IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91120

Machine Learning Based Algorithm For Predictive Analysis Using Python Language

Sumit Koul[1]

Assistant Professor, Department of Mathematics,

Lovely Professional University, Phagwara, India[1]

Abstract: To study the instructions which are being governed by a procedure, when the firstly machines are to follow a learning pattern and then to try for various rules and knowledge within which it is found that how the machine is to perform, that is called as Machine Learning. In this paper the various abilities and working implementations of the machine learning as a tool is studied and analysed with the help of case study in linear regression and logistic regression using python language. This also shows the coherence and application of machine learning algorithms by the prediction of various events that are considered in this study.

Keywords: M LAlgorithms, sampling, train and test, Python language.

I.Introduction

According to the nominal 1956 initialization of the branch of artificial intelligence (AI) or, as per various definitions the aim the creating non-human brain power is the area of concern for a long time. According to literature it has been very logical to launch the aim of "creating" brainpower as compared to that of "designing" it, like as in 17th century it was practical as Leibnitz was scripting about reasoning as a form of computation, to consider in order to create AI would be similar to create a waterwheel or a pocket watch. This means that firstly one is to comprehend the principle behind it and then to utilize the human brains for devising a pattern dependent on the ideology, then lastly to build a system according to that pattern. William Paley in the beginning of the 19th century had built certain assumptions discussing that intellectual creation is essential for the construction of complex systems. Darwin analyzed how complex and adaptive systems can take place naturally from a procedure of choice selection for random variation and the procedure was explained that how intricate and adaptive plan is formed by not including a human interference. Based on the facts from various evolutionary theories, and other fields it can be inferred that mostly all of the remarkable characteristics of genetic representative came into picture more or less by Darwinian evolutionary processes (with certain modification).

Certain definitions are coated in literature that would explain AI as a field. Several of the proffered definitions are clearly oriented in the direction concerning the design parameter. For example, Dean et al. [6] have described AI as "the design and study of computer programs that behave intelligently". It is also taken into consideration that the design consideration of the developed system is utmost important for understanding the foundation for future designs. The significance of these philosophical theories possesses and illustrious background in AI, and contributions of these analysis persists to play a vital responsibility in the development of efficient computational utilization of argumentation technology. Among which the Machine Learning (ML) is one of such AI techniques which works, by providing access to the right data, and the machine may learn itself how to solve the particular problem.

In today's world the Machine Learning (ML) is undoubtedly influential and powerful technologies among the tools being used nowadays and it is going to remain as the vital component in future also. The procedure of transformation of ideas into facts by using ML. since during past few decades there is a huge increase in the data in various sectors. This accumulation of data becomes futile until it is analyzed properly and a useful conclusion is drawn from it. The ML techniques play a vital role in locating the fundamental model from the composite data which otherwise is a tedious task. To forecast upcoming events and accomplish all types of intricate decision making the unseen patterns and ideas of a problem are utilized. The interaction with the ML is very common in our day to day life, like we always take the help of Google in order to search anything. ML is the main building block behind it, as it is also continuously learning and refining from each interface. ML also is playing a vital role in global advancements like detecting certain deadly disease like cancer, in producing new medications, automated vehicles etc. ML involves two types of tasks which are supervised machine learning and unsupervised learning.

In this case then supervised learning, the aim is to study the mapping (the rules) among the group of inputs and outputs. e.g. the input values can be the forecasting weather, with the output can be the number of persons visiting to enjoy the weather. The objective in this type of learning is to study the mapping which defines the connection among the temperature and people visiting. Thus, ML is a pre-defined set of training-examples which enables to attain a precise conclusion. In case of unsupervised learning, it is the input data only which is coated in the examples. This type of learning algorithms doesn't possess a categorized example outputs for targeting. It is also astonishing to find the composite patterns concealed in data without any labels. It can be understood with an example in real-time is to sort various coloured coins in the different piles. This kind of segregation is not made known to anybody but can just be isolated from each other by just seeing to the characteristics which is colour in this case and hence they can be separated in this way.

