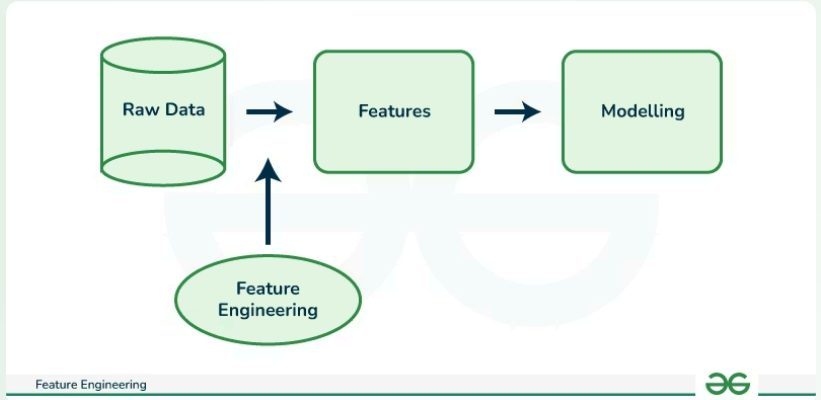


Feature Engineering

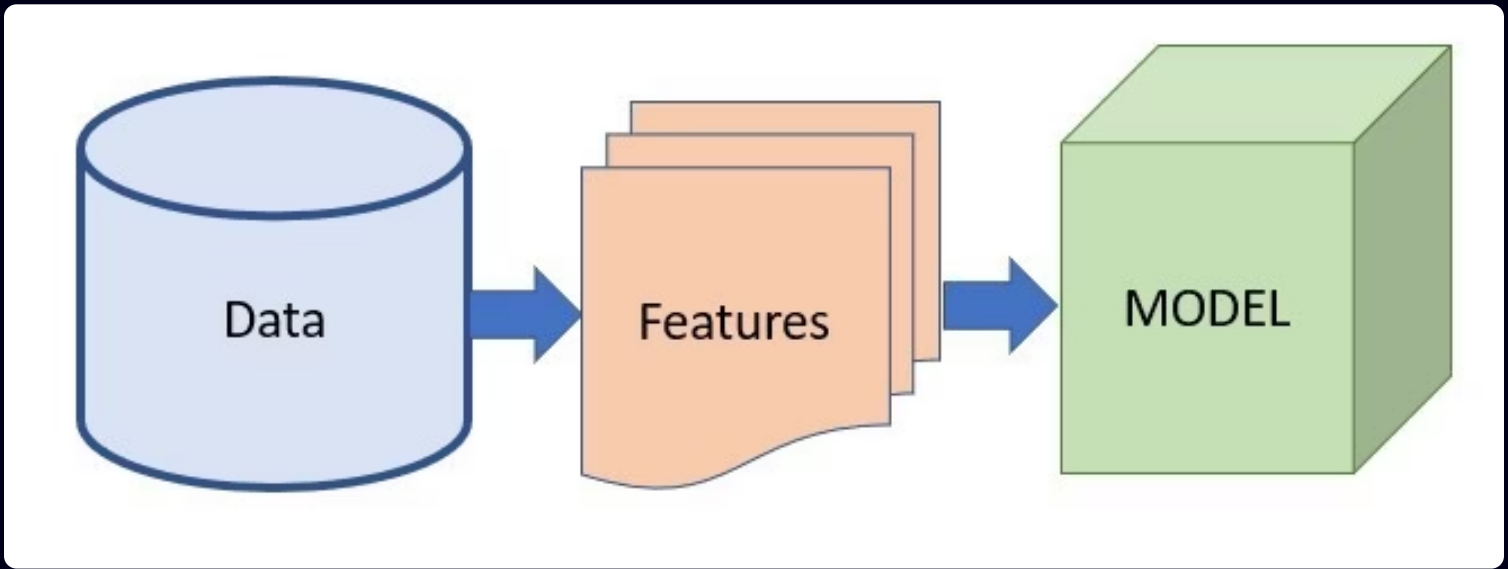
Made with GAMMA

Attendance: "Manual Feature Engineering"

