CMI-Problematic Internet Usage Report

Contents

1 Problem Statement and Objective

The primary aim of this competition is to predict the `sii` (a measure related to problematic internet use) by utilizing both static and time-series data. Participants are provided with two types of datasets: labeled training data to build and train models and unlabeled test data to evaluate predictions. The challenge lies in effective feature engineering, comprehensive data preprocessing, and the implementation of machine learning models to achieve high predictive accuracy. The evaluation metric for the competition is the Quadratic Weighted Kappa (QWK), a robust measure of agreement that accounts for ordinal nature.

2 Data Overview

2.1 Datasets Provided

The data provided for this competition comprises static and time-series datasets. A summary is given below:

- **Static Data:**
  - `train.csv`: This dataset contains labeled data with a variety of features and the target variable `sii`. Features include demographic information, clinical scores, and physical metrics.
  - `test.csv`: This dataset is similar to `train.csv`, but the target variable `sii` is not provided. Predictions are required for these samples.
- **Time-Series Data:**
  - `series_train.parquet`: Contains time-series data linked to each ID in `train.csv`. Measurements were taken at regular intervals.
  - `series_test.parquet`: Contains corresponding time-series data for `test.csv` IDs.
- **Sample Submission File:**
  - `sample_submission.csv`: This file serves as a template for organizing and submitting predictions in the correct format.

2.2 Feature Categories

- **Static Features:** These features encompass demographic data (e.g., age, gender), clinical scores, fitness metrics, and other individual attributes.
- **Time-Series Features:** These include temporal measurements collected over specific intervals. For example, sensor readings, periodic tests, or activity levels recorded over time.

2.3 Target Variable

The `sii` variable is the target for prediction. It is an ordinal variable with values ranging from 0 to 3, representing the severity levels of problematic internet use. The goal is to predict this variable as accurately as possible.

3 Data Preprocessing

3.1 Static Data Preprocessing

- **Feature Selection:** Relevant features were identified using a predefined `featuresCols` list. Columns that were not informative, such as IDs, were excluded to reduce noise and computational overhead.
- **Handling Missing Values:**
  - Categorical columns with missing data were filled with a placeholder value `"Missing"`. These were later encoded into numerical representations.
  - Continuous columns with missing values were imputed using the median, ensuring that outliers did not skew the data.
- **Encoding Categorical Variables:** All categorical variables were mapped to integer values using a dictionary-based approach. This transformation ensured compatibility with machine learning models.

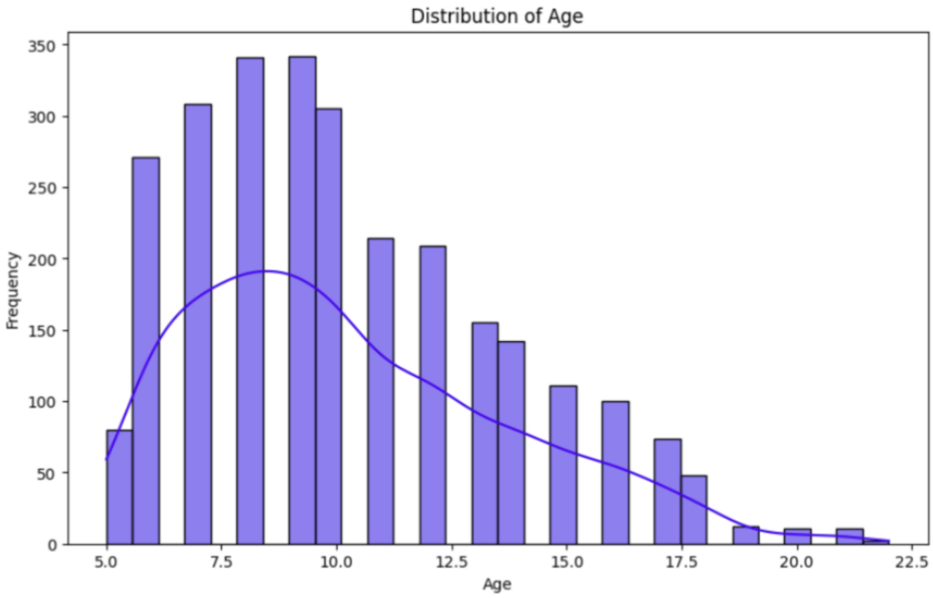
3.2 Time-Series Data Preprocessing

- **Loading Time-Series Data:** Time-series datasets were processed using the `polars` library for efficient data handling. Statistical features such as mean, standard deviation, and trend were extracted for each ID.
- **Merging Time-Series Features:** Extracted features were merged with static data based on unique IDs. This step ensured a unified dataset for training and testing.
- **Final Dataset:** The final dataset combined both static and time-series features. Rows with missing target values in `sii` were dropped.