ML is thus technique which deals with the initiative that, by providing access to the right data, machines themselves can learn how to solve a particular problem without any human interference. ML algorithms are capable to solve the complex mathematical problem into the simple problem. For example if we have to find the rank of matrix of order two it very easy to solve manually, but if order of matrix increases i.e. (10*10) or (n*n) order then it becomes complicated to find it's rank but with help of machine (computer language c language, c++ and matlab) it can be solved easily. So, machine is saving our time in these mathematical calculations likewisethe machines also help in the field of health-care.As now a days the advancement and automation in the field of biotechnology is gaining pace day by day and we get different types of heath checkup machine like (CT scanning, x-ray EEG etc.). All these advance machines in healthcare helps to detect a problem in our body and finally doctors give us treatment to recover in normal mode. Training data are sample data which obtain through algorithm based on mathematical model are used by machine learning. There are many problems which are complex in nature cannot be solved by human hands for doing this their occur in machine which use algorithm to solve the define problem i.e. complex problem. Some of the examples are shown on Fig. 1. This paper provides a computational study with respect to the statistical models to analyze the ML concepts.

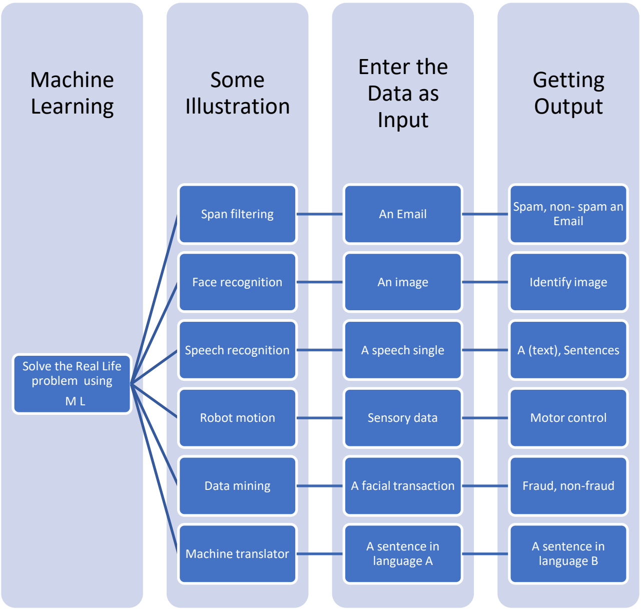

Fig. 1. Machine Learning in real world

II.Literature Review

Since its inception Spector [20] has described the preliminary theories for the explanation of Artificial intelligence. Das *et al.*[4] has studied new technologies used in our daily life through machine learning. For instance, more than a dozen of people use Facebook which recognizes friends associated with the individual person by their photo, this is done by learning algorithms. Also, many more application has been introduced in this paper about machine learning. Musumeci*et al.*[12] has explained the concept of optical exchanges and networking via. Machine learning with its future aspects.

Simeone [17] has studied the application of machine learning in telecom sector Also utilized the concept of supervised and unsupervised learning for developing the conventional engineering result. Das and Behera [5] have studied the ideas and initialization of machine learning with some important algorithm. They have used the data set containing 140 observations and they have been used for train data with different algorithms. Silmon *et al.* [18] has discussed the recent advantages of deep learning which is a part of the machine learning. Also, big data as an application of it. Ye and Zhang [24] has considered the concept ANN with its approximation technique for derivative pricing. They have use the BSDEJs and PIDEs to calculate the conditional expectation of price problem. Wang [21] has considered the modified

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91120

machine learning for improving the accuracy by prediction the bankruptcy on the bases of real-life data. Also, they have compared it with other machine learning methods that are frequently used. Buczakand Guven [3] has discussed the cyber analytics using machine learning and data mining, surveyed few literatures based upon them. Also has explained the new challenges to tackled the cyber security through machine learning and data mining. Wang *et al.*[22] has considered the mobile service traffic classification, mobile network management, mobile devices and security using machine learning. Mitchell *et al.*[11] has discussed the emerging trend in the machine learning and its association with sciences and to the general public. Sreeja *et al.* [19] has studied a short audit and future viewpoint of the immense operation of machine learning has been made, Classify the different types of machine learning with good definition and applications are given in this paper. Kaur *et al.* [9] has studied instruction detection, it is mainly two types, first one misuse or signature-based detection and another one is Anomaly detection .in this paper, he is discussing Anomaly detection techniques based on machine learning method. Reshmi *et al.* [14] has studied the machine learning algorithm with presumption and panorama and real life applications of machine learning are written in this paper. Wiese *et al.* [23] has studied the Credit Card Transactions, scam exposure, and Machine Learning Modeling Time with LSTM Recurrent Neural Networks. Shen *et al.* [16] has discussed the new algorithm that exploits the time correlation between worldwide share exchange and different monitory product the next-day share exchange drift with model of machine learning algorithm support vector machine.

