Step５　特徴量エンジニアリング

アイデア５：名前の敬称

16T4056S　溝口賢治

2019/4/18

- Name（名前）の乗客乗員の文字列データには名前に付随して敬称もデータとして入っている[Mr, Ms, Dr, Capt, etc.]

- データセットのNameから敬称のみを摘出する

  Train_set[‘Title’] = train_set.Name.str.extract(r’ ,¥s*([^¥.]*)¥s*¥.’, expand=False)

  Pandasのstr.extractを使うと、特定の文字列を正規表現で抽出できる

- Nameから継承の文字列のみ摘出して新たにTitleとしてデータフレームに格納する

Train_set['Title'].unique()

Pandasのunique()はユニークな要素の値をリストで戻す

- 摘出した敬称の特色を考慮して、下記の５つのグループに分類する

G1（Mr）--- Mr

G2（Crew1）--- Don, Rev, Capt

G3（Crew2）--- Major, Col, Dr

G4（Women-Masters）--- Mrs, Miss, Master

G5（Affluence）--- Mme, Ms, Lady, Sir, Mile, the Countess, Johkheer

- 敬称をグループごとにリストへまとめるコード

```
Mr = ['Mr']
Crew1 = ['Don', 'Rev', 'Capt']
Crew2 = ['Major', 'Col', 'Dr']
Women_Masters = ['Mrs', 'Miss', 'Master']
Affluence = ['Mme', 'Ms', 'Lady', 'Sir', 'Mile', 'the Countess',
                Jokheer]
```

- Nameから抽出したTitleをグループ分けして、Title_Groupとしてデータフレームへ追加するコード

```
train_set[‘Title_Group’] = np.where(train_set[‘Title’] == Mr[0], ‘Mr’, ‘Affluence’)
train_set[‘Title_Group’] = np.where(train_set[‘Title’].isin(Crew1),  ‘Crew1’, train_set[‘Title_Group’])
train_set[‘Title_Group’] = np.where(train_set[‘Title’].isin(Crew2),  ‘Crew2’, train_set[‘Title_Group’])
train_set[‘Title_Group’] = np.where(train_set[‘Title’].isin(Women_Masters),  ‘Women_Masters’, train_set[‘Title_Group’])
train_set[‘Title_Group’] = np.where(train_set[‘Title’].isin(Affluence),  ‘Affluence’, train_set[‘Title_Group’])
```

- 敬称の各グループのターゲット（Survived）の平均値を表示するコード

```
Titleplot = train_set['Survived'].grouphy(train_set['Title_Group']).mean()
Titleplot
```

- 各グループのターゲット（Survived）の平均値

```
Title_Gruop
Affluence        0.875000
Crew1            0.000000
Cre2             0.454545
Mr               0.156673
Woman_Masters    0.717579
Nome: Survived, dtype: float64
```