"Credit Card Fraud Detection *using* Logistic Regression *in* SPSS"

(A Machine Learning Approach)

Likitha { 21-375-015 }; Nissi Divija { 21-375-023 }; Bhanuprakash { 21-375-061 }

----------***----------

Abstract:

This work is focused on Credit Card Fraud Detection (CCFD),a serious problem in real world scenarios. As compared to the earliest times ,nowadays credit card frauds are drastically increased in numbers associated with the increasing in use of credit cards.

This effect was due to fraudsters are using fake identities to trap users in order to get money .So to eradicate or to control these frauds ,we came up with the idea of supervised machine learning model "Logistic Regression" ,which detects illegal and illicit transactions. With this we can trace the behavior and pattern of criminal activities by automating the process using respective algorithms.

From the conclusion of this entire work ,the performance of the model is based upon the Accuracy, Precision, Sensitivity and Specificity. So , our model logit classifier resulted in the best accuracy of ***97.20%*** respectively.

Keywords: Logistic regression, criminal transactions, illegal and illicit transactions.

1. Introduction : Credit Card Fraud

In this case study, we will explore how machine learning algorithms in SPSS were used to detect fraudulent transactions. The dataset used for this study is the Credit Card Fraud Detection dataset, which contains information on credit card transactions, including fraudulent and non-fraudulent transactions.

Credit card fraud is a major issue for financial institutions, businesses, and consumers that affects millions of people around the world each year. As the use of credit cards continues to grow, so does the prevalence of fraud.

In order to combat this issue, financial institutions and other organizations use various tools and techniques to detect and prevent fraudulent activities. With the increasing reliance on electronic transactions, credit card fraud is becoming more common and sophisticated. In such a way that statistical analysis techniques can help mitigate the risks associated with it.

- ✓ ***What are the key factors that contribute to credit card fraud?***

The key factors that contribute to credit card fraud can vary depending on the specific circumstances of the fraudulent activity.

However, some common factors that have been identified as Stolen or compromised card information, Skimming, Phishing, Hacking, Unauthorized use of a legitimate card, Counterfeit cards, High-risk transactions ,Poor security measures, etc..

Identifying these factors and developing effective measures to prevent them can help reduce the incidence of credit card fraud.

- ✓ ***What was the statistical hypothesis being tested in this analysis & level of significance?***

Null hypothesis (H_0): There is no significant difference in transaction patterns between legitimate and fraudulent credit card transactions.

Alternative hypothesis (H_1): There is a significant difference in transaction patterns between legitimate and fraudulent credit card transactions.

If the p-value is less than the chosen level of significance (usually 0.05), the null hypothesis will be rejected in favor of the alternative hypothesis.

Overview : Binary Logistic Regression

A logistic regression (often referred to simply as Binary logistic regression), predicts the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical.

If, on the other hand, your dependent variable is a count, then Poisson regression is to be use. Alternatively, if you have more than two categories of the dependent variable, we chose multinomial logistic regression.

For example,

- To understand whether exam performance can be predicted based on revision time, test anxiety and lecture attendance (i.e., where the dependent variable is “Results", measured on a dichotomous scale – “arrears" or “no arrears")

Assumptions :

The following are some of the assumptions for logistic regression:

Linearity, Independence, Sample size, No multicollinearity, No influential outliers, Binary response, Absence of interactions, No perfect separation, Model specification are the one must obey in order to perform the analysis for model building.

2. Data Source - Overview:

- Credit Card Fraud Detection dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud

In the Credit Card Fraud Detection dataset, the variables v1, v2, v3, ..., v28 are numeric input variables that represent the result of a principal component analysis (PCA) transformations,as they were anonymized for privacy reasons.

These variables may represent various features related to the credit card transactions, such as transaction amounts, locations, timestamps, or other characteristics. In which it consists of *284807* instances with *31* attributes respectively.Finally, the *Class* variable is a binary variable indicating whether a transaction is fraudulent (Class = 1) or legitimate (Class = 0).

Procedure:

Step 1: ***Import the dataset***

First, we need to import the Credit Card Fraud Detection dataset into SPSS.

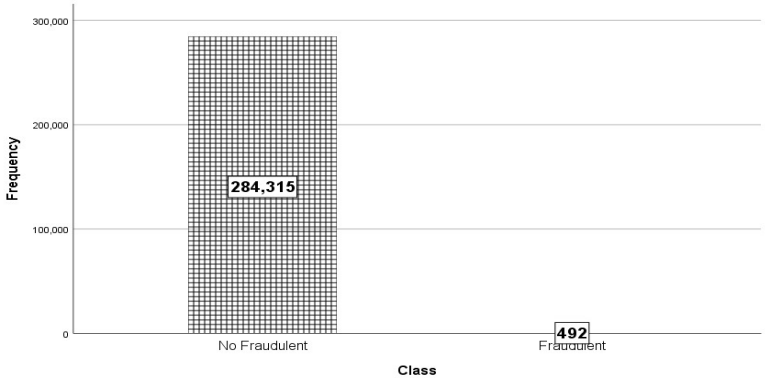
- ✓ Open SPSS and go to File > Import Data > CSV Data.
- ✓ Import>select the first option "Read variable names from the first row of data."
- ✓ Click on OK button on Read CSV file dialog box.

