NFBT Credit Card Analytics Project

MKTG 6620 PROJECT 2

ADESH VALECHA, SUSHRUTI ACHARYA

BUSINESS PROBLEM

New Federal Bank and Trust Company (NFBT) is providing two types of accounts i.e. checking and savings accounts. The bank offers debit cards and credit card to its customers. The Bank's goal is to expand their credit card business. Customers are using the debit card routinely issued by the NFBT, but these same customers are not applying for the credit card offered by NFBT.

We have been hired to use a data analytics approach to enhance the credit card business. We have a dataset of 1200 customer observations which contains information about NFBT's existing debit card users and whether they had opted for the credit card.

DEFINE THE METHOD

We started by building the logistic regression on all the observations to determine what impacts "offer_success", i.e. which variable influence that customer will accept a credit card from NFBT. Data set provided is standardized, which means that the variables have been centered on the means. As a result, we do not know the true means of each variable. We observed that none of the variables are correlated so we are not removing any variable from our analysis. We have used glm on our standardized data set with "offer_success" as the outcome (dependent) variable and the other variables as predictor (independent) variables. The reason we are using logistic regression is because the outcome is binary that is the customer will either accept a credit card or they will not.

We suspect that there may be some unique customer profiles in our data set, possibly some outliers so we want to run an analysis to detect these outliers. If we can identify outliers and separate them from the inliers, running our logistic regression on both data sets will help us make recommendations to NFBT so the bank can approach differently to each customer base to increase their credit card enrollment.

To achieve the goal of separating outliers, we are using Fast MCD, which is a very effective algorithm used to detect outliers. We have number of observation n=1200 and subsample size h=1006. We need to calculate the determinant of potential ${}^{1200}C_{1006}$=1.162785243E+229 combinations. Using the conventional MCD algorithm for such a large value is quite hard, so we used Fast MCD algorithm.

After identifying outliers, we are splitting the data into two data sets: Inliers and Outliers. After splitting the dataset, we ran the logistic regression on each data set. We included all the variables i.e. amount spend, years, transaction frequency, credit score for the analysis. We assume that a customer accepting the credit card offer is synonymous with them applying for the credit card.

RESULTS, CONCLUSIONS & RECOMMENDATIONS

We have included all the variables of the data set for our analysis to detect the outlier and their influence on our whole analysis. Including every variable will also include any data exception in the data set.

The results of our logistic regression on the entire data set (n=1200) is shown below. The output conveys that the amount spent by a customer and the years they have been with NFBT are highly significant when determining if a customer will apply for a credit card. Both of these relationships are also positive, meaning that the more a customer spends in a month, and the longer they have been with NFBT, the more likely they are to accept an NFBT credit card.

We also observe that credit score is not significant and transaction frequency is less significant and have a negative impact on whether or not the customer will apply for a credit card. As transaction frequency increases, log odds of offer success decreases.

```
Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)         -5.32436    0.38028 -14.001 < 0.0000000000000002 ***
transaction_frequency -0.18213    0.09082  -2.005               0.0449 *
amount_spent         0.84326    0.10134   8.321 < 0.0000000000000002 ***
credit_score        -0.08852    0.05985  -1.479               0.1391
years                1.18957    0.18014   6.604      0.0000000000401 ***
```

Our recommendations based only on the output of the logistic regression on the full data set would be to target marketing efforts and offers to customers who has high monthly spending's and who are loyal, long-term customers of NFBT and are not using their debit card much for transactions. For example, the bank can set a threshold for years, such as upon reaching 3 years with NFBT they are offered credit card. Additionally, the bank can also set a threshold for amount spent with the incentive that by reaching a specific monthly spending amount, customers will be offered credit card with NFBT.

After we ran outlier analysis using Fast MCD, we grouped our data into two sets: one of the outliers, and one of the inliers. We then performed the same logistic regression analysis on each of these data sets to make observations about how the variables impact the outcome for outliers vs. inliers.

For the outliers (194 observations), our results are somewhat similar to those of when we observed the data set as a whole. Here all the variables are statistically significant. Years and amount spent have a positive impact on offer success, and the variable 'year' has the largest impact. Transaction frequency and credit score both have a negative impact on offer success.

