```python
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

```python
movie  = ['movie_id','movie_name','genre']
rating = ['user_id','movie_id','rating','time_stamp']
user = ['user_id','gender','age','occupation','zip_code']
```

```python
df_movies = 
pd.read_csv('C:/Users/Lenovo/Downloads/Compressed/movies.dat',header=N
one,delimiter ='::',names=movie,encoding='windows-1251')
df_ratings = 
pd.read_csv('C:/Users/Lenovo/Downloads/Compressed/ratings.dat',header=
None,delimiter ='::',names=rating,encoding='windows-1251')
df_users = 
pd.read_csv('C:/Users/Lenovo/Downloads/Compressed/users.dat',header=No
ne,delimiter ='::',names=user,encoding='windows-1251')
```

```python
df_movies.head(3)
```

```
   movie_id               movie_name                         genre
0         1         Toy Story (1995)   Animation|Children's|Comedy
1         2           Jumanji (1995)  Adventure|Children's|Fantasy
2         3  Grumpier Old Men (1995)                Comedy|Romance
```

```python
df_ratings.head(3)
```

```
   user_id  movie_id  rating  time_stamp
0        1      1193       5   978300760
1        1       661       3   978302109
2        1       914       3   978301968
```

```python
df_users.head(3)
```

```
   user_id gender  age  occupation zip_code
0        1      F    1          10    48067
1        2      M   56          16    70072
2        3      M   25          15    55117
```

```python
df_movies.head(3)
```

```
   movie_id               movie_name                         genre
0         1         Toy Story (1995)   Animation|Children's|Comedy
1         2           Jumanji (1995)  Adventure|Children's|Fantasy
2         3  Grumpier Old Men (1995)                Comedy|Romance
```

```python
# merging 3 dataframes
df1 = df_movies.merge(df_ratings,how='outer',on='movie_id')
df = df1.merge(df_users,how='outer',on='user_id')
```

```python
df
```

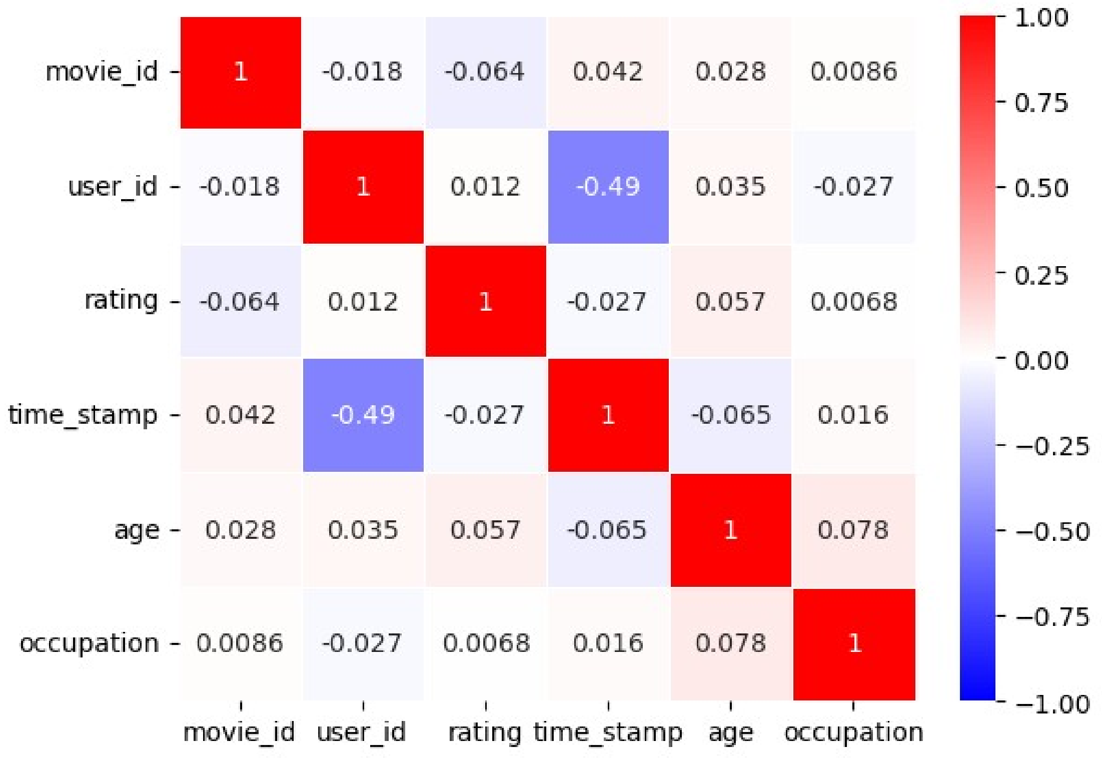
```
         movie_id                                movie_name  \
0               1                          Toy Story (1995)   
1              48                         Pocahontas (1995)   
2             150                          Apollo 13 (1995)   
3             260  Star Wars: Episode IV - A New Hope (1977)   
4             527                   Schindler's List (1993)   
...           ...                                       ...   
1000381      3513               Rules of Engagement (2000)   
1000382      3535                   American Psycho (2000)   
1000383      3536                 Keeping the Faith (2000)   
1000384      3555                             U-571 (2000)   
1000385      3578                         Gladiator (2000)   

