Detecting Financial Fraud Using Machine Learning: Winning the War Against Imbalanced Data

D.Rajanipriya[1]

[1]Department of Computer Science, Texas Tech University, Texas, United States

***ABSTRACT*-Credit card fraud events take place frequently and then result in huge financial losses. Criminals can use some technologies such as Trojan or Phishing to steal the information of other people's credit cards. Therefore, an effective fraud detection method is important since it can identify a fraud in time when a criminal uses a stolen card to consume. One method is to make full use of the historical transaction data including normal transactions and fraud ones to obtain normal/fraud behavior features based on machine learning techniques, and then utilize these features to check if a transaction is fraud or not. The data used in our experiments come from an e-commerce company in Europe. In this project, we divide the data-set into train and test set by making use of the different machine learning algorithms like smote, under sampling, over sampling, both under and over sampling, rose. We have compared the accuracy and performance of these models using confusion matrix and roc curve**

Keywords:* *Detecting Financial Fraud(DFU),Smote(Synthetic Minority Over Sampling Technique)

I. Introduction

Credit cards are widely used due to the popularization of e-commerce and the development of mobile intelligent devices .Card-not-present transactions (i.e., online transaction without a physical card). It is more popular, especially all credit card operations are performed by web payment gateways, e.g., PayPal and Alipay.Credit card has made an online transaction easier and more convenient. However, there is a growing trend of transaction frauds resulting in a great losses of money every year. It is estimated that losses are increased yearly at double digit rates by 2030. Since the physical card is not needed in the online transaction environment and the card's information is enough to complete a payment, it is easier to conduct a fraud than before. Transaction fraud has become a top barrier to the development of e-commerce and has a dramatic influence on the economy. Hence, fraud detection is essential and necessary. Due to these frauds many of the common people and sometimes countries revenue may lose. Fraud detection is a process of monitoring the transaction behavior of a cardholder in order to detect whether an in-coming transaction is done by the cardholder or others. Generally, there are two kinds of methods for fraud detection. Misuse detection and anomaly detection. Misuse detection uses classification methods to determine whether an incoming transaction is fraud or not. Usually, such an approach has to know about the existing types of fraud to make models by learning the various fraud patterns. Anomaly detection is to build the profile of normal transaction behavior of a cardholder based on his/her historical transaction data, and decide a newly transaction as a potential fraud if it deviates from the normal transaction behavior. However, an anomaly detection method needs enough successive sample data to characterize thenormal transaction behavior of a cardholder .This paper is about Anomaly method.

II. CREDIT CARD FRAUD PROBLEM

Billions of dollars of loss are caused every year by the fraudulent credit card transactions. Fraud is old as humanity itself and can take an unlimited variety of different forms. The Pwc global economic crime survey of 2020 suggests that approximately 48% of organizations experienced economic crime. Therefore, there is definitely a need to solve the problem of credit card fraud detection. Moreover, the development of new technologies provides additional ways in which criminals may commit fraud. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years. Hugh Financial losses has been fraudulent affects not only merchants and banks, but also individual person who are using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that, though difficult to quantify in the short term, may become visible in the long period. For example, if a cardholder is victim of fraud with a certain company, he may no longer trust their business and choose a competitor.

