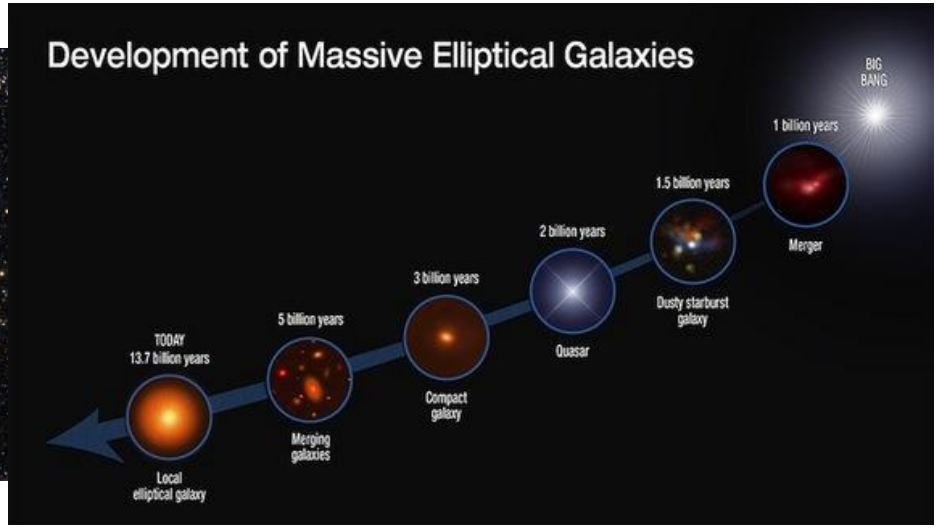
UseR! Kaggle Titanic and The Conditional Random Forest

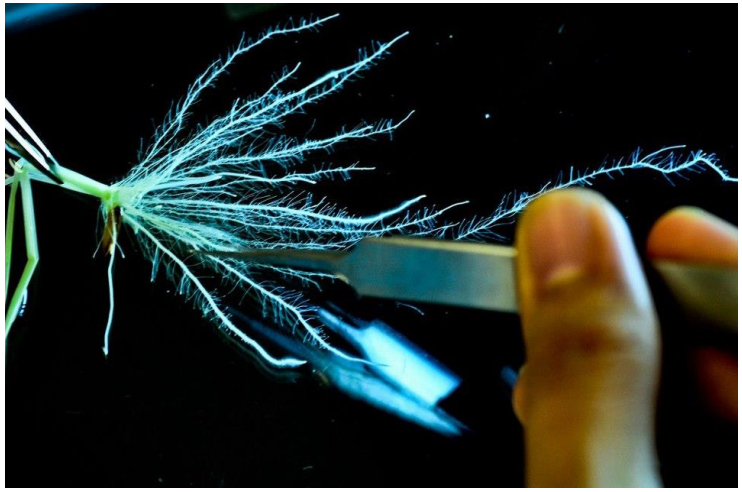
Bryan R. Balajadia

https://github.com/bbalajadia

DataSciencePH

Motivation


Development of Massive Elliptical Galaxies
BIG BANG
1 billion years
Merger
1.5 billion years
Dusty starburst galaxy
2 billion years
Quasar
3 billion years
Compact galaxy
5 billion years
Merging galaxies
TODAY
13.7 billion years
Local elliptical galaxy