                                genre  user_id  rating  
time_stamp  \
0            Animation|Children's|Comedy      1.0     5.0  
978824268.0   
1       Animation|Children's|Musical|Romance      1.0     5.0  
978824351.0   
2                                  Drama      1.0     5.0  
978301777.0   
3          Action|Adventure|Fantasy|Sci-Fi      1.0     4.0  
978300760.0   
4                              Drama|War      1.0     5.0  
978824195.0   
...                                  ...      ...     ...  
...   
1000381                   Drama|Thriller   5727.0     4.0  
958489970.0   
1000382             Comedy|Horror|Thriller   5727.0     2.0  
958489970.0   
1000383                   Comedy|Romance   5727.0     5.0  
958489902.0   
1000384                  Action|Thriller   5727.0     3.0  
958490699.0   
1000385                     Action|Drama   5727.0     5.0  
958490171.0   

        gender   age  occupation zip_code  
0            F   1.0        10.0    48067  
1            F   1.0        10.0    48067  
2            F   1.0        10.0    48067  
3            F   1.0        10.0    48067  
4            F   1.0        10.0    48067  
...        ...   ...         ...      ...  
1000381      M  25.0         4.0    92843  
1000382      M  25.0         4.0    92843  
1000383      M  25.0         4.0    92843  
1000384      M  25.0         4.0    92843  
1000385      M  25.0         4.0    92843  
```

```
[1000386 rows x 10 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1000386 entries, 0 to 1000385
Data columns (total 10 columns):
 #   Column      Non-Null Count    Dtype  
---  ------      --------------    -----  
 0   movie_id    1000386 non-null  int64  
 1   movie_name  1000386 non-null  object 
 2   genre       1000386 non-null  object 
 3   user_id     1000209 non-null  float64
 4   rating      1000209 non-null  float64
 5   time_stamp  1000209 non-null  float64
 6   gender      1000209 non-null  object 
 7   age         1000209 non-null  float64
 8   occupation  1000209 non-null  float64
 9   zip_code    1000209 non-null  object 
dtypes: float64(5), int64(1), object(4)
memory usage: 84.0+ MB
```

