**PROJECT REPORT**

**ON**

**Predicting credit card**

**Defaulters**

TEAM MEMBERS:

Pratyusha Reddy Velama (pxv150130) Manwitha Yarradoddi (mxy150130)

Manaswi Kothi (mxk154430)

**INTRODUCTION:**

* This project aimed at the case of predicting default payments in Taiwan and compares the predictive accuracy of the probability of default among different data mining methods.
* When a person applies for and receive a credit card, they are taking on a huge responsibility.
* The credit card company evaluates their credit worthiness and gives them a line of credit that they feel they can be responsible for.
* While most people will use their card to make purchases and then diligently make payments on what they charge, there are some people who, for one reason or another, do not keep up on their payments and eventually go into credit card default.
* Credit card default is the term used to describe what happens when a credit card user makes purchases by charging them to their credit card and then they do not pay their bill.
* The probability that a customer will default on their credit card debt is important to know for banks, with an accurate predictive model, banks can avoid issuing credit cards to customers who will later default, and also banks can avoid turning down customers who are actually likely to pay on time
* The basic approach used is if there is a payment due for corresponding bill statement and if the payment is not done in correct time then corresponding delay flag is set.
* So if there are delay flags set and the payment is not done by the end then the ID is classified as Defaulter else ID is not classified as defaulter.
* Here, we classify the data using different classifiers and compare their accuracy, precision, recall, f-score values, to find the best classifier for the data.
* We are also using 10-fold cross validation to avoid overfitting of data.

**2. PROBLEM DEFINITION AND ALGORITHM:**

**2.1 Task Definition:**

* It focuses on the relationship between default and the outcomes of financial choices consumers make within the constraints of the contract terms set by credit card issuers.
* Our data set contains the most detailed information yet available on a regular basis on the behavioral aspects of credit card use.

There are about 23 features which are

X1 (LIMIT\_BAL): Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

X2(SEX):Gender(1=male;2=female).

X3 (EDUCATION): Education (1 = graduate school; 2 = university; 3 = high school;4=others).

X4 (MARRIAGE): Marital status (1 = married; 2 = single; 3 = others).

X5(AGE):Age(year).

X6 - X11 (PAY\_1 - PAY\_6): History of past payment i.e. the repayment status in September, August,….., April, 2005 respectively. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17 (BILL\_AMT1 - BILL\_AMT6): Amount of bill statement (NT dollar) i.e. amount of bill statement in September, August,………., April, 2005 respectively.

X18-X23 (PAY\_AMT1 - PAY\_AMT6): Amount of previous payment (NT dollar) i.e. amount paid in September, August, ……, April, 2005 respectively.

* The output class variable (default payment next month) is binary i.e.
* 1 - Defaulter
* 0 - Not a Defaulter
* The reason why this is more important is that Credit risk analysis (finance risk analysis, loan default risk analysis) and credit risk management is important to financial institutions which provide loans to businesses and individuals.
* Credit can occur for various reasons: bank mortgages (or home loans), motor vehicle purchase finances, credit card purchases, installment purchases, and so on.
* Credit loans and finances have risk of being defaulted or delinquented.
* To understand risk levels of credit users, credit providers normally collect vast amount of information on borrowers.
* Banks often rely on credit prediction models to determine whether to approve a loan request.
* To a bank, a good prediction model is necessary so that the bank can provide as much credit as possible without exceeding a risk threshold.

**2.2 Algorithm Definition**

**Pseudocode:**

A = List of algorithms

Read “default of credit card clients” dataset

Create 10 folds // for 10-fold cross validation

for a in Algorithms A

for i in 1:10 // for all folds

Create training data // all the data except data in ith fold

Create testing data // data in ith fold

// The createModel method returns the predictive accuracy, Precision, Recall, //F-Score using algorithm ‘a'

acc <- createModel(a, train, test)

totalAcc <- totalAcc + acc[1]

totalPreci <- totalPreci +acc[2]

totalRecall <- totalRecall + acc[3]

totalFscore <- totalFscore +acc[4]

next i

avgAcc <- totalAcc/10 //divide by number of folds

avgPreci <- totalPreci /10

avgRecall <- totalRecall/10

avgFscore <- totalFscore/10

next a

**3. EXPERIMENTAL EVALUATION:**

**3.1 Methodology:**

In this project, classification is done using Logistic regression, Decision tree, Naïve Bayes, SVM, Artificial Neural Network, kNN, Random Forest, Boosting and Bagging.

