Using Machine Learning, the Random Forest Algorithm and Logistic Regression to Predict Default Loan Approval.

P. Bhargav[1], K. Malathi[2*]

[1] Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, Pincode:602105.

ABSTRACT

Aim: The objective of this work is to determine an approach in machine learning for default loan approval prediction by comparing Random Forest algorithms with Logistic Regression. **Materials and Methods:** Accuracy and loss are performed with loan risk datasets from kaggle library. The total sample size is 20. The two groups considered were Random Forest (N=10) and Logistic Regression (N=10). The computation is performed using G-power as 80%.**Results:**The accuracy of the Random Forest algorithm is 80.8920% and loss is 19.1080%, which appears to be better than Logistic Regression is 81.2030% and loss is 19.0970% respectively. Finally, the Random Forest algorithm appears significantly better than Logistic Regression. The two algorithms Random Forest (RF) and Logistic Regression (LR) are statistically insignificant with an independent sample T-test value is p=0.317 (p>0.05) shows that two groups are statistically insignificant with confidence level of 95%. **Conclusion:** Determining loan prediction significantly seems to be better in Random Forest than Logistic Regression.

Keywords: Loan Prediction, Loan Risk, Machine Learning, Logistic Regression, Novel Random Forest Classifier, Accuracy.

INTRODUCTION

Even though loan defaults result in significant losses for banks, they pay close attention to this issue and employ a variety of strategies to detect and predict consumer default patterns (“An Empirical Study on P2P Loan Default Prediction Model” 2020). Customer segmentation and profitability, high-risk loan applicants, predicting payment default, marketing, credit analysis, ranking investments, fraudulent transactions, optimising stock portfolios, cash management and forecasting operations, most profitable credit card customers, and cross-selling are some of the main applications in the financial sector (Stahringer 2012). When it comes to borrowing money, there are many different sorts of loans to choose, and it's critical to be aware of your alternatives. The process of reviewing loan collections and assigning loans to groups or grades based on perceived danger and other related loan features is the most important aspect of loan categorization (Hassan and Abraham 2013). The continuous evaluation and classification of loans allows you to keep track of the credit quality of your loan portfolios and take action to prevent them from deteriorating. Banks are required to utilise more complex internal classification schemes than the

more standardised schemes that bank managers use for reporting purposes and that are designed to facilitate monitoring and interbank review (Madaan et al. 2021).

The impact of default risk is discussed in different forms compared to this paper analysis. Around 18 articles were published in IEEE Digital explore and 12 articles were published in research gate on default loan prediction.The goal of this study is to see if an individual's credit risk can be predicted using psychometric tests that assess areas like effective economic decision-making, self-control, selflessness and a giving attitude, neuroticism, and perception toward money(Gramespacher and Posth 2021). One of the most important sources of income for banks is the loan industry. Loan default issues, on the other hand, are a big concern for the loan industry. Other than human processing, there are now more possibilities for classifying and predicting loan default thanks to the growth of the big data era and the development of machine learning techniques. Demonstrate that the AdaBoost model can forecast loan default with 100% accuracy using a real-world dataset from a famous multinational bank, beating other models such as XGBoost, random forest, k closest neighbours, and multilayer perceptrons(Lai 2020).This study established the key elements determining default risks using the Logit model for regression analysis of data provided by the LendingClub platform, and developed a model that can be used to The research gap identified is that the Logistic Regression algorithm has less accuracy. Determining the default loan prediction is shown at a very low percentage while analysing and manual input is not possible to add to dataset.The check borrowers' default risks in advance (Deng 2019). Databases offer a wealth of hidden data that may be used to make informed decisions. Classification and prediction are the two types of data analysis. It's used to explain major data classes or forecast data patterns in the future. Data is classified into preset categories or classes through classification. Because this is supervised learning, the classes are established before the data is examined (Reddy, Jagannatha Reddy, and Kavitha 2010).This paper introduces a credit score performance metric that excludes the impact of "situational factors." Examine the role and effectiveness of credit scoring that underpinned subprime securities during the mortgage boom of 2000-2006 using this metric. Credit score performance is measured both parametrically and nonparametrically, revealing divergent trends, particularly for low-credit-score originations (Bhardwaj and Sengupta 2011). Given that current assessments for P2P borrowers are insufficient and sensitive to unknown causes, a wide range of personal data should be validated, as well as if the data is accurate (income, job, identification, credit situation, etc.) (Xu, Lu, and Xie 2021).(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)
classification of default loan prediction in the existing model is less accurate due to selection of minimal functions. The aim of research work is to improve accuracy of determining the default loan risk, and to reduce loss of data while training and

testing dataset. Novel Random Forest Classifier algorithm used to achieve accuracy.