```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.heatmap(df.corr(),annot=True,cmap='bwr',vmin=-
1,vmax=+1,linewidth=0.5)
plt.show()
```



# top 25 movies by viewership

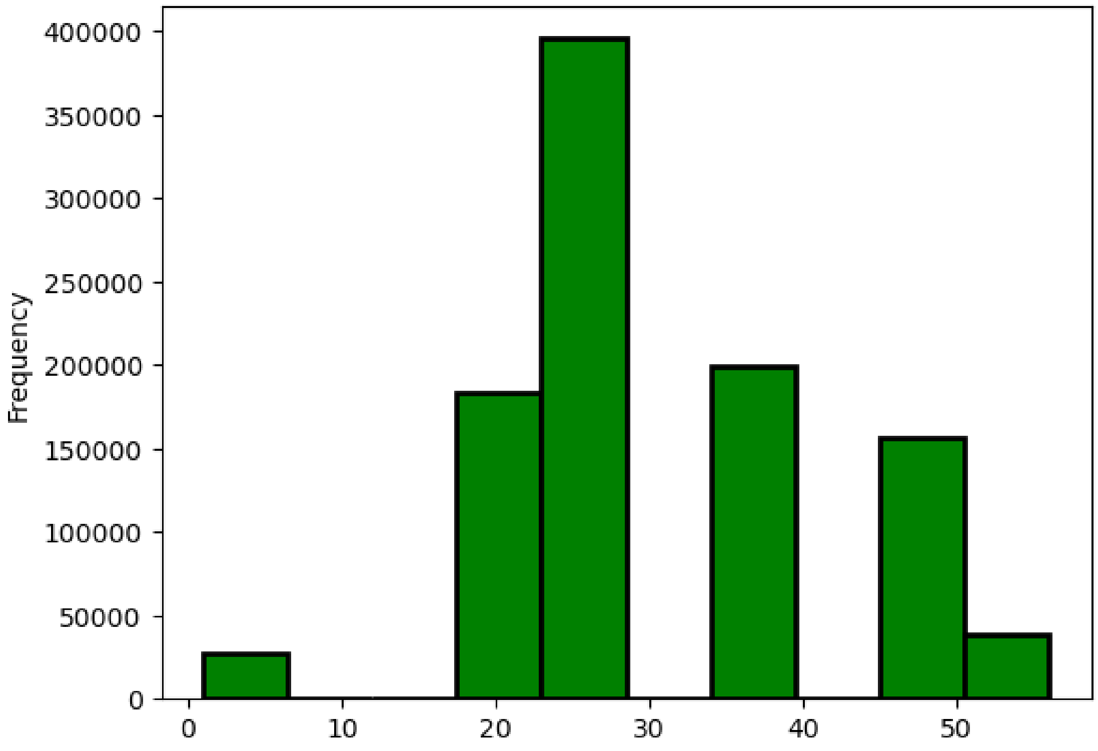
```
df.groupby('movie_name')
[['user_id']].nunique().sort_values(by='user_id',ascending=False)[:25]
```

```
                                                    user_id
movie_name
American Beauty (1999)                                 3428
Star Wars: Episode IV - A New Hope (1977)              2991
Star Wars: Episode V - The Empire Strikes Back ...     2990
Star Wars: Episode VI - Return of the Jedi (1983)      2883
Jurassic Park (1993)                                   2672
Saving Private Ryan (1998)                             2653
Terminator 2: Judgment Day (1991)                      2649
Matrix, The (1999)                                     2590
Back to the Future (1985)                              2583
Silence of the Lambs, The (1991)                       2578
Men in Black (1997)                                    2538
Raiders of the Lost Ark (1981)                         2514
Fargo (1996)                                           2513
Sixth Sense, The (1999)                                2459
Braveheart (1995)                                      2443
Shakespeare in Love (1998)                             2369
Princess Bride, The (1987)                             2318
Schindler's List (1993)                                2304
L.A. Confidential (1997)                               2288
Groundhog Day (1993)                                   2278
```

```
E.T. the Extra-Terrestrial (1982)                          2269
Star Wars: Episode I - The Phantom Menace (1999)           2250
Being John Malkovich (1999)                                2241
Shawshank Redemption, The (1994)                           2227
Godfather, The (1972)                                      2223
```

user age distribution