**Artificial Neural network:** Artificial neural networks (ANNs) are non-linear statistical modeling based on the function of the human brain. They are powerful tools for unknown data relationship modeling. ANNs able to recognize the complex pattern between input and output variables then predict the outcome of new independent input data.

**Bayesian classifier:** A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions and is particularly suited when the dimensionality of the inputs is high. A naive Bayes classifier assumes that the existence (or nonexistence) of a specific feature of a class is unrelated to the existence (or nonexistence) of any other feature. The major disadvantage of this model is that the predictive accuracy is highly correlated with this assumption. An advantage of this method is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

**Logistic regression:** Logistic regression is a form of linear regression. This model can predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous. The relationship between the predictor and response variables is not a linear function in logistic regression. The advantages of this method are that the logistic regression does not assume linearity of relationship between the independent variables and the dependent, does not require normally distributed variables and the weakness of the model is that the independent variables be linearly related to the logit of the dependent variable.

**K-nearest neighbor:** K-nearest neighbor is a nonparametric classifier based on learning by similarity. A training data set is collected, for this training data set, a distance function is introduced between the explanatory variable of observations. For each new observation this method explores the pattern space for the K nearest neighbors that are closest to the new observation in term of distance between the explanatory variables. The new observation is assigned to the class which its most KNN belong to that class.

**Decision tree:** A classification tree is a tree-like graph of decisions and their possible consequences. Topmost node in this tree is the root node which a decision is supposed to take on it. In each inner node, it is done a test on an attribute or input variable. Each branch which follows the node lead to the result of the test, and the classes are represented by leaf nodes. Classification trees are used when the response variable is quantitative discrete or qualitative. CT is based on maximizing purity measure of the response variables of the observations. The advantage of this method is that it is a white box model and so it is simple to understand and explanation, but the limitation of this model is that, it cannot be generalized a designed structure for one context to the other contexts.

**Support vector machine:** Support vector machine is a classifier technique. This method involves these elements, an objective function which considers both training and test samples to optimize the classification of new data, an optimizing algorithm for determining the optimal parameters of training sample objective function. The advantages of the method are that, in the nonparametric case, SVM requires no data structure assumptions such as normal distribution and continuity. SVM can perform a nonlinear mapping from an original input space into a high dimensional feature space and this method is capable of handling both continuous and categorical predictions. The weaknesses of this method are that, it is difficult to interpret unless the features interpretable and standard formulations do not contain specification of business constraints.

**Boosting:** Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. Informally, [the hypothesis boosting] problem asks whether an efficient learning algorithm that outputs a hypothesis whose performance is only slightly better than random guessing [i.e. a weak learner] implies the existence of an efficient algorithm that outputs a hypothesis of arbitrary accuracy [i.e. a strong learner]. Algorithms that achieve hypothesis boosting quickly became simply known as "boosting".

**Bagging:** :Bootstrap aggregating, also called as bagging, is an ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting, it is a special case of the model averaging approach.

**Random Forests:** Random forests is a popular ensemble method invented by Breiman and Cutler. It was chosen for this project because of the many advantages it offers. Random forests can run efficiently on large databases, and by its ensemble nature, does not require much supervised feature selection to work well. Most importantly, it supports not just classification, but regression outputs as well. However, there are still performance parameters that need to be tuned to improve the performance of the random forest.

**10-fold cross validation:** In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).[[4]](https://en.wikipedia.org/wiki/Cross-validation_(statistics)#cite_note-Newbie_question:_Confused_about_train.2C_validation_and_test_data.21-4) The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the [validation dataset](https://en.wikipedia.org/wiki/Validation_set)), in order to limit problems like [overfitting](https://en.wikipedia.org/wiki/Overfitting), give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc.