Outlier Coefficients:

```
Coefficients: (1 not defined because of singularities)
                    Estimate Std. Error z value   Pr(>|z|)
(Intercept)         -3.67416    0.89082  -4.124 0.00003716 ***
x.mcd.mcd.wt              NA         NA      NA         NA
transaction_frequency -0.22080    0.09271  -2.382    0.01723 *
amount_spent         0.59584    0.18699   3.187    0.00144 **
credit_score        -0.12797    0.06022  -2.125    0.03359 *
years                0.93080    0.20739   4.488 0.00000718 ***
```

Our recommendation from these results would be similar as what we recommended in the original logistic regression: focus marketing efforts and offers on customers with high monthly spending

and loyalty to NFBT. Additionally, the bank should take into consideration the credit score of the customer while offering credit card. As the credit score increases the chances of customer accepting the credit card decreases.

For the inliers (1006 observations, output shown below), our results were quite different from the output with all the data and from the output of just the outliers. One main difference is that all the variables showed a positive impact on the dependent variable, offer success. While amount spent and years are still impactful and significant, their magnitude of impact is larger than the other two outputs. Concisely, the significance of every variable have changed except 'years', It still remains very significant. Unlike, 'years' the significance of variable 'amount_spent' have declined. Here we also observer that variables transaction frequency and credit score are not significant.

Inlier Coefficients:

```
Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)           -6.2556     0.7110  -8.798 < 0.0000000000000002 ***
x.mcd.mcd.wt               NA         NA      NA                   NA
transaction_frequency  0.1291     0.3328   0.388              0.69798
amount_spent           0.8464     0.3067   2.760              0.00578 **
credit_score           0.4717     0.2848   1.656              0.09769 .
years                  1.6447     0.3870   4.250            0.0000214 ***
```

Our recommendation for the bank NFBT with the inlier data set would be to continue offering credit cards to existing customers who have been with the bank for some time period, and who spend a high monthly amount. Secondly, amount_spent also has significant impact on the independent variable i.e. offer_success when running the regression on whole data set, but when running the regression with outlier and inlier data set, the significance of the amount spent in particular month declines. So, variable 'years' should be on higher priority than the variable 'amount_spent' while considering the customer to offer the credit card. Lastly, number of transaction and credit score are not statistically significant so these variables can be avoided.

We can interpret that, while the outliers make up only 16% of the data, they are impacting the magnitude and direction of influence each of the variables have on offer success. We know this because the logistic output for all the observations was more similar to that of the outliers than of the inliers. By separating the two data sets we can make more accurate conclusions which will help drive our business decisions for the majority of our customers, the inliers.

We also analyzed statistical output of the two data sets (shown below) to identify differences between the data sets and understand more about the customers. From the below output we see that Outliers have a higher success rate of accepting the credit card offer. We can now recommend that NFBT target customers with a profile similar to that of the Outlier data set.

```
> summary(data_Inlier)
  x.mcd.mcd.wt offer_success     transaction_frequency  amount_spent        credit_score          years
 Min.    :1   Min.    :0.00000   Min.    :-3.003659   Min.    :-3.325844   Min.    :-3.136567   Min.    :-3.01795
 1st Qu.:1   1st Qu.:0.00000   1st Qu.:-0.609368   1st Qu.:-0.683398   1st Qu.:-0.631041   1st Qu.:-0.63891
 Median :1   Median :0.00000   Median : 0.007403   Median : 0.004294   Median :-0.009934   Median : 0.07614
 Mean   :1   Mean   :0.01193   Mean   : 0.032923   Mean   : 0.025087   Mean   : 0.040284   Mean   : 0.03573
 3rd Qu.:1   3rd Qu.:0.00000   3rd Qu.: 0.651001   3rd Qu.: 0.730949   3rd Qu.: 0.735896   3rd Qu.: 0.69393
 Max.   :1   Max.   :1.00000   Max.   : 2.894185   Max.   : 3.398854   Max.   : 2.991570   Max.   : 2.95857
> summary(data_Outlier)
  x.mcd.mcd.wt offer_success   transaction_frequency  amount_spent     credit_score        years
 Min.    :0   Min.    :0.000   Min.    :-3.5917      Min.    :-3.170   Min.    :-4.448   Min.    :-1.3738
 1st Qu.:0   1st Qu.:0.000   1st Qu.: 0.7551      1st Qu.: 3.423   1st Qu.: 1.175   1st Qu.: 0.4162
 Median :0   Median :0.000   Median : 2.0190      Median : 4.006   Median : 3.395   Median : 1.0597
 Mean   :0   Mean   :0.299   Mean   : 1.9521      Mean   : 3.961   Mean   : 3.137   Mean   : 1.0878
 3rd Qu.:0   3rd Qu.:1.000   3rd Qu.: 3.1902      3rd Qu.: 4.688   3rd Qu.: 5.229   3rd Qu.: 1.7863
 Max.   :0   Max.   :1.000   Max.   : 7.0514      Max.   : 6.825   Max.   :10.363   Max.   : 3.7364
```

