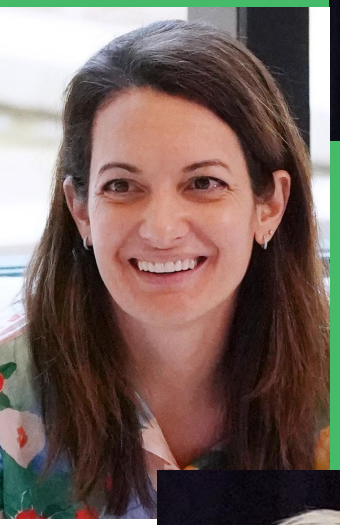
Women in Data Science Worldwide

WiDS Datathon 2025
#WiDSDatathon

Kylie Cancilla
Caterina Ponti
Liana Mendoza

Workshop Goals

- Building the Multi-output Model
- Feature Selection
- How to Evaluate the Model
- Explaining the F1 Score



Women in
Data Science
Worldwide

WiDS Mission

To change the field of data science across the globe by elevating, educating, and empowering women to achieve 30% representation of women in data science by 2030.

Women in Data Science Worldwide

WiDS Datathon Challenge Task

This datathon challenge aims to answer this question:

What brain activity patterns are associated with ADHD; are they different between males and females, and, if so, how?

To work towards the answer to this question, participants will be tasked with building a **multi outcome model to predict both an individual's sex and their ADHD diagnosis** using functional brain imaging data of adolescents and their socio-demographic, emotions, and parenting information.

A multi-outcome model is designed to predict multiple target variables simultaneously using a single machine learning model.

(female:1, male:0, ADHD: 1 ; no ADHD 0)

Multi-Outcome Model

Women in Data Science Worldwide

Reminder of the Challenge Question:

To predict both target variables simultaneously, we will use a multi-outcome machine learning model. This type of model is designed to predict multiple dependent variables (or outcomes) at the same time, rather than one at a time. For this task, our Y_train dataset will include two target columns—gender and ADHD diagnosis.

Benefits:

- Lets us take into account ADHD and gender simultaneously since we are trying to notice the pattern in ADHD predictions in females. Using the same feature variables to predict both targets allows us to examine this connection in a systematic way.
- Streamlines Workflow: Data preprocessing, model fitting, training and testing only needs to occur once to predict two variables

Cross-Validation to Evaluate the Model

Women in Data Science Worldwide

How Cross-Validation Works:

- The dataset is divided into k equally-sized subsets or folds.
- The model is trained on k - 1 folds (the training set) and evaluated on the remaining fold (the validation set)
- This process is repeated k times, with each fold being used once as the validation set
- The evaluation metric, we will use accuracy score, computed for each iteration

Why are we using cross-validation score:

- Since we don't have the true labels for the test data, we cannot calculate the actual accuracy score. The accuracy score is the difference between the true y and predicted y. Cross-validation provides a reliable way to evaluate our model's performance using the training data without needing the test set answers.
- Cross-validation ensures that a model is evaluated on multiple subsets of the data

