Assignment 1

Kaggle Competition

Tahmidul Islam
Student ID: 24587139

7/10/2023

36120 - Advanced Machine Learning Application
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Executive Summary

This project had a core objective of developing a predictive model to assess the likelihood of college basketball players being selected in the NBA draft, using their current season's statistics as input. This endeavor held immense significance in the world of sports analytics, as it aimed to provide valuable insights for NBA teams, scouts, and players, potentially impacting career trajectories and team performance.

The central challenge addressed in this project was the prediction of NBA draft outcomes, a pivotal juncture in the lives of college basketball players. The NBA draft represents a high-stakes decision, and the project was undertaken in the context of data-driven decision-making in sports. It sought to leverage machine learning and data analysis techniques to transform raw statistical data into actionable information, aiding stakeholders in making more informed draft selections.

The project involved experimenting with Polynomial Logistic Regression Models and SVM Models which successfully predicted draft outcomes. Extensive hyperparameter tuning, primarily focusing on the regularization parameter (C), fine-tuned the models for optimal performance. These models demonstrated promising predictive capabilities, offering insights into the potential success of college players in the NBA draft.

The achieved outcomes signify a significant step in the realm of sports analytics, potentially assisting NBA teams, scouts, and players in the decision-making process during the draft. Future iterations of the project may explore advanced techniques like feature engineering, ensemble models, and data augmentation to further enhance predictive accuracy and contribute to the ongoing evolution of the NBA draft selection process.

2. Business Understanding

a. Business Use Cases

The primary application is assisting NBA teams in making well-informed draft selections. Predicting which college basketball players are likely to be drafted enables teams to optimize their picks, improving roster quality and team performance.

NBA scouts and recruiters benefit by streamlining their talent assessment efforts. Predictive models help identify prospects with higher draft potential, saving time and resources in the scouting process.

College basketball players can leverage their insights to make informed career decisions. Understanding their draft prospects allows them to focus on skill development and visibility enhancement, potentially impacting their professional careers positively.

b. Key Objectives

The project aims to factor in all the statistics and attributes of individual NBA basketball players in games that they have played before, and use the data to train a machine learning model that can accurately predict the probability of a player that he can be drafted.

The primary stakeholders are the Team Coaches who play a vital role in player development. They closely monitor player performance and make reald time decisions in games. Then there are sponsors and advertisers who associate their bands with successful players and teams. Player performance can influence endorsement deal and sponsorship contracts, making it essential for sponsors to monitor the players they are affiliated with. Finally, there is the entire NBA League that is concerned with brining entertainment value in the games and score high TV ratings to earn global popularity. Apart from their individual goals mentioned, all these 3 bodies are ultimately concerned with winning games and increased competition, which can only be done by selecting powerful players.

For coaches, the project identifies players with higher predicted draft chances, saving time and resources and ensuring a more efficient talent assessment process.

For sponsors, they can identify and target college players who are more likely to be drafted into the NBA. By sponsoring players with higher predicted draft potential, they increase the chances of aligning their brand with future NBA stars, which can lead to more significant exposure and brand recognition.

3. Data Understanding

- Provide insights into the dataset used for the project.

The dataset used for this project contains information on college basketball players and their current season's statistics, as well as whether or not they were drafted into the NBA. The dataset is instrumental in predicting the likelihood of a college player being selected in the NBA draft based on their performance metrics.

The limitation of this dataset is that because it has such a high number of features, we cannot create any pairplots to see an overview of which features is best correlated with the Target Variable ‘drafted’.

```
30  year                 56091 non-null  int64
31  type                 56091 non-null  object
32  Rec_Rank             17036 non-null  float64
33  ast_tov              51901 non-null  float64
34  rimmade              50010 non-null  float64
35  rimmade_rimmiss      50010 non-null  float64
36  midmade              50010 non-null  float64
37  midmade_midmiss      50010 non-null  float64
38  rim_ratio            46627 non-null  float64
39  mid_ratio            46403 non-null  float64
40  dunksmade            50010 non-null  float64
41  dunksmiss_dunksmade  50010 non-null  float64
42  dunks_ratio          25298 non-null  float64
43  pick                 1386 non-null   float64
44  drtg                 56047 non-null  float64
45  adrtg                56047 non-null  float64
46  dporpag              56047 non-null  float64
47  stops                56047 non-null  float64
48  bpm                  56047 non-null  float64
49  obpm                 56047 non-null  float64
```

In addition from the list here, we can see that many features have high amount of missing values which can impose a problem on our Machine Learning model. We should figure out a way to handle this problem. Further more, there are type ‘object’ for two features; they need to be removed or type_casted.

Finding mean, median, mode and outlier values for such a huge number of features is also very challenging for this dataset.

4. Data Preparation

- Describe the steps taken to prepare the data for modeling.

In the first step of the data, it was found that columns 'Rec_Rank', 'pick', and ‘'dunks_ratio' had excessive missing values. Imputing them with mean values would thus be a wrong approach for which they were dropped from the dataset.

Following this the nominal columns 'team', 'conf', 'ht', 'yr', 'type', and 'num' had too many unique values. Thus, trying to apply any form of encoding would leading to the creation of new columns (about 20 for each feature). Thus, they were dropped as well.

In the next step, we performed imputations on the remaining columns that had missing values. The missing values were replaced by the total mean value of that column.

Once done, the dataset was split into train, validation and test after which they were all standardized under a standard scaler. This ensured the variation among the columns were removed otherwise, the model would have underfitted.

Finally, because the dataset had a huge number of columns, we applied PCA for dimensionality reduction so that the model may not overfit. Regarding outliers, there were no steps taken as we chose models that were outlier-resistant. For imbalanced data, we assigned more weights to the class that was low in number and assigned low weights to the class that was high in number before training the model.

5. Modeling

The machine Learning algorithms used were polynomial Logistic Regression and Support Vector Machines.

Polynomial Logistic Regression was chosen because it can capture non-linear relationships by including polynomial terms of the predictor variables. It is especially suited for binary classification. No hyperparameter was tuned for this model except that the polynomial degree was changed. This provided a relative comparison to show which degree performed better.

Support Vector Machines was chosen because it is particularly effective when dealing with high-dimensional data where the number of features is large. In addition, it inherently provides regularization through the margin concept, which can help prevent overfitting, especially when dealing with high-dimensional data. SVM is also robust to outliers. The hyperparameter tuned was Regularization (C), and class_weight parameter to deal with overfitting and imbalance data.

a. Approach 1: Polynomial Logistic Regression

For the Logistic Regression model, degree of 2 was chosen first. The model used a quadratic hypothesis function to classify the categories provided. For pre-processing, PCA was applied which reduced the number of features to 15 so that the model didn't have to overfit.

At this stage of the experiment, no steps to handle imbalance data was done so that it could be set as a control for Approaches 2 and 3..

b. Approach 2: Support Vector Machine (class_weight = balanced)

In approach 2, SVM was used with absolute default parameters without changing anything. The model was run and evaluation was done.

In our second approach in Approach 2, the SVM model was run by setting class_weight to 'balance', which was used to deal with the imbalance data. Doing this assigned higher weight to the low-abundance class and assigned lower weights to the higher-abundant class. PCA was also done for this model.

c. Approach 3: Support Vector Machine (C = 0.5, 0.8)

In approach 3, SVM was used with class_weight = balanced and C = 0.5. This was an attempt to reduce overfitting.

In our second approach in Approach 3, we set class_weight = balanced and C = 0.8 to check if there's any change to the fit. No further preprocessing was done.

6. Evaluation

a. Evaluation Metrics

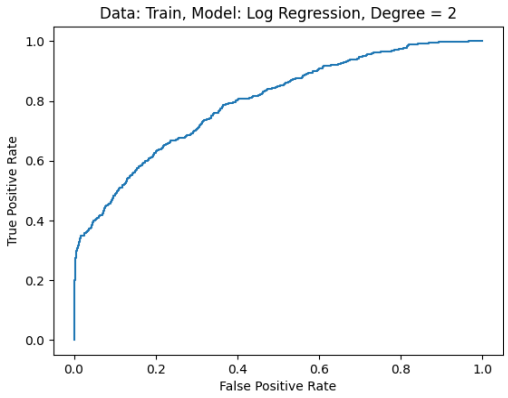
The evaluation metrics used was "Compute Area Under the Receiver Operating Characteristic Curve" (ROC AUC) Score. The AUROC score summarizes the ROC curve into a single number that describes the performance of a model.

The wisdom behind choosing this metric is that AUROC is robust when dealing with imbalanced datasets, where one class significantly outnumbers the other. This is particularly important in scenarios where the project aims to classify rare events or anomalies. By considering the entire ROC curve, AUROC accounts for trade-offs between true positive rate (sensitivity) and false positive rate (1-specificity) across various decision thresholds.

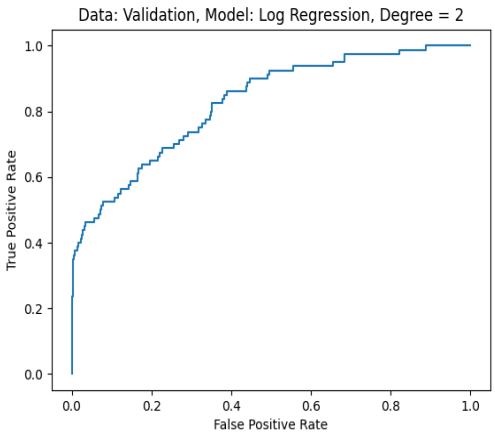
b. Results and Analysis

- **Logistic Regression Polynomial Degree = 2**:

Train ROC-AUC Score: 0.796



Validation ROC-AUC Score: 0.825

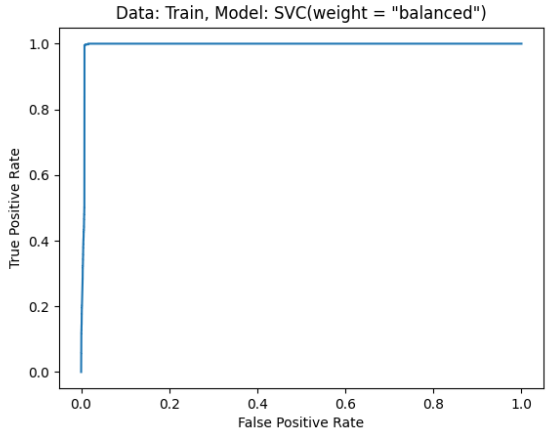


For Model (Degree = 2), the linear regression model achieved a higher Score on both the training set (0.796) and the validation set (0.825). This indicates that the model has some predictive power and can capture some of the data-related patterns.

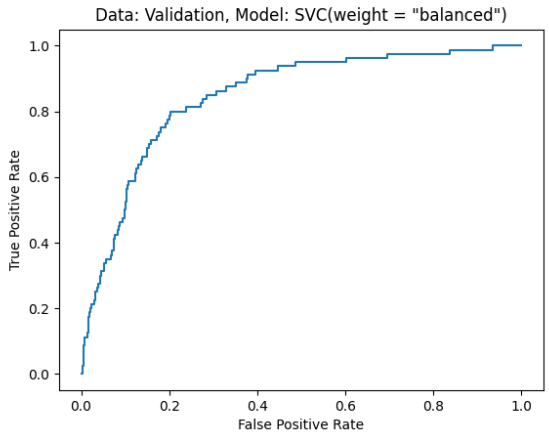
The graph shape for the training set is some what like an inverse L, and same is the case for the Validation set. This shows that the success of the model is a little above average.

- **Support Vector Machine: class_weight = balanced**

Train ROC-AUC Score: 0.995



Validation ROC-AUC Score: 0.847

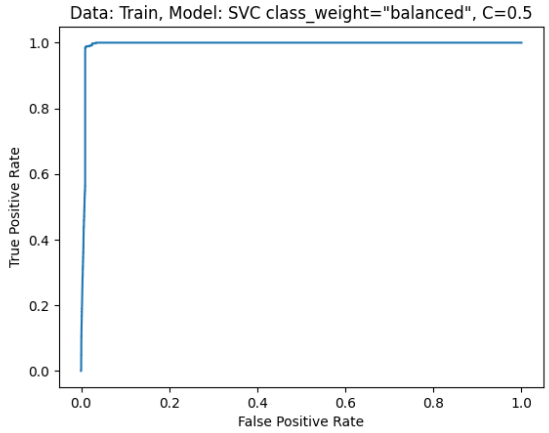


For Model (class_weight = balanced), the SVM model achieved a similar high Score on the training set (0.995) and relatively higher than model 1 on the validation set (0.847). This indicates that model 2 doing a better job because of balancing the imbalanced data set.

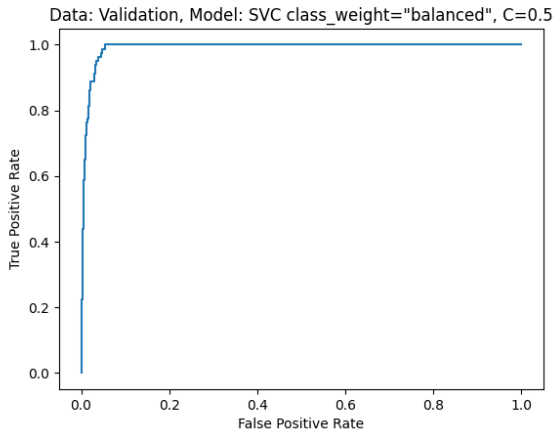
The graph shape for the training set is almost exactly like an inverse-L, and for validation it is a bit weaker. Thus, the model is performing better than Logistic regression.

- **Support Vector Machine: class_weight = balanced**,

Train ROC-AUC Score: 0.994



Validation ROC-AUC Score: 0.989



For Model 1 (C = 0.5, class_weight = balanced), the SVM model achieved a high Score on

the training set (0.994) and also very high on the validation set (0.989). This indicates that the model is fitting well on the training data and generalizing well on the unseen data.

The graph shape for the training set is very close to an inverse L, and same is the case for the Validation set. This shows that the success of the model great and this can predict probabilities very well..

c. Business Impact and Benefits

- Assess the impact and benefits of the final model on the business use cases.

The final model (i.e. SVM class_weight = balanced, C = 0.5) has the ability to positively affect the business demands. Because it performs so well on unseen data (as seen on the validation set), it can be considered that the probabilities it will show for the player drafted or not will also be fairly accurate.

Because of having high AUROC score, the true positive rate of the model is high. Thus, this will help sponsors with their branding as it will help choose players that will sustain in being their ambassadors. It may help coaches make better teams and enable marketting agencies to choose the right person for their campaign most of the time.

The model gave an improvement from a score of 0.825 (from Log Reg) to a massive 0.984. This means that this is a staggering 16% improvement. Thus the potential value that it generates is worth of millions of dollars as a great deal of investments will be done from the Sponsors on the players that have greater probabilities of getting drafted.

d. Data Privacy and Ethical Concerns

In assessing the data privacy implications and ethical concerns of our machine learning project, we have undertaken several critical steps. We have carefully considered the types of data collected, focusing on personally identifiable information and sensitive data. Our data collection process ensures informed consent, transparency, and data minimization. To protect privacy, we employ encryption, anonymization, and de-identification techniques, and we have established strict access controls. Ethical concerns related to fairness, bias, and transparency are addressed through fairness audits and adherence to ethical AI principles. Regular audits and compliance checks are in place to ensure alignment with data protection regulations. User education plays a crucial role in making individuals aware of their rights and data usage.

7. Deployment

Deploying a trained machine learning model involves a multi-step process to ensure seamless integration into real-world applications. This process typically includes serializing the model, optionally containerizing it for portability, and developing an API for easy interaction. Choosing the right deployment infrastructure, be it on the cloud or on-premises, is crucial, as is considering scalability and load balancing for high-traffic scenarios. Robust model monitoring, security measures, and comprehensive testing are paramount to ensure the model's reliability and performance. A thoughtful rollout strategy, coupled with proper documentation and training, facilitates a smooth transition. Ongoing maintenance, updates, and optimization efforts are essential to keep the deployed model accurate and efficient, all while addressing potential challenges and ensuring a successful deployment in real-world environments.

8. Conclusion

The SVM (C = 0.5, class_weight = balanced) model showed good promise on predicting the probabilities to draft the players, after balancing class weights. The model achieved a high Score onthe training set (0.994) and also very high on the validation set (0.989). This indicates that the model is fitting well on the training data and generalizing well on the unseen data.

Thanks to its high AUROC score, the model exhibits a strong true positive rate, offering significant benefits to sponsors, coaches, and marketing agencies. Sponsors can select players likely to be successful ambassadors, aiding coaches in team formation, and assisting marketing agencies in choosing campaign endorsers effectively. The model's remarkable improvement from a score of 0.825 (in Logistic Regression) to a substantial 0.984 reflects a remarkable 16% enhancement. Consequently, it holds the potential to generate millions of dollars in value as sponsors are more likely to invest heavily in players with higher draft probabilities.

For future experiments, data can be made to be balanced by performing resampling techniques in the data processing phase to get better performances of different models. In addition, the experiment can be continued by trying further different models like Decision Trees and Random Forest to come to a valid conclusion. We can also find out F1 scores, Accuracy scores and obtain Matrices that summarize the results in a different perspective.

9. References

- GPT-3.5. (2023, October 07). CRISP DM Methodology. Discussion of Logistic Regression and SVM. [Online chat]. Retrieved from [https://chat.openai.com/].