Fraud Classifier Using Logistic Regression on Credit Card Transaction Data

Project Report 960:563 Fall 2023
Name: Udayveer Singh Andotra
NetID: ua118

Table of Contents

1.Abstract

Modelled the frauds in a transaction dataset using Logistic regression. The regressors used were continuous as well binary. The binary variables were created from continuous and categorical variables in the dataset by selecting a suitable partition to differentiate the frauds from non-frauds. The regressors for the model were selected on the basis of significance level of 0.5 and non-multicollinearity. The model also utilizes lasso regression and cross validation to get the model coefficients. Finally, the accuracy of the model in detecting frauds was calculated to be over 99% for training as well as validation.

2.Introduction

Fraud is not a rare event when it comes to financial transactions. When it is reported that a case of 'fraud' has been observed, it is the duty of the concerned financial institution to check the validity of such a claim. Hence, it would be robust to have a classifier for 'fraud' using the transaction data of the customers/organizations.
The following shows the frauds reported and the losses observed for US states in 2023. It is predicted that the financial frauds will see an increase in the future not just in US but globally.

			Fraud Reports		Fraud Loss	
Rank	State	Ranking Index	Per 100,000 Residents	Total	Median Loss	Total Loss
1	California	100.00	244	46,999	$700	249M
2	Florida	99.63	380	29,602	$620	99.9M
3	New Jersey	91.57	295	11,229	$593	44.8M
4	Texas	91.20	274	29,856	$540	119.6M
5	Georgia	89.75	437	13,931	$600	33.4M
6	Arizona	87.92	262	9,890	$650	33.5M
7	Maryland	87.31	360	9,239	$580	34.7M
8	New York	85.25	258	22,688	$500	64.9M
9	Illinois	85.07	285	14,528	$511	42.3M
10	Pennsylvania	84.64	310	15,358	$500	38.8M

Note: All figures reflect Q1 2023 data from the Federal Trade Commission.

Source: Federal Trade Commission [Analyzed by Forbes Advisor] • Get the data • Embed

Forbes ADVISOR

The financial institutions have their risk management wings to deal with this (amongst other responsibilities) and hereby reduce the loss due to fraud by –

1. Accurately identifying the cases of frauds.
2. Deploy necessary framework to mitigate the frauds under similar circumstances.

I would be addressing the 'identification' part of the 'loss due to fraud reduction'. This project is my attempt at building a classifier predominating using variable creation, logistic regression and variable/model selection. The following sections describe the dataset, the methodology used and showcase the results. The full code and output can be found in the Appendix.

3.Materials and Methods

Dataset

The Dataset is a simulated credit card transaction dataset containing legitimate and fraud transactions from the duration 1st Jan 2019 - 31st Dec 2020. It covers credit cards of 1000 customers doing transactions with a pool of 800 merchants.
Development/Training: Jun2020-Dec2020
OutOfBag/Validation: Jan2019-Dec2019

The training data has 555719 transactions and validation has 1296675 transactions. The transactions have not been rolled-up to customer ID level. The scope of this project is to classify the transactions and not to take into account the multiple transactions per customer. In other words, each transaction is considered to be linked to a unique customer. This can be modified to incorporate the rolled-up results depending on the problem statement.

Variables

trans_date_trans_time: Date

amt: Continuous. Transaction Amount

gender: Categorical. Male → M, Female → F

state: Categorical. US State code

city_pop: Continuous. Population Number

job: Categorical. Job Type

dob: Date. Date of Birth

category: Categorical. Transaction Category

merchant: Categorical. Merchant Description

is_fraud: Binary. Fraud Indicator. Fraud → 1

maleF: Binary. (gender == 'M') → 1

popG500: Binary. (city_pop > 500) → 1

catShop_net: Binary. (category == "shopping_net") → 1

catMisc_net: Binary. (category == "misc_net") → 1

stateHF: Binary. state belongs to ("AK", "CT") →1

stateLF: Binary. state belongs to ("UT", "VT", "WV", "NV", "RI") →1

merch: Binary. merchant belongs to

```
("fraud_Heathcote, Yost and Kertzmann","fraud_Lemke-Gutmann",
"fraud_Mosciski, Ziemann and Farrell","fraud_Kilback LLC",
```