III. System Architecture

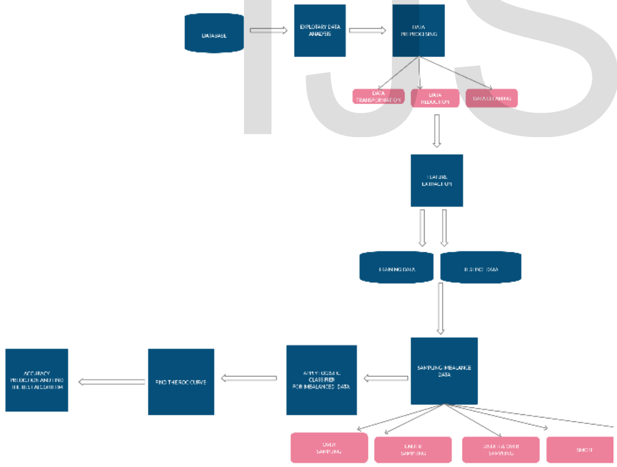
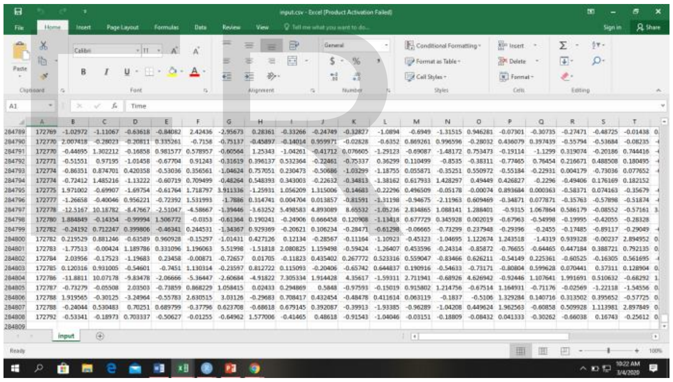


Fig 1 Block Diagram

Framework design is the decided model that depicts the structure, direct, and more perspectives on a framework. A planning portrayal is a standard outline and delineation of a framework, filtered through with the ultimate objective that supports thinking about the structures and practices of the framework. A structure arrangement can contain framework parts and the sub-frameworks created that will team up to finish the general framework. There have been attempts to formalize vernaculars to portray structure working; as a rule these are called plan depiction tongues.

