Machine Learning I – Group 2 Project Summary

This document provides a comprehensive summary of the steps, decisions, and results carried out during the Machine Learning I group project for the Master in Business Analytics & Data Science.

The objective of the project was to predict solar energy production for the ACME station, using the dataset provided in the "AMS 2013–2014 Solar Energy Prediction Contest" on Kaggle. The following sections describe the rationale behind dataset selection, data cleaning and preprocessing, exploratory analysis, model development, and the final predictions.

1. Dataset Selection

The project used the solar_dataset.csv derived from the AMS 2013–2014 Solar Energy Prediction Contest, containing daily observations from 1994 to 2012. The ACME station was selected as the prediction target following project requirements.

To enhance predictive performance, we integrated the additional_variables.csv file, which includes real Numerical Weather Prediction (NWP) data such as temperature, humidity, wind speed, and pressure. These dynamic weather variables were merged by date to capture day-to-day atmospheric conditions that influence solar production.

Conversely, station-specific features like latitude, longitude, and elevation were excluded, as they remain constant and therefore do not contribute meaningful temporal variance. This streamlined the dataset, focusing on features that truly drive solar energy fluctuations.

2. Data Cleaning and Preprocessing

The preprocessing phase was fundamental to ensure model accuracy and robustness. First, we identified that solar production values for all stations were missing from 2008 onwards. Since the target variable (ACME) was only available until 2007-12-31, we defined this date as the cutoff point to separate training and prediction periods. All records after this date were reserved for future predictions.

We removed irrelevant or redundant columns and focused on the variables that carried the most predictive power:

- ACME production values
- The Principal Component variables (PCs)
- Weather-related variables (from additional variables)
- Additional calendar features (e.g., year, month, day, seasonality) were also included to help capture temporal patterns.

To handle missing values, numerical features were imputed using the mean of their respective columns or grouped averages by hour or period, depending on the context.

2.1 Scaling & Normalization

We standardized all numeric predictors using StandardScaler (zero mean, unit variance). This choice keeps features on comparable scales, important for our linear/regularized baselines, and preserves coefficient interpretability in standard-deviation units. To prevent leakage, the scaler was fit only on the training folds and applied to validation/test via a scikit-learn Pipeline within TimeSeriesSplit. The target (ACME) was not scaled. Principal Components (PCs) were re-standardized along with the remaining features for consistency. For completeness, tree-based models (e.g., HistGradientBoosting) do not require scaling, but we kept a uniform preprocessing pipeline for reproducibility across models.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand data distributions, temporal trends, and correlations between features. We inspected descriptive statistics to detect potential outliers or inconsistencies and plotted several time-series visualizations to examine the evolution of solar production over time.

We also analyzed missing value patterns and correlations among PCA and weather features to ensure data consistency.

This analysis confirmed that solar production was heavily dependent on weather-related principal components, which justified their inclusion as predictors.

The EDA visualizations provide an overview of the dataset’s temporal structure and variable relationships. The time-series plots of solar production reveal strong seasonal patterns, with higher energy values during mid-year months and lower ones during winter, confirming the influence of sunlight intensity and weather conditions. The correlation bar chart shows that several principal components (PCs) and weather variables are moderately to highly correlated with solar output, indicating they capture key meteorological dynamics. Distribution plots highlight that most features are approximately centered but contain mild skewness and outliers, which justified scaling and imputation steps. Overall, the EDA confirms that weather-derived variables are reliable predictors of solar energy generation and that a time-based modeling approach is appropriate given the temporal dependencies observed.

4. Model Development and Optimization

Model development began with Lasso and Ridge Regression to establish a baseline and identify relevant predictors. While Lasso helped highlight key variables through regularization, both models were limited by their linear assumptions and could not fully capture the non-linear relationships between weather conditions and solar energy output.

To improve accuracy and robustness, we advanced to a Gradient Boosting Regressor using the Huber loss function, an ensemble technique that sequentially builds decision trees to reduce prediction errors. The Huber loss combines the sensitivity of Mean Squared Error to outliers with the stability of Mean Absolute Error, making it well-suited for energy data that contain occasional extreme values.

A preprocessing pipeline was implemented using a ColumnTransformer with median imputation for missing numerical values. We used TimeSeriesSplit cross-validation to respect chronological order and RandomizedSearchCV for hyperparameter optimization. The objective was to minimize the Mean Absolute Error (MAE) while ensuring temporal generalization.

The final model configuration was:

- learning_rate = 0.0186
- n_estimators = 1376
- max_depth = 5
- max_features = 0.5
- min_samples_leaf = 28
- min_samples_split = 101
- subsample = 0.79
- alpha = 0.88
- loss = 'huber'.

This setup achieved the best trade-off between bias and variance. Gradient Boosting Regressor (Huber Loss) outperformed all linear baselines, showing higher temporal stability and generalization capacity across validation folds. It effectively captured the non-linear dynamics between weather patterns and solar production, confirming its suitability for this forecasting problem.

The model was evaluated using the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE), the same metrics employed in the Kaggle competition. During validation, the final model achieved an MAE of approximately 2.09, which represented a substantial improvement over the baseline Lasso model.

The results indicate that the model effectively captured the underlying structure of the data, producing reliable forecasts for the ACME station. Despite minor seasonal fluctuations, the Hist Gradient Boosting Regressor demonstrated strong temporal stability and robustness to noise.

6. Final Predictions and Output Generation

Once the model was finalized, it was retrained using all available data prior to 2008 to maximize information usage.

Predictions were then generated for the period between 2008-01-01 and 2012-11-30. These predictions were exported to a CSV file titled `predictions_ACME_2008_onwards_group2.csv`, which constitutes the final output required for submission.

All predictions were post-processed to ensure no negative energy values were present, aligning with physical plausibility.

7. Conclusions and Key Learnings

The final Gradient Boosting Regressor (Huber Loss) achieved the best trade-off between accuracy and robustness, with a test MAE of approximately 2.1 million and an sMAPE near 17%. This confirms the model’s suitability for long-term solar forecasting tasks.

This project reinforced the importance of rigorous preprocessing, time-aware validation, and thoughtful model selection in real-world forecasting. By prioritizing MAE optimization and leveraging gradient boosting, the team achieved both accuracy and interpretability.

The final workflow, from feature integration to cross-validated tuning, provides a replicable framework for similar time-series prediction challenges involving environmental or energy data.