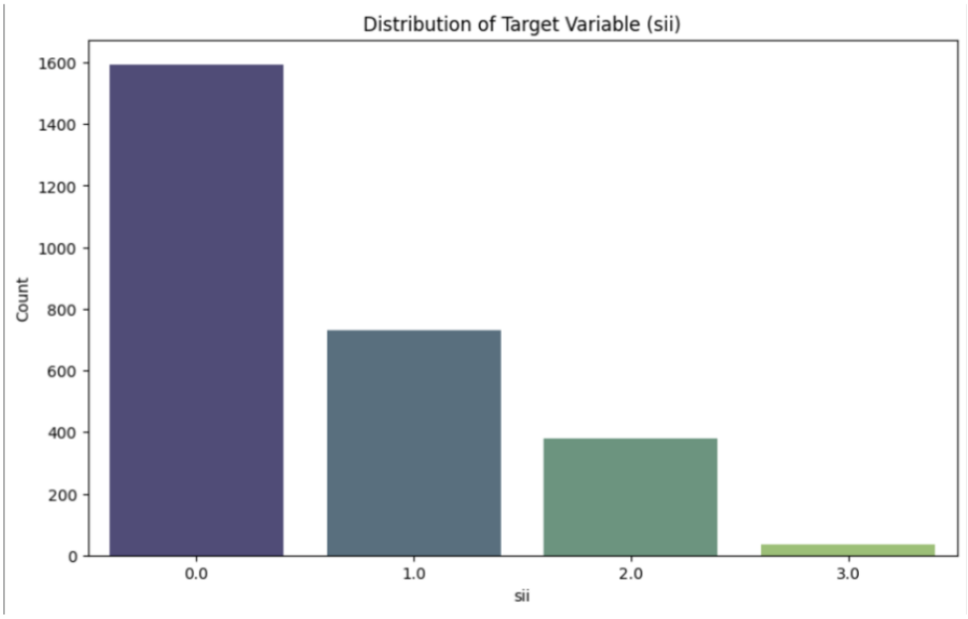
3.3 Visualizations

To understand the data distribution and relationships, the following visualizations were created:

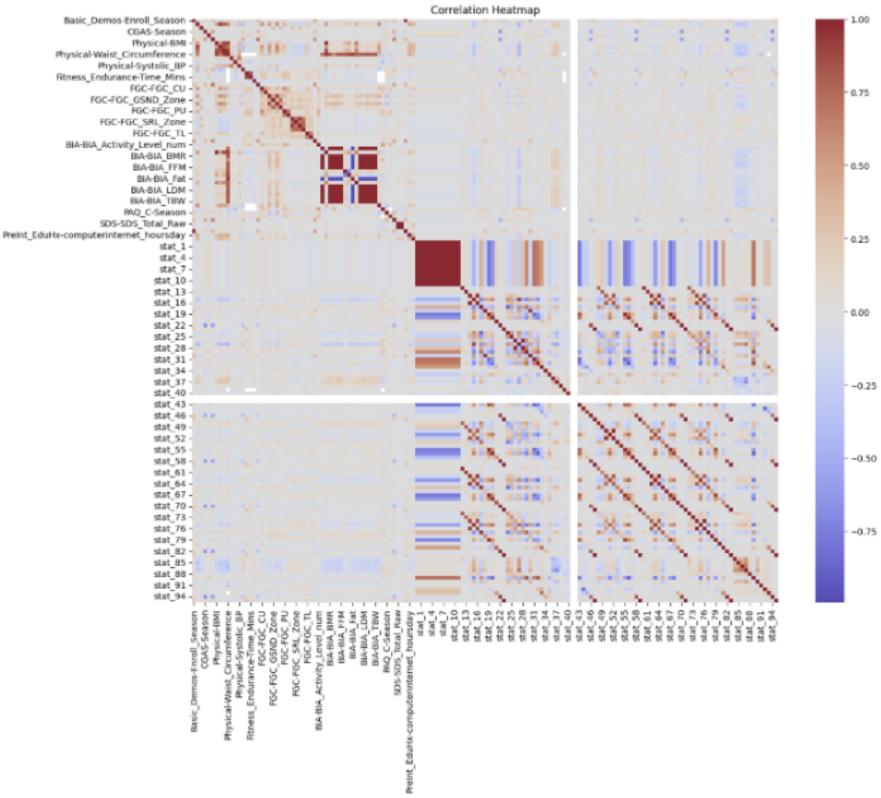
- **Age Distribution:** Shows the spread of participant ages in the dataset.



- **Target Variable Distribution:** Displays the proportion of each `sii` class.



- **Correlation Heatmap:** Highlights relationships between static features.



4 Model Architecture

4.1 Selected Models

An ensemble of advanced regression models was employed to predict `sii`. Each model contributed unique strengths:

- **LightGBM (`LGBMRegressor`):** Optimized for speed and efficiency.
- **XGBoost (`XGBRegressor`):** Provides robust handling of overfitting and missing data.
- **CatBoost (`CatBoostRegressor`):** Specifically designed for handling categorical variables without explicit encoding.
- **Random Forest:** Offers high interpretability and resistance to overfitting.
- **Gradient Boosting:** Captures non-linear patterns effectively.

4.2 Ensemble Strategy

- A **Voting Regressor** was used to combine predictions from all models. Equal weight was assigned to each model to ensure balanced contributions.

5 Training Process

5.1 Cross-Validation

- A `StratifiedKFold` approach with 5 splits was employed to maintain a balanced distribution of `sii` across training and validation datasets.

5.2 Metrics

- The Quadratic Weighted Kappa (QWK) metric was calculated for each fold to monitor performance.

6 Threshold Optimization

- Thresholds were fine-tuned using the Nelder-Mead optimization technique to maximize QWK scores. Adjustments were made to convert continuous predictions into ordinal classes.

7 Observations and Insights

- Combining static and time-series data significantly enhanced the overall model performance by providing us information from both types of data. The static data provided consistent baseline characteristics, while the time-series data captured temporal variations. This integration allowed the model to achieve a better understanding of the underlying patterns, resulting in improved accuracy across various test scenarios.

- Threshold optimization was an important step in fine-tuning the model's output to better reflect the ordinal structure of the `sii` variable. By carefully adjusting the decision boundaries, we ensured that the predicted classes were not only accurate but also was compatible with our target variable. This alignment reduced misclassification errors and improved the interpretability of the model predictions.