Eller College of Management

MIS 580: NFL 2001 Knowledge Discovery

Final Report

Kaijia Bao • Abhijit Kumar • Vishal Rupani

TABLE OF CONTENTS

INTRODUCTION

The team composition for the MIS580 course in Knowledge Management was a good balance of technology and business. As part of the MBA curriculum, we had already worked on a project with Honeywell Aerospace in patent research and analysis. In addition, with the Super-Bowl being hosted in Arizona and the rising interest in football, the choice for sports as the topic for the KM project became an easy one. The team got in touch Dr. Lisa Ordonez, professor of statistics and researcher of sports data at Eller, to discuss ideas on working with complex data related to football.

The professor is currently pursuing active research in the field of behavioral analysis in football based on an entire 2001 season of football plays. The dataset is a comprehensive set of every play made that season with player details, yardage, direction and outcomes amongst others. (Refer to Appendix A for detailed breakdown of data columns in our dataset). This is a detailed report on the study of NFL play data for the 2001 season using statistical and data mining techniques. In addition, we also contacted Steve Aldrich, who is a football expert, and is currently working on “Moneyball in Football” – a paper on using data analysis in developing better football strategies. Lisa and Steve, played the part of subject experts in this exercise by validating our research, analysis and results.

OBJECTIVES

Two key objectives were identified in drafting the proposal for this project. During the course of the semester, these objectives were successfully realized. One of the key aspects of these objectives was to develop a solution set of probable patterns and then focus on the strongest variables. During the analysis phase, our team came across several interesting findings, but in order to establish relevant findings the focus was streamlined to "PlayDecision” and “PlayDirection". Results achieved towards each objective have been outlined in the “Data Mining” section of this report. Below is a list of the two main objectives identified for our project:

- **Pattern Identification:** After a detailed understanding of the dataset and their impact on the eventual outcome of that play, the team was able to identify some trends in play patterns with changing goals (yardage, down, etc.) and their impact on player performance (play decision, play direction).
- **Prediction:** Another key objective of this project was to predict likely player behavior under different situations. Initially, the team had planned to analyze an entire season's play data and associate it with player information, coach statistics and team metrics, in order to define a relation between these key parameters. However, during the course of the semester, the focus was restricted to using play-by-play data exclusively as our dataset.

KNOWLEDGE DISCOVERY

Our team followed the Knowledge Discovery process that was taught in class. Each of the stages of this process has been described in detail below. We have also illustrated our system design in Appendix B.

DATA COLLECTION

Our team was able to collect data from two different data sources:

1. **Professor Lisa Ordonez:**

Professor Lisa has access to a proprietary play-by-play football database for the entire 2001 season. This contains raw quantitative and descriptive data. It does not have any aggregate data. It spans 1 table, 90 columns and 50,415 rows. The column names and respective descriptions can be found in Appendix A.

2. **Pro-football-reference:**

This database contains aggregate information about teams and players. It spans 3 tables, 40 columns and 82,345 rows. Although we collected this data, we did not use it in our analysis. A detailed explanation for this can be found in the “Future Directions” section below.

Both the above data were available in Microsoft Excel format. The tool we used to import the data into the database was SQL 2005 Integration Service.

DATA TRANSFORMATION

During this stage we cleaned, selected and transformed the data into a format that can be consumed by statistical analysis and data-mining models.

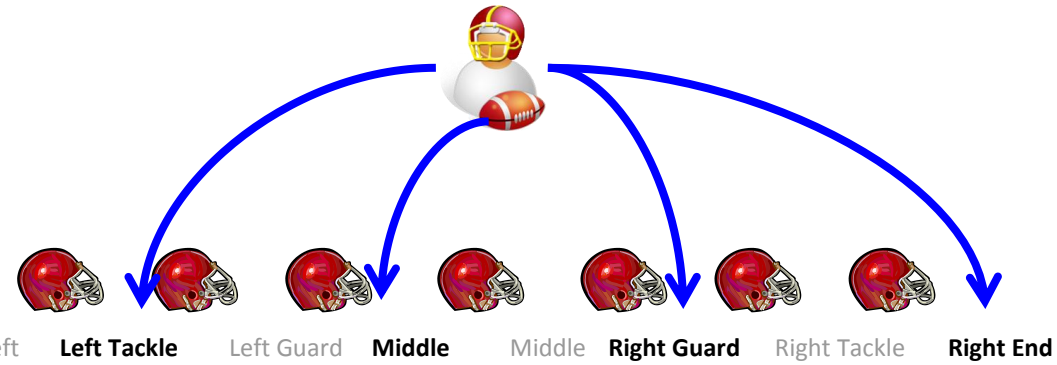
Cleaning: We initiated the data transformation process by cleaning the data first. The following is a summary of the cleaning process undertaken:

1. We started with 50,415 rows of data.
2. We had about 117 rows of data that contained partial duplicate entries (totaling 234 rows). These occurred whenever a specific play was "challenged" by an offense or a defense coach. Depending on the decision taken for the challenged play, a new entry was created using the same identifier (unique key) variable. Since the duplicate entries were a very small portion of the total size of the dataset, in order to save time, we removed these entries all together.
3. In the remaining 50,181 rows, there were 3,148 nullified plays that were removed.
4. Finally we had 47,033 rows of data that we used for the knowledge discovery process.