XBOX 360
KINECT

Overview

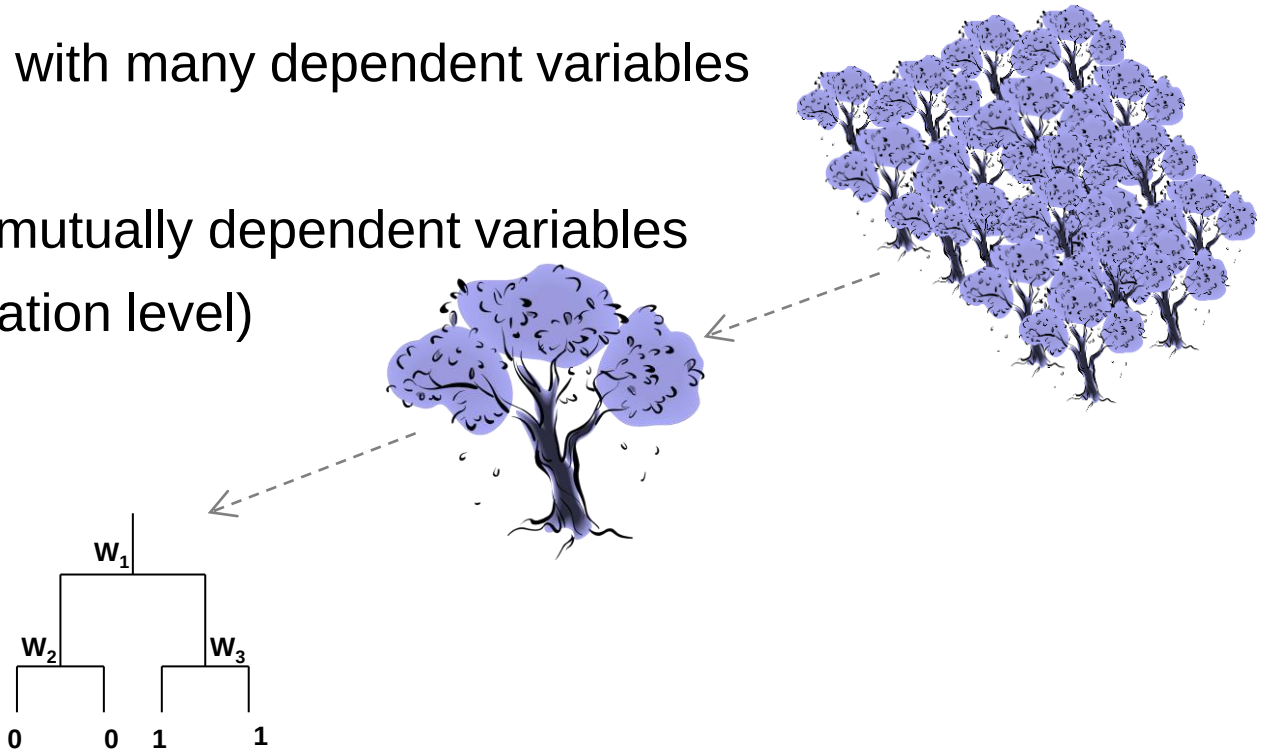
- Basics
  - Random Forests
  - Variable importance *(Gini vs. Permutation)*
  - R implementation
- Practical Example
  - Titanic: Machine Learning from Disaster
- Summary

What are Random Forests?

Breiman and Cutler's

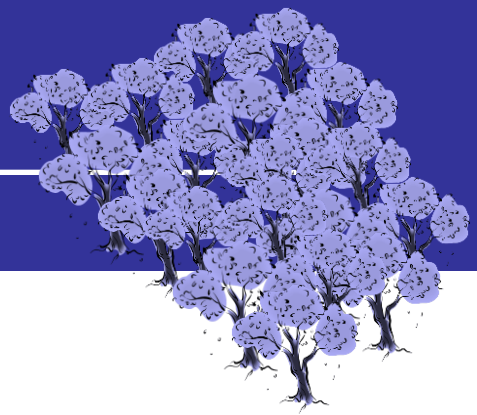
- Supervised Learning
- Ensemble of multiple independent decision trees
- Small sample sizes with many dependent variables
- Applicable even to mutually dependent variables

(e.g., wealth and education level)



Random Forests

Basic Algorithm for Classification



- Let ntree be the number of trees to build
- For each of ntree iterations
    1. Select a new bootstrap sample from training set

        Bootstrap sampling:

        "Bag" = 2/3 of the data; train

        "Out-Of-Bag" = the remaining 1/3; test
    2. Grow an un-pruned tree on this bootstrap
    3. At each internal node, randomly select mtry predictors and determine the best split using only these predictors.

"Overall prediction = <u>Majority vote</u> from all individually built trees."

Example of (Small) Random Forests

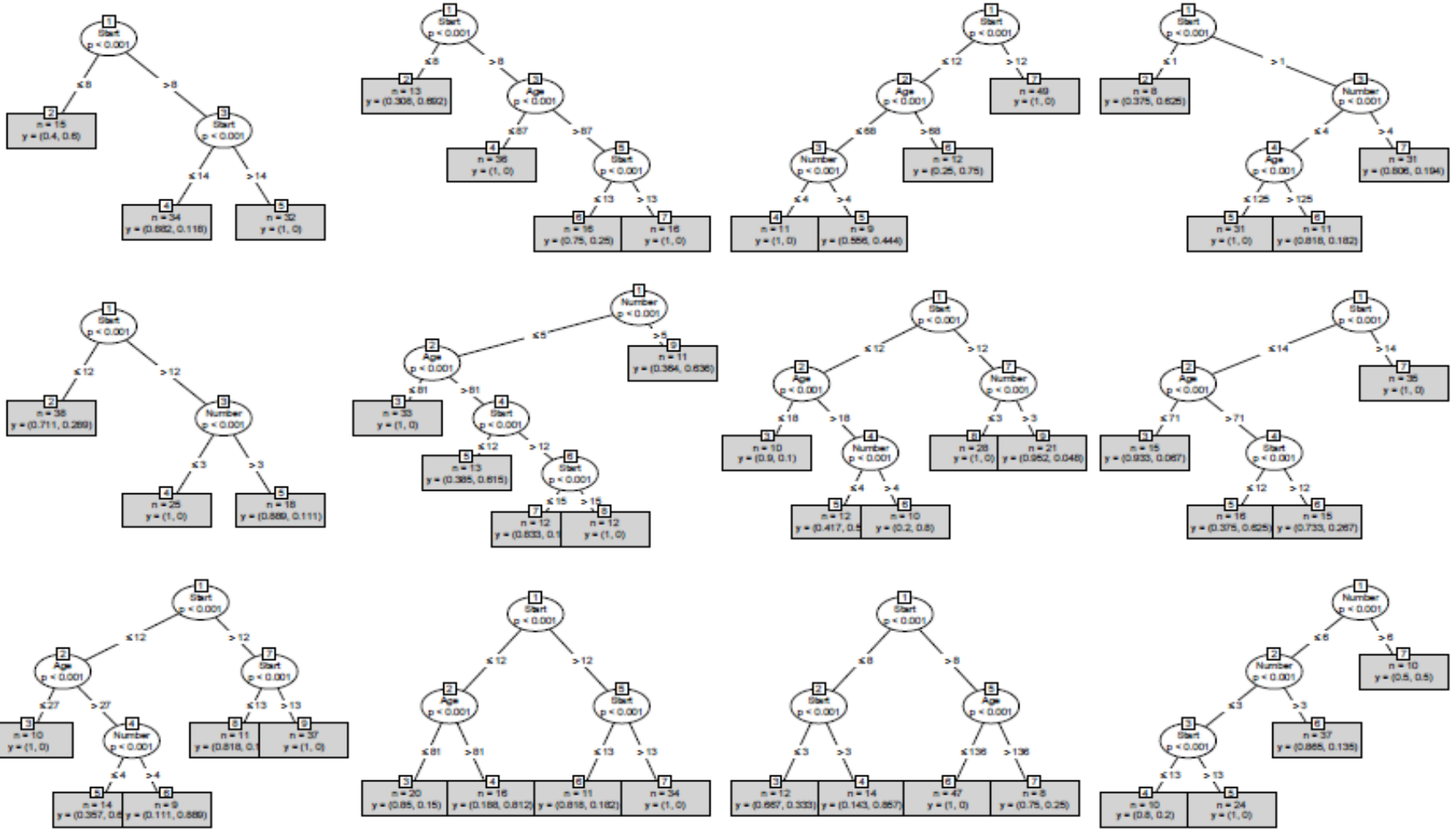


Image: Carolin Strobl

Random Forests in R

`randomForest` (pkg: `RandomForest`)

Employs information metrics (e.g., Gini coefficient) for selecting the variable at specific split

Variable-selection bias: biased in favor of continuous variables and variables with many categories

`Cforest` (pkg: `party`)

Utilizes conditional inference trees to avoid selection bias in `randomForest`

Practical example

Titanic: Machine Learning from Disaster

Titanic : Machine Learning From Disaster

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew.

