Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Predicting Fee Defaulters Using Logistic Regression

Sudha Ramkumar A.

Associate Professor, Department of Computer Science,
Sri Kanyaka Parameswari Arts & Science College for Women, Chennai
Emailed:sudharam99@gmail.com

Predictive analytics is used to determine the future outcome using techniques like data mining, data modelling and machine learning. Predictive analytics uses the existing data of an organization to make predictions about future outcomes. Using predictive analytics, one can identify the risk as well as opportunities which will affect the organization in the future. In today's fast-growing environment, it is essential for organizations to anticipate the outcome, identify opportunities, and prevent loss. Predictive analytics plays a vital role in overcoming challenges and making organizations aware of the opportunities in advance. Students drop out ratio is increasing in the recent years; the predictive analytics techniques can be used to predict fee defaulter in educational institutions to prevent them from dropping their education at the earliest possible time. In this paper, Logistic regression model is used to predict the fee defaulter. Precision, recall, f1-score and accuracy are used as evaluation metrics to evaluate the performance of the regression model.

Keywords: predictive analytics, regression, logistic regression.

I Introduction

Every organization has set its goal and works towards that goal. Organizations can use predictive analytics in the early stage to identify the risk in achieving that goal. Predictive analytics uses regression analysis to predict the future outcome based on the available data. The organization has to decide what type of technique can be used to achieve the organization's goal. The type of predictive analytics technique to be used depends on the organization's goal. For the smooth operation of educational institutions, timely fee payment is very important. Late payments or fee defaults will increase the financial burden and it challenges the financial sustainability of the institutions. To address this issue, predictive analytics can be used in the institutions to predict the fee defaulter in advance, so that the institutions can take proactive actions.

The percentage of students pursuing higher education is less when compared to the students who have completed schooling education. During the pandemic time, most of the parents are finding it very difficult to pay the fees on time. Educational institutions can use predictive analytics to find fee defaulters using features like parents' income and they can improve their financial stability by reducing the fee defaults. This paper proposed predictive analytics for identifying fee defaulters using students' details. The predictive analytic model

will be evaluated by confusion matrix and evaluation metrics such as precision, recall, f1 score and accuracy to calculate its capability to predict the fee defaulters.

In the recent scenario, all the students who have completed schooling did not continue their higher education. The percentage of students completing the degree course is also decreasing because of drop outs. It is very important for the educational organizations to predict the fee defaulter, so that the drop outs can be minimized. Predictive analytics methods can be used to minimize the drop outs in educational institutions. The Logistic regression model is used in this paper to predict fee defaulter.

Most of the research in predictive analytics focuses on marketing and manufacturing of a business, health care, stock market and real estate business. This paper contributes to the educational institution by offering practical insights into how logistic regression can be used to enhance the fee management system.

This paper is organized as follows. Section II presents the classification of predictive analytics model and Section III provides the related work. Section IV explains the methodology used and Section V provides the results and discussion. Section VI concludes with the future scope.

II Classification of Predictive Model

The predictive analytics model can be classified into 4 types

1. Regression Model,
2. Classification Model,
3. Clustering Model,
4. Time Series Model.

2.1 Regression Model,

The regression model estimates the outcome based on the relationship between two variables. Regression model estimates the outcome based on the relationship between two variables. For example, a regression model can be used to identify the performance of a student based on her/his hours of study per day. If the study hours increase, the performance also increases. Regression model is simple when it uses one independent variable and one dependent variable. Regression model is used to study how actions i.e the independent variable affects the outcome i.e dependent variable and uses this information to predict the future outcome. Regression model can also be used for multiple independent variables. The independent variable is the value to be changed and the dependent variable is how it reacts to that change in the independent variable. For example, to predict the price of a flat by using the location, accessibility to the city, square feet of the flat.

2.2 Classification Model

This model places the data into categories based on the past available knowledge. classification model starts with the training dataset where the data has been already labelled. the

classification algorithm learns the correlation between the data and the label and categorizes new data. Some of the widely used classification models include decision trees, random forest, and text analytics. Organizations use a classification model to identify whether the customer is good or bad. Banking sectors use a classification model to find fraudulent customers and they will use this information to alert customers when any suspicious activity has been made in their account in the future.