Dependent variables: Prof. Lisa and Steve recommended looking at the following four variables that they believed were critical to the game and could result in interesting findings. These dependent variables could be objectively used by a coach to formulate game strategies.

1. **PlayType:** Decisions such as passing, rushing, punting, field goal and kneel taken by an offense team during a play. We calculated a new dependent variable called "PlayDecision" from this which has been defined in the next section.

2. **Direction:** This is defined as the path taken by the person rushing the ball at a strategic position on the line of scrimmage. As you can see in the figure below there are several possible directions in which the ball can be rushed.



We calculated a new dependent variable called "PlayDirection" from this and this has been defined in the next section.

3. **Intended:** This is the name of the player from the offense team to whom a pass is intended.
4. **Yards:** This is the number of yards covered during a play.

Calculated variables: We calculated nine variables from the collected data and these are specified in Appendix A. However, we used only the following three calculated variables in order to correctly undertake statistical analysis and data mining:

1. PlayDecision: This was created from the "Decision" variable which originally included penalties. Penalties were excluded in this calculated variable. This is our new dependent variable.
2. PlayDirection: This was created from the "Direction" variable which originally included penalties. Penalties were excluded in this calculated variable also. This is our new dependent variable. In the case of data mining analysis, we also aggregated all the "left tackle" and "left guard" to "left" and all the "right guard" and "right tackle" to "right".
3. HalfTimeLeft: This is the time left until the end of "second quarter" or "fourth quarter" or "extra time".

Independent variables: After the dependent variables were defined, we needed to define a list of independent variables that best predict the selected dependent variables. We used SQL Server 2005 Analysis Service to help us define the list of independent variables. The analysis service samples the dataset and uses entropy (not correlation) to calculate a score (between 0.0 and 1.0) that measures how well variables predict the selected dependent variable. Unfortunately, we were not able to find how the score is calculated exactly, but the documentations stated that the score is based on entropy calculation. Appendix C shows an example output of how well a set of variables predict "PlayDecision". The entropy analysis was repeated for all the dependent variables and an initial list of independent variables was generated by

1. Combining a set of variables that had a good prediction score.
2. Removing any duplicated derived variables. For example, since "PlayDecision" is derived from "PlayType", "PlayType" is removed from the set of independent variables.
3. Removing any other outcome variables (penalties, scores gained for the play, etc.) that are not specified as our dependent variable.

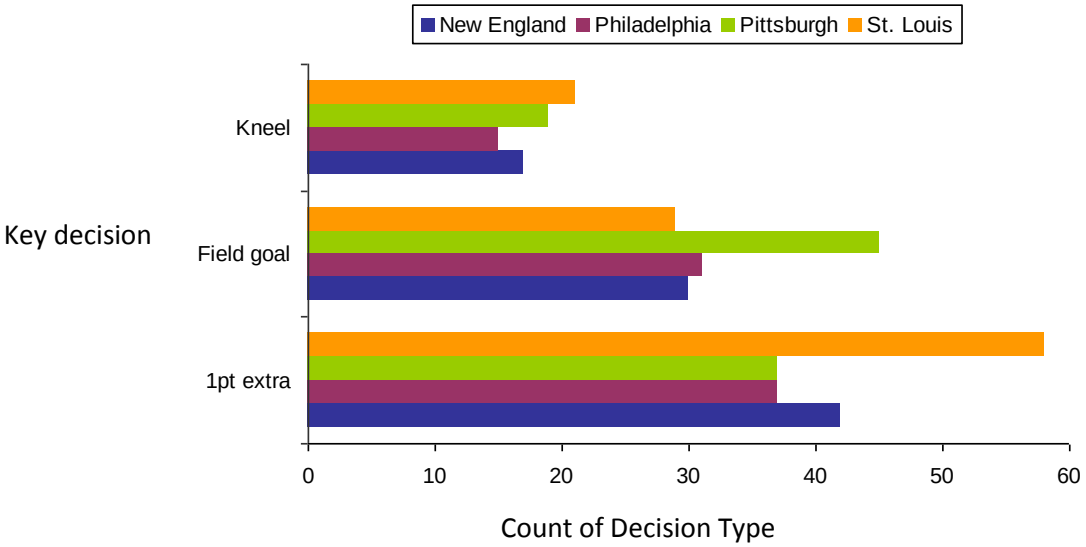
By using the calculated scores for all dependent variables, a Dependency Network diagram can be generated to graphically represent the relationship between variables. Appendix D shows a Dependency Network Diagram with "PlayDecision" node selected. Using the dependency diagram, we were able to further refine our list of independent variables to ensure there are no two-way dependencies between the dependent variable and any of the independent variables. For example, from Appendix D, you can see that there is a two-way dependency between "PlayDecision" and "Intended". Therefore, "Intended" was removed as an independent variable. Using the tools provided

in the analysis service, we identified ten independent variables which be found in Appendix A. These are classified as “Independent Used”. It also lists other independent variables which were not used.