Bernardi*et al.* [1] has discussed the iterative, supposition process is conducted, incorporated with other disciplines tobuild150 successful products enabled by Machine Learning. Boyarshinov [2] has studied algorithm for an artificial intelligence application in three-part i) building a data are in trained form, ii) for recognizing the pattern algorithm are to trained and iii) As far as concerned about the overfilling of data that are trained some additional information kept a side. Sajda [15] has discussed the different methodologies and, for each, provide examples of their application to specific domains in biomedical diagnostics. Øland [13] has studied some general features, as well as some popular models, and algorithms. And he was also discussed the application of ML in musical field. Husain *et al.* [8] has discussed machine learning algorithm is applying financial sector.

III.SUPERVISE LEARNING

Supervised means obvious or direct certain activity runs correctly. In these types of learning, machine learns things under some guidance or algorithm. For example, Teacher is guiding us or taught us correctly same as in supervised learning. Machine learns by feeding of explicitly data and machine is telling us, this is input and this is output. So, Teacher is a trained data for student. There are some examples and their description shown in Table1 and Working rule of supervised learning shown on Fig. 2 and the types of supervised learning is classification and regression analysis as shown on Table1 below.

TABLE IEXAMPLES OF SUPERVISED LEARNING

Examples of supervised learning	Description
Image detection	To evaluate the data through algorithm and given precision
Sorting	Identifying those things which can sort with the help of algorithm
Making of decision	How one can reduce the risk less in treading by estimate it

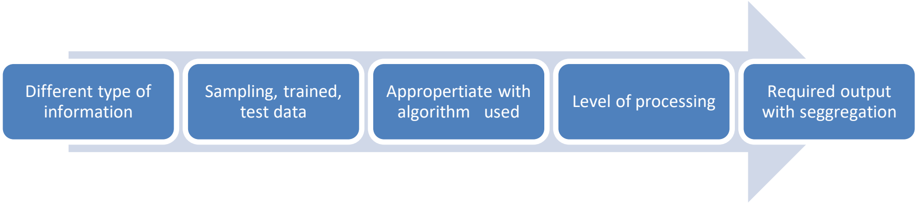
Different type of information
Sampling, trained, test data
Appropertiate with algorithm used
Level of processing
Required output with seggregation

Fig. 2. Working rule of supervise learning

IV.ALGORITHMS IN MACHINE LEARNING

In machine learning there are various algorithms which are used in predictive modeling as represented in the Fig. 3. In this paper, linear regression and logistic regression is explained with a case study using Python language.

DOI 10.17148/IJARCCE.2020.91120

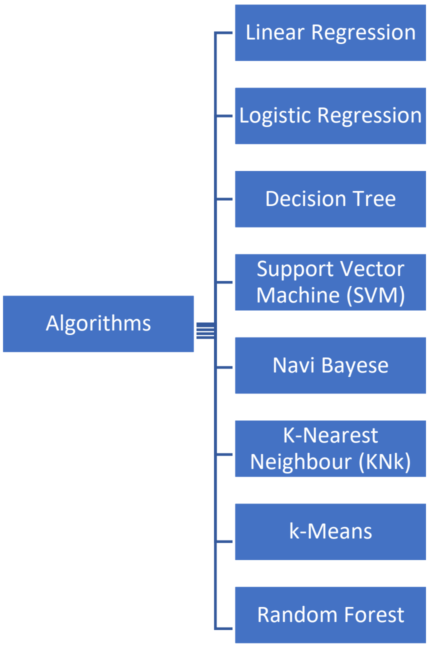

Fig. 3. Machine Learning Algorithms

4.1 REGRESSION

Regression analysis is a statistical tool for forecasting the predictions in the real time problems. There are various models that are used in machine learning for variety of statistical analysis to evaluate the input data and their features. There are two types of regression analysis first one linear regression and other is logistic regression. In linear regression only one line is drawn to separate the dependent and independent variable, independent variable is as input dependent variable as getting optimal solution. The mathematical model in linear regression is defined as

$$W = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \cdots + \delta_n Z_n, \qquad (1)$$

