The Default of Credit card client

Hetarth Bhatt
Master Engineering
Electrical & Computer
Dept.
Western University,
London, Canada
hbhatt7@uwo.ca

Vatsal Shah
Master Engineering
Electrical & Computer
Dept.
Western University,
London, Canada
vshah56@uwo.ca

Rajaraman Ganesan
Master Engineering
Electrical & Computer
Dept.
Western University,
London, Canada
rganesa@uwo.ca

Khushali Patel
Master Engineering
Electrical & Computer
Dept.
Western University,
London, Canada
kpate372@uwo.ca

***Abstract*-- In modern day's credit card plays an important role in every person's daily activity. Customer purchases their needs with their credit cards and online transitions. Banks and financial institutes consider denying the credit applications of customers to avoid the risk of defaulters. Credit risk is the rise of debt on the customer who fails to make the billing payment for some period. The purpose of the project is how to reduce the defaulters among the list of customers, and make a background check on whether to provide the loan or not and to find the promising customers. These predictive models would benefit the lending institutions and to the customers as it would make them more aware of their potential defaulting rate. The problem is a binary classification problem whether a customer will be defaulting to pay next month payment. The dataset is unbalanced so the focus was on the precision and recall more than the accuracy metrics. After comparison with precision-recall curve, logistic regression is the best model based on the False Negative value of confusion metrics. Moreover, after changing the threshold value of the logistic regression, GUI (Graphical user interface) implemented and predicted whether a customer is defaulter or not-defaulter.**

Keywords-- *Defaulters, Credit risk, Logistic regression, SVM, Decision Tree, Naïve bayes, Feed forward Neural Network, KNN, Ensembled learning, Voting classifier, Precision, Recall, Accuracy*

I. Introduction

Credit card is a physical card used for paying our bills easily. The cardholder could use it to give a paying promise as a requital to the cost of service and goods. There is a brief explanation of algorithms to define term credit scoring [1], which determines the relation between defaulters and loan characteristics. It is a useful information for financial institution to maintain financial statement and customer transaction list to reduce the uncertainty. Yeh and Lien (2009) compared the predictive accuracy of probability of default among six data mining methods (specifically, K-nearest neighbor classifier, logistic regression, discriminant analysis, naive Bayesian classifier, artificial neural networks, and classification trees) using customers default payments data in Taiwan. Their experimental results indicated that only artificial neural network could accurately estimate default probability. The use of Taiwan data is beneficial for us because the sample size of the default payment data in Taiwan is 30,000. [2]

Currently, a variety of Machine Learning approaches used to detect fraud and predict payment defaults. Some of the more common techniques include Logistic Regression, K Nearest Neighbor, Decision Tree, Naive Bayes, Support Vector Machine, Feed Forward Neural Networks and Ensemble approaches like Voting Classifier. The dataset contains information on 24 variables, obtained from the UCI Machine Learning Repository. Here we categorized the dataset based on independent variables such as credit amount, age, sex, education, marital status, and their past loan repayment history of last 6 months, History of their past payments made (April to September), amount of bill statement, amount of previous payment. The dependent variable is default, which means whether the customer will pay their next month payment, or not. We can reduce the cost, make a good decision for a potential customer and help in reducing the time consumption for processing loan application and more.

A. *Aims & Objectives*

The problem is to classify the defaulters and non-defaulters on the credit payment of the customers. This project is helpful for solving the real problem by using various classification techniques. Moreover, any user can access GUI and add their gender, education, marital status and payment details to check next month in which category they fall (defaulter or non-defaulter).

The core objectives: Find whether the customer could pay back his next credit amount or not and Identify some potential customers for the bank who can settle their credit balance.

The steps followed to manage these goals:

- Selection of dataset
- Display some graphical information and visualize the features.
- Check Null values in the dataset
- Data pre-processing using one-hot encoding and remove extra parameters
- Train with classifiers
- Evaluate the model with test data
- Compare the accuracy, precision and recall finding the optimal model.
- Created a Graphical User Interface to check with real time customer data and predict defaulter for their next month payment

B. *High-level overview*

The major purpose of risk prediction is to use information, such as financial statement, customer transaction and repayment records to predict individual customer's credit risk and to reduce the damage and uncertainty. Many methods, including Logistic Regression, SVM, KNN, Decision Tree, Naive Bayes and Feed Forward Artificial Neural Networks used to develop models of risk prediction. [4]

The remainder of this paper organized as follows. Section 2 summarizes the basic properties of applied models and accuracy

3 explores the methodology with data preprocessing. Section 4 comprises of evaluation process. Section 5 presents summary.

