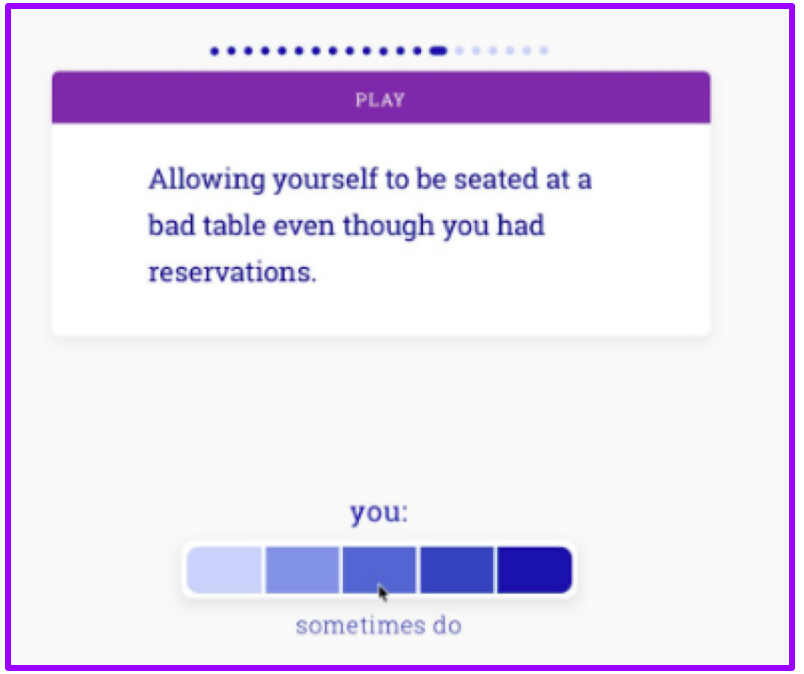
whodo

Parsing Big Data with Machine Learning

Baljinder Ghotra
Bruno Yamashita
Megha Asnani
Sakshi Sharma

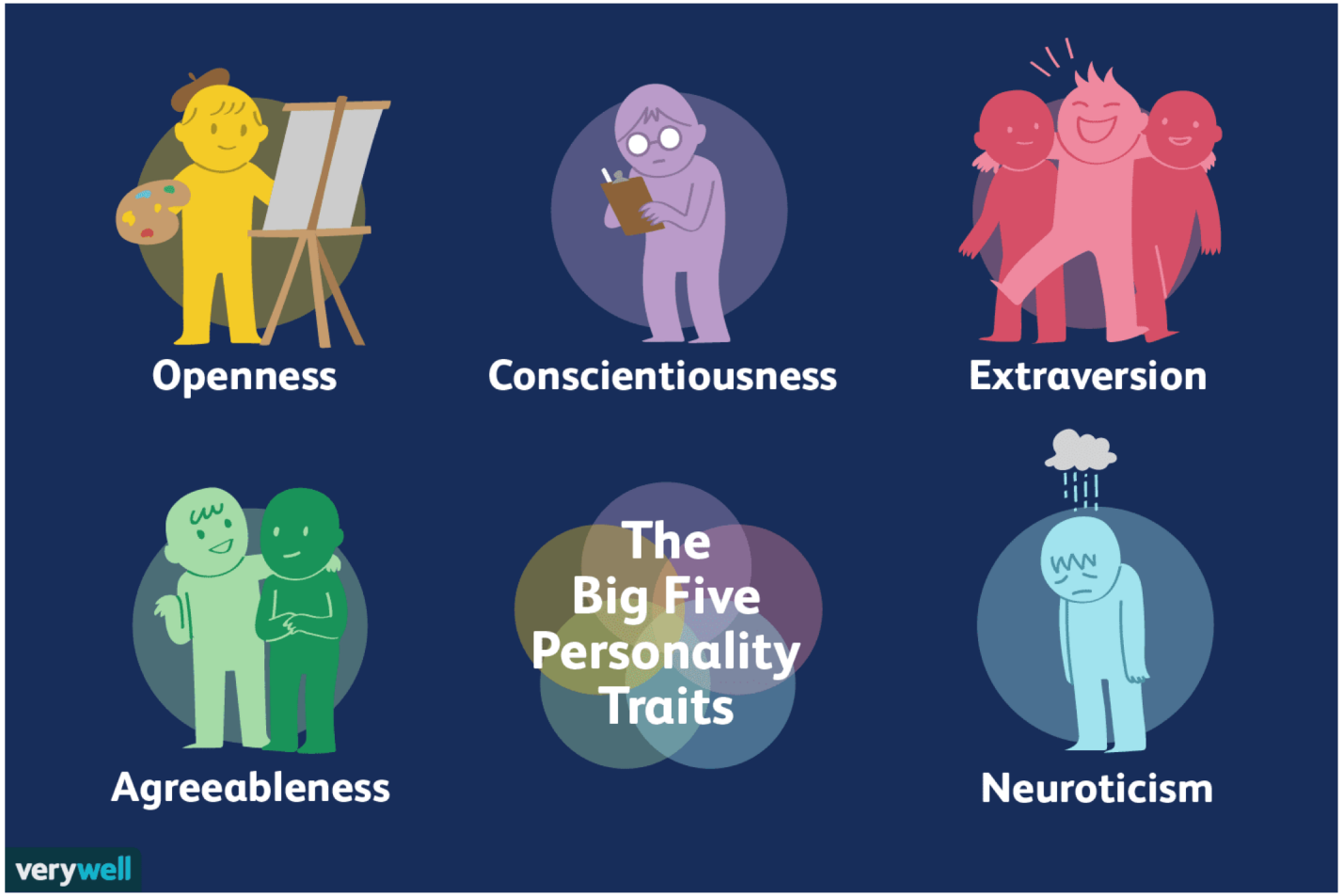
About Whodo

- ★ Whodo is building an app where users take personality quizzes, each containing about 20 questions.
- ★ The target audience is young women who would like to know more about themselves and how they fit into the world.
- ★ Each question in the quiz corresponds to 2 facets of the Big Five Personality traits (explained later).
- ★ Test taker's answers are later "anonymously" reviewed by friends and family members.
- ★ Users can volunteer additional data such as age, sex, race, relationship status, etc.



