# EXPERIMENT REPORT

| | |
|---|---|
| **Student Name** | Werner Schott |
| **Project Name** | NBA Rookie Career Prediction |
| **Date** | 14/02/2021 |
| **Deliverables** | Repo Link: https://github.com/yidaveding/adsi_g5_kaggle_nba |

---

| **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** | **Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?**<br><br>Assuming I work for a betting company (the business). The goal of the project is building a model that is fed NBA rookie player statistics and predicts which players will last at least 5 years in the league.<br><br>If the predictions are accurate, they will be useful to provide better betting odds on players and insight into the nature of what the NBA looks for in rookies. This information would be highly valuable to NBA teams as selecting rookies is a strategically important decision due to the limited selection (two picks per draft) and the difference this outcome could make to their team.<br><br>**Relative Metric Importance to Rookie Selection**<br><br>An important consideration regarding the use of outcomes by such a model to select rookies is that while the model will base its outcome on a select number of features (perhaps even all), the team looking to select will likely be looking to supplement and enhance a team. This means that some aspects of the player's game will likely be more important than others.<br><br>For example, if the team seeks a strong defensive player, metrics like steals, blocked shots, and rebounds, will likely be more important than field goals made. As such, it would be unwise to use such a model as the sole source of intelligence in selecting a rookie. |

<table>
<tr><td>1.b. Hypotheses</td><td>
Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it.<br><br>
Main:<br><br>
There is a significant statistical relationship between the rookie statistics and the target variable (career length longer than 5 years in the league).<br><br>
Experiment 1:<br><br>
Last week I included all features in the model. I assumed that a better knowledge of the variables (a business-context knowledge) would help me to improve the subset of variables I use as features. This will include new features.<br><br>
Hypothesis 1: Feature selection/reduction will improve XG Boost model AUC by more than 3%.<br><br>
Last week’s models had no hyperparameter tuning. Empirically, tuning parameters generally yields better results.<br><br>
Experiment 2:<br><br>
Hypothesis 2: The tuning of hyperparameters will improve AUC on XG Boost model by more than 3% AUC.<br><br>
Experiment 3:<br><br>
The first step here will be to get a realistic split from statistics. What has been the past trend? This will provide a good place to start adjusting this split.<br><br>
Hypothesis 3: The balancing of the training target variable will improve AUC on XG Boost model by more than 3% AUC.
</td></tr>
<tr><td>1.c. Experiment Objective</td><td>
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.<br><br>
Experiment 1:<br><br>
<u>Objective</u>:<br><br>
Feature Reduction: Use analysis of the variables and feature importance XG Boost model function to reduce variables. Compare score with baseline model, including full set.<br><br>
Feature Supplementation: Add two calculated variables TOV% (Turnover Percentage) and PPG (Points per game) and compare score to baseline. Check feature importance.<br><br>
<u>Expected</u>:<br>
XG Boost model AUC by more than 3%.<br><br>
<u>Scenarios</u>:<br>
The following scenarios are possible and would show the efficacy of the feature selection.<br>
1. Score reduces from baseline.<br>
2. Score stays the same or similar (within 1% of baseline).
</td></tr>
</table>

3. Score improves <1%<3%.
4. Score improves >3%.
5. Score improves significantly above 3% - e.g. >10%.

**Experiment 2:**

<u>Objective</u>:
Tune hyperparameters to improve AUC.

<u>Expected</u>:
XG Boost model AUC by more than 3%.

<u>Scenarios</u>:
The following scenarios are possible and would show the efficacy of the hyperparameter tuning.

1. Score reduces from baseline.
2. Score stays the same or similar (within 1% of baseline).
3. Score improves <1%<3%.
4. Score improves >3%.
5. Score improves significantly above 3% - e.g. >10%.

**Experiment 3:**

<u>Objective</u>:
Balance train target variable to improve AUC.

<u>Expected</u>:
XG Boost model AUC by more than 3%.

<u>Scenarios</u>:
The following scenarios are possible and would show the efficacy of balancing the training target variable.

1. Score reduces from baseline.
2. Score stays the same or similar (within 1% of baseline).
3. Score improves <1%<3%.
4. Score improves >3%.

Score improves significantly above 3% - e.g. >10%.