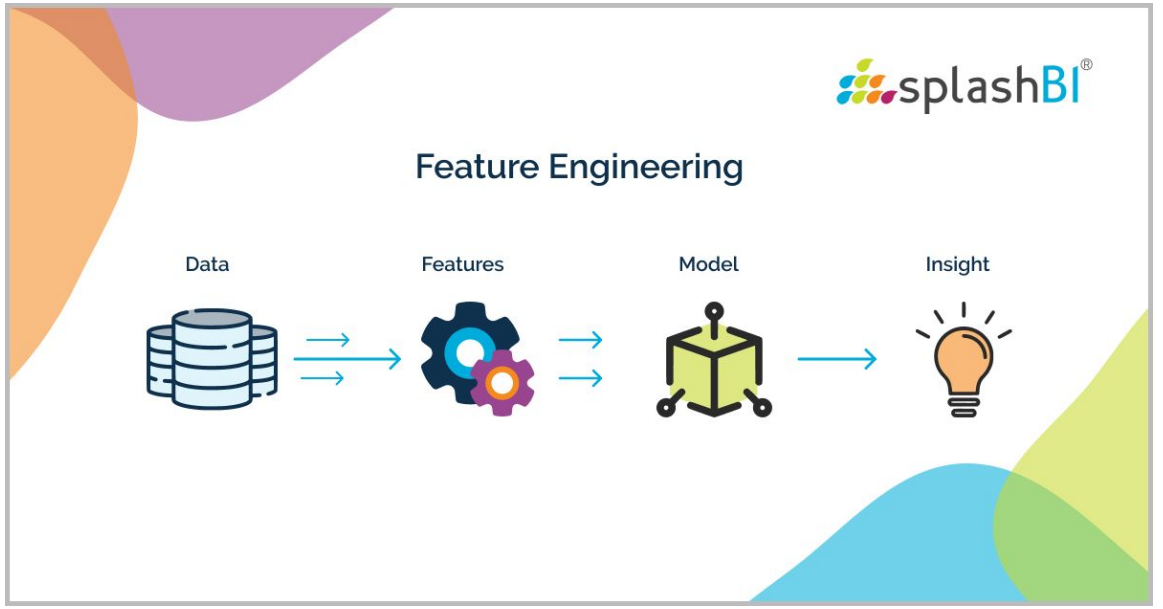
Features Engineering

Women in Data Science Worldwide

Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. In other words, it is the process of **selecting, extracting, and transforming the most relevant features** from the available data to build more accurate and efficient machine learning models.

Processes Involved in Feature Engineering:

1. Feature Creation
2. Feature Transformation
3. Feature Extraction
4. Feature Selection
5. Feature Scaling



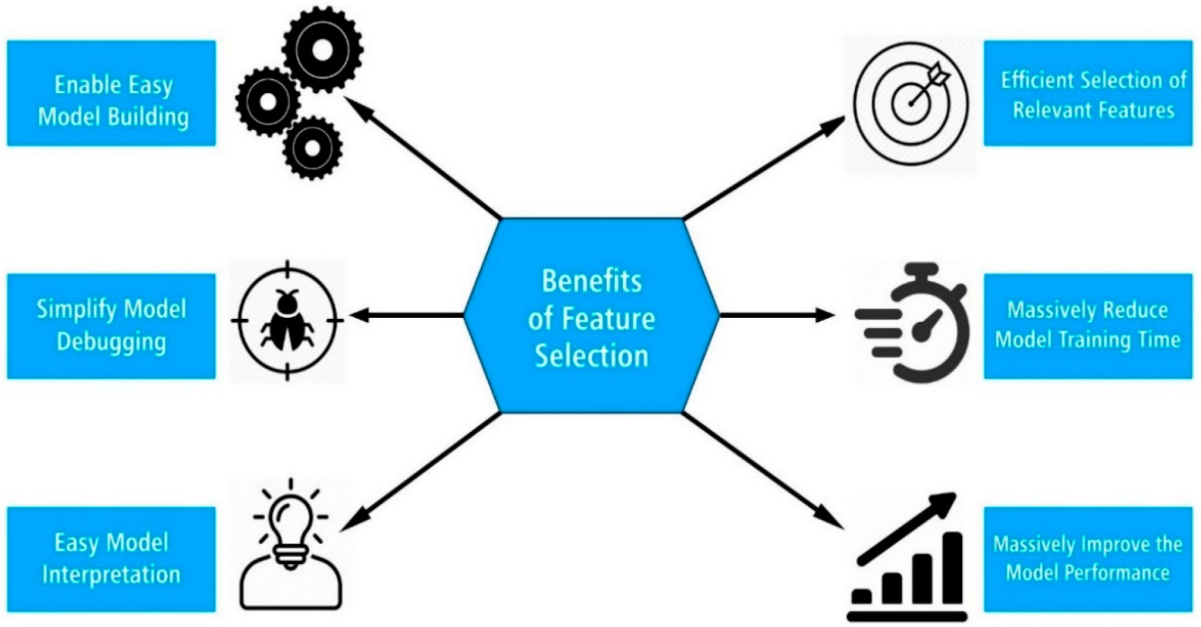
What is Feature Selection?

