NIB 7072: Machine Learning

Course Work: 2024 Batch

Coventry University/NIBM: MSc Data Science

****

Questions 1:

Problem Statement

You are required to identify, acquire, and analyze your own dataset (from an open-source repository such as Kaggle, UCI ML Repository, government portals, or other reliable sources).

The dataset must meet the following requirements:

- At least 10,000 records (rows).
- A mix of categorical and numerical features.
- Must allow you to define a clear supervised prediction task (classification or regression).
- Must have sufficient complexity (e.g., imbalanced classes, missing values, noise, high dimensionality).

The goal is to design, implement, evaluate, and interpret machine learning models that solve a meaningful problem, while situating your work in existing literature.

Tasks

1. Dataset Justification & Literature Review (15%)

- Describe the dataset source, size, and structure.
- Define the prediction problem (classification/regression) and explain its real-world significance.
- Literature Requirement:
    - If the dataset is commonly used, conduct a literature survey of at least 5 peer-reviewed studies or industry reports that used the same dataset. Summarize their approaches and highlight how your work differs or improves upon them.
    - If the dataset is new or less explored, conduct a literature review on similar use cases/domains (minimum 5 references) to situate your study in a broader research context.

2. Data Exploration & Preprocessing (10%)

- Perform exploratory data analysis (EDA) with relevant visualizations.

- Handle missing values, outliers, and imbalanced data.
- Engineer at least three new features relevant to the problem domain.
- Justify preprocessing and feature engineering choices.

3. Model Development (20%)

- Implement at least four machine learning models across different families:
  - One linear model (e.g., Logistic Regression, Linear Regression).
  - One tree-based model (e.g., Decision Tree, Random Forest).
  - One boosting model (e.g., XGBoost, LightGBM, CatBoost).
  - One advanced model (e.g., Neural Network, AutoML pipeline).
- Tune hyperparameters using cross-validation.
- Justify your model choices in relation to dataset characteristics.

4. Evaluation & Comparison (20%)

- Use multiple metrics:
  - Classification: Accuracy, Precision, Recall, F1-score, ROC-AUC.
  - Regression: RMSE, MAE, R^2.
- Conduct error analysis (e.g., which classes are misclassified most, or which records have high residuals).
- Compare models in a results table and plots (confusion matrix, ROC curve, precision-recall curve, error distributions).
- Generate the best precision, recall, accuracy, and F1-score achieved and explain why they are optimal for the use case.
- Apply adjustments if needed (e.g., class weighting, threshold tuning, SMOTE, regularization).
- Track all experiments with MLflow (or an equivalent tool) and provide:
  - Parameters and hyperparameters used
  - Recorded metrics
  - Model artifacts (saved models, plots)

5. Interpretability & Insights (10%)

- Apply model explainability techniques (e.g., SHAP, LIME, permutation importance, partial dependence plots).
- Identify the most influential features and explain their impact.
- Translate findings into real-world insights (e.g., business decisions, risk mitigation, policy recommendations).

6. Critical Reflection (5%)

- Discuss dataset limitations, ethical implications, bias, and generalizability of your model.
- Suggest future extensions (e.g., inclusion of new features, larger datasets, deep learning).

7. Deployment (20%)

Suggest a deployment solution using appropriate technologies (e.g., Docker, Kubernetes, MLflow, Azure ML, AWS Sagemaker, or on-premises solutions). Include considerations for versioning, CI/CD pipelines, and model monitoring.

Questions 2:

Problem Statement

You are required to identify, acquire, and analyze a time series dataset. The dataset can be from any domain (finance, retail, energy, weather, IoT, or healthcare), and must meet the following criteria:

- Students should select an appropriate dataset with sufficient historical records (at least 2–5 years where possible).Must have temporal patterns such as seasonality, trend, or irregularity.
- Should allow you to forecast a target variable and evaluate predictive performance.

The goal is to develop multiple forecasting models, including models discussed in class (Facebook Prophet, Amazon Forecast/Chronos, LSTM) as well as at least one novel or advanced time series model not covered in class.

