EXPERIMENT REPORT

Student Name	Priyanka Srinivasa
Project Name	MDSI ADSI Assignment 1
Date	07-02-2020
Deliverables	srinivasa_priyanka_13684182_week1_log_regression_feature_elimination.ipynb **model**: log_regression_feature_elimination https://github.com/roger-yu-ds/assignment_1/tree/priya

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.

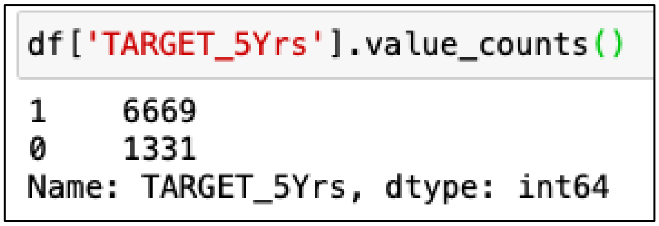
1.a. Business Objective	The goal of the project is to predict if a rookie player in the National Basketball Association (NBA) will remain at least 5 years in the league. Measuring the success of rookie players and learn about the future possibilities on how long these young talents will last is an important question in sports analytics as they help the business side of sports to secure a competitive edge.
1.b. Hypothesis	With the help of statistics, and Machine Learning algorithms we aim to predict the career length of NBA rookie players. The question we want to answer is, • If a rookie player will last at least 5 years in the league based on his performance statistics. By predicting the career performance of a rookie player with the help of their performance statistics, the team management can make better decisions which improves their business by providing the organization an opportunity to win a championship.
1.c. Experiment Objective	The outcome expected from this experiment is to find out if the end result agrees or differs from the hypothesis. The expected goal from the first experiment of this project is to use a classification algorithm, run multiple trials to collect data, and interpret results to accurately predict how many of the rookie players will remain in the NBA league for at least 5 years. The steps undertaken during the course of this experiment are, • Train a simple logistic regression with all variables – to get a baseline model. • Perform logistic regression using SMOTE – to account for class imbalance. • Perform logistic regression by implementing PCA – to train the model by gathering only the important features. • Perform logistic regression by Feature Elimination – to get rid of features that do not aid in improving model performance.

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

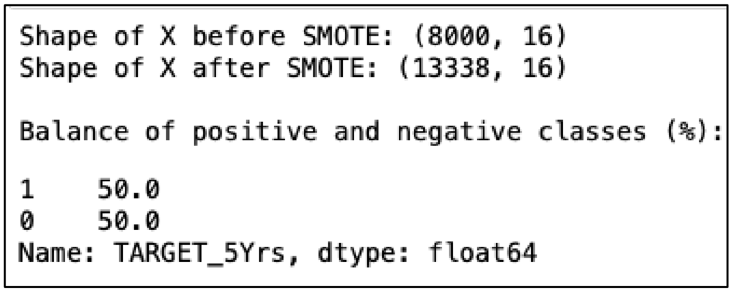
- **Removed Id_old / Id** – these columns were removed as it captures the uniqueness of the players.
- **Missing / NULL values** – the data was examined for missing / NULL values to ensure the completeness of the data.
- **Handling Duplicates** – there were no duplicate values present in the data.
- **Data Transformation** – all the numerical values were converted to a standard scale to make sure the data is in a standard normal distribution.
- **Class Imbalance** – Checked for class imbalance as it can influence the accuracy of the model.



```
df['TARGET_5Yrs'].value_counts()

1    6669
0    1331
Name: TARGET_5Yrs, dtype: int64
```

Figure 1

- **Target variable** – the 'Target_5YRS' feature was assigned to a new variable, and removed from the main dataset.
- **SMOTE** – Synthetic Minority Oversampling Technique was implemented to account for class imbalance.



```
Shape of X before SMOTE: (8000, 16)
Shape of X after SMOTE: (13338, 16)

Balance of positive and negative classes (%):

1    50.0
0    50.0
Name: TARGET_5Yrs, dtype: float64
```

Figure 2

Some of the steps that were not executed for this experiment were,

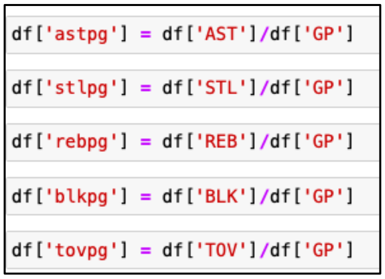
- **Principal Component Analysis** – PCA was performed initially to identify the most important features. However, this was not included in the final Logistic Regression model as the accuracy, and the AUC score reduced after performing PCA.
- **Normalization** – the data was not normalised for this experiment as it is planned to be considered it for the future experiments.

