DSAI 5101 Report

Team 7

November 7, 2025

1 Introduction

1.1 Research Background

Taking League of Legends as the research focus, this study delves into the correlation between 2021 World Championship and players' performance in Solo Rank during the month preceding the 2021 World Championship. Professional matches not only depend on team collaboration; data such as individual players' champion usage and win rates in Solo Rank matches may also reflect their competitive form. However, there is currently a lack of clear verification regarding the effectiveness of such data in predicting World championship outcomes.

1.2 Research Question

In League of Legends, does a pro player's rank influence their match performance, and can rank data be utilized to predict match win rates?

1.3 Significance of the Research

For team analyst, players' rank data is of great significance. Analysts can use players' rank data to objectively assess players' form, champion proficiency, and the strength of champion. This allows for timely adjustments to BP (Ban/Pick) strategies, which helps the team achieve victories.

For viewers, prediction models can evaluate the strength of the champion selected by both sides. This enables viewers to better understand the strength of the OP champions of both teams during the BP, and gives people who are not familiar with the OP champions of the current meta a more accessible way to understand .

2 EDA(Explore Data Analysis)

2.1 Data Background and Description

Pick rate (professional pick rate and mixed pick rate), Win rate (Professional win rate and mixed win rate), difference value (professional data - mixed data) and correlation(correlation coefficient between professional solo rank pick rates and World Championship pick rates)

2.2 Analysis of Champion Win Rate Differences by Role

- **Support:** Lulu (+25.10) and Karma (+25.76). Their protective kits lean heavily on team coordination—this is exactly why their win rates are way higher in pro play than in pub matches.
- **Jungle:** Xin Zhao (+25.08). Professional teams utilize his gank rhythm and team-fight control far better than non-professional players.
- **Mid:** Orianna (+14.88) and Ryze (+13.83). Professional advantages in team-fight positioning and macro management are core to their performance.
- **Top:** Kennen (+11.07) and Aatrox (+13.06). Their mechanical complexity is better exploited through professional teamwork.

2.3 Correlation Analysis Results between World Championship Pick Rate and Solo rank Pick Rate

- **Mid: 0.7454**, **Top:0.6906** , **Jungle:0.6876**, **ADC:0.5770**, **Support:-0.0556**
- **Correlations:**

  R-value close to +1: High Pick rate champion in rank are positively correlation with champion in the World Championship

  R-value close to 0: no linear relationship

  R-value close to -1: The two patterns show a negative correlation

3 Methods

3.1 Data Overview

- **Data Source:** The dataset combines professional match data from the *2021 League of Legends World Championship* (**https://oracleselixir.com**) with solo queue performance statistics of the same professional players.

- **Original Datasets:**
  - *2021_worlds_kda_with_result.csv:* All the champion picks in each game of the 2021 League of Legends World Championship with results and statics like KDA
  - *2024_worlds_kda_with_result.csv:* All the champion picks in each game of the 2024 League of Legends World Championship with results and statics like KDA
  - *solo_stats.csv:* Each player's pick rate, win rate, and total number of games for champions they used in ranked matches.
  - *side_winrate.csv:* Each side's win rate (blue & red)
- **Sample Size:** The dataset contains **approximately 200 team match instances** (each representing a single team's performance in one game).
- **Data Preprocessing:**
  - *Data cleaning:* Translate champion names to English and standardized player identifiers. Remove % symbols ans extra spaces.
  - *Feature aggregation:* Aggregated player-level data to team-level means and calculated inter-team differences.
- **Key Variables:**
  - **Dependent variable:** `A_win`, a binary variable indicating whether Team A won (1) or lost (0).
  - **Independent variables:**
    - *pickrate_diff:* Difference in average champion pick rate between Team A and Team B under the rank environment.
    - *winrate_diff:* Difference in average champion win rate between Team A and Team B under the rank environment.
    - *picks_diff:* Difference in average champion picks (count) between Team A and Team B under the rank environment.
    - *solo_pickrate_diff:* The pick rate difference based on player×champion usage.
    - *solo_winrate_diff:* The win rate difference based on player×champion usage.
- **Feature Sets:**

- **features_3:** [solo_pickrate_diff, solo_winrate_diff, picks_diff]
- **features_5:** [pickrate_diff, winrate_diff, picks_diff, solo_pickrate_diff, solo_winrate_diff].

Constructing two feature sets allows the model to capture both macro-level meta trends and micro-level individual proficiency. It provides more interpretable insights into the balance between 'champion strength' and 'player skill', enabling a more robust and realistic representation of the factors influencing professional match outcomes.

3.2 Model Training

- **Method Selection:** To assess whether solo queue performance influences professional match outcomes, two ensemble learning models were employed: Random Forest (RF) and XGBoost (XGB). These models capture nonlinear relationships and feature interactions, suitable for tabular datasets.
- **Model Description:**
  - *Random Forest (RF):* An ensemble of 800 trees with balanced class weights and bootstrapped samples.
  - *XGBoost (XGB):* A gradient boosting tree model with 1200 estimators, learning rate 0.05, and max depth 5. Class imbalance was corrected using `scale_pos_weight`.
- **Evaluation Procedure:** Due to the limited availability of real-world professional match data, the size of the dataset is relatively small. Consequently, a **10-fold stratified cross-validation** is approached to ensure the robustness and generalization of model performance. Metrics: Accuracy and F1-score.

4 Results

4.1 Cross Validation

Feature_3 included only professional match data, while feature_5 combined professional match data with solo performance data from ranked games.

In Figure 1 and Figure 2, a noticeable improvement can be observed for the Random Forest model when the solo features were added: the mean accuracy increased from 0.592 to 0.649, and the F1-score

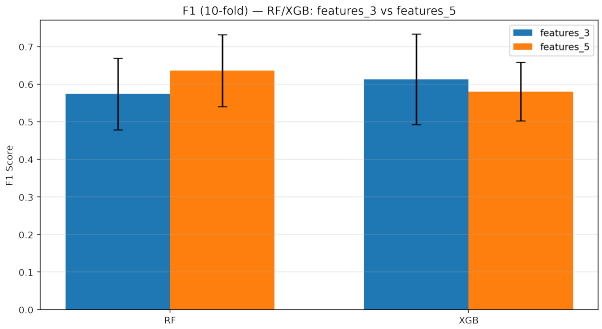

Figure 1: Comparison of F1 Score

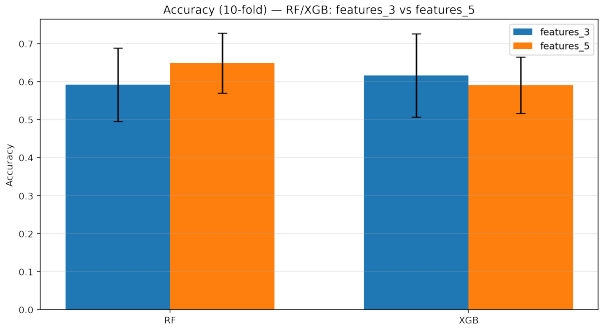

Figure 2: Comparison of Accuracy

from 0.574 to 0.636. However, the improvement was not statistically significant, which means the enhancement is not conclusive.

And the XGBoost model showed slightly lower performance after including solo features — the mean accuracy dropped from 0.616 to 0.591, and the F1-score from 0.613 to 0.580. This indicates that XGBoost is more sensitive to these variables but does not gain a predictive advantage from them.