whodo

Big Five Personality Traits and Facets



Big Five Personality Traits and Facets

The Big Five Factors	The 30 Personality Facets that make up each Factor	
Click on the Factor heading to go to an in-depth description. Or scroll down - browse around and read descriptions of the Facets relating to the Factor.		
Openness to Experience	Imagination	Willingness to Experiment
	Artistic Interests	Intellectual Curiosity
	Depth of Emotions	Tolerance for Diversity
Conscientiousness 'Work Ethic'	Sense of Competence	Achievement Striving
	Orderliness	Self-Discipline
	Sense of Responsibility	Deliberateness
Extraversion	Warmth	Activity Level
	Gregariousness	Excitement-Seeking
	Assertiveness	Positive Emotions
Agreeableness	Trust in others	Compliance
	Sincerity	Modesty
	Altruism	Sympathy
Natural Reactions	Anxiety	Self-Consciousness
	Angry Hostility	Self-Indulgence
	Moodiness/Contentment	Sensitivity to Stress

http://www.testsonthenet.com/Factors-facets.htm

Project's Objective

- Whodo requires a software that will look at users responses and discern patterns or clusters of behavior based on various demographics variables such as age, gender, ethnicity, income, etc.

- Their goal is to remove human bias from the process as much as possible so that these clusters present themselves based on a statistical finding by the software, not by the preconceived notions of how to group people is expected to behave.

- Taking Whodo's goal one step further we wanted to see if we could use ML to build prediction models of demographics and results based on the data provided.

Data Description (Raw Data)

item	topic01	topic02	topic03	topic04	primary_factor	facet	reverse	secondary_factor	facet1	reverse1	type1	type2	type3
Saying "thank you" when someone holds the door.	In Public	Service	NaN	NaN	Extraversion	Positive Emotions	True	Agreeableness	Compliance	True	NaN	NaN	NaN
Low talking, or talking at a volume that's har...	Personal	Friends	Habits	NaN	Natural_Reactions	Self-Consciousness	True	Extraversion	Assertiveness	False	NaN	NaN	NaN
Standing close to a stranger in public.	In Public	NaN	NaN	NaN	Extraversion	Gregariousness	True	Agreeableness	Modesty	False	NaN	NaN	NaN

QUESTION

TOPICS

PRIMARY PERSONALITY TRAIT AND FACETS

SECONDARY PERSONALITY TRAIT AND FACETS

QUESTION TYPE

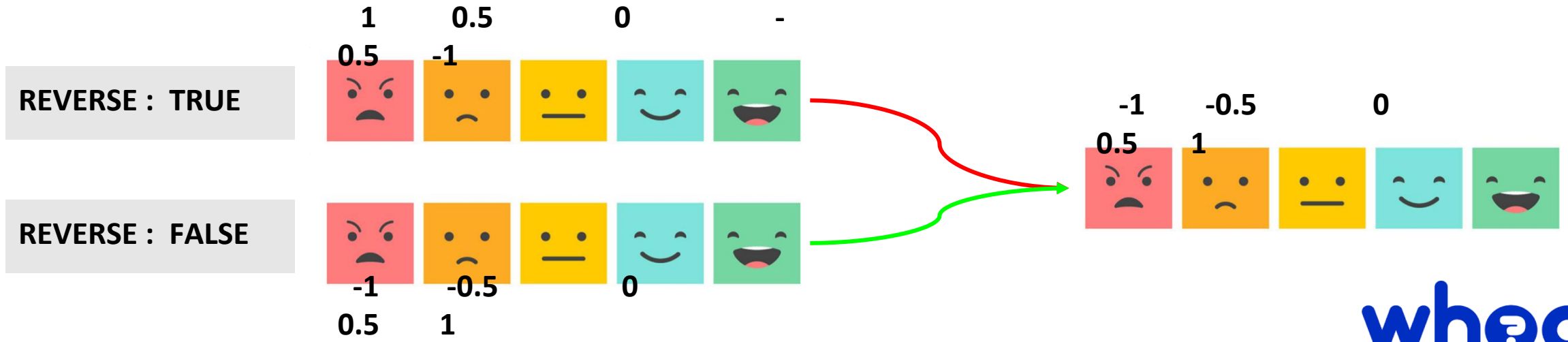
whodo

Data Description (Raw Data)

	WX.1	NX	WY	WX.2	WX.3	WY.1	NY	WX.4	NX.1	WY.2	WY.3
0	1.0	0.5	0.5	1.0	1.0	0.5	0.0	1.0	0.5	0.5	0.5
1	0.0	0.0	-0.5	0.0	-0.5	-0.5	-0.5	0.0	0.0	-0.5	-0.5
2	1.0	0.5	0.5	1.0	0.0	0.5	0.0	1.0	0.5	0.5	0.5
3	0.0	-0.5	0.5	0.0	-1.0	0.5	0.0	0.0	-0.5	0.5	0.5
4	-0.5	0.0	1.0	0.5	-0.5	1.0	0.5	0.5	0.0	1.0	1.0

DEMOGRAPHIC INFORMATION

RESPONSE



https://blog.leadquizzes.com/wp-content/uploads/2019/08/New-blog-graphic-3-1.jpg

whodo

Data Transformation

```
item  topic01  topic02  topic03  topic04  primary_factor        facet  reverse  secondary_factor       facet1  reverse1      type1  type2  type3
```

```
WX.1   NX   WY   WX.2   WX.3   WY.1   NY   WX.4   NX.1   WY.2   WY.3
```

