Original Article

An Innovative Approach to Predict Loan Eligibility of a Customer in Bank by Comparing Random Forest Algorithm over Logistic Regression in terms of Accuracy

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

[1]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.
[2]Project Guide, Corresponding Author, 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 predict the loan of the person using Random Forest Algorithm (RF) over Logistic Regression Algorithm (LR). **Materials and Methods:** The existing model uses a logistic regression algorithm. The 20 sample values are used to find out the mean, std. deviation, std. error means. The proposed Novel Random forest algorithm uses 20 sample values where various statistical metrics are evaluated per group. A total of 40 samples are used to find out the mean, Std. deviation, std. error means between the groups. 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 sample size was measured as 20 per group using G power (80%). **Results**: The graph explains the comparison of the mean accuracy value with algorithms Novel Random Forest and Logistic Regression where the mean accuracy of the decision tree is about 70.5% and the mean accuracy value of the Logistic Regression is about 69.5%. The statistical significance $p>0.05$, since $p=1.0$ previls insignificance 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 Logistic Regression which is having around 69.5% mean accuracy. This suggests the proposed system provides an accurate analysis for loan approval.

Keywords: Logistic Regression, Novel Random Forest, Machine Learning, Mean Accuracy, Loan Prediction, Analysis.

DOI: 10.47750/pnr.2022.13.S04.210

INTRODUCTION

Credits are crucial for all sectors, from students to small and medium-sized businesses. However, it is difficult for banks to collect on all of their loans. Defaults are unavoidable owing to a variety of circumstances such as bankruptcy, family obligations, several dependencies, and so on. Analyzing trends and behaviors based on real-time data provides an advantage throughout the loan approval process (Kleinbaum 2013). This study shows the significance of using analytical algorithms to analyze and apply prediction techniques to categorize data as defaulters or not. This study provides (Kelleher, Namee, and D'Arcy 2020) a comprehensive understanding of how to perform a fair assessment of numerical (Kumar 2016) data from consumers using Novel Random forest over the logistic regression algorithm. Applications are building an effective predictive model in (Siddig, Ibrahim, and Elkatatny 2021) banking and computation of accuracy for the credit card approval and calculation of the classification accuracy for a loan.

On IEEE Xplore, there are around 9400 publications published since 2017 that are relevant to the issue and approximately 82 articles on ScienceDirect. The current technique, Logistic Regression, a prominent statistical machine learning algorithm, is used to categorize data based on the outcome variable of extreme fringes through predictive modeling, with the endpoints separated by something like a logarithmic line. Data is taken from Kaggle and utilized for predictions for (Hilbe 2009) research and analysis. The sigmoid function is employed to

construct the model since the output is binary. Multiple characteristics, such as age, creditworthiness, (Kelleher, Namee, and D'Arcy 2020) term, amount, scores, account information, Business values, Customer assets, and so on, are used to determine the chance of default per person. The algorithm's primary objective is to guarantee that the data is cleaned in order to avoid missing values in the numerical data set. Following that, the model begins to be trained using relevant data sets. Because pre-processing is such an important part of the model, it becomes (Kelleher, Namee, and D'Arcy 2020; Singla 2021) time-consuming and costly. The model also loses authenticity since it does not tackle the concerns of class imbalance, and it is hard to analyze non-stationary settings. Novel Random Forests, a supervised learning method based on Decision trees, is proposed. It begins by picking random characteristics, say k out of m. (Koning and Smith 2017) The best split criteria are then utilized to locate the root node among the randomly chosen k characteristics. The fittest 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 as the target. This method is repeated in order to generate a number of randomly generated forests. In this approach, the number of trees and the random variables is critical factors. The term Information Gain (G) refers (Kelleher, Namee, and D'Arcy 2020) to the process of dividing data in a tree into daughter nodes. This method decreases overall costs and processing time while also 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 clear when considering the growing demand for new algorithms to address the shortcomings of present models (Pavlov 2019). Because the Novel Random Forest method has an increasing advantage over the logistic regression approach, which is restricted to non-stationary situations. The study's goal is to find the best performing algorithm among Novel Random Forest and Logistic Regression models for properly analyzing consumer's real-time numerical data.

MATERIALS AND METHODS

The setup of the research has been performed in the Data Analytics Laboratory, Department of CSE in Saveetha School 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%) with an alpha value of 0.05 and a beta value is 0.95.

