Fuel blend properties prediction challenge

Max. score: 100

Introduction - Welcome to the sixth edition of the Shell.ai Hackathon for Sustainable and Affordable Energy. Shell.ai Hackathon brings together brilliant minds passionate about digital solutions and AI, to tackle real energy challenges and help build a lower-carbon world where everyone can access and afford energy. In the previous five editions, we addressed some of the digital challenges around the energy transition: windfarm layout optimisation (2020), irradiance forecasting for solar power generation (2021), optimal placement of electric vehicle (EV) charging stations (2022), supply chain optimisation for biorefineries (2023), and fleet decarbonisation (2024). This year, we focus on blend properties estimation for sustainable fuel.

Challenge - The global call for sustainability is reshaping every industry, including mobility, shipping and aviation. For example, Sustainable Aviation Fuels (SAFs) are pivotal in this transformation, offering a powerful lever to significantly reduce the sector's environmental footprint. However, integrating these innovative fuels into the existing ecosystem presents a sophisticated challenge. Crafting the optimal fuel blend – mixing various sustainable fuel types sourced from diverse pathways with each other or with conventional fuels – is an intricate science. It demands a delicate balancing act: ensuring adherence to rigorous safety and performance specifications while maximizing environmental benefits and maintaining economic viability. In this hackathon, you will immerse yourselves in the critical field of fuel blending. Your challenge is to develop models capable of predicting the final properties of complex fuel blends based on their constituent components and proportions. By exploring datasets rich with complex interactions, you will decipher the hidden relationships that dictate fuel performance, safety, regulatory and environmental characteristics. The endgame is to engineer powerful predictive tools that can guide the industry in formulating the next generation of sustainable fuels, accelerating the transition to a net-zero future, without compromising on excellence.

Problem Statement - In the fuel industry, blending different fuel components to achieve desired properties is both an art and a science. Crafting the optimal fuel blend—mixing various components from diverse pathways with each other or with conventional fuels—is an intricate science. The relationships between component fractions and final blend properties are highly complex, involving linear and non-linear interactions, synergistic effects, and conditional behaviours that vary based on component combinations. This complexity makes accurate prediction a challenging, high-dimensional problem. Your challenge is to develop models capable of accurately predicting the properties of fuel blends based on their constituent components and their proportions. These predictions must be precise enough to guide real-world blending decisions where safety, performance, and sustainability are paramount. By harnessing the power of data science and machine learning, you will help accelerate the adoption of sustainable aviation fuels by providing tools that can:

- Rapidly evaluate thousands of potential blend combinations.
- Identify optimal recipes that maximize sustainability while meeting specifications.
- Reduce the development time for new sustainable fuel formulations.
- Enable real-time blend optimization in production facilities.

Your work will help engineer powerful predictive tools to guide the industry in formulating the next generation of sustainable fuels, accelerating the transition to a

net-zero future without compromising on excellence.

Data description -

The dataset contains the following:

- **train.csv**: This file contains the data for training your predictive models. Each row represents a unique fuel blend with a total of 65 columns, which are organized into three distinct groups: Blend Composition (first 5 columns): These columns specify the volume percentage of each base component in the blend. Component Properties (next 50 columns): This section simulates a real-world Certificate of Analysis (COA), providing the properties for the specific batch of each component used in a blend. The column names are structured using the format {component_number}*{property_number}. For example, the Property1 for the Component1 component is found in the column Component1* Property1. This structure applies to all 5 components and their 10 respective properties. This suite of properties provides a holistic assessment of the fuel, detailing its core physical and chemical nature, critical safety and operational limits, and its full lifecycle environmental impact. Final Blend Properties - Targets (last 10 columns): These are the 10 target variables your model must predict. They correspond to properties of the blend which will be used as a 'drop-in' replacement fuel. The column names for these target properties are Blend{property_number} (e.g., BlendProperty1, BlendProperty2 etc.).
- **test.csv**: This file is used to evaluate the performance of your model. It contains the input features for 500 blends that are not present in the training data. The structure of this file is identical to train.csv, containing the 55 input feature columns (5 for blend composition (volume fractions) followed by 50 for component properties). However, it does not contain the 10 target property columns. Your model must predict these values as part of the submission file. The test set is split for leaderboard evaluation.
- **sample_submission.csv**: This file demonstrates the required format for your submission. Your submission file must follow this structure precisely to be scored correctly. The file should contain your predictions for the samples in test.csv. Each row in your submission must correspond to the same row in the test.csv file. The order must be preserved for correct evaluation. It must contain exactly 10 columns (excluding ID column), one for each of the 10 target properties your model is expected to predict. Here also, the ordering of the columns must be preserved as ID, BlendProperty1, BlendProperty2, …,BlendProperty10.

Evaluation -

There will be 2 rounds of evaluation which are as follows:

- Public - First half of the samples (250) will be used to calculate your score on the public leaderboard, which is visible throughout the competition.
- Private - The remaining half of the samples (250) will be used for the private leaderboard. Your final score and ranking will be determined by your model's performance on this private subset, which will be revealed only after the competition closes.

You are expected to provide the solution in a .csv format file. The column names that should exist in the .csv along with “valid” entries are provided in the table below.

Columns	Valid Entries

ID	Should be among list of IDs provided in test.csv
BlendProperty1 - BlendProperty10	Should have floating point numbers

Like in previous years, we want to inspire our participants with a computing challenge vital for sustainable and affordable energy. This year's challenge is to estimate the properties of blended fuels. The evaluation metric we will be using is Mean Absolute percentage Error (MAPE) which is guided by the formula:

$$MAPE\left(y, \hat{y}\right) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{\left|y_i - \hat{y}_i\right|}{\left(\varepsilon, \left|y_i\right|\right)}$$

Where:

$$y_i = ground - truth\ predictions$$
$$\hat{y}_i = model\ predictions$$
$$n_{samples} = number\ of\ data\ samples$$

We recommend you use the scikit learn API of MAPE to calculate the efficacy of your solution. The output of the MAPE API will be directly used to calculate the final leaderboard scores. The scores on the leaderboard are calculated using:

$$LeaderboardScore = max\left[10, \left(100 - \frac{90 \times cost}{Reference\ Cost}\right)\right]$$

Scores between 0 and 10 are reserved for error codes. In the above formula, “cost” is the MAPE produced by the solution, and “Reference Cost” is set differently for public and private leaderboard. For public leaderboard, the reference cost is: 2.72 and for private leaderboard, the reference cost is: 2.58.

Note: Error scores - if the candidate gets any integer score between 0-3, it means it encountered an error during evaluation.

0 - Not a .csv file **1 -** Some property column with the specified name structure doesn't exist **2 -** Solution contains non-floating point numbers. **3 -** Solution should have 500 rows and 10 columns (excluding ID column)