**# Pokémon Legendary Status Prediction**

## Overview

This project aims to predict whether a Pokémon is legendary or not based on various attributes. It employs a hybrid approach, combining traditional machine learning with deep learning techniques. The project involves data preprocessing, resampling to address class imbalance, and an ensemble of Logistic Regression and Long Short-Term Memory (LSTM) models with an attention mechanism.

## Data Preprocessing

### Resampling

The initial dataset suffers from class imbalance, with significantly fewer legendary Pokémon compared to non-legendary ones. To address this, a custom function `sample\_data()` is implemented. This function utilizes both oversampling (ADASYN - Adaptive Synthetic) and undersampling (NearMiss) techniques to create a more balanced dataset.

Key steps in the resampling process:

1. Separate features and target variable.

2. One-hot encode categorical variables.

3. Apply ADASYN and NearMiss in a pipeline.

4. Combine resampled features with the target variable.

5. Save the resampled data to a new CSV file.

### Feature Engineering

After resampling, the data undergoes further preprocessing:

1. Missing values are imputed using the most frequent strategy.

2. Categorical variables are one-hot encoded.

3. The dataset is split into training and testing sets.

4. Numerical features are standardized, and categorical features are one-hot encoded again (for model-specific preprocessing).

## Model Architecture

The project employs an ensemble of two models:

1. \*\*Logistic Regression\*\*: A traditional machine learning model known for its interpretability and efficiency with binary classification tasks.

2. \*\*LSTM (Long Short-Term Memory)\*\*: A type of recurrent neural network capable of learning long-term dependencies, often used for sequential data but adapted here for its powerful feature learning capabilities.

The predictions from both models are combined using an attention mechanism, which assigns dynamic weights to each model's predictions based on their perceived importance for each instance.

## Training Process

1. The Logistic Regression model is trained on the preprocessed data using a pipeline that includes imputation and scaling/encoding steps.

2. The LSTM model is trained with the following configuration:

- 64 LSTM units

- Dropout layer (20% dropout rate) for regularization

- Dense output layer with sigmoid activation

- Binary cross-entropy loss

- Adam optimizer

- Custom learning rate scheduler (reducing the learning rate after 10 epochs)

3. Both models make predictions on the test set.

4. An attention mechanism is applied to the predictions:

- The predictions are concatenated.

- A dense layer with tanh activation computes attention scores.

- Another dense layer with softmax activation computes attention weights.

- The final prediction is a weighted sum of both models' predictions.

## Evaluation

The ensemble model's performance is evaluated using several metrics:

Certainly! I'll incorporate the output you provided into the documentation. Here's the updated version with the results included:

# Pokémon Legendary Status Prediction

[Previous sections remain the same...]

## Evaluation

The ensemble model's performance is evaluated using several metrics. The results are as follows:

### Confusion Matrix

```

[[ 34 2]

[ 3 246]]

```

This shows that:

- True Negatives (correctly predicted non-legendary): 34

- False Positives (incorrectly predicted as legendary): 2

- False Negatives (incorrectly predicted as non-legendary): 3

- True Positives (correctly predicted legendary): 246

### Performance Metrics

1. \*\*Sensitivity (True Positive Rate)\*\*: 0.9880

- This indicates that 98.80% of the actual legendary Pokémon were correctly identified.

2. \*\*Specificity (True Negative Rate)\*\*: 0.9444

- This shows that 94.44% of the non-legendary Pokémon were correctly identified.

3. \*\*Accuracy\*\*: 0.9824561403508771

- Overall, the model correctly predicted the legendary status for 98.25% of the Pokémon.

4. \*\*Area Under the Receiver Operating Characteristic Curve (AUC-ROC)\*\*: 0.9991075412762159

- This extremely high value (close to 1) indicates that the model has an outstanding ability to discriminate between legendary and non-legendary Pokémon.

5. \*\*Precision\*\*: 0.9919354838709677

- Of all the Pokémon predicted as legendary, 99.19% were actually legendary.

6. \*\*Recall\*\*: 0.9879518072289156

- This is the same as sensitivity, confirming that 98.80% of all legendary Pokémon were correctly identified.

7. \*\*F1 Score\*\*: 0.9899396378269618

- The harmonic mean of precision and recall, with a high value of 0.99, indicates a well-balanced model.

Classification Report:  
precision recall f1-score support

0 0.92 0.94 0.93 36

1 0.99 0.99 0.99 249

accuracy 0.98 285

macro avg 0.96 0.97 0.96 285

weighted avg 0.98 0.98 0.98 285 ```

This report breaks down the performance by class:

- Class 0 (Non-Legendary):

- Precision: 0.92

- Recall: 0.94

- F1-score: 0.93

- Class 1 (Legendary):

- Precision: 0.99

- Recall: 0.99

- F1-score: 0.99

The high scores across all metrics for both classes demonstrate that the model performs exceptionally well, even with the initial class imbalance.

## Interpretation of Results

1. The model shows high performance across all metrics, with accuracy, precision, recall, and F1-score all above 0.98. This indicates that it's highly reliable in predicting both legendary and non-legendary Pokémon.

2. The AUC-ROC score of 0.999 is particularly noteworthy. It suggests that the model has learned to rank legendary Pokémon higher than non-legendary ones with near-perfect discrimination.

3. Despite the initial class imbalance, the resampling techniques and the ensemble approach have resulted in a model that performs well on both classes. This is evident from the high specificity (ability to identify non-legendary Pokémon) and sensitivity (ability to identify legendary Pokémon).

4. The slightly lower performance on the minority class (non-legendary, in this case, due to resampling) is common in imbalanced datasets. However, with precision and recall both at 0.92 and 0.94 respectively for this class, the model still performs robustly.

5. The confusion matrix shows that misclassifications are minimal: only 2 non-legendary Pokémon were incorrectly classified as legendary, and only 3 legendary Pokémon were misclassified as non-legendary.

## Conclusion

The ensemble model, combining Logistic Regression and LSTM with an attention mechanism, has demonstrated exceptional performance in predicting the legendary status of Pokémon. The comprehensive preprocessing, including addressing class imbalance and feature engineering, contributed significantly to this success. The high scores across various evaluation metrics underscore the model's reliability and effectiveness, making it a valuable tool for Pokémon researchers and enthusiasts alike.

.