where,
$W$is the predicted value.
$\delta_0$is the bias term.
$\delta_1 + \delta_2 + \cdots + \delta_n$are the model parameters.
$Z_1 + Z_2 + \cdots + Z_n$are the feature values.
In order to minimized the error, the cost function is defined as the residual sum of square of the residuals and the cost function is mathematically expressed as below

$$k(\delta) = \frac{1}{2r}\sum_{j=1}^{r}(d(z^j) - w^i)^2, \qquad (2)$$

where, the hypothesis function $d(z)$is denoted by

$$d(z) = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \cdots + \delta_n z_n, \qquad (3)$$

and r is the total number of training examples in our data-set.RMSE is the square root of the average of the sum of the squares of residuals.RMSE is defined by

$$\text{RMSE} = \frac{1}{2r}\sum_{j=1}^{r}(d(z^j) - w^i)^2. \qquad (4)$$

The aim of the suggested model is the minimization of the cost function by evaluating the model parameters. In order to measure the accuracy of the proposed model certain parameters are defined in order to do so. R-square, adjusted R-Square, Root mean squared error (RMSE) and MAE have been calculated in the proposed model.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 11, November 2020

DOI 10.17148/IJARCCE.2020.91120

4.2 Logistic Regression

Logistic regression is used in ML which is basically a tool of statistical method which is modeled for prediction of data depending on previous information of the data-set. This approach is vital component of machine learning algorithms as method is used to classify the information based on the previous information. The accuracy of existing data is a important parameter for predicting the exact value of data set. Logistic regression being a classification algorithm is used to get the prior uncertain values of a binary variables. These binary variables are the dependent variables which can be defined as one that means true information and zero that means false information. Furthermore, the logistic regression model predicts P(U=1) as a function of Z.Considering a situation when it is to be classified that a patient is diabetic or not. When the linear regression is to be used for such problem, a threshold is essential on which the classification can be grounded. Consider if the actual class is malicious, having a predicted value of 0.3 and 0.6, as the threshold value then the particular data point is marked as non-malignant this kind of situation can lead to the stern results in real-times. With this instance, it is concluded that the linear regression does not hold good in case of classification problem. As the Linear regression is unbounded, it can give importance to the existence of logistic regression. This can range Their value firmly between zero and one. The mathematical model is delineate as

$$U=\frac{1}{1+e^{-(\delta_0+\delta_1 Z_1+\delta_2 Z_2+\cdots+\delta_n Z_n)}}. \qquad (5)$$

Following is the description of the parameters used
U is the response variable.
Z is the predictor variable.
δ_0 and δ_1 are the coefficients which are numeric constants.

V.Result and Discussion

In this paper, separate case studies have been taken to analyse the impact of linear regression and logistic regression as machine learning tools through python programming language to predict the future values which is briefly explained as follows. Case study using lung cap real data set is obtain from [10]. In this data six parameters are taken such as lung_cap, Age, Age, Height, Smoke, Gender, Caesarean. Lung_cap is depending upon the other parameters. Our aim is to give the best fit model using machine learning. For this purpose, python 3.0 is used. Methodology used is represented in Fig. 4. According to the model given in the regression theory, the model (6) is again delineate asand the computed value of coefficient of the model is shown in Table 2.

$$\text{Lung\_Cap} = \delta_0+\delta_1 Age+\delta_2 Height+\delta_3 Smoke+\delta_4 Gender+\delta_5 Caesarean. \qquad (6)$$

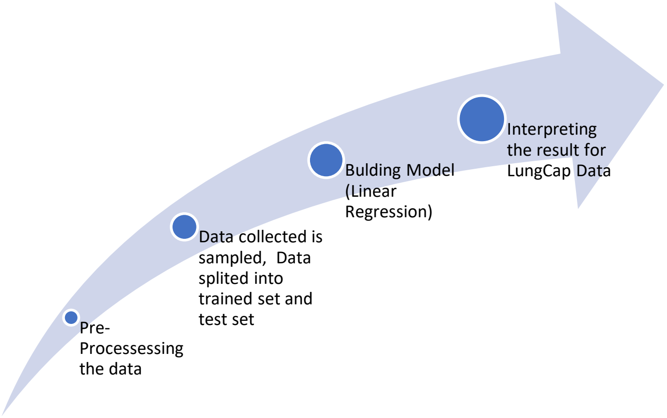

Fig. 4. Methodology for the Lung_Cap Data

R-square value is 0.85, adjusted R-square is 0.83, RMSE is 0.94, RMSE is 0.97, MAE is 0.77. Histogram of the given data is shown in Fig. 5 and Fig. 6 is representing the scattered plot predicted vs actual values.

