# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ivan Cheung |
| **Project Name** | Kaggle 1 |
| **Date** | 27.08.2023 |
| **Deliverables** | cheung\_ivan-13975420-week2\_log\_reg\_smote.ipynb  Logistic Regression with upsampling and pipeline |
| **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** | A new data modelling pipeline was created in this experiment. This pipeline performs the following feature engineering steps:    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** | 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** | Based on the results of experiment 1, two changes were made:   1. Inclusion of categorical features (model enhancement), and 2. Applying upsampling. The application of upsampling was used to address the skewness of the target data.   These changes saw a significant improvement in the false negatives (0%), but there was an increase in false positives (+1.8%). |
| **3.b. Business Impact** | 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** | The creation of the pipeline and preprocessor steps will greatly improve experiment efficiency moving forward.  The use of upsampling is a positive direction for the experiments. |
| **4.b. Suggestions / Recommendations** | Next Steps: - Apply adaboosting to the experiment  - Apply hypertuning parameters. |