Contents

[PRE-BOOTCAMP 2](#_Toc465237127)

[DURING BOOTCAMP 6](#_Toc465237128)

[DAY 1: Data Exploration 6](#_Toc465237129)

[DAY 2: Basic Building and Evaluating Predictive Model 9](#_Toc465237130)

[DAY 3: 12](#_Toc465237131)

[Questions 13](#_Toc465237132)

# PRE-BOOTCAMP

What is data?

Data = Objects defined by attributes

Types of *attributes*:

* Discrete = integer variables or binary (e.g. click counts)
* Continuous = real numbers (e.g. temp, weight)
* Categorical – nominal or ordinal (has an order)
* Interval
* Ratio

Types of *data*:

* Record
  + Transaction = each record is a set of items (more info than count associate)
* Matrix
* Document = counts number of appearance
* Graph
* Ordered data (e.g. time series data)

Data quality problems:

* Noise and outliers
* Missing values 🡪 SOLN eliminate/estimate/ignore/replace
* Repeated data

Data Preprocessing

* Aggregation = combining two or more attributes to a single attribute > data reduction, change of scale, decrease data variability
* Sampling (sample needs to be **representative**)

Types of *sampling*:

* Simple random
* Stratified
* Sampling without replacement
* Sampling with replacement

Dimensionality reduction

* Principle Component Analysis
* Singular Value Decomposition
* Others: Various supervised and non-linear techniques
* Feature subset selection (remove dimensionality that does not add information)
  + Brute force
  + Embedded = natural feature selection
  + Filter = feature pre-selected
  + Wrapper = black box to find best subset
* Feature creation
  + Extraction
  + Construction
  + Mapping data to new space
* Attribute transformation
  + Simple functions
  + Standardization
  + Normalization

Similarity = measure of how alike two data objects are [0 – 1]

Dissimilarity = how different [0 – upper limit varies]

* Euclidean distance
* Cosine similarities
* Correlation
* Visual evaluating

Data exploration and visualization **\*DAY 1**

* Visualization and calculation to better understand characteristics of data
* Clustering and anomaly detection

Summary statistics = number that summarize property of data

Visualization techniques: histogram, box plot, scatter plot, contour plot

What is R? Programming language for data science

|  |  |
| --- | --- |
| <- | Assignment |
| ? | Provide example of function |
| ?? | Help |
| c( , ,… , ) | vector |
| as.factor( ) / levels() | Take unique characters and index |
| rbind() / cbind() | Matrix to row or column bind |
| colnames() rownames() |  |
| is.nan() / is.na() |  |
| library() required() |  |
| dir() |  |
| read.csv() |  |
| read.table(‘’, header=T) |  |

Basic data types:

* Numeric
* Integer (e.g. a <- 132L)
* Complex
* Character (string)
* Logical (Boolean)(e.g. TRUE, FALSE, T, F)

R restrictions:

* Vectors must all be of the **same class**
* First vector index is [1]
* . is a valid name e.g. vec.a

Lean six sigma

Steps to Machine Learning

1. Get in Data
2. Data exploration
3. Clean data
   * Categorical casting
   * Filter/Transform out strings
   * Cleaning missing data 🡪 Replace with median or mode; Remove entire row; Substitute
4. …

|  |  |
| --- | --- |
| **MODULE** | **FUNCTION** |
| Metadata Editor | Transform data, e.g. change into categorical |
| Descriptive statistics | Check for which feature has missing data |
| Clean missing data | Make adjustment to missing data |
| Project columns | Exclude columns |
| Apply SQL transformation |  |
| Execute R script or Python |  |

# DURING BOOTCAMP

## DAY 1: Data Exploration

Type of data science:

* Classification
* Regression (numbers)
* Un-supervised learning
* Recommender system
* Ensemble Method