"fraud_Heathcote LLC","fraud_Boyer PLC",

"fraud_Medhurst PLC","fraud_Bashirian Group",

"fraud_Schultz, Simonis and Little","fraud_Miller-Hauck",

"fraud_Romaguera, Cruickshank and Greenholt") → 1

jobHF: Binary. job belongs to

("Accountant, chartered certified", "Buyer, retail", "Commissioning editor",

"Conservator, furniture", "Designer, television/film set", "Horticultural consultant",

"Hydrogeologist", "Investment banker, operational", "Nature conservation officer",

"Surveyor, hydrographic", "TEFL teacher", "Television camera operator",

"Tour manager", "Visual merchandiser") → 1

age: Continuous. Calculated using dob and trans_date_trans_time

ageHF: Binary. (age > 60) →1

pred_val: Continuous. Fitted values from model (0 to 1)

pred_fraud: Binary. (pred_val > 0.5) → 1

Transforming the output response to binary (Fraud or not Fraud)

The green highlighted variables are present in the raw dataset and red highlighted variables are derived from them. The contingency tables for the derived binary variables indicate the partition's ability to differentiate frauds from not frauds. The partition was selected in a way that-

1. Fraud Rate is noticeably different for the two binary responses 0 and 1.
2. The population sample for the partitions are not too small. Failure will result in introducing a bias which is not ideal.

##	is_fraud		
## catShop_net	0	1	Fraud Rate
## 0	512301	1639	0.319%
## 1	41273	506	1.211%

##	is_fraud		
## catMisc_net	0	1	Fraud Rate
## 0	526474	1878	0.355%
## 1	27100	267	0.976%

##	is_fraud		
## popG500	0	1	Fraud Rate
## 0	98384	257	0.261%
## 1	455190	1888	0.413%

##	is_fraud		
## merch	0	1	Fraud Rate
## 0	552546	2132	0.038%
## 1	1028	13	1.249%

##	is_fraud		
## stateHF	0	1	Fraud Rate
## 0	551592	2118	0.383%
## 1	1982	27	1.344%

##	is_fraud		
## stateLF	0	1	Fraud Rate
## 0	548947	2145	0.389%
## 1	4627	0	0.000%

The fraud rate for catShop_net, catMisc_net, popG500, merch and state_HF is strikingly different. However, for merch the sample size for indicator equals 1 is quite small which can result in bias. state_LF on the other hand does not statistically differentiate significantly. We can drop them right now or later after seeing the results from logistic regression. I chose the latter.

Methodology

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). 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. It uses the sigmoid function is referred to as an activation function for logistic regression and is defined as:

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

here,

- e = base of natural logarithms
- x = numerical value one wishes to transform

The logistic regression equation is:

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

here,

- x = input value
- y = predicted output
- b0 = intercept
- b1 = coefficient for input (x)

For more than one regressor x, the exponential becomes b0 + b1x1 + b2x2 + b3x3…...

logit(y)=ln(y/1-y) = b0 + b1x1 + b2x2 + b3x3…...

Dependent variable/response: is_fraud
Independent variables/regressors: city_pop, amt, age, maleF, popG500, catShop_net, catMisc_net, merch, stateHF, stateLF, jobHF and ageHF

Modeling the binary response to a simple multinomial model would result in the predicted response not bounded to (0,1). The adj. R squared for such a model would be extremely small.

```
##
## Call:
## lm(formula = is_fraud ~ city_pop + amt + age + maleF + popG500 +
##     catShop_net + catMisc_net + merch + stateHF + stateLF + jobHF +
##     ageHF, data = fraudTest)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62948 -0.00535 -0.00212  0.00049  1.00369
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.011e-03  3.775e-04 -13.276  < 2e-16 ***
## city_pop    -1.173e-09  2.756e-10  -4.257 2.07e-05 ***
## amt          7.166e-05  5.216e-07 137.387  < 2e-16 ***
## age          3.611e-05  7.641e-06   4.726 2.30e-06 ***
## maleF        9.324e-05  1.645e-04   0.567   0.5708
## popG500      1.617e-03  2.178e-04   7.425 1.13e-13 ***
## catShop_net  8.175e-03  3.118e-04  26.216  < 2e-16 ***
## catMisc_net  6.179e-03  3.785e-04  16.325  < 2e-16 ***
## merch        3.491e-03  1.896e-03   1.841   0.0656 .
## stateHF      9.913e-03  1.362e-03   7.276 3.43e-13 ***
## stateLF     -3.513e-03  8.999e-04  -3.903 9.48e-05 ***
## jobHF        2.291e-02  3.838e-03   5.970 2.37e-09 ***
## ageHF        2.251e-04  3.280e-04   0.686   0.4926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06091 on 555706 degrees of freedom
## Multiple R-squared:  0.0352, Adjusted R-squared:  0.03518
## F-statistic:  1690 on 12 and 555706 DF,  p-value: < 2.2e-16
```

The variables are selected by step-wise logistic using significance level as the criteria as well as lasso regression by looking at the regressor coefficients.
The coefficients are determined for the final model using K-cross validation using 10 folds. Model accuracy is finally calculated based on the no. of correct predictions.