II. BACKGROUND

There is much research on credit card lending. It is a widely researched subject. Many statistical methods have applied to developing credit risk prediction, such as Logistic Regression, Support Vector machine, K-nearest neighbor classifiers, probabilistic classifiers such as Bayes classifiers and neural networks and ensembled classifiers such as Voting Classifier.

A. K-nearest neighbor:

K-nearest neighbor (KNN) is one of the simplest supervised classifiers. The vision is to define K centroids, one for each cluster. These centroids placed inappropriately because of different location causing different results. Selecting the value of K is more critical part as a small value of K means that noise will have a higher influence on the result (probability of overfitting is high) and on another side, the higher value of K defeat idea to find the nearest value and lead to a greater amount of time & underfitting of model. When given an unknown data, the KNN classifier searches the pattern space for the KNN, which are the closest to this unknown data.

B. Gaussian naïve Bayes (GNB)

The Bayesian classifier is a probabilistic classifier based on Bayes theorem. Naïve Bayes classifier assumes that all features are unrelated to each other and more useful for predictive modeling. In practice, however, dependences can exist between variables. Advantages of naïve bayes are easy and fast to predict the class of test dataset. It performs well in multi-class prediction. Limitation of naïve bayes is the assumption of independent predictors, is hard to get a set of predictors, which are completely independent. [5]

C. Logistic Regression:

Logistic Regression is a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predictors. Binary logistic regression used when the dependent variable is binary or has two levels. In logistic regression, the purpose of the analysis is to assess the effects of multiple explanatory variables, which can be numeric or categorical. The goal is to find the best fitting model to describe the relationship between the binary characteristic of interest and a set of independent variables. [5]

D. Decision tree:

A decision tree is a flowchart-like structure in which each inward node represents a "test" on an attribute, each branch represents the consequence of the test, and each leaf node represents a class label. In decision analysis, a decision tree and the closely related influence diagram used as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated.

E. Support Vector Machine

Support vector machine is a popular machine learning classification algorithm. SVM used as supervised learning when the dataset has features and class labels. The Linear classifier implemented in a code. A focus is to maximize the distance from hyperplane to the nearest data point of either class in SVM; the maximum-margin hyperplane determined by the dataset lies nearest to it. These data points which influences hyperplane knows as Supper vector. When data separated linearly, draw two parallel hyperplanes, which separate two classes of data. The Distance between two hyperplanes is 2/||w||, to maximize this distance denominator value should be minimized i.e., ||w|| should be minimized. (Shows in image 1.1). The different kernel is available as linear, poly, sigmoid and wrong choice of the kernel can lead to an increase in error percentage. [6]

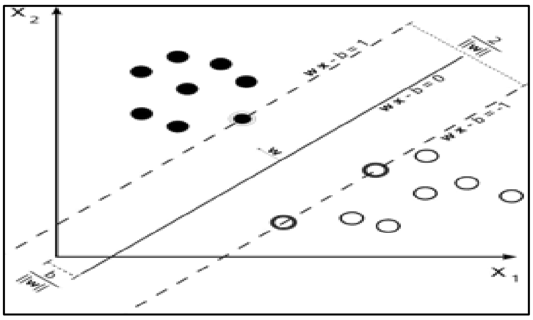


Fig (1.1) Support Vector Machine with hyperplane and two classes

F. Feed Forward Artificial Neural Networks:

A feedforward neural network is a type of artificial neural network. It can perform several classification tasks at once, although commonly each network performs only one. The best result is usually to train separate networks for each output, then to combine them into an ensemble so that they can run as a unit. In Feedforward neural network, all nodes connected in a network. Predictions based on the input nodes and weights. As the name suggests, activation flow is from the input layer to the output layer. There is one or more hidden layer between the input and output layer and no cycle or loop into the network. One neuron called a perceptron. A perceptron consists of one input layer and one neuron. A total node in the input layer is the same as total features in the dataset. Each input multiplied with random weight value where a weight is in the range of 0 to 1. An activation function knows as summed weighted input to the output of neurons. Neurons have activation function such as a step function, sigmoid, relu, softmax or tanh function. Bias added to the sum of input and weight to avoid null values. [7]