Bad model 🡪 Wrong model, wrong move e.g. Amazon recommender

DON’T GET ATTACHED TO A LANGUAGE

Not above 90% accuracy, w/ historical data you want to be predictive

R Programming

* Levels 🡨 used with categorical
* Important to understand the levels in factors

**Hot**key

* Data$Col\_name

DATA SCIENTIST STEPS

1. Frame qns
2. Obtain data
3. Scrubbing data
4. Exploring data
5. **FEATURE ENGINEERING** = create features that make machine learning algorithm works
6. Model
7. Deploy model
8. Interpret

Data beats algorithm… Extract useful feature from data

Segment different models for different type of stock

Common graphical parameters 🡨 main = “title”, xlab = “x-axis”, col FOR colors

Interquartile tells the middle 50%

GRAPH

> png(“bla.png”) # Save Plot

> Dev.off

* SCATTER 🡨 should not have categorical value
* BOXPLOT 🡨 The outlier percentage is dependent on the chosen interquartile range
* PIECHART 🡨 needs table()
* HISTOGRAM 🡨 increase bins with breaks = #no
* DENSITY 🡨 peaks at most dense plots and decrease ala Gaussian curve as density decrease e.g. towards zero

**HOT**key

Scatterplot Matrix:

* Pairs()
* GGally 🡨 ggpairs()
* Library(lattice) 🡨 splom()
  + Xyplot( ~ , groups= , auto.key=TRUE)

Primary key 🡨 unique ID, useless in titanic data in predicting survivors

Foreign key 🡨 keeps coming back e.g. student ID, anticipates Timmy’s needs

Composite key 🡨 e.g. address, driver license ID

Mean is misleading as outliers can pull values. Median is more honest.

What to predict?

Numeric 🡨 regression

Categorical

* Ordinal 🡨 ordinal regression
* Nominal 🡨

Historical Data 🡨 70% training, 30% testing

**HW HINT**

> g + geom\_point(aes(color=color)) +facet\_wrap(cut ~ clarity)

> g + geom\_point(aes(color=color)) +facet\_grid(cut ~ clarity)

**SOLN**

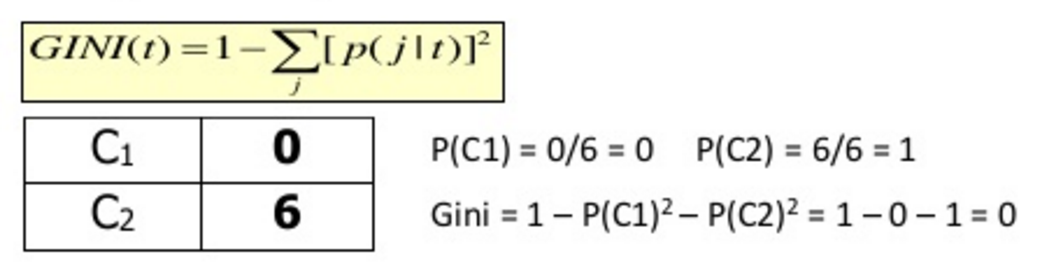
>ggplot(titanic,aes(x=Embarked,y=Age)) + geom\_bloplot(aes(color=Sex)) + facet\_wrap(Pclass)