4.Results and Discussion

- Null Model vs Full Model

```
lr.stat<-lrtest(mod00,mod0)
lr.stat
```

```
## Likelihood ratio test for MLE method
## Chi-squared 12 d.f. =  2794.565 , P value =  0
```

p-value < 0.05 implies that the Null Hypothesis can be rejected. There is at least one variable which is statistically significant for the logistic regression.

- Full Model Summary

```
##
## Call:
## glm(formula = is_fraud ~ city_pop + amt + age + maleF + popG500 +
##     catShop_net + catMisc_net + merch + stateHF + stateLF + jobHF +
##     ageHF, family = binomial(logit), data = fraudTest)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.968e+00  1.133e-01 -61.499  < 2e-16 ***
## city_pop    -4.296e-07  1.026e-07  -4.188 2.82e-05 ***
## amt          1.701e-03  4.143e-05  41.043  < 2e-16 ***
## age          9.848e-03  2.111e-03   4.666 3.07e-06 ***
## maleF        8.993e-03  4.489e-02   0.200   0.8412
## popG500      5.260e-01  6.901e-02   7.622 2.51e-14 ***
## catShop_net  1.375e+00  5.477e-02  25.106  < 2e-16 ***
## catMisc_net  1.279e+00  6.783e-02  18.859  < 2e-16 ***
## merch        4.845e-01  2.868e-01   1.689   0.0912 .
## stateHF      1.357e+00  1.977e-01   6.866 6.62e-12 ***
## stateLF     -1.285e+01  9.327e+01  -0.138   0.8904
## jobHF        1.852e+00  3.918e-01   4.727 2.28e-06 ***
## ageHF       -3.358e-02  8.790e-02  -0.382   0.7025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28122  on 555718  degrees of freedom
## Residual deviance: 25327  on 555706  degrees of freedom
## AIC: 25353
##
## Number of Fisher Scoring iterations: 17
```

maleF, merch, stateLF and ageHF have p-values greater than 0.05. These would be dropped by the stepwise logistic.

- Full Model logistic stats

```
##
## Logistic regression predicting is_fraud
##
##                      crude OR(95%CI)             adj. OR(95%CI)              P(Wald's test) P(LR-test)
## city_pop (cont. var.) 1 (1,1)                     1 (1,1)                     < 0.001        < 0.001
##
## amt (cont. var.)      1.0018 (1.0017,1.0019)      1.0017 (1.0016,1.0018)      < 0.001        < 0.001
##
## age (cont. var.)      1.0071 (1.0047,1.0096)      1.0099 (1.0057,1.0141)      < 0.001        < 0.001
##
## maleF: 1 vs 0         1.0245 (0.9409,1.1155)      1.009 (0.924,1.1018)        0.841          0.913
##
## popG500: 1 vs 0       1.59 (1.39,1.81)            1.69 (1.48,1.94)            < 0.001        < 0.001
##
## catShop_net: 1 vs 0   3.83 (3.47,4.24)            3.96 (3.55,4.4)             < 0.001        < 0.001
##
## catMisc_net: 1 vs 0   2.76 (2.43,3.14)            3.59 (3.15,4.1)             < 0.001        < 0.001
##
## merch: 1 vs 0         3.28 (1.89,5.67)            1.62 (0.93,2.85)            0.091          0.113
##
## stateHF: 1 vs 0       3.55 (2.42,5.2)             3.89 (2.64,5.72)            < 0.001        < 0.001
##
## stateLF: 1 vs 0       0 (0,9.26246775532108e+75)  0 (0,6.48430949173778e+73)  0.89           < 0.001
##
## jobHF: 1 vs 0         7.39 (3.48,15.69)           6.37 (2.96,13.73)           < 0.001        < 0.001
##
## ageHF: 1 vs 0         1.23 (1.11,1.36)            0.97 (0.81,1.15)            0.702          0.732
```

stateLF has an extremely small confidence region and its statistical significance is in question. It's insignificant by Wald's test but significant by the LR-test at 0.05 significance level.