Women in Data Science Worldwide

Feature selection, one of the main components of feature engineering, is the process of **selecting the most important features to input in machine learning algorithms**. The main goal of feature selection is to **improve** the **performance** of a predictive model and reduce the computational cost of modeling.

Benefits:

- Simpler models
- Shorter training times
- Variance reduction
- Avoid the curse of high dimensionality



Logistic Regression for Features Selection

Women in Data Science Worldwide

Logistic Regression is a statistical method used for predicting the probability of a binary outcome (e.g., yes/no, 1/0) based on one or more input features. It's widely used in classification problems.

Why Use Logistic Regression for Feature Selection?

Logistic Regression not only predicts outcomes, but it also provides useful information about how each feature affects the prediction.

How Does Logistic Regression Help in Feature Selection?

- **Coefficients:** Each feature in the model has a coefficient that indicates how strongly that feature is related to the outcome. Larger absolute values of coefficients indicate more important features.
- **Regularization:** Using L1 regularization can further improve feature selection by shrinking less important feature coefficients to zero.

Two Ways to Implement Logistic Regression for Feature Selection

Women in Data Science Worldwide

1. Using Coefficients:

- **Fit the Model:** Train a logistic regression model to predict the outcome.
- **Extract Coefficients:** The model generates coefficients (numbers) for each feature. These coefficients represent the strength and direction of the relationship between each feature and the target variable.
- **Select Top Features:** The absolute value of each coefficient tells you how important that feature is for the prediction. Larger absolute values indicate more important features. You can select the top features by picking the ones with the highest absolute coefficients.

2. Using L1 Regularization:

- **Train the Model with L1 Regularization:** L1 regularization helps simplify the model by shrinking less important features' coefficients toward zero.
- **Select Non-Zero Features:** After training, features with non-zero coefficients are considered important. These are the features you want to keep for prediction.

F1 Score

Women in Data Science Worldwide

What is the F1 score metric?

- An evaluation metric that will be utilized in order to determine the performance of your model
- Highest possible score is 1, lowest possible score is 0
- Combines two parameters: precision and recall

What is precision?

- Ratio of true positives among predicted positives (false positives and true positives)
- Diagnosed individuals among predicted diagnosed individuals
- Measures the accuracy of positive predictions

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{\text{True Positives}}{\text{All PREDICTED Positives}}$$

F1 Score

Women in
Data Science
Worldwide

What is recall?

- Ratio of true positives among actual positives
- Ratio of actual diagnosed individuals among actual positives (true positives and false negatives)
- Measures the ability of your model to predict true positives

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{\text{True Positives}}{\text{All ACTUAL Positives}}$$

F1 Score

Women in Data Science Worldwide

What is harmonic mean?

- Equally combines the recall and precision factors
- Harmonic mean ensures that these two factors are weighed together equally or symmetrically, despite their differences in the denominator.

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

The weighted F1 Score is calculated on each column, and those two individual scores are averaged to get the final Kaggle leaderboard score. ADHD is harder to predict in females, thus female ADHD cases have higher weight.

Women in Data Science Worldwide

Timeline

- **January 7, 2025** Start Date. Register here with WiDS Worldwide to participate in the datathon.
- **April 16, 2025** Leaderboard Prize form opens. Fill out a brief form for students, high schoolers & first-timer participants to be eligible for prizes.
- **April 27, 2025** Last day participants may join or merge teams.
- **April 30, 2025** Final Submission Deadline. Prize form also closes at this time
- **May 7, 2025** Winners announced. Top performers may also be recognized throughout the competition.

All deadlines are at **11:59 PM UTC** on the corresponding day unless otherwise noted.

THANK YOU!

Questions?

Women in Data Science Worldwide