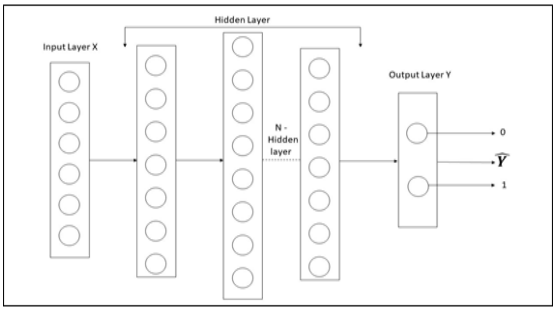


Fig (1.2): Neural Network of this dataset

G. Ensemble Learning:

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Voting Classifier is one of the simplest ways of combining the predictions from multiple machine learning algorithms by first creating two or more standalone models from your training dataset. A Voting Classifier used to wrap your models and average the predictions of the sub-models when asked to make predictions for new data. The predictions of the sub-models weighted, but specifying the weights for classifiers manually or even heuristically is difficult

H. Accuracy Measures:

a. Accuracy:

The accuracy of a model is usually determined after the model parameters learned and fixed and no learning is taking place. Then the test samples fed to the model and the number of mistakes (zero-one loss) the model makes recorded, after comparison to the true targets. Then the percentage of misclassification is calculated. In our dataset, accuracy determine how often the model predicts defaulters and non-defaulters correctly.

b. *Precision:*

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. If the precision is high, then there will be low false positive rate. Here precision tells us that whenever our models predicts it is defaulter how often it is correct.

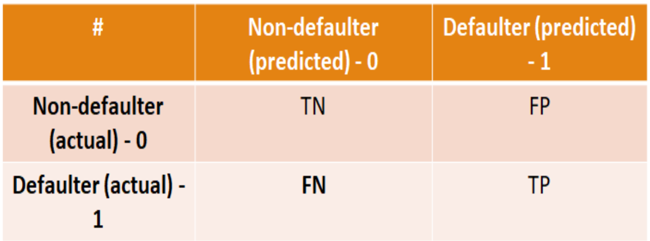
c. *Recall:*

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. In other words, out of all positive class how much we have predicted correctly. When we apply this in our dataset, it shows the actual defaulters that the model will actually predict.

d. *Precision Recall Curve:*

It will measure the success of prediction, when classes are imbalanced. It will show the tradeoff between precision and recall threshold. [8]

Table (1.1) Precision Recall Curve



#	Non-defaulter (predicted) - 0	Defaulter (predicted) - 1
Non-defaulter (actual) - 0	TN	FP
Defaulter (actual) - 1	**FN**	TP

Loss: Loss functions let the optimization function know how well it is doing. Loss functions used in the output layer, Layers that support unsupervised layer wise pre-training.

a. *Cross Entropy loss:* Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label.

b. *Binary Cross Entropy:* In binary classification, where the number of classes equals 2 either 0 or 1, then it is known as binary cross entropy. [9] Binary cross-entropy calculated as:

$$-y(\log(p) + (1 - y)\log(1 - p)$$

III. METHODOLOGY

In methodology, data description, independent variable and dependent variable described with scale of variables. Moreover, in the process data preprocessing and feature engineering described as below.

A. Data Description:

This dataset consists of 30000 total instances and 25 features including-

Table (1.2) attributes of the selected dataset

Independent Variable	Description	Scale of variable
Limit_ BAL	Amount of the given credit (NT dollar)	Continuous Interval
Sex	Gender (1 = male, 2 = female)	Categorical Nominal
Education	Education (1 = graduate school, 2 = university, 3 = high school, 4 = others)	Categorical Nominal
Marital Status	Marital status (1 = married, 2 = single, 3 = others)	Categorical Nominal
Age	Age (year)	Continuous Interval
PAY_0 to PAY_6	April to September	Categorical
Bill_AMT1 to Bill_AMT6	Amount of bill statement (NT dollar)	Continuous Interval
Pay_AMT1 to Pay_AMT6	Amount of previous payment (NT dollar)	Continuous Interval

Dependent Variable	Description	Scale of variable
is default	Default payment (Yes = 1, No = 0)	Binary

The total number of customer based on defaulter and non-defaulter from a dataset.

