ISYE 6501 - Introduction to Analytics Modeling: Course Project

Prompt

This project should be done individually.

The web sites for SAS data and AI solutions (SAS), International Business Machines (IBM), and Institute for Operations Research and the Management Sciences (INFORMS) contain brief overviews of some major Analytics success stories (SAS[1], IBM[2], and INFORMS[3]). In this course project, your job is to think carefully about what analytics models and data might have been required.

1. Browse the short overviews of the projects. Read a bunch of them – they're really interesting. But don't try to read them all unless you have a lot of spare time; there are lots!

2. Pick a project for which you think at least three different Analytics models might have been combined to create the solution.

3. Think carefully and critically about what models might be used to create the solution, how they would be combined, what specific data might be needed to use the models, how it might be collected, and how often it might need to be refreshed and the models re-run. ***DO NOT find a description online (or elsewhere) of what the company or organization actually did.*** I want this project to be about your ideas, not about reading what someone else did.

4. Write a short report describing your answers to (3).

[1] SAS Customer Success Stories
[2] IBM Client and Partner Success Stories
[3] O.R. & Analytics Success Stories

Background

Considering National Basketball Association (NBA) teams, the Orlando Magic is recognized as one of the most analytically savvy teams. For over a decade, the Orlando Magic have been creating a path to utilize a data-driven strategy to improve the team and fan experience. According to the SAS customer success story, the team has a proven track-record, with a seven-figure surge in ticket sales and sponsorships by utilizing a new mobile app and a winning analytics strategy [1].

The Orlando Magic initiated the data-driven process by agreeing to a three-year deal with SAS Sports & Entertainment in 2010 [2]. SAS or Statistical Analysis System is a software suite that is used for statistical analysis, data visualization and advanced analytics. SAS has capabilities to retrieve, alter, and manage data followed by performing statistical analysis. In the first season, ticket revenue increased 50 percent and since 2013-2014, single-game ticket revenue is up 91 percent. With SAS, the Orlando Magic have implemented a data-driven culture that is crucial for anything from ticket utilization to lineup rotation. This success story has come in result of improved pricing, revenue management, and marketing strategies driven in large part by analytics.

Based on the Orlando Magic success story, a data-driven approach is discussed for marketing campaigns and player analytics.

Player Analytics

Player analytics is modeled for the Orlando Magic using big data with cloud deployment of SAS for Machine Learning and Deep Learning on AWS. The Magic have reported that they utilize over 20 years of data from more than 18 data sources on 100,000 players. This allows creating models for player fitness, player acquisition, and lineup optimization.

The goal is to create an analysis using 3 separate models and combining them to help come to a conclusion. To form an analysis using all 3 models, there are several considerations. You could use a stacking or layered approach to train individual models and combine their outputs into a meta-model using logistic regression or neural networks. This is helpful to learn relationships between outputs of the base models. A weighted averaging approach could also be implemented to assign weights to predictions from each model based on validation performance. This is a simple approach that is easily to interpret and a final output is given with a weighted average of predictions. If classification problems need to be analyzed, ensemble voting should be implemented to combine the 3 models. The models vote on categorical outcomes such as “fit” or “at risk” which reduces sensitivity to outlier predictions. Last, one could consider Bayesian Model Averaging (BMA) to combine models probabilistically. BMA handles uncertainty well and can incorporate prior domain knowledge.

Player Fitness

- Predictive models for injury risk using supervised learning (e.g., Random Forest, Gradient Boosting).
- Fatigue modeling using time series (e.g., LSTM, ARIMA).
- Clustering to group players by fitness profiles and recovery rates.

For player fitness, data can be utilized to analyze injury risk, fatigue, and training decisions. To model injury risk, predictive models can be implemented with supervised learning.

First, relevant data is needed historically and real-time. The historical data can include injury records, past performance, and recovery timelines. For real-time data, the player could be tracked with to use metrics such as heart rate and speed, and workload data can also be utilized such as minutes played and effort level. External factors could also be included such as sleep, time traveling, game schedule, and environmental data.

Next, a hierarchical framework is leveraged for each model’s strength. This allows for diverse perspectives and more robust predictions. For 3 models, the model combination strategy would include:

- **Injury Risk Model**: Predicts the likelihood of injuries using features like workload, prior injuries, and biomechanical stress.
- **Fatigure Prediction Model**: Estimates player fatigue levels using time series data on heart rate, speed, and workload
- **Recovery Model**: Predicts recovery timelines and optimal return to play

The *injury risk model* can utilize a random forest model to create an ensemble of decision trees based on subsets of data. Injury risk can then be quantified with the model predicting the probability of injury between 0 and 1. Using feature importance, high-impact factors can be determined to understand the features that put players at higher risk. The *fatigue prediction model* can use an Autoregressive Integrated Moving Average (ARIMA) model for forecasting or predicting future outcomes based on historical data. The *recovery model* will utilize logistic regression to predict recovery timelines.

To combine injury risk, fatigue prediction, and a recovery model, a hierarchical framework can be formed to train all models separately. Next, the model outputs are combined with raw features. Last, the 3 models can be utilized to train a meta-model using combined features and provide a target variable for overall score.