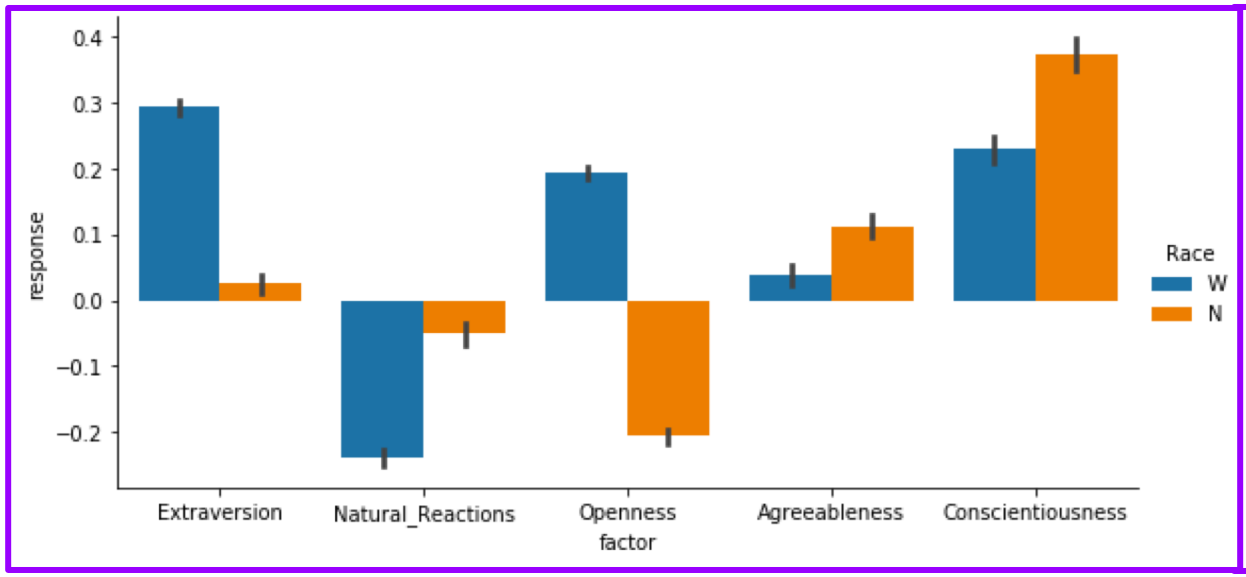
```
item  topic01  topic02  topic03  topic04  primary_factor        facet  reverse
```

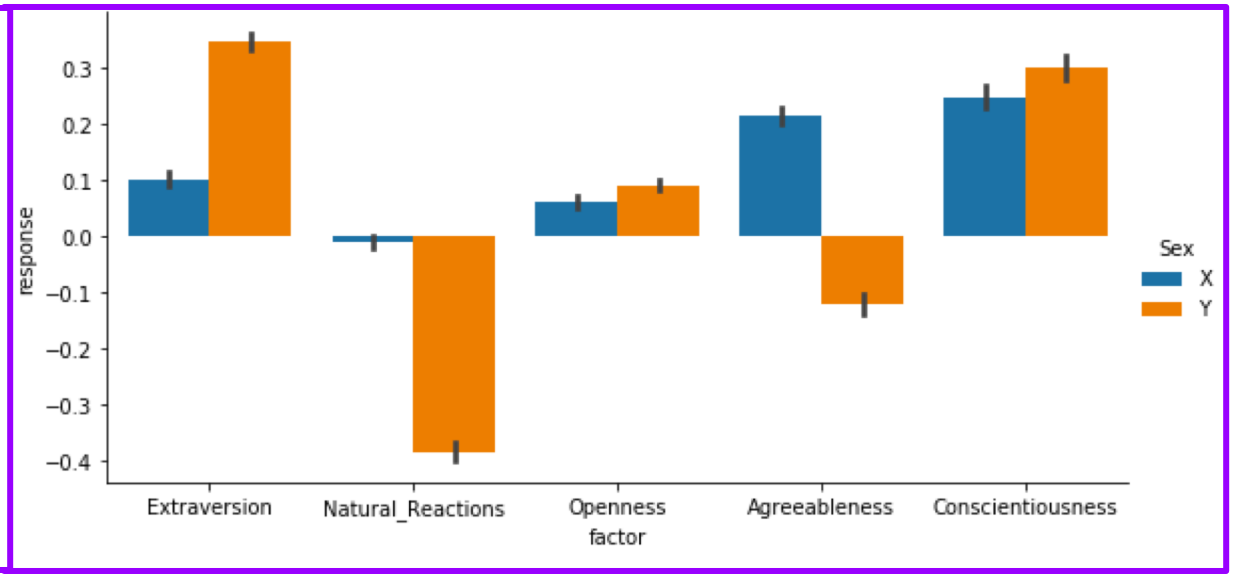
```
secondary_factor       facet1  reverse1
```

	topic	factor	facet	type	response	Race	Sex
0	In Public	Extraversion	Positive Emotions	na	1.0	W	X
1	Personal	Natural_Reactions	Self-Consciousness	na	0.0	W	X
2	In Public	Extraversion	Gregariousness	na	1.0	W	X
3	Love & Sex	Openness	Willingness to Experiment	Relationship	0.0	W	X
4	In Public	Extraversion	Activity Level	na	0.5	W	X