2.3 Clustering Model

This model places the data into groups based on similar attributes. A clustering model groups the data into clusters where each cluster has similar characteristics of data and data between two clusters have distinct characteristics. The clustering model uses a matrix where attributes are columns and groups the data based on similar features. Organizations use clustering to classify customers into good or bad and they will apply some personalized strategies on the customer segmentation. For example, Amazon classified customers into prime customers and it provides a lot of benefits to the prime customers.

2.4 Time Series Model

This model groups the data in relation to time because most of the real-world data are modelled as a time series, being time as the independent variable. A model will use the last year's data to analyse a metric and then predict that metric for the upcoming week. Time series model is well suited for predicting the stock market prices, predicting temperatures, to forecast energy consumption and to predict the future product sales.

Choosing the right Predictive model category for a given problem depends upon the nature of the data, the problem and the outcome. Regression model is used for predicting the continuous outcome whereas classification is used for predicting the categorical outcome. Clustering is used to group data based on the similarity whereas time series model focuses on the time dependent patterns. Each and every category of predictive model consists of a collection of algorithms and methods are available for the different types of data and outcome.

III Related Work

Shin, Seung-Jun et al. presented the design of a big data analytics model accompanied with the identification of its functional architecture. This work presented contributes towards 1. the big data model for the machining process as a starting point for manufacturing process analytics.2. The expansion of the composite analytic model to enable data driven planning and control with faster decision making and 3. The utilization of open platform tools makes SM possible for small and medium sized manufacturers [7].

Mohanty, A., and P. Ranjana described the development of predictive models in an industry using two of the existing algorithms such as time series and logistic regression algorithm. This predictive model is implemented in the Python and R language. Predictive analytics can be useful for demand forecasting, defect detection, maximizing equipment value, preventive maintenance, optimize marketing strategies, retain customer and connected aftermarket service in industry [5].

Meyberg, Camilo et al., developed a model to predict rental prices per square meter based on the accommodation's features and location within the city. This paper used the housing sale data from Ames Iowa for the time frame 2006-2010 to construct the various models to predict the final price of a flat [4]. This paper detected offers which appear in both portals by means of statistical matching and removed duplicate offers. Missing values were treated by multiple imputation. The prediction model is a semi-parametric approach where the postal districts are used to describe the location effect. Comparisons with micro census results and the local rent index reveal significant differences between the market of online flat offers and the stock of existing flat contracts. Interested readers will find the commented programming code in the internet supplement.

Swani, Lakshay, and Prakita Tyagi provided predictive modelling analytics with trends, techniques and. Data mining applications through data mining. Data mining leads to predictive analytics and is becoming key to every organization as it can be applied under various circumstances so as to highlight the growth of the organisation. The growth of the organization is possible with the help of expansion of business using predictive analytics aid and as well as presents the degradation through analysis of fraudulent activities [10].

Jeon, Seungwoo, Bonghee Hong et al. suggested a new complex methodology to find the optimal historical dataset with similar patterns according to various algorithms for each stock item and provides a more accurate prediction of daily stock price. This paper used a Dynamic Time Warping algorithm to find patterns with the most closely similar situation adjacent to a current pattern and generated an artificial neural network model with selected features as training data for predicting the best stock price. This paper also used Jaro-Winkler distance with Symbolic Aggregate approXimation (SAX) as prediction accuracy measure to verify the model [2].

Boukenze, Basma, Hajar Mousannif et al. presented an overview of big data in the health care sector with the application of predictive analytics. The application of predictive analytics in healthcare is very essential because predictive analytics predict life threatening diseases earlier. In this work, c4.5 algorithm is used to predict the patients with chronic kidney failure disease. The classifier proved its performance in predicting with best results in terms of accuracy and minimum execution time[1].

Sohrabi, Babak, Iman Raeesi Vanani et al. examined the information system articles in order to design a predictive solution to cluster a group of similar documents together and predict the area of k of a given set of articles, based on the previously validated learning. K-Means clustering has been used to reach the first objective and different classification algorithms have been used to reach the second objective of results. The results have shown great promise on the actual trends of information systems and IT developments in the near future regarding the soft aspects of information technology [9].

