Machine Learning Approaches to Forecasting Surf Competitions.

---

1. Introduction

This project investigates the use of machine learning to predict the winners of surf competition heats. Three machine learning models, Artificial Neural Networks (ANNs), Random Forests, and XGBoost, are employed to analyse historical competition data, surfer statistics, and wave conditions. The project aims to determine the feasibility and effectiveness of machine learning in forecasting surf competition outcomes.

2. Methodology

2.1 Dataset: The project uses a dataset consisting of more than 19,000 Championship Tour heats conducted by the World Surfing League from 2012 to 2024. The dataset contains three primary sources of information:

- **Historical Competition Data:** This segment encompasses details such as outcomes of heats, locations of competitions, among other relevant metrics.
- **Surfer Statistics:** This category provides detailed information regarding the competitors. It includes metrics such as the surfers' rankings, stances, average heat totals, and historical performance at various competition venues.
- **Wave Conditions:** The dataset also includes quantifiable metrics related to the surfing environment, covering aspects like wave height, wind direction, and the type of waves encountered.

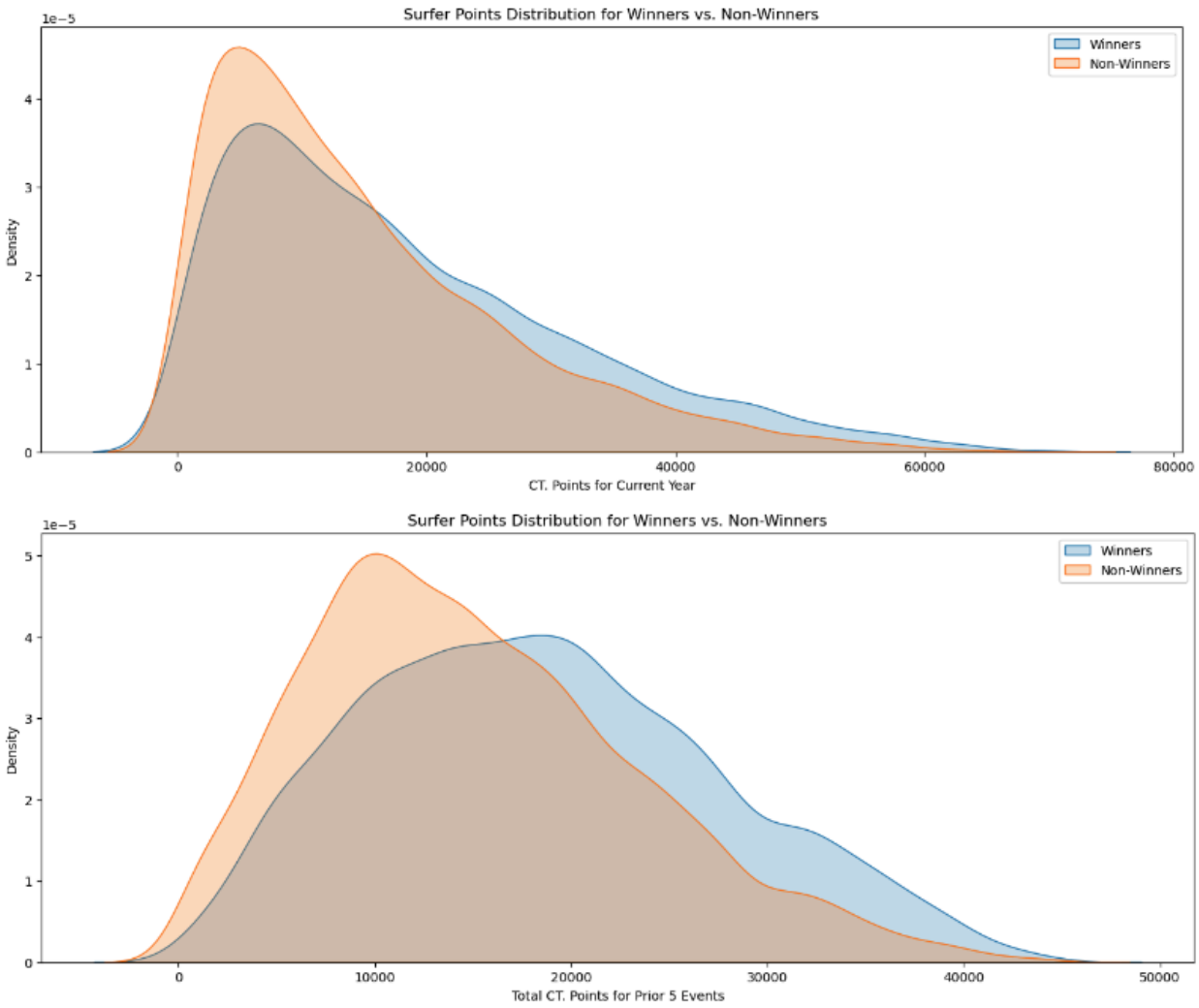

Fig 1. Surfer Points Distribution by Heat Outcome

During the initial exploratory analysis of the dataset, clear indicators of predictive capabilities were not immediately evident. As demonstrated in Figure 1, the distribution of Championship Tour points between winners and non-winners for the current year and prior events displayed a significant degree of overlap. This suggests that while higher points are generally associated with successful outcomes, they are not solely predictive of winning heats, indicating a more complex interplay of variables at work.

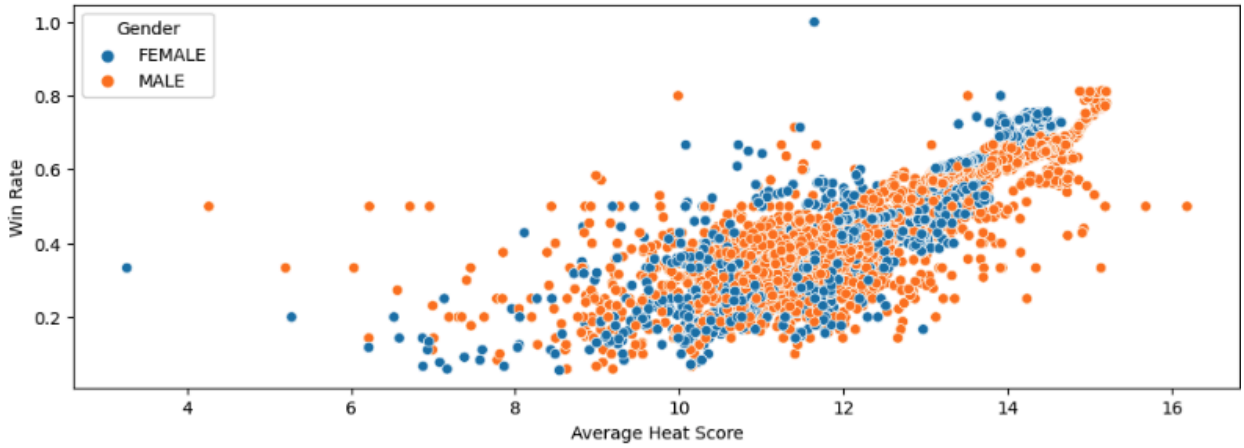

Fig 2. Correlation of Avg. Heat Score and Win Rate by Gender

Similarly, Figure 2 highlights a broad positive correlation between average heat scores and win rates across genders, yet it does not depict a stark differentiation that could be leveraged for straightforward predictions. These visualisations underscore the challenge in identifying distinct correlations within the competitive and multifaceted environment of surf competitions, necessitating a more nuanced approach to model development that can capture the subtleties and intricacies inherent in the data.

2.2 Data Processing: Redundant or irrelevant features were removed, alongside heats conducted in wave pools as the competition format is incompatible with that of regular events. Surfer-specific data were pruned to remove any superfluous information, retaining only the most relevant surfer statistics for the competition analysis. Binary features were created to indicate whether a surfer was competing in their home country, and a feature indicating whether a surfer was competing on their frontside was also generated based on the surfer's stance and the wave direction.

In dealing with missing age data within the dataset, median values were used for imputation based on gender categories. The choice of median over mean as a measure of central tendency was chosen to negate the skewed distribution of ages in the dataset.

For the numeric features, standard scaling was applied to normalise the data and ensure that all features contributed equally to the models' performance. A MinMaxScaler was also employed for features requiring scaling within a specific range. Categorical features were encoded using one-hot encoding, allowing the models to interpret these features numerically.

A training-validation split was performed to separate the dataset into a training set, which the models learn from, and a validation set, which is used to evaluate the models' generalisation performance. This split was conducted with a test size of 30% of the original data and a consistent random state to ensure reproducibility.

2.3 Model Training

Artificial Neural Network (ANN): For the prediction of surf competition heat winners, an ANN was constructed using Keras, a high-level neural networks API. The ANN architecture was designed with an input layer followed by multiple hidden layers and a dropout regularisation applied after every hidden layer. Specifically, the network consists of dense layers with 128, 64, 32, 16, and 8 neurons respectively, each followed by a dropout layer with a dropout rate of 0.5 to prevent overfitting. The activation function for all neurons in the hidden layers was the rectified linear unit (ReLU), chosen for its efficiency and effectiveness in non-linear modelling. The output layer contains a single neuron with a sigmoid activation function suitable for binary classification.

The network was compiled with a binary cross-entropy loss function, which is appropriate for binary classification tasks. The Adam optimizer was utilised for its adaptive learning rate capabilities. To avoid overfitting, an EarlyStopping callback was employed, monitoring the validation loss with a patience of 15 epochs, meaning the training would stop if the validation loss did not improve for 15 consecutive epochs. The model was trained on the training dataset for a maximum of 100 epochs with a batch size of 16, and the validation dataset was used to monitor performance after each epoch.

Random Forest: A Random Forest classifier with 750 estimators was implemented to analyse the same dataset. The Random Forest model operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. This model benefits from its ability to reduce overfitting without substantially increasing error due to bias. The random_state parameter was set to 42 to ensure reproducibility of the results.

XGBoost: The third model applied was XGBoost, an optimised distributed gradient boosting library designed to be highly efficient, flexible, and portable. The XGBoost classifier was configured for a binary classification problem with the objective set to 'binary:logistic'. The number of gradient boosted trees was set to 100 (n_estimators=100), with a maximum depth of each tree capped at 6. The learning rate was set to 0.1, and the subsample ratio of the training

instance was 0.8, which means 80% of the data was used for each tree's growth. Similar to the Random Forest model, the seed was set to 42 for consistent results across runs.

Ensemble Model with Log Loss Weighing: To enhance the predictive performance, an ensemble model was created by integrating the predictions of the three aforementioned models based on their log loss with the test dataset. Log loss provides a measure of accuracy, where a lower log loss indicates better performance. For each model, the inverse of the log loss was used as a weight, assuming that a model with lower log loss would contribute more to the final prediction. The weights for the ensemble were normalised so that they summed up to 1, creating a weighted average ensemble model that combines the individual strengths of the ANN, Random Forest, and XGBoost models.

3. Results

Each model (ANN, Random Forest, XGBoost) was trained on the training set evaluated on the validation set using the following classification metrics:

- **Accuracy:** Measures the overall proportion of correctly predicted heat winners.
- **Precision:** Indicates the ratio of true positives (correctly predicted winners) to all positive predictions by the model.
- **Recall:** Represents the ratio of true positives (correctly predicted winners) to all actual winners in the validation set.
- **F1-Score:** A harmonic mean of precision and recall, providing a balanced view of model performance.

		Precision	Recall	F1-Score	Overall Accuracy	Support
ANN	0	0.72	0.74	0.73	0.70	3135
	1	0.66	0.65	0.66		2396
Random Forest	0	0.71	0.74	0.73	0.70	3135
	1	0.69	0.64	0.67		2396
XGBoost	0	0.73	0.75	0.74	0.71	3135
	1	0.68	0.66	0.68		2396

Fig 3. Model Comparison by Classification Metrics

As stated in figure 3, the ANN model exhibited an overall accuracy of 70% with an F1-score of 0.73 for class 0 (surfer losing the heat) and 0.66 for class 1 (surfer winning the heat). The Random Forest model also achieved a 70% accuracy rate, with F1-scores of 0.73 for class 0 and 0.67 for class 1, indicating consistent performance across both classes. The XGBoost model performed marginally better, registering a 71% overall accuracy. Its F1-scores were 0.74 for class 0 and 0.68 for class 1, suggesting a slightly more balanced classification ability, particularly in correctly predicting heat winners.

The results highlight that while all models achieved a comparable level of accuracy, XGBoost demonstrated a superior balance between precision and recall, as evidenced by its F1-score. This indicates that XGBoost was marginally more effective at classifying both positive and negative classes in this specific dataset.

It is noteworthy that all models displayed a tendency towards higher precision and recall for class 0 predictions. This could suggest a potential bias in the models towards predicting the loss of a heat rather than a win, or it may reflect an inherent characteristic of the dataset.

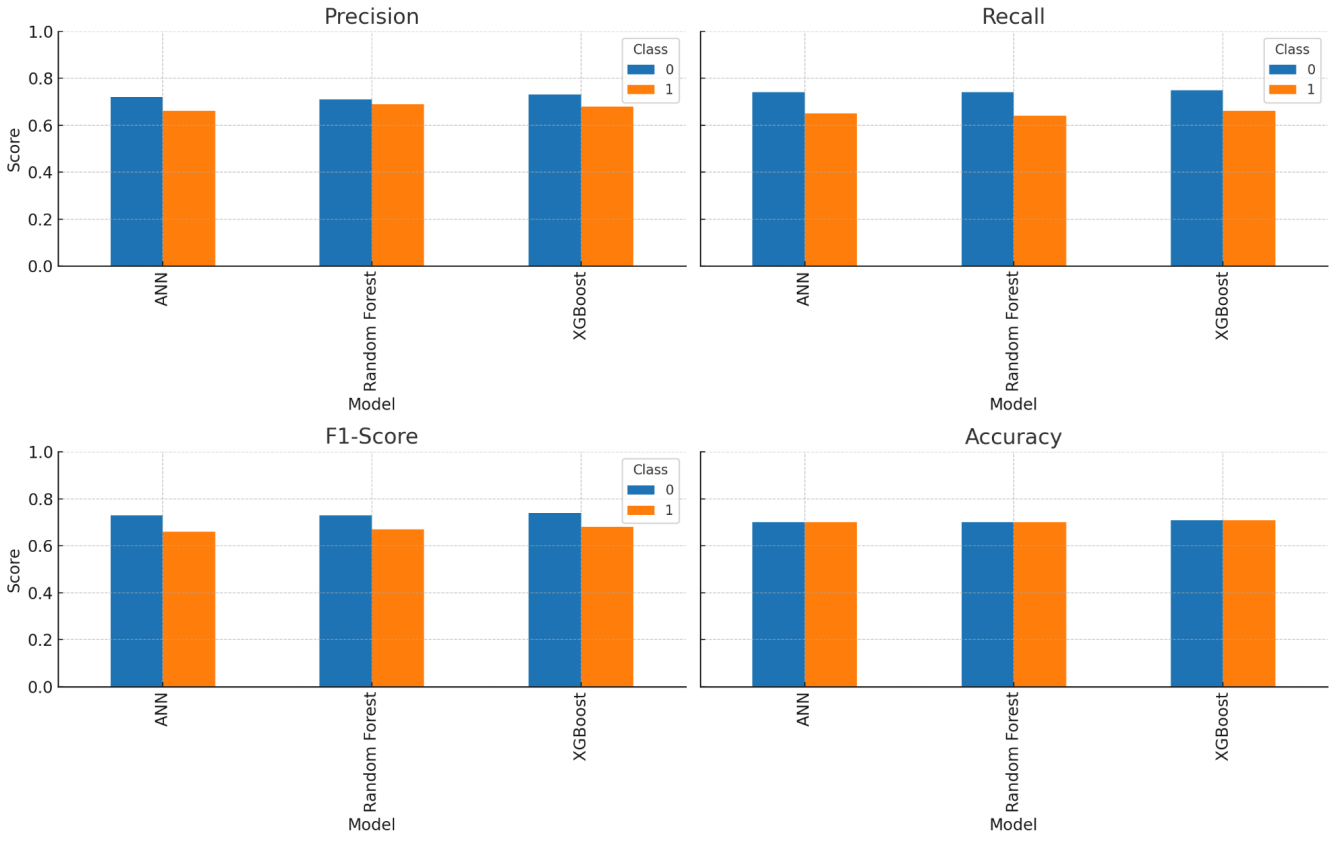

Fig. 4 Bar Model Comparison

3. Conclusion

The analysis demonstrates that among the models tested, XGBoost yielded the highest accuracy in addressing the complex and variable nature of competitive surfing. Although ANNs and Random Forest models displayed comparable levels of accuracy, XGBoost distinguished itself with marginally superior F1-scores. This indicates that machine learning, particularly XGBoost, has the capacity to navigate the unpredictable dynamics of surf competitions. To further refine the precision of these predictions, future work should focus on enlarging the dataset and integrating a broader spectrum of variables in order to capture the high variability characteristic of the sport.

The core objective of this project was to assess the feasibility and effectiveness of machine learning techniques in forecasting outcomes of surf competitions. Among the tested models, XGBoost showed a notable proficiency, achieving the highest accuracy and F1-scores, thus underscoring its suitability for this specific predictive task.

The study confirms that machine learning can be a viable tool for predicting the outcomes of surf competitions, with XGBoost providing the most reliable forecasts within the scope of the current dataset. Although ANNs and Random Forests performed adequately, the marginal superiority of XGBoost indicates that its algorithmic nuances are well-suited to the complex patterns present in competitive surfing data.

Whilst promising results, this study was limited by the dataset's size and variable scope. The choice to only incorporate Championship Tour heats forced the models to make predictions based on very little to no prior performance metrics for new athletes that had just qualified for the Championship Tour. A more comprehensive dataset could be achieved by incorporating heats from lower level competitions. Furthermore, the supplementation of a historical weather database would likely provide more reliable and detailed condition variables than those used in this project.

In closing, the findings highlight the potential of XGBoost in surf competition forecasting, establishing a benchmark for future work in the field. The next steps involve expanding the dataset and refining the models to further improve prediction accuracy. This project has laid the groundwork for such advancements, pointing the way for more in-depth and comprehensive approaches.