Step 2: ***Data Preparation***

Prepare the dataset by modifying datatypes, identifying missing values, duplicates,check the balancing of the data,etc..

- Analyze > Descriptive statistics > Frequencies > chose all desired variables > Charts

  (Bar charts) >uncheck Frequency tables and click OK.



Notice how *imbalanced* is our original dataset! Most of the transactions are non-fraud.We might get a *lot of errors* and our algorithms will probably *overfit* ,we want our model to *detect patterns* that give signs of fraud!

So, We can have a *SMOTE* (Synthetic Minority Oversampling Technique) is one of the most commonly used *oversampling* methods to solve imbalance problems. we make use of *PYTHON* to implement this.

Pycode…,

```
from imblearn.over_sampling import SMOTE

import pandas as pd

data = pd.read_csv("\\..\\Desktop\\Creditcard\\creditcard.csv")

# Split the data into predictor variables (X) and outcome variable (y)

X = data.drop('Class', axis=1)

y = data['Class']

smote = SMOTE(sampling_strategy='auto')

X_resampled, y_resampled = smote.fit_resample(X, y)

resampled_data = pd.concat([X_resampled, y_resampled], axis=1)

resampled_data.to_csv("\\.. \\Desktop\\Creditcard\\creditresampled_data.csv", index=False)

print("success")
```

On successful compiling ,it'll return a balanced dataset with this we will move forward in doing Logistic regression model building

Analysis on SMOTE resampled data **:**

- Now our dataset consists of **568630** instances and **31** attributes with " 0 " Missing observations.
  - Data > Identify Duplicate Cases >select all the desired variables for the label " Define matching cases by " > click OK

- Now all the unique values have code of (1) and Duplicate values have code of (0).

Indicator of each last matching case as Primary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Duplicate Case	8601	1.5	1.5	1.5
	Primary Case	560029	98.5	98.5	100.0
	Total	568630	100.0	100.0	

Now chose to separate the dataset without duplicates which can be used for analysis

- Data > Select Cases > chose the radio option " If condition is satisfied " >enter the if condition (=> PrimaryLast = 1) >click continue & OK.

Step 3: ***Model Building***

we will use logistic regression as the machine learning algorithm.

- Data > select cases >Random sample of cases (80%) > filter out unselect cases
- Create train and test datasets using *$filter variable* and for 80% code (1) ,for 20% code (0).

Step 4: **Feature-Selection** :

Correlation matrices helps in understanding our data. We want to know if there are features that influence heavily in whether a specific transaction is a fraud.

Negative Correlations: V10, V12, V14 and V17 are *negatively correlated*, the lower these values are, more likely results in a fraud transaction.

Positive Correlations: V2, V4, V11, and V19 are *positively correlated*. the higher these values are, the more likely results in a fraud transaction.

So we'll consider these variables as they are effective with the dependent variable.

- Go to File > open data > choose the training set > Click Open.
- Analyze > Regression > Binary Logistic.
- In the Logistic Regression dialog box, select "Class" as the dependent variable and choose correlation variables as the independent variables into "Covariates" label .
- Click on Save and choose predicted values as "Probabilities" and Continue…
- Choose Options and select Classification plots, Hosmer-Lemeshow goodness of fit, Case wise listing of residuals and CI for exp(B) .
- Choose the method as "Enter " .Click OK to run the analysis.

3. Result :

Logistic Regression

Block 0: Beginning Block

- This part of the output describes a "null model", which is model with no predictors and just the intercept.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.023	.003	58.610	1	.000	.977

Block 1: Method = Enter

- Its an overall test of the model,From Omnibus Tests of Model Coefficients: Step,Block,Model have 8 df with sig.val of 0.00 along with chisq val of 546426.024.
- *Cox & Snell R Square* and *Nagelkerke R Square* values are referred to as *pseudo* R^2 values. so,our model ranges from **70.5% to 94.0%**.Therefore, The overall model is statistically significant, $\chi 2(8) = 546426.024$, $p < 0.05$ with deviance of 74500.376.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	74500.376[a]	.705	.940

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.513 E+ 12	8	.320

Hosmer-Lemeshow goodness-of-fit test is a useful method to assess the fit of a logistic regression model. A non-significant p-value (typically greater than 0.05) indicates that the model fits the data well.

Category prediction:

If the estimated probability of the event occurring is greater than or equal to 0.5,classifies the event as occurring (e.g., fraud is being present). If the probability is less than 0.5,classifies the event as not occurring (e.g., no fraud).