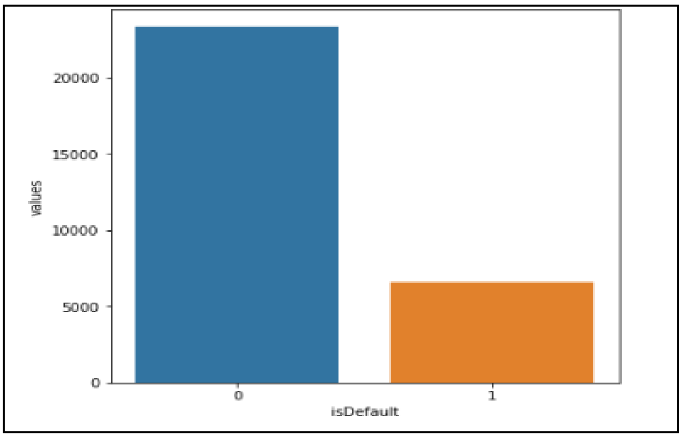


Fig (1.3) No: of defaulters and non-defaulters

B. Process:

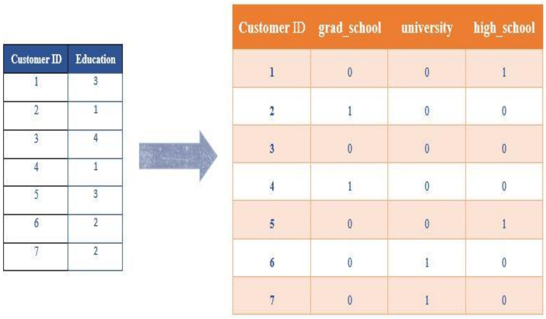
The first step is data preprocessing. Data preprocessing used to convert the raw data into a clean data set.

- ID column dropped as its unnecessary for our modeling.
- The attribute name ‘PAY_0’ converted to ‘PAY_1’ for naming convenience.
- Numeric attributes converted to nominal.
- One hot encoding which is a process by which categorical variables converted into a dummy form that provided to algorithms to do a better job in prediction. One hot encoder used to perform linearization of data. For instance, value in the ‘EDUCATION’ variables were grouped such that the values ‘0, 4, 5, 6’ was combined to one value and assigned a value ‘4’.

Table (1.3) one hot encoding

EDUCATION	ENCODING
Grad School	1
University	2
High School	3
Others	4

Converting categorical features into (n-1) features. Customer ID 1 has education value 3 which is converted to 0,0,1 as 1 is assigned to high school. Likewise, for gender Male and Female respectively 0 and 1. For Marital status, there are 4 categorical values as 1 means married, 2 means single and 3 means others. As in the dataset, there is no description about value 0, so we converted to value 3 as others. So, One-hot encoding is applied to education, gender and marital status.



Customer ID	Education
1	3
2	1
3	4
4	1
5	3
6	2
7	2

Customer ID	grad_school	university	high_school
1	0	0	1
2	1	0	0
3	0	0	0
4	1	0	0
5	0	0	1
6	0	1	0
7	0	1	0

Fig (1.4) One-Hot encoding for categorical column education

- Robust Scaler is used which converts all the variables in the same scale so if the data contains many outliers, scaling using the mean and variance of the data is likely to not work very well then in such cases Robust Scaler is used. For example, in Limit Balance column there are different range of values, which are converted, in proper scale.
- For all classification tasks, target variable converted to numeric.
- Next step is data preparation or feature selection where features selected by declaring the independent and target variable. Different graphs like count plots and pair plots are plotted with the reference to the target variable to check the default (=0) and non-default (=1).
- Before applying algorithms on train data, dataset is split into a ratio of 60:40, which is 60% train data and 40% is test data
- Next step is to train data by applying different algorithms as Support Vector Machine, K-Neighbors Classifier, logistic regression, Gaussian Naïve Bayes and artificial neural network.

