News_Words

July 19, 2020

1 Feature Extraction

[1]:
```
import numpy as np
import pandas as pd
pd.options.display.max_columns = 999
```

[2]:
```
#loading the Boston Dataset
data = pd.read_csv('/Users/kyle/Documents/Virtual_Intern/Tencent/data/words.
 ↪csv', low_memory=False)
df = pd.DataFrame(data)
df
```

[2]:

	label	qid	title
0	5	g5 高速	G5 京昆高速,瓦厂坪大桥路段山体险情,建议大家推迟出行!
1	5	g5 高速	G5 京昆高速雅西段拖乌山突降暴雪 部分路段积雪深达 1 米
2	5	g5 高速	12 月起,G5 京昆高速开始“冬管”,这些地方需特别注意!
3	4	g5 高速	G50 沪渝高速部分路段进行对接施工,这些车辆全天禁止通行!
4	5	g5 高速	G5 京昆高速因大雪、路面结冰,继续交通管制!
…	…	…	…
413347	3	赤脚医生	还记得 H7N9 吗? 绍兴“赤脚医生”李兰娟领衔完成的项目获国家科技进步特等奖
413348	2	赤脚医生	赤脚医生和她的小背篓
413349	2	赤脚医生	情洒杏林路的赤脚医生
413350	3	赤脚医生	邯郸冀南新区赤脚医生补助名单正在审核中
413351	3	赤脚医生	疑似传播 HIV 病毒 印度一名“赤脚医生”被捕

[413352 rows x 3 columns]

[3]:
```
#Feature Engineering
# print(df_words[:10])
import string
def remove_punctuation(text):
    try: # python 2.x
        text = text.translate(None, string.punctuation)
```

```
    except: # python 3.x
        translator = text.maketrans('', '', ',?!《》.。、“”-+*/:……()@:" " ,’‘— ')
        text = text.translate(translator)
    return text
```

[4]:
```
qid_without_punc = []
# Remove punctuation for qid
for i in range(len(df['qid'])):
    qid_without_punc.append(remove_punctuation(df['qid'][i]))
qid_without_punc = pd.DataFrame(qid_without_punc).rename(columns={0:"qid"})
qid_without_punc['qid'] = qid_without_punc['qid'].str.strip()
qid_without_punc = qid_without_punc['qid'].replace(' ', '')
df['qid_clean'] = qid_without_punc
df
```

[4]:
```
        label   qid                                    title qid_clean
0           5  g5 高速          G5 京昆高速,瓦厂坪大桥路段山体险情,建议大家推迟出行!      g5 高速
1           5  g5 高速           G5 京昆高速雅西段拖乌山突降暴雪 部分路段积雪深达1 米      g5 高速
2           5  g5 高速          12 月起,G5 京昆高速开始“冬管”,这些地方需特别注意!      g5 高速
3           4  g5 高速         G50 沪渝高速部分路段进行对接施工,这些车辆全天禁止通行!      g5 高速
4           5  g5 高速               G5 京昆高速因大雪、路面结冰,继续交通管制!  ␣
 ↪    g5 高速
…       …   …                                     …       …
413347      3  赤脚医生  还记得 H7N9 吗? 绍兴“赤脚医生”李兰娟领衔完成的项目获国家科技进步特等奖      赤脚医生
413348      2  赤脚医生                                赤脚医生和她的小背篓      赤脚医生
413349      2  赤脚医生                                情洒杏林路的赤脚医生      赤脚医生
413350      3  赤脚医生                   邯郸冀南新区赤脚医生补助名单正在审核中␣
 ↪     赤脚医生
413351      3  赤脚医生                 疑似传播 HIV 病毒 印度一名“赤脚医生”被捕  ␣
 ↪    赤脚医生

[413352 rows x 4 columns]
```

[5]:
```
# Remove punctuation for title
title_without_punc = []
for i in range(len(df['title'])):
    title_without_punc.append(remove_punctuation((df['title'][i])))
title_without_punc = pd.DataFrame(title_without_punc).rename(columns={0:
 ↪"title"})
```

```
title_without_punc['title'] = title_without_punc['title'].str.strip()
title_without_punc = title_without_punc['title'].replace(' ', '')
df['title_clean'] = title_without_punc
df
```

