COMPARATIVE ANALYSIS OF CONSUMER CLAIM DATASET
BY LOGISTIC REGRESSION AND LINEAR REGRESSION
ALGORITHMS

Dhyan Chandra Yadav
S.N.P.G College, Narahi, Balia (U.P)
Department of Computer application
Email Id-dc9532105114@gmail.com

ABSTRACT:

Every bank has lot of records of employ and account holders but there some general problems arises as like dispute of consumers. In this paper we use Logistic regression and Linear regression by which to describe consumer claim data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, ratio-level independent variables.

Keywords: Classifiers Function: Logistic Regression; Decision Tree, Weka Tools.

[1] INTRODUCTION

Bureau is financial protection government agency in America. Some consumer suffer in financial problem as a differ types of loan claims in the bank. Credit reporting companies must investigate your dispute, forward all documents to the furnisher, and report the results back to you unless they determine your claim is frivolous. If the consumer reporting company or furnisher determines that your dispute is frivolous, it can choose not to investigate the dispute so long as it sends you a notice within five days saying that it has made such a determination. If the furnisher corrects your information after your dispute, it must notify all of the credit reporting companies it sent the inaccurate information to, so they can update their reports with the correct information [1].
A decision tree is a flowchart like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values of competing alternatives are calculated [2].

D A Freedman introduced about logistic regression. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In statistics logistic regression, or logit regression, or logit model is a regression model where the dependent variable is categorical. This article covers the case of a binary dependent variable that is, where it can take only two values, "0" and "1", which represent outcomes such as pass/fail [3].

For example we visualize the Logistic Function consumer claim data set in Fig.1.

![Figure 1: Representation of Logistic Function by Weka.](image)

[2] RELATED WORK:

H, K analyzed that Credit cards fraudulence arises at very high level scale so we cannot easily detect and predict the related attributes but by the help of data mining classifier tool to prevent the activity of fraudsters in the misuse of credit cards uses the algorithms of neural networks. This system predicts the probability of fraud on an account by comparing the current transactions and the previous activities of each holder [4].

D C Yadav and S Pal discussed that classifier algorithms provide very accurate result in software error detection by J48, ID3 and Naïve Bayes data mining algorithms correctly classified instances will be partition in to numeric and percentage value, kappa statics, mean absolute error and root mean square error will be at numeric value only ID3 and J48 time taken to build model: 0.2 seconds and test mode: 10 fold cross validation. Here Weka compare all required parameters on given instances with the classifiers respective accuracy
and prediction rate based on highest accuracy of J48 is 100% without error also Naïve Bayes 100% correctly classified but with some error and ID3 95% correctly classified, so it is clear that J48 is the best in three respective algorithms so it is more accurate [5].

D C Yadav and R Kumar discussed that association algorithms provide very accurate result in the frequent and relationship between data object and find the percentage of confidence, support, of data object by the help of apriori, predictive apriori and filtered associate algorithms. Therefore these algorithms can be used in other domains to bring out interestingness among data present in the origin [6].

D C Yadav and R Kumar discussed that three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms. Author compared using a clustering tool and find result: K-Means algorithm is better than Hierarchical Clustering and Make density based algorithm because all the algorithms have some ambiguity in some (noisy) data when clustered [7].

R Sukanya and K Prabha discussed that back propagation Neural Network, Support Vector Machine is used for rainfall prediction. ANN improves the efficiency of Rainfall prediction by analyzing the historical and current facts to make accurate predictions about future [8].

R S, S M, N E, S P and V Kirand discussed that the huge volume of warranty data for segregating the fraudulent warranty claims using pattern recognition and clustering .Survey of automotive industry shows up to 10% of warranty costs are related to warranty claims fraud, costing manufacturers several billions of dollars. The existing methods to detect warranty fraud are very complex and expensive as they are dealing with inaccurate and vague data, causing manufacturers to bear the excessive costs [9].

D C Yadav analyzed that in statical analysis of binary classification, the F1 score is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. In this analysis author computed the best score for F1 by the help of data mining classifier algorithms and choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest F1 score and take less time to build a mode [10].

D C Yadav analyzed that the Matthews correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifiers. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. Author computed the MCC is in essence a correlation coefficient between the observed and predicted binary classifications by the help of data mining classifier algorithms and ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest MCC value and minimum number of time in second 0.00 to build a model [11].