TABLE 2 Coefficient of the model

Intercept δ_0	δ_1	δ_2	δ_3	δ_4	δ_5
-12.18	0.15	0.26	0.65	0.41	0.15

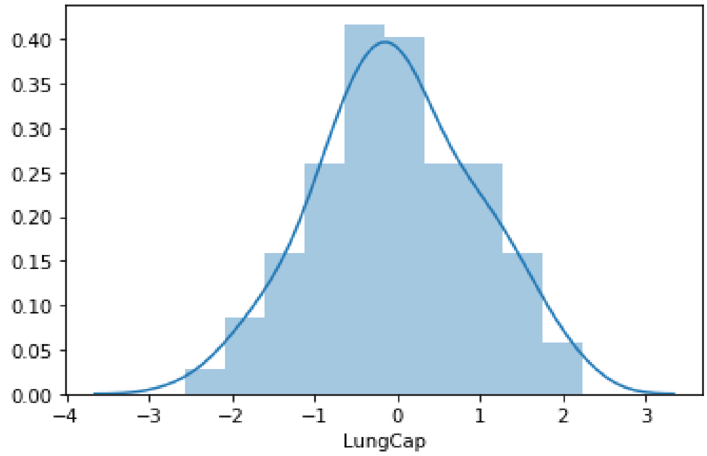

Fig. 5. Histogram of the Lung_Cap data

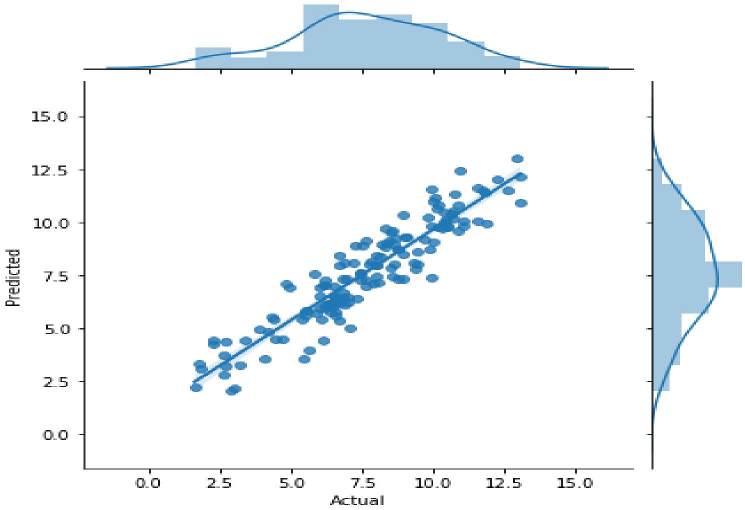

Fig. 6. The Predicted vs actual scattered plot of Lung_Cap data

Case study for logistic regression. In order to study the influence of logistic regression a case study has been done, to predict the loan applicant's nature for the sanction of loans [7]. It's a classification problem, given information about the application we have to predict whether the they'll be to pay the loan or not. We'll start by exploratory data analysis, then preprocessing, and finally we'll be testing different models such as Logistic regression and decision trees. The methodology used as shown in Fig. 7. Firstly, the consumer has to apply for such loans then-after the concerned loan sanctioning authority authenticates the entitlement of loan against the consumer. The automation for sanctioning the loan is done by the loan approving authority that depends on various real-time data which includes the customer particulars given by the concerned consumer during the filling application form online or offline. Such particulars are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History etc. In this process, the consumers such segments are analyzed, studied and those which are eligible in such circumstances are specifically targeted and approved for the next step. Another interesting variable is credit history, to check how it affects the Loan Status we can turn it into binary then calculate it's mean for each value of credit history.