MATERIALS AND METHODS

The study setting of the proposed work is done at the Data Analytics Lab at Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. Two groups were identified for the study setting: group one is Random Forest algorithm and group two is Logistic Regression algorithm. The sample size is taken as 10 for each group. The computation is performed using G-power as 80% with a confidence interval at 95% and alpha value is 0.05 and beta value is 0.2 (Goedecke 2018).

The datasets are downloaded from kaggle and named as default loan risk dataset.The extracted data sets contain information about the Variable and Description with the unique features like Loan_ID, Gender, Married, Dependents.The dataset was splitted into two parts namely the training part and testing part. 70% of the data was used for training and the remaining 30% was used for testing. The algorithm was implemented by evaluating the train and test datasets. For exhibiting this research work, a jupyter notebook is used along with a laptop with AMD RYZEN 5 processor, 8GB RAM, along with 512 SSD with 64 bit operating system.

RANDOM FOREST ALGORITHM

Random Forest algorithm falls under the background of supervised algorithm learning. It is thrown into ensemble techniques. Its purpose is to create an optimal prediction model by combining numerous basic models. As the name implies, it is a novel Random Forest classifier that consists of a number of decision trees on distinct subsets of data and uses mean accuracy to increase the dataset's predictive accuracy. Instead than relying on a decision tree, Random Forest takes each tree's forecasts and predicts the ultimate result based on the majority votes of the predictions (Vaidya 2017).Table 1 describes the complete pseudo code for Random Forest Algorithm.

LOGISTIC REGRESSION

Logistic Regression falls under supervised algorithms. It is used to predict dependent variables with the use of a set of independent factors. The yield of a dependent variable is specified by the Logistic Regression algorithm as 'yes' or 'no' and 'true' or 'false,' implying that there are only two classes. The sole difference between logistic regression and linear regression is that in logistic regression, the results are in the form of a curve, whereas in linear regression, the results are in the form of a straight line. The count vectorizer, TFIDF transformer, and other sets of words in this approach are utilised for vectorization. These are used to convert data imported from articles into 3 of 15 models in coded format, after which the data is delivered for training, as shown in table 3 (Zhou and Wang 2012).Table 2 describes pseudo code for Logistic Regression Algorithm.

STATISTICAL ANALYSIS

For statistical implementation, software used here is IBM SPSS V22.0. Statistical Package for Social Sciences (SPSS) is used for calculating statistical calculations such as mean, standard deviation, and also to plot graphs etc. The independent variables are the Variable Name, Description, Type. The dependent variable

is ‘accuracy’. The dataset is prepared using 10 as sample size for each group and accuracy is given as a testing variable.

RESULTS

The experimental results are carried out on Random Forest algorithm and Logistic Regression algorithm where performance is measured based on accuracy. Table 3 describes Datasets for loan approval prediction. The classification accuracy and corresponding loss obtained with both the classifiers in the number of test runs shown in Table 4 and Table 5. Table 6 represents comparison of the Random Forest algorithm and Decision Tree algorithm. The accuracy of the Random Forest algorithm is 80.8920% and the Logistic Regression algorithm is 81.2030%.

Table 7 represents the independent sample test that has been performed on Random Forest model and Logistic Regression algorithm for calculating the equal variance assumed and equal variance not assumed and it also shows mean difference, standard error differences with a confidence level of 95%. An independent sample T-test value is $p=0.317$ ($p>0.05$) shows that two groups are statistically insignificant

A simple bar means graph of accuracy by a group of Random Forest model and Logistic Regression algorithms as shown in Fig. 1. It is observed that a novel Random Forest Classifier algorithm has a higher significance when compared to Logistic Regression algorithm. The error bars are shown in graphs and the error rate is less for Random Forest algorithm compared to Logistic Regression.

