Bob Marshall Wilderness Center

Trent Gillingham

Objectives:

- Introduction
- Analysis
- Results
- Future Steps

Introduction: Why I Chose This Project

- Ecology Dataset
  - USDA.gov: Forestry Research
- Find the right machine learning model with the highest accuracy

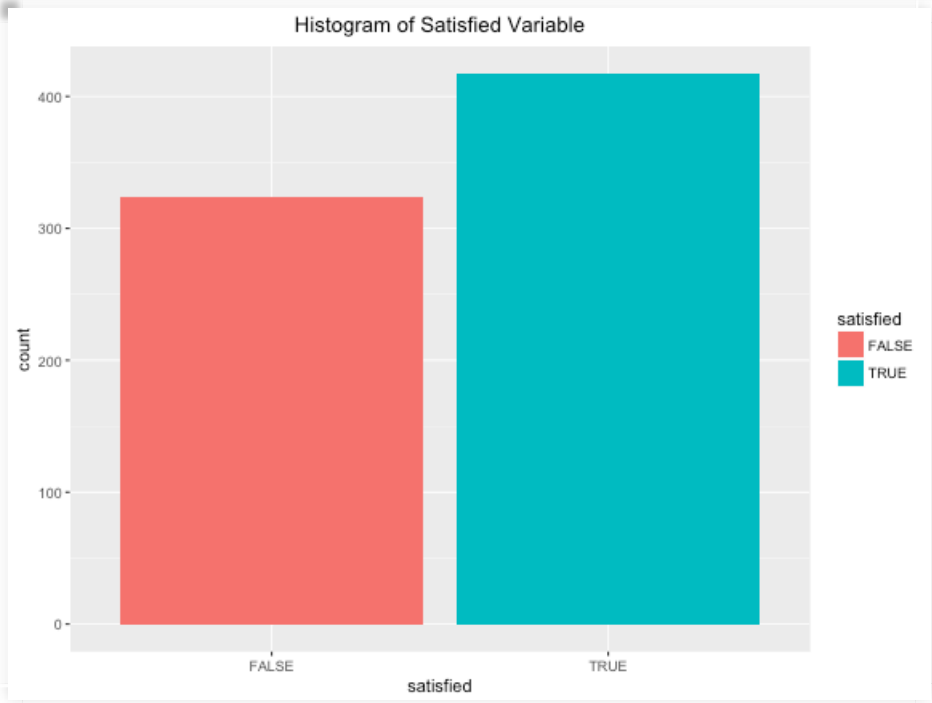
Analysis: Overview of Data

```
library(tidyverse)
BMWC_1982_Data <- read_csv("RDS-2017-0015/Data/BMWC_1982_Data.csv")
Marshall <- BMWC_1982_Data
Marshall %>%
  dplyr::select(c(SATIS, PHOTO, NUMHORS, NATUR, HIKE, TRAVEL, FISH)) %>%
  head()
```

```
# A tibble: 6 x 7
  SATIS PHOTO NUMHORS NATUR  HIKE TRAVEL  FISH
  <int> <int>   <int> <int> <int>  <int> <int>
1     1     1      NA     2     2      1     1
2     1     2      99     2     2      3     1
3     2     1      NA     2     2      1     1
4     1     2       8     1     2      5     2
5     3     1       4     1     1      4     2
6     2     1      NA     1     2      1     1
```

Analysis: Output Variable

```
Marshall%<>%
  mutate(satisfied = as.factor(ifelse(SATIS == 1, TRUE, FALSE)))%>%
  filter(SATIS != 9)
Marshall%>%
  ggplot(aes(satisfied, fill = satisfied))+geom_histogram(stat = "count")+
  ggtitle("Histogram of Satisfied Variable")+
  theme(plot.title = element_text(hjust = 0.5))
```



Analysis: More Data Cleaning

```
library(caret)
Marshall_complete <- read_csv("/Users/trent/Bob_Marshall/Marshall_complete.csv")
masscoercion <- function(x){
  if (nlevels(as.factor(x))>1 & nlevels(as.factor(x))<8) {
    x <- as.factor(x)
 } else {
   x
   }
}

Marshall_complete<- as.data.frame(map(Marshall_complete, masscoercion))

set.seed(3000)
trainindex <- createDataPartition(Marshall_complete$satisfied, p = 0.8,
                                  list = FALSE,
                                  times = 1)

Marshall_train <- Marshall_complete[trainindex,]
Marshall_test <- Marshall_complete[-trainindex,]
```

- Needed to designate variables as factors
- Filled missing data with MICE package
- Split data