Classification Table[a]

Observed			Predicted Class: No Fraudulent	Predicted Class: Fraudulent	Percentage Correct
Step 1	Class	No Fraudulent	223716	2819	98.8
		Fraudulent	9723	211688	95.6
	Overall Percentage				97.2

- **Accuracy :** The percentage of cases that can be correctly classified as "no" fraud with the independent variables added .The model now correctly classifies 97.2% of cases .
- **Sensitivity :** The percentage of cases that had the observed characteristic (for fraud) which were correctly predicted by the model is 95.83 % of fraud cases were also predicted by the model to have fraud

The Wald test ("**Wald**" column) in the "Variables in the Equation" is used to determine statistical significance for each of the independent variables. From these results we can see that ***V4,V10, V11,***

V12, V14,V17,V19 are added significantly to the model/prediction and remaining variables along with constant did not add significantly to the model.

Step 4: ***Model Evaluation***

We will use confusion matrix to evaluate the performance of the model.

- Go to Analyze > Descriptive Statistics > Crosstabs.>select "Class" as the row variable and compute a variable for "Predicted Probability > 0.5" as the column variable.
- Click OK to run the analysis.

Class * Prob Crosstabulation

Count

		Prob		
		.00	1.00	Total
Class	No Fraudulent	223716	2819	226535
	Fraudulent	9723	211688	221411
Total		233439	214507	447946

Evaluation Metrics :

Few commonly used statistical measures in binary classification problems to evaluate the performance of a predictive model.

- **Accuracy**: It is the proportion of correctly classified samples out of the total number of samples. Its 97.2 % for train set and 97.19% for test set.
- **Misclassification**: It is the proportion of incorrectly classified samples out of the total number of samples. Its 0.028 % for train set and 0.0281% for test set.
- **Sensitivity** (***True Positive Rate***): It is the proportion of true positives out of all actual positives. Its 95.83 % for train set and 95.85 % for test set.
- **Specificity** (***True Negative Rate***): It is the proportion of true negatives out of all actual negatives. Its 98.69 % for train set and 98.64% for test set.
- **Positive Predictive Value** (***PPV***): It is the proportion of true positives out of all predicted positives. Its 98.7556 % for train set and 98.71% for test set.
- **Negative Predictive Value** (***NPV***): It is the proportion of true negatives out of all predicted negatives. Its 95.6086 % for train set and 95.62% for test set.
- **Positive Likelihood Ratio** (***PLR***): It is the ratio of true positives to false positives. It is calculated as Sensitivity / (1 - Specificity). Its 72.92 % for train set and 70.47% for test set.
- **Negative Likelihood Ratio** (***NLR***): It is the ratio of false negatives to true negatives. It is calculated as (1 - Sensitivity) / Specificity. Its 0.028 % for train set and 0.0281 for test set.
- **F1-Score :** 2*((PPV*Recall) / (PPV +Recall)) **.** Its 0.04221 % for train set and 0.04209% for test set.

Since our dataset is unbalanced one we should use RECALL which is Sensitivity instead of ACCURACY .Therefore ,our model had an efficiency of 95.83 %,which process best bit for our dataset.

4. Conclusion:

In conclusion, the main objective of this assignment was to find the most suited model in credit card fraud detection in terms of the machine learning techniques. Finally this model helps in detecting fraud transactions

For Train set & For Test set:

The logistic regression model was statistically significant at $\chi^2(8) = 546426.064$,

$p=0.00<0.05$. The model explained 94.00% (Nagelkerke R^2) of the variance and correctly classified 97.2% of cases accurately with only 12542 misclassified instances for 80% trainset data respectively.

Similarly, the model was statistically significant at $\chi^2(8) = 136775.109$, $p=0.00<0.05$. The model explained 94.00% (Nagelkerke R^2) of the variance and correctly classified 97.2% of cases accurately with only 3155 misclassified for 20% of test set data.

ROC curve:

From the ROC curve it is determined that, area under the curve is 0.798 along the asymptotic significance which is less than p=0.05 And the lesser the deviance of the model more the model is efficient ,for an true area of null hypothesis 0.5 respectively.

References :

- ✓ Bhanusri, A., Valli, K. R. S., Jyothi, P., Sai, G. V., & Rohith, R. (2020). Credit card fraud detection using Machine learning algorithms. Journal of Research in Humanities and Social Science, 8(2), 04-11.
- ✓ A machine learning based credit card fraud detection using the GA algorithm for feature selection Emmanuel Ileberi1*, Yanxia Sun1 and Zenghui Wang2
- ✓ Safa, M. U., & Ganga, R. M. (2019). Credit Card Fraud Detection Using Machine Learning. International Journal of Research in Engineering, Science and Management, 2(11).
- ✓ Credit Card Fraud Detection Using Machine Learning, Mr. Thirunavukkarasu.M1 ; Achutha Nimisha2 ; Adusumilli Jyothsna3