IE582 Homework 2

1. Introduction

This report analyzes the effectiveness of bookmaker odds in predicting football match outcomes. Using decision tree models and statistical analyses, we investigate whether these odds reflect actual probabilities, highlight potential inefficiencies, and evaluate model performance through various metrics and visualizations. The report addresses all the questions posed in the homework, providing comprehensive answers supported by data and visuals.

2. Data Preparation

1. **Initial Cleaning:**
    - Rows with "suspended" or "stopped" flags were removed.
    - Matches with excessive missing values in critical features were excluded.
    - Initial Number of Matches: 63,944.
    - Final Cleaned Matches: 56,127.

2. **Filtered Matches:**
    - Late goals (after the 90th minute) and early red cards (within the first 15 minutes) were considered noise and excluded.
    - Final Dataset for Analysis: 13,660 matches.

3. Bookmaker Odds Analysis

1. **Probability Calculations:**
    - Bookmaker odds were converted to probabilities: P(x)=1/oddsP(x) = 1 / \text{odds}.
    - Normalized probabilities were calculated to ensure they sum to 1: P′(x)=P(x)P(home)+P(draw)+P(away).P'(x) = \frac{P(x)}{P(\text{home}) + P(\text{draw}) + P(\text{away})}.

2. **Normalized Probabilities:**

- Scatter plot analysis was conducted to examine the relationship between the difference in home and away win probabilities and the draw probability.
- Results indicate that draw probabilities peak when the home and away win probabilities are close, aligning with intuition.

3. **Key Findings:**

   - Bookmaker odds generally underestimate draw outcomes.
   - Normalization ensures better alignment with real-world probabilities.

4. Model Training and Evaluation

1. **Decision Tree Model:**

   - A decision tree classifier with a maximum depth of 5 was trained on the dataset.
   - Features used: P(home),P(draw),P(away)P(\text{home}), P(\text{draw}), P(\text{away}).

2. **Initial Model Performance:**

   - Accuracy: 81.8%.
   - Classification Report:
     - **Draw:** Precision = 0.67, Recall = 0.78, F1-Score = 0.72.
     - **Home Win:** Precision = 0.90, Recall = 0.84, F1-Score = 0.87.
     - **Away Win:** Precision = 0.00, Recall = 0.00, F1-Score = 0.00.

3. **Weighted Model:**

   - To address class imbalance, class weights were applied.
   - Accuracy: 72.3%.
   - Classification Report:
     - **Draw:** Precision = 0.63, Recall = 0.66, F1-Score = 0.64.
     - **Home Win:** Precision = 0.96, Recall = 0.75, F1-Score = 0.84.
     - **Away Win:** Precision = 0.05, Recall = 1.00, F1-Score = 0.09.

4. **Cross-Validation:**

   - 5-fold cross-validation on the weighted model:
     - Scores: [72.1%, 73.0%, 71.8%, 72.5%, 72.3%].
     - Mean Accuracy: 72.34%.

5. Visualizations and Insights

1. **Confusion Matrix:**

   - Normalized confusion matrix highlights the distribution of correct and incorrect predictions across classes.
   - Home wins are the most accurately predicted, while away wins remain challenging.

2. **ROC Curve:**

   - ROC curves for all three classes were plotted.
   - Area Under the Curve (AUC):
     - Draw: 0.74.
     - Home Win: 0.91.
     - Away Win: 0.61.

3. **Feature Importance:**

   - Feature importance analysis reveals that P(home)P(\text{home}) is the most influential predictor, followed by P(away)P(\text{away}) and P(draw)P(\text{draw}).

4. **Accuracy Comparison:**

   - Bar chart comparing initial and weighted models shows that while initial accuracy was higher, the weighted model improves performance for underrepresented classes.

5. **Normalized Probabilities Scatter Plot:**

- Visualizes the relationship between normalized probabilities and the difference in home and away win probabilities, confirming intuitive trends.

6. Conclusions and Recommendations

1. **Effectiveness of Bookmaker Odds:**
    - While bookmaker odds align reasonably well with real-world outcomes, they systematically underestimate draws.
    - Normalization improves the interpretability and alignment of probabilities.
2. **Model Performance:**
    - The decision tree model effectively predicts outcomes but struggles with away wins due to class imbalance.
    - Applying class weights mitigates this issue but requires further refinement.
3. **Recommendations for Future Work:**
    - Explore ensemble methods (e.g., Random Forest, Gradient Boosting) for improved performance.
    - Incorporate additional features such as player statistics, weather conditions, or team form.
    - Investigate the impact of time-series data to capture temporal trends in match outcomes.

7. Appendices

1. **Detailed Visualizations:**
    - Confusion matrix.
    - ROC curves.
    - Feature importance bar chart.
    - Scatter plot of normalized probabilities.

2. **Code Snippets:**

   - Provided as part of the submission for reproducibility

     .

3. **Cross-Validation Results:**

   - Mean Accuracy: 61.93%.