Iv. Dataset

The datasets contains transactions made by credit cards in September 2013 by european cardholders.
This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
284789	172769	-1.02972	-1.11067	-0.63618	-0.84082	2.42436	-2.95673	0.28361	-0.33266	-0.24749	-0.32827	-1.0894	-0.6949	-1.31515	0.946281	-0.07301	-0.30735	-0.27471	-0.48725	-0.01438
284790	172770	2.007418	-0.28023	-0.20811	0.335261	-0.7158	-0.75137	-0.45897	-0.14014	0.959971	-0.02828	-0.6352	0.869261	0.996596	-0.28032	0.436079	0.397439	-0.55794	-0.53684	-0.08235
284791	172770	-0.44695	1.302212	-0.16858	0.981577	0.578957	-0.60584	1.25343	-1.04261	-0.41712	0.076605	-1.29123	-0.69087	-1.48172	0.753473	-0.19114	-1.1299	0.319074	-0.20186	0.744416
284792	172771	-0.51551	0.97195	-1.01458	-0.67704	0.91243	-0.31619	0.396137	0.532364	-0.22461	-0.75337	0.36299	0.110499	-0.8535	-0.38311	-0.77465	0.76454	0.216671	0.488508	0.180495
284793	172774	-0.86351	0.874701	0.420358	-0.53036	0.356561	-1.04624	0.757051	0.230473	-0.50686	-1.03299	-1.18755	0.055871	-0.35251	0.550972	-0.55184	-0.22931	0.004179	-0.73036	0.077652
284794	172774	-0.72412	1.485216	-1.13222	-0.60719	0.709499	-0.48264	0.548393	0.343003	-0.22632	-0.34813	-1.38162	0.617933	1.428297	0.49449	0.426827	-0.2296	-0.49406	0.176169	0.182152
284795	172775	1.971002	-0.69907	-1.69754	-0.61764	1.718797	3.911336	-1.25931	1.056209	1.315006	-0.14683	-0.22296	0.496509	-0.05178	-0.00074	0.893684	0.000363	-0.58371	0.074163	-0.35679
284796	172777	-1.26658	-0.40046	0.956221	-0.72392	1.531993	-1.7886	0.314741	0.004704	0.013857	-0.81591	-1.31198	-0.94675	-2.11963	0.609469	-0.34871	0.077871	-0.35763	-0.57898	-0.51874
284797	172778	-12.5167	10.18782	-8.47667	-2.51047	-4.58667	-1.39446	-3.63252	5.498583	4.893089	8.65532	-1.05236	2.834865	1.088141	1.288401	-0.9315	1.067864	0.586179	-0.08552	-0.57161
284798	172780	1.884849	-0.14354	-0.99994	1.506772	-0.0353	-0.61384	0.190241	-0.24906	0.666458	0.120908	-1.13418	0.677729	0.345928	0.002019	-0.67963	-0.54998	-0.19995	-0.42055	-0.28328
284799	172782	-0.24192	0.712247	0.399806	-0.46341	0.244531	-1.34367	0.929369	-0.20621	0.106234	-0.28471	-0.61298	-0.06665	-0.73299	0.237948	-0.29396	-0.2455	-0.17485	-0.89117	-0.29049
284800	172782	0.219529	0.881246	-0.63589	0.960928	-0.15297	-1.01431	0.427126	0.12134	-0.28567	-0.11164	-1.10923	-0.45323	-1.04695	1.122674	1.243518	-1.4319	0.939328	-0.00237	2.894952
284801	172783	-1.77513	-0.00424	1.189786	0.331096	1.196063	5.51998	-1.51818	2.080825	1.159498	-0.59424	-1.26407	0.453596	-0.24314	-0.85872	-0.76655	-0.64465	0.447184	0.388721	0.792135
284802	172784	2.03956	-0.17523	-1.19683	0.23458	-0.00871	-0.72657	0.01705	-0.11823	0.435402	0.267772	0.523316	0.559047	-0.83466	0.626211	-0.54149	0.225361	-0.60525	-0.16305	0.561695
284803	172785	0.120316	0.931005	-0.54601	-0.7451	1.130314	-0.23597	0.812722	0.115093	-0.20406	-0.65742	0.644837	0.190916	-0.54633	-0.73171	-0.80804	0.599628	0.070441	0.37311	0.128904
284804	172786	-11.8811	10.07178	-9.83478	-2.06666	-5.36447	-2.60684	-4.91822	7.305334	1.914428	4.35617	-1.59311	2.711941	-0.68926	4.626942	-0.92446	1.107641	1.991691	0.510632	-0.68292
284805	172787	-0.73279	-0.05508	2.03503	-0.73859	0.868229	1.058415	0.02433	0.294869	0.5848	-0.97593	-0.15019	0.915802	1.214756	-0.67514	1.164931	-0.71176	-0.02569	-1.22118	-1.54556
284806	172788	1.919565	-0.30125	-3.24964	-0.55783	2.630515	3.03126	-0.29683	0.708417	0.432454	-0.48478	0.411614	0.063119	-0.1837	-0.5106	1.329284	0.140716	0.313502	0.395652	-0.57725
284807	172788	-0.24044	0.530483	0.70251	0.689799	-0.37796	0.623708	-0.68618	0.679145	0.392087	-0.39913	-1.93385	-0.96289	-1.04208	0.449624	1.962563	-0.60858	0.509928	1.113981	2.897849
284808	172792	-0.53341	-0.18973	0.703337	-0.50627	-0.01255	-0.64962	1.577006	-0.41465	0.48618	-0.91543	-1.04046	-0.03151	-0.18809	-0.08432	0.041333	-0.30262	-0.66038	0.16743	-0.25612
284809																				

Fig 2 Screenshot of Dataset for Credit Card

V. Implementation

We are using Rstudio to implement algorithm and with R language we are writing the code.**RStudio** is an integrated development environment (IDE) for R,programming language for statistical computing and graphics. It is available in two formats: R Studio.Desktop is a regular desktop application while R Studio Server runs on a remote server and allows accessing R Studio using a web browser.An IDE that was built just for R. Syntax highlighting, code completion, and smart indentation. Execute R code directly from the sourceeditor. It Quickly jump to function definitions, as well as tools for plotting, history, debugging and workspace management.Brings your workflow together with Integrated R help and documentation. It Easily manage multiple working directories using projects, Workspace browser and data viewer. It has Interactive debugger to diagnose and fix errors. It is famous for quick Extensive package development tools.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures writtenin the C, C++, .Net, Python or FORTRAN languages for efficiency. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. R is free software distributed under a GNU-style copy left, and an official part of the GNU project called **GNU S**.R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University ofAuckland in Auckland, New Zealand. R made its first appearance in 1993.A large group of individuals has contributed to R by sending code and bug reports. Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive. As stated earlier,R is a programming language and software environment for statistical analysis, graphicsrepresentation and reporting. The following are the important features of R is a well-developed,simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities. R has an effective data handling and storagefacility, R provides a suite of operators for calculations on arrays, lists, vectors and matrices. Rprovides a large, coherent and integrated collection of tools for data analysis. R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers. As a conclusion, R is world’s most widely used statistics programming language. It's the1st choice of data scientists and supported by a vibrant and talented community of contributors.You can download the Windows installer version of R from R-3.2.x for Windows (32/64 bit) and save it in a local directory. As it is a Windows installer (.exe) with a name "R-version-win.exe".You can just double click and run the installer accepting the default settings. If yourWindows is32-bit version, it installs the 32-bit version. But if your windows is 64-bit, then it installs both the 32-bit and 64-bit versions. After installation you can locate the icon to run the Program in a directory structure "R\R3.2.x\bin\i386\Rgui.exe" under the Windows Program Files.Clicking this icon brings up the R-GUI which is the R console to do R Programming.There are five algorithms which we are implementing they are under sampling,over sampling,both under and over sampling,rose,smote for sampling the imbalance data.

VI .RESULTS AND DISCUSSION

Data Extraction: The data sets contains transactions made by credit cards in September 2013 by European cardholders. This data-set presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The data-set is highly unbalanced, the positive class(frauds) account for 0.172% of all transactions. Dueto confidentiality reasons, the data was anonymized variable names wererenamed to V1, V2, V3 until V28. Moreover, most of it was scaled, except for the Amount and Class variables.

Explotary Data Analysis: In the exploratory data analysis we will do five tasks

1)Calculate the Class Count and Class Percentage

2) Visualize theData

3)Split the transactions into days

4) Find the time of transaction of a particular day

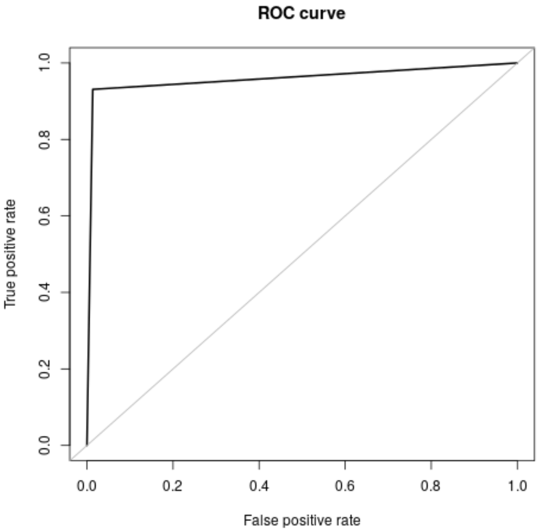
5)Find the time quarters for transaction

We will create a data frame which consists of a class group, count of records of each class, and calculates the percentage of each class. The input data is CSV format and the input filename "input.csv" and the output file is “output.csv”. After this we will visualize the data-set to check how many transactions are fraudulent and non-fraudulent plotting the graph(Using ggplot2). The input data is CSV format and the input filename "input.csv”, the output is an image file. split the transactions into day 1 and day 2 based on recorded the time. Add the column named day into the data-set and then find the time of transaction made for that particular day, based on the time and the day details the time quarter of each transaction which falls on 4 different time quarters like0 - 6 hr - 1st quarter ,7 - 12 hr - 2nd quarter ,13 - 18 hr - 3rd quarter ,19 - 24 hr - 4th quarter based on the time_day details and add the column named "Time_qtr".Plot the graph of time quarters for the fraudulent and non-fraudulent transactions.

Data Preprocessing:Split the dataset into train and test datasets, and scale all the fields except the Class field.The dataset should divide 70% of records as training data_set and rest records as test dataset(using catools library to split the data).Then find the logistic classifier on the train set andpredict the response on the test set(exclude the class field), and output the model's accuracy using the ROC curve, Use Generalized Linear Models(glm) function to predict with binomial. If the predicted value is less than 0.5 make it 0 else make it 1 with variable name "y_pred_normal"..Omit the na values.

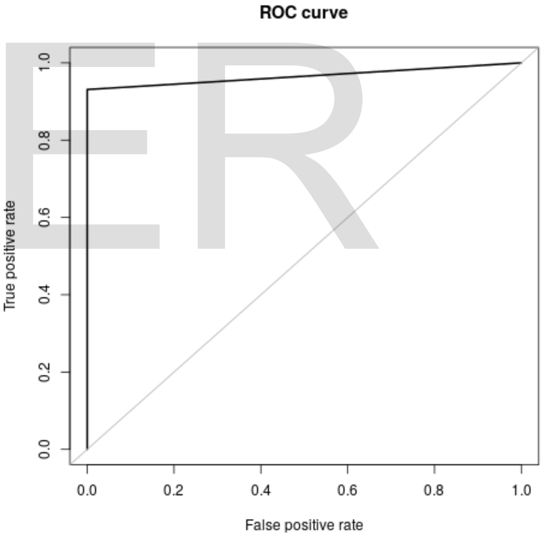
Sampling The Imbalance Data:Sample the imbalanced training data-set, using over method .”Oversample" means to artificially create observations in our data set belonging to the class that is under represented in our data.use the ovun.sample function for sampling on training dataset ,N = Double of the sample data,Method = over. Sample the imbalanced training data-set, using both method. use the ovun.sample function for sampling on the training dataset. seed = 222, Method = both ,N = Number of the sample data, Probability = 0.5. Sample the imbalanced training data-set, using the rose method. use the ovun.sample function for sampling on the training dataset, seed = 111.Sample the imbalanced training data-set, using smote method. se the smote function , and the size of Minority class – 100, the size of the Majority class – 200.

Implementing Algorithm : Find the auc value for oversampled data-set using logistic regression for imbalanced data.Re-Sample the data set with over method ,Set seed as 111 ,Omit the na's.If the predicted value is less than 0.5 make it 0 else make it 1 with variable name "y_pred_over". Find the auc value for under sampled data-set using logistic regression for imbalanced data. Re-Sample the data set with under method. Find the auc value for both sampled data-set using logistic regression for imbalanced data , re-Sample the data set with both method and find the auc value for rose ,smote and sampled data-set using logistic regression for imbalanced data.Analyzing the best algorithm by comparing all the auc curve..



