# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ivan Cheung |
| **Project Name** | Kaggle 1 |
| **Date** | 03.09.2023 |
| **Deliverables** | cheung\_ivan-13975420-week3\_adaboost\_searchgrid.ipynb  AdaBoost Decision Tree Classification with GridSearch. |
| **Github** | https://github.com/ivanutsmdsi/amla2023 |

|  |  |
| --- | --- |
| 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.  The objective of this experiment will be to accurately predict if a player will be drafted or not, while minimising the false positive and false negative predictions. |

|  |  |
| --- | --- |
| 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** | Updated from experiment 1, the data preparation has been programed into functions, to enable reusability and improved coding practice.  Two steps have been set up in data preparation:   * Remove unwanted features; several features were removed, with the following justifications:   # ftr - no description was given in the data dictionary  # yr – categorical (this will be included in future experiments)  # ht - player height data in source is corrupted  # num – the player jersey number is cosmetic and does not have any bearing on a player’s ability to be drafted.  # pfr - no description was given in the data dictionary  # type – this is a metadata field and is not relevant as a feature  # year – because our dataset is looking at a subset of 1 year, this feature is not relevant   * Replace null values. As the classifier model cannot handle null values, a placeholder value of ‘0’ has been used. However, this value will need to be further explored in future experiments. |
| **2.b. Feature Engineering** | This experiment expands upon the pipeline work initiated in the previous experiement. The model now uses the AdaBoost classifier instead of a logisitic regression. A GridSearch optimiser has also been included to iterate through different sets of hyperparameter tuning.    Standard Scaler: scales the numeric features to smaller values. This improves the performance of the modelling, by reducing value complexity.  OneHotEncoder: this converts the categorical values into numeric columns, which allows these features to be analysed by the classifier as a feature.  SMOTE: this function performs the upsampling. This is needed as the positive target has a ratio of 1:1000. |
| **2.c. Modelling** | For this experiment, the AdaBoost classifier was used instead of the logistic regression. The base estimator for the adaboost classifier was a Decision Tree Classification model.  The GridSearchCV model optimiser was used on top of AdaBoost to allow for several hyperparameters to be tested. For the parameter testing, two hyperparameters were modified:  n\_estimates: 100 and 200  learning\_rate: 0.5, 1 and 2  Analysis of the grid search results identified one set of hyper parameters as the best, based on the roc\_auc score: |

|  |  |
| --- | --- |
| 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** | The change to adaboost and grid search allowed the tuning of hyperparameters.  The performance of the best scoring model from the grid search CV produced an AUROC score of 0.9972383 and the following confusion matrix results: |
| **3.b. Business Impact** | This model has a very high accuracy rate with minimal false negative predictions.  Use of this model in a business context will not likely miss any true drafts. |
| **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** | The adoption of a hyperparameter optimiser function from sklearn made iterating through different permutations of the model very effective and easy.  Ideally, model validation should also be conducted to further improve the modelling predictions. |
| **4.b. Suggestions / Recommendations** | Next Steps: - Experiment with model validation through cross validation methods. |