whodo

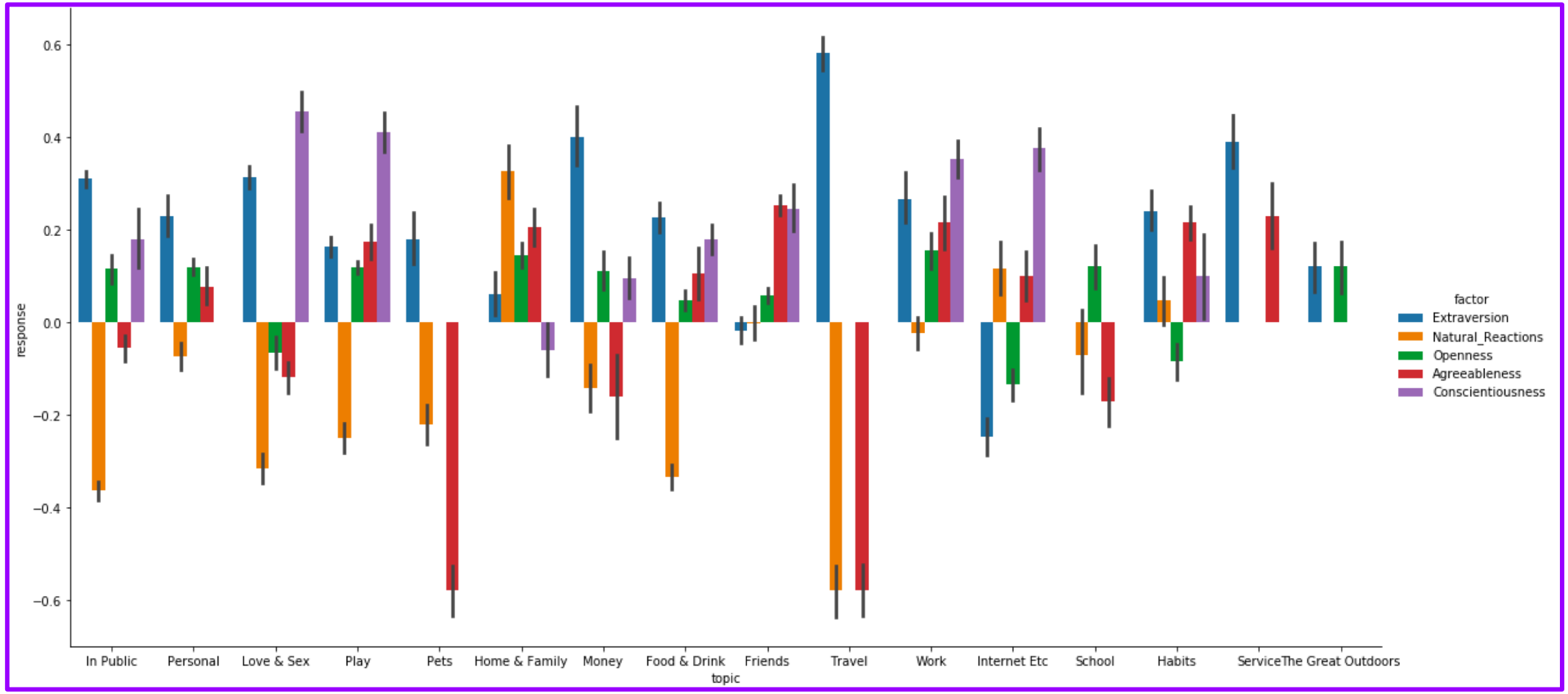
Data Visualization and Analysis



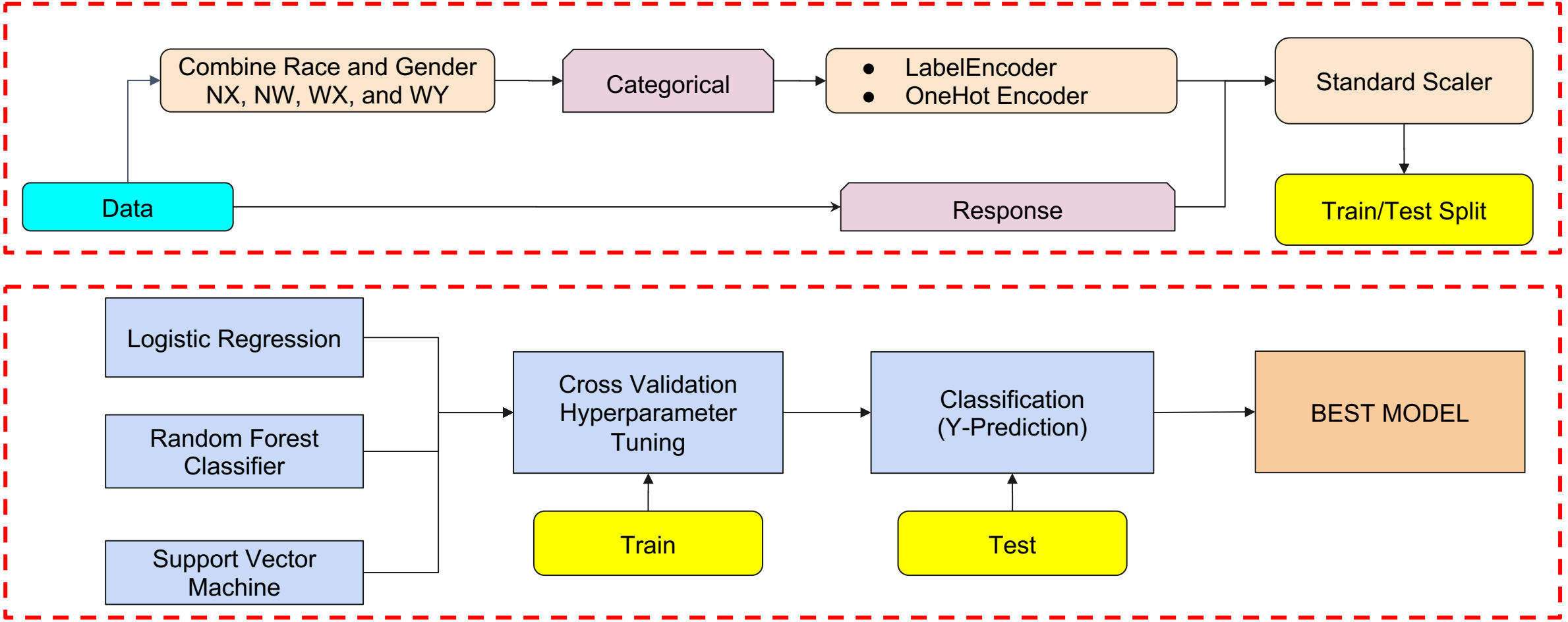


n = 50

Data Visualization and Analysis



Data Preparation for ML



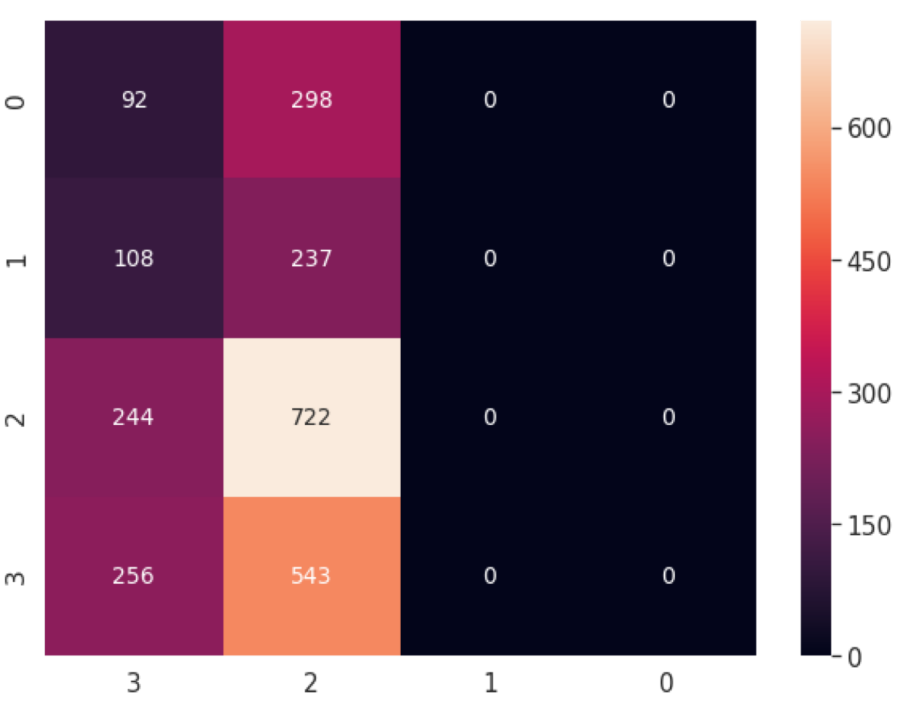
ML Models

- We wanted to see if we could come up with demographic & result prediction models
- Four (4) different approaches -
  - Predict race and gender of population based on their response to a question that is associated with a particular facet and topic.
    - X is response, topics and facets; Y is demographic.
  - Predict race and gender of population based on their response to a question that is associated with a particular primary factor and topic.
    - X is response, topics and primary_factor; Y is demographic.
  - Predict response of population based on their characteristics i.e. demography, topics and facets.
