feature-construction-and-spliting

May 26, 2025

0.1 Feature Construction ( )

Feature Construction

0.1.1 Feature Construction :

1. **Polynomial Features ( )**
   : `Age`, `Age`^{2}, `Age * Fare`
2. **Binning ( )** continuous :
   - 0–12 → "Child"
   - 13–59 → "Adult"
   - 60+ → "Senior"
3. **Interaction Features** (multiplication) : `Income * Education_Level`
4. **Datetime Features** , : year, month, day :

   ```
   df['Year'] = pd.to_datetime(df['Date']).dt.year
   ```

5. **Text** , CountVectorizer TF-IDF :

   ```
   from sklearn.feature_extraction.text import TfidfVectorizer
   ```

0.2 Feature Splitting ( )

Feature Splitting

0.2.1 :

1. **Full Name → First Name + Last Name**

   ```
   df['First_Name'] = df['Full_Name'].str.split().str[0]
   df['Last_Name'] = df['Full_Name'].str.split().str[1]
   ```

2. **Date → Day, Month, Year**

   ```
   df['Day'] = pd.to_datetime(df['Date']).dt.day
   df['Month'] = pd.to_datetime(df['Date']).dt.month
   df['Year'] = pd.to_datetime(df['Date']).dt.year
   ```

3. **Address → City, State, Zip Code**

4. **Name+Gender –> Maritaile Status**
5. **Marks,Study Hours –> IQ**

:

- **Feature Construction =**
- **Feature Splitting =**

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[ ]:
```
df=pd.read_csv('/content/Titanic-Dataset.csv')
df.head(3)
```

[ ]:
```
   PassengerId  Survived  Pclass  …     Fare Cabin  Embarked
0            1         0       3  …   7.2500   NaN         S
1            2         1       1  …  71.2833   C85         C
2            3         1       3  …   7.9250   NaN         S

[3 rows x 12 columns]
```

[ ]:
```
df=df.iloc[:,[3,5,4,6,7,2,1]]
df.head(3)
```

[ ]:
```
                                                Name   Age  … Pclass  Survived
0                            Braund, Mr. Owen Harris  22.0  …      3         0
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  38.0  …      1         1
2                             Heikkinen, Miss. Laina  26.0  …      3         1

[3 rows x 7 columns]
```

##Train model Without Feature Construction or Spliting

[ ]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(df.
  ↪drop(['Name','Survived'],axis=1),df['Survived'],test_size=0.2,random_state=2)
X_train.head(2)
```

[ ]:
```
     Age     Sex  SibSp  Parch  Pclass
30  40.0    male      0      0       1
10   4.0  female      1      1       3
```

[ ]:
```
from sklearn.preprocessing import OneHotEncoder
ohe=OneHotEncoder(sparse_output=False,dtype=np.int32)
X_train_age=ohe.fit_transform(X_train[['Sex']])
```

```
X_train_age=pd.DataFrame(X_train_age,columns=['sex_male','sex_female'])

X_test_age=pd.DataFrame(ohe.
  ↪transform(X_test[['Sex']]),columns=(['sex_male','sex_female']))
X_test_age,X_train_age
```

[ ]:
```
(     sex_male  sex_female
 0           0           1
 1           0           1
 2           1           0
 3           0           1
 4           1           0
 ..        …         …
 174         0           1
 175         0           1
 176         0           1
 177         0           1
 178         1           0

 [179 rows x 2 columns],
      sex_male  sex_female
 0           0           1
 1           1           0
 2           0           1
 3           0           1
 4           0           1
 ..        …         …
 707         1           0
 708         0           1
 709         0           1
 710         0           1
 711         0           1

 [712 rows x 2 columns])
```

[ ]:
```
new_X_train=pd.concat([X_train.reset_index(),X_train_age.reset_index()],axis=1).
  ↪drop('Sex',axis=1)
new_X_train
```

[ ]:
```
     index   Age  SibSp  Parch  Pclass  index  sex_male  sex_female
0       30  40.0      0      0       1      0         0           1
1       10   4.0      1      1       3      1         1           0
2      873  47.0      0      0       3      2         0           1
3      182   9.0      4      2       3      3         0           1
4      876  20.0      0      0       3      4         0           1
..     …    …    …     …      …     …        …         …
```

