Task 2 - Fortune Teller

Objective

- You are the data scientist behind an AI-powered fortune-teller! The fortune-teller predicts how much luck someone will have tomorrow based on their lifestyle habits. Unfortunately, the AI isn't doing a great job. what's wrong and make it better!

Problem Statement

- The AI fortune-teller was trained using 4 parameters - sleep hours, exercise time, coffee consumption and water intake

- **Residual Analysis**:
- The model tends to overestimate luck for people drinking more than 5 cups of coffee.
- It underestimates luck for those exercising over 2 hours per day Model

- **Performance Metrics**:
- Training MAE: 1.5 (scale: 0–10 luck score).
- Validation MAE: 3.0
- Training R^2: 0.80.
- Validation R^2: 0.50

Interpreting Results

Q. Why is the model struggling?

Residual Analysis:

- **Overestimation for high coffee drinkers (more than 5 cups):**
  The model is overestimating the luck score for people who drink more than 5 cups of coffee. This could indicate a non-linear relationship between coffee consumption and luck score.
- **Underestimation for heavy exercisers (more than 2 hours per day):**
  The underestimation of luck for people exercising more than 2 hours per day suggests the model may not capture the positive impact of high exercise on luck. This could indicate that the model's relationship between exercise and luck is too simplistic, again potentially assuming linear effects.

Model Performance Metrics:

- **Training MAE: 1.5, Validation MAE: 3.0**
  The training MAE is much lower than the validation MAE, indicating that the model is **overfitting** to the training data. It performs well on the data it was trained on, but struggles to generalize to new, unseen data. This suggests issues with model robustness and its ability to generalize.
- **Training R^2: 0.80, Validation R^2: 0.50**
  The R^2 score is high in training but drops significantly in validation. This further reinforces the overfitting

problem: the model has learned specific patterns from the training set, but these patterns do not generalize well to new data.

Model Struggles:

1. **Overfitting**
2. **Non-linearity and complex relationships:** The model struggles to handle non-linear relationships, particularly for coffee consumption and exercise time.
3. **Feature limitations:** The features (sleep hours, exercise time, coffee consumption, and water intake) may not fully capture all factors that influence a person's luck, leading to biased predictions.

Q. What are potential biases or issues with feature representation?

1. **Extreme Coffee Drinkers:** Coffee consumption above 5 cups could create a bias, with the model treating these extreme values inappropriately. The overestimation for high coffee drinkers is a sign of this issue.
2. **Heavy Exercisers:** The underestimation for people who exercise more than 2 hours could indicate that extreme values in exercise time aren't being handled well, possibly due to an inappropriate feature scaling or non-linear effects.
3. **Data Imbalance:** There could be an imbalance in the distribution of values for certain features (e.g., most people might consume fewer than 5 cups of coffee), causing the model to be poorly equipped to handle outliers or extreme cases.

Proposing Fixes

1. Adding or Removing Features:

- Add features that might have an impact on "luck", such as:
    - Mental health or stress levels
    - Dietary habits or nutrition intake
    - Sleep quality or consistency
- Remove features or preprocess them to reduce bias:
    - **Binning coffee consumption**: Instead of modeling coffee consumption directly, categorize it into bins (e.g. low, medium, high) to capture the non-linear effect.
    - **Feature transformations**: For features like exercise time or coffee consumption, non-linear transformations can help the model capture complex relationships.

2. Balancing the Data:

- **Handle outliers:** Apply techniques like **capping** or **winsorization** to limit extreme values for coffee consumption or exercise time, or use outlier detection methods to remove or adjust these points.
- **Re-sampling** (over-sampling the underrepresented classes or under-sampling the overrepresented ones) could help the model generalize better by ensuring a balanced representation of different values for each feature.

3. Adjusting the Model:

- **Non-linear models:** Using a more flexible model like **decision trees**, **random forests**, or **gradient boosting machines (GBMs)** would allow the model to capture these relationships better.
- **Hyperparameter tuning:** Fine-tuning the model's hyperparameters (e.g., depth of decision trees, regularization strength) would help prevent overfitting. This could involve techniques like **cross-validation** to ensure the model generalizes well to unseen data.
- **Regularization:** Use **L1** or **L2 regularization** to penalize the complexity of the model, reducing the risk of overfitting.

Design Experiments

Experiments to test fixes

1. **Baseline Model:**
    - Train the model using the current features and algorithms.
    - Measure MAE and R^2 on both the training and validation sets to understand the current performance.
2. **First Experiment: Feature Transformation & Binning:**
    - Apply feature transformations (e.g., logarithmic or polynomial) to coffee consumption and exercise time.
    - Categorize coffee consumption into bins (e.g., 0-1 cups, 2-4 cups, 5+ cups).
    - Measure performance again to see if the residual patterns improve (i.e., reduce overestimation for high coffee drinkers and underestimation for heavy exercisers).
3. **Second Experiment: Non-Linear Models:**
    - Train a decision tree or random forest model, and compare its MAE and R^2 with the baseline model.
    - Measure the improvement in validation metrics (aiming for validation MAE closer to training MAE and an increase in R^2).
4. **Third Experiment: Data Balancing & Outlier Handling:**
    - Apply data balancing techniques such as oversampling or winsorization to handle extreme coffee drinkers and heavy exercisers.
    - Evaluate the impact on performance metrics (validation MAE and R^2).

Measuring Success (e.g., improved MAE, better R^2 on validation)

- **Improvement in MAE:** A reduction in the **validation MAE** is a key indicator that the model is becoming more accurate and is generalizing better to unseen data.
- **Increase in R^2:** A higher **validation R^2** indicates that the model is explaining more of the variance in luck scores.
- **Residual analysis improvement:** Ideally, after applying fixes, the residuals should show less systematic bias for high coffee drinkers and heavy exercisers.

Success Threshold

- Aim for an **MAE reduction** of at least 1 point on the validation set (i.e., reducing validation MAE to ~2.0).
- Increase **validation R^2** closer to 0.65 or higher.
- Ensure that residual plots show fewer systematic errors and less bias related to extreme values of coffee consumption and exercise time.

A Quirky Habit to Consider

I think we can go for **how many times someone accidentally bumps into a door frame in a day**! It's quirky because it's totally random, but you could imagine a theory where, the more often it happens, the better your luck is that day—because, maybe, it's just a sign you're moving through the world at full speed, taking chances, and pushing forward. You could even track it like a personal "**doorframe bump ratio**" to see if there's any pattern.