Executive Summary

# Scenario

Soccer fans have long wanted to understand more of the information behind the beautiful game. Soccer is such a dynamic sport it is hard to predict the outcome of a match by simply observing trends over the course of the match. A game changing play could happen at any moment with no indication and that could be all that’s needed to win the match. With the help of data from soccer data analytics company, StatsBomb, I was able to analyze 17 seasons of data for FC Barcelona to find out what makes them most likely to win.

The objective of this project is to predict the likelihood of a win for FC Barcelona. This analysis is very beneficial to a team because they can develop a system of play that increases their chances of winning matches. To achieve this, I aggregated 2 million rows of event data from 519 soccer matches to see summaries of matches alongside their outcome. Event data is every on-ball action that occurs in a match. This can be anything from passes, shots, and dribbles all the way to defensive actions, goalkeeper involvement, and even disciplinary action by the referee. While some events are far too granular for meaningful high-level analysis, we can aggregate several key parts of the match to see overall summary metrics.

# Data Preparation

Since this dataset was so large there was a significant amount of data wrangling. First, I identified the primary metrics on which I wanted to report. Overall, these metrics were ball position, passing, dribbling, and shooting metrics for both FC Barcelona and their opponent. These are important metrics because they indicate possession of the ball and attacking threat, which are two of the most important factors in winning a match.

Most of the metrics were normally distributed which made for a useful dataset. There were some skewed variables that were not corrected by log transformations.

We wanted to check for collinearity as well, so we created correlation plots that show the relationship between some of the variables. One of the common areas where we saw collinearity was between attempts, completed attempts and completion rate (attempted\_passes, completed\_passes, pass\_completion\_rate). Attempts and completed attempts are nearly always going to increase and decrease together. This is evident by the correlation plots below.

Chart

Description automatically generated

With the help of total attempts, conversion rate explains the number of completed attempts therefore we removed all completed attempt variables.

# Analysis

There are a variety of different ways to play the game of soccer and the best playing styles have been debated for a long time. There are two elements of the game that are undoubtedly effective ways to win a match. Those ways are through ball control and quality goal scoring chance creation. When a team has the ball more, they can control more of the game and when a team can create goal scoring opportunities (shots) then they have a greater chance of scoring. Our data supports this as well. Looking at a Garson plot below we can see the two most important variables in the model are close\_shots and avg\_ball\_position\_x.

Chart, bar chart

Description automatically generated

What’s very interesting about this plot is that we don’t see opp\_avg\_ball\_position\_x (the average ball position lengthwise) on this chart, but avg\_ball\_position\_x is the most important variable. But with an understanding of FC Barcelona’s playing style this makes sense. Barcelona usually dominates possession so a match where they struggle to maintain possession would be an indicator of a poor performance possibly leading to a loss. Having more possession of the ball would lead to the average ball position being higher up the field (assuming Barcelona tries to moves the ball forward). However, some teams willingly concede possession to Barcelona. It does not mean they are badly beaten, it is just a different playing style. Often, teams prefer to counterattack when Barcelona loses the ball. This would lead to more opp\_close\_shots which is the third most important variable.

# The model

When fitting the model I decided to use logistic regression and an artificial neural network. The first step was to identify the best logistic model selections. I used the regsubsets function on the full model and scaled using both BIC (Bayesian Information Criterion) and Adjusted R-Squared. Below we can see the graphs showing us the most performant model for each scale type for the given variable limit of 20.

A picture containing chart

Description automatically generated

The most performant model by Adjusted R-Squared is our response variable “win” as a function of opp\_pass\_completion\_rate, attempted\_dribbles, dribble\_completion\_rate, opp\_attempted\_dribbles, opp\_dribble\_completion\_rate, attempted\_through\_balls, through\_ball\_completion\_rate, opp\_attempted\_through\_balls, avg\_ball\_position\_x, opp\_avg\_ball\_position\_x.

A picture containing calendar

Description automatically generated

The most performant model by BIC is our response variables “win” as a function of opp\_pass\_completion\_rate, through\_ball\_completion\_rate, avg\_ball\_position\_x, close\_shots, midrange\_shots, opp\_close\_shots, opp\_midrange\_shots. Finding the lowest BIC and the highest Adjusted R-Squared gives us two model selections to train, test, and fit.

We went on to fit these models using 10-fold Cross Validation. We also tested every possible threshold (.1 - .99, by .01) against our two model models to determine their accuracy. The BIC model showed its highest accuracy of 77.84% at a threshold of .63. Graphical user interface, chart

Description automatically generated

The Adjusted R-Squared model showed its highest accuracy of 76.88% at a threshold of .63.

Graphical user interface, chart

Description automatically generated

The last major step was to perform an outer layer of double cross validation around our models to determine which model was best. At this point we created a logistic regression model and artificial neural net model each fit on three different model subsets (full dataset, best by BIC, and best by adjusted r-squared). Double cross validation allowed us to compare all the models’ accuracy and selection the best model. The most accurate model was the original BIC logistic regression model with an accuracy of 76.71% as determined by the double cross validation. But this wasn’t higher than our original model tuned to the threshold of .63 so is it really our best model? The best model overall is the original BIC logistic regression model tuned to a threshold of a .63 probability. When we finally fit this model on the full dataset it performs better than on the test set. Its final accuracy once fit on the full dataset is 78.03%. And it seems as though there is some justice in the end because the model selected by the double cross validated process matches the BIC logistic regression model once fit to the full dataset.

# Conclusion

Utilizing double cross validation allowed us to compare several models against each other. While they were all effective, some stood above the rest and exceeded expectations once it came time to fit to the full dataset. The BIC logistic regression model that used 10-fold cross validation is able to predict wins for FC Barcelona with 78% accuracy. This is important information for the team as they can focus on ball creating scoring opportunities close to goal to increase their chances of winning matches.