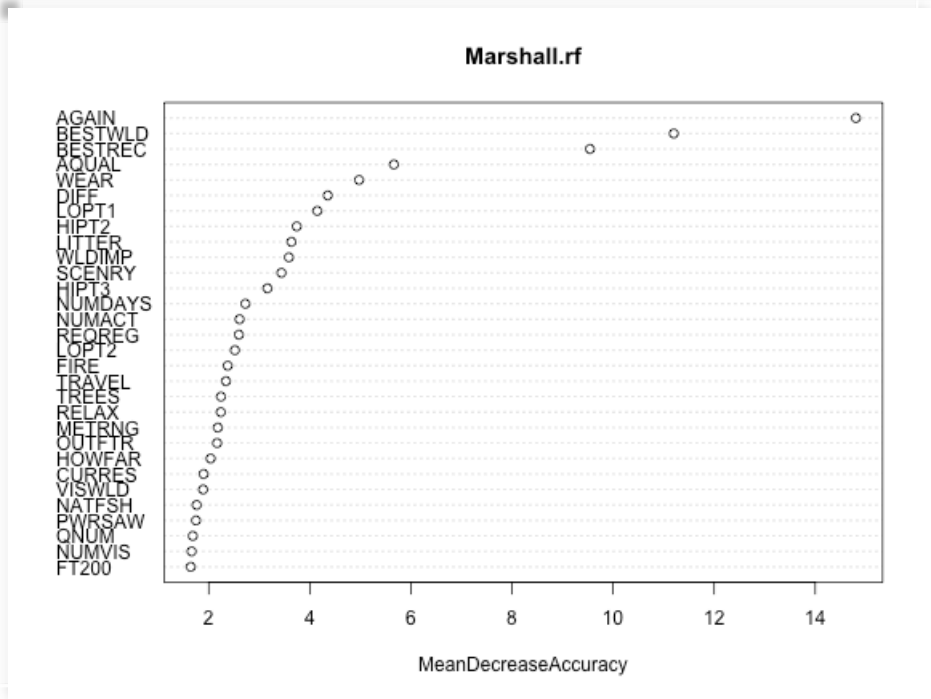
Analysis: First Random Forest

```
library(randomForest)

set.seed(3000)

Marshall.rf <- randomForest(satisfied~., data = Marshall_train, importance = TRUE)

varImpPlot(Marshall.rf, type = 1)
```



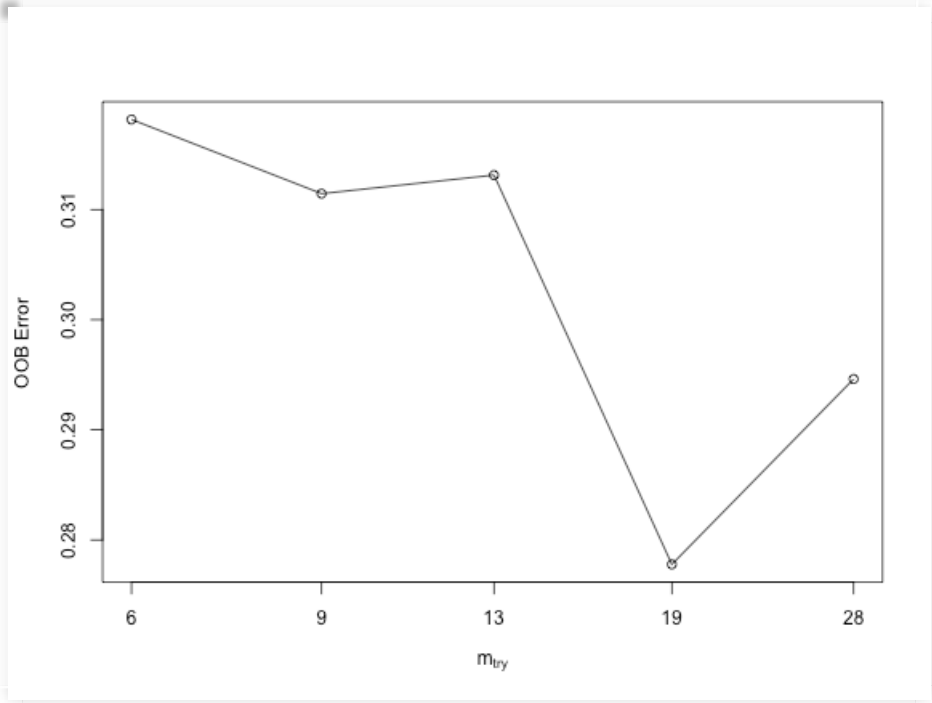
- Accuracy of 73%

Analysis: Second Random Forest

```
set.seed(3000)

bestmtry <- tuneRF(Marshall_train[,-181], Marshall_train$satisfied, stepFactor=1.5, improve=1e-5, ntree=500, doBest = TRUE)
```