The group 1 which is considered the existing model uses a logistic regression algorithm (LR). Logistic Regression uses an equation for the representation similar to linear regression. It is basically a supervised classification algorithm. The 20 sample values are used to find out the Mean Std. deviation, std. error means. 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 which is considered the proposed Novel Random forest algorithm uses 20 sample values where various statistical metrics are evaluated to get to 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 individual tree to create an uncorrelated forest of trees.

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

The project is mainly a comparison of two algorithms one being Novel Random Forest and the other being Logistic Regression , where a data set named DataSet of customer1 containing 20 rows is used to find the mean accuracy of both algorithms. The dataset used for the existing model has been imported by Kaggle by downloading the dataset which has records of around 20 sample values and different attributes related to the output of the data.

Statistical Analysis

The measurable programming which is utilized for investigating IBM SPSS rendition 22 (64 bit), which is an examination programming that is finished by transferring a dataset to the product, which gives the output as independent variables N, means, std. deviation, std. error means with the mean accuracy as the output for the given models Novel Random forest and logistic regression. The dependent variables are output of prediction categories 1 / 0. The independent variables are the time period of experience(Wang et al. 2022).

RESULTS

Table 1 explains the pseudocode for supporting Random Forest Algorithm. Initially, the data is stored temporarily in a given memory space, then all the factors that determine the process are taken into consideration. The final result is obtained only after storing output at the data allocation.
Table 2 explains the pseudocode for supporting the Logistic Regression Algorithm by comparing the document's accuracy and providing the loan status. It verifies all the factors that are related to the documentation before approving the loan.
Table 3 explains the group statistics of the model by comparing the algorithm and accuracy using sample values = 20 for Random Forest (RF) and Logistic Regression (LR), Mean = 70.5000 and 69.5000. The Std.Deviation = 3.02765, Std. Error Mean = 0.9574.
Table 4 shows the independent sample T-test analysis which defines the Equality of the Variances with the significance p = 1.0 for the confidence interval 95%.
Figure 1 represents a bar graph created to compare loan status with the amount of interest paid to the total amount taken. The graph shows the variation between the short term and long term loans. It also shows the house mortgage within the period.
Figure 2 represents the comparison of the Random Forest Algorithm over Logistic Regression with the X-axis of RF vs LR Algorithm and Y-axis of Mean accuracy of detection ± 1 SD.

DISCUSSION

Based on the results obtained by independent T-test analysis, the significance value is determined as p = 1.0, where p>0.05 states there does not have significance among the groups due the incompatible dataset. The accuracy of 70.5% for RF which is higher than the 69.5% mean accuracy of LR.
The analysis of both (G and Manoj 2021) algorithms has been done with Table1 representing the group statistics (Dimitriadis, Liparas, and Alzheimer's 2018) and Table 2 representing the independent variables, and a bar graph that represents the comparison of the two algorithms with the accuracy percentages of 70.5% for Novel Random Forest and logistic regression with an accuracy of 69.5%. Defaulters are quite prevalent in the financial sector in uncertain times like these. It is critical to authorize (Gholamnezhad, Broumandnia, and Seydi 2020) loans based on an accurate assessment of real-time data (Hilbe 2009) gathered from consumers. The study demonstrates how the Novel Random Forest is more accurate than the Logistic Regression Algorithm, which is used (Naeem et al. 2021) to categorize data sets to locate suitable clients for loan approval.
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 (Ross Quinlan 1993) 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 the 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

Background verification primarily falls into one of the major buckets of evaluation. The mean accuracy rate of the Random Forest (RF) algorithm has been improved to 70.5% compared to Logistic Regression which is having around 69.5% mean accuracy. This suggests the proposed system provides an accurate analysis for loan approval.

DECLARATIONS

Conflicts of interests
No conflicts of interests in the manuscript.
Authors Contribution
Author CHVS was involved in data collection, data analysis, and manuscript writing. Author SPC 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. Saveetha University
2. Saveetha Institute of Medical and Technical Sciences
3. Saveetha School of Engineering
4. Sree Vidya High School, Nandigama

