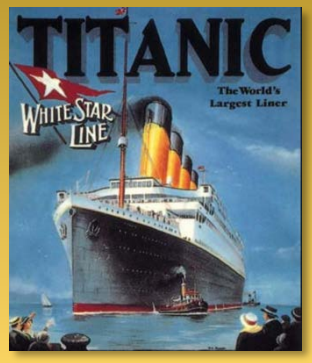
R Script Feature Engineering

R SCRIPT FEATURE ENGINEERING



R Script Feature Engineering

In this session

- What is the Feature?
- What is Feature Engineering?
- The process of Feature Engineering
- Where is FE in ML?
- Preparing for experiment
- Adding family size feature
- Adding Age*Class and Fare per person feature
- Adding Deck feature
- Adding Title feature

R Script Feature Engineering

What is the Feature?

What is the Feature?

- A piece of information
- Might be useful for prediction
- Any useful attribute to the model
- Is measurable property
- Feature is input; label is output.
- Is one column of the data

R Script Feature Engineering

What is Feature Engineering?

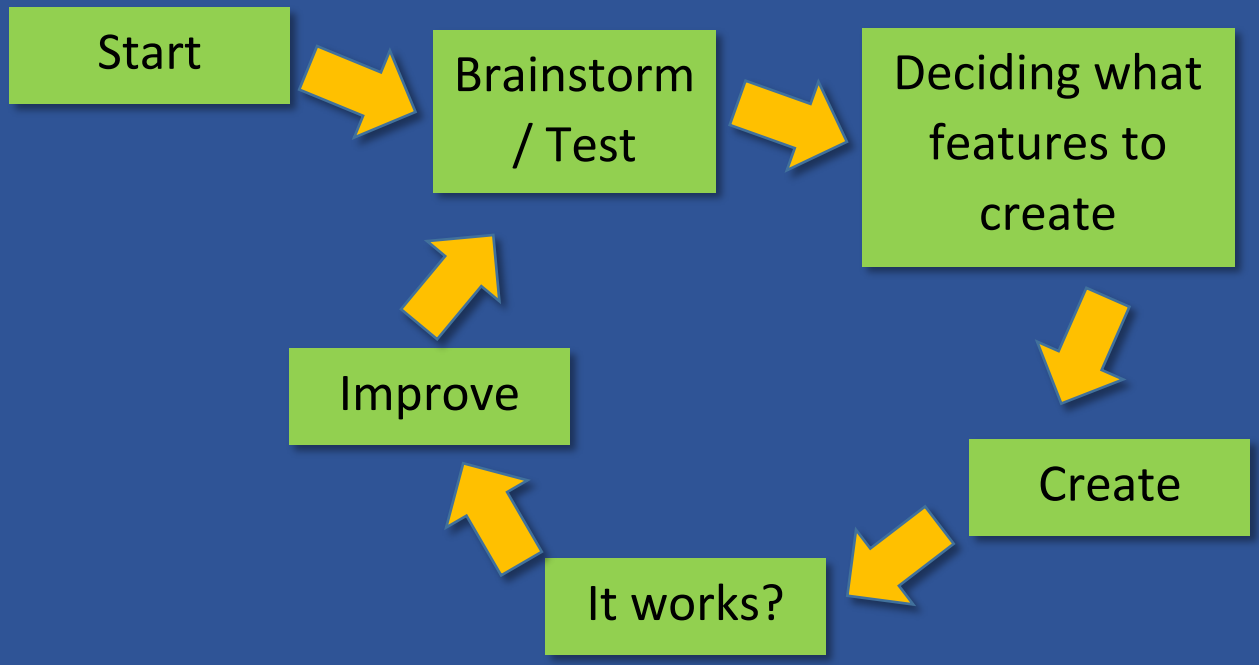
What is Feature Engineering?