Tasks

1. Dataset Justification & Literature Review (15%)

- Describe the dataset: source, frequency, size, and features.
- Define the forecasting target and horizon (e.g., next 30 days).
- Conduct a literature survey:
    - If the dataset is public and commonly used, summarize at least 5 studies that have forecasted it.
    - If it is a new dataset, survey literature on similar problem domains to situate your approach.

2. Exploratory Analysis & Preprocessing (15%)

- Plot the time series and its decomposition (trend, seasonality, residuals).
- Identify missing values, anomalies, or outliers and explain how you handle them.
- Create additional temporal features (lag features, rolling statistics, seasonal indicators) as appropriate.

3. Model Development – Class Discussed Models (25%)

- Implement at least the **models discussed in class**:
    1. Facebook Prophet
    2. Amazon Chronos / Forecast
    3. LSTM (or other RNN variant)
- Tune hyperparameters for each model.
- Evaluate forecasting performance using metrics such as MAE, RMSE, MAPE, and R^2.

4. Model Development – Novel or Advanced Models (20%)

- Implement at least one new model not discussed in class, for example:
    - Temporal Fusion Transformer (TFT)
    - N-BEATS or N-HiTS
    - SARIMAX with external regressors
    - DeepAR, Transformer-based, or other recent research models
- Justify your choice of model and its expected advantages over class-discussed models.
- Evaluate and compare its performance against class-discussed models.

5. Comparison, Error Analysis & Insights (15%)

- Compare all models' forecasting performance using multiple metrics.
- Conduct error analysis: identify periods of high prediction error and possible causes.
- Discuss model strengths, limitations, and suitability for the problem domain.
- Include visualizations: forecast vs. actual plots, residuals, error distributions.

6. Critical Reflection (10%)

- Discuss data limitations, challenges in modeling, and ethical considerations (if any).
- Suggest possible improvements or extensions (e.g., multi-step forecasts, ensemble methods, additional exogenous variables).

Questions 3:

Tasks

1. Problem Definition & Literature Review (15%)

- Clearly define the **optimization problem**: objective function, decision variables, and constraints.
- Provide a **mathematical formulation**:
  - Linear / nonlinear objective
  - Equality and inequality constraints
  - Integer/binary decision variables where applicable
- Conduct a **literature review** of at least **5 related studies** using similar problems or solution approaches (genetic algorithms, MIP, heuristics).

2. Data Exploration & Preparation (10%)

- Describe the dataset, decision variables, and constraints.
- Identify challenges: large solution space, combinatorial complexity, or data inconsistencies.
- Preprocess / generate additional features if necessary (e.g., distance matrices for routing, skill matrices for workforce allocation).

3. Model Implementation – Genetic Algorithm (30%)

- Design and implement a Genetic Algorithm (GA) for the problem.
  - Include chromosome representation, fitness function, selection, crossover, and mutation.
  - Tune hyperparameters (population size, mutation rate, crossover rate, number of generations).
- Track GA convergence and visualize fitness over generations.
- Evaluate the GA solution with objective value and constraint satisfaction.

4. Model Implementation – Mixed-Integer Programming (MIP) (25%)

- Formulate the same problem as a MIP problem.
- Solve using a solver such as Gurobi, CPLEX, or PuLP.
- Compare solution quality, computational time, and scalability with GA.

5. Comparison, Analysis & Insights (15%)

- Compare GA and MIP results based on:
  - Objective function value
  - Constraint satisfaction / feasibility
  - Computational time and scalability

- Analyze strengths and weaknesses of each approach in the context of the problem.
- Discuss how problem complexity, size, or constraint tightness affects performance.

6. Critical Reflection (5%)

- Discuss limitations of your models and assumptions made.
- Suggest future improvements (e.g., hybrid GA-MIP approaches, parallel computing, metaheuristics like simulated annealing).
- Comment on real-world applicability and potential business impact.