Loan Payment Prediction using Classification Models

Shahrar Nizam

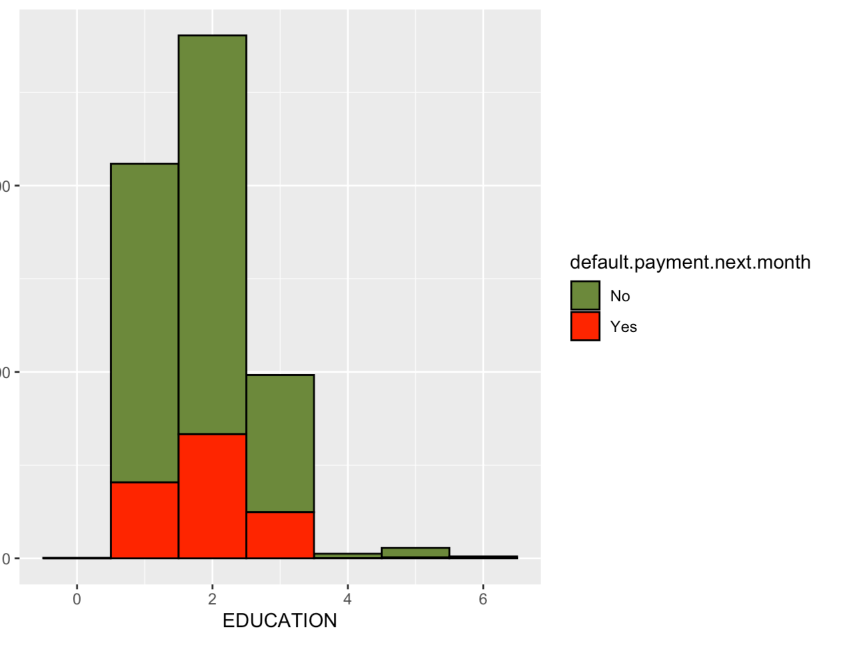
A. Data Preparation

A.1. Data Loading and Initial Transformation

```r
library(ggplot2)
cc <- read.csv("UCI_Credit_Card.csv")
cc$default.payment.next.month <- factor(cc$default.payment.next.month,levels=c(0,1), labels=c("No","Yes"))
```

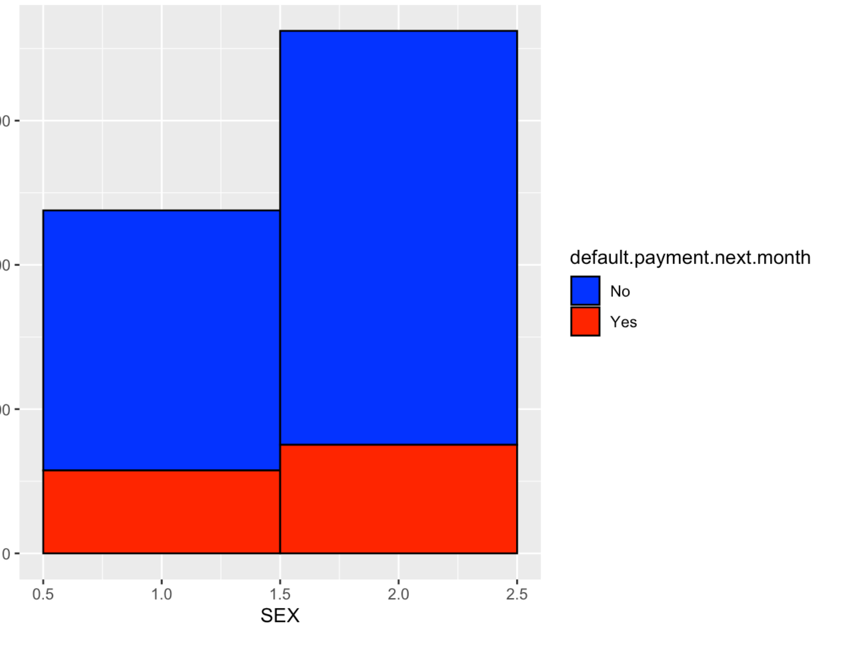
A.2. Demographic Variables

```r
ggplot(cc, aes(x=EDUCATION, fill=default.payment.next.month, color=default.payment.next.month)) +
  geom_histogram(binwidth=1, position="stack") +
  scale_color_manual(values=c("black","black")) +
  scale_fill_manual(values=c("darkolivegreen4", "red"))
```



From the above histogram it is clear that that most participants in the dataset belong to the 2 critera which is University graduates. But it cannot be assumed that they will default the most on payment because while defaulters for other educations are lower, the number of data collected with educations are also lower. Therefor the ratio of defaulters to non-defaulters might be the same. To reach a conclusion on the relationship between the numer of defaulters and education, equal number of data must be collected for all education levels.

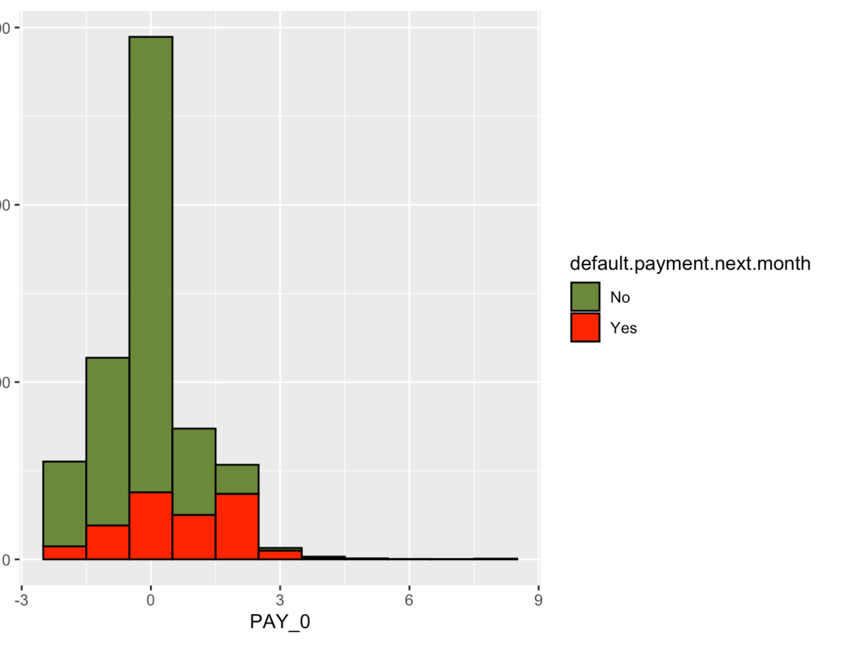
```r
ggplot(cc, aes(x=SEX, fill=default.payment.next.month, color=default.payment.next.month)) +
  geom_histogram(binwidth=1, position="stack") +
  scale_color_manual(values=c("black","black")) +
  scale_fill_manual(values=c("blue", "red"))
```



From the above histogram it can be determined that males (1) default more than females. Because even though the number of female defaulters is higher, the number of females in the dataset is also significantly higher. The ratio of female defaulters to non-defaulters is much lower than the ratio of male defaulters to non-defaulters. Therefore it can be assumed that males will default more than females.

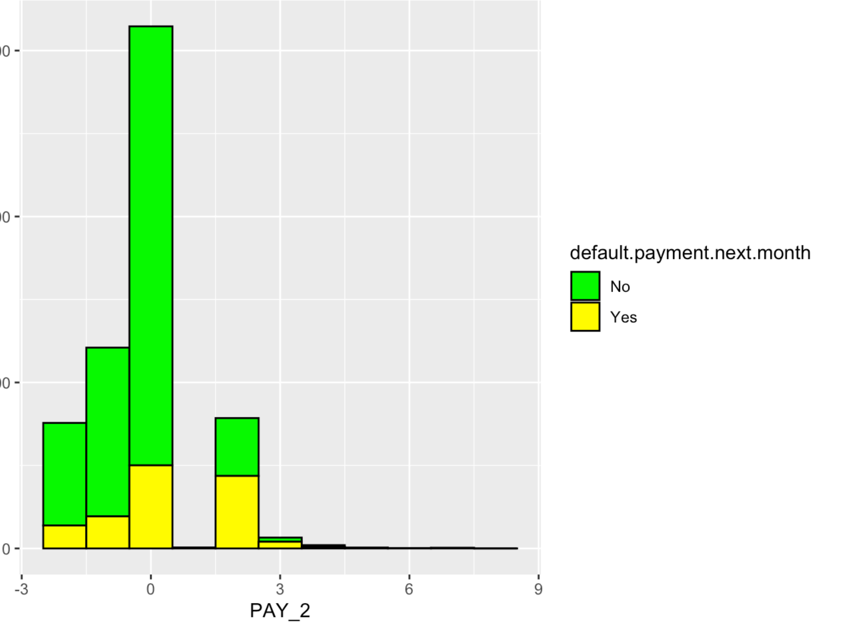
A.3. Payment Status Variables

```r
ggplot(cc, aes(x=PAY_0, fill=default.payment.next.month, color=default.payment.next.month)) +
  geom_histogram(binwidth=1, position="stack") +
  scale_color_manual(values=c("black","black")) +
  scale_fill_manual(values=c("darkolivegreen4", "red"))
```



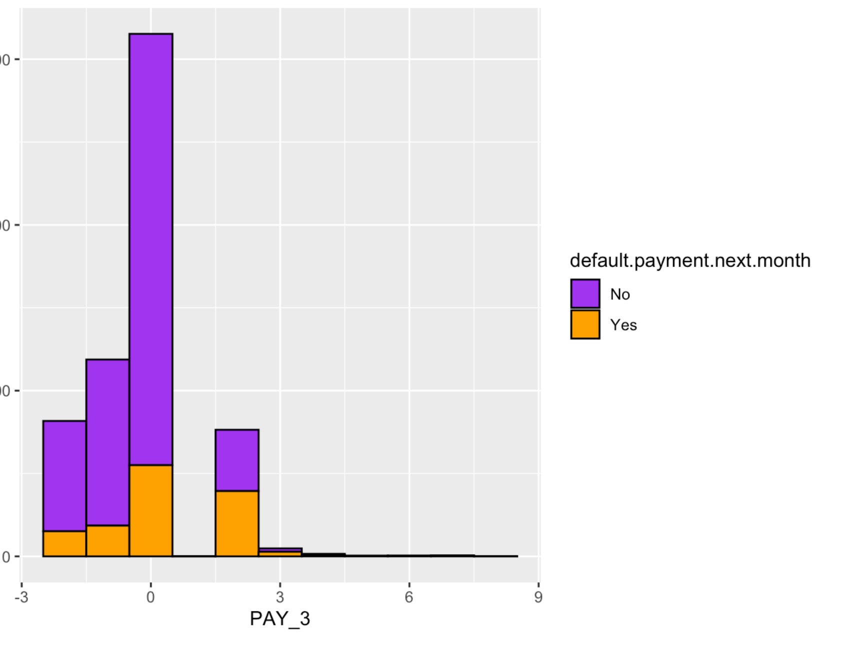
From the aove plot it can be predicted that the longer people delay the payment the more chances of defaulting next months payment. Even though the people with revolving credit (0) have a higher spike of default than people wit one month delay on payment, the number of participants with revolving credit is also higher. Therefore it would be wrong to assume that they will default payment more than people with delayed payments.

```r
ggplot(cc, aes(x=PAY_2, fill=default.payment.next.month, color=default.payment.next.month)) +
  geom_histogram(binwidth=1, position="stack") +
  scale_color_manual(values=c("black","black")) +
  scale_fill_manual(values=c("green", "yellow"))
```



The histogram shows that the data is inconsistent. Because in Pay_0 (September) we see that there are people with 2 months delay in their bills which means that in Pay_2(August) they are supposed to have 1 month delay but we see that the number for 1 month delay is significantly lower than 2 month delay on PAY_0. I do not think this can help predict the future payment.

```r
ggplot(cc, aes(x=PAY_3, fill=default.payment.next.month, color=default.payment.next.month)) +
  geom_histogram(binwidth=1, position="stack") +
  scale_color_manual(values=c("black","black")) +
  scale_fill_manual(values=c("purple", "orange"))
```



We see a similar situation with PAY_3 data.

A.4. Transforming Nominal Variables

```r
cc$MARRIAGE <- factor(cc$MARRIAGE,levels=c(1,2,3),labels= c("Married", "Single", "Others"))
cc$SEX <- factor(cc$SEX,levels=c(1,2),labels= c("Male", "Female"))
cc$EDUCATION <- factor(cc$EDUCATION,levels=c(1,2,3,4,5,6),labels= c("Graduate", "University", "High-School", "Oth
ers", "Unknown", "Unknown"))
cc$PAY_0 <- factor(cc$PAY_0,levels=c(-2,-1,0,1,2,3,4,5,6,7,8),labels= c("No-Consumption", "Paid-in-full", "Revolv
ing-Credit", "1-Month-delay", "2-month-delay", "3-month-delay", "4-month-delay", "5-month-delay", "6-month-delay"
, "7-month-delay", "8-month-delay"))
cc$PAY_2 <- factor(cc$PAY_2,levels=c(-2,-1,0,1,2,3,4,5,6,7,8),labels= c("No-Consumption", "Paid-in-full", "Revolv
ing-Credit", "1-Month-delay", "2-month-delay", "3-month-delay", "4-month-delay", "5-month-delay", "6-month-delay"
, "7-month-delay", "8-month-delay"))
cc$PAY_3 <- factor(cc$PAY_3,levels=c(-2,-1,0,1,2,3,4,5,6,7,8),labels= c("No-Consumption", "Paid-in-full", "Revolv
ing-Credit", "1-Month-delay", "2-month-delay", "3-month-delay", "4-month-delay", "5-month-delay", "6-month-delay"
, "7-month-delay", "8-month-delay"))
```

A.5. Selection of Training Data

```r
train <- cc[sample(nrow(cc), 5000), ]
```

A.6. Selection of Testing Data

```r
test <- cc[c(20,20000),]
test
```

```
##          ID LIMIT_BAL    SEX  EDUCATION MARRIAGE AGE         PAY_0
## 20       20    180000 Female   Graduate   Single  29 1-Month-delay
## 20000 20000    240000 Female University  Married  37  Paid-in-full
##                PAY_2          PAY_3 PAY_4 PAY_5 PAY_6 BILL_AMT1 BILL_AMT2
## 20    No-Consumption No-Consumption    -2    -2    -2         0         0
## 20000  2-month-delay   Paid-in-full    -1    -1     0      1769       842
##       BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2 PAY_AMT3
## 20            0         0         0         0        0        0        0
## 20000     14015         0      1317       566        0    14015        0
##       PAY_AMT4 PAY_AMT5 PAY_AMT6 default.payment.next.month
## 20           0        0        0                         No
## 20000     1317        0        0                         No
```

B. Data Classification

B.1. Naive Bayes using Demographic Variables

```r
library(e1071)
nbDem <- naiveBayes(default.payment.next.month ~ SEX + EDUCATION + MARRIAGE, train)
nbDem
```

```
## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
## 
## A-priori probabilities:
## Y
##     No    Yes 
## 0.7736 0.2264 
## 
## Conditional probabilities:
##      SEX
## Y          Male    Female
##   No  0.3875388 0.6124612
##   Yes 0.4540636 0.5459364
## 
##      EDUCATION
## Y         Graduate   University  High-School       Others      Unknown
##   No  0.3605794102 0.4614588722 0.1546818417 0.0059493016 0.0173305742
##   Yes 0.3021201413 0.5185512367 0.1740282686 0.0008833922 0.0044169611
## 
##      MARRIAGE
## Y         Married      Single      Others
##   No  0.453133092 0.537804247 0.009062662
##   Yes 0.496466431 0.487632509 0.015901060
```

```r
predict(nbDem, test[1,])
```

```
## [1] No
## Levels: No Yes
```

Although the prediction above is correct, from the above data it seems that while females have a higher chances of paying the bill, they also have a higher chance of not paying the bill. There is a similar prediction with education and we see that University student have the highest probability of both paying and not paying. Only in Marriage data we see that married people have higher chances of defaulting compared to singles. This is most likely because the dataset is biased with higher numbber of female university students and so i do not think this is a good model for prediction for our dataset. Maybe we should use different attributes.

B.2. Naive Bayes using Payment Status

```r
nbPay <- naiveBayes(default.payment.next.month ~ PAY_0 + PAY_2 + PAY_3, train)
nbPay
```

```
## 
## Naive Bayes Classifier for Discrete Predictors
## 
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
## 
## A-priori probabilities:
## Y
##     No    Yes 
## 0.7736 0.2264 
## 
## Conditional probabilities:
##      PAY_0
## Y     No-Consumption Paid-in-full Revolving-Credit 1-Month-delay 2-month-delay
##   No    0.1010858325 0.2063081696     0.5498965874  0.1036711479  0.0341261634
##   Yes   0.0538869258 0.1404593640     0.2906360424  0.1969964664  0.2641342756
##      PAY_0
## Y     3-month-delay 4-month-delay 5-month-delay 6-month-delay 7-month-delay
##   No   0.0036194416  0.0005170631  0.0007755946  0.0000000000  0.0000000000
##   Yes  0.0459363958  0.0053003534  0.0008833922  0.0017667845  0.0000000000
##      PAY_0
## Y     8-month-delay
##   No   0.0000000000
##   Yes  0.0000000000
## 
##      PAY_2
## Y     No-Consumption Paid-in-full Revolving-Credit 1-Month-delay 2-month-delay
##   No    0.1336608066 0.2233712513     0.5659255429  0.0002585315  0.0718717684
##   Yes   0.1033568905 0.1457597173     0.3772084806  0.0008833922  0.3339222615
##      PAY_2
## Y     3-month-delay 4-month-delay 5-month-delay 6-month-delay 7-month-delay
##   No   0.0033609100  0.0015511892  0.0000000000  0.0000000000  0.0000000000
##   Yes  0.0273851590  0.0079505300  0.0017667845  0.0017667845  0.0000000000
##      PAY_2
## Y     8-month-delay
##   No   0.0000000000
##   Yes  0.0000000000
## 
##      PAY_3
## Y     No-Consumption Paid-in-full Revolving-Credit 1-Month-delay 2-month-delay
##   No    0.1452947260 0.2140641158     0.5529989659  0.0000000000  0.0824715615
##   Yes   0.1227915194 0.1342756184     0.4231448763  0.0000000000  0.2835689046
##      PAY_3
## Y     3-month-delay 4-month-delay 5-month-delay 6-month-delay 7-month-delay
##   No   0.0041365047  0.0005170631  0.0002585315  0.0000000000  0.0000000000
##   Yes  0.0220848057  0.0070671378  0.0035335689  0.0000000000  0.0035335689
##      PAY_3
## Y     8-month-delay
##   No   0.0002585315
##   Yes  0.0000000000
```

```r
predict(nbPay, test[1,])
```

```
## [1] No
## Levels: No Yes
```

