# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Roger Yu |
| **Project Name** | MDSI ADSI Assignment 1 Part C |
| **Date** | 2020-02-21 |
| **Deliverables** | yu\_roger-10906675-week3\_early\_stopping.ipynb  xgb\_top\_8\_features\_early\_stopping.joblib  <https://github.com/roger-yu-ds/assignment_1/tree/roger> |

|  |  |
| --- | --- |
| 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** | Predict the probability of a rookie NBA player, given certain traits, having a career in the NBA that is greater than 5 years. |
| **1.b. Hypothesis** | Predicting more negative classes Reducing the scale\_pos\_weight from the default of 1 would make the model predict more negative classes and improve the validation/test AUC scores. Further reduction of overfitting  * Early stopping is used in all experiments * Reduction in the number of features  Adversarial validation Previous experiments indicated that the distributions of the training and test sets are different, this was hypothesised because the test AUC proved to be much lower than the validation AUC.  If the training set and test sets were indeed different to a significant extent, then a classification algorithm would be able to tell them apart, i.e. if the observation came from the training set or from the test set. |
| **1.c. Experiment Objective** | * + - 1. Improve the validation AUC beyond 0.70       2. Reduce overfitting as indicated by the difference between the training and validation AUC, i.e. 0.13       3. Determine if there are significant differences in the training and test sets |

|  |  |
| --- | --- |
| 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** | Original classification No notable preparation steps in addition to those in the previous reports. Adversarial validation  * + - 1. Columns are dropped.  |  |  |  | | --- | --- | --- | | Column | Training Set | Test Set | | Id Old | drop | drop | | Id | drop | drop | | TARGET\_5Yrs | drop | na |  * + - 1. A new column “dataset” is added, with values of “train” and “test”, for the training and test sets respectively. This column is the target column of the adversarial validation       2. The data is sampled without replacement to produce an adversarial training set and an adversarial testing set; the sizes of the sets are the same as that of the original training and test sets. |
| **2.b. Feature Engineering** | NA |
| **2.c. Modelling** | scale\_pos\_weight ([10906675\_xgb\_es\_spw.joblib](https://github.com/roger-yu-ds/assignment_1/blob/roger/models/10906675_xgb_es_spw.joblib))  The XGB classifier parameters were the same as the best model from the previous experiment, from the model [randomised\_xgb.joblib](https://github.com/roger-yu-ds/assignment_1/blob/roger/models/randomised_xgb.joblib) The investigation starts with a coarse search space in the range of (0, 1) with a step size of 0.1, which resulted in an optimal weight of 0.5. After which a finer grid is searched with the range (0.4, 0.6) with a step size of 0.01, the results on the validation set are shown in the table below. Further reduction of overfitting A GridSearchCV was run over the range (1, 7) of numbers of PCA components; 8 components was the random search previous found to be optimal. The base estimator is the model in the previous section ([10906675\_xgb\_es\_spw.joblib](https://github.com/roger-yu-ds/assignment_1/blob/roger/models/10906675_xgb_es_spw.joblib)). Adversarial validation |

|  |  |
| --- | --- |
| 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** | scale\_pos\_weight The finer grid search achieved a scale\_pos\_weight value of 0.55, which has slightly higher validation results, see table below.   |  |  |  |  | | --- | --- | --- | --- | | scale\_pos\_weight | AUC | Class 0 f1-score | Class 1 f1-score | | 1 | 0.6989 | 0.01 | 0.91 | | 0.55 | 0.6991 | 0.04 | 0.91 | | 0.5 | 0.6963 | 0.04 | 0.91 |   Even with the scale\_pos\_weight value of 0.55, reduced by almost half, there are still very few predictions of the negative class, as shown by the low f1-score, and the confusion matrix of the validation set below.   |  |  |  | | --- | --- | --- | |  | Predicted 0 | Predicted 1 | | Actual 0 | 5 | 258 | | Actual 1 | 5 | 1332 |   This model ([10906675\_xgb\_es\_spw.joblib](https://github.com/roger-yu-ds/assignment_1/blob/roger/models/10906675_xgb_es_spw.joblib)) resulted in the highest test AUC of 0.69390. Further reduction of overfitting As expected, the fewer the components the less overfitting observed, with one component resulting in the smallest difference between the training and validation AUC, but also with the lowest validation AUC; the table below shows the   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **n\_components** | **train\_auc** | **val\_auc** | **diff\_auc** | **ratio\_auc** | | 1 | 0.685 | 0.685 | 0.001 | 0.001 | | 2 | 0.706 | 0.681 | 0.025 | 0.035 | | 3 | 0.706 | 0.682 | 0.024 | 0.035 | | 4 | 0.719 | 0.691 | 0.029 | 0.040 | | 5 | 0.708 | 0.678 | 0.031 | 0.043 | | 6 | 0.744 | 0.683 | 0.061 | 0.082 | | 7 | 0.770 | 0.694 | 0.076 | 0.098 |   The class 0 and 1 f1-scores are 0.08 and 0.91; a doubling of the class 0 f1-score compared to the previous model. The confusion matrix below shows an increase in the predictions of the negative class, although the false negative has also increased.   |  |  |  | | --- | --- | --- | |  | Predicted 0 | Predicted 1 | | Actual 0 | 12 | 251 | | Actual 1 | 17 | 1320 |  Adversarial Validation Using an XGB Classification algorithm, the AUC on the adversarial test set was close enough to 0.5 to conclude that there wasn’t any difference in the data sets, the ROC curve below shows that the observed curve (orange) is just as skilful as an algorithm predicting the mode (dashed line). |
| **3.b. Business Impact** | The issues from the previous experiment still persists, namely, misclassifications on the positive side.  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)  Sponsors and teams would like to support players that are likely to have a career greater than 5 years. There is likely to be a lot of upfront costs in training that are not recuperable if that player stops their career early, so the cost of a false positive is high. Too many of such cases could cause the company to shutdown, as initial investments are not recuperated.  On the other hand the cost of a false negative is foregone chance of hiring a well performing player for basketball teams or a player that produces a lot of marketing income for sponsoring companies. While this is unlikely to bankrupt companies/teams, they are also unlikely to overcome their competitors. |
| **3.c. Encountered Issues** | Bias towards the positive class The XGB Classifier seems to be heavily weighted to the prediction of the positive class and while reducing the pos\_scale\_weight would mitigate this effect, the drop in the validation AUC is such that the optimal value is 0.55, which also doesn’t increase the predictions of the negative class by a significant amount. This is due to the imbalanced data set; SMOTE and undersampling did not seem to have solved this problem |

|  |  |
| --- | --- |
| 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** |  |
| **4.b. Suggestions / Recommendations** | Data investigation Adversarial validation to discover differences in the training and test sets. Modelling Search through the space of scale\_pos\_weight for an XGB Classifier. Best practices Agree on a naming convention for different types of objects  Save objects often  Consider good function design |