REFERENCES

1. 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.
2. 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.
3. Dimitriadis, Stavrosi, Dimitris Liparas, and D. N. I. Alzheimer's. 2018. "How Random Is the Random Forest? Random Forest Algorithm on the Service of Structural Imaging Biomarkers for Alzheimer's Disease: From Alzheimer's Disease Neuroimaging Initiative (ADNI) Database." Neural Regeneration Research. https://doi.org/10.4103/1673-5374.233433.
4. Gholamnezhad, Pezhman, Ali Broumandnia, and Vahid Seydi. 2020. "An Inverse Model-Based Multiobjective Estimation of Distribution Algorithm Using Random-Forest Variable Importance Methods." Computational Intelligence. https://doi.org/10.1111/coin.12315.
5. G, Manoj Kumar, and Kumar G. Manoj. 2021. "Accuracy Analysis for Logistic Regression Algorithm and Random Forest Algorithm to Detect Frauds in Mobile Money Transaction." Revista Gestão Inovação E Tecnologias. https://doi.org/10.47059/revistageintec.v11i4.2182.
6. Hilbe, Joseph M. 2009. Logistic Regression Models. CRC Press.
7. 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.
8. Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2020. Fundamentals of Machine Learning for Predictive Data Analytics, Second Edition: Algorithms, Worked Examples, and Case Studies. MIT Press.
9. Kleinbaum, David G. 2013. Logistic Regression: A Self-Learning Text. Springer Science & Business Media.
10. Koning, Mark, and Chris Smith. 2017. Decision Trees and Random Forests: A Visual Introduction for Beginners. Independently Published.
11. 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.
12. Naeem, Muhammad, Jian Yu, Muhammad Aamir, Sajjad Ahmad Khan, Olayinka Adeleye, and Zardad Khan. 2021. "Comparative Analysis of Machine Learning Approaches to Analyze and Predict the COVID-19 Outbreak." PeerJ. Computer Science 7 (December): e746.
13. 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.
14. 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.
15. Pavlov, Yu L. 2019. Random Forests. Walter de Gruyter GmbH & Co KG.
16. 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.
17. 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.
18. Ross Quinlan, J. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.
19. 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.
20. 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.
21. Singla, Saurav. 2021. Machine Learning for Finance: Beginner's Guide to Explore Machine Learning in Banking and Finance (English Edition). BPB Publications.
22. 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.
23. 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.

24. Wang, Jianan, Xiaoxian Gong, Hongfang Chen, Wansi Zhong, Yi Chen, Ying Zhou, Wenhua Zhang, Yaode He, and Min Lou. 2022. "Causative Classification of Ischemic Stroke by the Machine Learning Algorithm Random Forests." Frontiers in Aging Neuroscience 14 (April): 788637.
25. 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.
26. D, Brahmananda Reddy, Reddy D. Brahmananda, and P. R. Kuber Gupta. 2020. "AN OPENCV BASED EFFICIENT FACE RECOGNIZATION APPROACH FOR AUTOMATED ATTENDANCE UPDATION." EPRA International Journal of Research & Development (IJRD). https://doi.org/10.36713/epra4323.
27. Ganidisastra, Asep Hadian Sudrajat, and Yoanes Bandung. 2021. "An Incremental Training on Deep Learning Face Recognition for M-Learning Online Exam Proctoring." 2021 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob). https://doi.org/10.1109/apwimob51111.2021.9435232.
28. Kamencay, Patrik, Tibor Trnovszky, Miroslav Benco, Robert Hudec, Peter Sykora, and Andrej Satnik. 2016. "Accurate Wild Animal Recognition Using PCA, LDA and LBPH." 2016 ELEKTRO. https://doi.org/10.1109/elektro.2016.7512036.

TABLES AND FIGURES

Table 1. Pseudocode for supporting Random Forest Algorithm over Logistic Regression in terms of accuracy based on the initialize, compute, update, sum of vectors compares the values.

Input: Graph dataset $x_{i,}$ labels y_i
Output: Sum of vectors ,a array, b and RF
Procedure: 1: Initialize:$a_{i,}$=0,f_i = -y 2:Compute:dataframe['Bankruptcies'] = dataframe['Bankruptcies'].fillna(0) 3:Update:Loan status 4:Compute:list(find graph dataset.columns) 5: Until dataframe['Annual Income'].isnull(),'Annual Income'] = 0 6:Update the threshold b 7:Store the status value 8:Update the data entry 9: Determine the data set amount.

Table 2. Pseudocode for Logistic 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 parameter.

Input: L: Mean accuracy U: Unique terms in all documents
Output: Accuracy
Procedure: for sns.pairplot(dataframe[Continuous], hue='Loan Status') for df_dummies = pd.get_dummies(dataframe[Category], drop_first= True) wij= accuracy of approval Often t_i in document d_j End for document End for of term

Table 3. Comparison between significance level for Random Forest Algorithm over Logistic Regression, the accuracy pertains 70.50% and 69.50% respectively.

