Logistic Regression, Examples from ’Hands On Machine Learning with R,

Notes from Hands On Machine Learning with R, Chapter 5 which covers Logistic Regression models.

```
library(dplyr)     # for data wrangling
library(ggplot2)   # for awesome plotting
library(rsample)   # for data splitting
library(caret)     # for logistic regression modeling
library(vip)       # variable importance
library(modeldata) # has Attrition data set
library(tidyverse) #Tidyverse
```

Logistic Regression Model

A basic logistic regression model, predicting an output (attrition) based on one variable (Monthly Income).

```
#Use the attrition dataset from model date
data("attrition")
df <- attrition %>% mutate_if(is.ordered, factor, ordered = FALSE)

# Create training (70%) and test (30%) sets
set.seed(123)  # for reproducibility
churn_split <- initial_split(df, prop = .7, strata = "Attrition")
churn_train <- training(churn_split)
churn_test  <- testing(churn_split)

#Use glm() to create a Logistic Regression model
model1 <- glm(Attrition ~ MonthlyIncome, family = "binomial", data = churn_train)

#Model results are easier to interpret using tidy()
tidy(model1)
```

```
## # A tibble: 2 x 5
##   term           estimate std.error statistic       p.value
##   <chr>             <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   -0.924    0.155         -5.96 0.00000000259
## 2 MonthlyIncome -0.000130 0.0000264     -4.93 0.000000836
```

Logistic Regression Model, Confidence Intervals

We can say with 95% confidence that monthly income will impact the attrition rates between the 2.5% value and 97.5% value.

```
#Confidence Intervals using Estimated Standard Error
confint(model1)
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %        97.5 %
## (Intercept)   -1.2267754960 -6.180062e-01
## MonthlyIncome -0.0001849796 -8.107634e-05
```

Logistic Regression Model, Model Accuracy

Find classification accuracy by using aret::train() to fit 3, 10 fold cross validated logistic regression models.

```
#Cross Validation for Logistic Regression with 1 variable MonthlyIncome predicting Attrition
set.seed(123)
cv_model1 <- train(
  Attrition ~ MonthlyIncome,
  data = churn_train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 10))

cv_model1
```

```
## Generalized Linear Model
##
## 1030 samples
##    1 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 926, 926, 927, 928, 928, 927, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8388478  0
```

Multiple Logistic Regression Model

Logistic Regression model predicting Attrition based on variables MonthlyIncome and OverTime.

```
#Multiple Logistic Regression
model3 <- glm(
  Attrition ~ MonthlyIncome + OverTime,
  family = "binomial",
  data = churn_train
)
#View the model3 results
tidy(model3)
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -1.43     0.176         -8.11 5.25e-16
## 2 MonthlyIncome -0.000139 0.0000270     -5.15 2.62e- 7
## 3 OverTimeYes    1.47     0.180          8.16 3.43e-16
```

Logistic Regression Model, Model Accuracy

Find classification accuracy by using aret::train() to fit 3, 10 fold cross validated logistic regression models.

```
set.seed(123)
cv_model3 <- train(
  Attrition ~ .,
  data = churn_train,
  method = "glm",
  family = "binomial",
  trControl = trainControl(method = "cv", number = 10)
)

cv_model3
```

```
## Generalized Linear Model
##
## 1030 samples
##   30 predictor
##    2 classes: ’No’, ’Yes’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 926, 926, 927, 928, 928, 927, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8757893  0.476267
```

Comparing Models for Accuracy

Compares the preformance of Model1 to Model3. Most of the time we will run multiple models to determine which model preforms the best.

```
# extract out of sample performance measures
summary(
  resamples(
    list(
      model1 = cv_model1,
      model3 = cv_model3)))$statistics$Accuracy
```

```
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA’s
## model1 0.8349515 0.8349515 0.8365385 0.8388478 0.8431373 0.8446602    0
## model3 0.8365385 0.8495146 0.8792476 0.8757893 0.8907767 0.9313725    0
```