Seyedan, Mahya, and Fereshteh Mafakheri presented a study of predictive analytics in supply chain management. This study performed a thorough review for application of

predictive analytics. The survey overviewed the Big data analytics methods applied to supply chain demand forecasting and provided a comprehensive categorization of them [6].

Singh, Archana et al. used big data analytics to predict the price of the real estate. The housing Sale Data from Ames, Iowa is considered for the timeframe 2006–2010 with a view to construct relevant models to estimate the final sale price of a house. Due to the high number of explanatory variables several models such as Logistic regression, random forest and gradient boosting models have been used as tools for feature selection to determine the statistically significant characteristics that influence the final sale price of a house. It has been observed that out of all the models, the gradient boosting model returned the efficient results [8].

IV Methodology

Predictive analytics model is built once the past dataset is collected. Using existing predictive algorithms and statistical models, the dataset is trained. This paper uses Logistic regression because the problem is to find the fees defaulters. The Logistic regression model was chosen for the reason is to predict the fee defaulters, the variable parents' annual income is used. Here, the dependent variable is fee defaulter and the independent variable is parent income. The Logistic regression model stimulates the relationship between these variables. Predictive analytics model consists of the following phases,

1. Data Collection
2. Data preparation
3. Exploratory data analysis
4. Developing Predictive model
5. Evaluating the Model
6. Deployment of the Model
7. Comparing the prediction with Actual

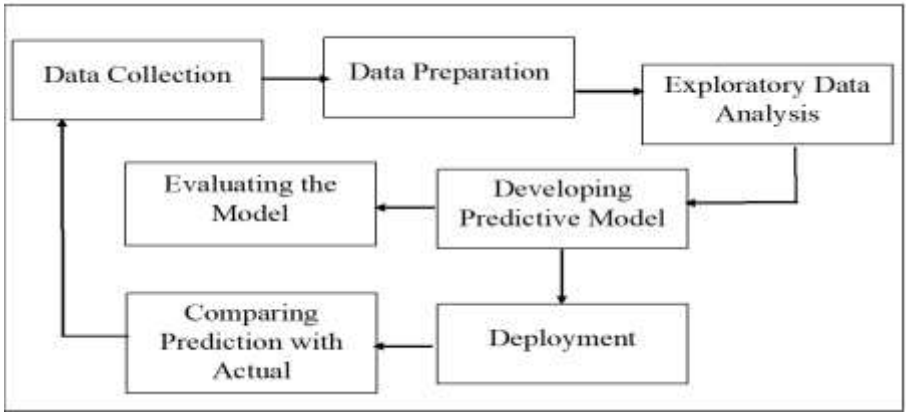

Figure 1. Methodology

Data collection step is to collect the required information for the analysis. This paper used the students' details. Students' details consist of the student's name, age, father occupation, father income etc. The students' details have been stored in the excel file and it is saved in the csv format. This file consists of 1088 students' data. The fees paid or not variable is having two values 1 or 2 where 1 represents fees paid and 2 represents fee defaulter.

Data preparation phase deals with the data cleansing process. Here, the redundant and identical data are removed from the details. This phase also removes the incomplete data or incorrect data from the dataset.

Exploratory data analysis is the important phase of predictive analytics. This phase helps to understand the variables and because of that, variables can be chosen for further analysis. A lot of insights into the variable is generated during this step and it improves the understanding of predictive analytics model.

Developing the predictive model phase deals with the implementation part of the model. In this phase, data visualization tools are used to find the hidden relationships between variables and it is used to set up the predictive model. This paper used Python to implement the Logistic regression predictive model.

Evaluating the model is used to assess the model's accuracy for the given set of students' details. Series of test runs are executed to check how well the model predicts the outcome. Validation is a very important feature in this phase. Confusion matrix is displayed for the logistic regression model for both the test data and train data. Confusion matrix consists of four terms such as true positive, true negative, false positive and false negative. The true positive means correctly predicted positive instances whereas true negative means correctly predicted negative instances. False positive means incorrectly predicted as positives whereas false negative means incorrectly predicted as negatives.

