Comprehensive Strategy for Women's Shoe Price Prediction Model

1. **Data Loading and Splitting**:
    - The training data is loaded, and unnecessary columns are removed to simplify the dataset.
    - The data is split into three subsets:
        - **Validation Set**: 10% of data reserved for final model testing.
        - **Tuning Set**: 30% of the remaining data used for model tuning.
        - **Feature Engineering Set**: 60% of the remaining data for feature creation.
2. **Feature Engineering**:
    - **Cleaning Existing Features**: Missing values are imputed, and categorical variables are standardized. Specific fields like heel_height and origin are transformed to more usable formats.
    - **Text Feature Extraction**: Key words from the product title are identified and tested for statistical significance in predicting price. Significant words are transformed into binary features to capture their presence in each listing.
    - **Text Features with Embeddings**: The textfeatures package is used to generate numerical embeddings (e.g., sentiment scores) from the title, adding deeper insights into textual content.
3. **Model Tuning**:
    - **Cross-Validation**: An XGBoost model is set up with a parameter grid, tuning tree count, learning rate, and sample size.
    - **Grid Search**: The code performs grid search using 5-fold cross-validation to identify the best-performing hyperparameters for the model.
4. **Baseline Model Comparison**:
    - A baseline linear regression model is created to set a performance benchmark.
    - The RMSE (Root Mean Square Error) of the tuned XGBoost model is compared against this baseline to determine improvement.
5. **Final Model Training and Prediction**:
    - The full training set is used to retrain the final XGBoost model with the best parameters identified, and then tested on the test set.

Summary

The strategy involves carefully preparing data, engineering both numeric and text features, tuning an advanced model, and comparing it to a baseline to ensure meaningful improvements in prediction accuracy. This approach maximizes predictive power by leveraging both structured data and text-based insights.