D C Yadav analyzed that the informedness of a prediction method as captured by a contingency matrix is defined as the probability that the prediction method will make a correct decision as opposed to guessing and is calculated using the bookmaker algorithm. Their correlation is the generated by LAD Tree, ID3, and J48 data mining algorithms and find ID3 is the best data mining classifier algorithms to be applied over selected datasets [12].
D C Yadav analyzed that the FDR-controlling procedures provide less stringent control of Type I errors compared to class wise errors. In this analysis we choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has minimum time to build a model [13].

D C Yadav analyzed that all analysis on the basis of dependable variables for overall performance and Predicts categorical class level classifiers based on training set and the values in the class level attribute use the model in classifying new data. Author analyzed between AD Tree, LAD Tree, J48 and Naïve Bayes for correctly classify and incorrectly classify with kappa static model and choose the LAD Tree is the best data mining classifier algorithms to be applied over selected datasets. Because LAD Tree has highest correctly value 83.333% and minimum number of unclassified instances is 0.00. Also Lad tree have highest value 0.3576 of metric for accuracy [14].

T, R and Liu discussed that a framework was presented on the base of security systems and Case based reasoning for fraud detection. First, a set of normal and fraud cases are made from labeled data. Then, the primary detectors are made with random or genetic algorithms. Then, negative selection and clonal selection operations are applied on primary detectors in order to obtain a set of detectors with different algorithms that can detect a variety of frauds [15].

M,S,B and Saira discussed that many fraud detection systems that have been presented so far, have used data mining and neural network approaches. While no fraud detection system with the combination of anomaly detection, misuse detection and decision making system have been used so far for fraud detection in credit cards. Then, a system was proposed that used Hidden Markov Model to detect the fraudulent transactions [16].

A John analyzed that hybrid feature selection and anomaly detection algorithm in order to detect fraud in credit cards. The authors have noted that fraud detection on the internet must be done online and immediately. Since the use of credit card by card holders follows a fixed pattern, this fixed pattern can be extracted from a usual legal activity of card holders in 1 or 2 years. Thus, this pattern is compared to the use of process of card holder and in case of non-similarity in the pattern, the activity is considered illegal. It should be noted that the neural networks were used to teach the patterns detection in the model in this study [17].

A P, M K and A N discuss that data mining as one of the most efficient tools of data analysis has attracted the attention of many people. The use of different techniques and algorithms of this tool in various fields like customer relationship management, fraud management and detection, medical sciences, sport and etc. Due to the large number of data in banks, data mining has had lots of functions in financial and monetary affairs so far. Credit risk management, fraud detection, money laundering, customer relationship management and banking services quality management are some examples of data mining function in banks [18].

In this paper we use Logistic regression by which to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
[3] METHODOLOGY:

Our research approach is to use logistic regression and decision tree on consumer claim data set. The research methodology is divided into 5 steps to achieve the desired results:

Step 1: In this step, prepare the data and specify the source of data.

Step 2: In this step select the specific data and transform it into different format by weka.

Step 3: In this step, implement data mining algorithms and checking of all the relevant dispute is perform.

Step 4: The decision is taken on the presence of dispute in source code. If dispute is present then proceed further, otherwise it will stop. We classify the relevant dispute using logistic regression and decision tree.

Step 5: At the end, the results are display and evaluated.

Now we have study about consumer complaints of bank different type’s loans and relate dispute.

A. Data Preparation:

Bureau is financial protection government agency in America. Some consumer suffer in financial problem as a differ types of claims in the bank. Your card has a high interest rate. Just do the math and you will see why it’s a problem if your credit card has a high interest rate. The simple fact is that this card will cost you more than one with a lower interest rate. There’s just no reason to have a card with a high interest rate if you have a really good credit score. Credit reporting companies must investigate your dispute, forward all documents to the furnisher, and report the results back to you unless they determine your claim is frivolous. The Consumer Complaint Database shows the consumer’s original product, sub-product, issue, and sub-issue selections consistent with the options available on the form at the time the consumer submitted the complaint. The Consumer Complaint Database is a collection of complaints on a range of consumer financial products and services, sent to companies for response. We don’t verify all the facts alleged in these complaints, but we take steps to confirm a commercial relationship between the consumer and the company.

Table 1. Representation of Computational Variables of Consumer Claim
B. Data Selection and transform

The database generally updates daily, and contains certain information for each complaint, including the source of the complaint, the date of submission, and the company the complaint was sent to for response. The database also includes information about the actions taken by the company in response to the complaint, such as, whether the company’s response was timely and how the company responded. If the consumer opts to share it and after we take steps to remove personal information, we publish the consumer’s description of what happened. Companies also have the option to select a public response. Data from those complaints helps us understand the financial marketplace and protect consumers.