DISCUSSION

Based on the above study it is observed that the Random Forest algorithm has achieved a result of 80.8920% accuracy. This method has achieved better accuracy compared to the Logistic Regression algorithm.

To support this research work, it is shown that the Random Forest algorithm produced better accurate results (Wood 1975). Fortunately, thorough credit records from our cooperative bank give us a one-of-a-kind opportunity to remedy the issue. Conduct a set of empirical tests to confirm the viability of our hypothesis, i.e. In networked loans, the default of one firm (node) is influenced by the default of another company (adjacent node) (Cheng, Niu, and Zhang 2020). However, after their applications are granted, some consumers behave in an unfavourable manner. To avoid this, banks must develop mechanisms for anticipating client behaviour (Gültekin and Şakar 2018).To oppose this research work, Customers who are in default have failed to meet their contractual responsibilities and may be unable to repay their loans. As a result, there is a desire to obtain a model that can forecast defaulting clients (Turiel and Aste 2020). Statistical learning approaches have recently been used by statisticians and data scientists to try to automate this process in order to reduce risk and boost profitability. In this paper, we use a framework and an application to investigate a framework by using tree-based approaches on publically available datasets (Alaradi and Hilal 2020). Many troubled loans passed typical underwriting norms, indicating that, in addition to LTV and DSCR ratios, other factors such as included property

characteristics should be considered (Lux 2019).

Random Forest algorithm got better accuracy compared with previous research articles. The suggested model's shortcoming is that it only uses a small number of attributes to determine loan approval predictions. Predicting loan amounts is beneficial to both bank workers and applicants. The loan prediction system can determine the weight of each characteristic involved in loan processing automatically, and the same features are processed according to their associated weight on new test data. This paper work can be extended to a higher level in future.

CONCLUSION

This research work indicates that the Random Forest algorithm based model for detection of default loan prediction using the Random Forest algorithm performs better than Logistic Regression with improved accuracy of 80.8920%.

DECLARATION

Conflict of Interests

The author declares no conflict of interest.

Authors Contribution

Author PB was involved in the data collection, data analysis, Manuscript writing. Author PSP assisted in conceptualization, data validation and critical review of the manuscript.

Acknowledgement

The author wishes to express his gratitude to the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, (Formerly known as Saveetha University) for providing the opportunities and facilities needed to carry out research studies.

Funding

We thank the following organisations for providing financial support that enabled us to complete the study.

1. Insysness Technology Pvt. Ltd. Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Alaradi, Mohamed, and Sawsan Hilal. 2020. “Tree-Based Methods for Loan Approval.” *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. https://doi.org/10.1109/icdabi51230.2020.9325614.
2. “An Empirical Study on P2P Loan Default Prediction Model.” 2020. *Financial Engineering and Risk Management*. https://doi.org/10.23977/ferm.2020.030102.
3. Bhardwaj, Geetesh, and Rajdeep Sengupta. 2011. “Credit Scoring and Loan Default.” https://doi.org/10.20955/wp.2011.040.
4. Cheng, Dawei, Zhibin Niu, and Liqing Zhang. 2020. “Delinquent Events Prediction in Temporal Networked-Guarantee Loans.” *IEEE Transactions on Neural Networks and Learning Systems* PP (October). https://doi.org/10.1109/TNNLS.2020.3027346.

