Analysis of Credit Card Fraud

Michaela Bodie, Russell Chan, Joseph Giannantonio

2023-05-19

The Problem

The world around us continues to advance technologically, but with the good also comes the bad. One notable technological advancement is the evolution of digital payments. With this widespread progression however, has also come the evolution of cyber criminals. Credit card fraud has emerged as a prevalent and continuously evolving threat in the present day world. According to the Data Breach Index, at least 5 million records are being stolen on a daily basis. The use of online transactions and expanding dependence on electronic payment systems has lead to fraudulent activities targeting credit cards becoming more and more complex and elusive. This report delves into the analysis of the different factors and variables that play a significant role in being able to detect fraudulent activity regarding online transactions.

The Data

Kaggle maintains records of critical features for a fraudulent transaction documented into one csv file. There are a total of 1,000,000 files recorded. Each record is as follows:

distance_from_home: the distance from home where the transaction happened

distance_from_last_transaction: the distance from the last transaction that happened

ratio_to_median_purchase_price: ratio of purchased price transaction to median purchase price

repeat_retailed: is the transaction that happened from similar retailers

used_chip: did the transaction happen through credit card chip

used_pin_number: did the transaction happen using PIN number

online_order: is the transaction an online order

fraud: is the transaction fraudulent

To validate the data we will sample 10,000 observations of fraud, and 10,000 observations of non fraud activity.

```
setwd("C:/Users/macho/Desktop/sfsu/spr23/MATH449")
fraud1 <- read.csv("card_transdata.csv", header = TRUE)
head(fraud1)
```

```
##   distance_from_home distance_from_last_transaction
## 1          57.877857                      0.3111400
## 2          10.829943                      0.1755915
## 3           5.091079                      0.8051526
## 4           2.247564                      5.6000435
## 5          44.190936                      0.5664863
## 6           5.586408                     13.2610733
##   ratio_to_median_purchase_price repeat_retailer used_chip used_pin_number
## 1                     1.94593998               1         1               0
## 2                     1.29421881               1         0               0
## 3                     0.42771456               1         0               0
## 4                     0.36266258               1         1               0
## 5                     2.22276730               1         1               0
## 6                     0.06476847               1         0               0
##   online_order fraud
## 1            0     0
## 2            0     0
## 3            1     0
## 4            1     0
## 5            1     0
## 6            0     0
```

Logistic Regression Model

Beginning the analysis, a generalized logistic regression model was ran utilizing all of the variables in the dataset to determine which predictors may be significant. As seen in the output below, the corresponding p-values of the variables present the assumption that all variables are significant. In the next section, other inferences are performed to truly determine the best subset of variables for this dataset.

```
#stratify data for equal fraud and non-fraud observations
I = which(fraud1$fraud == 1)
J = which(fraud1$fraud == 0)

I1 = sample(I,10000)
J1 = sample(J,10000)

new_data = rbind(fraud1[I1,],fraud1[J1,])

# run a logistic regression model with all variables
fraudLogit <- glm(fraud ~ ., data = new_data, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

fraudulent <- summary(fraudLogit)
fraudulent

##
## Call:
## glm(formula = fraud ~ ., family = binomial, data = new_data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.2547   0.0000   0.2715   4.0121
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.484e+00  1.546e-01  -48.41   <2e-16 ***
## distance_from_home              2.763e-02  6.325e-04   43.68   <2e-16 ***
## distance_from_last_transaction  5.217e-02  1.731e-03   30.13   <2e-16 ***
## ratio_to_median_purchase_price  1.162e+00  1.875e-02   61.96   <2e-16 ***
## repeat_retailer                -1.366e+00  9.150e-02  -14.93   <2e-16 ***
## used_chip                      -1.137e+00  6.492e-02  -17.52   <2e-16 ***
## used_pin_number                -1.215e+01  5.398e-01  -22.50   <2e-16 ***
## online_order                    4.916e+00  1.222e-01   40.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27725.9  on 19999  degrees of freedom
## Residual deviance:  8720.2  on 19992  degrees of freedom
## AIC: 8736.2
##
## Number of Fisher Scoring iterations: 10
```