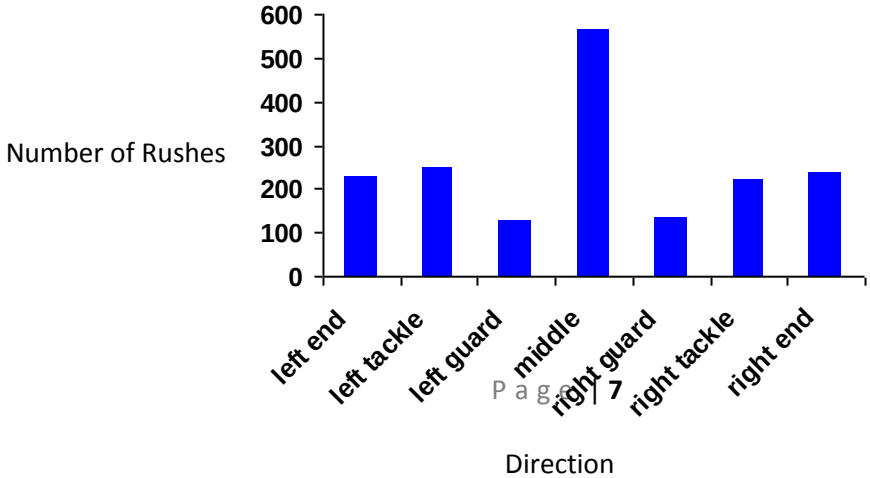
DATA PROCESSING

Excel was used to conduct simple statistical analysis for the 2 new dependent (and calculated) variables.

PlayDecision: The following graph takes a look at three key decisions taken by the top 4 teams in the 2001 season. On the Y axis is the decision type i.e. Kneel, 3pt Field Goal and a 1pt extra. On the X axis is the count of each of the decision types. From the below graph a defense coach would know that when the offense team in St. Louis they will often go for 1pt extra to gain a majority of their points.



PlayDirection: The graph below shows Direction of Rushes for all plays in 2001 season. The number of rushes is represented on the Y axis. Each of the different directions is represented on the X axis.



As evident, overall most rushes happen in the middle. Such statistical analysis, if performed for each specific team, can be used by a defense coach to better plan and possibly place its best defense players in the middle. At this stage we do not know how accurate our predication is. Data mining tools are required to determine the accuracy.

DATA MINING

SQL Server 2005 Analysis Services provided us with a plethora of data mining algorithms for sports predictions, which includes clustering, association, decision tree, regression, naïve bayes, neural network and time series. Our team chose to use data mining algorithm to predict "PlayDecision", "PlayDirection" and "Intended" because these variables capture the set of decisions that a coach has to make during a play. Our team also chose to use only decision tree and neural network algorithm in our prediction for the following reasons:

1. Some of the algorithms have restrictions on the type of variables that it could use for prediction.
2. Results of decision tree could be easily interpreted and viewed by users (i.e. coaches) of our analysis.
3. Neural network is an algorithm that simulates "brute force memorization" which was used to compare accuracy with the decision tree model.

"PlayDecision" Prediction: Appendix E shows a small fraction of a seven-level decision tree in predicting "PlayDecision". As you can see from the appendix, the outcome of "PlayDecision" can be easily interpreted in a graphical format as each node in the tree will have a percentage breakdown of the probability that a particular actions will be taken. For example, the decision tree shows that during 4th down, if the offensive yard line is greater than 91 yards, and the offensive team has 10 or more yards to go until the next down; there is a 99.91% chance that the offensive team will punt the ball. However, if

the offensive team has less than 10 yards to go until the next down, there is a 67.18% chance that the team will perform a field goal, 8.21% chance that the team will perform a pass, 13.33% chance that the team will perform a punt and 9.74% chance for a rush.

The follow chart shows the accuracy of the two algorithms used to predict "PlayDecision". As you can see from the chart, the overall accuracy of our data mining model is 68% for decision tree and 62% for neural network. However, the accuracy of each model varies depending on what types of decision it is trying to predict. The chart shows that the decision tree is very good at predicting "1pt extra" plays with 94% accuracy. However the algorithm is not so good at predicting "rush" plays with only 60% accuracy. Please note the blank cells denote that there are not enough predictions to create an accuracy measure for the particular "PlayDecision". Our team believes that with additional data, the models will more accurately predict play decisions and perhaps even discover team specific strategies.