5. Deng, Tiannan. 2019. "Study of the Prediction of Micro-Loan Default Based on Logit Model." *2019 International Conference on Economic Management and Model Engineering (ICEMME)*. https://doi.org/10.1109/icemme49371.2019.00058.
6. 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.
7. 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.
8. Goedecke, Jann. 2018. "Contagious Loan Default." *Economics Letters*. https://doi.org/10.1016/j.econlet.2018.05.028.
9. Gramespacher, Thomas, and Jan-Alexander Posth. 2021. "Employing Explainable AI to Optimize the Return Target Function of a Loan Portfolio." *Frontiers in Artificial Intelligence* 4 (June): 693022.
10. Gültekin, Başak, and Betül Erdoğdu Şakar. 2018. "Variable Importance Analysis in Default Prediction Using Machine Learning Techniques." *Proceedings of the 7th International Conference on Data Science, Technology and Applications*. https://doi.org/10.5220/0006872400560062.
11. Hassan, Amira Kamil Ibrahim, and Ajith Abraham. 2013. "Modeling Consumer Loan Default Prediction Using Neural Netware." *2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE)*. https://doi.org/10.1109/icceee.2013.6633940.
12. 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.
13. Lai, Lili. 2020. "Loan Default Prediction with Machine Learning Techniques." *2020 International Conference on Computer Communication and Network Security (CCNS)*. https://doi.org/10.1109/ccns50731.2020.00009.
14. Lux, Nicole. 2019. "Relevance of Loan Characteristics in Probability of Default Prediction for Commercial Mortgage Loans." *26th Annual European Real Estate Society Conference*. https://doi.org/10.15396/eres2019_86.
15. Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. 2021. "Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study." *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899x/1022/1/012042.

16. 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.
17. Parakh, Mayank K., ShriraamUlaganambi, 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.
18. 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.
19. Pham, Quoc Hoa, SupatChupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, AmirzhanKassenov, Zeinab Arzehgar, and WanichSuksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
20. Reddy, M. V. Jagannatha, M. V. Jagannatha Reddy, and B. Kavitha. 2010. "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis." *2010 International Conference on Signal Acquisition and Processing*. https://doi.org/10.1109/icsap.2010.10.
21. 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.
22. Stahringer, Tim. 2012. *Logistic Regression Applied on Loan Default Prediction Using Peer-to-Peer Lending Data*.
23. 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.
24. Turiel, J. D., and T. Aste. 2020. "Peer-to-Peer Loan Acceptance and Default Prediction with Artificial Intelligence." *Royal Society Open Science* 7 (6): 191649.
25. 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.
26. Vaidya, Ashlesha. 2017. "Predictive and Probabilistic Approach Using Logistic Regression: Application to Prediction of Loan Approval." *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt.2017.8203946.

27. Wood, John Harold. 1975. *Commercial Bank Loan and Investment Behaviour*. John Wiley & Sons.
28. Xu, Junhui, Zekai Lu, and Ying Xie. 2021. "Loan Default Prediction of Chinese P2P Market: A Machine Learning Methodology." *Scientific Reports* 11 (1): 18759.
29. Zhou, Lifeng, and Hong Wang. 2012. "Loan Default Prediction on Large Imbalanced Data Using Random Forests." *TELKOMNIKA Indonesian Journal of Electrical Engineering*. https://doi.org/10.11591/telkomnika.v10i6.1323.

TABLES AND FIGURES

Table 1. Pseudo code for Random Forest Algorithm

S.No	Steps
1	Input all required packages
2	Load the data into program using Pandas Library
3	Duplicate values,Stop words in dataset are removed
4	Word count package is imported and count the repetition of words
5	Prepare the model and train the data
6	Import the Machine learning algorithms and predict output
7	OUTPUT: loan approval prediction as an output.

Table 2. Pseudo code for Logistic regression

S.No	Steps
1	Input dataset records
2	Assign variable name to dataset
3	Clean data i.e removing stopwords, punctuation etc.
4	Importing word cloud to count occurrences of a word
5	Prepare model and train data
6	By using function pipeline () , activate model
7	Predicted Target is returned as output
8	Output: Default loan prediction

Table 3. Datasets for default loan prediction

Variable Name	Description	Type
Loan_ID	Unique Loan ID	Integer
Gender	Male/ Female	Character
Maritial_Status	Applicant married (Y/N)	Character
Dependents	Number of dependents	Integer
Education_Qualification	Graduate/ Undergraduate	String
Self_Imployed	Self Employed (Y/N)	Character
Applicant_Income	Applicant income	Integer
Co_Applicant_Income	Co Applicant income	Integer
Loan_Amount	Loan amount in thousands	Integer
Loan_Status	Loan Approved(Y/N)	Character

Table 4. Accuracy of loan approval prediction classification using Random forest algorithm (Mean Accuracy=80.8920%, mean Loss= 19.1080%)