Selecting the Best Subset of Variables

After running the model above that included all independent variables, Wald Test and Chi-Square Goodness of Fit test were ran to test the significance of the predictors as well. The ANOVA and Wald Tests demonstrated that all variables are indeed significant towards the model, and the Chi-Square Goodness of Fit test produced a p-value of 0, indicating that the model utilized above was a perfect fit for the data.

```
wald.test(Sigma = vcov(fraudLogit), b = coef(fraudLogit), Terms = 1:7)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 4098.8, df = 7, P(> X2) = 0.0
```

```
# compare to the null:  p-value for the chi-squared test of the model's
goodness of fit
pvalue <- 1-pchisq(fraudulent$null.deviance - fraudulent$deviance,
fraudulent$df.null - fraudulent$df.residual)
pvalue
```

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

The Final Model

With regards to the inferences ran above, the final logistic regression model is as follows:

$$logit[P(Y = 1)] = -7.533 - 0.02742(X_1) + 0.05224(X_2) + 1.175(X_3) - 1.26(X_4) - 1.16(X_5) - 11.16(X_6) + 4.862(X_7)$$

β0 = -10.36 / The log-odds of Credit Card Fraud is -10.36 if all independent variables are set to 0

β1 = 0.1522 / When increasing the factor distance_from_home by 1 unit, the probability of credit card fraud increases by approximately exp(.01522)

β2 = .02526 / When increasing the factor distance_from_last_transaction by 1 unit, the probability of credit card fraud increases by approximately exp(.02526)

β3 = .8623 / When increasing the factor ratio_to_median_purchase_price by 1 unit, the probability of credit card fraud increases by approximately exp(.8623)

β4 = -.6215 / When increasing the factor repeat_retailer by 1 unit, the probability of credit card fraud decreases by approximately exp(.6215)

β5 = -1.049 / When increasing the factor used_chip by 1 unit, the probability of credit card fraud decreases by approximately exp(1.049)

β6 = -13.74 / When increasing the factor used_pin_number by 1 unit, the probability of credit card fraud decreases by exp(13.74)

β7 = 6.651 / When increasing the factor online_order by 1 unit, the probability of credit card fraud increased by exp(6.651)

Confusion Table

The resulting Confusion Matrix is as follows:

	0	1
0	2785	151
1	215	2849

The corresponding sensitivity is 0.9283 and specificity is 0.9497. Based on these values and the confusion matrix, it can be assumed that the model is more than adequate at correctly classifying a fraudulent credit card transaction. Our confusion matrix has an accuracy rate of 0.939, with a significant p value. Specifically, the proportion of credit card transactions that were correctly classified as not fraudulent is 0.9497, and the proportion of credit card transactions that were correctly classified as fraudulent is 0.9283. Our confusion matrix supports the significance of our logistic regression model.