Cross-validation: Cross-Validation used to assess the predictive performance of the models and to judge how they perform outside the sample to a new dataset also known as test data the reason to use cross-validation techniques is that when we fit a model, we are fitting it to a training dataset. Without cross-validation, we only have information on how our model performs in-sample data. Ideally, we would like to see how the model performs when we have new data of customers. [10]
In cross-validation process, K-fold cross validation is used. In K-fold cross validation all observations are used for both training and validation process. Normally 10-fold cross validations process is used. (Step 10). The general process of K-fold validations is to Shuffle the dataset randomly and Split the dataset into k groups (k=10)

For Neural Network, the following are tuning parameters:

Epochs: One epoch is when an entire dataset passed forward and backward via NN once. Here epoch value is set to 100.
Activation function:
ReLU: ReLU is commonly used activation function for deep learning. This has value range from zero to infinity.

$$R(x) = \max(0, x)$$

Sigmoid: A sigmoid function is a differentiable, real function that defined for all real input values and has a non-negative derivative at each point.

$$F(x) = \frac{1}{(1+e^{-x})}$$

SGD: Stochastic gradient descent is an iterative method for optimizing a differentiable objective function. Adam optimizer used in this project.
Input layer: Input layer is the very beginning of the workflow for neural network. 26 neurons used in input layer.
Hidden Layer: Hidden layer is in between Input Layer and Output Layer. 2 hidden layers are used after applying 1,2,3 hidden layer and found overfitting issue as we increased hidden layers.
Output Layer: It is a predicted feature value or output variables. It is an outcome. In this dataset, there are 2 neurons in an output layer.

IV. EVALUATION

We have applied various supervised algorithm techniques for the dataset; we have tabulated the value of accuracy, precision, recall, and confusion matrix for every algorithm respectively shown below:
Table (1.4) Tabulation for accuracy, precision, recall for various algorithms

#	Algorithms	Accuracy	Precision	Recall	Confusion Metrix
-	Null	78	-	-	-
1	Logistic Regression	81.45	66.92	35.95	$\begin{bmatrix} 8927 & 419 \\ 1806 & 848 \end{bmatrix}$
2	KNN	78.86	53.47	34.17	$\begin{bmatrix} 8557 & 789 \\ 1747 & 907 \end{bmatrix}$
3	Naïve Byes	76.68	47.65	52.23	$\begin{bmatrix} 7736 & 1610 \\ 1188 & 1466 \end{bmatrix}$
4	Classification Tree	78.46	52.09	32.81	$\begin{bmatrix} 8545 & 801 \\ 1783 & 871 \end{bmatrix}$
5	SVM	81.66	63.99	39.11	$\begin{bmatrix} 8762 & 584 \\ 1616 & 1038 \end{bmatrix}$
6	Feed Forward NN	75.65	33.91	40.74	$\begin{bmatrix} 8927 & 419 \\ 1806 & 848 \end{bmatrix}$
7	Voting Classifier	83.95	67.49	32.83	$\begin{bmatrix} 8842 & 504 \\ 1822 & 832 \end{bmatrix}$

The graphical representation shown below to have a better understanding of the accuracy, precision and recall we have achieved using various algorithms.

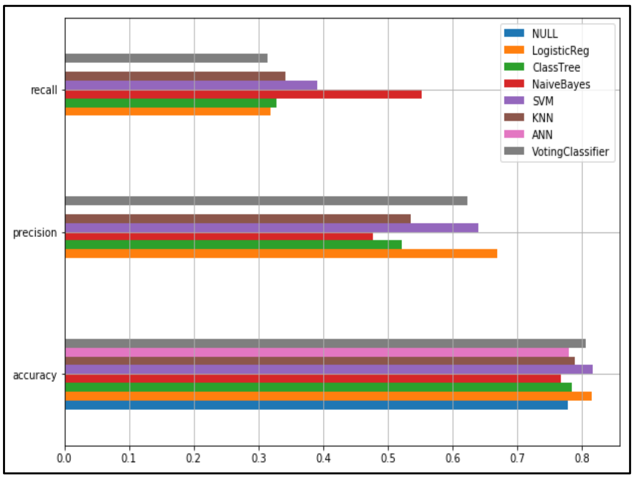


Fig (1.5) Accuracy, precision and recall for various algorithms

A. *Precision-Recall Curve comparison:*

The below graphical representation PRC comparison of various algorithms. By comparing algorithms, a Voting classifier has good accuracy but when we draw PRC, it shows that Logistic regression has good Precision-Recall value at threshold 0.5. So, while changing threshold values, it improves the Precision and Recall values.

