PYTHON FEATURE ENGINEERING

Python 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

Python 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

Python Feature Engineering

What is Feature Engineering?

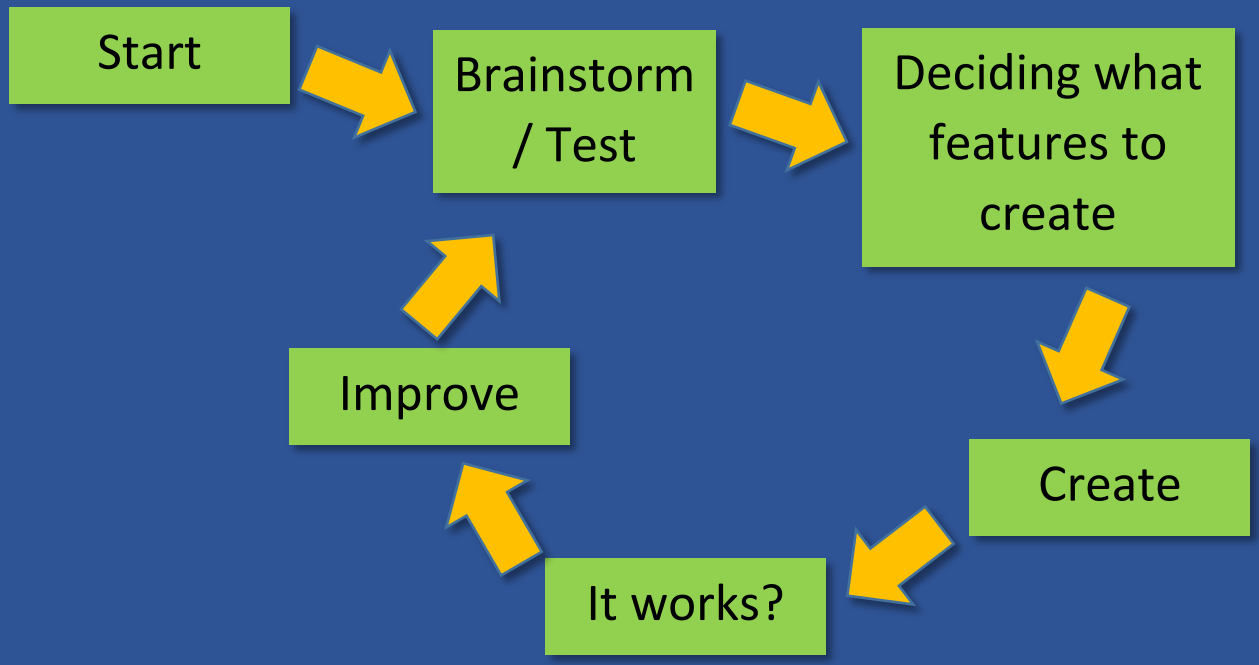
What is Feature Engineering?

- 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
- Is Yak shaving

Python Feature Engineering

The process of feature engineering

The process of feature engineering

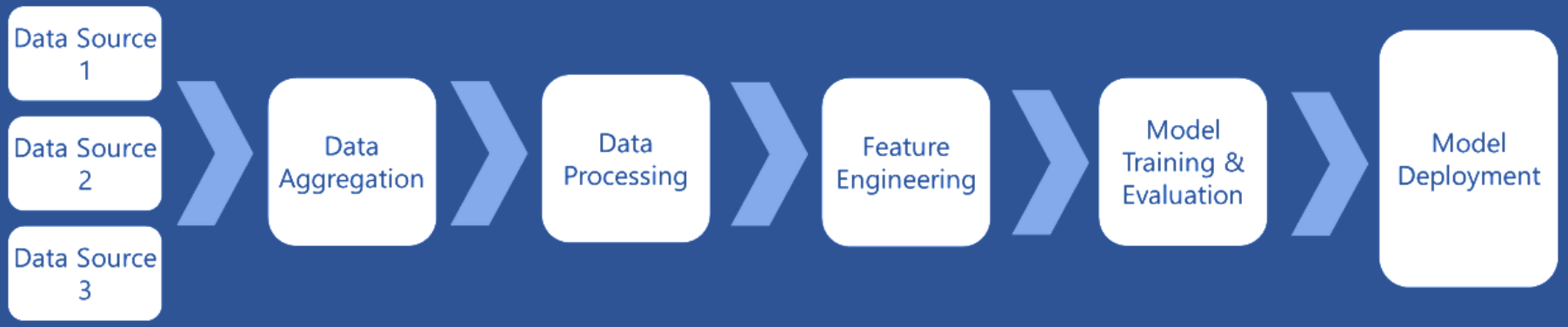


Python Feature Engineering

Where is FE in ML?

Where is FE in ML?

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



Python 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 Pycharm
4. Create New project name = c:\temp\fe
5. Right click project / click Add… / New Python file
6. File name = 100 familySize

Python Feature Engineering

Add family size feature

File name = 100 familySize

```
 9      import pandas as pd
10
11      df = pd.read_csv('d:\\temp\TitanicData.csv')
12      print(list(df))
13      # Create Family Size
14      df['Family_Size'] = df['SibSp']+df['Parch']
15      print(df[['SibSp', 'Parch', 'Family_Size']].head(10))
16      df.to_csv("d:\\temp\\output.csv")
17      print("end")
18      print("***")
```

	SibSp	Parch	Family_Size
0	1	0	1
1	1	0	1
2	0	0	0
3	1	0	1
4	0	0	0
5	0	0	0
6	0	0	0
7	3	1	4
8	0	2	2
9	1	0	1

Python Feature Engineering

Add Age*Class and Fare per person feature

File name = 101 ageClass

```
 9    import pandas as pd
10
11    df = pd.read_csv('d:\\temp\TitanicData.csv')
12    print(list(df))
13    # Create age per class
14    df['Age*Class'] = df['Age']*df['Pclass']
15    print(df[['PassengerId', 'Age', 'Age*Class']].head(10))
16    # Create fare per person
17    df['Family_Size'] = df['SibSp']+df['Parch']
18    df['Fare_Per_Person'] = df['Fare']/(df['Family_Size']+1)
19    print(df[['PassengerId', 'Family_Size', 'Fare_Per_Person']].head(10))
20    print("end")
```

	PassengerId	Age	Age*Class
0	1	22.0	66.0
1	2	38.0	38.0
2	3	26.0	78.0
3	4	35.0	35.0
4	5	35.0	105.0
5	6	NaN	NaN
6	7	54.0	54.0
7	8	2.0	6.0
8	9	27.0	81.0
9	10	14.0	28.0

	PassengerId	Family_Size	Fare_Per_Person
0	1	1	3.62500
1	2	1	35.64165
2	3	0	7.92500
3	4	1	26.55000
4	5	0	8.05000
5	6	0	8.45830
6	7	0	51.86250
7	8	4	4.21500
8	9	2	3.71110
9	10	1	15.03540

Python Feature Engineering

Add Deck feature

File name = 102 addDeck

```
 9     import pandas as pd
10
11
12     # function to extract title from name
13     def get_deck(main, sub):
14         if type(main) != str: return float('nan')
15         for s in sub:
16             if main.find(s) != -1:
17                 return s
18         return float('nan')
19
20     cabin_list = ['A', 'B', 'C', 'D', 'E', 'F', 'T', 'G', 'Unknown']
21
22     # Turning cabin number into Deck
23     df = pd.read_csv('d:\\temp\TitanicData.csv')
24     print(list(df))
25     df['Deck'] = df['Cabin'].map(lambda x: get_deck(x, cabin_list))
26     print(df[['PassengerId', 'Cabin', 'Deck']].head(10))
27     print("end")
```

```
   PassengerId Cabin Deck
0            1   NaN  NaN
1            2   C85    C
2            3   NaN  NaN
3            4  C123    C
4            5   NaN  NaN
5            6   NaN  NaN
6            7   E46    E
7            8   NaN  NaN
8            9   NaN  NaN
9           10   NaN  NaN
end
```

Python Feature Engineering

Adding Title feature

File name = 103 addTitle

```
 9    import pandas as pd
10    import numpy as np
11
12    title_list = ['Mrs', 'Mr', 'Master', 'Miss', 'Major', 'Rev',
13                  'Dr', 'Ms', 'Mlle','Col', 'Capt', 'Mme', 'Countess',
14                  'Don', 'Jonkheer']
15
16
17    # function to extract title from name
18    def get_title(main, sub):
19        for s in sub:
20            if main.find(s) != -1:
21                return s
22        return np.nan
```

Python Feature Engineering

Adding Title feature

```
25      # function to  replacing all titles with mr, mrs, miss, master
26      def replace_titles(x):
27          title = x['Title']
28          if title in ['Don', 'Major', 'Capt', 'Jonkheer', 'Rev', 'Col']:
29              return 'Mr'
30          elif title in ['Countess', 'Mme']:
31              return 'Mrs'
32          elif title in ['Mlle', 'Ms']:
33              return 'Miss'
34          elif title == 'Dr':
35              if x['Sex'] == 'Male':
36                  return 'Mr'
37              else:
38                  return 'Mrs'
39          else:
40              return title
```

Python Feature Engineering

Adding Title feature

```
43      # here comes the main program
44      df = pd.read_csv('d:\\temp\TitanicData.csv')
45      print(list(df))
46      df['Title'] = df['Name'].map(lambda x: get_title(x, title_list))
47      df['Title'] = df.apply(replace_titles, axis=1)  # 1 = row
48      print(df[['Name', 'Title']].head(10))
49      print("end")
```

	Name	Title
0	Braund, Mr. Owen Harris	Mr
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	Mrs
2	Heikkinen, Miss. Laina	Miss
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Mrs
4	Allen, Mr. William Henry	Mr
5	Moran, Mr. James	Mr
6	McCarthy, Mr. Timothy J	Mr

Python Feature Engineering

More information

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/fe