# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Tahmidul Islam |
| **Project Name** | Adv MLA AT1 |
| **Date** | 7/10/2023 |
| **Deliverables** | Notebook:  Islam\_Tahmidul-24587139-Week1\_LogisticRegression.ipynb  Model:  Logistic Regression  Git Repo: https://github.com/tahmislam21/adv\_mla\_at1 |

|  |  |
| --- | --- |
| 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** | The goal of this project for the business is to create a predictive model that evaluates the probability of a college basketball player being drafted into the NBA based on their current season's statistics.  The results will be invaluable in several ways. NBA teams can make more informed draft choices, potentially improving team performance. Scouts and college coaches can identify promising talents more effectively.  Accurate predictions enhance resource allocation, save time, and optimize player development programs.  Conversely, incorrect results could lead to missed opportunities for both players and NBA teams, affecting talent scouting and team competitiveness, highlighting the critical importance of model accuracy in this context. |
| **1.b. Hypothesis** | The Hypothesis of this experiment is that Logistic Regression Model will give us a high ROC Accuracy Score, and also generate a AUROC Graph that is close to the inverse-L shaped graph.  It is worth considering because Logistic regression provides probabilities as output, allowing for clear decision thresholds. This is particularly useful in scenarios where understanding the likelihood of an event is important. |
| **1.c. Experiment Objective** | The expected outcome is that the model will have a roc score close to 1. The goal is to select the degree of polynomial, to 2 and then to 3, for our Logistic Regression model, to understand which degree works better.  The possible outcomes are that Degree 2 will perform better in both Training and Validation set than Degree 3. The opposite could also be true for Degree 2, but both Training and Validation scores will perform worse if that is the case. |

|  |  |
| --- | --- |
| 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  To start off with, columns 'Rec\_Rank', 'pick', 'dunks\_ratio', were removed because of having too many missing values and columns 'team', 'conf', 'ht', 'yr', 'type','num' and 'player\_id' were removed because they had nominal values.  I decided to not perform any encoding on those nominal columns because they had a high number of unique values. Thus, they would give rise to an excessive amount of extra columns.  Following this, imputations were performed on the remaining null values with mean() values of the columns.  The data were then Scaled/Standardized.  Feature Scaling may be important for future experiments, otherwise future models will not generalise well to the dataset. |
| **2.b. Feature Engineering** | No new features were generated. Rather, PCA was applied to the standardized data to reduce the number of features. There were a total of 60+ rows, and having them all under the running model would lead to overfitting.  The features that were removed and the rationale are mentioned in 2.a. |
| **2.c. Modelling** | Model 1:  Polynomial Logistic Regression (Degree 2)  This model was chosen to capture potential nonlinear relationships between the input features and the binary outcome (drafted or not drafted). The degree 2 polynomial terms allow for quadratic relationships, providing flexibility in modeling.  Model 2:  Polynomial Logistic Regression (Degree 3)  By using a degree 3 polynomial, we aimed to capture even more complex, cubic relationships in the data. This higher-degree polynomial can better fit intricate patterns in the statistics and improve predictive performance if such relationships exist.  Polynomial degree was the only parameter changed. Further degrees were not tested because it could lead to overfitting of patterns.  For future experiments, the regularisation parameter may be an important one which can be varied to find the optimal balance between bias and variance. |

|  |  |
| --- | --- |
| 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** | Model 1:  Train ROC-AUC Score: 0.796  Validation ROC-AUC Score: 0.825  Test: *(Target variable was not available in test dataset)*  Model 2:  Train ROC-AUC Score: 0.628  Validation ROC-AUC Score: 0.645  Test: *(Target variable was not available in test dataset)*  For Model 1 (Degree = 2 ), the linear regression model achieved a higher Score on  both the training set (0.796) and the validation set (0.825) . This indicates that the model has some predictive power and can capture some of the  data-related patterns.  For Model 2 (Degree = 3 ), the linear regression model achieved a lower Score on  both the training set (0.628) and the validation set (0.645). This indicates that the model is actually underfitting on the training data and failing to generalize on the testing data when compared to model 2. |
| **3.b. Business Impact** | Based on the technical results, it can be found that Polynomial Logistic Regression with the degree of order 2 is a good predictor of drafting the basketball players into the team. This can be concluded because it shows good scores on both training and validation set. Thus it will also give a good prediction on the test set.  Accurate predictions enhance resource allocation, save time, and optimize player development programs.  Conversely, incorrect results could lead to missed opportunities for both players and NBA teams, affecting talent scouting and team competitiveness, highlighting the critical importance of model accuracy in this context. |
| **3.c. Encountered Issues** | Missing values:  One of the issues faced during the data preparation phase was  handling missing values. One solution is to impute missing values using techniques such as replacing by mean.  Too many features: The dataset had too many features (60+) which can make models overfit. Thus feature selection or Dimensionality Reduction may be applied to resolve this problem.  Future experiments: Issues that may have to be dealt with in future experiments include  addressing bias in the data, identifying more relevant features, and improving the  regularisation of the model. |

|  |  |
| --- | --- |
| 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** | The Logistic Regression model showed good promise on predicting the probabilities to draft the players.  However, this model can be sensitive to outliers, which can disproportionately influence the model’s predictions. Further experiments can be designed to work with models that are more resistant to this phenomenon.  In addition, when dealing with imbalanced datasets (where one class significantly outweighs the other), logistic regression may produce biased results. |
| **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.  Rank 1:  Training dataset with overfitting-resistant models like SVM: This way, the trends in the dataset can be generalized more strongly without being much influenced by outliers and anomalies. Not only will this produce a stronger prediction but also give more accurate probabilities of drafting of players.  Uplift: is that accuracy scores will be higher and more accurate probabilities will come out.  Rank 2:  Tune Regularisation Parameters of SVM:  Experiment with different values of C parameter of SVM.  Rank 3: Balance out the weights of SVM:  Upliift: The balance of bias-variance trade-off will be greater.  If model is performing well, the steps to deployment are as follows:  (Train, evaluate, and serialize the model) -> Containerize -> Deploy to Production -> Implement Load Balancing -> Monitor performance -> Provide Documentation. |