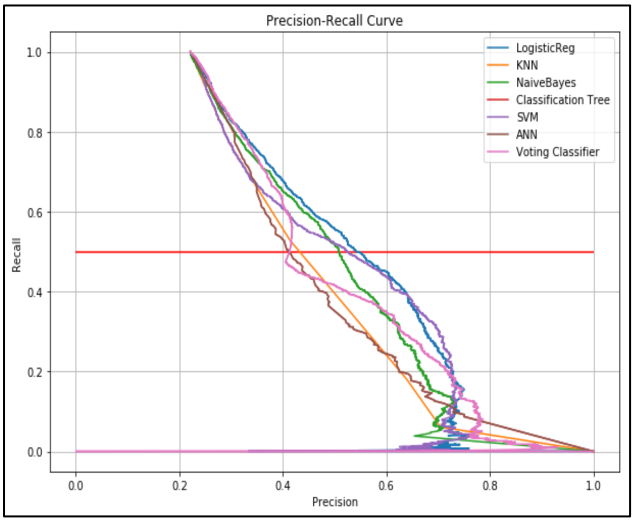


Fig (1.6) PRC comparison

Logistic Regression Classifier to check threshold value: To check threshold value and Precision, recall values at different threshold, we draw Logistic Regression classifier diagram. Here, we shown good precision and recall value at threshold 0.2. So, updated a model with threshold value 0.2 and the improvement was approx. 44% in precision and recall value of a model. As, it decreases False Negative value which means defaulters are predicted as non-defaulter. False Negative value is changed approximately 1800 to 1000 and the confusion matrix was:

$$\begin{bmatrix}7487 & 1859\\1074 & 1580\end{bmatrix}.$$

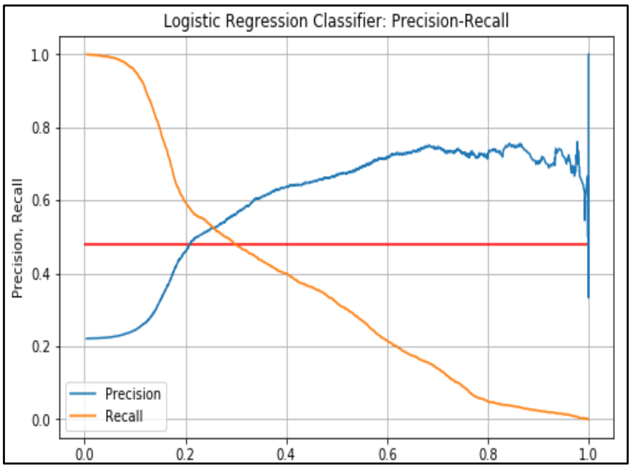


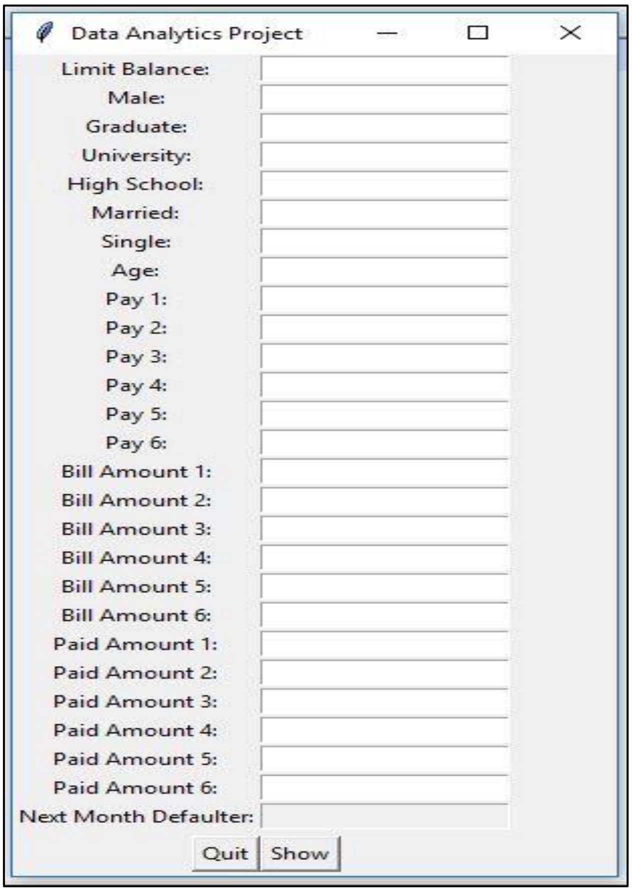
Fig (1.7) LR classifier

B. *Graphical User Interface:*

We created Graphical User Interface using python and tkinter, we trained a model and set threshold value at 0.2 in logistic regression. When user will submit below mentioned parameters value, model will predict whether a user will be defaulter or non-defaulter next month in payment.

The general steps are mentioned below:

1. Choose the best model and parameters
2. Save to .json file
3. Load a file from disk to predict data
4. Call a function on button submit and load data to a model
5. Check probability and result on GUI


Data Analytics Project

Limit Balance:
Male:
Graduate:
University:
High School:
Married:
Single:
Age:
Pay 1:
Pay 2:
Pay 3:
Pay 4:
Pay 5:
Pay 6:
Bill Amount 1:
Bill Amount 2:
Bill Amount 3:
Bill Amount 4:
Bill Amount 5:
Bill Amount 6:
Paid Amount 1:
Paid Amount 2:
Paid Amount 3:
Paid Amount 4:
Paid Amount 5:
Paid Amount 6:
Next Month Defaulter:

Fig (1.8) GUI for checking defaulters

C. *Libraries used for implementation:*

To implement the code, we have made use of Python 3.6 version. We had made use of numpy library for multidimensional array used to store of same datatype. Pandas library provide high performance and used for data analysis tools. Tensorflow library implemented in order to implement Artificial Neural Network. Sklearn library is used which has various features of classification, regression including SVM, KNN, gradient boosting etc. We make use of matplotlib.pyplot for comprehensive 2D/3D plotting and displaying in understandable manner. Keras is a high-level API to build and train deep learning model. It is user friendly and composable. Seaborn is a visualization library based on matplotlib. Graphviz is not a python package; it simply put the graphviz files into our virtual directory. Tkinter is used to create a GUI.

V. Summary

This would inform the issuer's decisions on who to give a credit card to and what credit limit to provide. We investigated the data, checking for data unbalancing, visualizing the features and understanding the relationship between different features. We used both train-validation split and cross-validation to evaluate the model effectiveness to predict the target value, i.e. detecting if a credit card client will default next month. We then investigated five predictive models: We started with Logistic Regression, Naïve bayes, SVM, KNN, Classification Tree and Feed-forward NN and Voting classifier accuracy is almost same. We choose based model Logistic regression based on minimum value of False Negative from confusion metrix.

References

[1]. Li, Xiao-Lin, and Yu Zhong. An overview of personal credit scoring: techniques and future work. Journal: International Journal of Intelligence Science ISSN 2163-0283. 2012.

[2]. Yeh, I-C. and C-H. Lien, 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systerm with Applications,36: 2473-2480.

[3]. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[4]. Taiwo Oladipupo Ayodele. (2010) "Types of Machine Learning Algorithms", New Advances In Machine Learning, Yagang Zhang (Ed.), Intech

[5]. NH Niloy, MAI Navid. Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients. American Journal of Data Mining and Knowledge Discovery. Vol. 3, No. 1, 2018, pp. 1-12. doi: 10.11648/j.ajdmkd.20180301.11

[6]. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36, 3302–3308.

[7]. Davis J., Goadrich M. The Relationship Between Precision-Recall and ROC Curves.ACM New York, NY, USA 2006. ISBN:1-59593-383-2.

[8]. Christopher M. Fraser(2000), "Neural Networks: A Review from a Statistical Perspective", Hayward Statistics.

[9]. Shie Mannor, Dori Peleg and Reuven Rubinstein. ICML '05 Proceedings of the 22nd international conference on Machine learning. ACM New York, NY, USA 2005. ISBN: 1-59593-180-5

[10]. Arlot, Sylvain, and Alain Celisse. A survey of cross-validation procedures for model selection. eprint arXiv:0907.4728. DOI:10.1214/09-SS054.