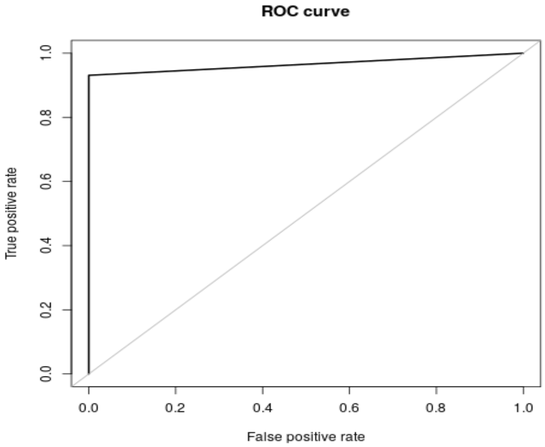
Area under the curve (AUC): 0.960

Fig 3 Over Sampling



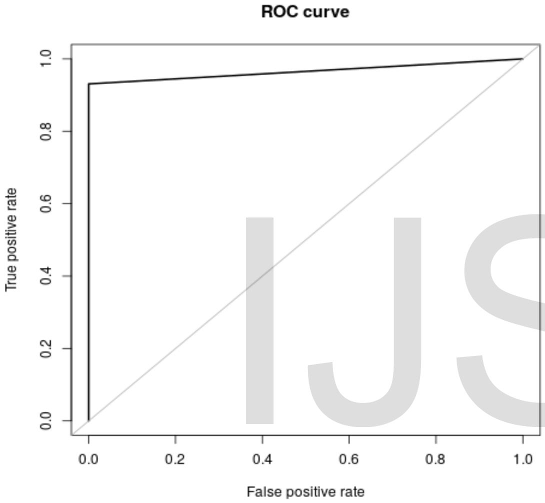
Area under the curve (AUC): 0.960

Fig 4 Under Sampling



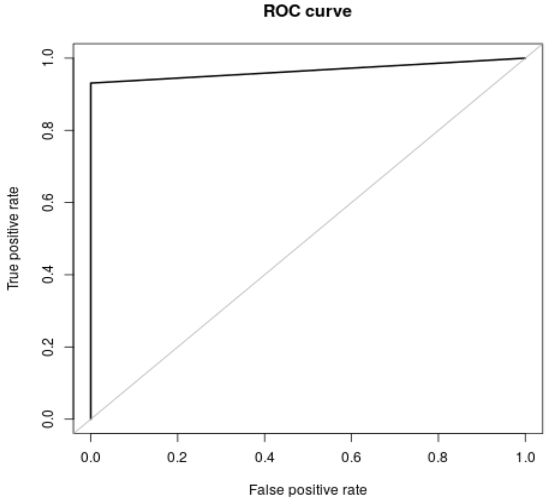
Area under the curve (AUC): 0.960

Fig 5 Both Under and Over Sampling



Area under the curve (AUC): 0.960

Fig 6 Rose



Area under the curve (AUC): 0.972

Fig 7 Smote

VII. Conclusion

Thus the Smote algorithm works far better than the other algorithms in predicting theCredit Card Financial Fraud.

VIII. Future Enhancement

The future upgrade of this work is tobe based on new multiple models with varying access pattern needs attention to improve the effectiveness. Privacy preserving techniques applied in distributed environment resolves the security related issues preventing private data access.

REFERENCES

[1]“BLAST-SSAHA Hybridization for Credit Card Fraud Detection” (AmlanKundu, SuvasiniPanigrahi, ShamikSural,) IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 6, NO. 4, OCTOBER-DECEMBER 2009.

[2] “ Research on Credit Card Fraud Detection Model Based on Distance Sum”(Wen-Fang-YU, Na Wang)

[3]“Detecting Credit Card Fraud by Decision Trees and Support Vector Machines” (Y. Sahinand E. Duman)

[4]” Fraudulent Detection in Credit Card System Using SVM &amp; Decision Tree” (VijayshreeB. Nipane, Poonam S. Kalinge, DipaliVidhate, Kunal War, Bhagyashree P. Deshpande)

[5]“Credit Card Fraud Detection Using Decision Tree Induction Algorithm” SnehalPatiletal, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April-2015

[6]“Data Mining Techniques for Credit Card Fraud Detection: Empirical Study “(MarwanFahmi, AbeerHamdy, Khaled Nagati).

[7] Gupta, Shalini, and R. Johari. ”A New Framework for Credit Card Transactions Involving Mutual Authentication between Cardholder and Merchant.” International Conference on Communication Systems and Network Technologies IEEE, 2011

[8] Y. Gmbh and K. G. Co, “Global online payment methods: Full year 2016,”