[5]:
```
        label   qid                                          title qid_clean  \
0           5  g5 高速           G5 京昆高速,瓦厂坪大桥路段山体险情,建议大家推迟出行!      g5 高速
1           5  g5 高速            G5 京昆高速雅西段拖乌山突降暴雪 部分路段积雪深达 1 米      g5 高速
2           5  g5 高速           12 月起,G5 京昆高速开始“冬管”,这些地方需特别注意!      g5 高速
3           4  g5 高速          G50 沪渝高速部分路段进行对接施工,这些车辆全天禁止通行!      g5 高速
4           5  g5 高速                G5 京昆高速因大雪、路面结冰,继续交通管制!  ␣
↪     g5 高速
…       …   …                                          …       …
413347      3  赤脚医生  还记得 H7N9 吗? 绍兴“赤脚医生”李兰娟领衔完成的项目获国家科技进步特等奖       赤脚医生
413348      2  赤脚医生                                赤脚医生和她的小背篓       赤脚医生
413349      2  赤脚医生                                情洒杏林路的赤脚医生       赤脚医生
413350      3  赤脚医生                  邯郸冀南新区赤脚医生补助名单正在审核中␣
↪      赤脚医生
413351      3  赤脚医生                疑似传播 HIV 病毒 印度一名“赤脚医生”被捕  ␣
↪     赤脚医生

                                title_clean
0                 G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟出行
1               G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深达 1 米
2                  12 月起 G5 京昆高速开始冬管这些地方需特别注意
3              G50 沪渝高速部分路段进行对接施工这些车辆全天禁止通行
4                      G5 京昆高速因大雪路面结冰继续交通管制
…                                          …
413347  还记得 H7N9 吗绍兴赤脚医生李兰娟领衔完成的项目获国家科技进步特等奖
413348                             赤脚医生和她的小背篓
413349                             情洒杏林路的赤脚医生
413350                 邯郸冀南新区赤脚医生补助名单正在审核中
413351                 疑似传播 HIV 病毒印度一名赤脚医生被捕

[413352 rows x 5 columns]
```

[6]:
```
df = df.drop(["qid",'title'], axis = 1)
df['qid_clean'] = df['qid_clean'].apply(lambda x: x.upper())
df['title_clean'] = df['title_clean'].apply(lambda x: x.upper())
```

```
df
```

[6]:

	label	qid_clean	title_clean
0	5	G5 高速	G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟出行
1	5	G5 高速	G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深达 1 米
2	5	G5 高速	12 月起 G5 京昆高速开始冬管这些地方需特别注意
3	4	G5 高速	G5O 沪渝高速部分路段进行对接施工这些车辆全天禁止通行
4	5	G5 高速	G5 京昆高速因大雪路面结冰继续交通管制
…	…	…	…
413347	3	赤脚医生	还记得 H7N9 吗绍兴赤脚医生李兰娟领衔完成的项目获国家科技进步特等奖
413348	2	赤脚医生	赤脚医生和她的小背篓
413349	2	赤脚医生	情洒杏林路的赤脚医生
413350	3	赤脚医生	邯郸冀南新区赤脚医生补助名单正在审核中
413351	3	赤脚医生	疑似传播 HIV 病毒印度一名赤脚医生被捕

[413352 rows x 3 columns]

[7]:
```
import jieba.analyse as anls
def words_extract(sentence, topN):
    tot_terms = []
    for i in range(len(sentence)):
        seg = anls.extract_tags(sentence[i], topK = topN, withWeight = False)
        tot_terms.append(seg)
    return tot_terms
```

[8]:
```
df['qid_terms'] = words_extract(df['qid_clean'], 5)
df
```

```
Building prefix dict from the default dictionary …
Loading model from cache
/var/folders/pf/pypg0m0x0ng1g1w1mhrbr2cm0000gn/T/jieba.cache
Loading model cost 0.645 seconds.
Prefix dict has been built successfully.
```

