# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ivan Cheung |
| **Project Name** | Kaggle 1 |
| **Date** | 18.08.2023 |
| **Deliverables** | cheung\_ivan-13975420-week1\_log\_regression.ipynb  Logistic Regression Test |

|  |  |
| --- | --- |
| 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** | This experiment aims to determine if NBA draft picks of players from American and International colleges and international professional leagues can be determined by a player’s record of performances during basketball games, using the statistics of the players from the current season.  If this experiment shows that a player’s performance metrics can accurately predict if a player will be drafted, then teams which use this model to identify the best players to scout or shortlist for draft potential. |
| **1.b. Hypothesis** | This experiment hypothesizes that the likelyhood of a player being drafted is correlated to the performance of the player during the current season.  By this theory, then future draft picks of players can be predicted based on performance metrics for a given season. |
| **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. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS – LOGISTIC REGRESSION** | |
| 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** | For experiment 1, only numeric features were selected. Categorical features were excluded in this proof of concept. This simplifed the need to convert categorical features for modelling (OneHotEncoder). Future models may look to apply OHE to capture these features.  Many features from the dataset contained null values, for the purposes of initial testing, a 0 value was added. However not all features were numeric, and some features were floats containing negative values. In future models, evaluation needs to be made on the best possible placeholder value for each feature, if it should be a 0, a minimum negative value, or a cohort average value.  A correlation matrix was also performed, to identify highly correlated features. While the matrix did highlight a number of features with high correlation, no removal of features was carried out for this first experiment. This work will most likely be considered for future experiments. |
| **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  A standard scalar was used to clean the features, this step is necessary to improve performance (normalizing the numeric values), as each feature had a wide gamut of values along the number line. |
| **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  A logistic regression model was selected for this experiment. This model was chosen over a linear or polynomial regression due to the range of features available for testing. I believe, given previous experience that this model would likely perform better than either a linear or polynomial model simply by the feature set at hand.  For the initial experiment, no hyperparameter tuning was carried out as I wanted to see how successful the modelling stands on it’s own merits. |

|  |  |
| --- | --- |
| 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** | S What I found, was that only a very small number of players (~100) are drafted each season, against a backdrop of thousands of players. Due to this, it is very easy to identify “true negatives” (players not drafted, predicted correctly). However, the aim should be to focus on reducing the “false negative” results (players drafted, not predicted correctly). |
| **3.b. Business Impact** | This model, although has a very high prediction success rate, in reality performs quite poorly in accurately predicting drafts:  39 out of 80 drafted players were incorrectly picked (50% success rate).  This model is not yet ready for business use and further work is required to improve it’s performance. |
| **3.c. Encountered Issues** |  |

|  |  |
| --- | --- |
| 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** | Because the postive target (drafted players) is a very small subset of the data, ie:  80 out of tens of thousands; research needs to be done on how to best improve modelling when there is a highly skewed target distribution. |
| **4.b. Suggestions / Recommendations** | Next Steps: - Look at how prediction models account for highly skewed target value distributions.  - Look at what hyperparameter tuning is likely to improve the performance of this model. |