	Decision Tree	Neural Network
Overall	68.02%	62.14%
punt	69.39%	67.96%
1pt extra	93.59%	100.00%
rush	60.00%	53.36%
field goal	77.48%	59.46%
2pt conversion		
risky	100.00%	
kneel	69.67%	
pass	73.14%	68.66%

"PlayDirection" Prediction: In addition to predicting play decisions, our team also used the model to predict play directions. Unfortunately, the accuracies of our predictions are 34.74% for decision tree and 37.65% for neural network, which is only marginally better than using random guesses. This indicates that our models are not accurate in predicting play directions. By examining the decision tree shown in Appendix F, Further analysis found that overall there are no specific patterns in the directions in which the ball is rushed. However, the model was able to find an outlier that does not conform to the average. The decision tree shows that while most other nodes contains a fairly even distribution between the different directions, when it is 2nd down and the defense team is "CHI" (Chicago), there is

significant higher tendency for the offense team to go right instead of middle. This finding could indicate that either Chicago have a strong middle and/or weak right defensive line. Our team believes that with additional data from other seasons, more interesting outliers can be found with our models.

"Intended" Prediction: Lastly, our team used the models to predict the "Intended" receivers for a pass. Unfortunately, similar to our "PlayDirection" prediction model, the accuracies of predicting intended receivers for decision tree and neural network are 24.41% and 23.78% respectively. Our team found that over one season, there are over 400 players that received a pass during a play. Only a fraction of the players received significant amount of passes, most of the player received less than 5 passes. Therefore our team concluded that there is just not enough data to construct an accurate model for all 400 potential receivers. However, with additional data from other seasons, our team believes that a model can be created for intended receivers.

FUTURE DIRECTIONS

Increase sample set: Increasing the sample set to include several seasons of play data would benefit this study, by providing multiple instances of different scenarios. This would help in analyzing complex predictor variables such as "Intended player" with reasonable accuracy.

Incorporate additional information: The team considered incorporating different supplementary data such as, player information, coach performance and team data (from Pro-football-Reference.com). In addition, to help with the predictor model, incorporating information about odds and bets in this model, would add a new layer that would further increase accuracy for predictions (such as the data at VegasInsider.com for favorites before the game begins).

Extend Analysis: Our team considered extending the analysis by using more complicated data mining

tools such as the Nested case analysis, which provides a historical perspective to the team, player and coach performance, which could aid the knowledge discovery process by bringing out several interesting and new patterns.

CONCLUSION

The team conducted extensive statistical and data mining techniques to discover interesting patterns and use them to base predictions with an accuracy that was better than a random guess model. During the course of the semester, the team was able to analyze the four independent variables and was able to discover patterns in the "Player Direction" variable that seemed most promising. The other variables (Play decisions, Play directions and Intended players) had varying degrees of accuracy. This was due to the fact that some of these variables did not have sufficient instances within one year's data, to help our team identify patterns. A summary of the variables analyzed and their results are as follows:

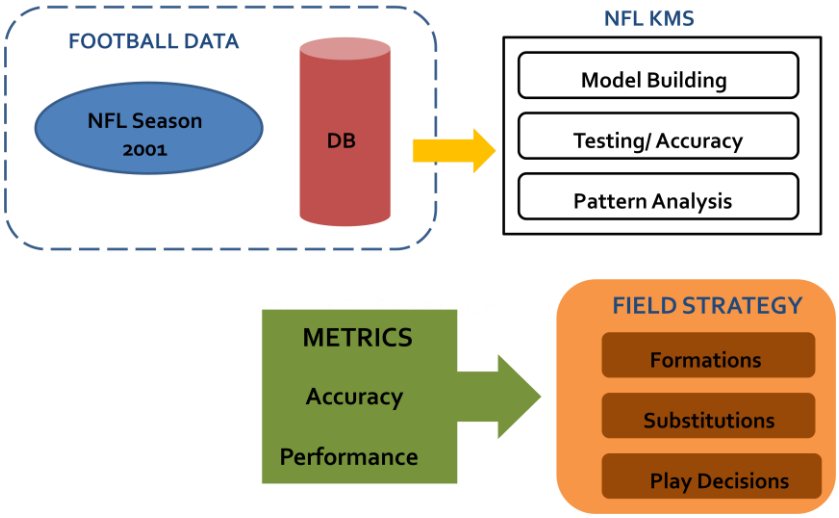
- **PlayDecision:** – Accurate model – with more than 68% in accuracy and over 90% accuracy in specific types of decision predictions.
- **PlayDirection**: – Fairly accurate – but provided enough data for the team to gain useful knowledge– e.g. upon analyzing Chicago Bears as the defense team, it was found that the probability of rushing right was considerably larger than other directions.
- **Intended player:** Not accurate: Not sufficient data to help us generate any knowledge. This was due to the lack of instances of intended player data in one seasons' play data.

