EXPERIMENT REPORT

Student Name	Vu Ngoc Anh Le - 14195743
Project Name	Kaggle Competition – NBA Draft
Date	25 Aug 2023
Deliverables	**Jupyter Notebook:** le_vungocanh_14195743_week2_baseline le_vungocanh_14195743_week2_svm **Github Repo:** https://github.com/anhlevn149/adv_mla_at1

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? The NBA draft is an annual event in which teams select players from their American colleges as well as international professional leagues to join their rosters. This project aims to build a machine learning model which can generate the probability of being drafted for each player. To do this, a number of machine learning models will be used. Ultimately, the best model with the most accurate probability will be selected. The results will be useful for sport commentators and NBA fans as they can have predictions of which players that will be drafted for the next season.
1.b. Hypothesis	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. I want to test a hypothesis if **Support Vector Machine Classifier** model is able to predict the probability of a player being drafted more accurately than the Logistic Regression in Experiment 1. In Experiment 1, although the AUROC scores are high (> 0.97), there is overfitting issue. In this Experiment 2, feature selection will be employed to eliminate irrelevant features. SMOTE oversampling technique will also be used to tackle the imbalanced dataset. In addition, to improve the robustness of the coding file, I will apply custom packages built on poetry and test-pypi.
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. This Experiment is expected to perform better than the Logistic Regression Classifier model in Experiment 1 with higher AUROC scores on training and validation datasets.

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
	
	1) Remove columns that are irrelevant to the prediction task, including 'ht', 'type', 'player_id'
	2) Column 'yr': Replace invalid values with NaN values
	3) Column 'num': Replace any non-numeric values with NaN values
	4) Remove columns with missing values containing > 50% of the total values, including 'Rec_Rank', 'dunks_ratio', 'pick'
	5) Apply LabelEncoder technique to categorical columns, including 'team', 'conf', 'yr'
	6) Fill missing values in the dataframe with their mean values
	7) Apply SMOTE oversampling technique to address the imbalanced dataset
	8) Feature elimination with correlation matrix to select the most relevant features of the dataset
	9) Scale data:
	   ▪ Apply StandardScaler method on X_train, X_val, X_test
	10) Split the training dataset:
	   ▪ Split X_data and y_data → X_train, X_val, y_train, y_val with test size of 20%

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 1) Below columns were removed from the dataset: • Column 'ht': As per the data dictionary, this column contains information of the player's height. However, the data is displayed in date format. It is not feasible to transform this type of data to the height, so we decide to remove this column. • Column 'type': There is only one value in this column, which is 'all'. Hence there is no need to keep this column. • Column 'player_id': This column does not play any role in the prediction task. • Columns 'Rec_Rank', 'dunks_ratio', 'pick': These columns contain > 50% of the total values 2) Apply LabelEncoder technique to categorical columns, including 'team', 'conf', 'yr' 3) Feature elimination with correlation matrix to select the most relevant features of the dataset → From 62 features in the original training dataset, after performing the feature selection, there are only 42 features left.

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
	I used the Support Vector Machine model to train the dataset. To improve the model performance, I also conducted manual hyperparameters tuning on class_weight and C to find the best hyperparameters.

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.
	1) **Baseline model:**
	AUROC score on training dataset is 0.5
	2) **Support Vector Machine Model without Default Hyperparameters:**
	AUROC score on training dataset is 0.9976931291723836
	AUROC score on validation dataset is 0.9972850000861887
	3) **Support Vector Machine Model with Hyperparameters Tuning class_weight='balanced':**
	AUROC score on training dataset is 0.9976931291723836
	AUROC score on validation dataset is 0.9972850000861887
	4) **Support Vector Machine Model with Hyperparameters Tuning class_weight='balanced', C=0.5:**
	AUROC score on training dataset is 0.9976931291723836
	AUROC score on validation dataset is 0.9972850000861887
	5) **Support Vector Machine Model with Hyperparameters Tuning class_weight='balanced', C=1:**
	AUROC score on training dataset is 0.9976931291723836
	AUROC score on training dataset is 0.9972850000861887
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)
	Support Vector Machine model by far was the most effective model in generating accurate predictions with high AUROC scores (more than 0.997) with minimal overfitting compared to the previous Logistic Regression model.

3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. 1) **Long execution time** when training the models due to large size of the dataset. The step of finding feature importance took more than half an hour to complete. 2) With this Experiment, I tried to **apply customised functions** to improve the code execution robustness. However, in this Experiment, some functions were not able to be employed, including those to save and load datasets, split dataset, assess baseline and model performance. The reason is because of the file structure.

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. With this experiment with Support Vector Machine model, the expected performance level was achieved. When applying the hyperparameter tuning of class_weight and C, the prediction accuracy remains unchanged compared to the default hyperparameters.
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. Next steps: ▪ Modify the custom packages so that they can be applied to this project ▪ Try another model to improve the prediction