This sensational tragedy shocked the international community and led to better safety regulations for ships.

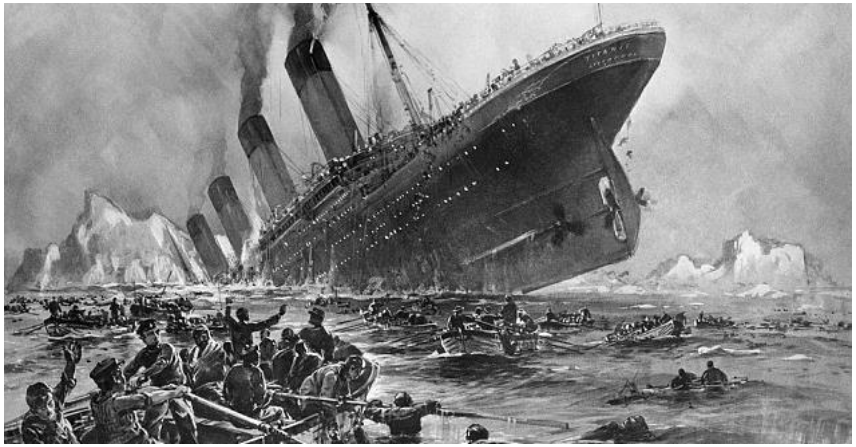

The Times Dispatch

Titanic, Giant White Star Liner, Sinks After Collision With Iceberg on Her Maiden Voyage, and 1,800 Lives Are Reported Lost in World's Greatest Marine Disaste

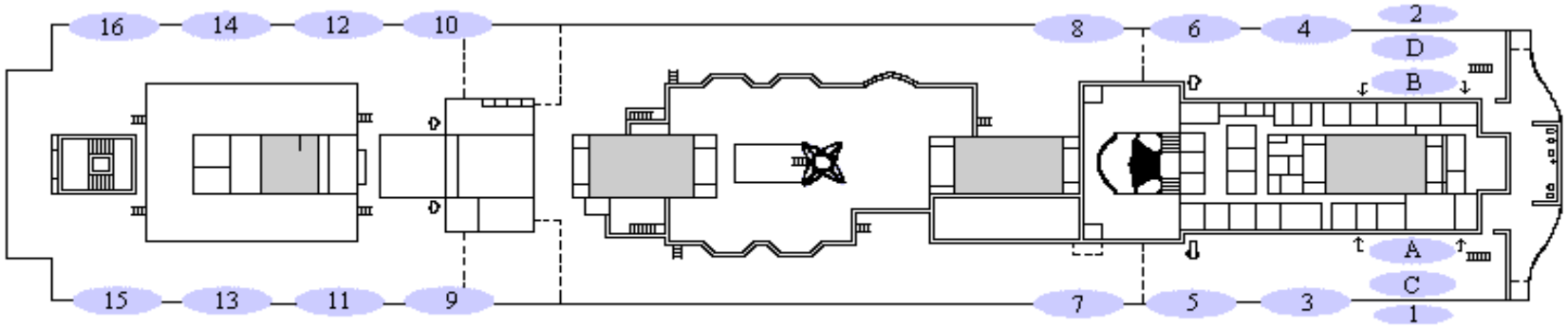
WIRELESS CALLS SEND VESSELS RUSHING TO AID OF SEA COLOSSUS

Through the Night They Drive With Full Speed to Reach Titanic.

MEANTIME, WORLD WAITS IN AGONY OF SUSPENSE



There were not enough lifeboats for everyone!



Copyright 2001: Titanic Inquiry Project

Boat Deck

64: The number of lifeboats the Titanic was capable of carrying

48: The number of lifeboats originally planned for Titanic by the chief designer Alexander Carlisle, 3 on each davit

20: The number of lifeboats actually carried aboard

Titanic Disaster: A priori knowledge

New York American

BUSINESS PROPERTY

J. J. ASTOR LOST ON TITANIC

1,500 TO 1,800 DEAD

John Jacob Astor was among the passengers who went down with the ship, according to a wireless dispatch received by Bradstreets last night from the liner Olympic. Mrs. Astor was saved and is being brought to shore by the Carpathia.

The Wireless Operator at Cape Race, Newfoundland, Flashes: "Eighteen Hundred Lives Have Been Lost in the Wreck of the Titanic."

Who were more likely to survive?

- Gender
- Age
- Class
- Fare
- Family

<u>More likely to survive</u>

- Females
- Children
- 1st Class Passengers
- Traveling with Family

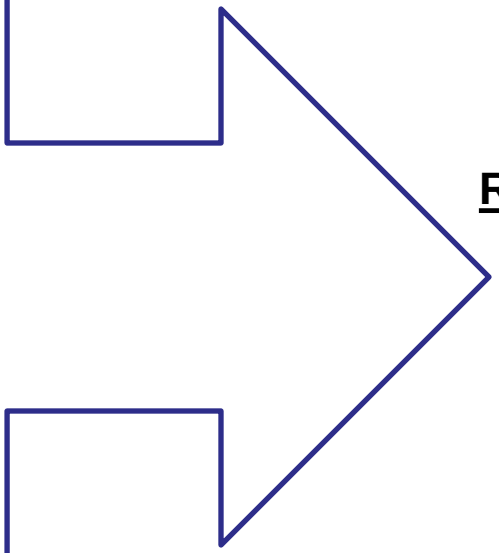
<u>More likely to perish</u>

- Males
- Adults
- 2nd and 3rd Class Passengers
- Traveling alone

Titanic Dataset

Predictor Variables

- Passenger Class (`Pclass`)
- Passenger Name (`Name`)
- Sex (`Sex`)
- Age (`Age`)
- No. of Sibling/Spouse aboard (`SibSp`)
- No. of Parent/Child aboard (`Parch`)
- Ticket number (`Ticket`)
- Fare (`Fare`)
- Passenger Cabin (`Cabin`)
- Port of Embarkation (`Embarked`)



Response Variable

Survived
(1 = Yes; 0 = No)

Feature Engineering

```
## Create Title variable using the Name info
## Hypothesis: Title is linked to AGE and SOCIAL STATUS

all$Name <- as.character(all$Name)
all$Title <- sapply(all$Name, FUN=function(x)
                    {strsplit(x, split='[,.]')[[1]][2]})
all$Title <- sub(' ', '', all$Title)
all$Title <- factor(all$Title)


## Family Size
all$FamilySize <- all$SibSp + all$Parch + 1
```

```
Model:
      Survived ~ Pclass + Sex + Age + Fare +
Embarked +  Title + FamilySize
```

Model: randomForest (pkg: randomForest)

```
library(randomForest)


set.seed(1992) #RANDOM Forest!


fit <- randomForest(Survived ~ Pclass + Sex + Age +
Fare + Embarked + Title + FamilySize, data=train,
ntree=1501, importance=TRUE)


## Create a prediction
pred <- predict(fit, OOB=TRUE, test)
```

KAGGLE RESULT:

0.79426

Conditional RF: cforest (pkg: party)

```
## Conditional Random Forest Model

library(party)
set.seed(1992) #RANDOM Forest!
```

```
data.controls <- cforest_unbiased(ntree=1501, mtry=3))

fit <- cforest(Survived ~ Pclass + Sex + Age + Fare +
Embarked + Title + FamilySize, data=train, controls =
data.controls)
```

