FOOTBALL PROJECT DEVELOPMENT PROCESS

Disclaimer: Analytics is not a purely linear process. It is expected that a practitioner will likely jump between different items outside of the order listed below.

1. Understand the analysis situation and core problem

*Describe the analysis situation*

Football is a sport of many variables; there are 11 players for each the offense and the defense on every play, and all players on a team must work together to find success. Each play has a specific situation, including factors like down, distance, quarter and distance to end zone. While capturing every variable is impossible, the structure of football nonetheless offers many variables that can be recorded. These variables come from overall statistics, play-by-play logs, online articles and game video.

Ultimately, we want to explain why teams as a whole are successful and why individual players are successful. Without intensive manual film study, these explanations will highlight explanatory areas without getting at a deep root cause. For example, we may identify that a team is ineffective at running the ball to the right on third and short, but we will not know if a particular offensive lineman (or multiple linemen) are causing the issues or if the running back has issues running to the right. Essentially, we can identify issues at a low level but one above the lowest, deepest level. Applications for this analysis include fantasy football, football strategy improvement, coaching decision-making improvement and roster formation.

*What is the core problem?*

What makes certain team units successful, and how is that related to individual performance?

*What is/are our target variable(s)? If none, what are we searching for?*

|  |  |  |
| --- | --- | --- |
| Supervised | Unsupervised | Mixed |

Our main target variables will be typical measures like points scored, yards, touchdowns, tackles, etc. In the play-by-play portion, we will need to create a measure of play success according to the situation. We could gain some insight from clustering teams/players according to their performance, and we may use those clusters as targets. We may need to create a custom metric for evaluating player performance overall.

*List the deliverables, include relative weights on a 100-1 scale*

|  |  |
| --- | --- |
| 95 | Datasets parsed from online statistics tables |
| 92 | Datasets parsed from play-by-play logs |
| 80 | Datasets parsed from online articles written about teams and players |
| 75 | Datasets parsed from game tape |
| 75 | Combined dataset at play granularity |
| 70 | Predictive models to explain team/player performance, built in R and Python |
| 65 | A D3.js intelligence tool for visually diving into the data |
| 60 | Evaluation metric for individual players, derived from play-by-play data and game tape data |
| 55 | Time series model to understand trends and variance over time |
| 50 | Java master software serving as a front-end to generate analyses |
| 45 | Optimization model selecting players for fantasy football and for a “real team” under a salary cap constraint |
| 45 | Strength-of-schedule-based model to predict performance given an opponent |
| 35 | Markov chain model to understand play-to-play performance and game-to-game performance |
| 30 | Clustering model to group similar players/teams/plays |
| 20 | Simulation model on team/player performance |
| 20 | Win-loss probability model using Bayesian statistics |

*What is the value proposition for this analytics project?*

This project will provide information for recommendations that can significantly improve a team’s performance on game day. It could also be used to gain a significant advantage in fantasy football by knowing which players are most probable to succeed.

*Link to and describe any external research*

None currently.

1. Exploratory Data Analysis and Data Preparation

*Link to or copy in the data dictionary*

*What are the issues with data quality (not including outliers)?*

*Specifically, are there outliers in the data? How did we determine they were outliers? What will we do with them?*

*Describe any other initial cleaning and related assumptions*

*What are the issues with data integrity (including merging and hierarchies)?*

*Describe any data modifications made for integrity purposes and related assumptions*

*Describe any transformations made, including variable type conversion, creating dummy variables, binning continuous variables and mathematical function transformations. How may these help the analysis?*

*Is time a dimension in this data? If so, give your initial impression on any time series considerations, issues and solutions*

*For target variables (if any), comment on scatterplots and correlations for continuous variables, mean and variance analysis if one is dichotomous, and category characterizations for categorical variables. Are there non-linear relationships?*

*Provide other notes on any variables of interest (if there are targets, must discuss those) after reviewing summary tables and histograms and/or box plots. Link to a separate document if necessary*

*Describe any feature selection or reduction methods used*

*Will you construct any custom metrics or create any custom classifications?*

\*\*\*COPY SECTION 3 BEFORE WRITING AND PASTE BELOW 3\*\*\*

1. Modelling – Round 1

*What is/are the general purpose(s) of modeling for this round?*

*What type(s) of model(s) are being developed?*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Descriptive | Diagnostic | Predictive | Prescriptive | Time Series |

*Select the model(s) developed*

|  |  |  |  |
| --- | --- | --- | --- |
| UNSUPERVISED | | | |
| Association Rules | | | |
| Clustering | | | |
|  | K-means | Hierarchical | Gaussian Mixed Model |
| Principal Component Analysis | | | |
| Collaborative Filtering | | | |
| SUPERVISED | | | |
| Artificial Neural Networks | | | |
|  | Single Layer | Deep Learning | Convolutional/Boltzmann |
| Decision Tree / Random Forest | | | |
|  | Single Tree | Boosted Tree | Bagged Tree |
|  | Random Forest |  |  |
| K-Nearest-Neighbors | | | |
| *Classification* | | | |
| Logistic Regression | | | |
| Support Vector Machine | | | |
|  | Radial | Polynomial | Custom |
| Naïve Bayes | | | |
| Discriminant Analysis | | | |
| *Regression* | | | |
| Linear Regression | | | |
|  | Standard | Interaction | Non-Linear Independents |
|  | Moderation | Mediation | Hierarchical |
|  | Transformed | Non-Normal Errors |  |
| MARS – Multivariate Adaptive Regression Splines | | | |
| Penalized Regression | | | |
|  | Ridge | LASSO | Elastic Net |
| TIME SERIES | | | |
| Linear Regression | ARIMA | GARCH | Fourier Analysis |
| Artificial Neural Networks | Simple Moving Average | Seasonal Adjustment | Simple Exponential Smoothing |
| Holt’s Exponential Smoothing | Holt-Winters Exponential Smoothing | Survival Analysis |  |
| PRESCRIPTIVE | | | |
| Optimization | | | |
|  | Linear Programming | Integer Programming | Stochastic Programming |
|  | Combinatorial | Convex |  |
| Simulation | | | |

*List any other models developed*

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

*Report any relevant diagnostics and/or model performance metrics*

*What new information did we learn about the variables?*

*What new information did we learn about the relationships?*

*Will any additional variables or data need to be brought into the analysis? If so, which variables and/or data?*

*Will any additional cleaning or transformations be required? If so, what cleaning and/or transformations?*

*Did the deliverables change? If so, what is new, modified or disregarded?*

*Is another modelling round necessary? Why?*

\*\*\*PASTE NEW MODEL ROUND HERE\*\*\*

1. Create Value, Provide Insights and Offer Recommendations

*List everything included in your deliverable package*

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

*What insights did you derive from your analysis? If your end product is an automated algorithm, what specific functionality does it provide?*

*How are your insights actionable? What should a user do to put this information into practice? Are the actions operational or strategic? If your end product is an automated algorithm, how specifically can the algorithm be operationalized?*

*Otherwise, what value does your analysis/code provide to the user? How does it match the requested deliverables?*

*How is your analysis/code innovative or spark innovation in others?*

*Given more time, what could be done to expand upon this analysis/code?*