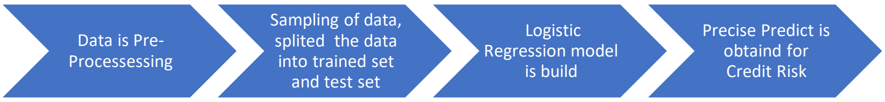
Data is Pre-Processessing
Sampling of data, splited the data into trained set and test set
Logistic Regression model is build
Precise Predict is obtaind for Credit Risk

Fig. 7. Methodology for the Credit risk Data

DOI 10.17148/IJARCCE.2020.91120

To validate the proposed model few parameters are calculated using the algorithm and confirm the accurateness of Binary Logistic Regression model. The initial step is to find the accuracy of the model then followed by construction of confusion matrix. The confusion matrix is a perception from ML which comprises statistics regarding the real classifications and forecasted classifications. This matrix possesses dual dimensions, first one is indexed by the actual type of an item, while the another is indexed by the type that the classifier forecasts. Secondly other key parameters are calculatedand when the model classification effect is bestthese are used to increase the precision of the system further.The below are few expressions calculated in order to find how accurate the prediction of the model is, with the inclusion of five model outcome estimation statistics, which are as tabulated in Table 3 as Accuracy, Recall, Precision, F1 score. Predicting the test set results and calculating the accuracy of the model is 77.15.

TABLE 3 DIFFERENT PARAMETERS USED IN MODEL PERDITION

Outcome	**Precision**	**Recall**	**F1-Socre**	**Support**
0	0.64	0.49	0.55	57
1	0.81	0.89	0.85	140
Accuracy	0.77	0.77	0.77	197
Average	0.72	0.69	0.70	197
Weighted Average	0.76	0.77	0.76	197

The receiver operating characteristic (ROC) curve is another common tool used with binary classifiers. It is a technique for envisaging, organizing and choosing classifiers dependent on their operation. This ROC curve is a dual-dimension curve having false positive rate (FPR) as the X axis and true positive rate (TPR) as the Y axis, ranging between (0,0) to (1,1). The ROC curve of an ideal random classifier is represented with respect to the dotted line. In order to check whether the given classifier is a good one, it should be as far away from that line as possible and should be placed towards the top-left corner of the graph as is shown in Fig. 8. In linear regression model the objective is to model parameters influencing the lung capacity. In this study it has been analyzed that the parameters like age, gender, smoking and height influence Lung capacity of individual as is also concluded from the findings of the results above and also from the fitted model. In case of logistic regression model, it can be inferred from the results that the loan was chosen by many of the applicants and the maximum of them have applied loan for liability consolidation. Both the predictive models are performing good prediction values and the accuracy is nearly close to one, which is required for a better performance.

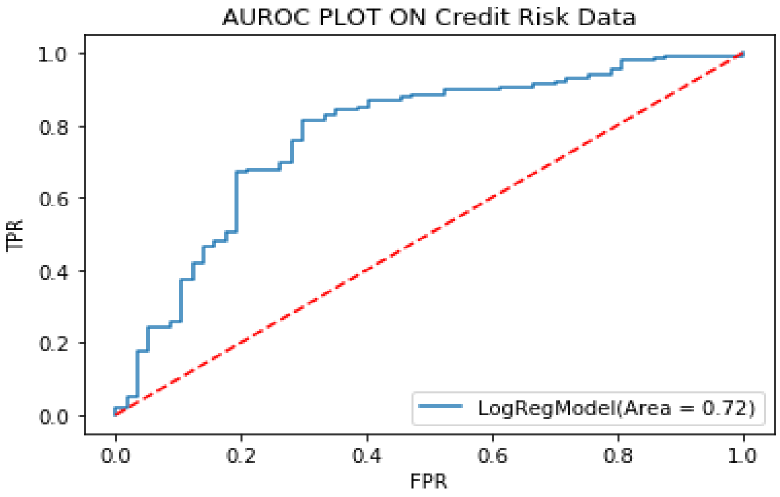

Fig. 8. ROC curve Plotting FRP vs TPR

VI.CONCLUSION

In this paper, ML algorithms are proposed for predicting the credit risk of consumers those have applied for loan and also for the prediction of factors influencing the lung capacity. In these training algorithms, the paper presents a collective approach for loan predications by using several parameters like accuracy for comparison. The objective of this paper is to assess the accuracy of models and give the best predictive values of the problems taken for analysis.