    - X is demographic, topics and facets; Y is response.
  - Predict response of population based on their characteristics i.e. demography, topics and primary factor.
    - X is demographic, topics and primary factor; Y is response.

Model 1- Predicting Demographic

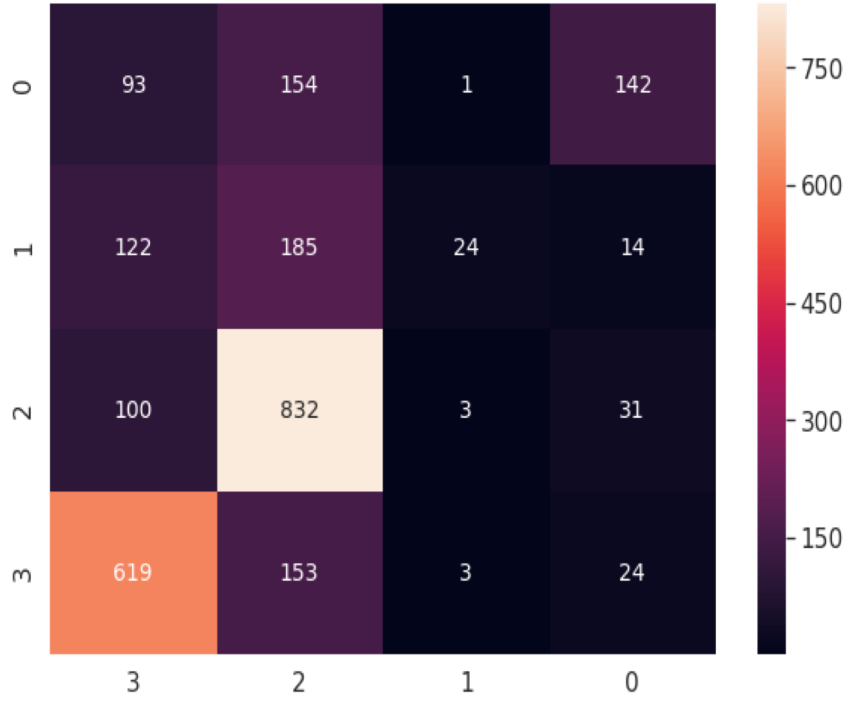
Predict demographic based on response, topic and facet.

- Best Model: {'solver': 'liblinear', 'penalty': 'l1', 'C': 0.033598182862837 81}
- Balanced accuracy on test set: 0.267



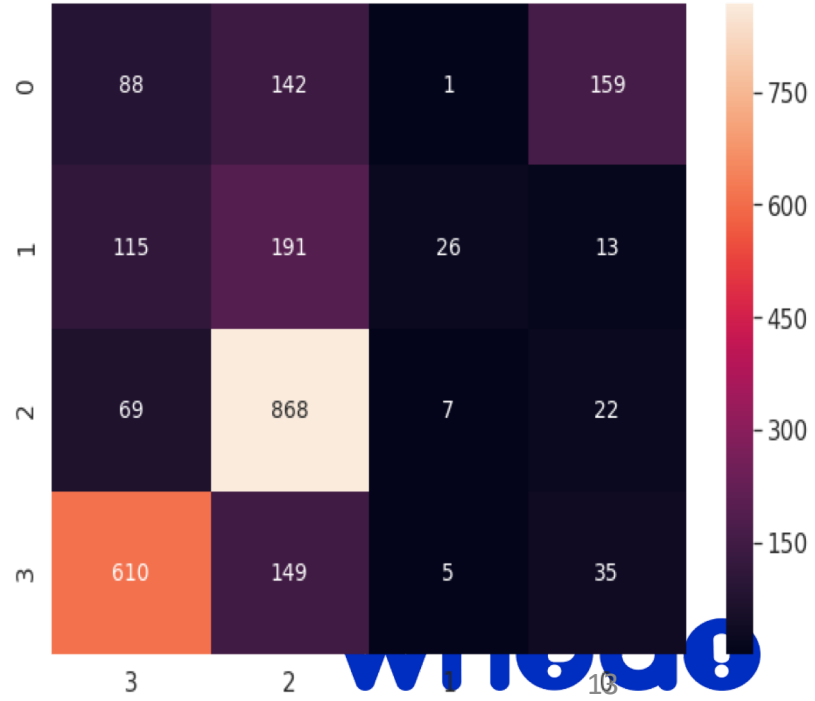
Logistic Regression

- Best Model: {'gamma': 0.1, 'C': 10}
- Balanced accuracy on test set: 0.517



SVC

- Best Model: {'n_estimators': 100, 'min_samples_split': 8, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 800, 'criterion': 'gini'}
- Balanced accuracy on test set: 0.536