```
df['age'].plot(kind='hist',color='g',edgecolor='k',linewidth=2)
plt.show()
```

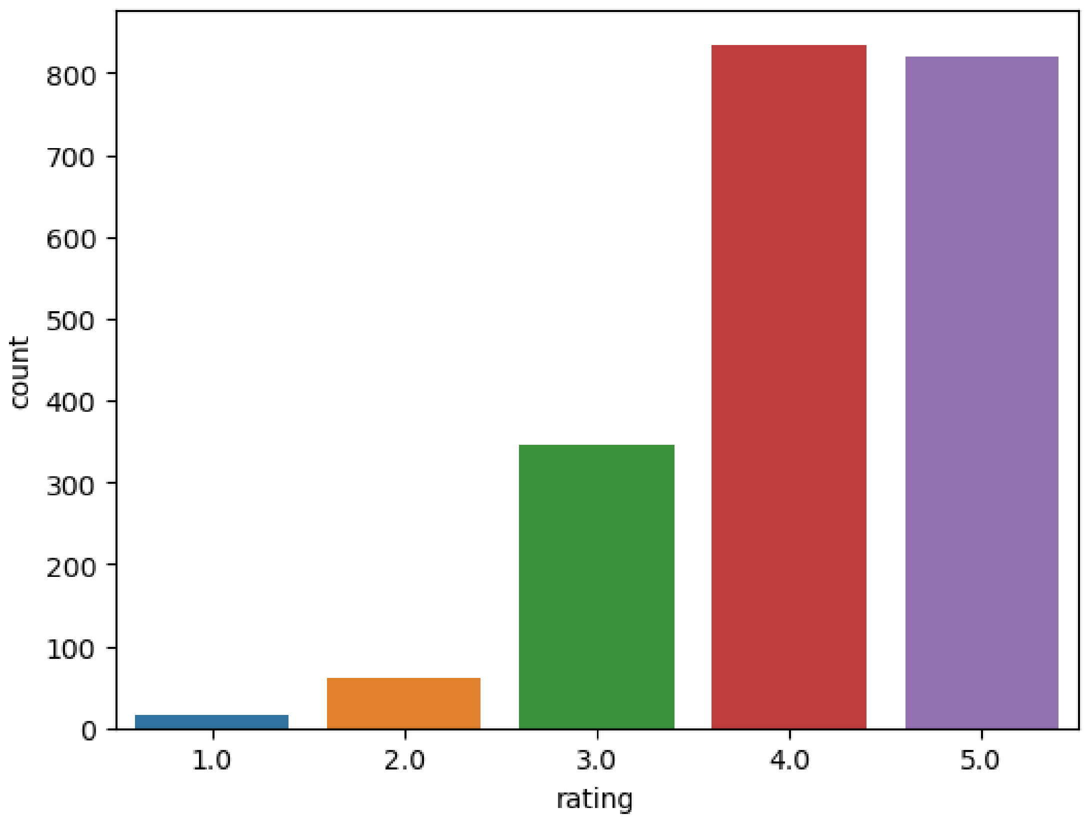


user ratings for movie named : Toy Story

```
zx = df[df['movie_name']=='Toy Story (1995)']['rating'].value_counts()
print(zx)
sns.countplot(df[df['movie_name']=='Toy Story (1995)']['rating'])
```

```
4.0    835
5.0    820
3.0    345
2.0     61
1.0     16
Name: rating, dtype: int64
```

```
<AxesSubplot:xlabel='rating', ylabel='count'>
```



ratings for all movies by user id 2696

```
df[df['user_id']==2696]
[['movie_name','rating']].sort_values(by='rating',ascending=False).T

                      991213                        991222
991224  \
movie_name  Lone Star (1996)  Devil's Advocate, The (1997)  Palmetto
(1998)
rating                   5.0                           4.0
4.0

                           991214                            991230  \
movie_name  Basic Instinct (1992)  Talented Mr. Ripley, The (1999)
rating                        4.0                              4.0

                         991216         991228
991226  \
movie_name  Shining, The (1980)  Psycho (1998)  Perfect Murder, A
(1998)
rating                      4.0            4.0
4.0

                              991219           991220
```

```
991225  \
movie_name  L.A. Confidential (1997)  Game, The (1997)  Wild Things
(1998)
rating                           4.0               4.0
4.0

                                                    991223  \
movie_name  Midnight in the Garden of Good and Evil (1997)
rating                                                 4.0

                        991212           991218  \
movie_name  Client, The (1994)  Cop Land (1997)
rating                     3.0              3.0

                                       991215  \
movie_name  E.T. the Extra-Terrestrial (1982)
rating                                   3.0

                                            991221  \
movie_name  I Know What You Did Last Summer (1997)
rating                                        2.0

                                                  991227  \
movie_name  I Still Know What You Did Last Summer (1998)
rating                                              2.0

                               991217            991229      991231

movie_name  Back to the Future (1985)  Lake Placid (1999)  JFK (1991)

rating                            2.0                 1.0         1.0
```

creating a profile report

```
import pandas_profiling as pf
profile = pf.ProfileReport(df)
profile.to_file('movie lens profile.html')
```

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

```
movie_id      0
movie_name    0
genre         0
user_id       0
rating        0
time_stamp    0
gender        0
```