## 2. EXPERIMENT DETAILS

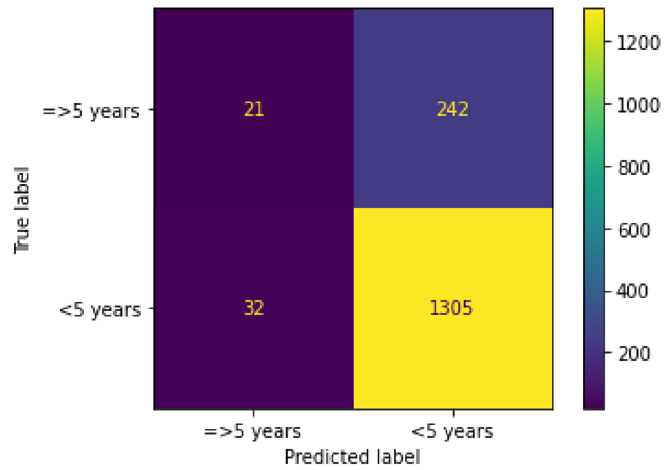
Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

| | |
|---|---|
| **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.**<br><br>There are no radical changes from last week's approach here. The treatment was to remove variables useless to the modelling (id variables) and to scale variable values to a similar scale. To this end, standardisation is applied, a common requirement for many scikit-learn machine learning estimators. If individual features are not standard normally distributed data, the model may produce suboptimal outputs. Here the standardisation is scaling features to lie between a given minimum and maximum value. I also checked for null values, but here are none.<br><br>The only changes will take part as experiment 3 applies a balancing technique to balance the training target variable, dimensionality reduction, and the introduction of two calculated features: TOV% and PPG.<br><br>As there is no categorical data, there was no need to on-hot-encode these variables. Also, there was no need to exclude variables because they have useless values e.g. a single or few values not adding to the model. There is also no incomplete/bad quality data so there is no need to deal with this. For instance, some variables can have multiple versions of the value e.g. 'Australeea', 'Austruliah', 'Aus', AU, etc. Also spaces or rubbish characters with numeric variables e.g. '1 1' instead of 11, or '45^.'<br><br>Also, no missing data. If there were, these values would be set to 0 to optimise for XG Boost model processing.<br><br>Finally a check was made to ensure the target variables is only composed of the expected binary outputs 1 and 2, which it is. |
| **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.**<br><br><u>Feature Education</u><br><br>Not understanding much about NBA stats or basketball in general, I created a matrix with descriptions to help me assess and understand the research. I followed the below steps:<br><br>1. Collated a glossary with definitions and descriptions of the abbreviated metrics. I references several sources as not all had definitions I understood.<br>2. Reviewed the definitions/descriptions to ensure understanding. Where the description did not make total sense, I supplemented with one that did. Sometimes the definitions included other unknown definitions.<br>3. I wanted to try some sort of ranking, so I searched for some contextual importance ranking reference source i.e. how are these ranked in importance to player value – perhaps a key driver for rookie NBA survival. I found the answer to this is a highly dynamic, speculative, biased, and contextualized, so I gave this up.<br>4. Searched for metrics not included in the data set but that may be calculated and perhaps offer better value than its constituents. The two I came across were TOV% and PPG, which I will include as part of the feature engineering |

| | |
|---|---|
| | experiment.<br><br><u>Feature Selection</u><br><br>**Qualitative Reduction**<br><br>1. The correlation matrix shows multicollinearity to be prevalent in the dataset. Unfortunately, it was difficult to find a definitive answer as to the degree this affects tree-based models. Certainly less than logistic regression. That said, reducing could begin by excluding metric constituents e.g. keep 'Field Goal Percentage' (FG%), exclude 'Field Goals Attempted' and 'Field Goals Made', both of which are used to calculate FG%. Of course, not having intimate knowledge of the game, I acknowledge that the constituent may be of more use than their function, but this is something I will try in establishing a dimensional reduction criteria.<br><br>2. Also as mentioned above, I will be adding two calculated features (TOV% and PPG).<br><br>**Quantitative Reduction**<br><br>1. Reduce some highly correlated variables that may be represented in other variables.<br>2. Examine 'feature importance' output and try excluding some variables with lower importance.<br><br>**Qualitative Addition**<br><br>1. Add two calculated measures: TOV% and PPG. |
| **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 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.**<br><br>**Model Used**<br><br>As explained above, I wanted to stick to the same model as last week XG Boost to compare the performance of the baseline model to a model whose features have been compared/reduced, whose hyperparameters have been tuned, and whose target variable has been balanced.<br><br>**Hyperparameter Tuning**<br><br><u>Qualitative/Intuitive</u><br><br>The first thing I tried was to read about the hyperparameters and how they apply to the data set. I mainly referenced the below article, which seems respected by practitioners and quoted as a reliable source.<br><br>[https://towardsdatascience.com/fine-tuning-xgboost-in-python-like-a-boss-b4543ed8b1e](https://towardsdatascience.com/fine-tuning-xgboost-in-python-like-a-boss-b4543ed8b1e)<br><br>Doing so increased the model's AUC score slightly by 0.32%. Not a great improvement |

