

EXPERIMENT REPORT

Student Name	Sudarat Sukjaroen
Project Name	Experiment on Logistic Regression model
Date	17 August 2023
Deliverables	sukjaroen_sudarat-24667255-week1_firsttrial.ipynb
Github repository	https://github.com/sudarat-pom/AdvanceML_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

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. Moving to the NBA league is a big deal for any basketball player.

Sport commentators and fans are very excited to follow the careers of college players and guess who will be drafted by an NBA team.

Data science is tasked to build a model that will predict if a college basketball player will be drafted to join the NBA league based on his statistics for the current season.

1.b. Hypothesis

Find the best solution with the best probability value within 4 weeks, submit 1 version per week and finalise the best one on the last week. In the first trial, I start with the Logistic Regression model.

Evaluate the model regarding the accuracy, precision, recall, and F1 score on the player statistics data and split the train and test data set (80:20). Predict with test data set to calculate probability value for each player_id and submit predictions file in Kaggle to check the score.

1.c. Experiment Objective

The project expects to apply machine learning techniques to calculate the best probability value for each player_id and submit a prediction file in Kaggle to check the score. The project expectation is to improve the score every week.

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

Data exploration

1. Read train and test data from a CSV file.
2. Display data information.
3. Display data description.
4. Display the data top 10 rows.
5. Count distinct value of drafted column in Train data set and found
Number of distinct values in drafted: 2
0.0 = 55,555
1.0 = 536

Data cleansing and data preparation

1. Check the number of records and columns.
Train data set: number of record = 56,091, Number of column = 64
Test data set: number of record = 4,970, Number of column = 63
2. Identify, remove duplicate records, and not find duplicates in a data frame.
3. Check the Null value and found as following column names, replace them with 0 and not find the Null target value.

Train data

Columns with null values in Train data set:

yr	274
ht	80
num	4669
Rec_Rank	39055
ast_tov	4190
rimmade	6081
rimmade_rimmiss	6081
midmade	6081
midmade_midmiss	6081
rim_ratio	9464
mid_ratio	9688
dunksmade	6081
dunksmiss_dunksmade	6081
dunks_ratio	30793
pick	54705
drtg	44
adrtg	44
dporpag	44
stops	44
bpm	44
obpm	44
dbpm	44
gbpm	44
mp	38
ogbpm	44
dgbpm	44
oreb	38
dreb	38
treb	38
ast	38
stl	38
blk	38
pts	38
dtype:	int64

Test data

Columns with null values in Train data set:

ht	6
num	88
Rec_Rank	3536
ast_tov	537
rimmade	248
rimmade_rimmiss	248
midmade	248
midmade_midmiss	248
rim_ratio	646
mid_ratio	697
dunksmade	248
dunksmiss_dunksmade	248
dunks_ratio	2717
pick	4921
drtg	1
adrtg	1
dporpag	1
stops	1
bpm	1
obpm	1
dbpm	1
gbpm	1
ogbpm	1
dgbpm	1

dtype: int64

2.b. Feature Engineering

Mapping columns from text to number

1. Team to team_number
2. Conf to conf_number
3. Yr to yr_number
4. Ht to ht_number
5. Num to num_number
6. Player_id to player_number

