Movies are Enjoyable?

Haoda Song

Brown University

Oct.15.2020

Github https://github.com/Haoda1860/Data1030MidtermProject

Introduction

- Motivation: Did you sometimes feel not satisfy with your movie recommendation list on Netflix/Hulu/HBO/Amazon Video?

- Goal: Improve satisfaction of movie recommendation list based on users' personalities

- Data Goal: Classify if the movies on the recommendation list are enjoyable based on the personalities of users

Introduction

- Data!Kaggle:

https://www.kaggle.com/arslanali4343/top-personality-dataset

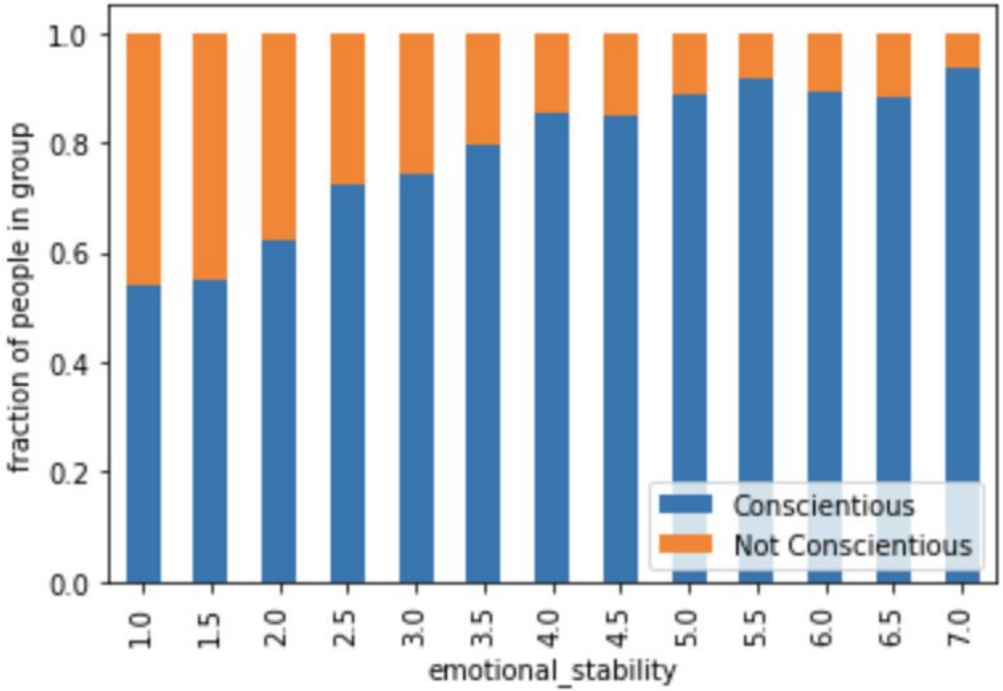
	openness	agreeableness	emotional_stability	conscientiousness	extraversion	assigned metric	assigned condition	is_personalized	enjoy_watching	avg_ratings
0	5.0	2.0	3.0	2.5	6.5	serendipity	high	4	Enjoyable	4.252363
1	7.0	4.0	6.0	5.5	4.0	all	default	2	Not Enjoyable	4.173935
2	4.0	3.0	4.5	2.0	2.5	serendipity	medium	2	Not Enjoyable	4.764654
3	5.5	5.5	4.0	4.5	4.0	popularity	medium	3	Not Enjoyable	4.444313
4	5.5	5.5	3.5	4.5	2.5	popularity	medium	2	Not Enjoyable	4.444313

Notes:

- Target Variable: “Enjoy_Watching”
- 1834 Observations with 10 features

Exploratory Data Analysis

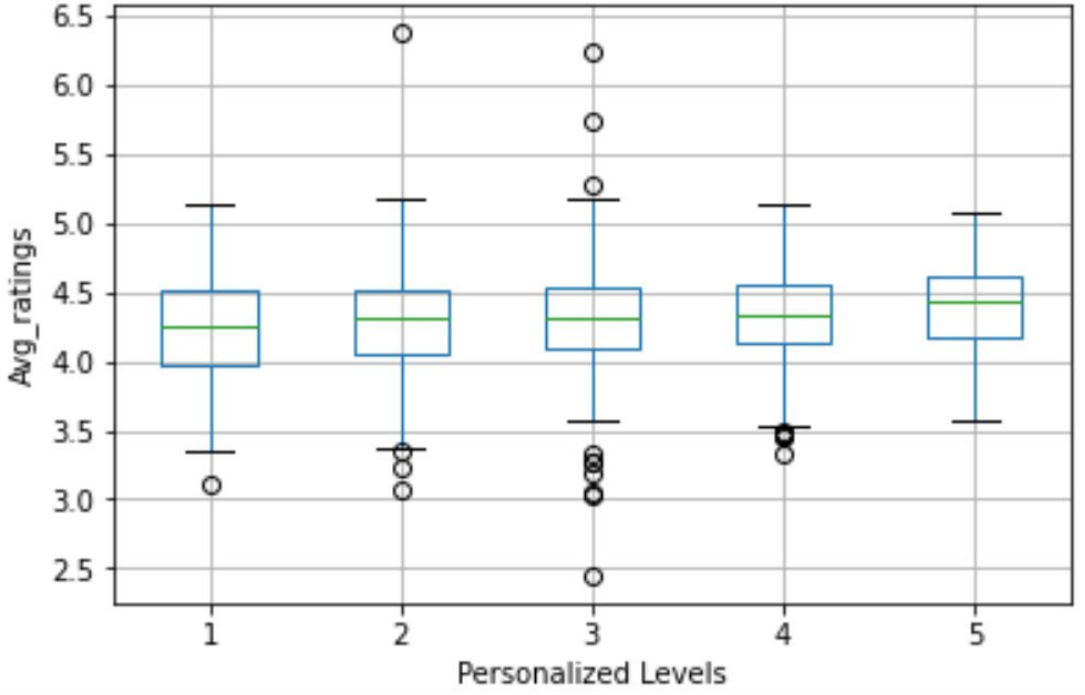
Emotional Stability highly and positively correlates with conscientiousness.