Test	Accuracy	Loss
Test 1	81.20	18.80
Test 2	80.45	19.55
Test 3	80.95	19.05
Test 4	81.45	18.55
Test 5	81.10	18.90
Test 6	81.75	18.25
Test 7	80.12	19.88
Test 8	80.55	19.45
Test 9	81.10	18.90
Test 10	80.25	19.75

Table 5. Accuracy of loan approval prediction classification using Logistic regression algorithm (Mean Accuracy=81.2030%, mean Loss= 19.0970%)

Test	Accuracy	Loss
Test 1	82.11	17.89

Test 2	82.00	18.00
Test 3	81.75	19.25
Test 4	81.87	19.13
Test 5	80.10	19.90
Test 6	81.25	18.75
Test 7	80.75	19.25
Test 8	81.10	19.90
Test 9	80.10	19.90
Test 10	81.00	19.00

Table 6. Group Statistics Random forest and Logistic regression algorithm with the mean value of 80.89% and 81.20%

	GROUPS	**N**	**MEAN**	**std.Deviation**	**std.Error Mean**
ACCURAY	RF	10	80.8920	.53258	.16842
	LR	10	81.2030	.73744	.23320
LOSS	RF	10	19.1080	.53258	.16842
	LR	10	19.0970	.72819	.23028

Table 7. Independent Samples T-test-shows significance value achieved is p=0.317 (p>0.05), which shows that two groups are statistically insignificant.

		F	**Sig**	**t**	**df**	**sig(2-trailed)**	**Mean difference**	**Std Error difference**	**Lower**	**Upper**
ACCURACY	Equal variance assumed	1.060	0.317	-1.081	18	<0.147	-0.31100	0.28766	-0.91534	0.29334
	Equal variance not assumed			-1.081	16.380	<0.148	-0.31100	0.28766	-0.91966	0.29766
	Equal variance assumed	0.459	0.507	0.039	18	<0.485	0.01100	0.28529	-0.58837	0.61037

LOSS	Equal variance not assumed			0.039	16.486	<0.485	0.01100	0.28529	-0.59234	0.61434

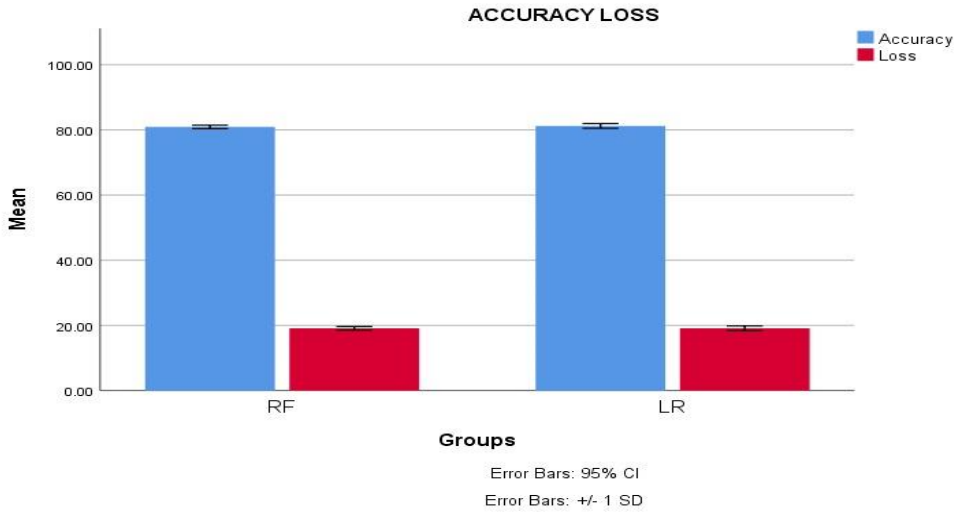

Fig. 1. Comparison of Random Forest algorithm and Logistic Regression algorithm in terms of mean accuracy. Mean accuracy of Random Forest is better than Logistic Regression and standard deviation of Random Forest is slightly better than Logistic Regression. X Axis: Random Forest vs Logistic Regression. Y Axis : Mean Accuracy of detection = +/- 1 SD.