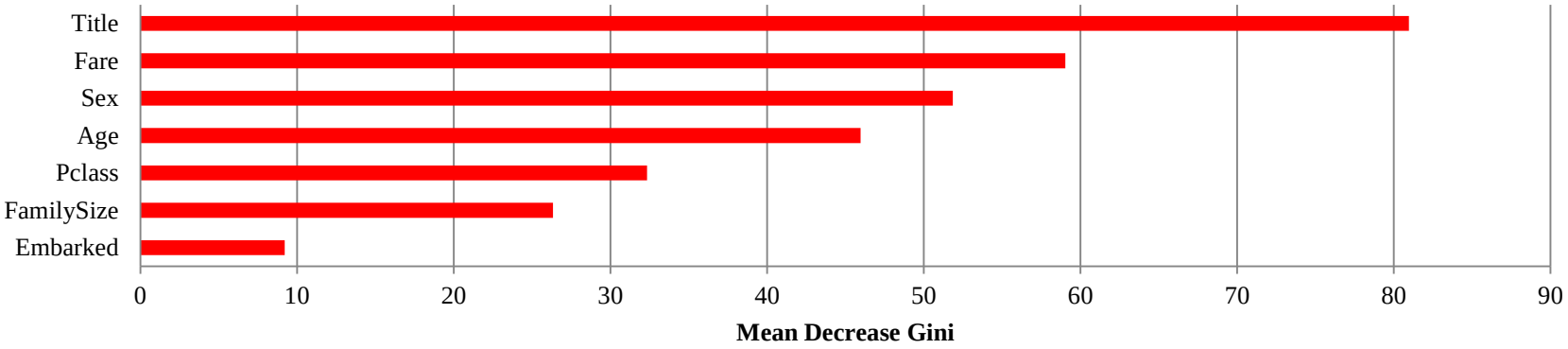
```
## Create a prediction and write a submission file
pred <- predict(fit, OOB=TRUE, test, type = "response")
```

KAGGLE RESULT:

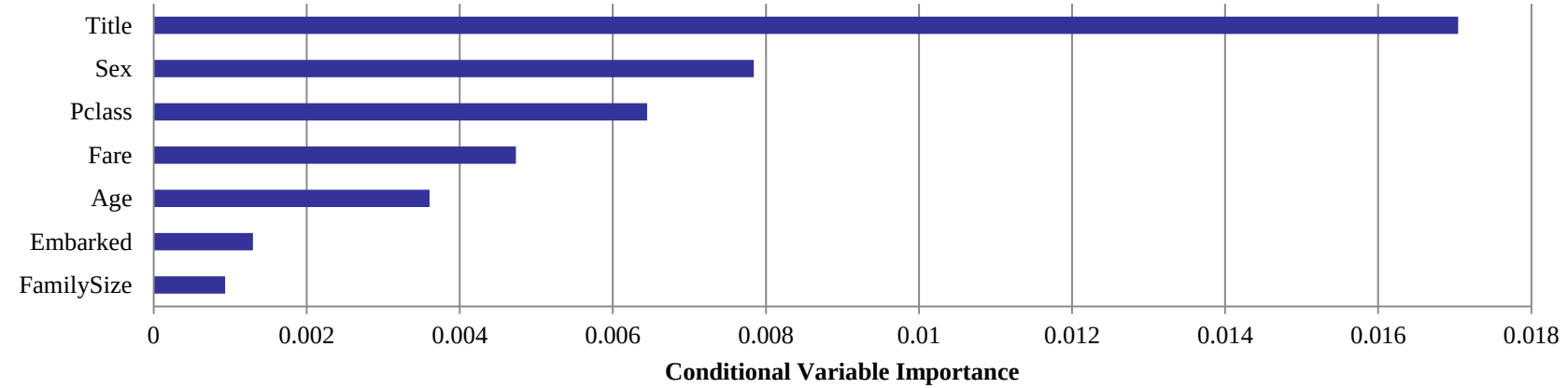
0.80383

Variable Importance

randomForest()
Prediction Accuracy: 79.426 %



cforest()
Prediction Accuracy: 80.383%



When to use what? RF vs Conditional RF

- If *predictor variables are of <u>different types</u>:*

  use `Cforest` (pkg: `party`)

  else feel free to use:

  `randomForest` (pkg: `RandomForest`)

- If *predictor variables are <u>highly correlated</u>:*

  use `Cforest` (pkg: `party`)

References

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics 15(3), 651–674.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. BMC Bioinformatics 9:307.

Altmann, A., L. Tolosi, O. Sander, and T. Lengauer (2010). Permutation importance: a corrected feature importance measure. Oxford University Press