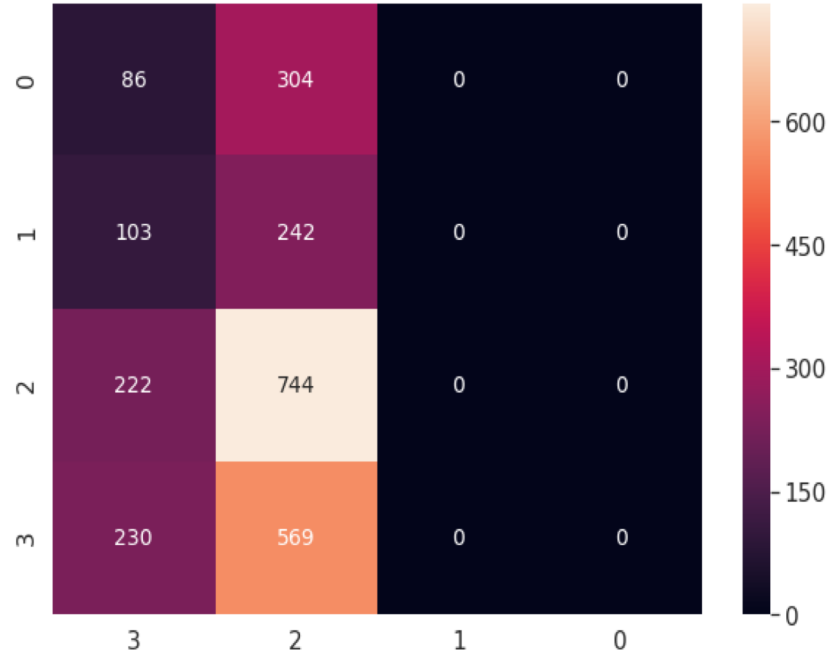


Random Forest

Model 2- Predicting Demographic

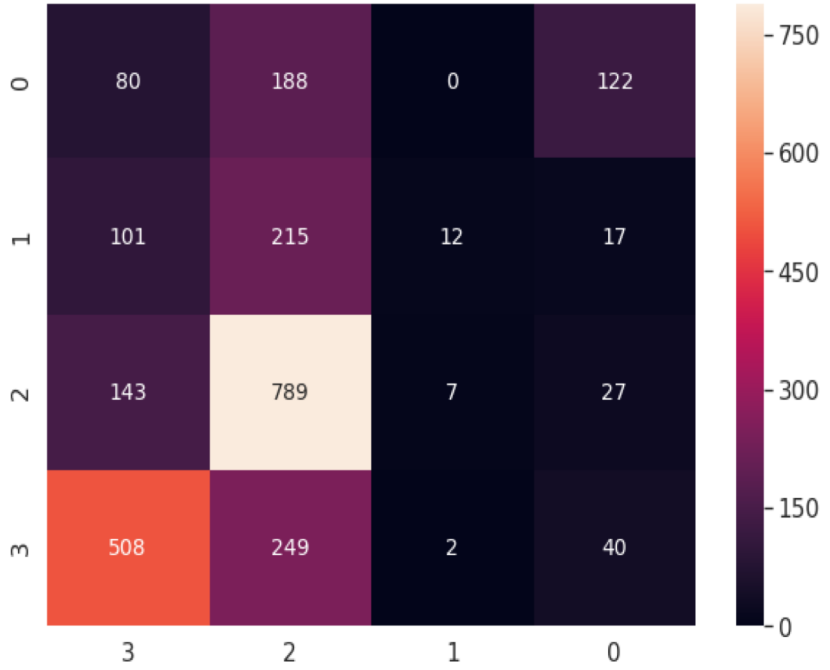
Predict demographic based on response, topic and primary factor.

- Best Model: {'solver': 'liblinear', 'penalty': 'l1', 'C': 0.08858667904100823}
- Balanced accuracy on test set: 0.264



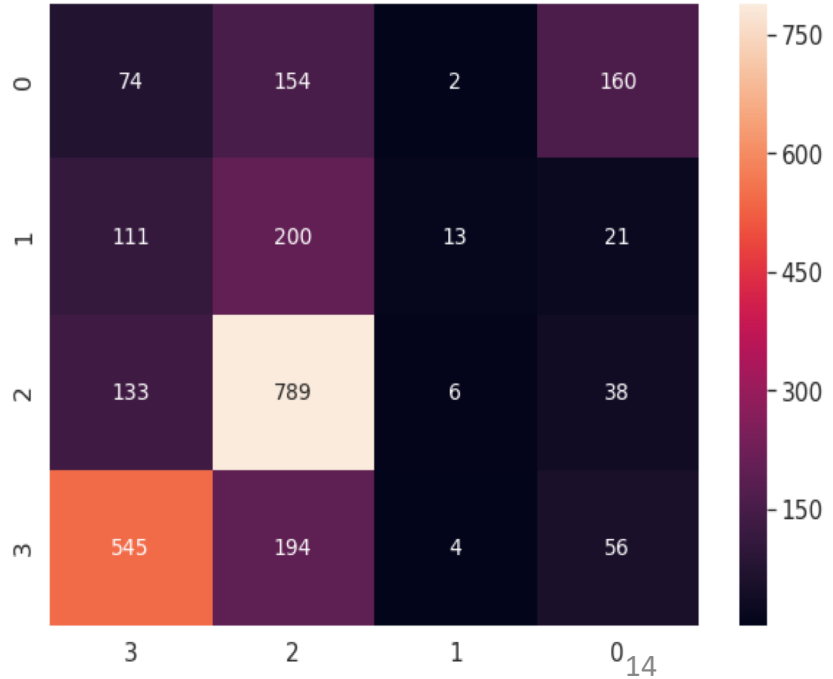
Logistic Regression

- Best Model: {'gamma': 0.1, 'C': 10}
- Balanced accuracy on test set: 0.450



SVC

- Best Model: {'n_estimators': 1050, 'min_samples_split': 4, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 1500, 'criterion': 'entropy'}
- Balanced accuracy on test set: 0.487