Using the four terms of confusion matrix, the evaluation metrics precision, recall, f1 score and accuracy can be calculated. In this paper, these four-evaluation metrics has been used for evaluating the performance of the regression model. The confusion matrix is shown in the table 1.

Detail	**Predicted 0**	**Predicted 1**
Actual 0	**TN**	**FP**

Actual 1	**FN**	**TP**

Table 1: Confusion Matrix

Precision provides the number of true positive predictions among all positive predictions made by the logistic regression model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{----------(1)}$$

The recall provides the number of true positive predictions out of all actual positive instances in the dataset.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad \text{----------(2)}$$

The f1 score provides a balanced measure for both false positive and false negatives.

$$\text{F1} - \text{Score} = \frac{\text{2XPrecisionXRecall}}{\text{Precision+Recall}} \qquad \text{----------(3)}$$

The accuracy measure provides the ratio of correct predictions that is true positives and true negatives to the total number of instances in the dataset.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+FN}} \qquad \text{----------(4)}$$

Deployment phase is used to test a model in a real-world situation which helps in practical decision making and makes it ready for implementation. In this paper, the model is built for the use of educational institution and it is implemented in python programming language.

Comparing the prediction with the actual phase is used to check the performance of the model constantly to ensure that the model receives the best future outcomes possible. It involves comparing model predictions to actual data sets. With the help of confusion matrix, the comparison of predicted data with actual data is easily interpreted along with the evaluation metrics.

V Results and Discussion

Logistic Regression is one of the most widely used regression model in the field of machine learning and statistics. The students' details are stored in csv file and it consists of students' details such as name, age, father name, annual income and fees paid or not. In this paper, the Logistic regression is implemented in python with student details as dataset. The Annual

income is given as independent variable and fees paid or not is given as dependent variable; the Logistic regression model is built for the students' details. Precision, recall, f1-score and accuracy are used as evaluation metrics to evaluate the performance of the Logistic regression model. The experimental results proves that the regression model built for predicting the fee defaulter is good. The confusion matrix is shown in the Figure 2.

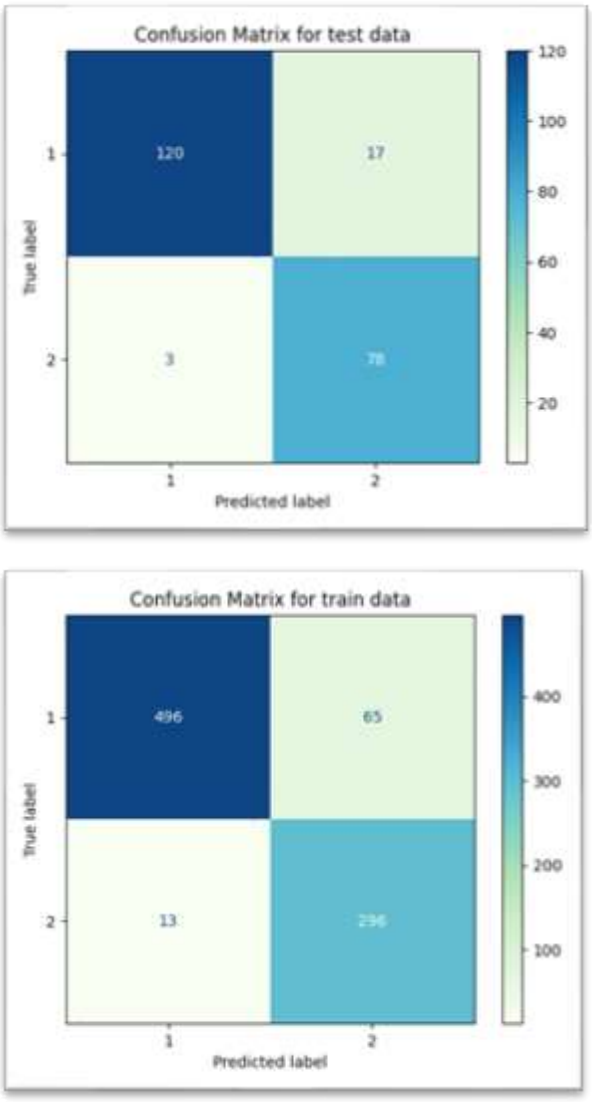

Figure 2: Confusion Matrix for the test data and train data

The model is implemented in Python program for the logistic regression model. The regression model provides the following results,