These results were discussed and validated by the experts – Steve and Lisa – who found them insightful and interesting. They have proposed a follow-up project to this exercise, to extend our analysis to more variables and techniques.

APPENDIX A – DATA SET: NFL SEASON 2001 PLAY DATA

NO.	FIELD	DESCRIPTION	TYPE
1	NEW DESCRIP	NULL	Independent
2	YEAR	Year of the season - this is 2001 for the entire dataset.	Independent
3	NUL	Nullified plays. Null penalty = this penalty nullifies previous play; Null TD = nullified TD, but play still counts; No Play = no play due to penalty (including nullified TDs where play is totally cancelled); Shortened play penalty = penalty shortens	Independent
4	CHAL	Result of Instant Replay Challenge: Coaches can challenge a limited number of calls made by referees. Upheld = original call of the ref upheld, so no change is made; Reverse = call made by the refs reversed; Fix=is the data from the previous line	Independent
5	XTRA NOTE	This column mentions any specific strategies that the team used, such as formations. Examples - Shotgun, which is a passing formation in which the quarterback stands 5 to 7 yards beind the center before the snap.	Independent
6	WEEK	Week of the season in which the game was played. The dataset has all the games played in the season, which goes from	Independent
7	QTR	One period of a football game. A NFL football game is divided into four quarters of 15 minutes each (but can go into over time, QTR=5).	Independent; Used
8	PLINE	Unique number for each play ordered by week, by game, by play sequence within a game. If two numbers are the same except one has a .5 decimal point, then that play is the revised play (see note on CHAL above)	Independent
9	HOME	The home team. 32 teams denoted individually with 2 or 3 letters	Independent
10	AWAY	The away team. 32 teams denoted individually with 2 or 3 letters	Independent
11	1/TD	Positive Outcomes: FD=First down, FG= Field Goal; SAF=Safety; TD=Touchdown; TWO=2 point conversion; XP=1 point conversion. Values for this field include FD or FG - which mean Field Goal (Generally, teams will attempt field goals on fourth down when they feel they are within reasonable distance of the goalpost in the opponent's end zone) or Field Goal	Independent
12	SC	Score on the play. FG=3, TD=6, XP=1, TWO=2, Safety (SAF)=2	Independent
13	HS	Home Score - this is the score for the home team, carried forward from the previous line item. Note that "Home" is stagnant for the game but this team can be offense or defense depending on if it has posesstion of the ball	Independent
14	AS	Away Score - this is the score for the away team, carried forward from the previous line item. Note that "Away" is stagnant for the game but this team can be offense or defense depending on if it has posesstion of the ball	Independent
15	GAP	Offense - Defense Score (Gap or difference). Note that this is not the same as subtracting the Home vs. Away scores since offense and defense changes during the game.	Independent; Used
16	OFF YDL	The five offensive players that line up on the line of scrimmage and block for the quarterback and ball carriers. The number of yards from the offensive goal line on that play (min=1 and max=100)	Independent; Used
17	OFFENSE	The team that has possession of the football and attempts to advance it toward the defense's goal line.The main goal of an offense is to pass or run the football into the opposing team's goal for a touchdown. The offense plays against the other	Independent; Used
18	DEFENSE	The unit that is responsible for keeping the opposition out of their end zone.	Independent; Used
19	ZONE	Zone type-- breaks up the field into 6 zones relative to offense from far away from goal line (DEEP) to within 10 yards of TD (RED2). DEEP= deep in own 1-19 yrdline; BACK=own 20-39 yrdline; MID=midfield 40-40; Front - Offense towards Goal	Independent
20	DECISION	The type of play. In general, the actions of the players following a snap or kickoff. More specifically, the type of action taken as part of a planned manuever. On offense, there are two basic types of plays; running and passing. On special	Dependent
21	PLAYER	Player involved in the play. Information includes player name, team and jersey number.	Independent
22	YARDS	The amount of yards gained or lost during the play.	Dependent
23	RECEPT	Mostly lists specific penalties (see code defs or penalties worksheets), but also lists what happened to pass. Pass can be complete (caught), incomplete, intercepted, QB sacked (tackled) before getting pass off, aborted snap from Center	Independent
24	TO	FK=fumble kept by offense; FL=fumble lost by offense; TO=interception; HM=Hail Mary play interception--i.e., throwing it far and praying to God that receiver catches it but it is intercepted at the end of a half where the result is really no different from an incomplete, or in the last couple minutes of a game where the offense is losing by a bunch of points and the game	Independent
25	DIRECTION	The direction when the play time was rushed (i.e. running play) or category of penalty (conversion, field goal, kickoff,punt,standard). Categories of run directions are: left end, left guard, left tackle, middle, right end, right guard, right	Dependent
26	INTENDED	In the case that a pass was completed successfully by a player other than the one intended for originally, the original	Dependent
27	KICKER	The player who kicks the ball on kickoffs, extra point attempts, and field goal attempts.	Independent
28	KICKYDS	Yards covered in the kick-off or the punt.	Independent
29	KICK YDL	Line from where they kicked from	Independent
30	CATCH YDL	The yardline at which the kick-off or the punt was caught	Independent
31	RETURN YDL	The yardline till which the player ran after catching the kick-off/punt	Independent
32	DOWN	The number of the down in the game - first, second, third or fourth	Independent; Used
33	TOGO	Yards to go according to the down to reach the next FD	Independent; Used
34	AT	just a word	Independent
35	YDTM	Yardline team for position of down	Independent
36	ACYDL	Actual Yardline - for position of down	Independent
37	TIME	Game clock	Independent
38	FirstDownType	The league counts the first downs. Number is # first down for that team in the game. P pass, R rush, X penalty.	Independent
39-90	Description	These are specific descriptive fields that were used to further define play data	Independent
91	GameNum	Used to indicate which game the current play belongs to.	Calculated
92	IsPlayChal	Used to indicate whether the current play is challenged by the team (not used)	Calculated
93	PlayZone	Used to indicate the zone of the play (combined Red1 and Red2 to form "Red")	Calculated
94	PlayDirection	The direction of a play (eliminated penalty information here)	Calculated & New Dependent
95	TotalOffTO	Indicates the total number of turn over for the offensive team for the current game	Calculated
96	PlayDecision	The play decision that team/coach have to make (eliminated penalty information here)	Calculated & New Dependent
97	QtrTimeLeft	The time left in the qtr	Calculated
98	HalfTimeLeft	The time left until half time	Calculated
99	GameTimeLeft	The time left until the end of the game	Calculated

