# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Kiran Shanker Das |
| **Project Name** | AT1- Kaggle Competition |
| **Date** | 18th August 2023 |
| **Deliverables** | <best\_model.pynb>  < XGBoost>< Logistic Regression>  Github Link: <https://github.com/kirandas-dev/NBAComp> |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?  **Goal and Utilization:** The project's core aim is to construct a predictive model that assesses the likelihood of a college basketball player being drafted by a professional team. By analyzing player performance metrics, the model aims to identify the key factors contributing to draft eligibility. The model's predictions will be crucial for basketball teams and scouts during the draft selection process, guiding their decisions on player choices. This data-driven approach will help teams allocate resources more effectively, identify hidden talent, and refine their draft strategies to align with team requirements.  **Impact of Results:** Accurate predictions will offer substantial advantages. Teams can efficiently focus their scouting efforts on players with higher predicted draft probabilities, uncover hidden gems, and make informed draft decisions. However, incorrect predictions could lead to missed opportunities, misallocation of scouting resources, and poor draft choices. Thus, the project's success could redefine player draft processes, enhancing decision-making through accurate predictions, while inaccuracies could result in suboptimal outcomes affecting team performance and resource efficiency. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it.  **Hypothesis:** Certain factors and attributes of college basketball players significantly influence their likelihood of being drafted by a professional team.  **Question**: What are the key factors and attributes of college basketball players that contribute to their likelihood of being drafted by a professional team?  Below are the reasons:  **Talent Identification:** Uncovering key attributes for draft success aids scouts, coaches, and players, guiding development programs and recruitment.  **Resource Allocation:** Pinpointing vital factors helps teams optimize resource use in scouting and focus on critical player traits.  **Informed Decisions:** Understanding draft factors empowers college players and advisors in career choices.  **Player Development:** Tailoring training to valued attributes readies college players for the transition to the pros.  **Objective Scouting:** Predictive models reduce bias, aiding scouts in assessing and comparing players fairly.  **League Competitiveness:** Draft insights improve team composition, elevating the competition and fan excitement.  **Fan Engagement:** Predictive insights fuel fan discussions about potential draftees and team suitability.  **Analytics Advancement:** Successful model construction showcases analytics' role in sports decisions, encouraging data-driven approaches in sports management. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  **Expected Outcome:**  The experiment aims to create a predictive model for a college basketball player's draft likelihood. The expected outcome is a model with around around 95% accuracy in predicting draft success.  **Possible Scenarios:**  **High Accuracy:** Model accurately predicts draft outcomes (95%+ accuracy), providing valuable insights.  **Moderate Accuracy:** Model has 70-90% accuracy, still offering useful guidance.  **Low Accuracy:** Model's accuracy is below 65%, prompting attribute reevaluation.  **Bias:** Model shows accuracy but highlights bias, emphasizing fairness importance.  **Fit Issues:** Model overfits/underfits, requiring adjustments.  **Surprising Attributes:** New insights from unexpected attribute correlations.  **External Factors:** Model reveals external factors' importance.  **Feedback Loop:** Stakeholder feedback iteratively improves the model. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  **Data Filtering:** We loaded data from a CSV and filtered out columns with excessive missing values.  **Missing Value Imputation:** Filled numerical missing values with medians and categorical with modes.  **Feature Dropping:** Removed the 'year' column and other irrelevant columns.  **Train-Validation-Test Split:** Split data into training, validation, and test sets, while keeping an eye on class distributions.  **Data Resampling:** Applied SMOTE for balancing and standard scaling for normalization.  **Saving Processed Data:** Saved processed datasets and resampling tools (Scaler, SMOTE) using joblib.  **Visualization:** Used correlation heatmap to identify highly correlated features.  **Highly Correlated Features:** Found pairs of features with strong correlations, crucial for addressing multicollinearity. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments  **Trained Models:**  **Logistic Regression:**  Hyperparameters tested: Regularization parameter (C) with values [0.001, 0.01, 0.1, 1, 10, 100].  Rationale: Logistic Regression is a simple and interpretable model suitable for binary classification. We used grid search to find the best regularization parameter through cross-validation.  **XGBoost Classifier:**  Rationale: XGBoost is an ensemble learning method known for its high performance. It can capture complex relationships and handle imbalanced datasets. We used default hyperparameters for this model.  **Not Trained Models:**  Decision behind not training certain models wasn't explicitly mentioned in the provided information. However, other models like Support Vector Machines, Random Forest, Neural Networks, etc., were not chosen due to their complexity, training time (it took 22 mins to run a Neural Network), or unsuitability for the given problem context.  **Important Model & Hyperparameters:**  **XGBoost Model:** The XGBoost model achieved very high AUROC scores on both validation (0.9987) and test (0.9979) sets, indicating a good discriminatory power. We have chosen a threshold of 1.2 for selecting important features after trying out multiple values. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  **Logistic Regression Model:**  Validation Accuracy: 0.9622  Test Accuracy: 0.9709  **Underperforming Cases/Analysis:** The model generally performs well but may struggle with certain instances.  **Potential Causes:**  **Feature Selection**: Reevaluate selected features' impact.  **Class Imbalance**: Explore advanced methods for class imbalance.  **Model Complexity**: Address if the relationship is non-linear.  **XGBoost Classifier Model:**  Validation Accuracy: 0.9761  Test Accuracy: 0.9669  **Underperforming Cases/Analysis:** High validation accuracy, slight drop on the test set.  **Potential Causes:**  **Overfitting**: Tune hyperparameters to avoid overfitting.  **Feature Importance**: Review feature importance.  **Hyperparameter Tuning**: Fine-tune parameters for generalization.  **Conclusion:** Both models show promise, but refinement can enhance performance. We are investigating causes, monitoring and testing for real-world deployment. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  The experiments revealed that both the logistic regression and XGBoost models performed well in predicting basketball player success. The logistic regression achieved around 96% accuracy on both validation and test sets, indicating its reliability. The XGBoost model achieved higher accuracy on the validation set (97.61%) but slightly dropped on the test set (96.69%), suggesting possible overfitting.  Potential Impacts:  **Recruitment Decisions**: Misclassifying successful players could lead to missing out on talent.  **Team Composition**: Incorrect predictions might affect team performance.  **Resource Allocation**: Wrongly allocating resources to players could hinder growth.  **Fan Engagement**: Misclassified players might impact fan excitement and sponsorships.  **Team Reputation**: Incorrect predictions could harm team reputation and fan loyalty.  **Opponent Strategy**: Mistaken player assessments could affect opponent strategies.  **Financial Consequences**: Poor player choices may result in financial losses.  Continuously refining models, addressing limitations, and validating predictions is crucial to make informed decisions and prevent negative business outcomes. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  Issues Faced During Experiments:  **Data Quality and Preprocessing**:  **Missing Values**: Solved by imputing with median/mode. Future: Explore advanced imputation techniques.  **Feature Selection**: Solved by domain knowledge. Future: Experiment with automated feature selection methods.  **Class Imbalance**:  **Imbalanced Classes**: Solved using SMOTE. Future: Explore other resampling techniques or ensemble methods.  **Model Selection and Performance**:  **Model Choice**: Explored logistic regression and XGBoost. Solved by evaluating multiple models. Future: Experiment with more complex models.  **Overfitting**: Observed in XGBoost on test set. Solved by hyperparameter tuning and feature selection. Future: Regularization techniques, more data.  **Hyperparameter Tuning**:  **Manual Tuning**: Solved by using GridSearchCV. Future: Implement more advanced hyperparameter optimization techniques.  **Interpretability**:  **Feature Importance**: Solved by analyzing coefficients and XGBoost's feature importance. Future: Explore SHAP values, LIME, or other interpretability methods.  **Scalability and Generalization**:  **Data Scale**: Solved by applying standard scaling. Future: Test model scalability with larger datasets.  **Model Deployment**:  **Saving Model**: Solved using joblib. Future: Explore deployment platforms, containerization, and monitoring.  **Evaluation Metrics**:  **Metric Choice**: Solved by using accuracy, AUROC, etc. Future: Consider domain-specific metrics or ensemble models.  Issues to Address in Future Experiments:  **External Validation**: Validate models on more recent or external data to assess generalization.  **Feature Engineering**: Explore creating new features, interactions, or transformation methods.  **Advanced Models**: Experiment with ensemble methods, neural networks, or more complex algorithms.  **Hyperparameter Optimization**: Implement advanced optimization techniques like Bayesian optimization.  **Explainability**: Utilize advanced explainability tools for better model interpretation.  **Temporal Data**: Incorporate time-series features for longitudinal analysis.  **Handling Unseen Data**: Plan for handling new player features not seen during training.  Future experiments should focus on enhancing model performance, interpretability, and robustness while addressing these challenges. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.  **Experiment Insights:**  **Model Performance:** Both Logistic Regression and XGBoost achieved high validation and test accuracies, with XGBoost slightly better. This indicates meaningful pattern learning.  **Feature Importance:** 'adrtg', 'Ortg', 'twoPA', 'AST\_per', etc., emerged as key predictors, aligning with player performance metrics.  **Imbalanced Classes:** SMOTE improved performance by addressing class imbalance, emphasizing class distribution handling.  **Interpretability:** Coefficients and feature importance provided valuable insights, aiding decision-makers.  **Future Experimentation:**  **Incremental Improvements:** Fine-tuning, ensemble methods, and advanced algorithms offer potential for enhanced results.  **Domain Alignment:** Models align with domain knowledge, showing promise for data-driven player selection insights.  **Interpretability:** Further exploring model interpretability methods can yield deeper insights.  **Robustness:** External validation and cross-validation will confirm generalizability.  **Business Impact:** Successful validation could enhance draft decisions and efficiency.  In conclusion, positive outcomes and insights support further experimentation. Refinement, alignment with basketball dynamics, and improved insights remain key focuses. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  **Potential Next Steps and Experiments:**  **Ensemble Models:** Combine Logistic Regression and XGBoost for improved accuracy and robustness.  **Hyperparameter Tuning:** Fine-tune model parameters to enhance performance.  **Advanced Algorithms:** Experiment with more complex algorithms like Random Forest for better predictions.  **Feature Engineering:** Create new features based on domain insights to enhance accuracy.  **External Data Integration:** Combine player rankings and college stats to enhance predictive power.  **Ranking:**  Ensemble Models  Hyperparameter Tuning  Advanced Algorithms  Feature Engineering  External Data Integration  **Deployment Steps (if Achieved Business Goals):**  **Validation:** Test models on new season's data.  **Setup and Integration:** Prepare production environment and integrate selected model.  **Data Pipeline:** Create real-time data feed for predictions.  **Monitoring and Updates:** Implement performance monitoring and continuous model updates.  **Domain Expertise:**  Continued experimentation benefits from domain expertise, aligning models with basketball player evaluation methods and trends, leading to more accurate predictions and actionable insights. |