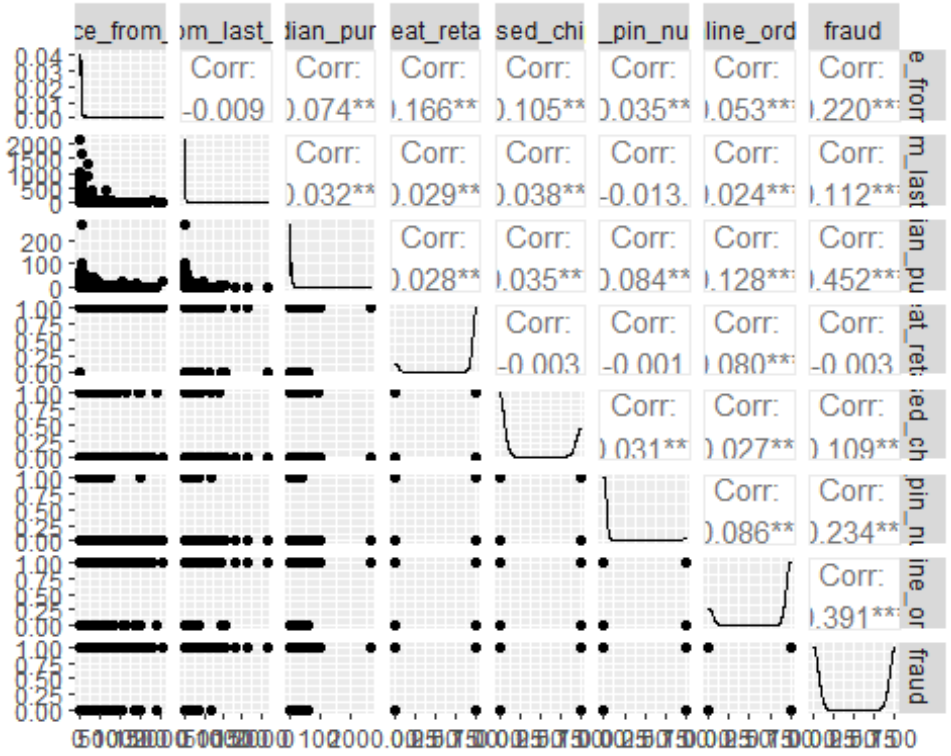
```
library(caTools)
```

```r
#use 70% of dataset as training set and 30% as test set
sample <- sample.split(new_data$fraud, SplitRatio = 0.7)
train  <- subset(new_data, sample == TRUE)
test   <- subset(new_data, sample == FALSE)

confusionModel <- glm(fraud~ ., data = train, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

predicted <- predict(confusionModel, test, type = "response")

predicted <- ifelse(predicted > 0.5, "1", "0")

predicted <- as.factor(predicted)
test$fraud <- as.factor(test$fraud)

confusionMatrix(data=predicted, reference=test$fraud)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2774  144
##          1  226 2856
##
##                Accuracy : 0.9383
##                  95% CI : (0.9319, 0.9443)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8767
##
##  Mcnemar's Test P-Value : 2.543e-05
##
##             Sensitivity : 0.9247
##             Specificity : 0.9520
##          Pos Pred Value : 0.9507
##          Neg Pred Value : 0.9267
##              Prevalence : 0.5000
##          Detection Rate : 0.4623
##    Detection Prevalence : 0.4863
##       Balanced Accuracy : 0.9383
##
##        'Positive' Class : 0
##

#sensitivity(test$fraud, predicted)
#specificity(test$fraud, predicted)
```