Made with GAMMA

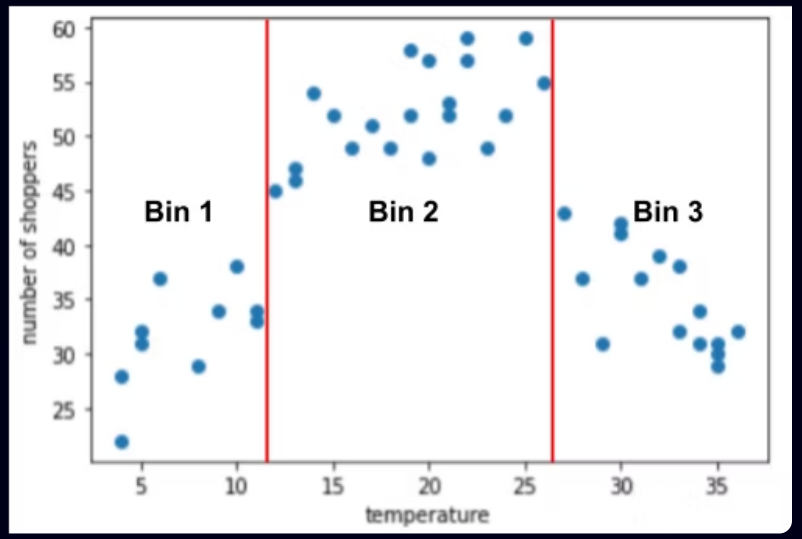
What is a Feature?

- **Features** are the individual, measurable properties or attributes of a dataset – essentially, the columns.
- They serve as the input variables that machine learning models use to learn patterns and make predictions.
- The quality and relevance of these features are critical, directly influencing a model's accuracy and effectiveness.



Made with GAMMA

What is Feature Engineering?



Data Transformation

The crucial process of selecting, creating, and transforming raw data into meaningful and predictive features

Enhanced Performance

It significantly improves accuracy, enhances interpretability, and boosts the overall efficiency

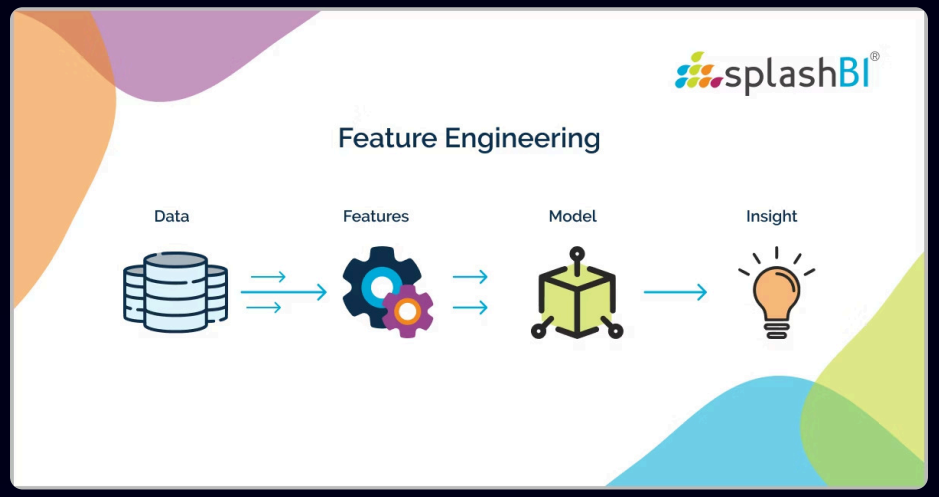
Diverse Techniques

Involves a wide array of techniques including encoding categorical variables, binning, feature splitting, lag, and sliding window features

Made with GAMMA

Why Manual Feature Engineering Matters

- Leverages deep domain knowledge to design impactful features that are often missed by automated approaches.
- Consistently outperforms automated methods, especially in niche or highly complex datasets where human intuition is invaluable.
- Provides better understanding and granular control over how data is represented to the learning algorithm.



Made with GAMMA

Core Manual Feature Engineering Techniques

1

Feature Creation

Deriving entirely new features by combining or manipulating existing ones (e.g., calculating ratios or differences).

2

Feature Transformation

Modifying existing features through techniques like encoding of categorical variables (One-Hot, Label Encoding).

3

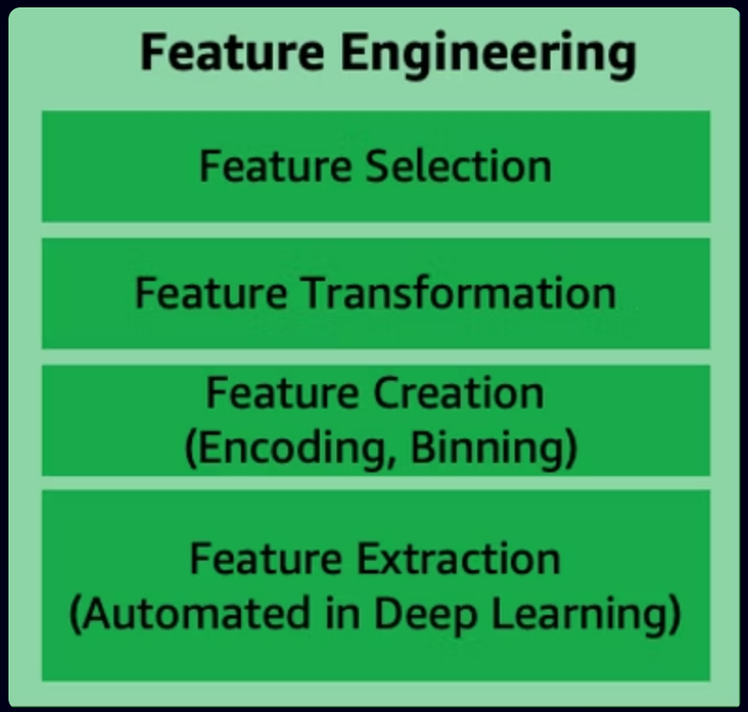
Feature Extraction

Reducing dimensionality while retaining crucial information (e.g., Principal Component Analysis), creating aggregated statistics, or generating lag features for time-series data.