2.b. Feature Engineering

Some of the feature engineered variables created were,

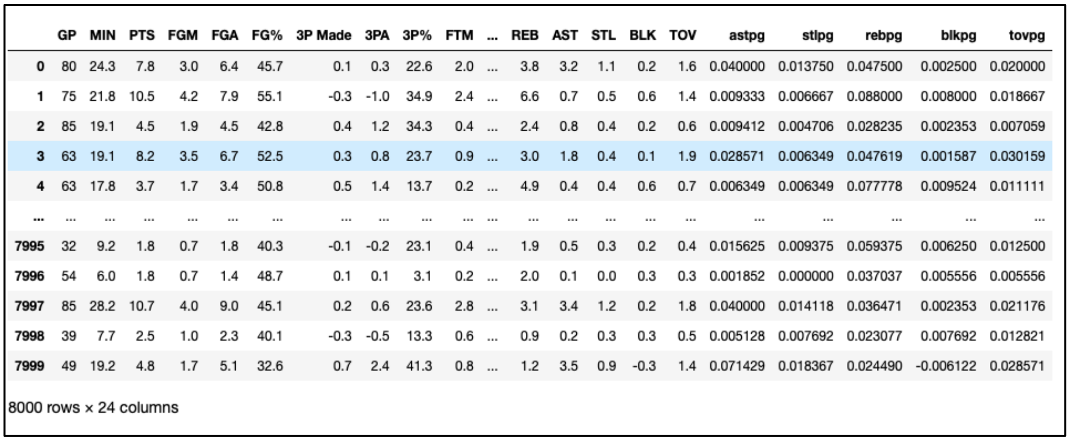
- astpg – assists per game.
- stlpg – steals per game.
- rebpg – rebounds per game.
- blkpg – blocks per game.
- tovpg – turnovers per game.

The dataset given had columns consisting of total ‘Assists’, ‘Steals’, ‘Rebounds’, ‘Blocks’, and ‘Turnovers’. An assumption was made that creating a dataset by calculating values for all these features per game would improve the performance of the model. However, upon discovering there was no high significance of the featured engineered variables, and they were not considered while building the model.



```
df['astpg'] = df['AST']/df['GP']

df['stlpg'] = df['STL']/df['GP']

df['rebpg'] = df['REB']/df['GP']

df['blkpg'] = df['BLK']/df['GP']

df['tovpg'] = df['TOV']/df['GP']
```

Figure 3



	GP	MIN	PTS	FGM	FGA	FG%	3P Made	3PA	3P%	FTM	...	REB	AST	STL	BLK	TOV	astpg	stlpg	rebpg	blkpg	tovpg
0	80	24.3	7.8	3.0	6.4	45.7	0.1	0.3	22.6	2.0	...	3.8	3.2	1.1	0.2	1.6	0.040000	0.013750	0.047500	0.002500	0.020000
1	75	21.8	10.5	4.2	7.9	55.1	-0.3	-1.0	34.9	2.4	...	6.6	0.7	0.5	0.6	1.4	0.009333	0.006667	0.088000	0.008000	0.018667
2	85	19.1	4.5	1.9	4.5	42.8	0.4	1.2	34.3	0.4	...	2.4	0.8	0.4	0.2	0.6	0.009412	0.004706	0.028235	0.002353	0.007059
3	63	19.1	8.2	3.5	6.7	52.5	0.3	0.8	23.7	0.9	...	3.0	1.8	0.4	0.1	1.9	0.028571	0.006349	0.047619	0.001587	0.030159
4	63	17.8	3.7	1.7	3.4	50.8	0.5	1.4	13.7	0.2	...	4.9	0.4	0.4	0.6	0.7	0.006349	0.006349	0.077778	0.009524	0.011111
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7995	32	9.2	1.8	0.7	1.8	40.3	-0.1	-0.2	23.1	0.4	...	1.9	0.5	0.3	0.2	0.4	0.015625	0.009375	0.059375	0.006250	0.012500
7996	54	6.0	1.8	0.7	1.4	48.7	0.1	0.1	3.1	0.2	...	2.0	0.1	0.0	0.3	0.3	0.001852	0.000000	0.037037	0.005556	0.005556
7997	85	28.2	10.7	4.0	9.0	45.1	0.2	0.6	23.6	2.8	...	3.1	3.4	1.2	0.2	1.8	0.040000	0.014118	0.036471	0.002353	0.021176
7998	39	7.7	2.5	1.0	2.3	40.1	-0.3	-0.5	13.3	0.6	...	0.9	0.2	0.3	0.3	0.5	0.005128	0.007692	0.023077	0.007692	0.012821
7999	49	19.2	4.8	1.7	5.1	32.6	0.7	2.4	41.3	0.8	...	1.2	3.5	0.9	-0.3	1.4	0.071429	0.018367	0.024490	-0.006122	0.028571

8000 rows × 24 columns

Figure 4

- Feature Elimination – a heatmap was plotted to identify the variables that are highly correlated, and aid to improve the results of the experiment. Variables - ‘FG%’, ‘3P%’, ‘FT%’ were removed as these features were less significant compared to other features.

2.c. Modelling

For week-1 experiment, I decided to use Logistic Regression model as the target variable was categorical (1 if career length >= 5 years, 0 otherwise), and since it had only 2 possible outcomes. The goal was to predict the impact of explanatory variables on the probability of a rookie player lasting in the league for 5 years based on the performance statistics. As the main objective of this experiment is to determine if an NBA rookie player will last in the league for at least 5 years, Logistic Regression model was suitable for this experiment as it works well with binary response variables.

The final model chosen for Week 1 experiment was Logistic Regression by implementing feature elimination of least significant variables. a confusion matrix was created to calculate Accuracy, Precision, Recall, F1 and AUC to help choose the best metric (Table.1).

Metric	**Value**
Accuracy	0.66
Precision	0.68
Recall	0.64
F1	0.66
AUC	0.72

Table 1

The models, and hyper parameter tuning techniques that will be considered for the future experiments,

- Random Forest Classifier.
- XGB classification algorithm.
- Decision Tree model.
- Support Vector Machine.

Hyper parameter tuning techniques

- Grid Search to select a grid of hyperparameter values, and return the best performing parameters by evaluating each of them.
- Random Search to evaluate the random samples of values in the grid.
- Cross Validation – to compare, and choose a suitable model to predict the career length of NBA rookie players.

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

AUROC (Area Under ROC) is the performance metric used to assess the model. The scores of the validation, and test set are proportional.

AUROC score for validation set	0.72707
AUROC score for test set on Kaggle	0.71034

Table 2

Underperforming observations

From fig,6, it was observed that the number false positives, and negatives are high. This can be due to the default threshold probability.

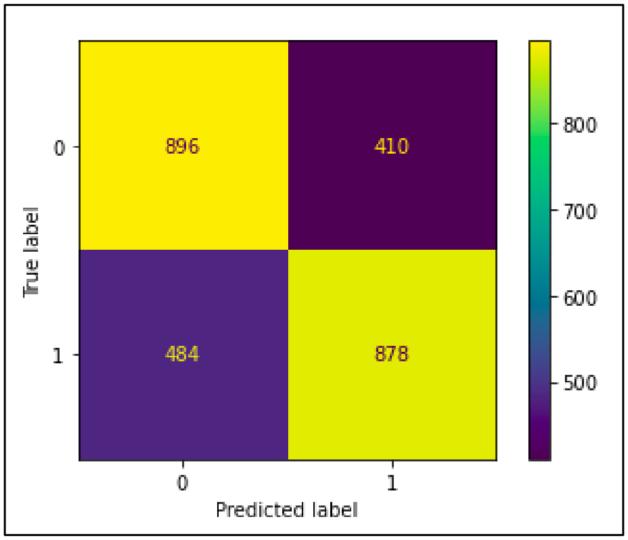


Figure 5

3.b. Business Impact

The results obtained from this experiment will be used to find out how well they will perform in the future, how many of the skilled players will be able to contribute to the success of their team, and gain more fan following as they grow in their career. Depending on the accuracy of the results, sports analysts and NBA franchises can focus on the players who are more capable of staying longer in the league and invest on improving the performance of those players. When the players perform well on the court it results in the team's victory which in turn results in increasing the popularity of the team.

However, inaccurate results provide an outcome which might not help the potential rookies players to have a successful future in the NBA league in comparison to other players. Since Basketball is a popularity game it is important that the results obtained have better accuracy to ensure the players, audience, and the sponsors have a better idea about the future of the game.

3.c. Encountered Issues

Issues	Solution
Imbalanced data	**Solved** : Used SMOTE to oversample the minority class, resulting in improved results.
Feature Selection	**Unsolved** : used PCA to get the most important features, but the results of the model performance was poor. Forward Feature selection, and backward feature elimination will be applied to gain a better understanding of the significant parameters.
Feature Elimination	**Solved** : correlation plot was used to get rid of the insignificant features. However, more attempts will be made in the future experiments to eliminate insignificant features.
Feature Engineering	**Unsolved** : a few feature engineered variables were created, but there is scope for creating better feature engineered variables. More experiments will be undertaken to create better variables.

Table 3

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

Outcome / Insights gained from the experiment

The logistic regression model performed better after eliminating less significant features. Using SMOTE to handle class imbalance provided better AUC score. In experiments where class imbalance was not taken into account, reduced the prediction accuracy of the model.

The logistic regression model performance can be improved by adding more featured variables and eliminating insignificant variables. Hyperparameter tuning is another option to be considered for future experiments to improve the ROC AUC score.

4.b. Suggestions / Recommendations

Next steps and experiments

- Research more on the NBA league to find out what factors actually affect the player's career length.
- I would like to explore other Machine learning models such as, Random Forest Classifier, XGB classification algorithm, and Support Vector Machine.
- Perform hyperparameter tuning and add more featured engineering variables to improve the ROC AUC score by 0.1.

If the experiment achieved the required outcome, the solution can be deployed into production using Docker, or TensorFlow framework.