

Model Development & Experiment Tracking Exercise


Overview:

Design, train, and optimize a machine learning model using Optuna for hyperparameter tuning, while systematically tracking experiments with MLflow. You will be expected to incorporate preprocessing and feature engineering techniques to enhance model performance. Ultimately, you will analyze the impact of hyperparameters on model outcomes, showcasing a complete pipeline from data preparation to model development, experiment tracking, and evaluation.

Objective:

Leverage predictive analytics to identify patterns in customer behavior, enabling proactive strategies to reduce churn. By analyzing comprehensive customer data, develop targeted retention programs to improve customer loyalty and maximize long-term value.

[Telco_churn.csv](#)

 Telco Assignment Template

Content:

Each row represents a customer, each column contains customer attributes described on the column Metadata.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
 - Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
 - Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
 - Demographic info about customers – gender, age range, and if they have partners and dependents
-

Task 1: Data Preparation and Preprocessing

Objective: Prepare and preprocess the data to ensure high-quality inputs for machine learning models.

Goal: Implement robust data cleaning, transformation, and feature engineering techniques to optimize data for model training.

Steps:

1. **Create a Github repository.**
2. **Make a commit to the github repository by:**
 - Selecting File>>Save a Copy In Github
 - Select the name of the repository you created.

- Add a commit message i.e a brief description of the task/step you have completed.
 - Confirm the status of the notebook from the repository.
 - 3. **Access and Load the Dataset:** Locate and load the dataset for the machine learning task.
 - Confirm that the dataset is complete and free of errors.
 - Address missing data through appropriate imputation techniques.
 - 4. **Perform Exploratory Data Analysis (EDA):**
 - Visualize the data to understand distributions, correlations, and trends.
 - Identify and address outliers or anomalies.
 - 5. **Apply Preprocessing and Feature Engineering:**
 - 6. Normalize or scale numerical features as needed.
 - 7. Encode categorical variables.
 - 8. Generate new features that could improve model performance (e.g., polynomial features, domain-specific features).
 - 9. **Deliverable:** A google drive link to your cleaned dataset ready for modeling. Provide a concise report documenting preprocessing steps and justifications (In Notebook). A new commit on github.
-

Task 2: Model Design and Training

Objective: Develop a machine learning model tailored to the dataset and task requirements.

Goal: Train a baseline model to evaluate performance before tuning.

Steps:

1. **Model Selection:**
 - Choose at least two appropriate algorithms for the problem.
 - Justify the choice of the algorithm based on the dataset characteristics.
 2. **Baseline Model Training:**
 - Split the data into training, validation, and test sets.
 - Train an initial model without hyperparameter tuning to establish a performance benchmark.
 3. Evaluate the baseline model using relevant metrics.
 4. **Deliverable:** A trained baseline model with performance metrics and visualizations of results. You should log your baseline and relevant artifacts in mflow. A new commit on github.
-

Task 3: Hyperparameter Tuning with Optuna

Objective: Optimize model hyperparameters using Optuna to achieve better performance.

Goal: Automate the search for optimal hyperparameters and analyze the tuning impact on model performance.

Steps:

1. **Set Up the Optuna Study:**
 - Define the hyperparameters to optimize (e.g., learning rate, number of estimators, regularization parameters).
 - Specify objective functions for performance evaluation.
 2. **Run the Optimization:**
 - Configure Optuna to track trials and log results systematically.
 - Run multiple trials to identify the best hyperparameters.
 3. **Analyze Results:**
 - Visualize the impact of hyperparameters on model outcomes using Optuna's analysis tools.
 - Compare the performance of the optimized model with the baseline.
 4. **Deliverable:** In-notebook report documenting the Optuna optimization process, best hyperparameters, and performance improvements. A new commit on github.
-

Task 4: Experiment Tracking with MLflow

Objective: Continue to track all experiments systematically using MLflow to maintain reproducibility.

Goal: Log cleaned dataset, model parameters, metrics, and other artifacts for each experiment.

Steps:

1. **Set Up MLflow:**
 - Configure the MLflow tracking server and integrate it into the pipeline.
 - Define the parameters, metrics, and artifacts to log (e.g., training loss, accuracy, confusion matrix).
 2. **Log Experiments:**
 - Track all training runs, and hyperparameter trials.
 - Record artifacts such as models and visualizations for comparison.
 3. **Deliverable:** An MLflow experiment tracking dashboard showcasing a history of experiments with detailed logs for reproducibility. A new commit on github.
-

Task 5: Model Evaluation and Analysis

Objective: Evaluate the best performing model's performance on test data.

Goal: Provide insights into the model's behavior and potential areas for improvement.

Steps:

1. **Model Evaluation:**
 - Use test data to compute different relevant performance metrics.
 - Generate relevant visualizations.
2. **Analyze Hyperparameter Impact:**
 - Review how specific hyperparameters influenced model performance.
 - Summarize key findings from the tuning process.

3. **Deliverable:** in-notebook evaluation report detailing the model's performance, strengths, and recommendations for future improvements. A new commit on github.
-

Final Notes

- Ensure that all pipeline steps are modular and well-documented for easy reproducibility.
- Proactively identify challenges encountered during preprocessing, tuning, or tracking and document solutions.
- Highlight how MLflow and Optuna streamlined experiment tracking and facilitated comparisons across different runs, ultimately improving the decision-making process.