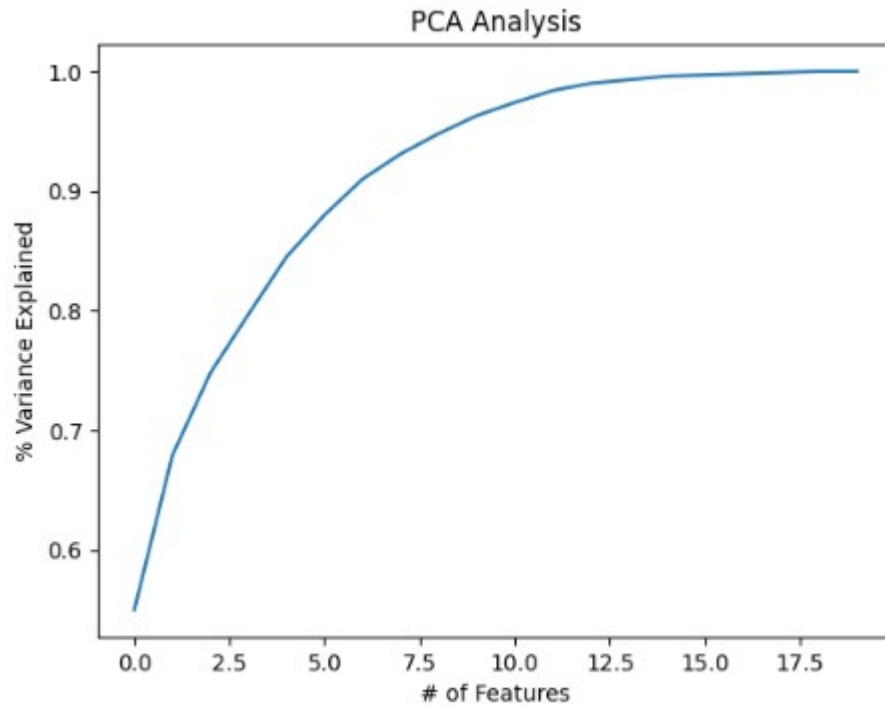
Calculate Information Gain (IG)

Calculate Information Gain from 69 features and find the top 20 influences as follows, but for the first trial, I decide to use all features to predict.

pick: 0.03795213649565099
dporpag: 0.023214999631525624
porpag: 0.022102364706616195
gbpm: 0.02139306348854475
bpm: 0.019259731075576658
stops: 0.018922022782925096
ogbpm: 0.01845620142715354
adjoe: 0.01842138862800269
Rec_Rank: 0.016650354956799007
twoPM: 0.016645421479337896
pts: 0.016634762050635632
twoPA: 0.01573814733641965
FTM: 0.01536598217175833
obpm: 0.015249632005036817
FTA: 0.014411160875968498
dreb: 0.013238482990499456
mp: 0.012233762570631468
team_number: 0.011951887246710147
rimmade: 0.011627495017502376
midmade_midmiss: 0.011453766614785255

Calculate Principal Component Analysis (PCA)

Calculate PCA 80%, 90%, and 100%, but the number of components for 80% and 90% is almost the same as 100%. Then I decide to use 100% to experiment.



[0.55 0.679 0.748 0.797 0.845 0.88 0.91 0.931 0.948 0.963 0.974 0.984
0.99 0.993 0.996 0.997 0.998 0.999 1. 1.]

2.c. Modelling

The first experiment is the Logistic Regression model.

1. Find the best set of hyperparameters

Split data percentage : 80:20

Define the set of hyperparameters to find which value is best from this list.

```
param_grid =  
{  
    'penalty': ['l1', 'l2'],  
    'C': [0.01, 0.1, 1, 10],  
}
```