[8]:

	label	qid_clean	title_clean	qid_terms
0	5	G5 高速	G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟出行	[G5, 高速]
1	5	G5 高速	G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深达 1 米	[G5, 高速]
2	5	G5 高速	12 月起 G5 京昆高速开始冬管这些地方需特别注意	[G5, 高速]

```
3           4      G5 高速          G50 沪渝高速部分路段进行对接施工这些车辆全天禁
止通行  [G5, 高速]
4           5      G5 高速                G5 京昆高速因大雪路面结冰继续交通管制 ␣
↪[G5, 高速]
…       …      …                                  …       …
413347      3      赤脚医生  还记得 H7N9 吗绍兴赤脚医生李兰娟领衔完成的项目获国家科
技进步特等奖    [赤脚医生]
413348      2      赤脚医生                          赤脚医生和她的小背篓     [赤脚
医生]
413349      2      赤脚医生                          情洒杏林路的赤脚医生     [赤脚
医生]
413350      3      赤脚医生                  邯郸冀南新区赤脚医生补助名单正在审核
中    [赤脚医生]
413351      3      赤脚医生                  疑似传播 HIV 病毒印度一名赤脚医生被捕␣
↪    [赤脚医生]

[413352 rows x 4 columns]
```

[9]:
```
df['title_terms'] = words_extract(df['title_clean'], 40)
df
```

[9]:
```
        label qid_clean                        title_clean qid_terms  \
0           5      G5 高速          G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟
出行  [G5, 高速]
1           5      G5 高速          G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深
达 1 米  [G5, 高速]
2           5      G5 高速             12 月起 G5 京昆高速开始冬管这些地方需特别注
意  [G5, 高速]
3           4      G5 高速          G50 沪渝高速部分路段进行对接施工这些车辆全天禁
止通行  [G5, 高速]
4           5      G5 高速                G5 京昆高速因大雪路面结冰继续交通管制 ␣
↪[G5, 高速]
…       …      …                                  …       …
413347      3      赤脚医生  还记得 H7N9 吗绍兴赤脚医生李兰娟领衔完成的项目获国家科
技进步特等奖    [赤脚医生]
413348      2      赤脚医生                          赤脚医生和她的小背篓     [赤脚
医生]
413349      2      赤脚医生                          情洒杏林路的赤脚医生     [赤脚
医生]
413350      3      赤脚医生                  邯郸冀南新区赤脚医生补助名单正在审核
中    [赤脚医生]
413351      3      赤脚医生                  疑似传播 HIV 病毒印度一名赤脚医生被捕␣
↪    [赤脚医生]

                                                title_terms
0           [大桥路, G5, 京昆, 瓦厂, 险情, 推迟, 出行, 山体, 高速, 建议, 大家]
```

```
1           [G5, 京昆, 乌山, 突降, 深达, 暴雪, 西段, 路段, 积雪, 高速, 部分]
2               [12, G5, 京昆, 冬管, 高速, 注意, 特别, 地方, 开始, 这些]
3         [G50, 禁止通行, 路段, 对接, 全天, 施工, 高速, 车辆, 部分, 这些, 进行]
4                     [G5, 京昆, 交通管制, 结冰, 大雪, 路面, 高速, 继续]
…                                                  …
413347  [H7N9, 李兰娟, 赤脚医生, 特等奖, 科技进步, 领衔, 绍兴, 记得, 完成, 项…
413348                                              [赤脚医生, 背篓]
413349                                          [情洒, 赤脚医生, 杏林]
413350                    [赤脚医生, 冀南, 邯郸, 补助, 新区, 审核, 名单, 正在]
413351                   [HIV, 赤脚医生, 疑似, 被捕, 病毒, 传播, 一名, 印度]

[413352 rows x 5 columns]
```

[10]:

```python
df['num_qid_terms'] = df["qid_terms"].apply(lambda s: len(s))
df['num_title_terms'] = df["title_terms"].apply(lambda s: len(s))
df
```

