Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	Gender Differences in Autism Screening Assessments
Requestor	Professor David Parent
Date of Request	July 5th 2022
Target Quarter for Delivery	Q3
Epic Link(s)	N/A
Impact	Improved methods of ASD assessment to increase accuracy of diagnosis.

1.0 Opportunity Brief

Autism Spectrum Disorder (ASD) is a neurological and developmental disorder that is characterized by difficulties in communication, learning, and social interaction. About 2.21% of adults in the U.S. have been diagnosed with this disorder, but this number is likely to be greatly underestimated as it is becoming clearer that many women with ASD have been incorrectly diagnosed with depression or anxiety, and it is theorized that autism manifests differently between males and females. According to some reports, women typically have high-functioning autism, allowing them to operate independently in the world, which is why their symptoms largely go unnoticed. Many of these women are highly successful and intelligent yet struggle to function in their everyday lives due to their undiagnosed disorders. In this analysis, we will be examining the data to determine if there are key differences between ASD in males and females that will indicate a need to develop more nuanced measures of assessment, specified for both genders.

……

The specific ask:

Using a variety of modeling techniques, I will be conducting an analysis to determine the behaviors that are most indicative of ASD and whether these behavioral patterns can be differentiated by gender, age, ethnicity, etc.

1.1 Supporting Insights

- 1 in 44 children are diagnosed with ASD.
- 2.21% of adults have ASD in the United States.
- Men are four times more likely to be diagnosed with autism than women.
- Research on ASD has focused on male participants, thus symptoms for diagnosis are biased towards men.
- ASD tests and models do not account for the possibility of gender differences in symptoms.
- Autistic women use “masking” a term that describes mimicking neurotypical behavior to fit in. This makes their condition less obvious than men.

1.2 Project Gains

ASD comes with many challenges. Individuals with this disorder struggle with social interactions, communication, and learning capabilities which makes it difficult for them to navigate society. These difficulties are exacerbated if the individual is unaware of their disorder. Getting a diagnosis and learning about how your brain differs from neurotypical individuals allows people to learn special techniques to deal with their problems and utilize their differences in a positive way. Unfortunately, many women with ASD have been misdiagnosed with ADHD, anxiety, or depression because their autism manifests differently from men on whom studies of ASD center around. By examining whether men and women respond differently to ASD assessments we can develop new methods of assessment that take gender into account, helping women with ASD receive the proper diagnosis.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.

2.0 Analytics Objective

Currently, ASD assessments are standardized for all individuals seeking an ASD diagnosis. The question I am trying to answer is whether we can develop more accurate methods of screening if we consider possible differences in gender or other demographics. To answer this, I will be attempting to solve the following questions throughout my analysis:

- Which behaviors measured in the AQ-10-Adult are most indicative of ASD?
- Do females display different combinations of behaviors than men?
- If there are gender differences, can they further be differentiated by age, ethnicity, country of residence, etc.?
- If there are no gender differences, can they still be differentiated by age, ethnicity, country of residence, etc.?

2.1 Other related questions and Assumptions:

Assumptions:

- Questionnaire responders are honest when completing their assessments and providing demographic information.
- Without an ID column we must assume all records are unique, even if they may seem like a duplicate.

2.2 Success measures/metrics

- Creating models with high accuracy above 80%.
- Determining key differences in questionnaire responses between males and females.
- Determining demographic factors that further differentiate responses on the questionnaire.
- Key insights into how to create a more specific and accurate screening test for ASD that accounts for gender differences and other demographic differences.

2.3 Methodology and Approach

Type of Analysis: *Decision Tree, Random Forest, Logistic Regression, Linear Regression, Forward Inclusion Regression, Backward Exclusion Regression, Neural Network, and Association Rules.*

The initial approach will be to use a Decision Tree to determine which ASD variables (gender and behavior score) are most significantly related to an ASD diagnosis. I will also use other techniques to verify my findings.