- Is the method if find X for input
- Is “Data Science”
- Is difficult
- Is expensive
- Is time-consuming
- Is require expert knowledge in domain
- Is applied machine learning

R Script Feature Engineering

The process of feature engineering

The process of feature engineering

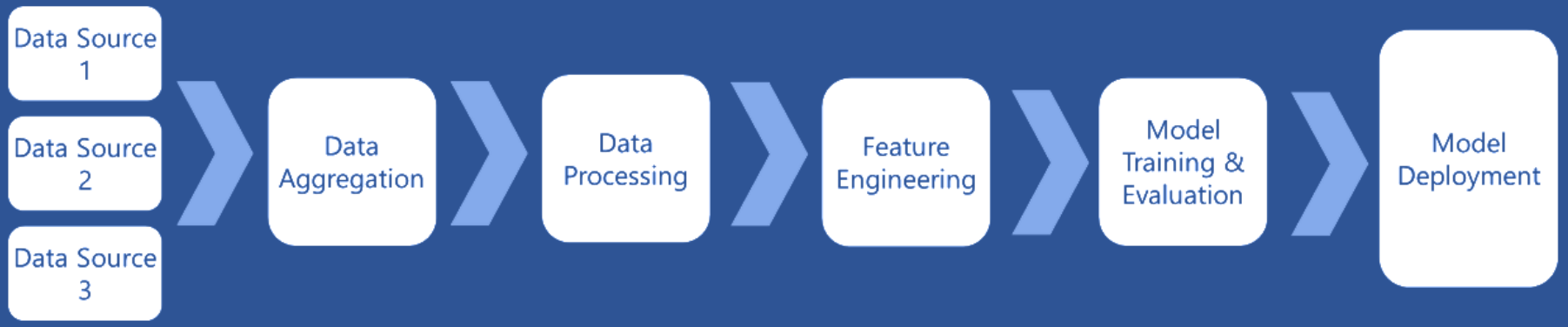


R Script Feature Engineering

Where is FE in ML?

Where is FE in ML?

- Data sources
- Data aggregation
- Data Processing
- Feature Engineering
- Model Training & Evaluation
- Model Deployment



R Script Feature Engineering

Preparing for experiment

Preparing for experiment

1. Go to https://github.com/laploy/ML
2. Right click TitanicData.csv and save link as to c:\temp
3. Open R Studio
4. Create New project name = c:\temp\rfe
5. Right click project / click Add… / New R Script file

R Script Feature Engineering

Add family size feature

File name = 100 familySize

```
 7  rm(list = ls()) # clear work space
 8  setwd("d:/temp") # set current work directory
 9  # import data file
10  dat <- read.csv("TitanicData.csv", na.strings = "")
11
12  # ----------------- Main program -------------------
13  # Create Family Size feature
14  dat$familySize <- dat$SibSp + dat$Parch
```

Ticket	Fare	Cabin	Embarked	familySize
A/5 21171	7.2500	NA	S	1
PC 17599	71.2833	C85	C	1
STON/O2. 3101282	7.9250	NA	S	0
113803	53.1000	C123	S	1
373450	8.0500	NA	S	0

R Script Feature Engineering

Add Age*Class and Fare per person feature

File name = 101 ageClass

```
7  rm(list = ls()) # clear work space
8  setwd("d:/temp") # set current work directory
9  # import data file
10 dat <- read.csv("TitanicData.csv", na.strings = "")
11
12 # ------------------ Main program --------------------
13 # 1. Create ageClass feature
14 dat$ageClass <- dat$Age * dat$Pclass
15 # 2. Create Family Size feature
16 dat$familySize <- dat$SibSp + dat$Parch
17 # 3. Create fare per person
18 dat$FarePerPerson <- dat$Fare / dat$familySize
```