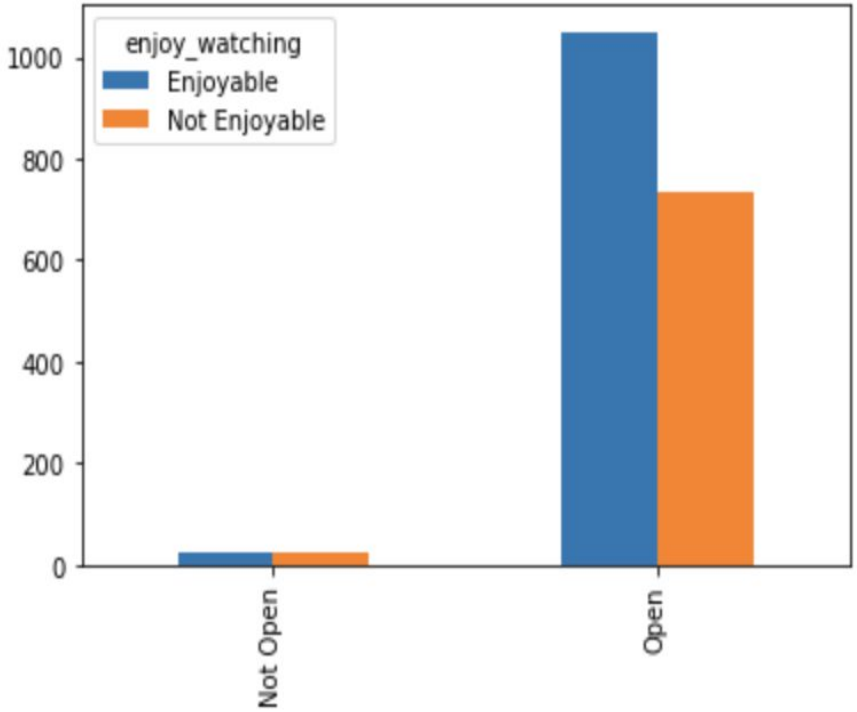
Exploratory Data Analysis

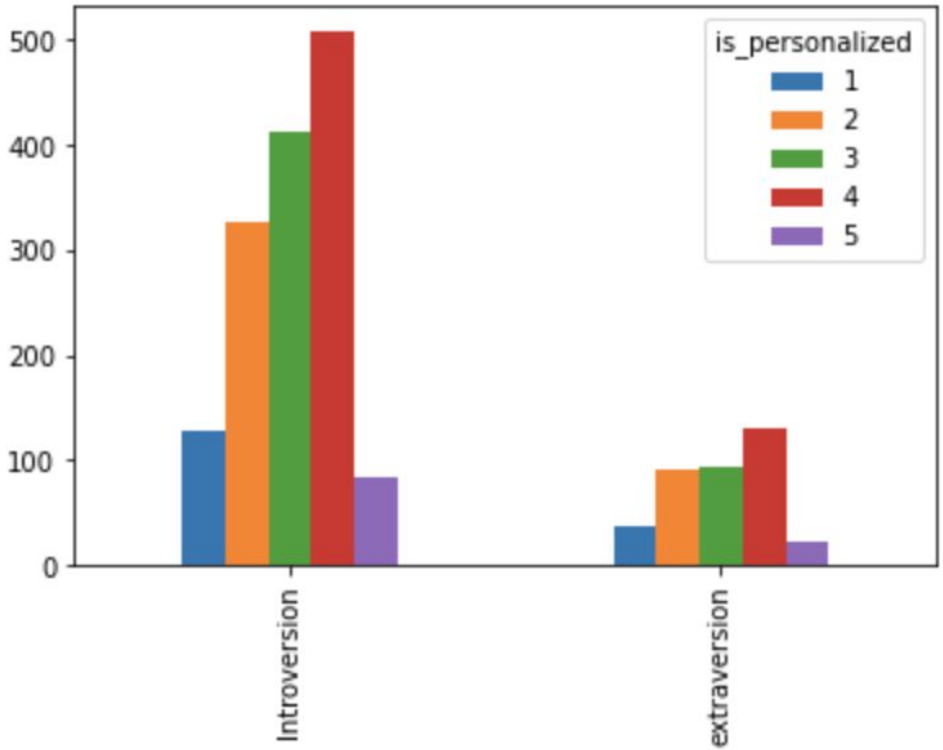
Interesting:

- No Obvious relationship
- Weak correlation (Range between the 25% to 75% percentiles)



Exploratory Data Analysis





Splitting Data and Preprocessors

Splitting Data:

- Basic Splitting: Train(64%);Validation(16%);Test(20%)
- Imbalanced data: Imbalanced Fraction of points in two classes, “Enjoy/Not Enjoy”.
- Stratified K Fold method

Preprocessing:

- No Missing Values!
- After preprocessing:
  - Train Set: (1174,12)(1174,)
  - Validation Set:(294,12)(294,)
  - Test Set:(367,12)(367,)

Splitting Data and Preprocessors

K Fold

```
train balance:
Enjoyable        0.5763
Not Enjoyable    0.4237
Name: enjoy_watching, dtype: float64
val balance:
Enjoyable        0.588435
Not Enjoyable    0.411565
Name: enjoy_watching, dtype: float64
train balance:
Enjoyable        0.568627
Not Enjoyable    0.431373
Name: enjoy_watching, dtype: float64
val balance:
Enjoyable        0.619048
Not Enjoyable    0.380952
Name: enjoy_watching, dtype: float64
```

Stratified K Fold

```
train balance:
Enjoyable        0.578858
Not Enjoyable    0.421142
Name: enjoy_watching, dtype: float64
val balance:
Enjoyable        0.578231
Not Enjoyable    0.421769
Name: enjoy_watching, dtype: float64
train balance:
Enjoyable        0.578858
Not Enjoyable    0.421142
Name: enjoy_watching, dtype: float64
val balance:
Enjoyable        0.578231
Not Enjoyable    0.421769
Name: enjoy_watching, dtype: float64
```

Preprocessing

- **Ordinal Encoder:** Seven *Ranked Categorical* Variables. (Openness, Agreeableness, Emotional Stability, Conscientiousness, Extraversion, Is Personalized, Assigned condition)
- **MinMax Encoder:** One *Continuous* Variable, Average Ratings (*Bounded* by 1 to 5)
- **One Hot Encoder:** One multilevel *unranked categorical* variable, Assigned metrics.
- **Label Encoder:** *Target Categorical* variable, “Enjoy Watching” with two levels: Enjoyable/Not Enjoyable

Thank YOU!!!

Questions?