Feature Engineering Report

Laavanjan

August 3, 2025

Dataset Overview

The objective is to predict whether a trip is a cutoff (`is_cutoff`) or not using various features derived from trip time, route, delays, and center pair combinations.

Exploratory Data Analysis (EDA)

Performed basic inspection using **shape**, **.info()**, **.describe()**, and **unique value checks**. This helps understand missing data, data types, and column distributions for preprocessing.

Feature Selection

- **Correlation matrix** and **VIF** were used to identify multicollinearity.
- Features with high correlation (> 0.9) and VIF > 10 were dropped to improve model generalization.
- **Mutual Information** and **Random Forest** were used to identify important features for classification.

Feature Creation

- Extracted **datetime parts** (hour, weekday, etc.) for time-based insights.
- Created **delay-related ratios**, time differences, and route-level aggregations.
- Created **domain-specific indicators** like `is_heavy_delay` and `delay_category` using thresholds.

Feature Transformation

- Applied **MinMaxScaler**, **StandardScaler**, **Log Transform**, **QuantileTransformer**, and **PolynomialFeatures**.
- Each transformation normalizes or enhances features for better learning behavior in ML models.

Feature Scaling

- Applied multiple scalers (Standard, Robust, MaxAbs, MinMax) to numeric features.
- Combined scaled results for experimentation with different ML algorithms.

Feature Reduction

- **PCA** reduced features while retaining 95% of the variance.
- Univariate methods (**ANOVA F-test**, **Chi-square**) and embedded methods (**LassoCV**, **RFE**) selected top features.
- This enhances model performance and reduces overfitting risk.

Saved Outputs

- `feature_engineered_dataset.csv`: Post custom feature creation
- `transformed_features_dataset.csv`: After all transformations
- `scaled_features_dataset.csv`: Various scalings
- `pca_reduced_dataset.csv`: Dimensionality reduced with PCA

Summary

This pipeline includes domain-specific and statistical feature engineering, scaling, and reduction techniques. It enables training high-performance machine learning models by improving data quality, relevance, and interpretability.