Fare	Cabin	Embarked	ageClass	familySize	FarePerPerson
7.2500	NA	S	66.00	1	7.250000
71.2833	C85	C	38.00	1	71.283300
7.9250	NA	S	78.00	0	Inf
53.1000	C123	S	35.00	1	53.100000
8.0500	NA	S	105.00	0	Inf
8.4583	NA	Q	NA	0	Inf
51.8625	E46	S	54.00	0	Inf
21.0750	NA	S	6.00	4	5.268750
11.1333	NA	S	81.00	2	5.566650

R Script Feature Engineering

Add Deck feature

File name = 102 addDeck

```
 7  rm(list = ls()) # clear work space
 8  setwd("d:/temp") # set current work directory
 9  # import data file
10  dat <- read.csv("TitanicData.csv", na.strings = "")
11  # create Deck feature
12  dat$Deck <- (substr(dat$Cabin,0,1))
13  print('end')
```

Parch	Ticket	Fare	Cabin	Embarked	Deck
0	A/5 21171	7.2500	*NA*	S	*NA*
0	PC 17599	71.2833	C85	C	C
0	STON/O2. 3101282	7.9250	*NA*	S	*NA*
0	113803	53.1000	C123	S	C
0	373450	8.0500	*NA*	S	*NA*
0	330877	8.4583	*NA*	Q	*NA*
0	17463	51.8625	E46	S	E
1	349909	21.0750	*NA*	S	*NA*

R Script Feature Engineering

Adding Title feature

File name = 103 addTitle

```
require(magrittr)
require(purrr)
rm(list = ls()) # clear work space
setwd("d:/temp") # set current work directory
# import data file
dat <- read.csv("TitanicData.csv", na.strings = "")

titleList = c('Mrs', 'Mr', 'Master', 'Miss', 'Major', 'Rev',
              'Dr', 'Ms', 'Mlle','Col', 'Capt', 'Mme', 'Countess',
              'Don', 'Jonkheer')

getTitle <- function(name){
  for(s in titleList)
    if(regexpr(pattern=s, name) != -1)
      return(s)
  return(NA)
}
```

R Script Feature Engineering

Adding Title feature

```
replaceTitles <- function(x){
  title = x['Title']
  sex = x['Sex']
  s = sex$Sex
  t = title$Title
  if(any(t == c('Don', 'Major', 'Capt', 'Jonkheer', 'Rev', 'Col')))
    return('Mr')
  if(any(t == c('Countess', 'Mme')))
    return('Mrs')
  if(any(t == c('Mlle', 'Ms')))
    return('Miss')
  if(t == 'Dr')
    if(s == 'male')
      return('Mr')
  if(s == 'female')
    return('Mrs')
  return(t)
}
```

R Script Feature Engineering

Adding Title feature

```
44 # ----------------- Main program --------------------
45 # Extract title from column Name and create Title column
46 dat$Title <- dat %>% .$Name %>% map(~ getTitle(.x))
47 # Replacing all titles with mr, mrs, miss, master
48 dat$x <- apply(dat[c('Title','Sex')], 1, replaceTitles)
49 print('end')
```

Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title	x
male	29.00	0	0	W./C. 14263	10.5000	NA	S	Mr	Mr
male	22.00	0	0	STON/O 2. 3101275	7.1250	NA	S	Mr	Mr
male	30.00	0	0	2694	7.2250	NA	C	Mr	Mr
male	44.00	2	0	19928	90.0000	C78	Q	Dr	Mr
female	25.00	0	0	347071	7.7750	NA	S	Miss	Mrs
female	24.00	0	2	250649	14.5000	NA	S	Mrs	Mrs
male	37.00	1	1	11751	52.5542	D35	S	Mr	Mr
male	54.00	1	0	244252	26.0000	NA	S	Rev	Mr
male	NA	0	0	362316	7.2500	NA	S	Mr	Mr
female	29.00	1	1	347054	10.4625	G6	S	Mrs	Mrs

R Script Feature Engineering

Feature engineering in data science
https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-data-science-create-features

Source code
https://github.com/laploy/rfe