## DAY 2: Basic Building and Evaluating Predictive Model

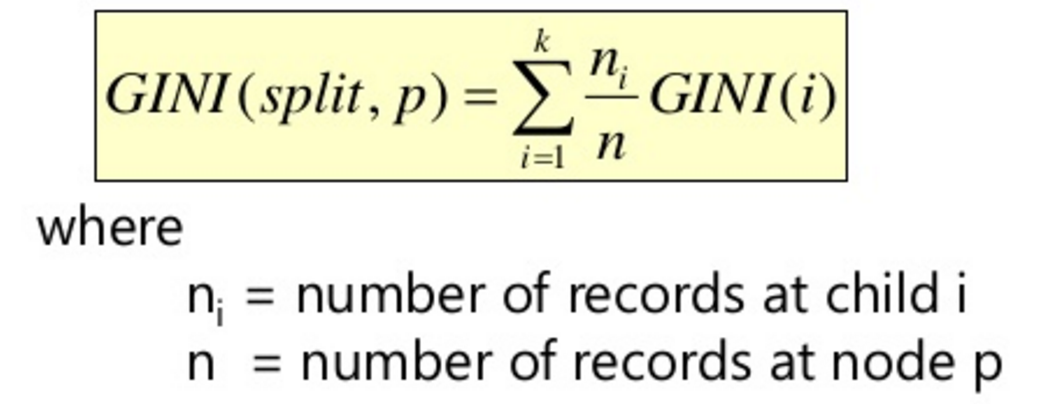
Measure of split 🡨 node impurity

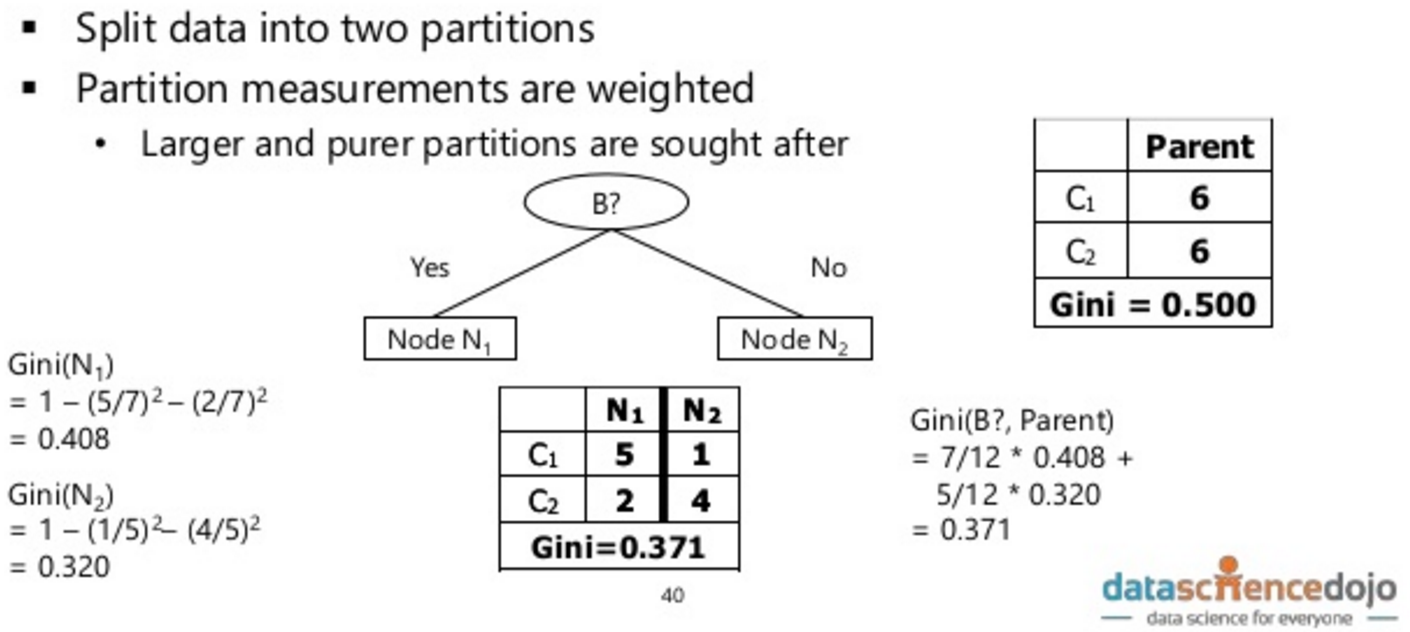
Measure of Node Impurity

* GINI index
* Entropy
* Misclassification error



Improvement is measured in decrease in GINI





The GINI of this particular split for this particular column is 0.371. Find best split, find lower GINI

Tuning parameters 🡨 to control span and growth of trees

Pruning 🡨 slash outer notes but not computationally feasible

Make sure that what you want to predict is a factor in r (categorical value)

sample(1:5) is different than sample(1:5, 1) The Latter randomly picks a number out of 1-5