B.3. Smoothed Naive Bayes using Payment Status

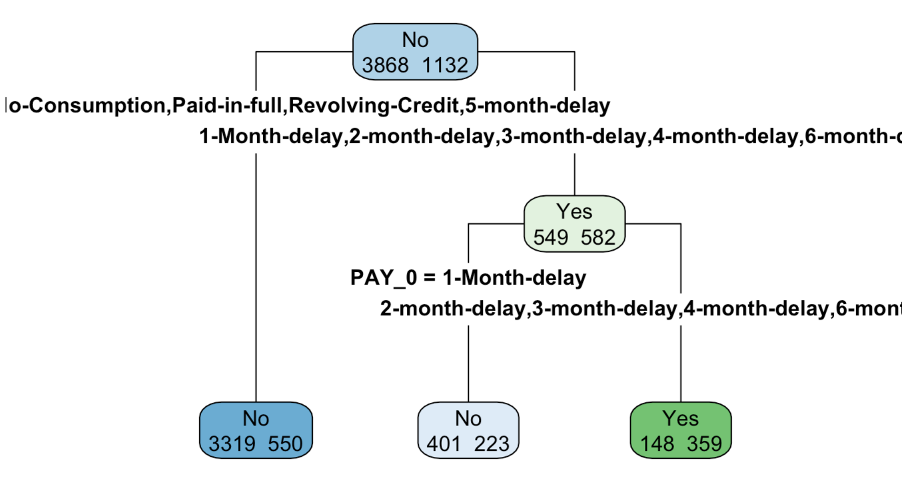
```r
nbPay <- naiveBayes(default.payment.next.month ~ PAY_0 + PAY_2 + PAY_3, train, laplace=1.5)
predict(nbPay, test[1,])
```

```
## [1] No
## Levels: No Yes
```

C. Classification with Decision Tree

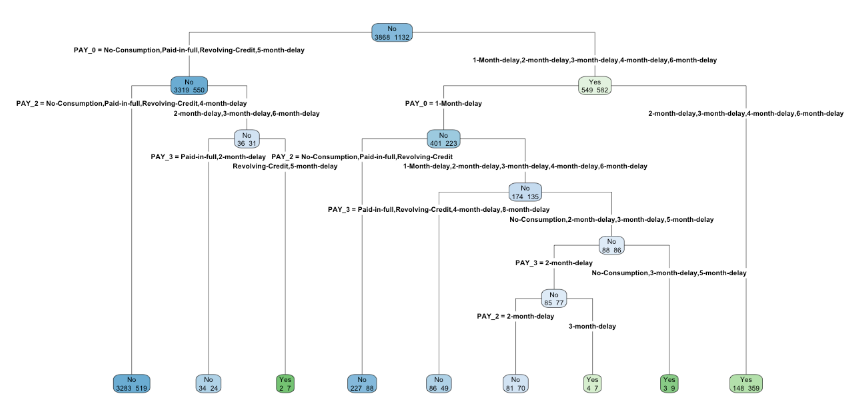
C.1. Basic Decision Tree

```r
library("rpart")
library("rpart.plot")
dtPay <- rpart(default.payment.next.month ~ PAY_0 + PAY_2 + PAY_3,
            method="class",
            data=train, parms=list(split='information'),
            minsplit=20, cp=0.02)
rpart.plot(dtPay, type=4, extra=1)
```



C.2. Decision Tree with a Different Complexity Parameter

```r
dtPay <- rpart(default.payment.next.month ~ PAY_0 + PAY_2 + PAY_3,
            method="class",
            data=train, parms=list(split='information'),
            minsplit=20, cp=0.001)
rpart.plot(dtPay, type=4, extra=1)
```



D. Conclusion

The decision tree model has performed better for the purpose of our analysis than Naive Bayes. This is because decision tree took into consideration dependent variables wheras naive bbayes assumes all variables are independent which in our case is not because PAY_0,PAY_2 and PAY_3 are dependent variables. Through the different entropies in the decsion tree we can see thatpeople who delay between 2-4 months in payments are most likely to default on next months payment and people with no-consumption, revolving credit, 1- month delay are less liekly. This is a reasonable prediction but what does not make sense is why people with 5, 6 and 8 month delays are predicted to not delay. Maybe they are data anamolies. Going forward i would discuss with the the business stakeholders to understand how they curated the data to label payments because some things in the dataset does not make sense. For example how can how can someone with PAY3..6 = -1 have PAY2 = 2? In other words, if the person duly paid their bill every month, how can they suddenly have 2 months delay the next month?