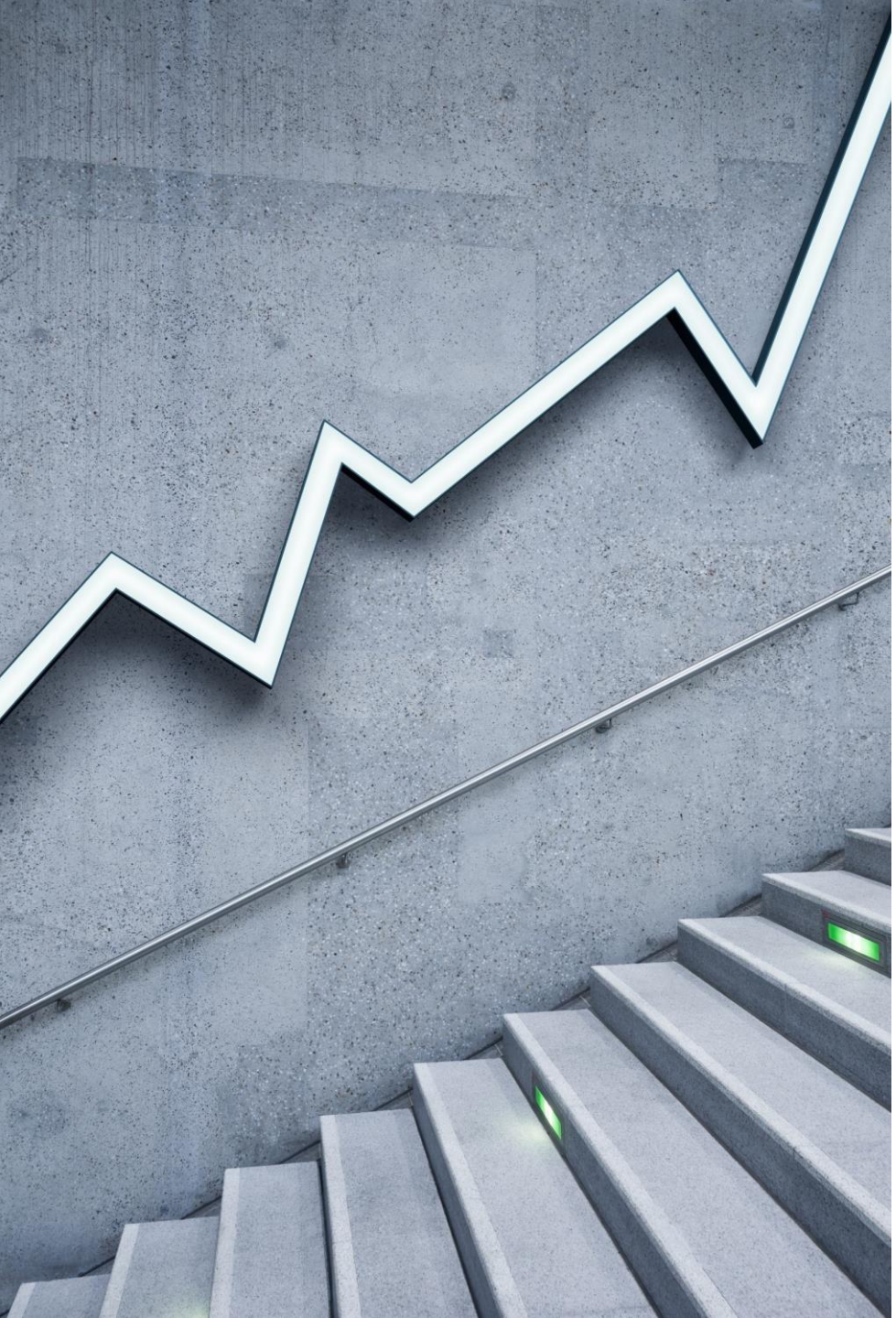


Sprint 2:
Predicting Depression Using Demographic and Lifestyle Data

Data Science Capstone Project

BrainStation

Gennaro Costantino

October 2024

Initial Approach & Dataset Shift

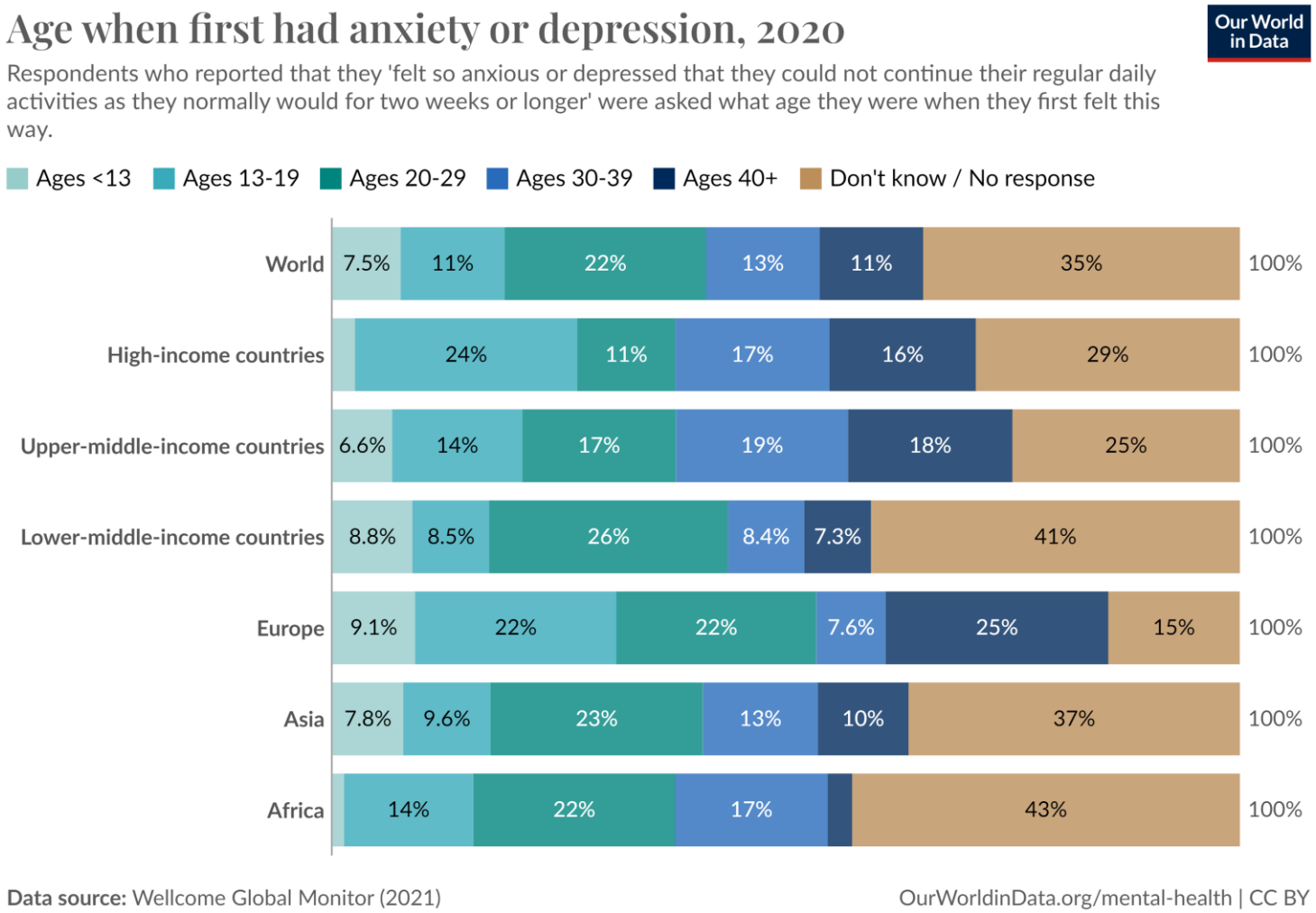
- Initially started with a text-based sentiment analysis approach using social media data.
- Challenge: The lack of a target variable led to a shift in the project direction.
- New Direction: Pivoted to using a structured depression dataset from Kaggle, which contain a target variable related to mental health outcomes.

New Direction - Key Features + Target:

- **Age**: The age of the respondent.
- **Income**: Standardized income data for the individual.
- **Smoking Status**: Whether the respondent is a current, former, or non-smoker.
- **Physical Activity Level**: Indicates whether the respondent has a sedentary, moderate, or active lifestyle.
- **Employment Status**: Reflects whether the individual is employed or unemployed.
- **History of Mental Illness**: Indicates whether the individual has a history of mental illness. (TARGET)
- **Family History of Depression**: Information about the presence of depression in the respondent's family.
- **Chronic Medical Conditions**: Whether the individual suffers from chronic health conditions.
- **Alcohol Consumption & Dietary Habits**: Data on the individual's alcohol consumption and dietary preferences.