Methodology: *Key questions from 'Analytics objective' will be tackled in ascending order as outlined in '5.0 Timelines and deliverable section'.*

I will start by conducting a blind analysis using the Decision Tree, Random Forest, Regression models, and Neural Network model to identify key trends in the data without bias. I will look specifically at which behaviors, as denoted by question scores, are most closely related to a positive ASD diagnosis and whether gender is a significant factor in assessment outcomes. I will then split the dataset into two subgroups by gender and rerun all the models to see how they compare with the full dataset analysis and how they compare to one another. I will then use Association Rules to see whether there are relationships between scores for any of the behaviors and whether these differentiate by gender as well. Finally, I will include other variables (age, ethnicity, jaundice…) into the models to determine whether they are significant for an ASD diagnosis.

Output: *The output will be a set of insights, rules, and strategic recommendations that will help us to create more effective and nuances screening methods by taking into account differences in demographics, specifically gender.*

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: Participants of the AQ-10-Adult questionnaire.

Observation window: 2017-12-24

Inclusions: Full Dataset

Exclusions: None

Data Sources: UCI Machine Learning Repository; Kaggle

Audience Level: Clinical Practitioners and Researchers

Variable Selection:

Table 1: Features and their descriptions

Attribute	Type	Role	Description
Age	Number	Input	Age in years
Age Description	String	Rejected	Description of age category (18 and above)
Gender	String	Input	Male or Female
Ethnicity	String	Input	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Input	Whether the case was born with jaundice
Who is completing the test	String	Input	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	Input	List of countries in text format
Used the screening app before	Boolean (yes or no)	Input	Whether the user has used a screening app
Question 1 Answer	Binary (0, 1)	Input	I often notice small sounds when others do not
Question 2 Answer	Binary (0, 1)	Input	I usually concentrate more on the whole picture, rather than the small details
Question 3 Answer	Binary (0, 1)	Input	I find it easy to do more than one thing at once
Question 4 Answer	Binary (0, 1)	Input	If there is an interruption, I can switch back to what I was doing very quickly
Question 5 Answer	Binary (0, 1)	Input	I find it easy to 'read between the lines' when someone is talking to me
Question 6 Answer	Binary (0, 1)	Input	I know how to tell if someone listening to me is getting bored
Question 7 Answer	Binary (0, 1)	Input	When I'm reading a story I find it difficult to work out the characters' intentions
Question 8 Answer	Binary (0, 1)	Input	I like to collect information about categories of things
Question 9 Answer	Binary (0, 1)	Input	I find it easy to work out what someone is thinking or feeling just by looking at their face
Question 10 Answer	Binary (0, 1)	Input	I find it difficult to work out people's intentions
Screening Score	Integer	Target	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
Class/ASD	String	Target	'YES' or 'NO' ASD Classification

Derived Variables: None

Assumptions and data limitations: Data may not be representative of the entire population as it was collected from a downloadable app. Therefore, the sample will only include individuals capable of downloading and operating the app.

4.0 Dependencies and Risks

There are no upstream or downstream dependencies for this project.

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact
Inability to find a peer for peer evaluation during the analysis.	*Low*	*2-3 days*	*Cannot continue analysis without feedback from a peer. This will delay progress by a few days.*

5.0 Deliverable Timelines

Item	Major Events / Milestones	Description	Days	Date
1.	Kick-off / Formal Request	*Present project approach and dataset to my academic advisor for approval.*	1	July 5th
2.	Project Approval	*Project approach approved by my academic advisor.*	1	July 7th
3.	Analytics Plan	*Create an outline of my analysis, including purpose, methodology, timelines for completion, etc. submitted as a document.*	3	July 15th
4.	Data Exploration & Analysis	*Assess data quality, data structure, variable properties, and plot the data using python.*	5	July 17th
5.	Data Modeling	*Model the data with python using Decision Trees, Random Forest, Regressions, Neural Network, and Association Rules.*	20	August 5th
6.	Governance	*Identify possible sources of risk or bias in data modeling. Define a set of logical rules to ensure the models are effective in interpreting the data. Submit as a document.*	7	August 5th
7.	Documentation	*Document each stage of the analysis using screenshots of the EDA and Models with clear*	20	August 8th

		descriptions. Submit as a document.		
8.	Create Presentation	*Compile my analysis and create a presentation of my work and findings. Format as PowerPoint.*	3	August 11th
	Peer Feedback	*Provide feedback on a peer's analysis and presentation. Receive feedback from a peer on my analysis and presentation.*	2	August 12th
9.	Presentation	*Present my analysis and results to my academic advisor. Submit as PowerPoint.*	1	August 24th