Lineup Optimization

To optimize the player lineup for each game as well as the lineup rotation in-game, one would need historical and real-time data. This data would be significant for on and off the court considering player chemistry metrics and game context. More specifically, data could be obtained by player tracking with camera to understand situational approaches and tendencies, as well as finding outliers where poor performance could be a result of the rotation. Player chemistry could be evaluated by assist networks and individual player statistics. Game situations could be included for opposing teams tendencies and player situational statistics.

Lineup optimization would be best determined by utilizing 3 models that account for optimizing rotations, lineup adjustments, and predicting lineup performance in specific scenarios.

- **Reinforcement Learning**: Optimize Rotations Dynamically
- **Bayesian Optimization**: Fine-Tune Lineup Adjustments
- **Neural Networks**: Predict Lineup Performance in Specific Scenarios

First, reinforcement learning can be implemented to optimize rotations dynamically. Reinforcement learning would be beneficial due to the capabilities for sequential decision-making, where the goal is to maximize long-term rewards by dynamically selecting lineups. This could include the basketball game considered with current score, time remaining, and foul situation. The game also could consider current player fitness, fatigue and matchups. Also, the model would include opponent lineup and strategy. With historical data, different game scenarios could be simulated with Deep Q-Learning (DQN).

Second, to identify optimal lineups while considering constraints, Bayesian Optimization (BO) would be implemented. BO will allow maximizing team performance metrics such as net rating and efficiency differential based on lineup configurations. The model would require lineup-specific features like scoring potential, defensive ratings, and player chemistry. Constraints should be included like minutes distribution, foul issues, or games on back-to-back nights. BO then iteratively selects lineups to evaluate balancing exploration and exploitation. Ultimately, the model identifies high-performing lineups for specific game situations or matchups while considering various constraints.

Third, neural networks can be used to predict lineup performance in specific scenarios. Neural networks are beneficial for modeling complex, non-linear relationships in data and will predict a lineup's performance. Neural network models can be formed considered historical lineup data such as points scored, defensive stop, and +/- values for player metrics. Next, data can also be considered for player-specific stats and opponent strategy. The neural network can then have an output layer that predicts performance metrics like expected points scored or net rating. This will allow for forecasting lineup effectiveness in real-time, assisting with pre-game strategy and in-game adjustments.

Player Acquisiton

For player acquisition a lot of data can be utilized to both evaluate the player as well as their fit for a team. Data included could include historical game data both on the collegiate and professional level. Second, the player positions, skillset, and playstyle can be evaluated while also considering advanced metrics such as Player Efficiency Rating (PER), Real Plus-Minus (RPM), and Wins above Replacement (WAR). Last, the data included must consider potential contract and salary data for cap management.

Using the data previously discussed, player acquisition models can be formed with models for:

- **K-Nearest-Neighbors (KNN)**: Classification of similar players
- **Regression**: Regression models to project career trajectories
- **Decision Tree**: Decision trees for evaluating trade/draft scenarios

Player acquisition models can be formed to find undervalued talent. Using KNN, players can be classified in similar groups to find similar talent that may be more affordable given the team's cap constraints. Regression models allow for projecting career possibilities and success. Decision trees will also be formed for evaluating trade/drafts scenarios.

With a classification model, regression model, and decision tree, the team can analyze gaps and desired skill sets. They can then rank players based on cost-effectiveness and potential. The acquisition scenarios can then be simulated using Monte Carlo methods. This will allow for smarter drafting and trading decisions, identifying under-the-radar players who fit team goals, and long-term financial sustainability.

Marketing Campaigns

Personalized marketing campaigns are formulated with the use of models with mobile app data and machine learning. Customer data is obtained by using 13 data sources and integrating streaming app data with historical data for more than 2 million customers. This database allows for considering ticketing, concessions, retail merchandise and app maker VenueNext is combined into an enterprise data warehouse. Individuals are then clusted into segments based on their behavior. This is all in an effort to increase fan engagement and satisfaction.

Data used to drive personalized marketing campaigns could include customer activity data, behavorial data, and transactional data. Customer activity data is based on app interactions such as time spent, pages browsed and clickstream. Behavioral data can consider game attendance and ticket purchasing behavior. Transactional data can include concessions and merchandise purchases. There could also be engagement metrics, demographics, sentiment data, and contextual data.

The first model to implement is a clustering model to identify individuals based on their behavior. Using KNN, fans can be segmented in categories such as "Super Fan", "Occassional Attendee" or "Merchandise Shopper". Predictive models can then be implemented such as logistic regression to predict interest in specific merchandise or ticket offers. Recommendation systems can then be formed with dynamic pricing models to optimize pricing for tickets or merchandise based on real-time and historical demand patterns. Finally, Natural Language Processing (NLP) models can be implemented to analyze survey responses and social media to gauge fan sentiment towards players, events or promotions.

References

[1] SAS, "Predictive analytics and AI deliver a winning fan experience." 2024, [Online]. Available: https://www.sas.com/en_us/customers/orlando-magic.html.

[2] E. Connolly, "Magic deal for SAS sports & entertainment." 2010, [Online]. Available: https://www.sportspromedia.com/news/magic_deal_for_sas_sports_entertainment/.