4

Handling Missing Values

Strategically addressing gaps in data through various imputation methods (mean, median, mode, predictive imputation) or appropriate deletion strategies.



Made with GAMMA

Color
Red
Green
Blue

→

Color	Red	Green	Blue
Red	1	0	0
Green	0	1	0
Blue	0	0	1

Manual Methods to Consider

One-hot encoding

Convert categorical features to a binary variable for each value of that feature

Label Encoding

Replace categorical feature values with numerical values (useful if the feature represents a quantity)

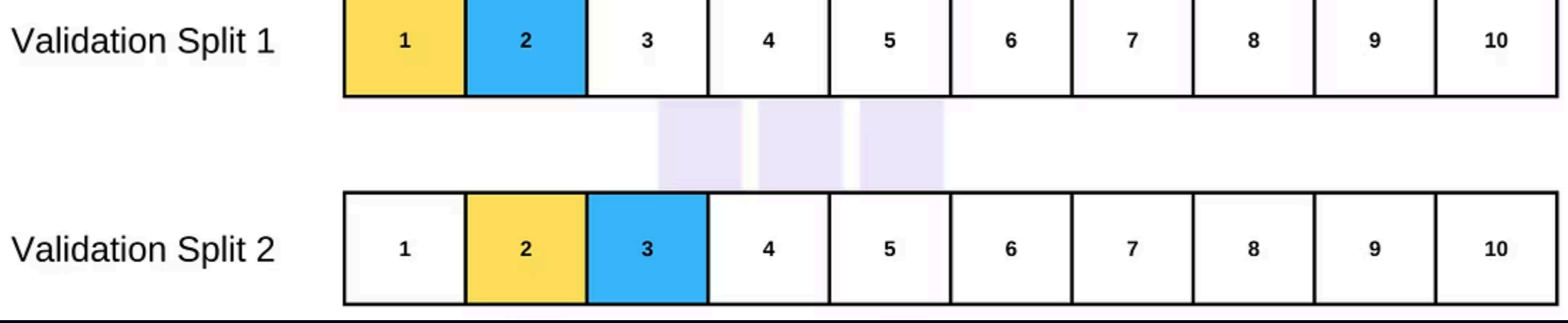
Binning

Group similar numbers into a single range (convert to a categorical variable)

Lag Features

In time-series data, incorporate past values of a variable as new features to model temporal dependencies and patterns over time.

Made with GAMMA



More Manual Methods to Consider

Feature splitting

Split a feature into multiple features

Sliding window

In time-series data, using a specified number of previous data points (the window) to predict the next one

Made with GAMMA

Challenges in Manual Feature Engineering

Overfitting Risk

Creating too many or overly complex features can lead to models that perform well on training data but poorly on unseen data.

Time & Expertise

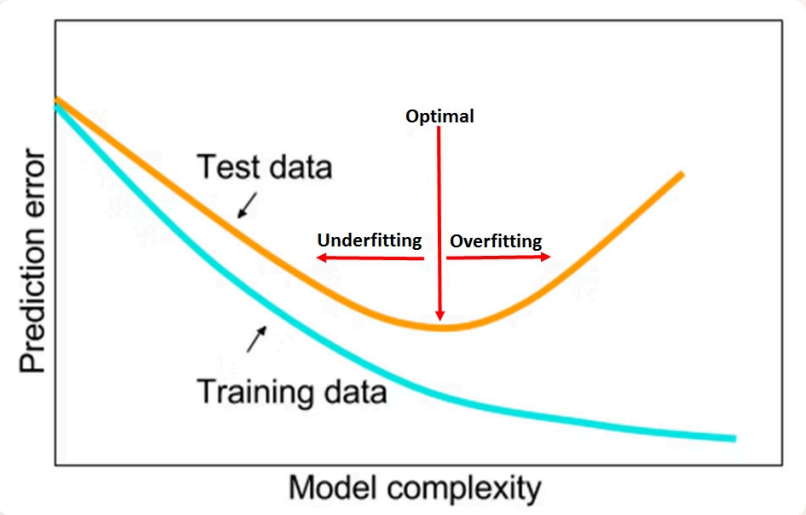
It's often a time-consuming process that demands deep domain expertise and a nuanced understanding of the problem space.

