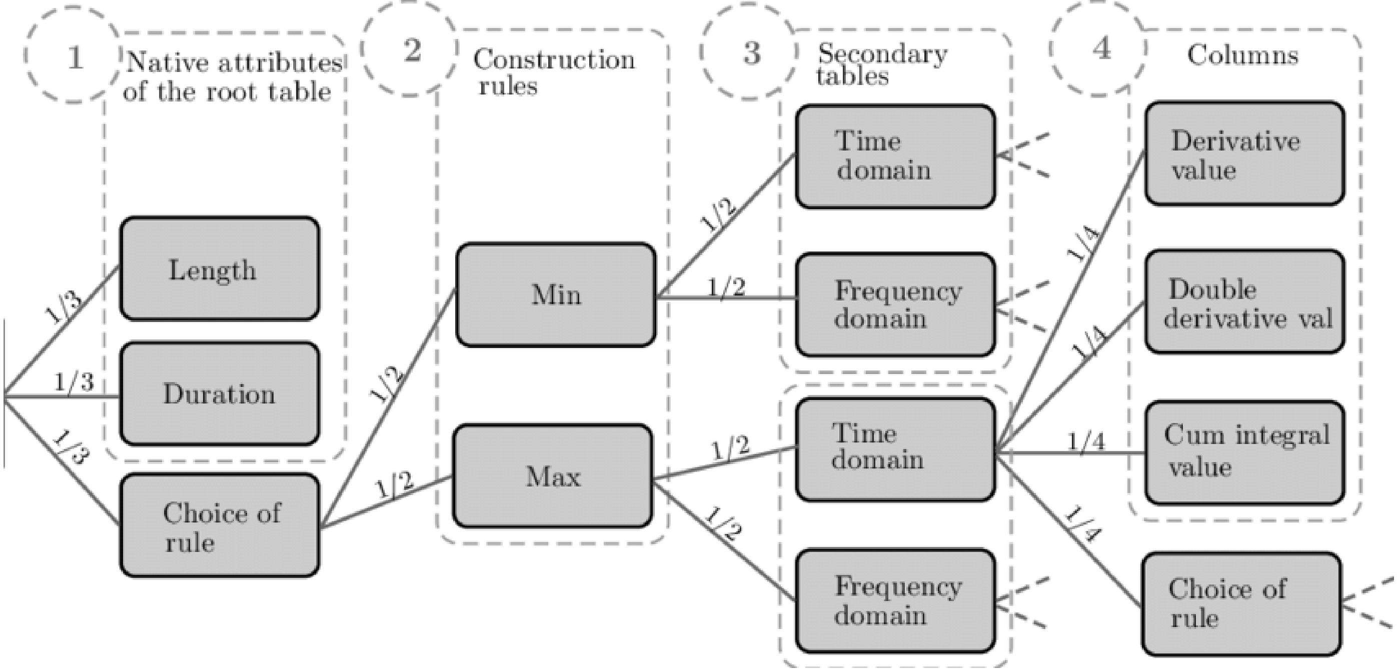
Feature Engineering 101

Topic - 11

Feature Construction



Feature Construction

In [1]:

```
import numpy as np
import pandas as pd

from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression

import seaborn as sns
```

In [2]:

```
df = pd.read_csv('train.csv')[['Age','Pclass','SibSp','Parch','Survived']]
```

In [3]:

```
df.head()
```

Out[3]:

	Age	Pclass	SibSp	Parch	Survived
0	22.0	3	1	0	0
1	38.0	1	1	0	1
2	26.0	3	0	0	1
3	35.0	1	1	0	1
4	35.0	3	0	0	0

In [4]:

```
df.dropna(inplace=True)
```

In [5]:

```
df.head()
```

Out[5]:

	Age	Pclass	SibSp	Parch	Survived
0	22.0	3	1	0	0
1	38.0	1	1	0	1

	Age	Pclass	SibSp	Parch	Survived
2	26.0	3	0	0	1
3	35.0	1	1	0	1
4	35.0	3	0	0	0

In [6]:

```
X = df.iloc[:,0:4]
y = df.iloc[:,-1]
```

In [7]:

```
X.head()
```

Out[7]:

	Age	Pclass	SibSp	Parch
0	22.0	3	1	0
1	38.0	1	1	0
2	26.0	3	0	0
3	35.0	1	1	0
4	35.0	3	0	0

In [8]:

```
np.mean(cross_val_score(LogisticRegression(),X,y,scoring='accuracy',cv=20))
```

Out[8]:

```
0.6933333333333332
```

Applying Feature Construction

In [9]:

```
X['Family_size'] = X['SibSp'] + X['Parch'] + 1
```

In [10]:

```
X.head()
```

Out[10]:

	Age	Pclass	SibSp	Parch	Family_size
0	22.0	3	1	0	2
1	38.0	1	1	0	2
2	26.0	3	0	0	1
3	35.0	1	1	0	2
4	35.0	3	0	0	1

In [11]:

```
def myfunc(num):
    if num == 1:
        #alone
        return 0
    elif num >1 and num <=4:
        # small family
        return 1
    else:
        # large family
        return 2
```

In [12]:

```
myfunc(4)
```

Out[12]:

```
1
```

In [13]:

```
X['Family_type'] = X['Family_size'].apply(myfunc)
```

In [14]:

```
X.head()
```

Out[14]:

	Age	Pclass	SibSp	Parch	Family_size	Family_type
0	22.0	3	1	0	2	1
1	38.0	1	1	0	2	1
2	26.0	3	0	0	1	0
3	35.0	1	1	0	2	1
4	35.0	3	0	0	1	0

In [15]:

```
X.drop(columns=['SibSp','Parch','Family_size'],inplace=True)
```

In [16]:

```
X.head()
```

Out[16]:

	Age	Pclass	Family_type
0	22.0	3	1
1	38.0	1	1
2	26.0	3	0
3	35.0	1	1
4	35.0	3	0

In [17]:

```
np.mean(cross_val_score(LogisticRegression(),X,y,scoring='accuracy',cv=20))
```

Out[17]:

```
0.7003174603174602
```