APPENDIX B – SYSTEM DESIGN

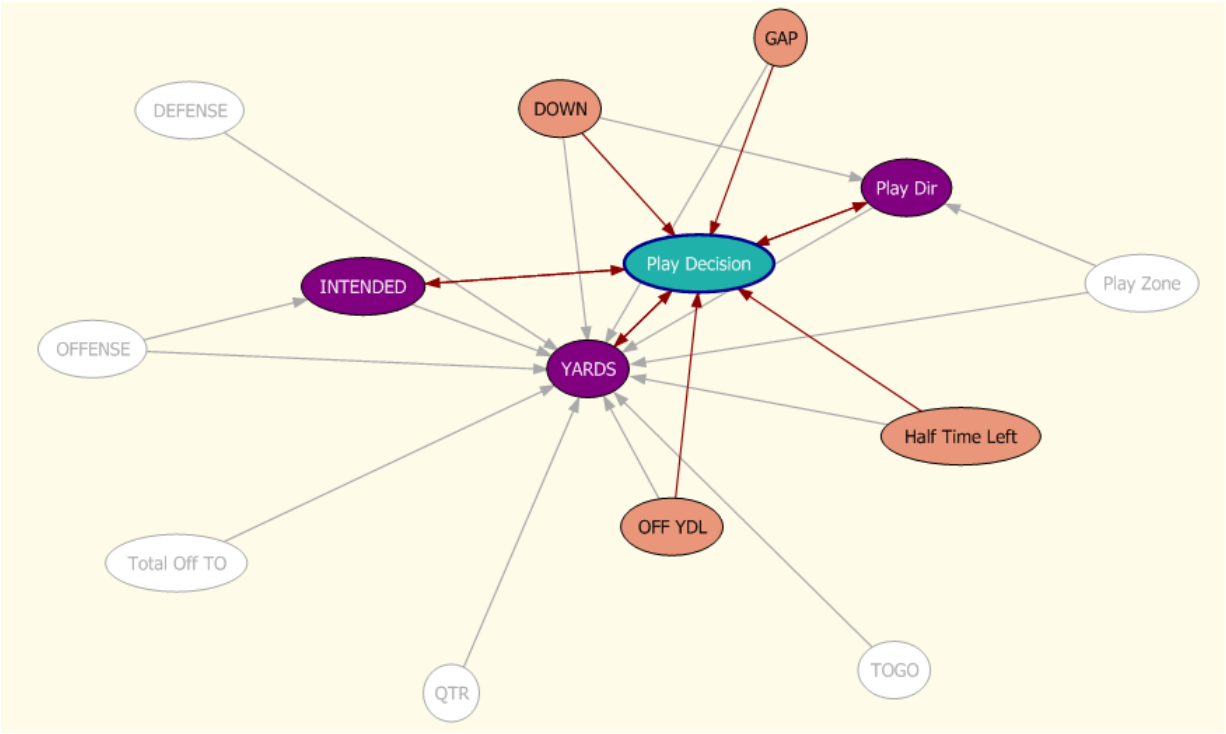


APPENDIX C - PREDICTION SCORE FOR "PLAYDECISION"