[10]:

```
        label qid_clean                       title_clean qid_terms  \
0           5      G5 高速          G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟
出行  [G5, 高速]
1           5      G5 高速          G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深
达 1 米  [G5, 高速]
2           5      G5 高速             12 月起 G5 京昆高速开始冬管这些地方需特别注
意  [G5, 高速]
3           4      G5 高速        G50 沪渝高速部分路段进行对接施工这些车辆全天禁
止通行  [G5, 高速]
4           5      G5 高速                   G5 京昆高速因大雪路面结冰继续交通管制 ␣
 ↪[G5, 高速]
…         …        …                                  …        …
413347      3      赤脚医生  还记得 H7N9 吗绍兴赤脚医生李兰娟领衔完成的项目获国家科
技进步特等奖     [赤脚医生]
413348      2      赤脚医生                          赤脚医生和她的小背篓     [赤脚
医生]
413349      2      赤脚医生                          情洒杏林路的赤脚医生     [赤脚
医生]
413350      3      赤脚医生                 邯郸冀南新区赤脚医生补助名单正在审核
中     [赤脚医生]
413351      3      赤脚医生                 疑似传播 HIV 病毒印度一名赤脚医生被捕␣
 ↪    [赤脚医生]

                                         title_terms  num_qid_terms  \
0          [大桥路, G5, 京昆, 瓦厂, 险情, 推迟, 出行, 山体, 高速, 建议, 大家]   ␣
 ↪           2
1           [G5, 京昆, 乌山, 突降, 深达, 暴雪, 西段, 路段, 积雪, 高速, 部分]   ␣
 ↪          2
```

```
2                [12，G5，京昆，冬管，高速，注意，特别，地方，开始，这些]                ␣
↪      2
3          [G50，禁止通行，路段，对接，全天，施工，高速，车辆，部分，这些，进行]          ␣
↪            2
4                       [G5，京昆，交通管制，结冰，大雪，路面，高速，继续]                ␣
↪      2
…                                                          …              …
413347   [H7N9，李兰娟，赤脚医生，特等奖，科技进步，领衔，绍兴，记得，完成，项…          ␣
↪           1
413348                                                [赤脚医生，背篓]                  1
413349                                              [情洒，赤脚医生，杏林]                1
413350                   [赤脚医生，冀南，邯郸，补助，新区，审核，名单，正在]              ␣
↪       1
413351                  [HIV，赤脚医生，疑似，被捕，病毒，传播，一名，印度]               ␣
↪      1

        num_title_terms
0                    11
1                    11
2                    10
3                    11
4                     8
…                   …
413347               11
413348                2
413349                3
413350                8
413351                8

[413352 rows x 7 columns]
```

[11]:
```
def num_matches(terms_a, terms_b):
    times = 0
    for i in terms_a:
        for j in terms_b:
            if i == j:
                times += 1
    return times
```

[12]:
```
tot_num_matches = []
for i in range(len(df)):
    tot_num_matches.append(num_matches(df['qid_terms'][i],df['title_terms'][i]))
len(tot_num_matches)
```

[12]: 413352

[13]:
```
df['tot_num_matches'] = tot_num_matches
df[0:10]
```

[13]:

	label	qid_clean	title_clean	qid_terms	title_terms	num_qid_terms	num_title_terms	tot_num_matches
0	5	G5 高速	G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟出行	[G5, 高速]	[大桥路, G5, 京昆, 瓦厂, 险情, 推迟, 出行, 山体, 高速, 建议, 大家]	2	11	2
1	5	G5 高速	G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深达 1 米	[G5, 高速]	[G5, 京昆, 乌山, 突降, 深达, 暴雪, 西段, 路段, 积雪, 高速, 部分]	2	11	2
2	5	G5 高速	12 月起 G5 京昆高速开始冬管这些地方需特别注意	[G5, 高速]	[12, G5, 京昆, 冬管, 高速, 注意, 特别, 地方, 开始, 这些]	2	10	2
3	4	G5 高速	G50 沪渝高速部分路段进行对接施工这些车辆全天禁止通行	[G5, 高速]	[G50, 禁止通行, 路段, 对接, 全天, 施工, 高速, 车辆, 部分, 这些, 进行]	2	11	1
4	5	G5 高速	G5 京昆高速因大雪路面结冰继续交通管制	[G5, 高速]	[G5, 京昆, 交通管制, 结冰, 大雪, 路面, 高速, 继续]	2	8	2
5	5	G5 高速	G5 京昆高速结冰严重栗子坪至彝海双向交通管制	[G5, 高速]	[G5, 京昆, 彝海, 交通管制, 栗子, 结冰, 双向, 高速, 严重]	2	9	2
6	1	G5 高速	吉利帝豪高速追尾比亚迪 G5 车主同样都是国产差距太大	[G5, 高速]	[帝豪, G5, 追尾, 比亚迪, 吉利, 车主, 国产, 差距, 高速, 同样]	2	10	2
7	4	G5 高速	突发 G5 高速绵广段厚坝方向一车辆起火燃烧原因暂不明	[G5, 高速]	[起火燃烧, G5, 绵广段, 厚坝, 突发, 不明, 高速, 车辆, 方向, 原因]	2	10	2
8	1	G5 高速	HOWOG5X 冷藏车高速高效宽箱体大容积	[G5, 高速]	[冷藏车, HOWOG5X, 箱体, 容积, 高效, 高速]	2	11	1
9	4	G5 高速	G5 京昆高速因降雪实施交通管制	[G5, 高速]	[G5, 京昆, 交通管制, 降雪, 高速, 实施]	2	8	2

```
5                 9                 2
6                10                 2
7                10                 2
8                 6                 1
9                 6                 2
```