Problem Statement

- Mental health disorders, particularly depression, are on the rise globally.
- Early prediction of individuals at risk can help in timely interventions.
- The goal of this project is to predict depression risks using demographic and lifestyle data.



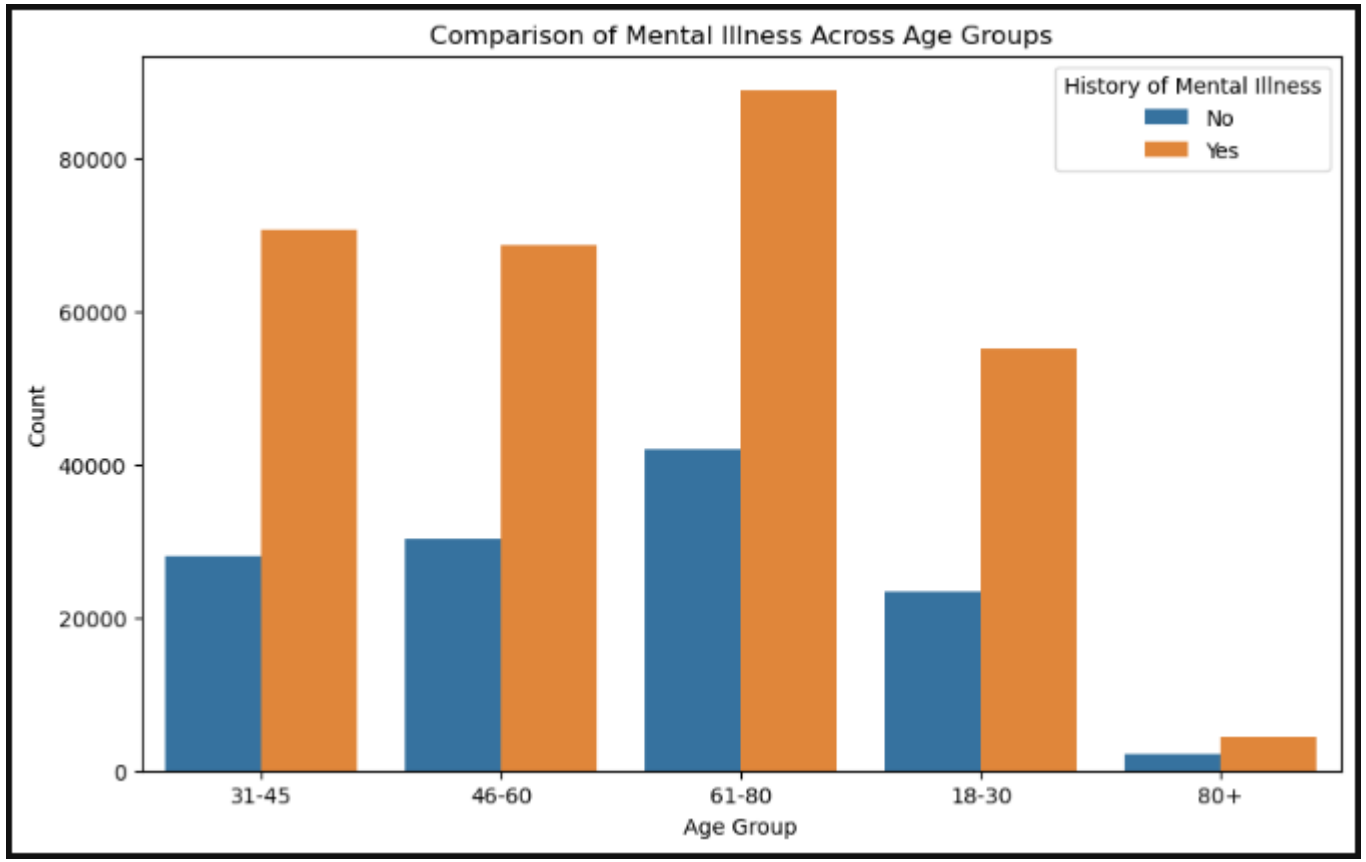
Key Insights from EDA

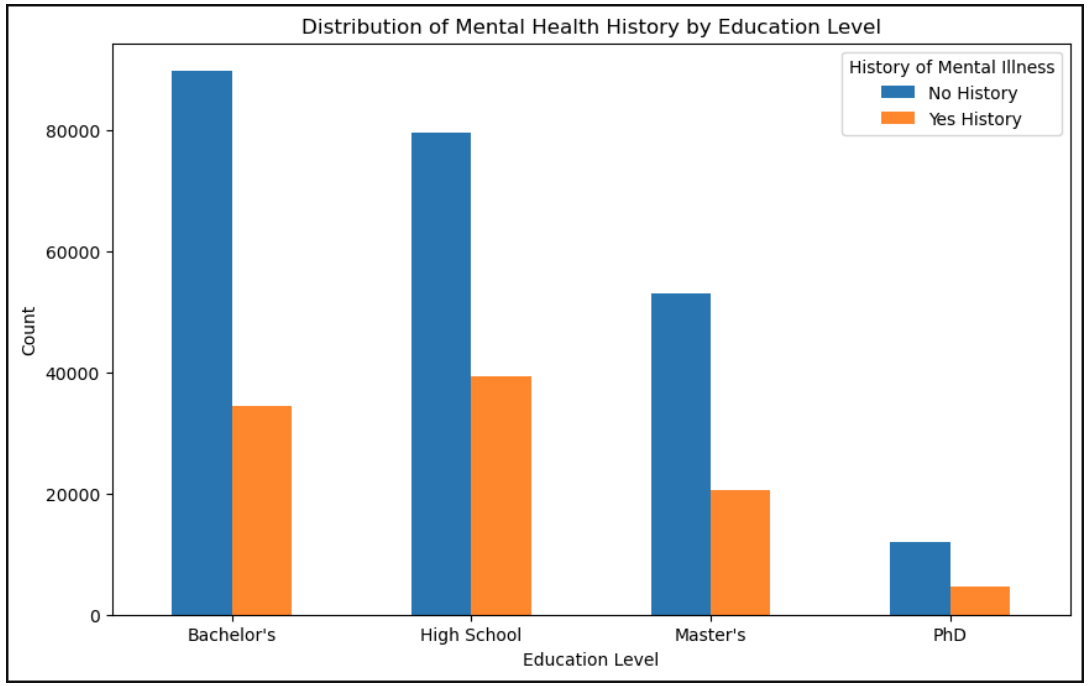
Age: Older individuals (61+) showed a higher likelihood of reporting mental health issues.

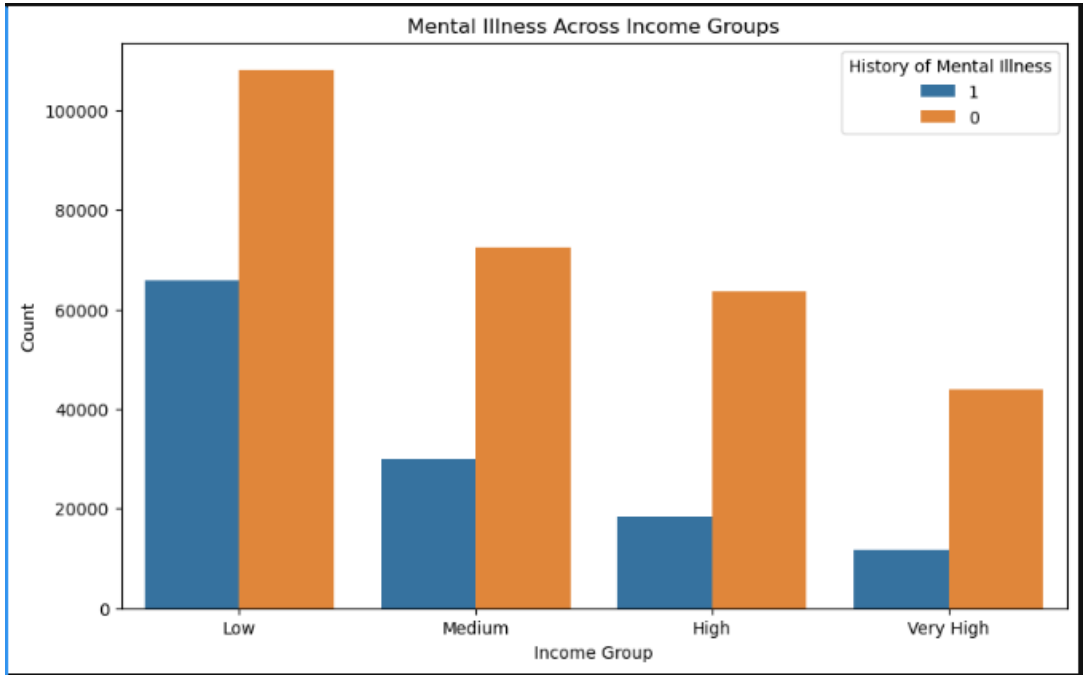
Income: Lower income groups had a higher proportion of mental illness.