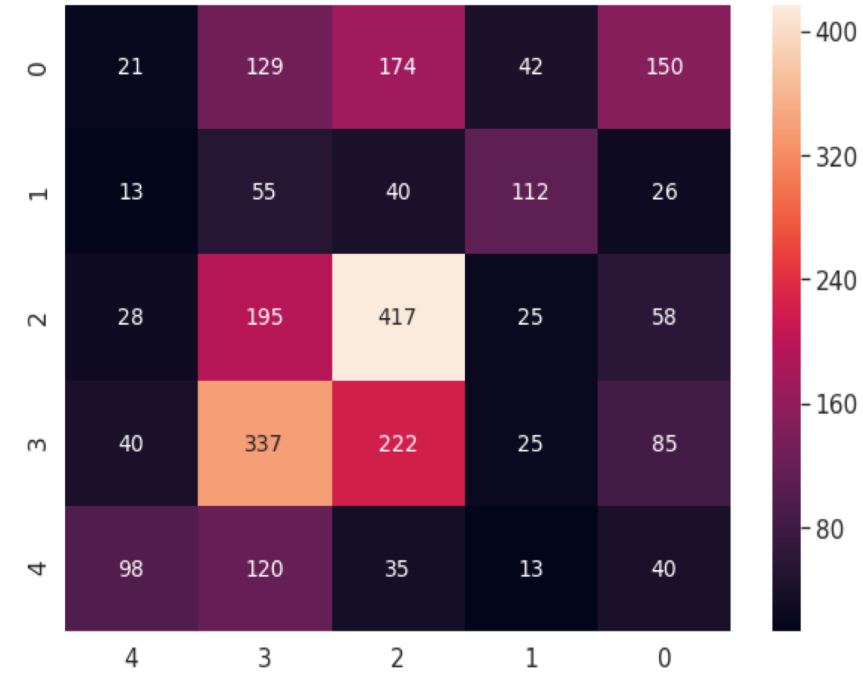


Random Forest

Model 3- Predicting Response

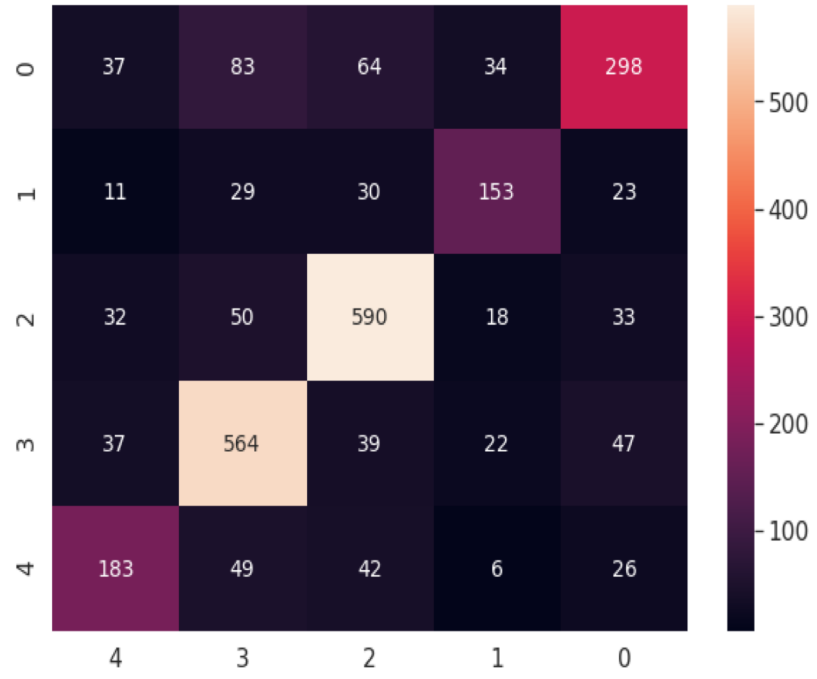
Predict response based on demographic, topics and facets.

- Best Model: {'solver': 'liblinear', 'penalty': 'l1', 'C': 11.288378916846883}
- Balanced accuracy on test set: 0.424



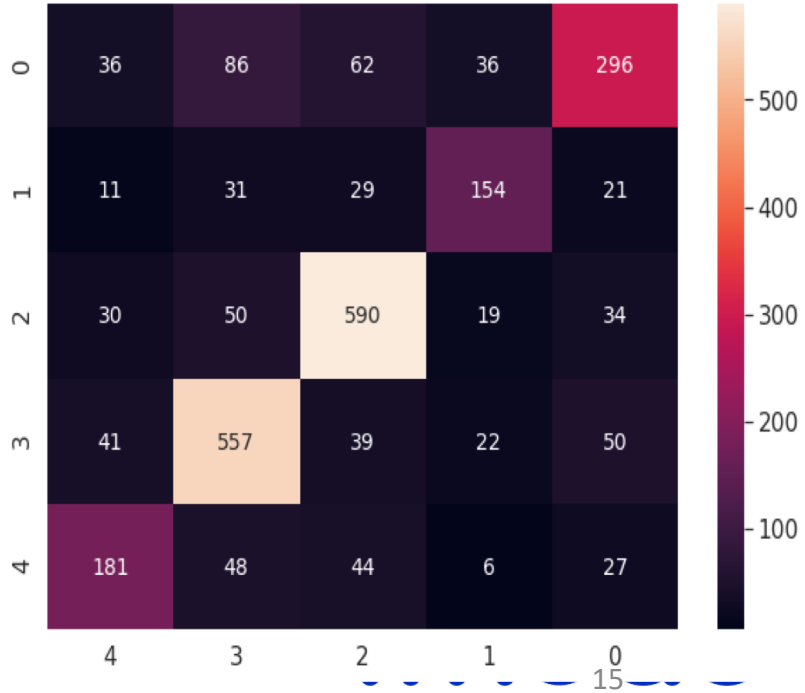
Logistic Regression

- Best Model: {'gamma': 0.1, 'C': 1}
- Balanced accuracy on test set: 0.682



SVC

- Best Model: {'n_estimators': 2000, 'min_samples_split': 8, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 1500, 'criterion': 'entropy'}
- Balanced accuracy on test set: 0.678