Columns related to PlayDecision:

Column Name	Score	Input
PLTYPE	1.000	x
DOWN	0.477	x
YARDS	0.473	x
RECEPT	0.466	x
OFF YDL	0.413	x
DIRECTION	0.409	x
PlayDir	0.403	x
TOGO	0.377	x
ACYDL	0.345	x
ZONE	0.285	x
PlayZone	0.281	x
YDTM	0.257	x
XTRA NOTE	0.236	x
KICKER	0.191	x
KICKYDS	0.190	x
KICK YDL	0.187	x
1/TD	0.177	x
CATCH YDL	0.171	x
RETURN YDL	0.161	x
FirstDownType	0.137	x
SC	0.127	x
UsedSpecialStrat	0.114	x
GAP	0.094	x
NEW DESCRIP	0.070	x
DEFENSE	0.042	
OFFENSE	0.041	
HOME	0.040	

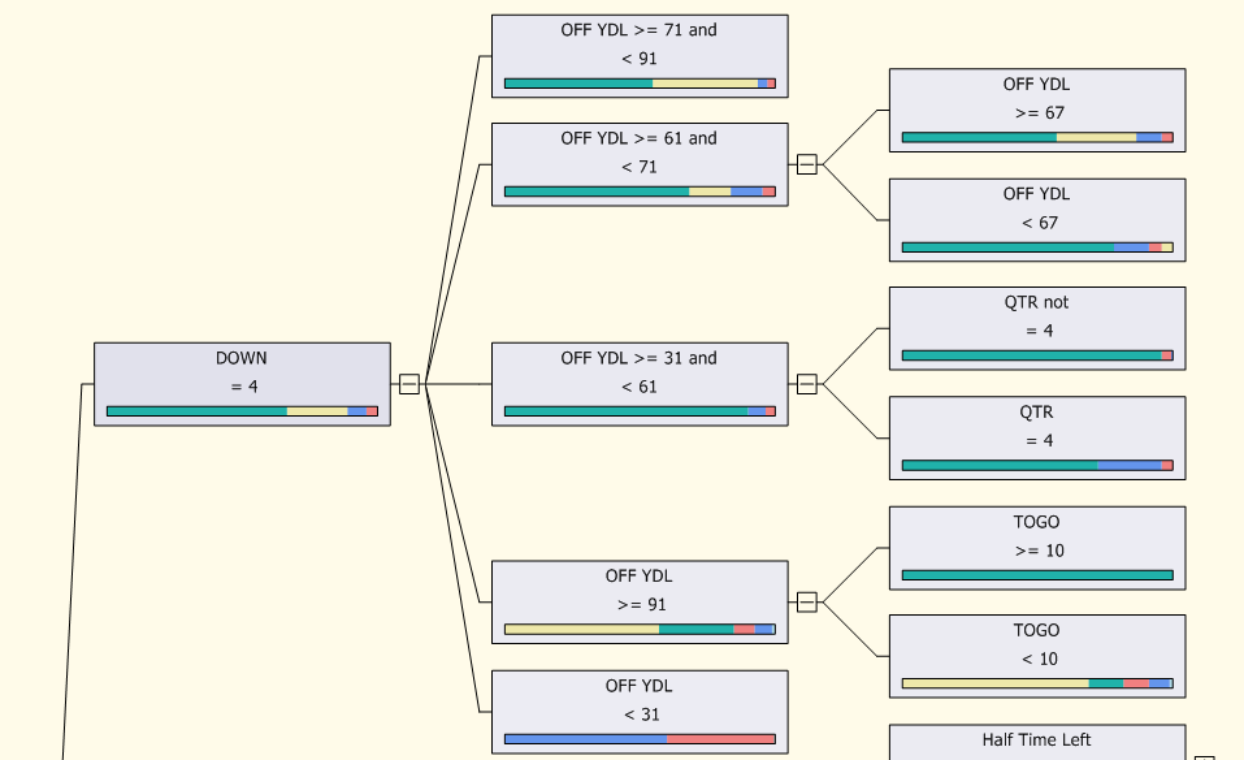
APPENDIX D - DEPENDENCY NETWORK: "PLAYDECISION"



Select a node in the network to highlight its dependencies.

- Selected node
- This node predicts the selected node
- Selected node predicts this node
- Predicts both ways

APPENDIX E - PARTIAL DECISION TREE: "PLAYDECISION"



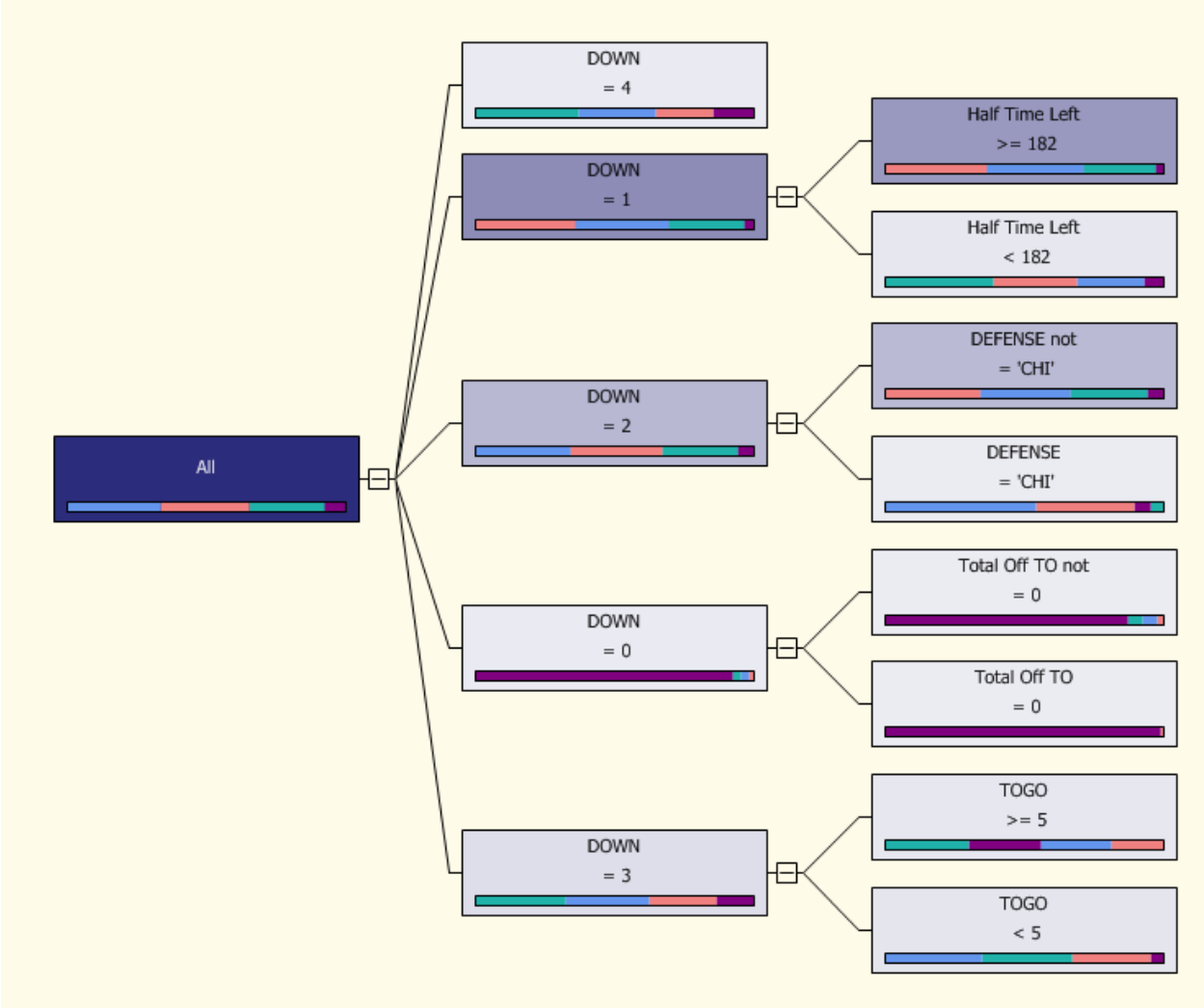
Down = 4 → OFF YDL >= 91 → TOGO >= 10

Value	Cases	Probability	Histogram
1pt extra	0	0.01%	
2pt conversion	0	0.01%	
field goal	0	0.01%	
kneel	0	0.01%	
Missing	0	0.01%	
pass	0	0.01%	
punt	40	99.91%	
risky	0	0.01%	
rush	0	0.01%	

Down = 4 → OFF YDL >= 91 → TOGO < 10

Value	Cases	Probability	Histogram
1pt extra	0	0.00%	
2pt conversion	0	0.00%	
field goal	131	67.18%	
kneel	3	1.54%	
Missing	0	0.00%	
pass	16	8.21%	
punt	26	13.33%	
risky	0	0.00%	
rush	19	9.74%	

APPENDIX F - DECISIONS TREE: "PLAYDIRECTION"



Down = 2 → DEFENSE = 'CHI'

Value	Cases	Probability	Histogram
left	30	35.70%	
middle	4	4.78%	
Missing	0	0.02%	
right	45	53.54%	
Standard	5	5.97%	