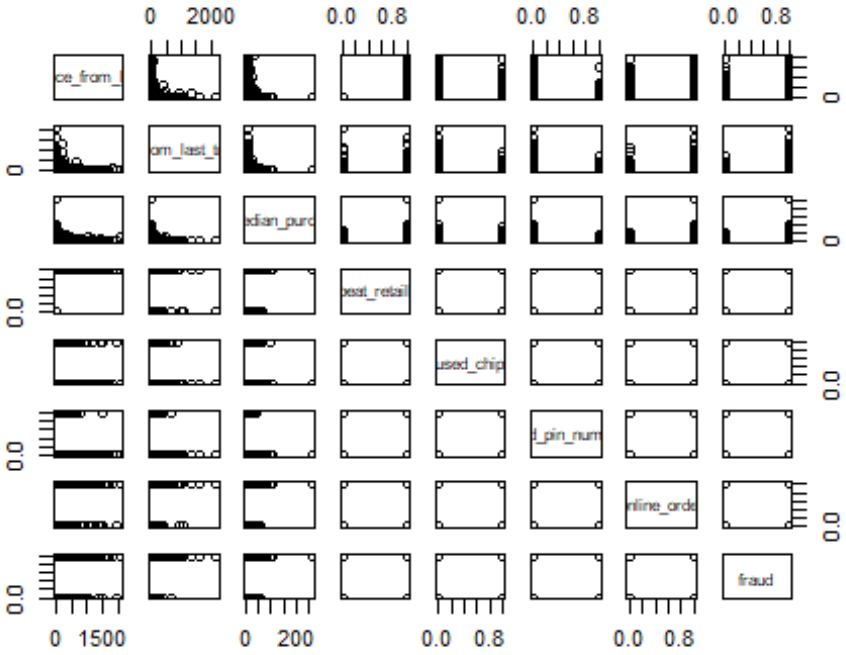
Data Visualization

We run pairs plot to see if there is relationship between predictors. The correlation between independent predictors is low, so there is no multicollinearity.

```
library(GGally)

## Warning: package 'GGally' was built under R version 4.1.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggpairs(new_data)
```




```
plot(new_data)
```



ROC Curve and AUC

Beta: The beta value of 0.439 indicates the likelihood ratio at a specific point on the ROC curve. It suggests that the odds of a positive prediction are approximately 0.439 times higher than the odds of a negative prediction at that particular threshold.

Sensitivity: The sensitivity of 96.4% indicates the proportion of true positive predictions correctly identified by the model. It suggests that the model has a high ability to correctly classify fraud cases.

Specificity: The specificity of 91.5% indicates the proportion of true negative predictions correctly identified by the model. It suggests that the model has a high ability to correctly classify non fraud cases.

PV+ (Positive Predictive Value): The PV+ of 3.8% represents the proportion of true positive predictions out of all positive predictions made by the model. It indicates that when the model predicts fraud, there is approximately a 3.8% chance that the prediction is correct.

PV- (Negative Predictive Value): The PV- of 8.1% represents the proportion of true negative predictions out of all negative predictions made by the model. It indicates that when the model predicts non fraud, there is approximately an 8.1% chance that the prediction is correct.

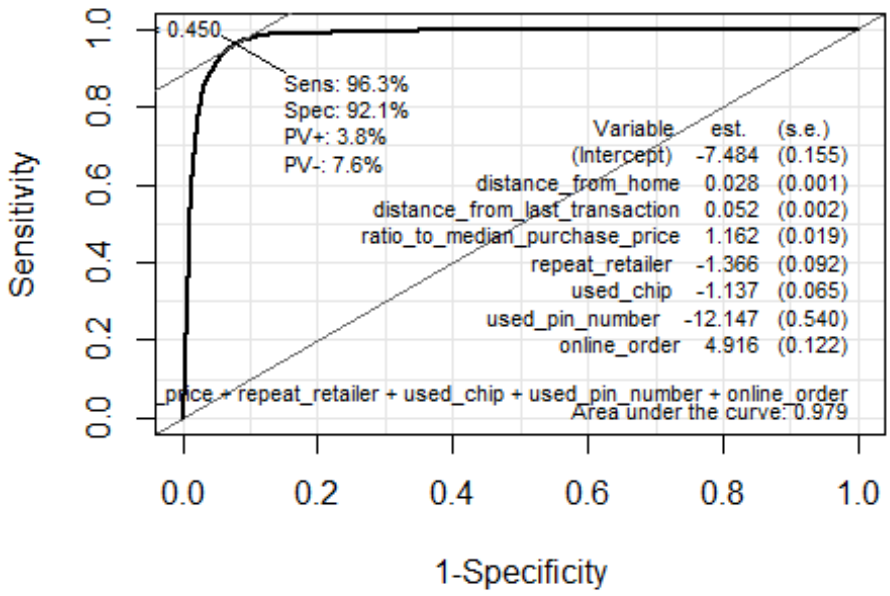
AUC (Area Under the Curve): The AUC of 0.978 indicates the overall performance of the model. The AUC ranges from 0 to 1, with a higher value indicating better discrimination between the positive and negative classes. An AUC of 0.978 indicates that the model has a

high probability of ranking a randomly chosen positive instance higher than a randomly chosen negative instance. This suggests that the model is effective in making accurate predictions and has strong discriminatory capabilities.