Test is usually the remainder of the train data set. More complication when splitting data > 2

Publish data in cran, needs predictor

LEFT 🡨 prediction wanted RIGHT🡨 predictors

Surrogate split 🡨 split performed if the first split is unavailable (e.g. due to missing data)

While accuracy is < , up the seed. But this is a bad model as it leads to overfitting

Sample stopping criteria

* Number of levels
* Number in Leaf Node
* Minimum samples per lead node

Minsplit 🡨 must have this much observation before split

Minbucket 🡨 let’s consider a split, children note needs at least 7 observation in the node

Max depth 🡨 e.g. four for Iris.data

Cp 🡨

Accuracy 🡨 (TP+TN)/(TP+TN+FP+FN) how many is correct

Precision 🡨 TP/(TP+FP) does not consider TN, measures relevant, How True?

Recall/Sensitivity 🡨 maximize finding TP, out of all cancer how many do you find? But you disregard FP and may increase

AUC 🡨 you want it high but not too high because it is suspicious

F1 score 🡨 the higher the number means the more optimum, recall/precision combination

Overfitting 🡨 lack of generalization, you learn rules that only pertain to the data

Bias/Variance tradeoff

Cross validation 🡨 only training and test data switch, everything else (e.g. model) is kept same

Choose the most stable data in terms of parameters (accuracy, F1)

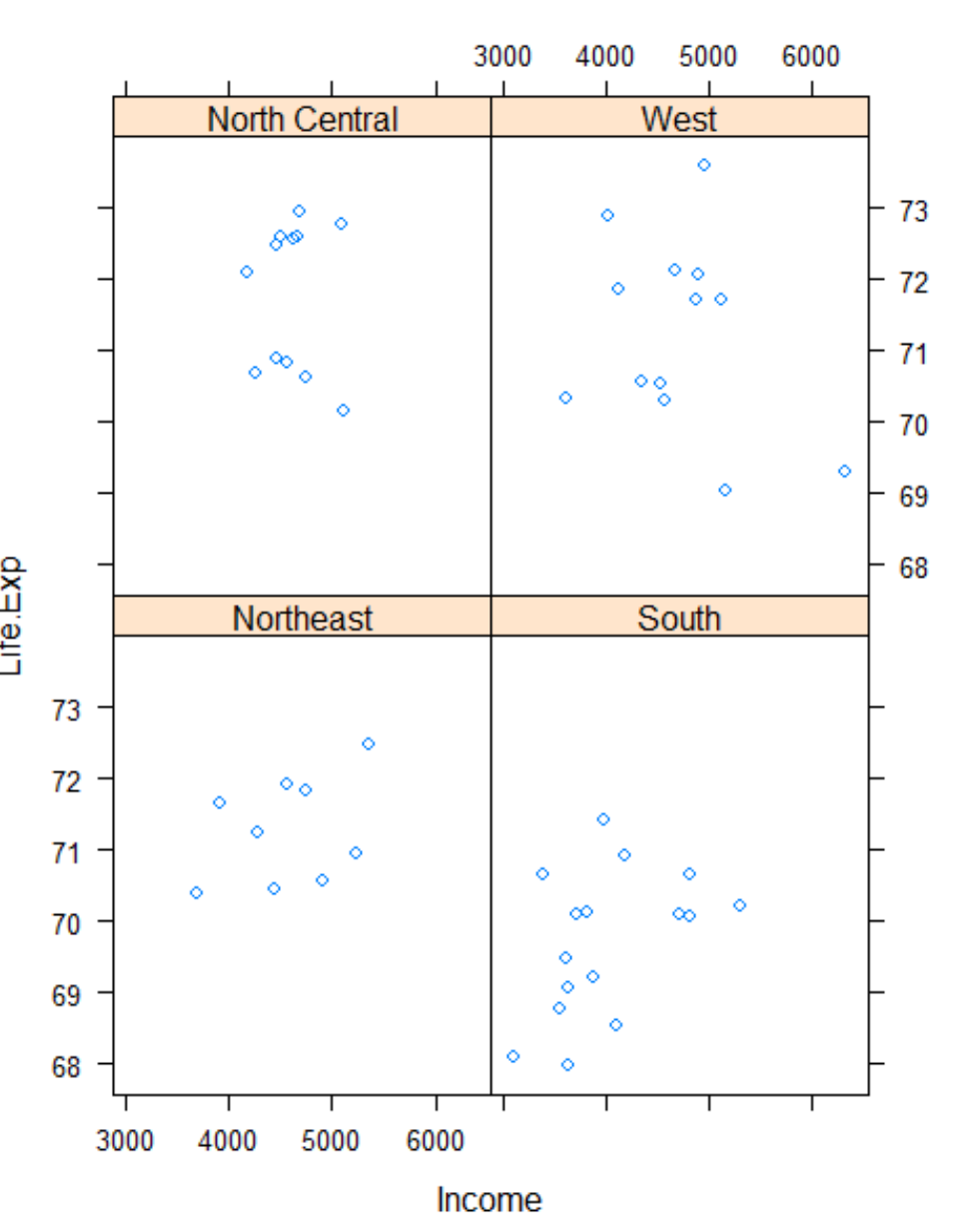
Maybe use the one with the set of data closest to the mean