[32]:
```
df['match_ratios'] = np.array(df['tot_num_matches']) * 1. / np.
↪array(df['num_qid_terms'])
df[0:10]
```

```
/Users/kyle/Library/Python/3.7/lib/python/site-packages/ipykernel_launcher.py:1:
RuntimeWarning: invalid value encountered in true_divide
  """Entry point for launching an IPython kernel.
```

[32]:
```
   label qid_clean                    title_clean qid_terms  \
0      5      G5 高速     G5 京昆高速瓦厂坪大桥路段山体险情建议大家推迟出行  [G5,␣
↪高速]
1      5      G5 高速    G5 京昆高速雅西段拖乌山突降暴雪部分路段积雪深达 1 米 ␣
↪[G5, 高速]
2      5      G5 高速       12 月起 G5 京昆高速开始冬管这些地方需特别注意  [G5, 高
速]
3      4      G5 高速  G50 沪渝高速部分路段进行对接施工这些车辆全天禁止通行  [G5,␣
↪高速]
4      5      G5 高速           G5 京昆高速因大雪路面结冰继续交通管制  [G5, 高速]
5      5      G5 高速       G5 京昆高速结冰严重栗子坪至彝海双向交通管制  [G5, 高
速]
6      1      G5 高速     吉利帝豪高速追尾比亚迪 G5 车主同样都是国产差距太大  [G5,␣
↪高速]
7      4      G5 高速     突发 G5 高速绵广段厚坝方向一车辆起火燃烧原因暂不明  [G5,␣
↪高速]
8      1      G5 高速          HOWOG5X 冷藏车高速高效宽箱体大容积  [G5, 高速]
9      4      G5 高速               G5 京昆高速因降雪实施交通管制  [G5, 高速]

                                      title_terms  num_qid_terms  \
0    [大桥路, G5, 京昆, 瓦厂, 险情, 推迟, 出行, 山体, 高速, 建议, 大家]          ␣
↪   2
1     [G5, 京昆, 乌山, 突降, 深达, 暴雪, 西段, 路段, 积雪, 高速, 部分]          ␣
↪  2
2         [12, G5, 京昆, 冬管, 高速, 注意, 特别, 地方, 开始, 这些]              2
3  [G50, 禁止通行, 路段, 对接, 全天, 施工, 高速, 车辆, 部分, 这些, 进行]          ␣
↪    2
4               [G5, 京昆, 交通管制, 结冰, 大雪, 路面, 高速, 继续]              2
5          [G5, 京昆, 彝海, 交通管制, 栗子, 结冰, 双向, 高速, 严重]             ␣
↪2
6        [帝豪, G5, 追尾, 比亚迪, 吉利, 车主, 国产, 差距, 高速, 同样]           ␣
↪ 2
```

```
7       [起火燃烧，G5，绵广段，厚坝，突发，不明，高速，车辆，方向，原因]                 ␣
↪    2
8                    [冷藏车，HOWOG5X，箱体，容积，高效，高速]                2
9                        [G5，京昆，交通管制，降雪，高速，实施]                 2
```

	num_title_terms	tot_num_matches	match_ratios
0	11	2	1.0
1	11	2	1.0
2	10	2	1.0
3	11	1	0.5
4	8	2	1.0
5	9	2	1.0
6	10	2	1.0
7	10	2	1.0
8	6	1	0.5
9	6	2	1.0

