EXPERIMENT REPORT

Student Name	Priyanka Srinivasa
Project Name	MDSI ADSI Assignment 1_Part B
Date	14-02-2021
Deliverables	srinivasa_priyanka_13684182_week2_SVM_SMOTE.ipynb model: Support_Vector_Machine_SMOTE 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. 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.b. Hypothesis	In Week_1 experiment, a Logistic Regression model was used to predict the career length of NBA rookie players. The AUROC score obtained from this experiment was, • AUROC score for validation dataset – 0.72707 • AUROC score for test set on Kaggle – 0.71034 In Week_2 experiment, a Support Vector Machine (SVM) model was used to check if it can outperform the Logistic Regression model and provide better results in predicting If a rookie player will last at least 5 years in the league based on his performance statistics.
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 reason for choosing SVM was, it attempts to find the best margin to separate the classes of the data using a hyperplane and in turn reduce the error on the data. In SVM, Hyperparameter tuning can be performed in order improve the results and obtain a better AUROC score.

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.
- **Class Imbalance** – Checked for class imbalance as it can influence the accuracy of the model.
- **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 and oversample the minority class.
- **Splitting dataset** – the training data was split into 80% training data and 20% validation dataset using sklearn's train_test_split function with using a 'random state' = 8.

2.b. Feature Engineering

StandardScaler – all the numerical values were converted to a standard scale using sklearn's 'StandardScaler' function to ensure the data is in a standard normal distribution. This was important since SVM maximizes the distance between the support vectors and the hyperplane, higher values tend to have a greater influence on the other features when calculating the best margin. Hence, converting the features to a standard scale ensures the slope of the linear decision boundary do not depend on the range of the variables but on the distribution instead.

2.c. Modelling

In week_2 experiment, Support Vector Machine (SVM) was used to determine if an NBA rookie will last in the league for at least 5 years. Since this is a classification problem, SVM model was considered suitable to predict the results. The reason being options to fine tune the model using different choices of kernel in SVM is more compared to Logistic Regression.

The results obtained from linear SVM model provided the best results on the training dataset (Table.1). The AUROC score obtained for Logistic Regression model (Week 1) on the training set was 0.72. The AUROC score on the training dataset using SVM was improved as compared to Logistic Regression model (AUROC – 0.72).

Metric	Value
Accuracy	0.66
Precision	0.68
Recall	0.62
F1	0.65
AUROC	0.73

Table 1

Hyperparameter tuning was performed using GridSearchCV. However, the results obtained after applying the best parameters provided lesser evaluation scores. Hence was not implemented further in this experiment. The reason for this will be further explored in the future.

3. EXPERIMENT RESULTS

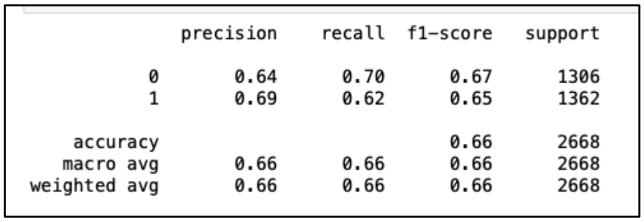
Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

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 were proportional. Although, the AUROC score obtained on Kaggle was lesser compared to the week 1 experiment using Logistic Regression model (0.71034).

AUROC for validation dataset – 0.73166
AUROC for test dataset on Kaggle – 0.70595

Classification report for linear SVM model.



```
              precision    recall  f1-score   support

           0       0.64      0.70      0.67      1306
           1       0.69      0.62      0.65      1362

    accuracy                           0.66      2668
   macro avg       0.66      0.66      0.66      2668
weighted avg       0.66      0.66      0.66      2668
```

Figure 1

3.b. Business Impact

Accuracy of the results helps the sports analysts and NBA franchises focus on the players who are more capable of staying longer in the league and invest on improving the performance of those players. Given the results of this experiment, the opportunities for new businesses might become lesser and lower the wealth of the economy if the model is predicting a potential rookie player to not last in the league for 5 years. This results in getting rid of an excellent player by looking at the prediction even before the player had a chance to perform and prove his potential.

3.c. Encountered Issues

Issues	Solution
Imbalanced data	**Solved**: Used SMOTE to oversample the minority class, resulting in improved results.
Features in different range	**Solved**: Converted all the features to a standard scale using StandardScaler function.
Feature Elimination	**Solved**: Removed Id and Id_old features in order to avoid capturing the uniqueness of the features.
Hyperparameter Tuning	**Unsolved**: The model performed poor after using GridSearchCV and applying the best parameters to predict the results. The reason behind this will be examined and try to find the best suitable parameters in the future experiment if decided to further explore SVM model.

Table 2

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 this experiment The SVM model performed better without hyperparameter tuning. The results of the model improved after converting all the features to a standard scale and then applying SMOTE to oversample the minority class providing a higher AUROC score compared to the AUROC score obtained after GridSearchCV and hyperparameter tuning. Since the prediction was not very accurate applying this for the business will result in poor prediction which will affect the revenue of the business and the sponsors.
4.b. Suggestions / Recommendations	Next Steps and experiments • I would like to try using polynomial Logistic Regression and eliminate more features which are of low significance. This approach might help improve the results of the model. • I would like to experiment with VotingClassifier model as a team member used this technique and the AUROC improved because of this approach. • If the experiment achieved the required outcome, the solution can be deployed into production using Docker, or TensorFlow framework.