FEATURE ENGINEERING FROM DATA TO ML

TITANIC DATASET CASE STUDY | VISTORA AI/ML ASSIGNMENT

WHAT IS FEATURE ENGINEERING ?

Feature Engineering is the process of transforming raw data into informative features that improve the performance and accuracy of Machine Learning models.

Types of Feature Engineering:

1) Feature Transformation
2) Feature Construction
3) Feature Searching
4) Feature Extraction

FEATURE TRANSFORMATION

Modifying the original values of features to enhance interpretability and model accuracy.

- Imputation of null or missing values
- Handling categorical values using Label and One Hot Encoding
- Feature Scaling using Normalization & Standardization
- Outlier Detections

FEATURE CONSTRUCTION

Feature Construction is the manual or automated process of creating new features from existing raw data. It is used introduce domain-specific insights or derived metrics that can boost model performance.

Example :
From Titanic dataset:
FAMILY_SIZE = SIBSP + PARCH + 1
ISALONE = 1 if FAMILY_SIZE == 1 else 0

FEATURE SEARCHING

Feature Searching involves systematically identifying the most relevant features from a potentially large pool — especially in automated pipelines,

<u>Example</u> :
Trying all combinations of AGE, FARE, and TITLE to find the best subset for prediction using a search strategy.

FEATURE EXTRACTION

Feature Extraction is about reducing dimensionality by transforming raw features into a lower-dimensional space while preserving information.

Techniques:

- PCA (Principal Component Analysis)
- LDA (Linear Discriminant Analysis)

TITANIC DATASET OVERVIEW

- The dataset comes from the Titanic shipwreck and is used for predicting survival outcomes.

  Goal: Predict if a passenger survived (Survived = 1) or not (Survived = 0).

- Input Features:

  PCLASS, SEX, AGE, SIBSP, PARCH, FARE, CABIN, EMBARKED
- Target:

  SURVIVED (Binary: 0 or 1)
- Use Case:

  Binary classification problem used in ML training and evaluation.

USING SNOWFLAKE FOR DATA STORAGE & PROCESSING

Snowflake as Cloud Data Warehouse

Snowflake provides scalable, secure, and high-performance storage for structured and semi-structured data. It supports SQL-based querying and integrates easily with data pipelines.

How CSV Was Uploaded

1. Created a stage and file format in Snowflake.
2. Uploaded CSV to the stage using SnowSQL or Web UI.
3. Used COPY INTO command to load data into table.

FEATURE STORE CONCEPTS

A centralized repository to store, manage, and serve engineered features for machine learning models consistently.

Benefits

- Consistency – Same features used for training and inference
- Reusability – Share features across multiple models
- Centralization – One source of truth for all features

Popular Feature Stores

- AWS SageMaker Feature Store
- Databricks Feature Store
- Snowflake Feature Store (used in this project)

FEATURE ENGINEERING WITH SNOWFLAKE & FEATURE STORE

<u>Extract Raw Data from Snowflake</u>

- Query Titanic dataset using SQL:
  SELECT * FROM ML_FEATURE_STORE
  .TITANIC_SCHEMA.TITANIC_FEATURES
- The data includes columns like PassengerId, Pclass, Sex, Age, SibSp, Parch, Fare, Cabin, Embarked, etc.

Transform: Feature Engineering in Python

Data Cleaning:

Fill missing Age with median

Cabin converted to binary Cabin_flag (1 = present, 0 = missing)

Encoding:

Sex encoded (Male → 1, Female → 0)

Embarked → One-hot encoded (Embarked_Q, Embarked_S)

Derived Features:

FamilySize = SibSp + Parch + 1

IsAlone = 1 if FamilySize == 1

Title extracted from Name (e.g., Mr, Miss, etc.) and encoded

Load Engineered Features into Feature Store

- A dedicated Snowflake table TITANIC_FEATURES was created to store only predictive and reusable features.
- Data inserted row-by-row using Python + Snowflake connector.

Access Features for ML Model Training

- Features fetched using SQL:
  SELECT * FROM TITANIC_FEATURES;
- Loaded into pandas DataFrame → used directly for training models like Random Forest, Logistic Regression, etc.

CONCLUSION

This project showcased how Snowflake and Python can be used together for efficient feature engineering and centralized feature storage. Engineered features from the Titanic dataset were stored in a Feature Store, enabling easy reuse and consistent model training. A Random Forest model trained on these features achieved 83% accuracy, validating the effectiveness of this approach.

Thank you