[31]:
```
data_after_extraction = df.
↪drop(['qid_clean','title_clean','qid_terms','title_terms'], axis = 1)
data_after_extraction
```

[38]:
```
# df.to_excel("total.xlsx", encoding = "UTF-8", index = 'False',␣
↪engine='xlsxwriter')
# data_after_extraction.to_csv("data.csv", index = False)
```

2 Model Set-up

[39]:
```
import lightgbm as lgb
from sklearn.datasets import load_breast_cancer, load_boston, load_wine
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import mean_squared_error, roc_auc_score, precision_score
```

[40]:
```
df_1 = data_after_extraction.drop(['label', 'num_title_terms'], axis = 1)
Y_1 = data_after_extraction['label']
df_1
```

[40]:

	num_qid_terms	tot_num_matches	match_ratios
0	2	2	1.0
1	2	2	1.0
2	2	2	1.0
3	2	1	0.5
4	2	2	1.0
…	…	…	…

```
413347              1               1           1.0
413348              1               1           1.0
413349              1               1           1.0
413350              1               1           1.0
413351              1               1           1.0

[413352 rows x 3 columns]
```

[41]:
```
#Scaling using the Standard Scaler
sc_1 = StandardScaler()
sc_1.fit(df_1)
X_1 = pd.DataFrame(sc_1.fit_transform(df_1))
```

[42]:
```
#train-test-split
X_train, X_test, y_train, y_test = train_test_split(df_1, Y_1, test_size=0.2,␣
↪random_state=0)
```

[43]:
```
#Converting the dataset in proper LGB format
d_train = lgb.Dataset(X_train, label = y_train)
```

[44]:
```
#setting up the prarmeters
params = {}
params['learning_rate'] = 0.1
params['boosting_type'] = 'gbdt' # GradientBoostingDecisionTree
params['objective'] = 'multiclass' #Multi-class target feature
params['metric'] = 'multi_logloss' #metric for multi-class
params['max_depth'] = 30
params['num_class'] = 6 #no.of unique values in the target class not inclusive␣
↪of the end value
```

[45]:
```
# trainning the model
clf = lgb.train(params, d_train, 100) #targeting the model on 100 epocs
```

3 Model Prediction

[46]:
```
#prediction on the test dataset
y_pred_1 = clf.predict(X_test)
```

[47]:
```
#printing the predctions
y_pred_1
```

[47]:
```
array([[1.59247800e-04, 3.40735000e-02, 9.07315510e-02, 1.76056811e-01,
        2.11787080e-01, 4.87191811e-01],
       [1.23389780e-06, 1.66646759e-01, 2.69983050e-01, 2.36427131e-01,
        1.06608158e-01, 2.20333668e-01],
```

```
       [8.56504662e-05, 1.02753799e-01, 1.72298600e-01, 2.23921238e-01,
        1.81965839e-01, 3.18974873e-01],
       …,
       [3.21046295e-04, 1.67815044e-01, 2.58268832e-01, 1.96599264e-01,
        1.66382173e-01, 2.10613640e-01],
       [8.84047550e-05, 7.40530792e-02, 1.85240366e-01, 2.60119147e-01,
        1.89937990e-01, 2.90561013e-01],
       [9.05087273e-04, 1.68175091e-01, 2.46730201e-01, 2.00043095e-01,
        1.08074526e-01, 2.76071999e-01]])
```

[48]:
```
#argmax() method
y_pred_2 = [np.argmax(line) for line in y_pred_1]
#printing the predictions
print(y_pred_2[0:50])
```

```
[5, 2, 5, 5, 5, 5, 5, 5, 5, 2, 5, 5, 3, 5, 3, 2, 5, 3, 5, 5, 5, 5, 3, 5, 2, 3,
5, 5, 3, 5, 5, 5, 5, 2, 5, 5, 5, 3, 5, 5, 5, 5, 5, 5, 5, 5, 5, 3, 5, 5]
```

[49]:
```
#using precision score for error metrics
prec_score = precision_score(y_pred_2, y_test, average=None).mean()
print(prec_score)
```

```
0.19978194932356466
```

[50]:
```
#using RMSE error metric
RMSE = np.sqrt(mean_squared_error(y_pred_2, y_test))
print(RMSE)
```

```
1.7398584568903015
```

[ ]:

[ ]: