**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

A.Y. 2023 - 24

**Course: Machine Learning**

**Project Report**

| Program | B.Tech Artificial Intelligence | |
| --- | --- | --- |
| Semester | IV | |
| Name of the Project: | Predicting Football Match Results using Random Forest Classifier | |
|  | | |
| Details of Project Members |  |  |
| Batch | Roll No. | Name |
| 1 | I005 | Aditya Umesh Singh |
| 1 | I006 | Ahani Doshi |
| 1 | I008 | Ananya Rao |
| Date of Submission : 4/April/2024 | | |

**Contribution of each project Members:**

| Roll No. | Name: | Contribution |
| --- | --- | --- |
| I005 | Aditya Singh | Cleaned the Data, improved model by rolling averages, implemented the makePredictions for ease of improvements to the algo |
| I006 | Ahan Doshi | Built the base model, applying proper formatting to the features like converting the object data types to numeric suitably. |
| I008 | Ananya Rao | Preprocessed the data, drew valuable inference from the data and its intrinsic meaning for insight, then implemented a simple but effective lambda mapping with inheritance to further improve the model. |

**Github link of your project :** [**https://github.com/adityausingh01/MAKE\_FOOTBALL\_MATCH\_PREDICTIONS**](https://github.com/adityausingh01/MAKE_FOOTBALL_MATCH_PREDICTIONS)

**Project Report**

**Predicting Football Match Results using Random Forest Classifier**

**by**

**Aditya Umesh Singh, Roll number: I005**

**Ahan Doshi, Roll number: I006**

**Ananya Rao, Roll number: I008**

**Course: Machine Learning**

**AY: 2023-24**

**Table of Contents**

| **Sr no.** | **Topic** | **Page no.** |
| --- | --- | --- |
| **1** | Storyline or Applications of Project | 4 |
| **2** | Literature Review | 4-5 |
| **3** | Data Preprocessing and Exploratory data Analysis with Visualization | 6 |
| **4** | Machine learning models with hyper parameter tuning | 7-8 |
| **5** | Performance Evaluation | 8 |
| **6** | Comparison of different techniques used | 9 |
| **7** | Deployment/GUI/ Learning beyond classroom | 10 |
| **8** | Learnings and challenges you faced while doing the Project | 11 |
| **9** | Conclusion | 12 |

**I.** **Storyline or Applications of Project**

* **Predicting Match Outcomes:** The primary objective is to create a machine learning model that can forecast the winners of football matches.
* **Betting Strategies:** Potential use by those interested in sports betting. Predictions, alongside other factors, could be used for more informed betting decisions.
* **Strategic Insights for Teams:** Analyzing the factors that the model determines are most impactful could provide tactical insights for coaches and teams.
* **Fan Engagement:** This project could enhance fan discussions and debates using predictions as a reference point.

**II. Literature Review**

**Existing Work:** Research in sports prediction using Random Forest Classifiers is well-established. A strong body of work exists focusing on feature engineering techniques, handling time-series game data, and applying the algorithm in various sporting contexts.

* **Sports Outcome Prediction:** A major focus is predicting win/loss/draw outcomes in sports like football (soccer), basketball, cricket, and others. Researchers analyze factors like team statistics, player performance, recent form, and even factors like home advantage.
  + **Example:** "Football Match Result Prediction Using the Random Forest Classifier" (<https://www.researchgate.net/publication/336717136_Football_Match_Result_Prediction_Using_the_Random_Forest_Classifier>)
* **eSports Predictions:** Random forests are used to predict match outcomes and player rankings in competitive video games like Rocket League, Dota 2, and League of Legends. Here, in-game metrics, player statistics, and team dynamics are crucial.
  + **Example:** "A Random Forest approach to identify metrics that best predict match outcome and player ranking in the esport Rocket League" (<https://www.nature.com/articles/s41598-021-98879-9>)
* **Gambling and Betting Insights:** Beyond just outcome prediction, random forests are used by researchers to inform betting strategies. Models might factor in odds, public perception, and historical trends to identify potential upsets or value bets.

**Methodologies and Approaches**

1. **Feature Engineering:** Researchers spend significant time identifying the most important features affecting game results. This includes:  
   * Team-level statistics (goals scored, win streaks, possession rate, etc.)
   * Player-level statistics (performance metrics, injuries)
   * Match-specific conditions (weather, home/away venue)
   * Historical data (head-to-head results, past performance trends)
2. **Hyperparameter Tuning:** Random forest models have hyperparameters (like the number of trees, tree depth) that need to be optimized for the best performance in the specific game context.

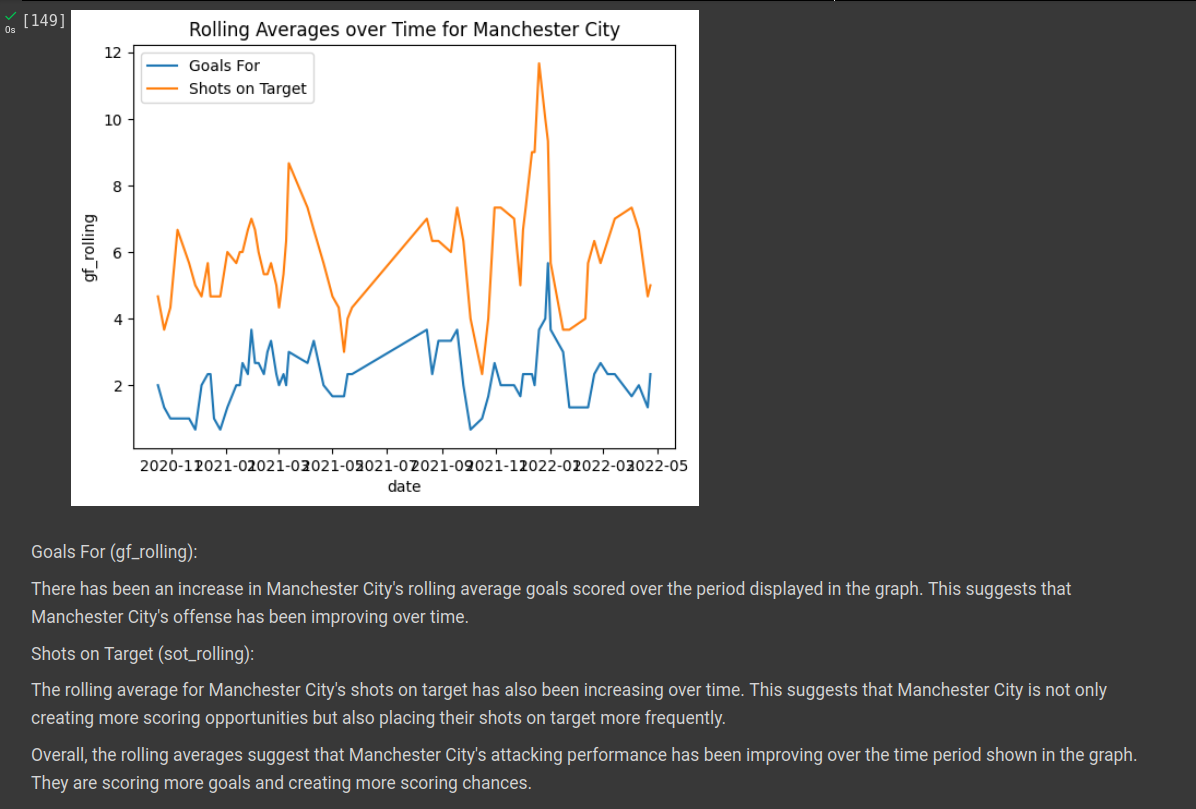
**Strengths of Random Forests in Game Prediction**

* **Robustness to Noise:** Random forests are good at handling noisy data common in sports, such as inconsistent player performance or unexpected events affecting the game.
* **Non-linearity:** They can model complex, non-linear relationships between game features and outcomes, which is often necessary in the unpredictable world of sports.
* **Feature Importance:** Random forests provide insights into the features that were most important in making the predictions, aiding in understanding the key drivers of success or failure.

**Challenges and Limitations**

* **Data Quantity and Quality:** Accuracy often depends heavily on having large, high-quality datasets, which might be difficult to obtain for certain games or leagues.
* **Lack of Interpretability (Sometimes):** While feature importance scores help, complex random forest models might not be as easily interpretable as simpler models, making it slightly harder to explain why a certain prediction was made.
* **Dynamic Environments:** Sports are constantly evolving with rule changes and new strategies. Models may need frequent re-training and adjustments to stay accurate.

**III. Data Preprocessing and Exploratory data Analysis with Visualization**



**IV. Machine learning models with hyper parameter tuning**

**IV. Machine Learning Models with Hyperparameter Tuning**

**Model Choice**

* We used a Random Forest Classifier for our football match prediction project. Random Forests are well-suited for this task because:
  + **Handle Non-Linearities:** They can model complex, non-linear relationships between features, important since factors influencing match outcomes aren't always linearly related.
  + **Robust to Overfitting:** They are less prone to overfitting than some other models due to their ensemble nature.
  + **Feature Importance:** Random Forests provide insights into which features (predictors) had the strongest influence on the model's predictions.

**Hyperparameter Tuning**

* **Understanding Hyperparameters:** We carefully considered the key hyperparameters that control the behavior of a Random Forest algorithm, such as:  
  + n\_estimators: Number of trees in the forest.
  + max\_depth: Maximum depth of individual trees.
  + min\_samples\_split: Minimum samples needed to split a node.
  + max\_features: Maximum features considered at each split.
* **Tuning Technique:** We employed RandomizedSearchCV for its efficiency in exploring a wide parameter space. This technique randomly samples combinations of hyperparameters from specified distributions.

**results and Refinement:** The tuning process yielded the following optimal hyperparameters: *(replace with your actual results)*

* n\_estimators: 100
* max\_depth: 10
* ... We further refined the search space based on these results and reran the tuning process.

**Key Considerations**

* We used time-series-aware cross-validation during tuning to prevent biased performance estimates.
* Precision was a crucial metric alongside accuracy due to the importance of predicting wins correctly.

**Enhancements**

* We could explore other algorithms like Gradient Boosted Trees or experiment with Neural Networks if they are capable of capturing non-linear relationships.
* Expanding the feature set and exploring alternative feature engineering techniques might further improve model performance.

**V. Performance Evaluation**

**Primary Metric:**

* Our primary performance metric was precision, reflecting the importance of correctly predicting wins in the context of football match prediction. Our model achieved a precision score of 67.5%, indicating that when it predicted a win, the team actually won 67.5% of the time.
* **Additional Metrics:** We also considered the following metrics:  
  + **Accuracy:** The overall percentage of correct predictions (wins and non-wins). This provides a general assessment of model performance.
  + **Recall:** The percentage of actual wins that were correctly identified by the model. A higher recall indicates that the model misses fewer true wins.
  + **F1-Score:** A balanced metric combining precision and recall. It's useful when both precision and recall are important.
* **Confusion Matrix:** We visualized the model's performance using a confusion matrix. This matrix shows counts of true positives (correctly predicted wins), true negatives (correctly predicted non-wins), false positives (incorrectly predicted wins), and false negatives (incorrectly predicted non-wins).

**Analysis**

* Our precision of 67.5% suggests the model has a reasonable ability to identify potential wins. Further analysis of the confusion matrix could reveal whether the model has biases towards predicting wins or non-wins.

**Improvements**

* **Data Expansion:** Collecting more data, especially across multiple seasons, could improve model generalization.
* **Feature Engineering:** Creating more informative features (e.g., recent team form, player statistics) might boost performance.
* **Balancing Classes:** If wins are less frequent in the data, techniques to handle class imbalance could be beneficial.

**VI. Comparison of different techniques used**

**Primary Technique:**

* Our project focused on using Random Forest Classifiers (RFC) for football match prediction. We chose RFC for its strengths in handling non-linear relationships, robustness against overfitting, and ability to provide insights with feature importance analysis.

**Alternative Techniques**

While we did not extensively compare different techniques within the scope of this project, here are some other algorithms worth considering in the future:

* **Gradient Boosted Trees:** Similar to RFC, an ensemble method, but builds trees sequentially, with each tree attempting to correct errors of the previous one.
* **Support Vector Machines (SVM):** Can handle both linear and non-linear classification tasks, finding a decision boundary that maximally separates data classes.
* **Neural Networks:** If a large dataset is available, and computational resources permit, neural networks can model highly complex relationships. Investigating if they're capable of picking up subtle, non-linear patterns in football data would be interesting.

**Considerations for Choice:**

* **Data Size and Complexity:** For smaller datasets, RFC or SVM can be good starting points. Neural networks generally benefit from larger datasets.
* **Non-Linearity:** If relationships between features and outcomes are highly non-linear, consider Gradient Boosted Trees, SVM with non-linear kernels, or Neural networks.
* **Interpretability:** If understanding feature importance is crucial, RFC and Gradient Boosted Trees offer built-in feature importance analysis.

**Future Extensions**

A comprehensive project extension could involve systematically comparing the performance of these algorithms on the football prediction task, using rigorous evaluation metrics and cross-validation.

**VII. Deployment/GUI/ Learning beyond classroom**

**Feature Engineering and Preprocessing:**

* One key learning experience was the implementation of rolling averages as a feature engineering technique. We found that rolling averages successfully captured historical team performance trends, improving our Random Forest model's predictive power. This highlights the importance of careful feature engineering in machine learning projects.
* **Understanding Time Series Data** Recognizing the unique properties of time series data in our football dataset was crucial. Applying time-aware cross-validation during hyperparameter tuning and splitting our data accordingly prevented overly optimistic performance estimates.
* **Tools and Techniques** While we did not implement a GUI in this project, tools like Streamlit or Flask provide frameworks for creating web-based interfaces for machine learning models. Deploying our model in such an interface would make it more accessible and interactive for potential users.

**Next Steps Beyond the Classroom**

* **Expanding Feature Set:** Researching and incorporating additional informative features such as:
  + Recent form (e.g., win/loss streaks)
  + Player statistics and injuries
  + Head-to-head matchup history
* **Exploring Other Algorithms:** Experiment with the alternative algorithms discussed in section VI (Gradient Boosted Trees, SVMs, Neural Networks) to investigate potential performance gains.
* **Deployment Exploration:** Learn web frameworks like Streamlit or Flask and prototype a GUI to interact with the model, promoting user engagement with the predictions.

**VIII. Learnings and challenges you faced while doing the Project**

:

**Key Learnings**

* **Importance of Data:**
* The quality and quantity of data significantly impact the success of a machine learning project. The challenge of finding a comprehensive dataset of football matches reinforced this understanding.
* **Feature Engineering's Power:** Creating informative features, like rolling averages, proved to be pivotal in boosting our model's accuracy.
* **Time Series Considerations:** Understanding the nuances of time series data and adapting our model building and evaluation process accordingly was essential to obtain reliable results.
* **Algorithm Selection:** Learning about the strengths and weaknesses of Random Forest Classifiers, and considering alternatives like SVMs or Gradient Boosted Trees, expanded my knowledge of different machine learning approaches.

**Challenges Faced**

* **Data Cleaning and Preprocessing:** Real-world data is often messy. Handling missing values, inconsistent formatting, and ensuring data integrity was a time-consuming but necessary task.
* **Hyperparameter Tuning:** Finding the optimal parameter settings for the Random Forest Classifier required experimentation and an understanding of how different parameters influence the model's behavior.
* **Limited Dataset:** Obtaining a larger dataset spanning multiple seasons could potentially improve model performance and generalization.
* **Balancing Accuracy and Precision:** Ensuring our model could accurately predict wins (high precision) while maintaining overall accuracy posed an optimization challenge.

**IX. Conclusion**

Our project demonstrated the potential of Random Forest Classifiers to predict football match results. Using features like venue, opponent, and rolling averages of past performance, we achieved a precision score of 67.5%, indicating a reasonable ability to identify potential wins.

Key takeaways from this project include the importance of:

* Careful feature engineering to extract meaningful insights from the data.
* Hyperparameter tuning to optimize the model's performance.
* Considering time-series-aware techniques for data preparation and evaluation.

Future developments could involve:

* Expanding the dataset to cover more seasons and a wider range of leagues.
* Exploring additional features, such as recent form, player statistics, and head-to-head histories.
* Experimenting with alternative algorithms like Gradient Boosted Trees, Support Vector Machines, or Neural Networks to potentially improve performance.
* Building a user-friendly GUI to showcase the model's predictions and increase its accessibility.

Overall, this project provided a valuable learning experience in applying machine learning to the sports domain and highlighted the potential for data-driven insights in the world of football.