Precision for test data: 0.976

Recall for test data: 0.876
F1 for test data: 0.923
Accuracy for test data: 0.908
Precision for train data: 0.974
Recall for train data: 0.884
F1 for train data: 0.927
Accuracy for train data: 0.910

The evaluation metrics such as precision, recall, f1 score and accuracy for the test data and train data is compared in the table 2.

Evaluation Metric	**Test Data**	**Train Data**
Precision	97.6	97.4
Recall	87.6	88.4
F1	92.3	92.7
Accuracy	90.8	91

Table 2: Comparison of test data and train data

From the table 2, it is very clear that the model predicts well and it is almost similar for both the test data and train data. The following figure 3. Shows the evaluation metric comparison of test data and train data.

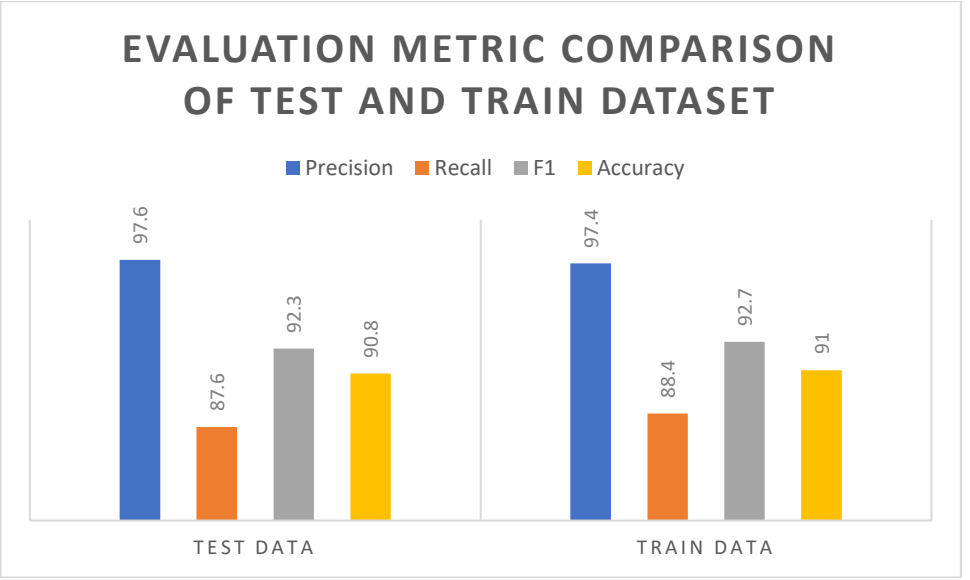

Figure 3: Comparison of evaluation metric of Test and Train dataset

The following figure 4. shows the logistic regression model for predicting the fee defaulter.