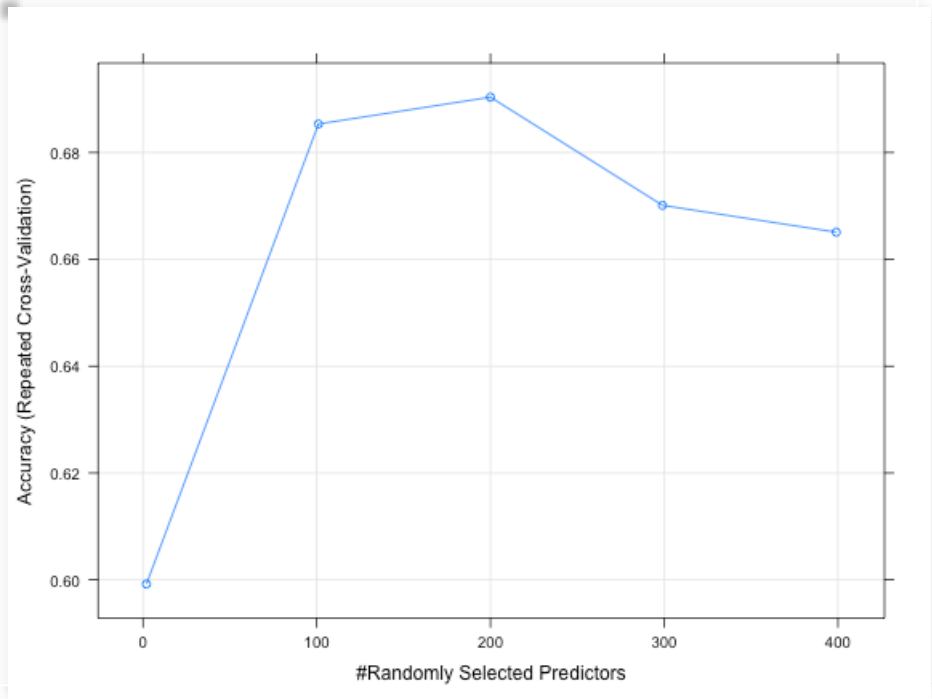
```
mtry = 13  OOB error = 31.31%
Searching left ...
mtry = 9     OOB error = 31.14%
0.005376344 1e-05
mtry = 6     OOB error = 31.82%
-0.02162162 1e-05
Searching right ...
mtry = 19    OOB error = 27.78%
0.1081081 1e-05
mtry = 28    OOB error = 29.46%
-0.06060606 1e-05
```



- Accuracy of 74%

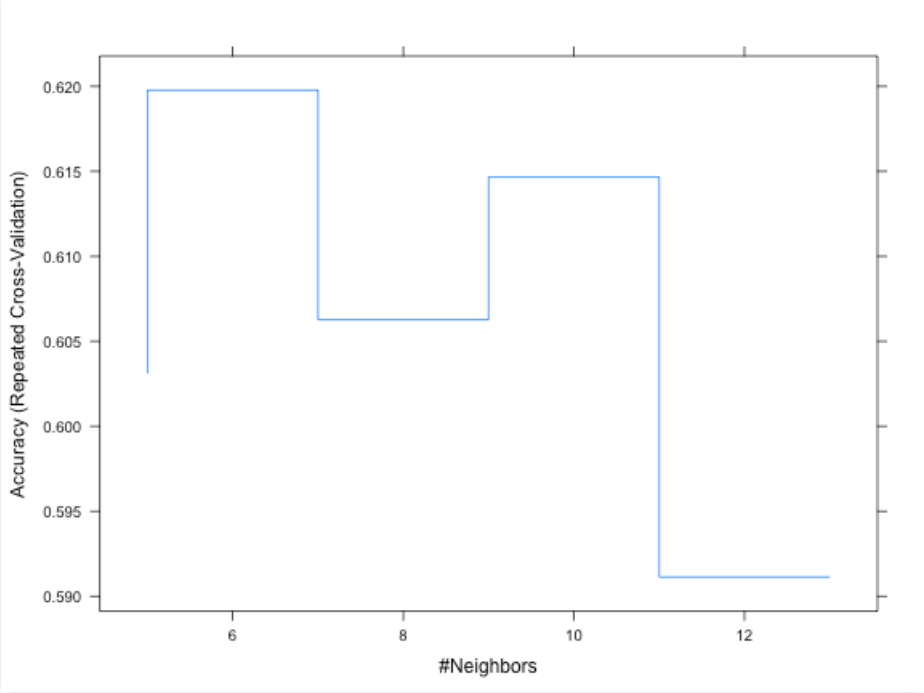
Analysis: Third Random Forest

```
set.seed(3000)
ctrl <- trainControl(method = "repeatedcv", number = 10, savePredictions = TRUE)
model_fit <- train(satisfied~., data = Marshall_train, method = "rf",
                   trControl = ctrl, tuneLength = 5)
plot(model_fit)
```



- Accuracy of 76%

Analysis:Bayesian and K-Nearest Neighbor



Results

- Accuracy for Each Model
  - Best Random Forest Model: 76%
  - Bayesian Logistic Regression: 65%
  - K-Nearest Neighbor: 56%

Next Steps

- Further Train Models
- Further Analysis into Important Variables