- Step-wise Logistic

```
##      Table 1. Summary of Parameters
## ===================================
##           Paramters            Value
## -----------------------------------
## Response Variable             is_fraud
## Included Variable             NULL
## Selection Method              backward
## Select Criterion              SL
## Stay Significance Level(sls)  0.05
## Variable significance test    LRT
## Multicollinearity Terms       NULL
## Intercept                     1
## ===================================
##
##                     Table 3. Process of Selection
## ================================================================
##  Step  EnteredEffect  RemovedEffect  DF  NumberIn        SL
## ----------------------------------------------------------------
## 1                                   13  13       1
## 2                    maleF          1   12       0.913497382380363
## 3                    ageHF          1   11       0.732067230114868
## 4                    merch          1   10       0.112979682867906
## ================================================================
```

Table 5. Coefficients of the Selected Variables

Variable	Estimate	StdError	t.value	P.value
(Intercept)	-6.94213554928317	0.0961222765029952	-72.2219219294796	0
city_pop	-4.29158421051173e-07	1.0247277314803e-07	-4.18802388056014	2.81393913399281e-05
amt	0.00170050975755561	4.14309625396404e-05	41.0444183122368	0
age	0.00921031088760975	0.00129086573242853	7.13498751747188	9.67959435817031e-13
popG500	0.527408602554556	0.0689357020426591	7.65073230454928	1.99837969480002e-14
catShop_net	1.38319960256441	0.0544189399870033	25.4176138472149	1.61084793167751e-142
catMisc_net	1.27822395351705	0.0678183060405209	18.8477717617029	3.06482550437153e-79
stateHF	1.35766499129452	0.197440334456748	6.8763304875374	6.14139340363604e-12
stateLF	-12.8473747396102	93.271492485327	-0.13774170861082	0.890444554850472
jobHF	1.84988703464586	0.391843542516338	4.72098384668091	2.34706580222554e-06

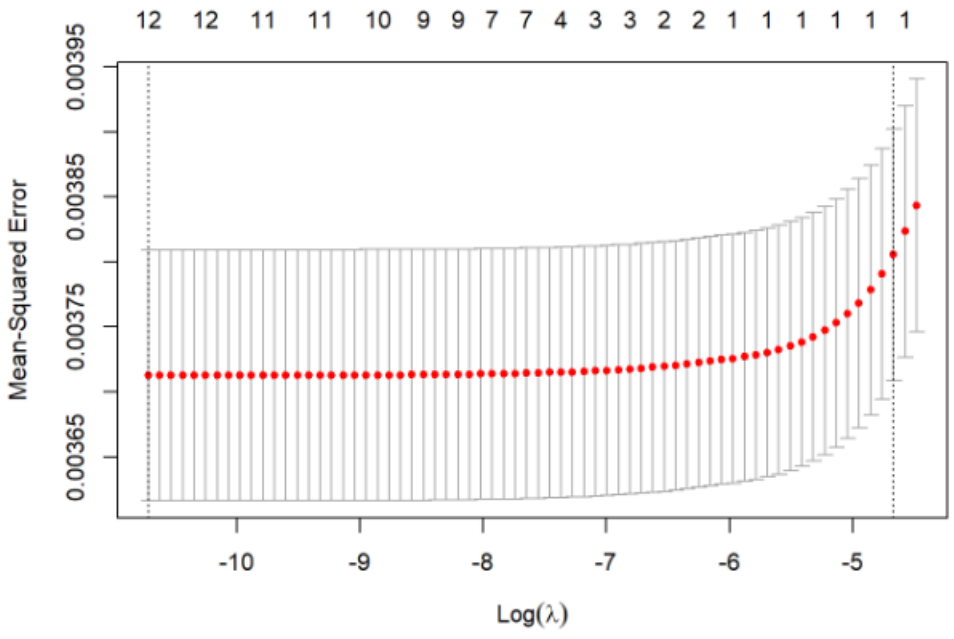
maleF, ageHF and merch are removed by the backward step-wise logistic regression at 0.05 significance level. stateLF is however still present.

