Credit Approval Analysis

Ryan Kuhn

May 1, 2015

Background

Objective:
To demonstrate the analytical techniques taught during the Special Topics in Audit Analytics course at Rutgers University

`**Data:**`
- Dataset- Credit Screening on Credit Card Applications
- Source- UCI Machine Learning Repository

Assumptions:
Field names and values changed to meaningless values. Made assumtions about what attributes the data represents.

Research Question - 1

Q: Is there a relationship between Age, Income, Credit Score, and Debt levels and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?

Research Question - 1

Q: Is there a relationship between Age, Income, Credit Score, and Debt levels and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?

A: Relationships exists between Prior default, Years employed, Credit score, and Income level. These variables are reasonable management strategies.

Research Question - 2

Q: Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the applicant's ethnicity. Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination?

H0: Ethnicity and approval are independent.
H1: Approval status is associated with the ethnicity and credit card company has a compliance risk.

Research Question - 2

Q: Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the applicant's ethnicity. Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination?

> *H0: Ethnicity and approval are independent.*
> *H1: Approval status is dependent on the ethnicity and credit card company has a compliance risk.*

A: A chi-squared test did not give evidence that ethnicity and approval status are dependent. We cannot reject the null hypothesis.

Analytic Methods Used

Methods used:

- Linear regression
- Descriptive Statistics and Normalization
- Association Rules
- Logistic regression
- Classification and Regression Tree
- Ensembling

Linear Regression

Used to fill in missing values in Age

```
#Download the data
myURL<- "http://archive.ics.uci.edu/ml/machine-learning-dat
if(!file.exists("Dataset.csv")){
    download.file(myURL,"Dataset.csv")
    }
rm(myURL)

# Load the data
Cols<- c(rep("character",2),"numeric",rep("character",4),"n
         "numeric",rep("character",3),"numeric","character'
Data<-read.csv("Dataset.csv",sep=",",colClasses=Cols)
rm(Cols)

# Give column names by letter
# names(Data)<- LETTERS[1:16]
names(Data)<-c("Male","Age","Debt","Married","BankCustomer'
```

Descriptive Stats and Normalization

Convert values to Z-Scores

```
#Convert to z score
SD.Age<-round(sd(Numeric$Age, na.rm=T),4)
Data$AgeNorm<- (Data$Age-mean(Data$Age, na.rm=T))/SD.Age
rm(SD.Age, Mean.Age)

# View the distribution
par(mfrow=c(1,2), oma=c(0,0,.75,0))
hist(Data$Age,main=NULL,xlab="Age",col="blue")
hist(Data$AgeNorm,main=NULL,xlab="AgeNorm",ylab=NULL,col="g
title("Distribution of Values Before and After Normalizatio
```

```
ggplot(Data) +
    aes(Approved,AgeNorm) +
    geom_boxplot(outlier.colour="red") +
    theme_bw() +
    coord_flip() +
    labs(title="Distribution of AgeNorm by Credit Approval
```

Association Rules

```
Data$Married<-ifelse(is.na(Data$Married),"u",Data$Married)
Data$BankCustomer<-ifelse(is.na(Data$BankCustomer),"g",Data
Data$Ethnicity<-ifelse(is.na(Data$Ethnicity),"v",Data$Ethni
Data$EducationLevel<-ifelse(is.na(Data$EducationLevel),"c",
Data$ZipCode<-ifelse(is.na(Data$ZipCode),"00000",Data$ZipCo
Data$Male<-ifelse(is.na(Data$Male),"b",Data$Male)
```

```
# Convert categorical variables to factors
Data[,1:10]<- lapply(Data[1:10],function(x) factor(x))

Data$Ethnicity<-relevel(Data$Ethnicity,"v")
```

```
# Generate rules
Rules<- apriori(Data[!incomplete,1:10],
                parameter=list(supp=0.1,
                               conf=0.75,
                               target='rules'))
```

Baseline Model

- Simple mean of results.
- Establish benchmark to measure model accuracy
-

```
Train %>%
    summarise(Total=n(), Success=sum(Approved==1)) %>%
    mutate(Percent=round(Success/Total*100,2))
```

Logistic Regression

```
set.seed(1234)

split<- sample.split(Data$Approved, SplitRatio=0.75)
Train<- subset(Data,split==TRUE)
Test <- subset(Data, split==FALSE)

rm(split)
```

```
LogFit<- glm(Approved~AgeNorm+DebtLog+YearsEmployedLog+Cred
summary(LogFit)
```

```
LogPred<- predict(LogFit,newdata=Train, type="response")
table(Train$Approved, LogPred>0.5)
```

CART Model

```
set.seed(1234)
TreeFit<-rpart(Approved~Male+Married+BankCustomer+Educatio
               data=Train,
               method="class",
               control=rpart.control(xval=10,cp=0.025))

TreeFit

prp(TreeFit,main="CART model", digits=6,
    extra=1,
    branch.col="blue",
    type=4,
    leaf.round=2,
    box.col=c("pink","palegreen")[TreeFit$frame$yval],
    ycompact=T)
```

References

Data:
`http://archive.ics.uci.edu/ml/machine-learning-databases/credit-screening`

Analytic Report:
`http://www.rpubs.com/kuhnrl30/CreditScreen`