```
age           0
occupation    0
zip_code      0
dtype: int64
```

unique genres

```
gg = df['genre'].tolist()

uniq_genre = set()
for i in gg:
    dd = i.split('|')
    for j in dd:
        uniq_genre.add(j)

print(list(uniq_genre))

# df.genre.str.get_dummies().columns
```

```
['Sci-Fi', 'Horror', 'Mystery', 'Film-Noir', 'Documentary', 'Crime',
'Animation', 'Comedy', 'Action', 'Adventure', 'Fantasy', 'Thriller',
'Western', 'War', "Children's", 'Musical', 'Romance', 'Drama']
```

```
df = pd.concat([df,df.genre.str.get_dummies()],axis=1)

df.head()
```

```
   movie_id                                  movie_name  \
0         1                            Toy Story (1995)
1        48                           Pocahontas (1995)
2       150                            Apollo 13 (1995)
3       260  Star Wars: Episode IV - A New Hope (1977)
4       527                     Schindler's List (1993)

                                 genre  user_id  rating   time_stamp
gender  \
0          Animation|Children's|Comedy      1.0     5.0  978824268.0
F
1  Animation|Children's|Musical|Romance      1.0     5.0  978824351.0
F
2                                Drama      1.0     5.0  978301777.0
F
3       Action|Adventure|Fantasy|Sci-Fi      1.0     4.0  978300760.0
F
4                            Drama|War      1.0     5.0  978824195.0
F

   age  occupation zip_code  ...  Fantasy  Film-Noir  Horror  Musical
\
0  1.0        10.0    48067  ...        0          0       0        0
```

```
1  1.0        10.0    48067  ...        0          0       0        1

2  1.0        10.0    48067  ...        0          0       0        0

3  1.0        10.0    48067  ...        1          0       0        0

4  1.0        10.0    48067  ...        0          0       0        0


   Mystery  Romance  Sci-Fi  Thriller  War  Western
0        0        0       0         0    0        0
1        0        1       0         0    0        0
2        0        0       0         0    0        0
3        0        0       1         0    0        0
4        0        0       0         0    1        0

[5 rows x 28 columns]
```

```
df.drop(['movie_name','zip_code','time_stamp','genre'],axis=1,inplace=
True)
df.head()
```

```
   movie_id  user_id  rating gender  age  occupation  Action 
Adventure  \
0         1      1.0     5.0      F  1.0        10.0       0
0
1        48      1.0     5.0      F  1.0        10.0       0
0
2       150      1.0     5.0      F  1.0        10.0       0
0
3       260      1.0     4.0      F  1.0        10.0       1
1
4       527      1.0     5.0      F  1.0        10.0       0
0

   Animation  Children's  ...  Fantasy  Film-Noir  Horror  Musical 
Mystery  \
0          1           1  ...        0          0       0        0
0
1          1           1  ...        0          0       0        1
0
2          0           0  ...        0          0       0        0
0
3          0           0  ...        1          0       0        0
0
4          0           0  ...        0          0       0        0
0

   Romance  Sci-Fi  Thriller  War  Western
```

```
0        0       0         0    0        0
1        1       0         0    0        0
2        0       0         0    0        0
3        0       1         0    0        0
4        0       0         0    1        0

[5 rows x 24 columns]
```

```
# df = pd.get_dummies(columns=['gender'],data=df,drop_first=True)
df.gender = pd.get_dummies(df.gender,drop_first=True)

df = pd.get_dummies(columns=['occupation'],drop_first=True,data=df)

df
```

```
         movie_id  user_id  rating  gender   age  Action  Adventure  \
0               1      1.0     5.0       0   1.0       0          0
1              48      1.0     5.0       0   1.0       0          0
2             150      1.0     5.0       0   1.0       0          0
3             260      1.0     4.0       0   1.0       1          1
4             527      1.0     5.0       0   1.0       0          0
...           ...      ...     ...     ...   ...     ...        ...
1000381      3513   5727.0     4.0       1  25.0       0          0
1000382      3535   5727.0     2.0       1  25.0       0          0
1000383      3536   5727.0     5.0       1  25.0       0          0
1000384      3555   5727.0     3.0       1  25.0       1          0
1000385      3578   5727.0     5.0       1  25.0       1          0

