

**UMD Baseball**

**Team 6 - Documentation**

**BUDT\_702\_06\_Documentation**

Shashank Bisht

Dhruvin Shah

Nilay Somani

Anudeep Koneru

**What our team used?**

### **Tools and Technologies Utilized:**

* Programming Language: Python
* Libraries: Utilized Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data manipulation, preprocessing, visualization, and predictive modeling.
* Environment: Worked in Google Colab for interactive data exploration, model development, and analysis.

### **Machine Learning Models and Techniques:**

* Linear Regression model - Using historical baseball data, we built a linear regression model to identify patterns and relationships between different variables such as location, opponent, and game outcomes for the same. By analyzing the impact of these factors from the historical trends, the regression model can provide insights into potential future outcomes.
* Logistic Regression: As a second model, we built a logistic regression model to classify the results into W/L as binary classification.
* Model Evaluation: Evaluated model performance using techniques and metrics like RMSE (Root Mean Squared Error), Confusion Matrix, Accuracy and Recall.

### **Data Preprocessing:**

* We processed the existing dataset segregating wins and losses for a year, Home and Away records, extracted year from game records, etc.
* On testing data of the 2024 games schedule, we first downloaded the dataset from UMTerps and cleaned and prepared that data completely for our model use.

### **Predictive Modeling for 2024:**

* Utilized Historical Data: Leveraged historical baseball data to train models capable of forecasting team performance against opponents at which locations and game outcomes for the 2024 season.
* Informed Model Development: Employed insights from exploratory data analysis to inform feature selection and model development.
* Thorough Evaluation: Conducted thorough model evaluation to ensure reliability and accuracy of predictions for the upcoming season.

**What our team searched?**

1. **Dataset for 2024 Baseball schedule**

We looked for the 2024 baseball dataset on UMTerps at <https://umterps.com/sports/baseball/schedule> and took the pdf of the schedule and results of the 2024 games till date.

1. **Model selection for predicting results on 2024 data**

We searched through the internet, course modules and existing work to understand what models will be the best ones to use for our mission objectives. We also checked on what model gives us best accuracy and prediction for our given dataset.

1. **Previous Activities and Assignments**

We searched through the past activities and assignments of our Python course for the pieces of information relevant to our project.

**What our team studied?**

* We studied course modules to understand why we use particular models and what is their effectiveness based on use-cases.
* In addition to the above, studying the assignments again, especially HW5, was really helpful to strengthen our domain knowledge.
* We referred pieces of our individual past projects to better our domain knowledge for us to better the flow of our prediction measures.

**What were the challenges?**

* Deciding the dataset for 2024 - Whether to evaluate our results based on existing data or work entirely on new dataset to be taken from UMTerps data source
* Data size & supporting details- In predicting games results, it's important to have a large dataset to estimate a result based on certain parameters. In addition to this, without player performance, it became slightly tricky to analyze the backend reasons behind the team’s performance dip/increase apart from the given dataset - whether it was addition of new players, experienced players leaving or injury management, etc.
* Model Selection - Choosing the appropriate machine learning algorithm or model architecture for the problem at hand can be challenging. Different algorithms have different strengths and weaknesses, and selecting the best one often requires experimentation and domain knowledge.
* Single ‘Draw’ record - As ‘draw’ or ‘tie’ is an important outcome of the game but there was only a single record in the entire 24 years of given data, we were skeptical whether to keep it or go ahead without it.

**What have we accomplished?**

* Data Preparation:

Preprocessed the dataset to handle outliers and inconsistencies, ensuring data quality and consistency.

* Exploratory Data Analysis (EDA):
  + Conducted extensive exploratory data analysis to gain insights into the dataset's characteristics, distribution, and relationships between variables.
  + Identified key features and trends that could potentially impact baseball game outcomes, such as team performance, performance against opponents, and game locations.
* Feature Engineering:
  + Engineered new features from the existing dataset, such as aggregating home/away statistics, calculating team performance metrics, and extracting relevant information from game records.
  + Selected and transformed features to improve model performance and capture meaningful patterns in the data.
* Model Development:
  + Built predictive models using machine learning algorithms including Linear Regression and Logistic Regression.
  + Trained the models on historical baseball data to predict outcomes such as team game results for the upcoming season (2024).
* Documentation and Reporting:
  + Documented the entire project workflow, including data preprocessing steps, model development, evaluation results, and insights derived from the analysis.
  + Stated sources of our findings, references and past work used in helping us navigate the built-in functions
  + Created comprehensive reports and visualizations to communicate the project findings effectively to stakeholders and team members.
* Future Work and Recommendations:
  + Identified areas for future improvement and refinement, such as collecting additional data sources, exploring advanced modeling techniques, and refining feature engineering strategies.
  + Provided recommendations for further research and development to enhance the predictive accuracy and applicability of the models in real-world scenarios.

**Ways to test and evaluate our model**

* Predicting results on 2024 season (till date) - Based on the results for different parameters obtained from 1999-2023 Baseball data, our trained model can predict the results for upcoming games given features like opponent, location, etc. The prediction can be evaluated by comparing our predicted results with the actual results of 2024 games played till now.
* Test on our training data itself (1999-2023) - The model prepared based on the improved features, can be tested on the training data (given data) itself to understand what accuracy it is giving on results of the games played from 1999-2023. The evaluation can be done by checking accuracy, confusion matrix and RMSE to understand how well our model has been trained with the given optimized parameters.

**Team members contribution**

* Shashank Bisht - Instrumental in setting up the presentation of the project.
* Dhruvin Shah - Instrumental in model preparation and evaluation.
* Nilay Somani - Instrumental in data processing, data cleaning of test data and project documentation.
* Anudeep Koneru - Instrumental in EDA, setting up mission objectives, mission statements, etc.

Above all, each of the team members did switch up tasks as per the requirements and collaborated effectively to manage workload and were not constrained to their delegated tasks.

**Project Learnings**

* Predictive Analytics in Sports:
  + Insights gained from applying predictive analytics techniques to sports data, including understanding team performance, team dynamics, and game outcomes.
  + Potential applications of predictive modeling in sports management, team performance optimization.
  + Baseball is a new sport to all of us in terms of understanding statistics, so the project helped us in learning how the stats are reported for the game and what major terminologies mean for the same.
* Collaboration and Communication:
  + Importance of effective collaboration and communication within the project team, including sharing insights, discussing findings, and resolving challenges.
  + Some of us had prior relevant projects which we referred to and shared our knowledge of how we did back then.
* Continuous Learning and Improvement:
  + Recognition of the iterative nature of data analytics projects and the importance of continuous learning and improvement.
  + Opportunities for further research and development, including exploring advanced modeling techniques, incorporating additional data sources, and refining predictive models for enhanced accuracy and applicability.