is classified as short term and long term based on their repayment and their annual income. Density changes for every time interval.

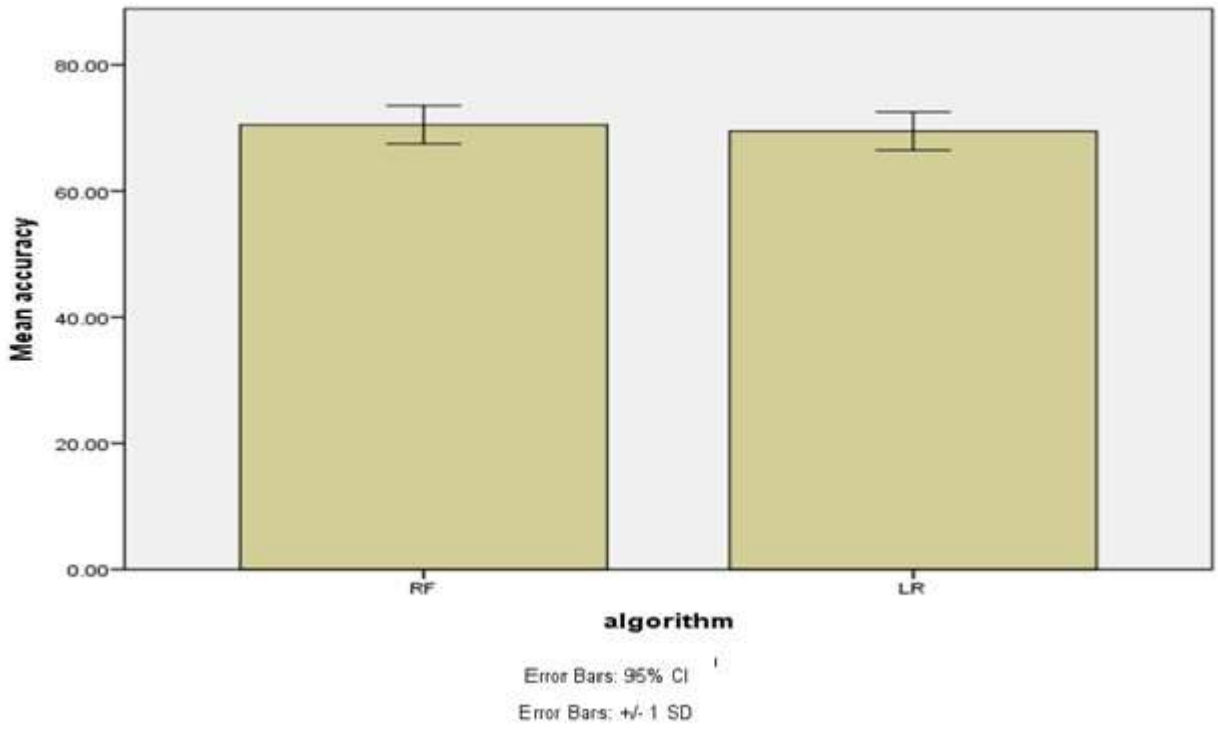

Fig. 2. The graph explains the comparison of the mean accuracy value with algorithms Random Forest (RF) and Linear Regression (LR) where the mean accuracy of random forest is about 70.5% and the mean accuracy value of the Linear Regression is about 69.5%. X axis: Random Forest over Linear Regression Algorithm., Y axis: Mean accuracy of detection + 1 SD.