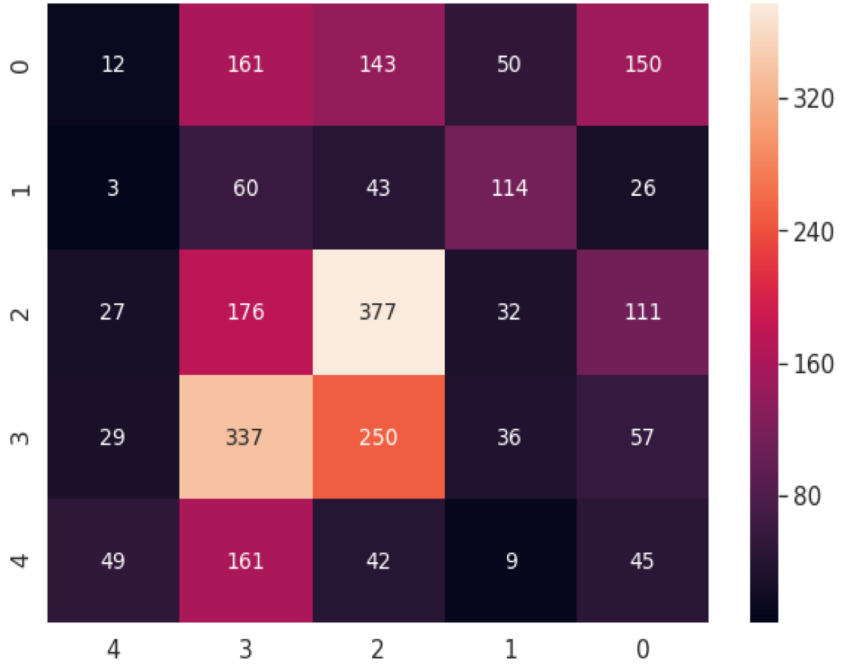


Random Forest

Model 4- Predicting Response

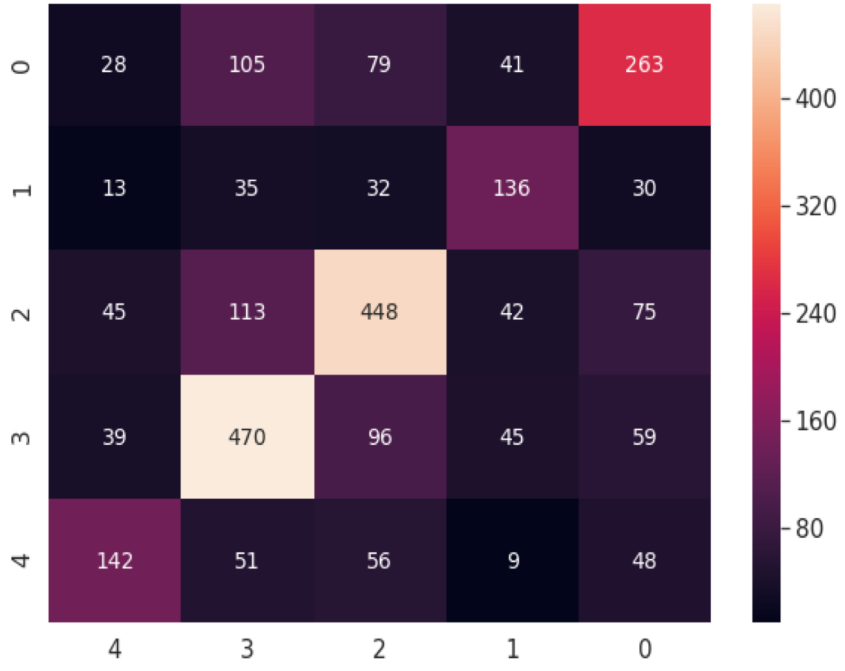
Predict response based on demographic, topics and primary factor

- Best Model: {'solver': 'liblinear', 'penalty': 'l2', 'C': 0.08858667904100823}
- Balanced accuracy on test set: 0.382



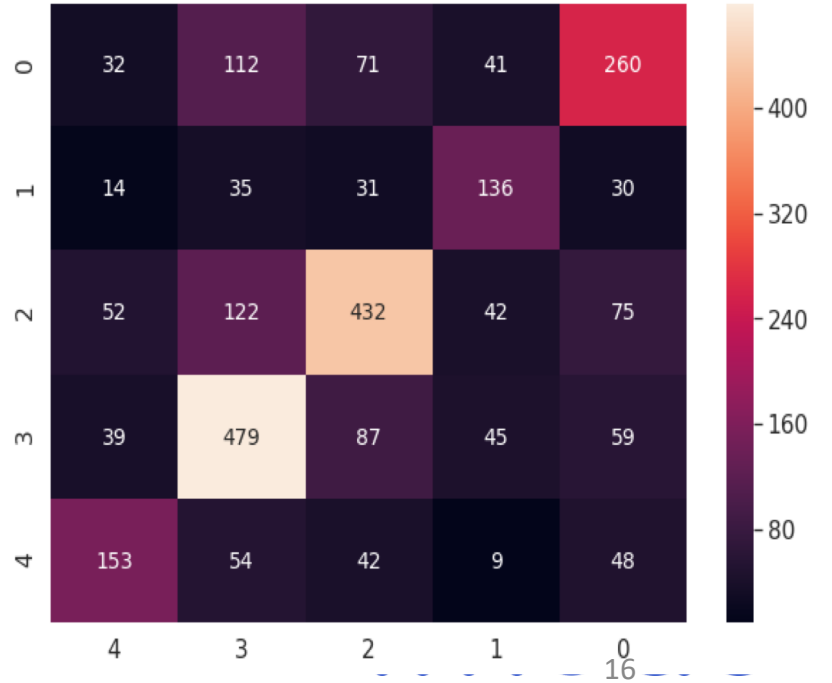
Logistic Regression

- Best Model: {'gamma': 0.1, 'C': 1}
- Balanced accuracy on test set: 0.562



SVC

- Best Model: {'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': None, 'criterion': 'gini'}
- Balanced accuracy on test set: 0.566



Random Forest

Results and Discussion

- Accuracy was the key element to define the best prediction:
    - For Race and Gender the Random Forest Classifier Model brought the best result (67% accuracy) based on responses associated with the following variables: Facet and Topic.
    - Even thought Random Forest and SVM Classifier showed similar accuracy, SVM Classifier was selected to predict the population responses based on the following characteristics: Demography, Topics, and Facets. Accuracy was of 71.5%.

Challenges

- Working with small sample size and many behavior classifications more combination of behavior classifications than samples;
- Defining what type of information is useful working with unusual database;
- Understanding customer demands and translating it to machine learning application;
- Defining the right parameters (features) and methodology to extract a good model;
- Settling the right and precise model.

Conclusion

- With the sample provided, we were able to build a model with at least 67% of accuracy considering the best scenario for each variable;

- If we increase the number of responses, improvements could be made in the performance of the model;

- Also with an increasing in samples size, other machine learning techniques could be applied and tested to try better results;

Thank You

whodo

whodo