         Animation  Children's  Comedy  ...  occupation_11.0
occupation_12.0  \
0                1           1       1  ...                0
0
1                1           1       0  ...                0
0
2                0           0       0  ...                0
0
3                0           0       0  ...                0
0
4                0           0       0  ...                0
0
...            ...         ...     ...  ...              ...
...
1000381          0           0       0  ...                0
0
1000382          0           0       1  ...                0
0
1000383          0           0       1  ...                0
0
1000384          0           0       0  ...                0
0
1000385          0           0       0  ...                0
```

```
0

         occupation_13.0  occupation_14.0  occupation_15.0
occupation_16.0  \
0                      0                0                0
0
1                      0                0                0
0
2                      0                0                0
0
3                      0                0                0
0
4                      0                0                0
0
...                  ...              ...              ...
...
1000381                0                0                0
0
1000382                0                0                0
0
1000383                0                0                0
0
1000384                0                0                0
0
1000385                0                0                0
0

         occupation_17.0  occupation_18.0  occupation_19.0
occupation_20.0
0                      0                0                0
0
1                      0                0                0
0
2                      0                0                0
0
3                      0                0                0
0
4                      0                0                0
0
...                  ...              ...              ...
...
1000381                0                0                0
0
1000382                0                0                0
0
1000383                0                0                0
0
1000384                0                0                0
0
1000385                0                0                0
```

```
0

[1000209 rows x 43 columns]
```

```
x = df.drop(columns=['movie_id','user_id','rating'])
y=df[['rating']]
```

```
x
```

```
         gender   age  Action  Adventure  Animation  Children's
Comedy  \
0             0   1.0       0          0          1           1
1
1             0   1.0       0          0          1           1
0
2             0   1.0       0          0          0           0
0
3             0   1.0       1          1          0           0
0
4             0   1.0       0          0          0           0
0
...         ...   ...     ...        ...        ...         ...     ..
.
1000381       1  25.0       0          0          0           0
0
1000382       1  25.0       0          0          0           0
1
1000383       1  25.0       0          0          0           0
1
1000384       1  25.0       1          0          0           0
0
1000385       1  25.0       1          0          0           0
0

         Crime  Documentary  Drama  ...  occupation_11.0
occupation_12.0  \
0            0            0      0  ...                0
0
1            0            0      0  ...                0
0
2            0            0      1  ...                0
0
3            0            0      0  ...                0
0
4            0            0      1  ...                0
0
...        ...          ...    ...  ...              ...
...
1000381      0            0      1  ...                0
0
1000382      0            0      0  ...                0
```

```
0
1000383      0           0      0  ...               0
0
1000384      0           0      0  ...               0
0
1000385      0           0      1  ...               0
0

         occupation_13.0  occupation_14.0  occupation_15.0
occupation_16.0  \
0                      0                0                0
0
1                      0                0                0
0
2                      0                0                0
0
3                      0                0                0
0
4                      0                0                0
0
...                  ...              ...              ...
...
1000381                0                0                0
0
1000382                0                0                0
0
1000383                0                0                0
0
1000384                0                0                0
0
1000385                0                0                0
0

         occupation_17.0  occupation_18.0  occupation_19.0
occupation_20.0
0                      0                0                0
0
1                      0                0                0
0
2                      0                0                0
0
3                      0                0                0
0
4                      0                0                0
0
...                  ...              ...              ...
...
1000381                0                0                0
0
1000382                0                0                0
```

```
0
1000383                 0                 0                 0
0
1000384                 0                 0                 0
0
1000385                 0                 0                 0
0

[1000209 rows x 40 columns]
```

```
from sklearn.model_selection import train_test_split as tts
x_train, x_test, y_train, y_test = tts(x,y,test_size=0.2,random_state
= 10,stratify=y)
```

The dataset size being very huge we use LGBM Classifier

```
from lightgbm import LGBMClassifier as lgbm_ , LGBMRanker as lgbmrank
from sklearn.metrics import accuracy_score

lgbm = lgbm_(n_jobs=-1,boosting_type='gbdt',objective='multiclass')
lgbm.fit(x_train,y_train)
```

```
LGBMClassifier(objective='multiclass')
```

```
print(lgbm.score(x_test,y_test))
```

```
0.3623139140780436
```