Minimum number of samples per leaf node

Maximum tree depth

**HOTkey**

> xyplot(Life.Exp ~ Income | region, data = state, layout = c(2,2))



par(mfrow,c(1,2)) #create two graphs

## DAY 3:

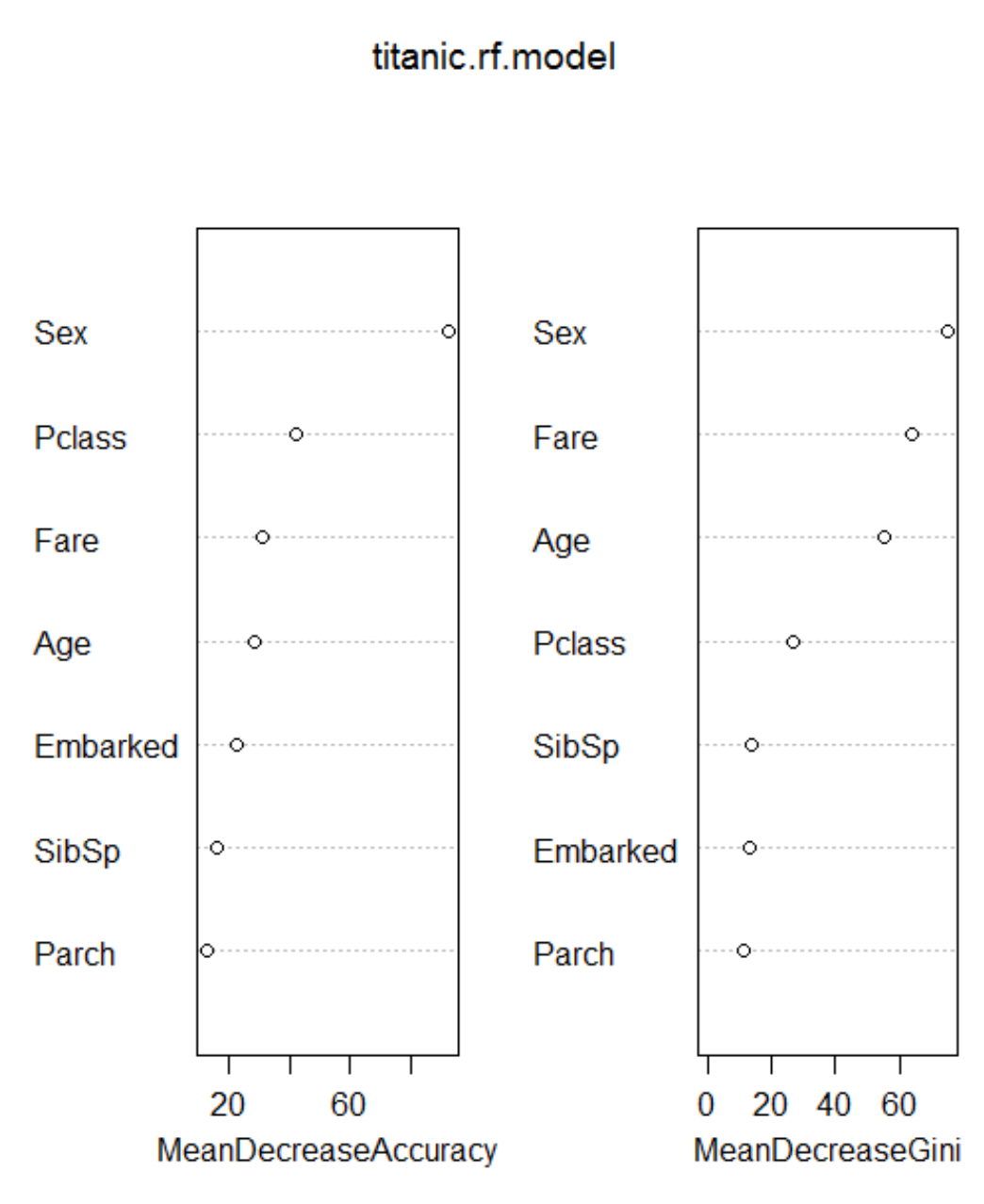
Binomial distribution 🡨p = p of event from all outcomes, n = # trials, k = succession