	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. P value is l.0 considered to be statistically insignificant and 95% confidence intervals. The p-value =1.0, Mean Difference = 1.00 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	1.00	.739	18	.470	1.00	1.35	-1.84	3.84
	Equal variance not assumed			.739	18.00	.470	1.00	1.35	-1.844	3.84

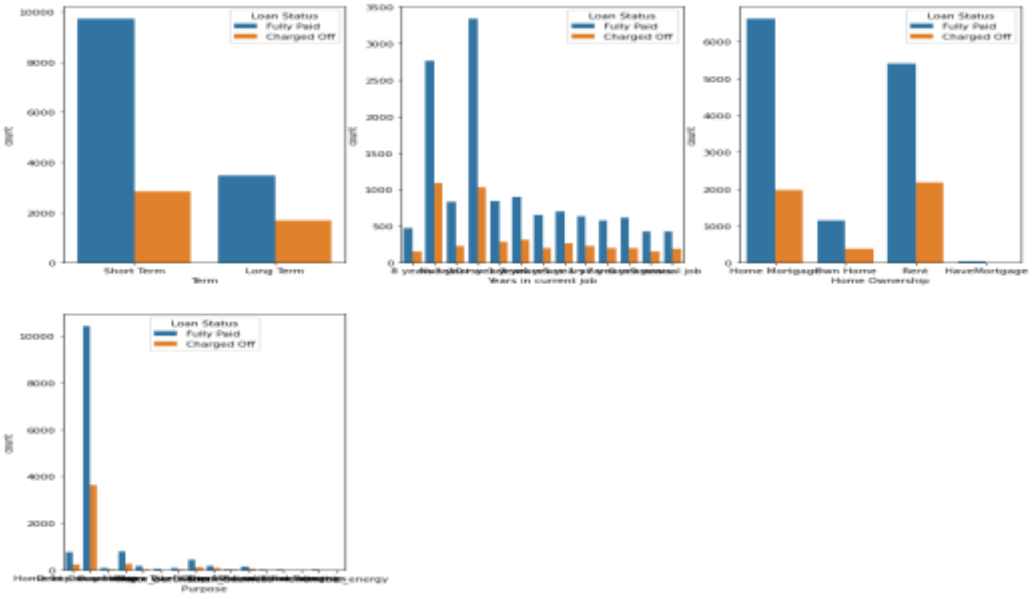

Fig. 1.The bar graph represents the status of a tube person by comparing all the amounts of a person and also it shows the amount within the time period. The x-axis represents the expenditure over time. The loan amount is classified short term and long term based on their repayment and their annual income.

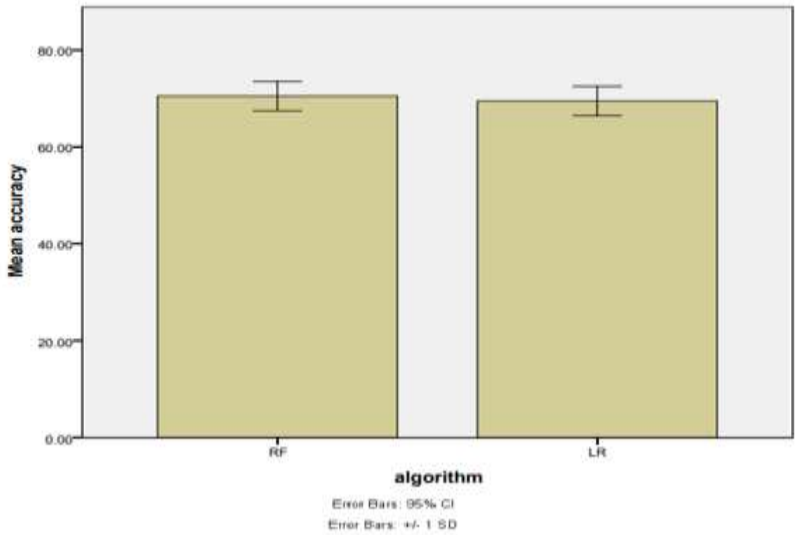

Fig. 2. The graph explains the comparison of the mean accuracy value with algorithms Random Forest (RF) and Logistic Regression (LR) where the mean accuracy of the Random Forest is about 70.5% and the mean accuracy value of the Logistic Regression is about 69.5%. X axis: Random Forest Algorithm over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