2. The best hyperparameters are these values.

Best Hyperparameters: {'C': 1, 'penalty': 'l2'}

3. The metrics values are as follows:

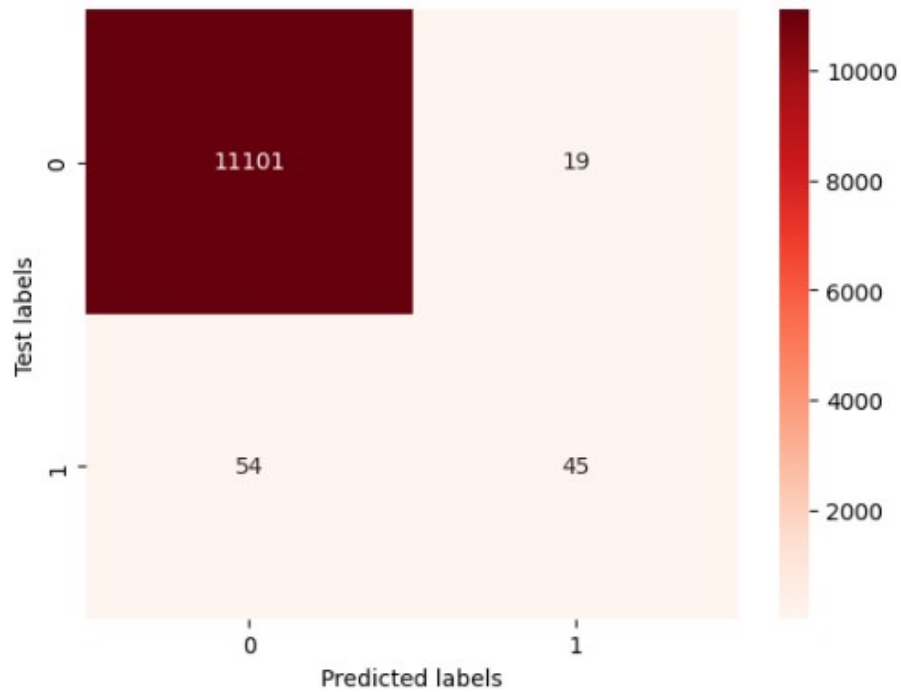
Accuracy: 0.9934931812104466

Precision: 0.703125

Recall: 0.45454545454545453

F1 Score: 0.5521472392638036

Confusion Matrix:



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

Execution time

No	Model name	Execution time Find best parameters	Execution time Best parameter
1	Logistic Regression	3 minutes	5 seconds

Score after submit prediction = 0.99365 (17 Aug 23 11.30pm)

Public Private

This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

#	Team	Members	Score	Entries	Last
1	Jeremy Herbert		0.99908	1	3h
2	Shiv07 Sharma		0.99882	2	3h
3	Sudarat Sukjaroen		0.99365	1	6m



Your First Entry!
Welcome to the leaderboard!

3.b. Business Impact	The impact in the NBA industry is to predict the possibility of the basketball player who has a high probability of joining the professional basketball team from their statistics value.
3.c. Encountered Issues	<p>The problem I found was based on running experiments Google Colab environment.</p> <p>Problem</p> <p>1. The execution time to find the best set of hyperparameters is very long. It is hard to test many times and many sets of hyperparameter possibilities within a time limit.</p> <p>Solution</p> <p><u>Short term solution</u></p> <p>1. Reduce the set of hyperparameters to only necessary values and cover a wide range of values.</p> <p>2. Reduce the split data percentage from 80:20 and 70:30 to only 80:20 based on the metric values.</p> <p><u>Long term solution</u></p> <p>1. Buy more resources from the Google Colab environment.</p> <p>2. I have more knowledge and experience in tuning the performance of models and execution times.</p>

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	<p>In the first trial, I started with standard techniques.</p> <ol style="list-style-type: none"> 1. Select from Support Vector Machine and Logistic Regression Model. I selected Logistic Regression Model as the first model because the execution time is faster than SVM. 2. Remove duplicate records in Train and Test data set. 3. Replace the Null value with 0 in Train and Test data set. 4. Map columns from text to number. 5. Find the best set of hyperparameters.
4.b. Suggestions / Recommendations	<p>In the following experiments, I plan to apply more machine-learning techniques.</p> <ol style="list-style-type: none"> 1. Run other classification models. 2. Replace the Null value with Means, Median, Mode etc. 3. Remove outlier values. 4. Replace missing value with Means, Median, Mode etc. 5. Adjust the density of the ratio of drafted = 0 and 1. 6. Apply Information Gain (IG). 7. Apply Principal Component Analysis (PCA). <p>All techniques will be considered from Accuracy, Precision, Recall and F1 Score values before run prediction.</p>