E.g. probability of getting 5 at least three times, given u roll it 4 times. P = 1/6, n = 4, k = 3

Which feature are used for learning?

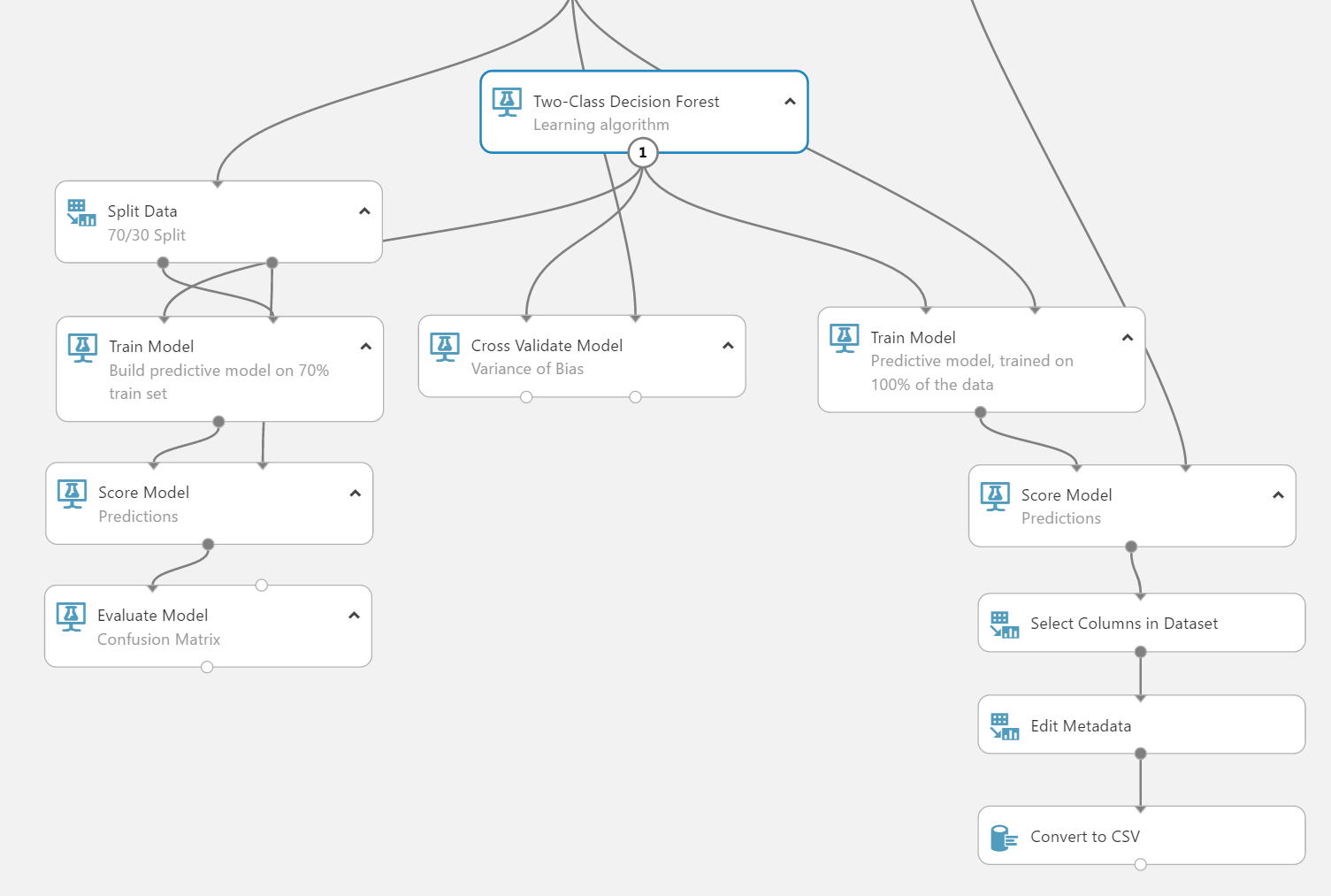
Feature selection

varImpPlot(titanic.rf.model)



Ensemble with two class decision forest with cross validation -> train model

Increasing number of trees <- reduces variance, less overlap e.g. for 32000 samples not much overlap for 30 trees



Feature select in the first branch

Test stability with cross validate – high bias and low variance

DAY 4

Stop when the function stop changing (similar to if GINI is unchanging)

Or when theta does not change

Gradient decent <- you know revenue and factors that affect it. Optimization technique

Ensure that the third and fourth order is smaller to not make overly complex model, but limits contribution to relevant ones

Token <- specific word

Term <- the version of a word set that is in the dictionary

Stemming <- convert tokens to terms by removing letters via heuristic

Lemming <- classify tokens into term in linguistic

Document <- unit of information (e.g. youtube vid, tweet, user profile, email)

Binary approach <- count whether the word exist or not and assign 0 or 1

TF-IDF Approach

Term Frequency: Measures how often a term appears (density in a document)

Inverse document frequency: Aims to reduce the weight of terms that appear in all document

A/B Testing Purpose

Know what the users wants subconsciously or otherwise

Nodes 🡨 how many computer

