WASHINGTON COMMANDERS QUANT ASSESSMENT

Part 2. Coding Exercise

The goal was to predict the number of yards gained ('GAIN') for plays marked as 'MISSING' in the dataset. The following steps were taken:

1. **Data Cleaning and Preparation:**
    - The 'GAIN' column was processed to replace 'MISSING' values with np.nan.
    - Missing values in the 'DOWN' and 'DIST' columns were input using their median values.
2. **Feature Engineering:**
    - Numerical and categorical features were identified.
    - Outliers in numerical features were handled using the Interquartile Range (IQR) method.
3. **Data Preprocessing:**
    - A preprocessing pipeline was created for numerical and categorical features. Numerical features were imputed and scaled, while categorical features were imputed and one-hot encoded.
4. **Model Selection and Training:**
    - Four regression models were evaluated: RandomForestRegressor, XGBRegressor, CatBoostRegressor, and GradientBoostingRegressor.
    - Stratified K-Fold cross-validation ensured the model generalizes well across different data subsets.
5. **Hyperparameter Tuning:**
    - RandomizedSearchCV performed hyperparameter tuning for each model, searching specified parameter grids to identify optimal settings.
6. **Model Evaluation and Ensemble Method:**
    - The best-performing models from hyperparameter tuning were combined using a stacking ensemble method (VotingRegressor) to improve prediction accuracy.
7. **Prediction and Uncertainty Estimation:**
    - The ensemble model predicted the 'GAIN' values for the missing rows.
    - Bootstrapping calculated a 90% prediction interval for the sum of the predicted 'GAIN' values by repeatedly resampling the predicted values.

Outcome

The final model produced accurate predictions with a validation RMSE of 5.607 and estimated the sum of predicted 'GAIN' values for the missing rows to be 695.31, with a 90% prediction interval between 591.19 and 801.13. This approach ensured reliable predictions and quantified uncertainty, providing a robust solution.

This methodology integrates data cleaning, preprocessing, model training with hyperparameter tuning, and ensemble techniques to achieve high prediction accuracy and reliable uncertainty estimation.