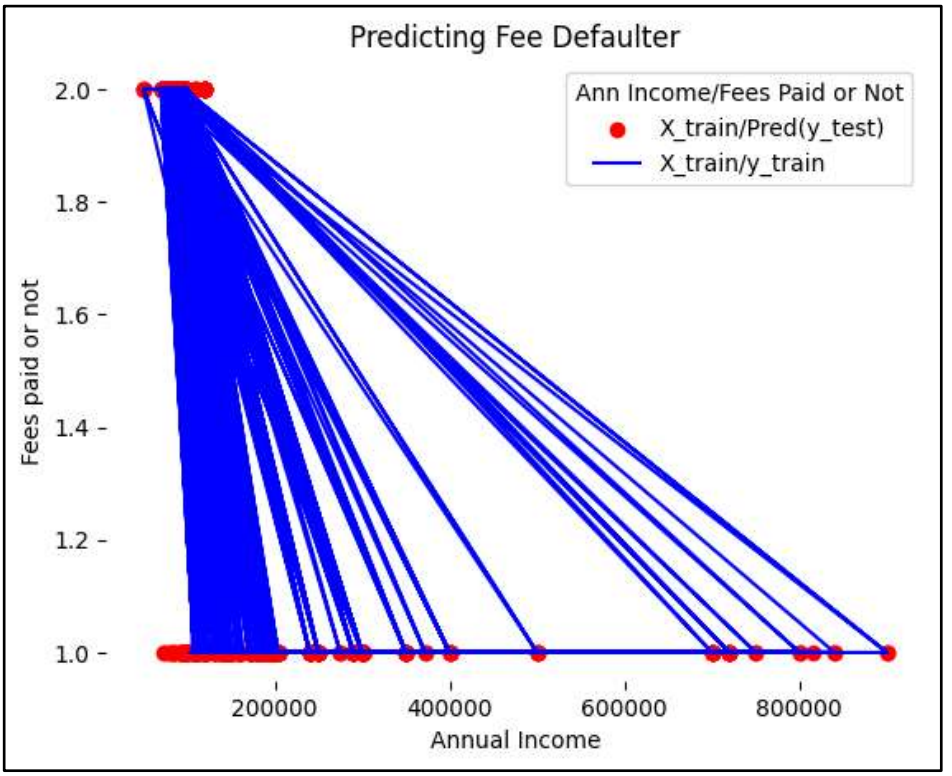

Figure 4: Logistic Regression Model for Predicting Fee Defaulter

VI Conclusion

In this paper, the Logistic regression model is used for predicting the fee defaulter with annual income as independent variable and fees paid or not variable as the dependent variable. The actual outcome is compared with the predicted outcome and it is clear that the model predicted the fee defaulter well and it was very clear with the evaluation metrics used in this paper. In future, the regression model can be built using multiple independent variables and it can be compared with the actual outcome with the predicted outcome. Regression model can be built using advanced techniques such as random forests and Neural Network to improve the accuracy of the prediction.

References

1. Boukenze, Basma, Hajar Mousannif, and Abdelkrim Haqiq. "Predictive analytics in the healthcare system using data mining techniques." Comput Sci Inf Technol 1 (2016): 1-9.
2. Jeon, Seungwoo, Bonghee Hong, and Victor Chang. "Pattern graph tracking-based stock price prediction using big data." Future Generation Computer Systems 80 (2018): 171-187.
3. Mahmoud, Fatimetou Zahra Mohamad. "The Application of Predictive Analytics: Benefits, Challenges and How It Can Be Improved." International Journal of Scientific and Research Publications 7.5 (2017): 549-566.
4. Meyberg, Camilo, Ulrich Rendtel, and Holger Leerhoff. "Flat rent price prediction in Berlin with web scraping." AStA Wirtschafts-und Sozialstatistisches Archiv (2024): 1-34.
5. Mohanty, A., and P. Ranjana. "Usage of predictive research on further business." International Journal of Innovative Technology and Exploring Engineering 8.11 (2019): 3464-3466.
6. Seyedan, Mahya, and Fereshteh Mafakheri. "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities." Journal of Big Data 7.1 (2020): 53.
7. Shin, Seung-Jun, Jungyub Woo, and Sudarsan Rachuri. "Predictive analytics model for power consumption in manufacturing." Procedia Cirp 15 (2014): 153-158.
8. Singh, Archana, Apoorva Sharma, and Gaurav Dubey. "Big data analytics predicting real estate prices." International Journal of System Assurance Engineering and Management 11 (2020): 208-219.
9. Sohrabi, Babak, Iman Raeesi Vanani, and Mohsen Baranizade Shineh. "Designing a Predictive Analytics Solution for Evaluating the Scientific Trends in Information Systems Domain." Webology 14.1 (2017).
10. Swani, Lakshay, and Prakita Tyagi. "Predictive Modelling Analytics through Data Mining." International Research Journal of Engineering and Technology (IRJET) 4.09 (2017): 2395-0056.