This supports our recommendations that NFBT should set a threshold for time the customer is with the bank and amount spent before offering a credit card. Targeting customers with these characteristics is more likely to enhance the credit card business at NFBT.

In conclusion, our customer group “outliers”, which form a small percentage of the total customer base, are more likely to apply for a credit card than the inliers. They also have a great customer profile which would be good for NFBT to target, with high spending and great credit scores compared to the other 84% of customers. We recommend NFBT offer credit cards to customers with high spending, and a sufficient amount of time spent being an existing account holder. This approach will maximize the enhancement efforts of their credit card business and will yield better results.

APPENDIX (code):

```
library(caret)
library(e1071)
library(kernlab)
library(dplyr)
library(arm)
library(robustbase)
library(chemometrics)

data<-read.csv(url("http://data.mishra.us/files/project-2.csv"))
View(data)
str(data)
summary(data)

#Correlation
colnames(data)
correlationMatrix <- cor(data[,c(2,3,4,5)])
highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.7)
print(highlyCorrelated)
#There is no correlation among variables

#Running Logistic regression on whole dataset
summary(model<-train(offer_success~.,
                     data=data,
                     method="glm",
                     family="binomial"))

coef(model$finalModel)

#From the model summary we can say the following:
```

```
#variable credit_score is not significant as p value is greater than 0.05.

#Variables amount_spent and years are highly significant.

#variable transaction_frequency is significant and has negative coefficient that decreases the probability of offer_success.

#In terms of probability

options(scipen=999)

invlogit(coef(model$finalModel))



#outlier detection


#Removing outcome variable (offer_success) from the dataset

data_new <- data[-1]


#Calculating the Minimum Covariance Determinant(MCD)

x.mcd=covMcd(data_new, alpha=.5)


#Estimator via the Fast MCD. Here alpha determines the values of h.

#Roughly h= sample size*alpha


#Prints robust mahalanobis distance

x.mcd$mah


# prints ‘‘outlyingness’’ of observations. here 0 means outlier and 1 inlier

x.mcd$mcd.wt


#Counting number of outliers

sum(x.mcd$mcd.wt == 0)


#Counting number of Inliers
```

```
sum(x.mcd$mcd.wt == 1)

#Creating a new data set where 0, 1 outlier scores are in the first column
new<-data.frame(x.mcd$mcd.wt,data)
new

#Creating a new dataset which excludes the outliers i.e. it is the clean data set.
data_Inlier <- new[-which(x.mcd$mcd.wt == 0), ]

#Create a new dataset which is the outliers i.e. it is the outlier data set.
data_Outlier <- new[-which(x.mcd$mcd.wt == 1), ]


summary(data_Inlier)
summary(data_Outlier)

#Running logistic regression on the new dataset - Inlier
summary(model<-train(offer_success~.,
                     data=data_Inlier,
                     method="glm",
                     family="binomial"))



#Inlier Data Analysis
#Variables amount_spent and years are significant and have positive impact on offer_success.
#Variables credit_score and transacion_frequency have p values greater than 0.05 and are not
significant.
```

```
#Running logistic regression on the new dataset - Outlier
summary(model<-train(offer_success~.,
                     data=data_Outlier,
                     method="glm",
                     family="binomial"))

#Outlier Data Analysis
#All variables are significant.
#Variables amount_spent and years have positive impact on offer_success.
#Variables transaction_frequency and credit_score have negative impact on offer_success.
```