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaint Type</td>
<td>Consumer Complaint Database Attributes for Financial Problem (Bank, Lender &amp; Company etc.)</td>
</tr>
<tr>
<td>Sample Size</td>
<td>199Total: 44 Consumer dispute and 155 non dispute</td>
</tr>
<tr>
<td>Dependent Variables</td>
<td></td>
</tr>
<tr>
<td>Dispute(YES)</td>
<td>Consumer Complaint dispute</td>
</tr>
<tr>
<td>Dispute(NO)</td>
<td>Consumer Complaint not dispute</td>
</tr>
<tr>
<td>Field name</td>
<td>Description</td>
</tr>
<tr>
<td>Date received</td>
<td>The date the CFPB received the complaint.</td>
</tr>
<tr>
<td>Tags</td>
<td>Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.</td>
</tr>
<tr>
<td>Date sent to company</td>
<td>The date the CFPB sent the complaint to the company.</td>
</tr>
<tr>
<td>Company response to consumer</td>
<td>This is how the company responded. For example, “Closed with explanation.”</td>
</tr>
<tr>
<td>Timely response?</td>
<td>Whether the company gave a timely response. For example, “Yes” or “No.”</td>
</tr>
</tbody>
</table>


C. Implementation-

Logistic regression is based on calculation the odds of the outcome as the ratio of probability handling the outcome divided by probability of not having it. The implemented classifier is proficient in reading and analyzing a number of populations in giving datasets.
The classifier is developed using Java programming language. The WEKA library provides built-in classes and functions to implement and utilize different mathematical and statistical algorithms for data classification. Based on the number of identified population, it estimates following results: kinds of species in a population (if there are more than 1), correctly classified instances, incorrectly classified instances, momentum and accuracy (optimized, weighted results). The classifier is capable of processing standard CSV dataset files, which describes the list of instances sharing a set of attributes, especially used to develop for machine learning projects. The classifier’s workflow starts with the analysis of inputted data and extraction of attributes, classes, instances and relationships. The example data set consumer claims classified by logistic regression database being processed using WEKA Explorer.

D. Result and Discussion-

Figures [2], [3] presents the distribution of consumer claims instances by linear regression and logistic. In the last steps, after cross validation, the individual and commutative weights of instances are calculated. The obtained results are validated and final output is presented to the user in the end. The measurement and prediction procedure can be repeated until the satisfactory results are achieved. Figure [4][5] Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal. Reducing an ordinal or even metric variable to dichotomous level loses a lot of information, which makes this test inferior compared to ordinal logistic regression in these cases. Logistic regression assumes linearity of independent variables and log odds.

```plaintext
Classifier model (full training set)

SimpleLogistic:

Class 0 :
0 +

Class 1 :
0 +

Class 2 :
0 +

Time taken to build model: 4.32 seconds
```

Fig.4 Representations of Linear Regression
From the Figure [4][5] it is clear that it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise the test underestimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Classification Accuracy</th>
<th>ROC</th>
<th>RMSE</th>
<th>RRSE</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression(Multi Nominal Logistic)</td>
<td>55.28</td>
<td>0.55</td>
<td>0.538</td>
<td>157.158</td>
<td>YES</td>
</tr>
<tr>
<td>Simple Logistic Regression(Linear Logistic)</td>
<td>76.88</td>
<td>0.51</td>
<td>0.409</td>
<td>119.14</td>
<td>YES</td>
</tr>
</tbody>
</table>

➢ In the above table give highest number of percentage for correctly classified instances value is 76.88%.
➢ Total minimum number of RMSE value is 0.409 for all classifier algorithms.
➢ The minimum number of RRSE value is 119.14.
➢ The ROC value are nearly equivalence 0.55 and 0.51.

CONCLUSION:
Linear Regression and Logistic Regression calculated a matrix which has classes and attributes with total instances. So in conclusion,

- These algorithms are really fast but Linear Regression give more accurate value for instances classification.
- In the experiment Linear Regression always give less error compare to Logistic Regression.
Logistic regression is intrinsically simple; it has low variance and so is less prone to over-fitting. Decision trees can be scaled up to be very complex, are more liable to over-fit. Pruning is applied to avoid this.

In the future the Meta Classifier algorithms can be experimented on the data set to obtain more efficient results. Also the Meta classifier can be calculated by using parameters such as the cross validation, percentage split and supplied test set.

REFERENCES


