ML Project Structure

1. Data Preparation

- **Uniting Tables:** Combine different data sources into a unified dataset, ensuring consistency and alignment of data points.
- **Reduce Large Categories:** Simplify datasets by reducing the number of categories in categorical variables where necessary. This might involve grouping smaller categories into an "Other" category.
- **Clean Text:** Perform text cleaning by removing stop words, punctuation, special characters, and normalizing text (e.g., lowercasing).
- **Transform/Manipulate data**

2. Exploratory Data Analysis (EDA)

- **Instant Reports:** Generate initial reports to get a quick overview of the dataset, including summary statistics and distributions.
- **Descriptive Analysis:** Use statistical methods and visualizations to understand the data, identify patterns, trends, and potential issues.
- **Correlation and other relationship Analysis:** between different features to identify multicollinearity and feature importance.

3. Data Cleansing

- **Outliers:** Detect and handle outliers which might skew the data analysis and model performance. This might involve removing, transforming, or capping outlier values.
- **Missing Values:** Address missing data through imputation methods (e.g., mean, median, mode) or by removing incomplete records, depending on the context and importance of the missing information.

4. Feature Engineering & Feature Selection

- **Enriching:** Create new features from existing ones to better capture the underlying patterns in the data. This can involve mathematical transformations, aggregations, or domain-specific knowledge.

- **Normalization/Standardization:** Scale numerical features to ensure they have similar ranges, which helps certain algorithms perform better.
- **Feature Selection** Choosing the most valuable features that predict the target value by running penalty models

5. One-Hot Encoding

- Convert categorical variables into a format that can be provided to ML algorithms to do a better job in prediction. This typically involves creating binary columns for each category in a categorical feature.
-

6. Model Selection and Fine Tuning

- **Model Selection:** Choose appropriate machine learning models based on the problem at hand (e.g., regression, classification, clustering).
- **Cross-Validation:** Use cross-validation techniques to assess model performance and ensure robustness.
- **Fine Tuning:** Optimize model hyperparameters to improve performance, typically using methods like grid search, random search, or Bayesian optimization.

7. Model Evaluation

- **Performance Metrics:** Evaluate the model using appropriate metrics such as accuracy, precision, recall, F1-score, ROC-AUC for classification, or RMSE, MAE for regression.
- **Validation:** Validate the model on a separate validation set to ensure it generalizes well to unseen data.

8. Model Deployment

* This phase is not required for the Academic ML Project

- **Packaging:** Package the model for deployment, ensuring it can be integrated into production environments.
- **API Development:** Develop APIs for model inference, making it accessible to other applications.

- **Monitoring:** Set up monitoring to track model performance and detect any degradation over time.

9. Documentation and Reporting

- **Documentation:** Maintain thorough documentation of the entire process, including data preparation, EDA, feature engineering, model selection, and evaluation.
- **Reporting:** Create detailed reports and visualizations to communicate findings and model performance to stakeholders.

Dataset

Spotify

https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs?select=spotify_songs.csv

Ecommerce

https://www.kaggle.com/datasets/ahmedaliraja/e-commerece-sales-data-2023-24?select=customer_details.csv

Basketball

https://www.kaggle.com/datasets/jacobbaruch/basketball-players-stats-per-season-49-leagues

https://www.kaggle.com/datasets/sujaykapadnis/basketball-dataset?select=wbb_rosters23_crosswalk.csv

Berlin Airbnb

https://www.kaggle.com/datasets/thedevastator/berlin-airbnb-ratings-how-hosts-measure-up

Uber

https://www.kaggle.com/datasets/shuhengmo/uber-nyc-forhire-vehicles-trip-data-2021

Social media advertisement

https://www.kaggle.com/datasets/alperenmyung/social-media-advertisement-performance?select=ad_events.csv

Trip Advisor EU restaurants

https://www.kaggle.com/datasets/stefanoleone992/tripadvisor-european-restaurants

USA Accidents

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Fraud detection

https://www.kaggle.com/datasets/karthikgangula/credit-card-fraud-mega-dataset

TV Shows

https://www.kaggle.com/datasets/asaniczka/full-tmdb-tv-shows-dataset-2023-150k-shows

New York Taxi

https://www.kaggle.com/datasets/arashnic/taxi-pricing-with-mobility-analytics

Covid-19

https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=covid_19_clean_complete.csv

Linkedin

https://www.kaggle.com/datasets/arshkon/linkedin-job-postings/data

Data Selection Criteria for the Final Project

Criteria for Choosing the Dataset

1. **Dataset Size**
   - **Minimum of 20,000 rows (observations)**.
   - Larger datasets are recommended, but ensure they can be managed within the given time and resources.
2. **Dimensionality**

- **Minimum of 40 columns (features/attributes)**.
- The dataset should include a variety of variable types (numeric, continuous, categorical, dates, text).

Project Objective

- The dataset must be complex enough to enable **research questions** and/or **predictive/classification modeling**.
- It should allow for **deriving insights** beyond simple descriptive statistics.

Documents to Submit (Github, or google Drive):

- Original data (Flat File)
- Automated reports
- Notebooks
- Project Overview document
- Presentation