The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times.

**Accuracy** is the number of correct predictions made divided by the total number of predictions but it can be misleading sometimes. Though a model can predict the value of the majority class for all predictions and achieve a high classification accuracy, this model is not useful in the problem domain. So, sometimes it may be desirable to select a model with a lower accuracy because it has a greater predictive power on the problem.

Precision, recall, F-score are additional measures required to evaluate a classifier.

**Precision** is a measure of a classifiers exactness. A low precision can indicate a large number of incorrect predictions (False Positives).

**Recall** is a measure of a classifiers completeness. A low recall indicates many incorrect predictions (False Negatives).

**F1 score** is the balance between the precision and the recall.

**3.2 Results**

No of instances: 30000

No of attributes: 24

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifiers | How many fold cross-validation | Accuracy (%) | Precision | Recall | F-Score |
| Decision Tree | 10 | 100 | 1 | 1 | 1 |
| Naïve Bayes | 10 | 98.93 | 0.99 | 0.98 | 0.99 |
| SVM | 10 | 100 | 1 | 1 | 1 |
| Perceptron | 10 | 79.1 | 0.54 | 0.77 | 0.63 |
| Neural Networks | 10 | 77.88 | 1.91 | 0.78 | 1.05 |
| Random Forests | 10 | 100 | 1 | 1 | 1 |
| K-NN | 10 | 76.38 | 0.6 | 0.54 | 0.57 |
| Ada boosting | 10 | 81.93 | 0.66 | 0.76 | 0.70 |
| Bagging | 10 | 81.59 | 0.66 | 0.74 | 0.7 |
| Gradient Boosting | 10 | 81.96 | 0.64 | 0.77 | 0.699 |

**3.3 Discussion**

* From the results obtained, Random forest, decision tree and SVM gave 100% accuracy.
* **Decision trees** are prone to overfitting. For the present data decision trees may give good accuracy but, when future data is added overfitting may occur. **Random forest** builds not a single biased tree but many biased trees with random subset of data and the decision is the majority of all the trees.
* **SVM** also gave good accuracy, as this is a two class problem and the data is reasonably clean and outlier free. But for data which contains outliers **Random forest** gives best results, because Random forest is more interpretable, feature importance can be estimated during training for little additional computation, plotting of sample proximities, visualization of output decision trees, it readily handles larger numbers of predictors, faster to train and has fewer parameters.
* So, we would like to conclude that the **Random forests** classifier is the best among all the classifiers we used.

**4. RELATED WORK**

A research was conducted which aimed at the case of customers default payments in Taiwan and compared the predictive accuracy among six data mining methods and Artificial Neural Networks gave the highest accuracy.

In their research ensemble methods were not used, but in this project we have compared simple classifiers and complex classifiers including ensemble methods. Our project is important because, ensemble methods can classify the data better than other classifiers.  
  
**5. FUTURE WORK**

* From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.
* Our work focuses only on binary classification whereas, in future the probability of default can be estimated.

**6. CONCLUSION**

Among all the classifiers, Random forest, SVM and Decision tree give the best results. As they handle high dimensional spaces as well as large number of training examples very well, we conclude that Random forest, SVM and Decision tree can be used to classify a customer as a defaulter or not a defaulter.

BIBLIOGRAPHY:

* Xiao, W., Zhao, Q., and Fei, Q., 2006, "a comparative study of data mining methods in

Consumer loans credit scoring management." journal of systems science and systems engineering, 15(4), 419-435.

* Yeh, I. C., and 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
* Kirkos, E., Spathis, C., and Manolopoulos., Y., 2007, "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4), 995-1003.
* 2005, Cutler, Random Forests, Encyclopedia of Statistics in Behavioral Science, pp 1665– 1667.
* 2002, Liaw A, Wiener M, Classification and Regression by Random Forest, R News, Vol 2/3, Dec
* <http://cs229.stanford.edu/proj2011/JunjieLiang-PredictingBorrowersChanceOfDefaultingOnCreditLoans.pdf>
* <http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>
* <http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf>