~~~
---
title: "Classification"
author: "Aidan Case"
output: html_notebook
---
Data was taken from [kaggle](https://www.kaggle.com/datasets/sanskar457/
fraud-transaction-detection).



Load data.

```{r}
ds <- read.csv("Final Transactions.csv", header=TRUE, nrows=1e6)
ds$TX_FRAUD <- factor(ds$TX_FRAUD, labels = c("NO_FRAUD", "FRAUD"))
ds$TERMINAL_ID <- factor(ds$TERMINAL_ID)
split <- sample(1:nrow(ds), .8*nrow(ds), replace=FALSE)
train <- ds[split,]
test <- ds[-split,]
```

Initial vis.

```{r}
head(ds)
tail(ds)
summary(ds)
str(ds[c(4:7,9)])
pairs(ds[c(4:7,9)])
```

Plots.

```{r}
par(mfrow=c(1, 2))
plot(ds$TX_FRAUD,ds$TX_AMOUNT)
plot(ds$TX_FRAUD,ds$TX_TIME_SECONDS)
```

```{r}
logm <- glm(train$TX_FRAUD~(train$TX_AMOUNT + train$TX_TIME_SECONDS
                         # impl'ing this factor actually increases
                         # computer usage significantly
                         # + ds$TERMINAL_ID
                         ), data = train, family = binomial)
summary(logm)
```

The null and residual deviance was calculated to be P ~ 0.0000, meaning that
the model has a very high confidence. The deviance is small between each, and
the log err between each (Pr(>|z|)) is small.

```{r}
library(e1071)
nbm <- naiveBayes(train$TX_FRAUD ~ (train$TX_AMOUNT + train$TX_TIME_SECONDS),
data=train)
nbm
```

The majority of the time, fraud does not happen. The average amount of
fraudulent transactions is 1466, while false positives are 2067. Meanwhile,
time taken between transactions seem to be similar.

```{r}
# glm
~~~

```
p <- predict(logm, newdata=test, type="response")
p <- ifelse(p>0.5, 2, 1)
print("")
print(paste("glm acc:", mean(p==as.integer(test$TX_FRAUD))))
# nbm
p <- predict(nbm, newdata=test, type="class")
table(p, test$TX_FRAUD)
print(paste("nbm acc:", mean(p==test$TX_FRAUD)))
```

The accuracy of naive bayes vs logistic regression seems to be that naive bayes has higher accuracy (86.6%) against logistic regression (77.2%), probably due to the amount of data trained on.

The main differences on logistic regression vs naive bayes is that logistic regression is computationally inexpensive, has nice probabilistic output, and finds classes easily that are linearly seperatable, while naive bayes works well with little data, is easy to implement and interpret, and handles higher dimensionality. The downsides of each is that linear regression may underfit, while naive bayes requires that the predictors are independent to reach it's full potential.

The main draw of of the table is that it directly shows the false positives vs the correct predictions. The drawback is that it may be useless or hard to use, as shown by my table produced. Accuracy is a good metric to show, well, how accurate the model is. Used incorrectly, it can show that a model is not overfitting when it actually is. P values show how good the model is at a significance level, though, it's only for the trained data.