Complexity vs. Performance

Striking the right balance between feature complexity, model interpretability, and predictive performance is a critical challenge.

Iterative Validation

The necessity for continuous, iterative experimentation and rigorous validation to confirm the true impact of engineered features.

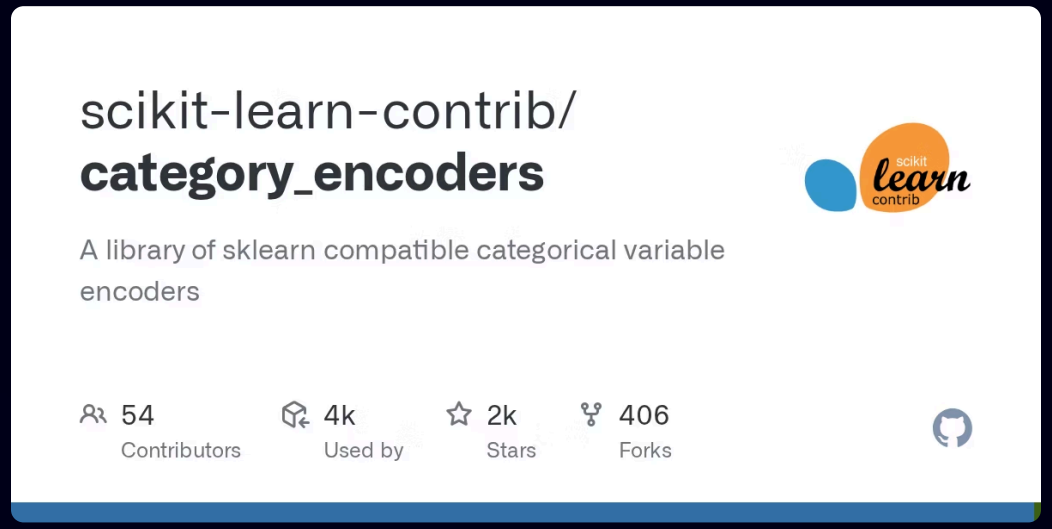


Made with GAMMA

Tools and Best Practices

Essential Tools:

- **Python Libraries:** Leverage **Pandas** for data manipulation, **Scikit-learn** for transformations and scaling, and **category_encoders** for robust categorical feature handling.
- **Documentation:** Maintain clear and thorough documentation of all feature transformations for reproducibility and future auditing.
- **Hybrid Approach:** Combine manual, domain-driven feature engineering with automated feature selection methods to achieve optimal results.
- **Continuous Evaluation:** Systematically evaluate the impact of each feature using various model metrics and feature importance scores to refine your approach.



Made with GAMMA

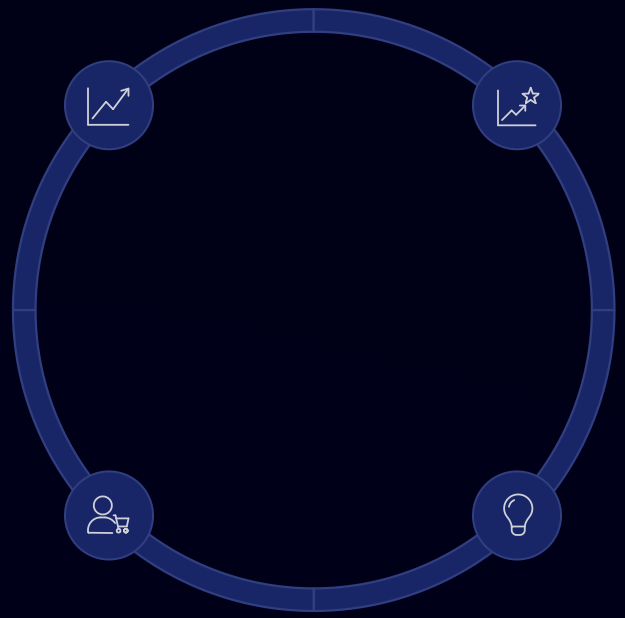
The Impact of Manual Feature Engineering

Superior Accuracy

Can dramatically elevate accuracy and robustness, leading to more reliable predictions and decisions.

Tailored Solutions

Empowers data scientists to craft highly specific solutions that address unique business problems and data characteristics.



Reduced resources

By focusing on the most relevant features, it helps reduce computational resources.

Enhanced Interpretability

Improves model interpretability, fostering greater trust among stakeholders and aiding in informed decision-making.

Made with GAMMA

Google Colab

TQRCG

Made with GAMMA

Thanks for attending!

Next Meeting:

Time Series Analysis & Visualization

2/8/26 @ Cybertorium

Made with GAMMA