Employment: Those unemployed or inactive had higher risks of mental illness.

Other factors: Physical activity and alcohol consumption patterns showed distinct differences. PENDING




Distribution of Mental Health History by Education Level
History of Mental Illness
No History
Yes History
Count
0
20000
40000
60000
80000
Bachelor's
High School
Master's
PhD
Education Level


Mental Illness Across Income Groups
History of Mental Illness
1
0
Count
0
20000
40000
60000
80000
100000
Low
Medium
High
Very High
Income Group

Feature Engineering & Preprocessing

- Handling missing values, encoding categorical features, and scaling numerical columns like income.
- One-hot encoding applied to marital status and education levels.
- Data split into training and testing sets for model building.

```
Data columns (total 16 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Name                        413768 non-null  object
 1   Age                         413768 non-null  int64
 2   Marital Status              413768 non-null  object
 3   Education Level             413768 non-null  object
 4   Number of Children          413768 non-null  int64
 5   Smoking Status              413768 non-null  object
 6   Physical Activity Level     413768 non-null  object
 7   Employment Status           413768 non-null  object
 8   Income                      413768 non-null  float64
 9   Alcohol Consumption         413768 non-null  object
 10  Dietary Habits              413768 non-null  object
 11  Sleep Patterns              413768 non-null  object
 12  History of Mental Illness   413768 non-null  object
 13  History of Substance Abuse  413768 non-null  object
 14  Family History of Depression  413768 non-null  object
 15  Chronic Medical Conditions  413768 non-null  object
dtypes: float64(1), int64(2), object(13)
memory usage: 50.5+ MB
```

```
a columns (total 21 columns):
    Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   Name                                  413768 non-null  object
 1   Age                                   413768 non-null  int64
 2   Number of Children                    413768 non-null  int64
 3   Smoking Status                        413768 non-null  int64
 4   Physical Activity Level               413768 non-null  int64
 5   Employment Status                     413768 non-null  int32
 6   Income                                413768 non-null  float64
 7   Alcohol Consumption                   413768 non-null  int64
 8   Dietary Habits                        413768 non-null  int64
 9   Sleep Patterns                        413768 non-null  int64
 10  History of Mental Illness             413768 non-null  int32
 11  History of Substance Abuse            413768 non-null  int32
 12  Family History of Depression          413768 non-null  int32
 13  Chronic Medical Conditions            413768 non-null  int32
 14  Marital Status_Married                413768 non-null  bool
 15  Marital Status_Single                 413768 non-null  bool
 16  Marital Status_Widowed                413768 non-null  bool
 17  Education Level_Bachelor's Degree     413768 non-null  bool
 18  Education Level_High School           413768 non-null  bool
 19  Education Level_Master's Degree       413768 non-null  bool
 20  Education Level_PhD                   413768 non-null  bool
dtypes: bool(7), float64(1), int32(5), int64(7), object(1)
memory usage: 39.1+ MB
```

Model Selection – 1st interaction

- Logistic Regression
- Random Forest

- Performance Metrics: Accuracy, Precision, Recall, F1-score

```
Logistic Regression Results:
Accuracy: 0.6959019100788683
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     86383
           1       0.00      0.00      0.00     37748

    accuracy                           0.70    124131
   macro avg       0.35      0.50      0.41    124131
weighted avg       0.48      0.70      0.57    124131
```

```
Random Forest Classifier Results:
Accuracy: 0.6631381363237225
              precision    recall  f1-score   support

           0       0.70      0.90      0.79     86383
           1       0.35      0.13      0.19     37748

    accuracy                           0.66    124131
   macro avg       0.53      0.51      0.49    124131
weighted avg       0.60      0.66      0.60    124131
```

Hyperparameter Tuning – NEXT STEPS

Parameters tuned for Logistic Regression:

- Regularization strength (C)
- Solver (lbfgs, liblinear)

Explore Neural Networks to improve prediction accuracy.

Generate new features from existing data, such as age and income brackets.

Deployment: Build a dashboard or web app for real-time monitoring of individuals at risk. EXTRA!

Continue to evaluate model generalization with new test sets.

Thank you

GC