Overall, the results indicate that the model has high sensitivity, specificity, and AUC, which suggests that it is performing well in distinguishing between positive and negative fraud cases. However, the relatively low PV+ and PV- values indicate that there is some room for improvement in accurately predicting positive and negative outcomes. Further evaluation and fine-tuning of the model may be necessary to enhance its predictive accuracy.

```
attach(new_data)
ROC(form=new_data$fraud ~ distance_from_home + distance_from_last_transaction
+ ratio_to_median_purchase_price + repeat_retailer + used_chip +
used_pin_number + online_order, plot="ROC")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```




```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.1.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

predicted <- predict(confusionModel, test, type = "response")
auc(test$fraud, predicted)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## Area under the curve: 0.9772
```

LOOCV and K-fold CV

We could not perform LOOCV or since our data set is too large to practically cross validate.

K fold cross validation results with an accuracy of 0.93675, and kappa value of 0.8735. The accuracy suggests that the model correctly classified approximately 93.675% of the samples. The kappa value is a measure based on the agreement between the predicted and actual responses. A kappa value of 0.8735 suggests a substantial level of agreement beyond chance, indicating good performance and reliability of the model.

```
library(caret)

train_control <- trainControl(method="cv", number=10)
kf_model <- train(fraud ~., data=new_data, trControl=train_control,
method="glm")

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to
do
## classification? If so, use a 2 level factor as your outcome column.

print(kf_model)

## Generalized Linear Model
##
## 20000 samples
##     7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 18000, 18000, 18000, 18000, 18000, 18000, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.3825031  0.4290988  0.326302
```

Probit and identity model

The original logistic regression model is the best model. It has the lowest AIC at 8736.2.

```
probit_model <- glm(fraud ~ distance_from_home +
distance_from_last_transaction +
                      ratio_to_median_purchase_price + repeat_retailer +
used_chip +
                      used_pin_number + online_order, data = new_data, family
= binomial(link = "probit"))

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(probit_model)

##
## Call:
## glm(formula = fraud ~ distance_from_home + distance_from_last_transaction
+
##     ratio_to_median_purchase_price + repeat_retailer + used_chip +
##     used_pin_number + online_order, family = binomial(link = "probit"),
##     data = new_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.3026   0.0000   0.3365   4.1437
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.9098477  0.0781355  -50.04   <2e-16 ***
## distance_from_home              0.0127865  0.0002921   43.78   <2e-16 ***
## distance_from_last_transaction  0.0261401  0.0008871   29.46   <2e-16 ***
## ratio_to_median_purchase_price  0.5699151  0.0083044   68.63   <2e-16 ***
## repeat_retailer                -0.7251560  0.0469899  -15.43   <2e-16 ***
## used_chip                      -0.6521130  0.0331407  -19.68   <2e-16 ***
## used_pin_number                -5.6430618  0.2608300  -21.64   <2e-16 ***
## online_order                    2.7514734  0.0640989   42.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27726  on 19999  degrees of freedom
## Residual deviance: 10116  on 19992  degrees of freedom
## AIC: 10132
##
## Number of Fisher Scoring iterations: 25
```

We could not check the model with identity link. Our data could not run with the model.

Conclusion

In conclusion, our analysis demonstrates that the logistic regression model applied to our data is highly suitable for predicting fraud cases. The model's predictors exhibit high significance, as confirmed by the Wald test and chi-square goodness-of-fit test. Furthermore, the ROC curve analysis further reinforces the efficacy of the model in distinguishing between positive and negative instances. These findings collectively support the claim that the logistic regression model is well-suited for accurately predicting fraudulent transactions based on our dataset.