DOI 10.17148/IJARCCE.2020.91120

REFERENCES

[1]. L. Bernardi, T. Mavridis and P. Estevez, “150 Successful Machine Learning Models: 6 Lessons Learned at Booking.com.” Applied DataScience, vol. 2, pp. 1743-1751, (2019).
[2]. V. Boyarshinov, “Machine Learning in Computational Finance,” Ph.D thesis, Graduate Faculty of Rensselaer Polytechnic Institute, Troy, NY, April, 2005.
[3]. A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," in IEEECommunications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, Secondquarter 2016, doi: 10.1109/COMST.2015.2494502.
[4]. S. Das, A. Dey,A. Pal andN. Roy, Application of Artificial Intelligence in Machine Learning: Review and Prospect. International Journal of Computer Applications, vol.115, pp. 31-41, 2016.
[5]. K. Das, R. N. Behera, A Survey on Machine Learning: Concept, Algorithms and Application, International Journal of InnovativeResearch in Computer and Communication Engineering,vol.5, pp.1301-1309, 2017.
[6]. T. Dean, J. Allen and Y. Aloimonos, Artificial Intelligence: Theory and Practice, Benjamin/Cummings, New York, 1995.
[7]. Lone Prediction Problem Dataset. Available: https;//www.medium.com/altuistdelhite04/lone-prediction-problem-dataset.
[8]. R. Husain andR. Vohra, Applying machine learning in The Financial Sector, International Education & Research Journal, vol.3, pp. 19--20, 2017.
[9]. H. Kaur, G. Singh andJ. Minhas, A Review of Machine Learning Anomaly Detection Techniques, International Journal of ComputerApplications Technology and Research, vol.2, pp. 185-187, 2013.
[10]. M. Marin.(2015) Marin Stats. Lecture:https//www.lectures.com/videos-tutorials.
[11]. T. M. Mitchell. (2006) The Discipline of Machine Learning. Lecture: https://www.researchgate.net/publication/268201693_The_Discipline_of_Machine Learning.
[12]. F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, M. Ruffini and M. Tornatore, An Overview on Application of Machine LearningTechniques in Optical Networks. In: IEEE Communications Surveys & Tutorials, vol.21, 2, pp. 1383-1408, 2019.
[13]. A. Øland,“Machine Learning and its Applications to Music.” 2012.
[14]. V. S. Reshmi, R. Priyadharshini, M. Ramana andM. Suraj, Enactment of Artificial Intelligence in Machine Learning with Presumptionand Panorama, Waffen-und kostumkunde journal, vol.11, pp. 321-328, 2020.
[15]. P. Sajda, Machine Learning for Detection and Diagnosis of Disease, Annu. Rev. Biomed. Eng., vol.8, pp.537-65, 2006.
[16]. S. Shen, H. Jiang and T. Zhang,“Stock Market Forecasting Using Machine Learning Algorithms.” 2012.
[17]. O. Simeone, A Very Brief Introduction to Machine Learning with Applications to Communication Systems, In: IEEE Transactions onCognitive Communications and Networking, vol. 4, pp. 648-664, 2018.
[18]. A. Simon, M. Deo, S. Venkatesan and D. R. R. Babu, An Overview of Machine Learning and Its Application, International Journal ofElectrical Sciences & Engineering, vol.1, pp. 22-24, 2015.
[19]. A. S. Sreeja, R. R. Pasnoor and P. Y.S. Sai, A Study on Machine Learning Algorithm, IJRAR- International Journal of Research andAnalytical Reviews, vol.5, pp. 649—653,, 2018.
[20]. L. Spector, Evolution of artificial intelligence, Artificial Intelligence, vol.170, pp.1251–1253, 2006.
[21]. N. Wang, Bankruptcy Prediction Using Machine Learning, Journal of Mathematical Finance, vol.**7**, pp. 908-918, 2017.
[22]. P. Wang, X. Chen, F. Ye and Z. Sun, A Survey of Techniques for Mobile Service Encrypted Traffic Classification Using Deep Learning, In: IEEE Access, vol.**7**, pp. 54024-54033, 2019.
[23]. B. Wiese and C.Omlin, Credit Card Transactions, Fraud Detection, and Machine Learning: Modelling Time with LSTM Recurrent Neural Networks. In: Bianchini M., Maggini M., Scarselli F., Jain L.C. (eds) Innovations in Neural Information Paradigms and Applications. Studies in Computational Intelligence, Springer, Berlin, Heidelberg, 2009.
[24]. T. Ye and L. Zhang, Derivative Pricing via Machine Learning. Journal of Mathematical Finance, vol.9, pp.561-589, 2019.