Random forest javacript? Probably not

64 gig ram

Dual core split the work

Haddop is more popular

Hive is language that takes diff data on diff computer

Diff security, diff cloudride, but for data scientist haddop is pretty equal unless ure acrchitect

HDinsight Microsoft flavor

No need data center, various hardware

Cloud computing

H base and mongo cluster <- if one goes down, all is lost

Why is storage not an issue in cloud

Install.packages”Hive”

Hadoop process and comes back to R as result

url username, passworld

if u ever ever hack excel to sql

how to calculate variance

eqn for standard dev

equation to mean

you want to hold off using sql database as much, maybe 20 GB is limit

labor, time hardware go up. Database administrator,

Cassandra cluster

felt disconnect to stats but now it makes sense

Hadoop ver 1 is update through programming olympics

each data node didn’t know other data node exist

turned hive into its own language

tez leverage yarn

the diff is execution engine = versions of Hadoop (E.g. impala vs tez)

Learn Pig language – yahoo and big company uses it, goes back to map reduce because it is a lot faster than hive. If you don’t want to use C and java

Mahout is machine learning inside of Hadoop

Spark cluster is good but more expensive

Almost every atm still run on windows xp

Spark is new, still bugs,

Hadoop is too slow, so they switch to spark. More expensive but needs company with set of coder team to sort the bug