```
707    534  30.0      0      0       3    707         1           0
708    584   NaN      0      0       3    708         0           1
709    493  71.0      0      0       1    709         0           1
710    527   NaN      0      0       1    710         0           1
711    168   NaN      0      0       1    711         0           1

[712 rows x 8 columns]
```

[ ]:
```
new_df=np.hstack([X_test,X_test_age])
new_df
#return Numpy Array
```

[ ]:
```
array([[42.0, 'male', 0, …, 1, 0, 1],
       [21.0, 'male', 0, …, 3, 0, 1],
       [24.0, 'female', 1, …, 2, 1, 0],
       …,
       [nan, 'male', 8, …, 3, 0, 1],
       [26.0, 'male', 0, …, 3, 0, 1],
       [29.0, 'female', 1, …, 3, 1, 0]], dtype=object)
```

[ ]:
```
new_X_test=pd.concat([X_test.reset_index(),X_test_age.reset_index()],axis=1).
  ↪drop('Sex',axis=1)
new_X_test
```

[ ]:
```
     index   Age  SibSp  Parch  Pclass  index  sex_male  sex_female
0      707  42.0      0      0       1      0         0           1
1       37  21.0      0      0       3      1         0           1
2      615  24.0      1      2       2      2         1           0
3      169  28.0      0      0       3      3         0           1
4       68  17.0      4      2       3      4         1           0
..     …    …    …     …     …      …         …
174     89  24.0      0      0       3    174         0           1
175     80  22.0      0      0       3    175         0           1
176    846   NaN      8      2       3    176         0           1
177    870  26.0      0      0       3    177         0           1
178    251  29.0      1      1       3    178         1           0

[179 rows x 8 columns]
```

[ ]:
```
from sklearn.tree import DecisionTreeClassifier
model=DecisionTreeClassifier()
model.fit(new_X_train,y_train)
```

[ ]:
```
DecisionTreeClassifier()
```

[ ]:
```
y_pred=model.predict(new_X_test)
y_pred
```

[ ]:
```
array([0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0,
       0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1,
       1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0,
       0, 0, 0])
```

[ ]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_pred,y_test)
```

[ ]:
```
0.7206703910614525
```

[ ]:
```
from sklearn.model_selection import cross_val_score
cross_val_score(model,new_X_train,y_train,cv=5,scoring='accuracy').mean()
```

[ ]:
```
np.float64(0.6926425686989066)
```

[ ]:

##With Feature Construction

[ ]:
```
new_X_test
```

[ ]:

	index	Age	SibSp	Parch	Pclass	index	sex_male	sex_female
0	707	42.0	0	0	1	0	0	1
1	37	21.0	0	0	3	1	0	1
2	615	24.0	1	2	2	2	1	0
3	169	28.0	0	0	3	3	0	1
4	68	17.0	4	2	3	4	1	0
..	…	…	…	…	…	…	…	…
174	89	24.0	0	0	3	174	0	1
175	80	22.0	0	0	3	175	0	1
176	846	NaN	8	2	3	176	0	1
177	870	26.0	0	0	3	177	0	1
178	251	29.0	1	1	3	178	1	0

[179 rows x 8 columns]

[ ]:
```
new_X_train['family']=new_X_train['SibSp']+new_X_train['Parch']
new_X_test['family']=new_X_test['SibSp']+new_X_test['Parch']
```

[ ]:
```
new_X_train
```

[ ]:

	index	Age	SibSp	Parch	Pclass	index	sex_male	sex_female	family
0	30	40.0	0	0	1	0	0	1	0
1	10	4.0	1	1	3	1	1	0	2
2	873	47.0	0	0	3	2	0	1	0
3	182	9.0	4	2	3	3	0	1	6
4	876	20.0	0	0	3	4	0	1	0
..	…	…	…	…	…	…	…	…	…
707	534	30.0	0	0	3	707	1	0	0
708	584	NaN	0	0	3	708	0	1	0
709	493	71.0	0	0	1	709	0	1	0
710	527	NaN	0	0	1	710	0	1	0
711	168	NaN	0	0	1	711	0	1	0

[712 rows x 9 columns]

[ ]:
```
new_X_train.drop(['SibSp','Parch'],axis=1,inplace=True)
new_X_test.drop(['SibSp','Parch'],axis=1,inplace=True)
```

[ ]:
```
new_X_test,new_X_train
```

[ ]:

	index	Age	Pclass	index	sex_male	sex_female	family
0	707	42.0	1	0	0	1	0
1	37	21.0	3	1	0	1	0
2	615	24.0	2	2	1	0	3
3	169	28.0	3	3	0	1	0
4	68	17.0	3	4	1	0	6
..	…	…	…	…	…	…	…
174	89	24.0	3	174	0	1	0
175	80	22.0	3	175	0	1	0
176	846	NaN	3	176	0	1	10
177	870	26.0	3	177	0	1	0
178	251	29.0	3	178	1	0	2

[179 rows x 7 columns],

	index	Age	Pclass	index	sex_male	sex_female	family
0	30	40.0	1	0	0	1	0
1	10	4.0	3	1	1	0	2
2	873	47.0	3	2	0	1	0
3	182	9.0	3	3	0	1	6
4	876	20.0	3	4	0	1	0
..	…	…	…	…	…	…	…
707	534	30.0	3	707	1	0	0
708	584	NaN	3	708	0	1	0
709	493	71.0	1	709	0	1	0
710	527	NaN	1	710	0	1	0
711	168	NaN	1	711	0	1	0

```
[712 rows x 7 columns])
```

[ ]:
```
model2=DecisionTreeClassifier()
model2.fit(new_X_train,y_train)
```

[ ]:
```
DecisionTreeClassifier()
```

[ ]:
```
y_pred2=model2.predict(new_X_test)
y_pred2
```

[ ]:
```
array([0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0,
       0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0,
       0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0,
       0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1,
       1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0,
       0, 1, 0])
```

[ ]:
```
accuracy_score(y_pred2,y_test)
```

[ ]:
```
0.7206703910614525
```

[ ]:
```
cross_val_score(model2,new_X_train,y_train,cv=10,scoring='accuracy').mean()
```

[ ]:
```
np.float64(0.6800078247261346)
```

[ ]:
```

```

##More feature Extraction

[ ]:
```
df.head()
```

[ ]:
```
                                                Name   Age  … Pclass  Survived
0                            Braund, Mr. Owen Harris  22.0  …      3         0
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  38.0  …      1         1
2                             Heikkinen, Miss. Laina  26.0  …      3         1
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  35.0  …      1         1
4                           Allen, Mr. William Henry  35.0  …      3         0

[5 rows x 7 columns]
```

[ ]:
```
df['nickname']=df['Name'].str.split(',').str[0]
df.head()
```

[ ]:
```
                                                Name   Age  … Survived
nickname
0                              Braund, Mr. Owen Harris  22.0  …        0
Braund
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  38.0  …        1
Cumings
2                               Heikkinen, Miss. Laina  26.0  …        1
Heikkinen
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  35.0  …        1
Futrelle
4                            Allen, Mr. William Henry  35.0  …        0
Allen

[5 rows x 8 columns]
```

[ ]:
```
df['name_extension']=df['Name'].str.split(',').str[1].str.split('.').str[0].str.
↪strip()
df.head()
```

[ ]:
```
                                                Name  …  name_extension
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 rows x 9 columns]
```

[ ]:
```
df['name_extension'].unique()
#ata akta important feature hote pare (Catagorical Column)
```

[ ]:
```
array(['Mr', 'Mrs', 'Miss', 'Master', 'Don', 'Rev', 'Dr', 'Mme', 'Ms',
       'Major', 'Lady', 'Sir', 'Mlle', 'Col', 'Capt', 'the Countess',
       'Jonkheer'], dtype=object)
```

[ ]:
```
df.groupby('name_extension').get_group('Miss')
#unmarid feamile
```

[ ]:
```
                                      Name  …  name_extension
2                     Heikkinen, Miss. Laina  …            Miss
10            Sandstrom, Miss. Marguerite Rut  …            Miss
11                   Bonnell, Miss. Elizabeth  …            Miss
14       Vestrom, Miss. Hulda Amanda Adolfina  …            Miss
22                 McGowan, Miss. Anna "Annie"  …            Miss
..                                         …  …             …
866              Duran y More, Miss. Asuncion  …            Miss
875          Najib, Miss. Adele Kiamie "Jane"  …            Miss
```

```
882                Dahlberg, Miss. Gerda Ulrika  …            Miss
887                Graham, Miss. Margaret Edith  …            Miss
888  Johnston, Miss. Catherine Helen "Carrie"  …            Miss

[182 rows x 9 columns]
```

[ ]: