Data Preparation

Zachary Levonian

11/02/2018

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(mice)  # for multiple imputation
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
train <- read.csv("../../data/raw/train.csv", stringsAsFactors=FALSE)
test <- read.csv("../../data/raw/test.csv", stringsAsFactors=FALSE)
```

```
test$Survived = 2
df <- rbind(train, test)
```

Build factors from data

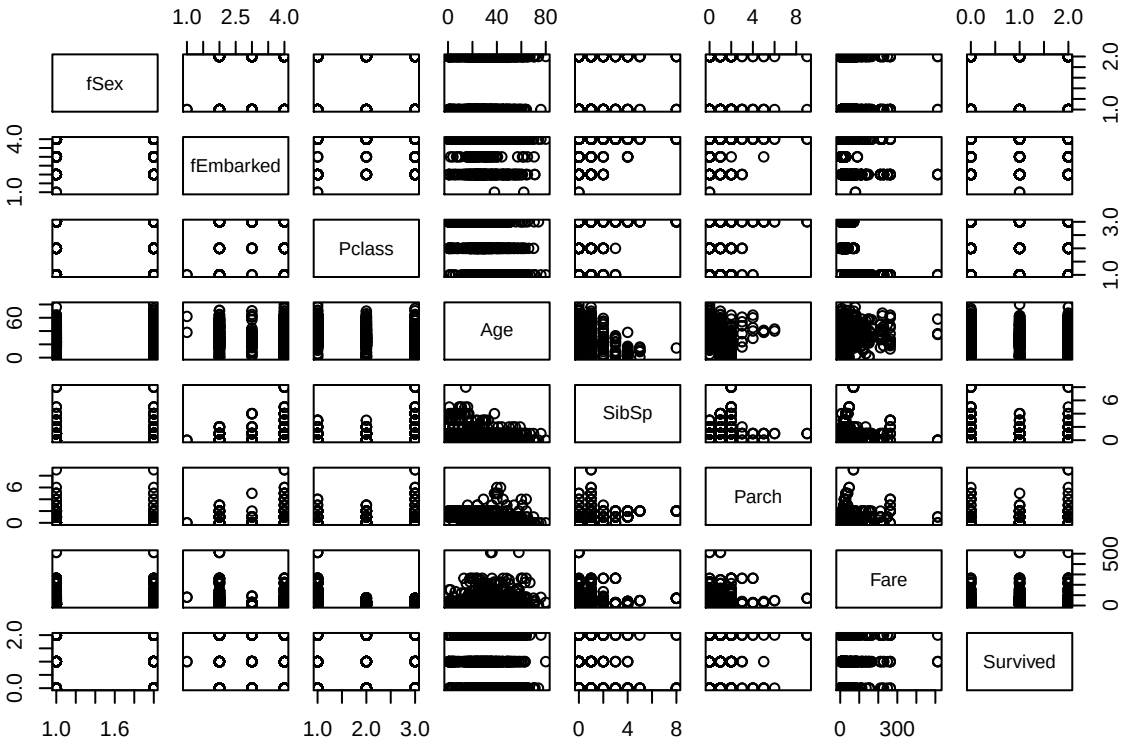
```
df$fSex = factor(df$Sex)
df$fEmbarked = factor(df$Embarked)
```

Now, we save all the columns to be used as potential features to a file.

```
toWrite <- df[,c("PassengerId", "fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Fare", "Survive
write.table(toWrite, file="../../data/derived/factorized_data.csv",
            row.names=FALSE, col.names=TRUE,
            sep=",", quote=FALSE)
```

Data Exploration

```
pairs(df[c("fSex", "fEmbarked", "Pclass", "Age", "SibSp", "Parch", "Fare", "Survived")])
```



Train a Model

The code below demonstrates: - Reading in the features - Splitting the data into the training and test data - Training a regression model - Predicting the test set based on the trained model - Saving the predictions in the Kaggle format for submission

```
df <- read.csv("../../data/derived/factorized_data.csv", stringsAsFactors=TRUE)
```

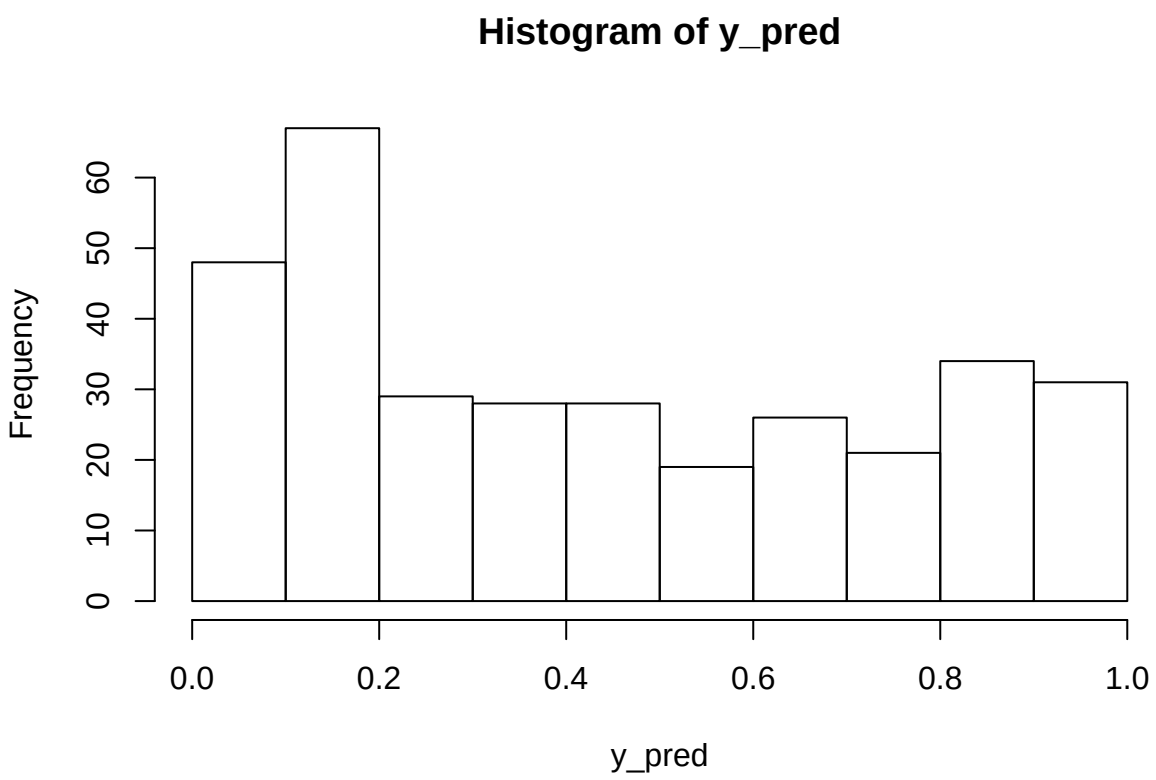
```
train <- df[df$Survived != 2, ]
test <- df[df$Survived == 2, ]
```

```
md <- glm(Survived ~ fSex + fEmbarked + Pclass + Age + SibSp + Parch + Fare, family="binomial", data=tra
summary(md)
```

```
##
## Call:
## glm(formula = Survived ~ fSex + fEmbarked + Pclass + Age + SibSp +
##     Parch + Fare, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7233  -0.6439  -0.3772   0.6288   2.4457
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  17.894850 607.855474   0.029  0.97651
## fSexmale     -2.638476   0.222256 -11.871  < 2e-16 ***
## fEmbarkedC  -12.257443 607.855250  -0.020  0.98391
## fEmbarkedQ  -13.080988 607.855453  -0.022  0.98283
## fEmbarkedS  -12.658656 607.855228  -0.021  0.98339
```

```
## Pclass       -1.199251   0.164619  -7.285 3.22e-13 ***
## Age          -0.043350   0.008232  -5.266 1.39e-07 ***
## SibSp        -0.363208   0.129017  -2.815  0.00487 **
## Parch        -0.060270   0.123900  -0.486  0.62666
## Fare          0.001432   0.002531   0.566  0.57165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.34  on 704  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 652.34
##
## Number of Fisher Scoring iterations: 13
```

```r
# predict on the test set
y_pred <- predict(md, test, type="response")
hist(y_pred)
```




```r
# save the predictions to a file that can be submitted to Kaggle
test$Survived = as.numeric(y_pred > 0.5)
# while there shouldn't be NAs in the test output,
# here we apply a nasty hack to ensure any NAs are numeric in the output
test[is.na(test)] <- 0
toWrite <- test[,c("PassengerId", "Survived")]
write.table(toWrite, file="../../data/derived/testModelPredictions.csv",
            row.names=FALSE, col.names=TRUE,
            sep=",", quote=FALSE)
```