Use to be rewritten to java and python

Revolution r, so you can use r and r studio directly to Hadoop cluster

Data base

Data warehouse

Eucladian is only applied now

Remove things by disabling cool features

Conditional inference forest

C forest

Features parch, sibsp

Caused noise if you left it

Pclass and sex made into numerical instead of

Have some effect on the probability

So when pclass became numerical, you can group it together

I had so many categorical so I changed stuff to numeric

Parameter

Do you need to do cross validation with trees? Theres a debate going on

Yea because my cross valid consistently showed

More trees didn’t mean less trees

The kaggle is the cross validation

Minimum skill set to be dangerous

Expose you to whatever is out there and possible

Kernel of the support function

Spatial stuff is happen. When you take the sin of smth and divided it by the cos of smth, magic happens

Popular event brokers: kafka

Rabbitmq 🡨 https usually has big header, faster than kafka, don’t adopt unless you have a developer team that knows java very well

Ibm was really big with atl in the 2000

Event hub for IoT 🡨 intercepting data and deliver it to people who needs it

Hyper replication of event hub

Each partition reads 2 megabyte

Message retention is high

Sell data sell these keys

Mouse signature

Google can only track you when youre in the website

What if they don’t have google or gmail

Data of everyone

Facebook pixel

Twitter pixel

Throw away enron

Overflow things, because its easier to find you

People who can grab the most comfort for the most people win

Google car initiative

Google everything is free because they sell data

Google can spy on you when you re online

Microsoft can spy on you offline

Stripe 🡨 does not need credit card, they already know all the cvs

Twitter have the best developers

Testimonial

Love the fun fact. Phuc is very knowleadgeable and passionate.

Financial forecasting

Linux box

120 analyst

Google finds hardtime to charge

Processing is the cost, storage is cheap now

Job topology

Event hub consumer group 🡨 e.g. Nordstrom and you can cut them out anytime

Dropbox pay for sinking

Data engineer 🡨 make sure it gets there on time and in good format

Select

From

Twitter\_feed

Group by TumblingWindow(minute, 5/0, hashtag

Sort by count(hastag) desc

Limit 5

Select

From

Demand

Supply

Group by window(minute, 30) location

Surge price for uber

Power Bi 🡨 free for company

Tablo and dato

Google business model

Hadoop definitive guide

Sensor to check inside

Predictive learning to optimize

## Questions

**TITANIC update**

Make the pclass numeric

**QNS today**

How come sweep parameter suggest depth>number of features? Is it recommended to use this feature?

When submitted to kaggle, sweep parameter’s high accuracy model performed worse

Raj say more feature better? But my accuracy decreases with feature, more feature and depth more overfitting?

If I make many features, will the algorithm sort it out to the best few that prevents overfitting

Can I add conditional statement when predicting just to improve accuracy?

Wanted to see what kills a woman, since being a woman is the best chance of surviving. I found number of sibling>2 will die. Should I rely on the age feature or make a new one. Why?

How can I do the feature prioritization with the TITANIC? So I CANT depend on the algo to take the features with the best order out of the selection of features

Webservice?

Running all script automatically like Matlab possible?

**YES** <https://support.rstudio.com/hc/en-us/articles/200484448-Editing-and-Executing-Code>

The only reason we split data, is to score model if we don’t have blind hold out data?

**YES**

Determining the right feature is important? E.g. embarked provide the same value as class with class being more powerful, same with cabin which is determined by class

**Collinearity**

Is it common practice modify test data and replace missing value to make it suitable to your model? So for Azure, have to fill the NA just like the train? Can I run with missing value?

**QNS for Detective Mannuel**

* Hadoop cluster?
* How to get data from the sites?
* FedEx predicts defection to a competitor with 65- 90%
* Square root method (intro to comp. science)
* Binomial distribution

**POST BOOTCAMP QNS**

* Readings?
* Python book?
* How to get data from the sites?