4.2 Confusion Matrix

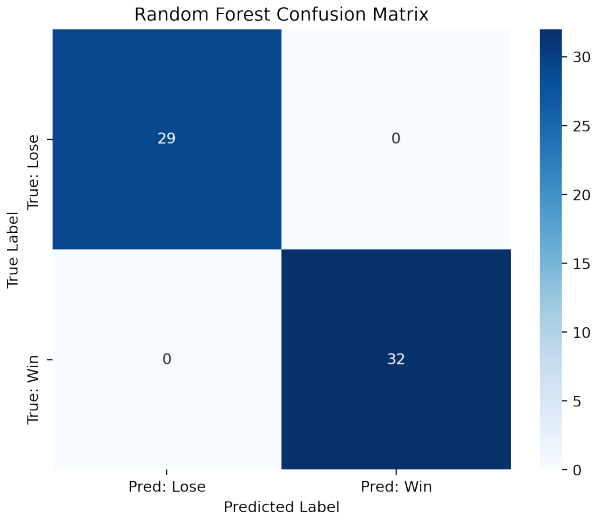

Figure 3: Confusion Matrix of the Random Forest Model

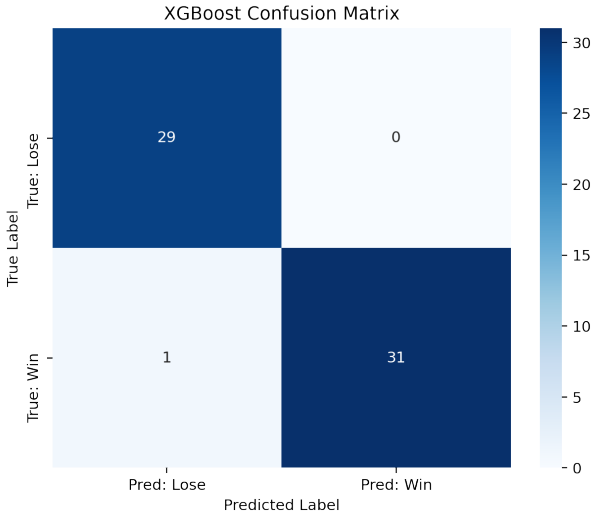

Figure 4: Confusion Matrix of the XGBoost Model

For the Random Forest model, Figure 3 shows that the model correctly predicted 32 wins, where the actual outcome was also a win, and 29 losses, where the actual outcome was a loss, with zero false positives and zero false negatives. For the XGBoost model, Figure 4 shows that the model correctly predicted 31 wins and 29 losses, with only one false positive and one false negative. Both models demonstrate strong classification ability, but such results may indicate overfitting.

The results are based on data from the 2021 World Championship, comprising a total of 81 matches. 75% of these match results are used for model training. Given that hero attributes can change with different patches, using data from the same patch version for model training ensures that the results are more consistent and convincing.

4.3 Permutation Importance

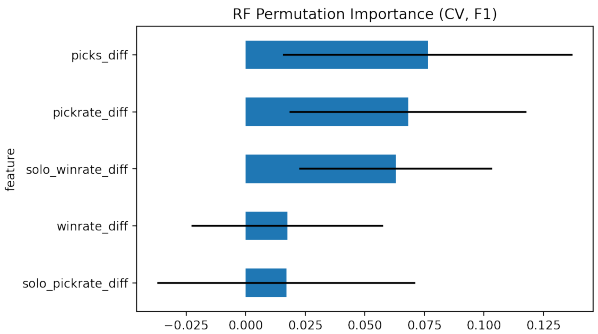

Figure 5: Permutation Importance of the Random Forest Model

Permutation importance was employed to evaluate the true contribution of each feature to model performance. This method assesses how the performance changes when the feature is randomly shuffled, thereby providing a robust measure of each feature's predictive power. For the Random Forest

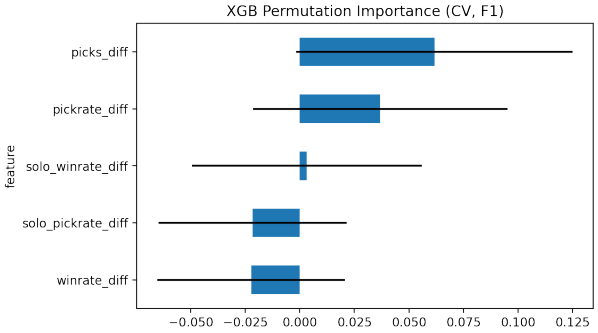

Figure 6: Permutation Importance of the XGBoost Model

model, the solo features together contributed about 33.1% of the total importance. Notably, the feature (solo_winrate_diff) had an influence comparable to major professional features (pickrate_diff and picks_diff). This suggests that hero performance in solo queue provides complementary signals for predicting professional match outcomes. However, for the XGBoost model, the permutation importance of the solo features was close to zero or slightly negative (overall share $\approx -31.9\%$), implying that these features may introduce noise rather than useful information.

4.4 Model Performance on 2021 and 2024 Datasets

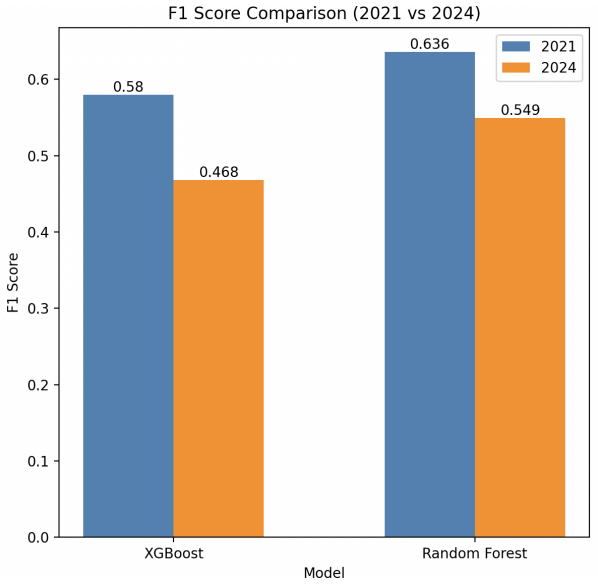

Figure 7: Comparison of F1 Scores Using different Datasets

We trained the models using the 2021 World Championship dataset and then validated the model performance using both the 2021 and the 2024 competition datasets. The results demonstrated that the models trained on 2021 dataset perform less effectively when applied to 2024 dataset. The differences in game rules, such as changes in hero attributes, strategies, and overall meta, between the 2021 and 2024 seasons significantly impact the model's predictions. These changes render the 2021-trained models less

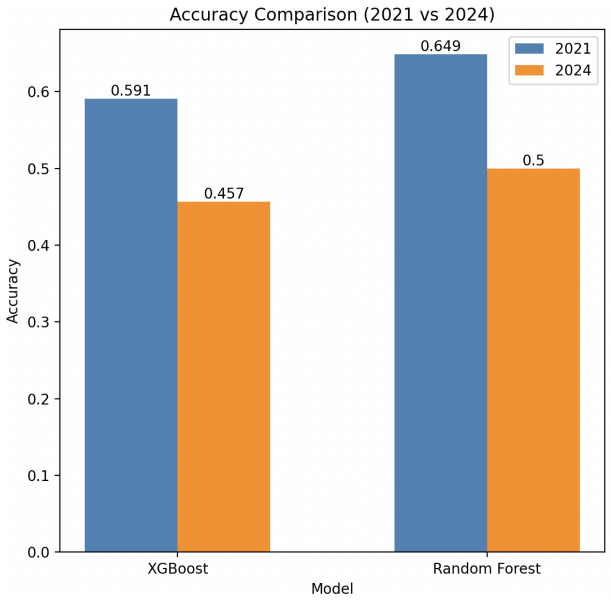

Figure 8: Comparison of Accuracy Using different Datasets

applicable, suggesting that the features derived from solo performance may no longer be as relevant or predictive in the current meta.

5 Conclusion

Our findings indicate that professional match outcomes are mainly determined by team-related factors — like drafting strategy, coordination— while individual hero performance in solo environments provides only marginal supplementary information. This observation aligns with intuitive understanding. Heroes that are strong in solo queue often reflect meta strength and public proficiency, whereas professional matches focus more on teamwork and strategy, which lessens the importance of solo performance data. Considering the limitations of the small dataset and the constantly changing League of Legends meta over different seasons, future research should consider using more varied datasets that cover multiple seasons and hero updates. This approach could enhance the accuracy and effectiveness of predictive models.

6 Team Member Contributions

- Huang Yanyao: The contributions involved supporting the team in code implementation, preparing selected parts of the presentation and slides, and contributing to the writing of the report.
- Shi Yusen: Responsible for data processing and model training.
- Wang Siqiang: Prepare selected parts of the presentation and slides,data analyse part and contribute to the writing of the report.

- Long Xingyu: Built a clean, match-level dataset of 300+ LoL Worlds games and hundreds of player records, giving the team reliable fuel for machine-learning analysis.
- Ji Yumo: Responsible for the results analysis part, including result analysis and visualization, preparing the presentation and slides, and writing the related documents.

References

[1] Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

[2] '2021 Season World Championship — Match History'. Leaguepedia - League of Legends Esports Wiki. From https://lol.fandom.com/wiki/2021_Season_World_Championship/Match_History

[3] "Tournament – Games Of Legends (gol.gg)". From https://gol.gg/tournament

[4] "Match Data Downloads – Oracle's Elixir". From https://oracleselixir.com/tools/downloads

[5] Vickz84259. (n.d.). Unofficial LoL Esports API Documentation. From https://vickz84259.github.io/lolesports-api-docs/