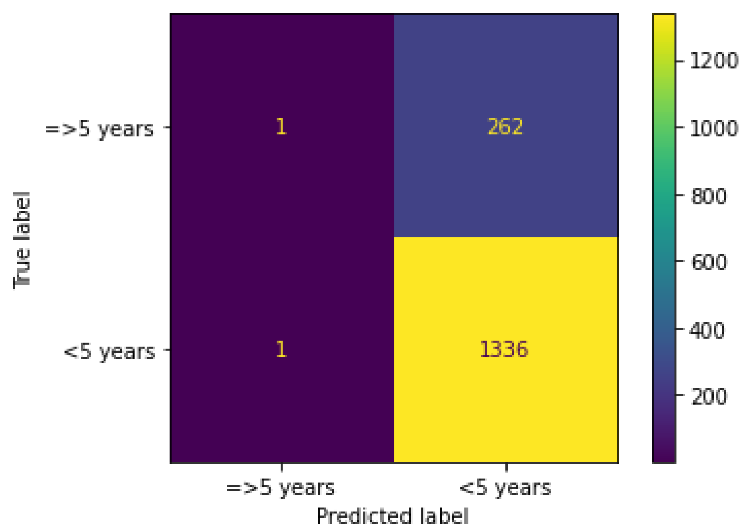
| | |
|---|---|
| | but a start.<br><br><u>Quantitative</u><br>Used hyperopt Trials to tune parameters and applied these to a new model.  This was done by defining a search space and fitting the model with values between the space ranges, by defined increments. |

| 3. 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** | **Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.**<br><br>**Experiment 1: Feature Reduction and Supplementation**<br><br><u>Remove Variables that are Constituents of Other Variables</u><br>The first model in this experiment removed features that were constituents of another feature e.g. a function. The AUC was 0.716, slightly lower than the baseline model.<br><br><u>Remove Variables that are Low in Importance</u><br>A feature importance graph was generated and the three features with the lowest importance were dropped. This lowered the AUC score more to 0.708.<br><br><u>Add Calculated Variables</u><br>A test was then made to run a model including two calculated variables: TOV% and PPG. The score was lower than baseline with 0.701.<br><br>**Conclusion:** The conclusion drawn is that the optimal set of features is the one that came in the dataset. Any taking from or adding to these will produce a sub-optimal result. **Hypothesis failed as AUC was not able to be increased by at least 3% following this method**.<br><br>**Experiment 2: Hyperparameter Tuning**<br><br>The result with hyperparameter tuning was 0.7256. Using hyperopt, hyperparameters were tuned and applied to a model. This scored an AUC score of 0.716.<br><br>**Conclusion:** The conclusion drawn is for whatever reason, understanding the hyperparameters in their context to the data and applying best judgement produced, in this case, a better outcome. **Hypothesis failed as AUC was not able to be increased by at least 3% following this method.**<br><br>**Experiment 3: Addressing Unbalanced Training Target Variable**<br><br>When a confusion matrix was generated on the baseline model, TP had a fairly poor outcome (8.7% TP to FP). Looking at the FN:TN, which has performed inversely (97.5% TN:FN), I assumed this is an issue that may be addressed by balancing the data. |




However, when I addressed this by setting ‘scale_pos_weight’ to 5 (as the ratio is 1:5 and this value is the inverse of the class distribution), the score was also inferior (0.712). The confusion matrix left me confused as while both the TP and AUC decreased, although TN have increased.

I also found it surprising that the balanced data outputs obtained a better score on Kaggle than the non-balanced.




**Conclusion:** The conclusion drawn is for whatever reason, understanding the hyperparameters in their context to the data and applying best judgement produced, in this case, a better outcome. **Hypothesis failed as AUC was not able to be increased by at least 3% following this method.**

## 3.b. Business Impact

**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).**

Like last week, as the results are inconclusive, using these models would yield an inaccurate estimation of which rookies would stay on, rendering the model useless.

| | |
|---|---|
| **3.c. Encountered Issues** | **List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Also highlight the issues that may have to be dealt with in future experiments.**<br><br>There were a lot of dependency issues to solve, which were mostly overcome by referencing sources on the internet such as stackoverflow. My main issue with future experiments is how to identify what drives, for this data, a better outcome. |

---

| **4. 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** | **Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.**<br><br>The main insight is that after all that experimentation, I'm no better off. This is very surprising as I entered the week positive something would change. I will now follow other methods, models, and practices to see if I can improve the outcome. Also, I would like to understand why worse-performing models in Colab cab be better performing in Kaggle. I understand Kaggle scores both the train and test data sets, and that the data I have does not have test target variables, but they can be large differences. Knowing how to create better scores in Kaggle will also be a priority for the next set of experiments. |
| **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.**<br><br>Up until now my method has been to set a hypothesis, prove/disprove the hypothesis, and document the findings. Unfortunately, this has proven slow and unfruitful.<br><br>This coming week I will try and get to my objectives quicker, focusing on a 'quick win/fail' approach, where I will explore without a specific hypothesis (the only one being that if I seek enough, I will find answers to my questions), until I find something to help me meet my objectives, iterating through these somethings until they make a collective difference.<br><br>I will then declare hypotheses, code them to show the proof, and document them. While using this method it is difficult to define the experiments I will be following, the overall approach will be to see what works in working with this type of data, focusing on improvements that I know will work because I have tested them along the way. |