- Lasso Regression for Variable Coefficients

```
best_lambda
```

```
## [1] 2.218624e-05
```

```
# produce plot of test MSE by lambda value
plot(cv_model)
```



```
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda, family = binomial(logit))
coef(best_model)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) -6.898614e+00
## city_pop    -3.884517e-07
## amt          1.697987e-03
## age          8.850904e-03
## maleF        .
## popG500      4.998302e-01
## catShop_net  1.363894e+00
## catMisc_net  1.263846e+00
## merch        4.490240e-01
## stateHF      1.335530e+00
## stateLF     -2.224019e+00
## jobHF        1.855456e+00
## ageHF        .
```

Determining the best lambda for lasso and looking at the coefficients indicates towards dropping maleF and ageHF. Lasso reduces the coefficients of multicollinear variables. The coefficients however do look almost identical from the stepwise logistic. From the results of stepwise logistic and lasso, variables dropped – ageHF, maleF and merch.

- ## Correlation Matrix

```
##             city_pop   amt   age popG500 catShop_net catMisc_net stateHF   stateLF jobHF
## city_pop        1.00  0.00 -0.09    0.14        0.00        0.00   -0.02    -0.02  0.00
## amt             0.00  1.00 -0.01    0.01        0.03        0.01    0.00     0.00  0.00
## age            -0.09 -0.01  1.00   -0.12       -0.02        0.00    0.01    -0.03  0.01
## popG500         0.14  0.01 -0.12    1.00        0.00       -0.01   -0.02    -0.02  0.01
## catShop_net     0.00  0.03 -0.02    0.00        1.00       -0.06    0.00     0.00  0.00
## catMisc_net     0.00  0.01  0.00   -0.01       -0.06        1.00    0.00     0.00  0.00
## stateHF        -0.02  0.00  0.01   -0.02        0.00        0.00    1.00    -0.01  0.00
## stateLF        -0.02  0.00 -0.03   -0.02        0.00        0.00   -0.01     1.00  0.00
## jobHF           0.00  0.00  0.01    0.01        0.00        0.00    0.00     0.00  1.00
```

Multicollinearity is not present. stateLF has small confidence interval, ambiguous significance value and low sample size when its indicator is 1. It is dropped also before the cross-validation model.

- ## Cross-Validation

```
## Generalized Linear Model
##
## 555719 samples
##      8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 500147, 500147, 500147, 500147, 500148, 500147, ...
## Resampling results:
##
##   RMSE        Rsquared     MAE
##   0.06286707  0.004340411  0.007570109
```

The 10-fold cross validation logistic regression has resulted in MAE = 0.0075 and the following model coefficients:

```
##
## Call:
## NULL
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.960e+00  9.604e-02 -72.468  < 2e-16 ***
## city_pop    -4.220e-07  1.023e-07  -4.123 3.73e-05 ***
## amt          1.702e-03  4.141e-05  41.104  < 2e-16 ***
## age          9.338e-03  1.291e-03   7.236 4.63e-13 ***
## popG500      5.310e-01  6.892e-02   7.705 1.31e-14 ***
## catShop_net  1.383e+00  5.441e-02  25.411  < 2e-16 ***
## catMisc_net  1.277e+00  6.782e-02  18.832  < 2e-16 ***
## stateHF      1.367e+00  1.974e-01   6.922 4.46e-12 ***
## jobHF        1.856e+00  3.919e-01   4.736 2.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28122  on 555718  degrees of freedom
## Residual deviance: 25362  on 555710  degrees of freedom
## AIC: 25380
##
## Number of Fisher Scoring iterations: 9
```

The p-value for model performance is 0. The model is significant

```
# p-value for model performance
1-pchisq(summary(mod2)$deviance, summary(mod2)$df[1]-1)
```

```
## [1] 0
```

- ## Model Accuracy

Training: 99.58%
Validation: 99.39%

The model is highly accurate for the detecting frauds from both the training as well as validation set. This was however a simulated dataset, real world datasets are more challenging in terms of missing data, outliers and require more rigorous variable creation and transformations.

5.Acknowledgements

The dataset used was generated using Sparkov Data Generation | Github tool created by Brandon Harris. This simulation was run for the duration - 1 Jan 2019 to 31 Dec 2020. The files were combined and converted into a standard format.

The dataset was made available on Kaggle by Kartik Shenoy.

6.References

1. Most Scammed States In America: Financial Fraud Statistics – Forbes Advisor
2. Introduction to Linear Regression Analysis, 6th Edition (Wiley, 6th Edition, 2021)

7.Appendix

R code file: 
ProjectFall23_ua118.Rmd

R output file: 
ProjectFall23_ua118.html