Feature Engineering 101

Topic - 12

Feature Splitting

Feature Splitting

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')
```

In [3]:

```
df.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [4]:

```
df['Name']
```

Out[4]:

```
0                                Braund, Mr. Owen Harris
1      Cumings, Mrs. John Bradley (Florence Briggs Th...
2                                 Heikkinen, Miss. Laina
3           Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                               Allen, Mr. William Henry
                             ...
886                                Montvila, Rev. Juozas
887                         Graham, Miss. Margaret Edith
888             Johnston, Miss. Catherine Helen "Carrie"
889                                Behr, Mr. Karl Howell
890                                  Dooley, Mr. Patrick
Name: Name, Length: 891, dtype: object
```

In [5]:

```
df['Name'].str.split(', ', expand=True)[1].str.split('.', expand=True)[0]
```

Out[5]:

```
0         Mr
1        Mrs
2       Miss
3        Mrs
4         Mr
       ...
886      Rev
887     Miss
888     Miss
889       Mr
```

```
890       Mr
Name: 0, Length: 891, dtype: object
```

In [6]:

```python
df['Title'] = df['Name'].str.split(', ', expand=True)[1].str.split('.', expand=True)[0]
```

In [7]:

```python
df['Name'].str.split(', ', expand=True)[1].str.split('.', expand=True)[0]
```

Out[7]:

```
0         Mr
1        Mrs
2       Miss
3        Mrs
4         Mr
       ...
886      Rev
887     Miss
888     Miss
889       Mr
890       Mr
Name: 0, Length: 891, dtype: object
```

In [8]:

```python
df[['Title','Name']]
```

Out[8]:

	Title	Name
0	Mr	Braund, Mr. Owen Harris
1	Mrs	Cumings, Mrs. John Bradley (Florence Briggs Th...
2	Miss	Heikkinen, Miss. Laina
3	Mrs	Futrelle, Mrs. Jacques Heath (Lily May Peel)
4	Mr	Allen, Mr. William Henry
...	...	...
886	Rev	Montvila, Rev. Juozas
887	Miss	Graham, Miss. Margaret Edith
888	Miss	Johnston, Miss. Catherine Helen "Carrie"
889	Mr	Behr, Mr. Karl Howell
890	Mr	Dooley, Mr. Patrick

891 rows × 2 columns

In [9]:

```python
(df.groupby('Title').mean()['Survived']).sort_values(ascending=False)
```

Out[9]:

```
Title
the Countess    1.000000
Mlle            1.000000
Sir             1.000000
Ms              1.000000
Lady            1.000000
Mme             1.000000
Mrs             0.792000
Miss            0.697802
Master          0.575000
Col             0.500000
Major           0.500000
Dr              0.428571
```

```
Mr                0.156673
Jonkheer          0.000000
Rev               0.000000
Don               0.000000
Capt              0.000000
Name: Survived, dtype: float64
```

In [10]:

```
df['Is_Married'] = 0
df['Is_Married'].loc[df['Title'] == 'Mrs'] = 1
```

```
C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\indexing.py:1732: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_single_block(indexer, value, name)
```

In [11]:

```
df['Is_Married']
```

Out[11]:

```
0      0
1      1
2      0
3      1
4      0
      ..
886    0
887    0
888    0
889    0
890    0
Name: Is_Married, Length: 891, dtype: int64
```

In [ ]: