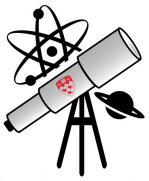
PHYSICS 321 FINAL PROJECT



Bayesian Inference for Penalty Kick Scoring Prediction

Michael Hodgins, Jordan Agathe, Minh Anh Trinh

McGill University Department of Physics

April 18, 2023

Abstract

Penalty kicks are one of the most high-pressure situations in soccer, as they can have a significant impact on the outcome of a game. Understanding the probability of a player scoring a penalty kick can provide valuable insights to coaches and players in their preparation for game situations. In this project, we aimed to model the probability of scoring a penalty kick using Bayesian Inference and a Markov-Chain Monte Carlo (MCMC) algorithm. To obtain the probability of scoring, we used logistic regression with the goal difference before the penalty kick and the goal keeper's historical save rate as predictors. We then applied MCMC algorithm to simulate the posterior distribution of the model parameters, which provided a probabilistic estimate of the player's scoring probability. Our results showed that when the goal keeper's historical save rate is 20% or lower, the probability of scoring a successful penalty kick is close to 90% or more. Conversely, if the goal keeper's save rate is 50% or higher, the ability to score a successful penalty kick drops significantly, with the probability of a failed penalty kick approaching 100%. Moreover, our analysis also revealed that the best outcome for a successful penalty kick is when the goal difference is positive and in favor of the team taking the penalty kick.

1 Introduction

Our project aims to determine the likelihood of a soccer player scoring a penalty kick. To achieve this, we used prior knowledge of the goalkeeper's historical save rate and the goal difference before the penalty kick. We adopted a statistical approach that combined linear regression, Bayesian inference and Markov-Chain Monte Carlo algorithm to simulate the posterior distribution of the model parameters. This technique enabled us to obtain a probabilistic estimate of the player's scoring probability while considering the predictor variables.

One of the key advantages of using logistic regression as a likelihood function is that it allows us to model the relationship between a binary response variable (whether or not a penalty kick is scored) and the predictor variables. This approach, coupled with Bayesian inference, helped us produce a posterior

distribution of the model parameters, which could be used to estimate the probability of scoring and to quantify the uncertainty associated with this estimate. The Markov-Chain Monte Carlo algorithm allowed us to display of our estimates of the probability of scoring by randomly sampling different values of our parameters and plotting the resulting logistic distribution. Doing this many times over allowed us to have a probabilistic view on the probability distribution

To illustrate our approach, we used the penalty kick statistics of Cristiano Ronaldo, one of the best soccer players in history, scraped from transfermarkt.us website. By leveraging Ronaldo's wealth of data, we gained insights into the probability of him scoring a penalty kick. Our project offers a practical framework that can be applied to evaluate the scoring probability of any soccer player while considering crucial predictor variables.

2 Methods

2.1 Logistic Regression

Our project focuses on investigating the factors that impact the success rate of penalty kicks in professional football. Given that the outcome of penalty kicks is binary, with either a goal or a miss, logistic regression is a suitable statistical method to model the probability of scoring a penalty kick.

To streamline our analysis, we have chosen to start with one predictor variable, which is the goalkeeper's historical save rate. We will use this variable to construct a logistic regression model that we can sample from to then calculate the probability distribution of the probability of scoring a penalty kick, where the probability of scoring is determined by the equation

$$P(\text{scoring}) = \frac{1}{1 + e^{-\beta(\text{save rate})}} \tag{1}$$

The β coefficient in this model represents the strength and direction of the relationship between the goalkeeper's historical save rate and the probability of scoring a penalty kick.

This is obviously not the correct way to perform a logistic regression analysis. The model above is incomplete but for the sake of simplicity we decided to start from an easy model and then gradually moving on to a more complicated but complete one.

We then decided to modify our model by adding a constant β value, which would allow us to better represent the relationship between the predictor variable and the probability of scoring. The revised model is expressed as

$$P(\text{scoring}) = \frac{1}{1 + e^{-\beta_1 - \beta_2(\text{save rate})}} \tag{2}$$

where β_1 represents the constant β value and β_2 represents the β coefficient for the goalkeeper's historical save rate.

In order to enhance the accuracy and completeness of our logistic regression model, we decided to incorporate a second predictor variable. We selected the goal difference immediately prior to the penalty kick as the second predictor, as it is a critical contextual factor that may impact the probability of scoring.

The modified logistic regression model now takes the form of

$$P(\text{scoring}) = \frac{1}{1 + e^{-\beta_1 - \beta_2(\text{save rate}) - \beta_3(\text{goal difference})}} \tag{3}$$

β_1 represents the constant beta value, β_2 represents the beta coefficient for the goalkeeper's historical save rate, and β_3 represents the beta coefficient for the goal difference immediately before the penalty kick.

By including the goal difference as a predictor variable, we can better account for the situational context in which the penalty kick occurs and thus more accurately predict the likelihood of scoring. This updated model will enable us to gain more comprehensive insights into the factors that impact the success rate of penalty kicks in professional football and provide a more informed basis for decision-making.

2.2 Bayesian Inference

We wanted to incorporate our existing understanding of the model parameters in our analysis. To achieve this, we used Bayesian Inference, which allowed us to integrate our prior knowledge about the model parameters into the analysis.

Specifically, we represented our prior knowledge about the βs by a linear function that spanned a range of values between our estimated minimum and maximum values for β. We assumed that every value of beta was equally likely, and we normalised it so that the total probability of all possible values of beta added up to 1.

We applied Bayes' theorem, which states that the posterior distribution is proportional to the product of the likelihood function and the prior distribution. This is expressed mathematically as

$$\text{Posterior distribution} \propto \text{Likelihood function} \times \text{Prior distribution} \tag{4}$$

We normalize our posterior distributions to ensure that the probabilities summed up to 1 and therefore the proportional sign does not matter. For the likelihood function [1], we used the formula

$$l = \sum_{k=1}^{k} [y_k * ln(p_k) + (1 - y_k) * ln(1 - p_k)] \tag{5}$$

where l is the natural logarithmic value of the likelihood function, $y_k = 1$ if the penalty was scored, and 0 if the penalty was missed. p_k is the probability of scoring given the βs and the data.

After estimating the posterior distribution using Bayesian Inference, we generated a 1D graph to visualize the distribution when the likelihood function was modeled with a logistic regression that utilized a single β coefficient. To visualize the posterior distribution for the likelihood function that employed two β coefficients,

we created a 2D density chart. However, when we attempted to analyze and plot the posterior distribution for a logistic regression model with three β coefficients, we found it to be too complex for the computer to handle in a reasonable time. Therefore, we decided to skip directly to using Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution and analyze it further.

2.3 Markov-Chain Monte Carlo

This is a method to sample parameters from a posterior distribution. It is mainly used when creating a posterior distribution would take too long as there would be too many points in parameter space to analyse completely. We used the Metropolis MCMC algorithm to analyse the posteriors. After running the MCMC algorithm, we plotted corner plots for the βs so that we could visulise their posterior distributions. We also plotted the trace plots so that we could be sure that the MCMC was working as desired and that there was good mixing. It also allowed us to get a general idea of the values of each β. We then went onto sample βs from our MCMC results, and plotted the logistic regression for a variety of different β values. This allows us to get a good view on how the probability distribution should look, and how much it varies leading us to know how uncertain it is.

3 Results and Discussion

Our first surface analysis of the probability of scoring consists of using an incomplete but simpler version of the logistic regression formula as show in equation 1, where β is the only parameter to be found.

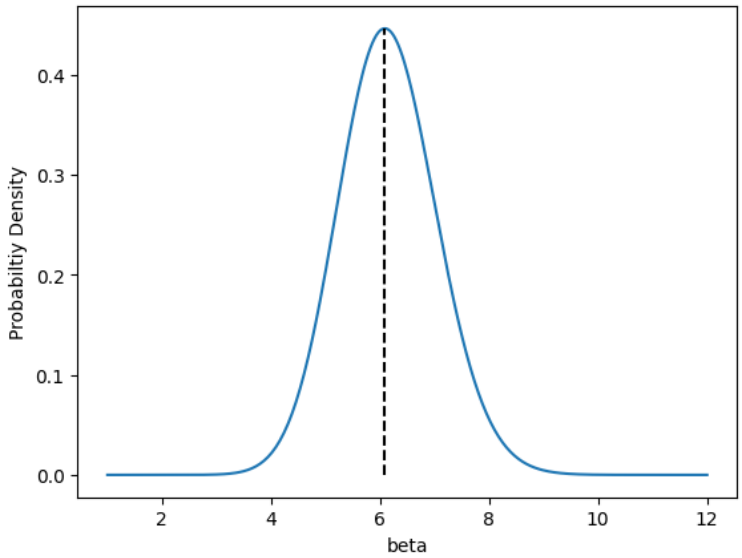
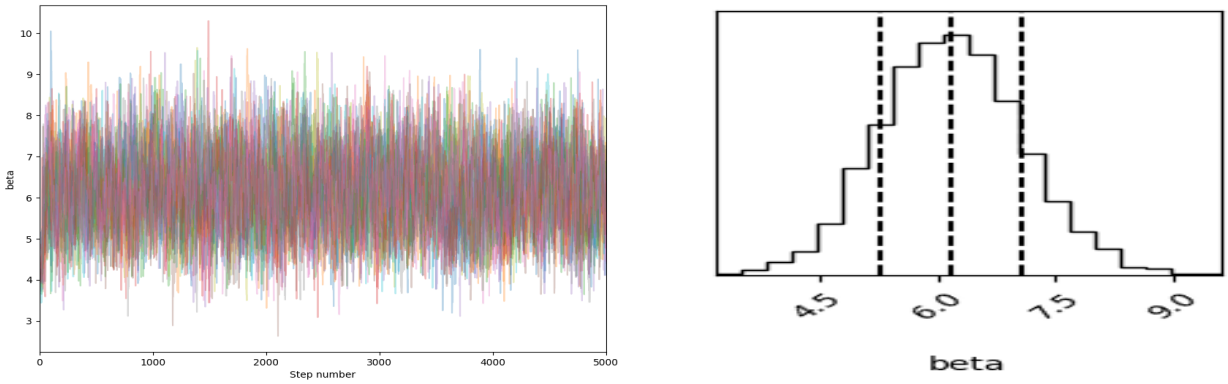


Figure 1: A simple posterior probability distribution for β obtained from Bayesian analysis of our data

As observed from Figure 1, the value of β that gives the highest probability is around 6.077. This method of estimating parameters is called the maximum likelihood estimation. These parameters will then be used to

visualize our logistic regression. Finally, using a Metropolis–Hastings algorithm, which is a type of MCMC, we obtained a trace and a corner plot (histogram-like) of β :



((a)) Trace Plot ((b)) Corner Plot

Figure 2: MCMC analysis of one parameter logistic regression

The results obtained from this are not really surprising since they agree with the Bayesian analysis but what this allows us to do is to finally plot the logistic regression curve for $P(\text{Scoring})$, the probability of scoring!

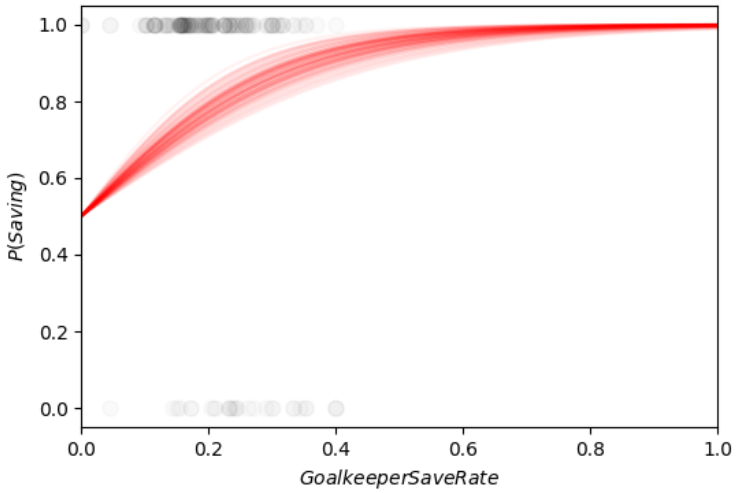


Figure 3: Logistic regression plot of our data fitted to one parameter

The red lines in Figure 3 represent the various curves fitted to equation 1 and the dots represent our data points. This plot clearly does not make much sense since you would expect that the probability of scoring would decrease with increasing goalkeeper save rate. This is because equation 1 is not a proper linear regression. There is not enough parameters to properly adjust the shape of the curve. It is like trying to fit linear data using $y = mx + c$ but you force $c = 0$. However, it's a good starting point to build upon as we'll now move to the next phase of our analysis which is to add more parameters to our model and make it more

complete.

Our new formula for our model will then be equation 2, and instead of a simple posterior plot, we used a heat map to represent the posterior of the range of the two parameters. This is because with two βs, we are now dealing with a two dimensional parameter space.

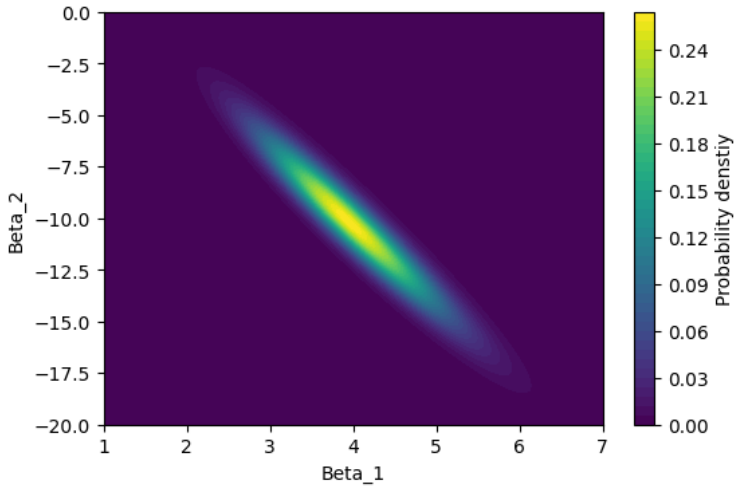


Figure 4: Heat map for β_1 and β_2 obtained from Bayesian analysis of our data

Again we can see here that this method gives us pretty good estimates for our parameters. Notice how the value of our parameter associated with the goalkeeper save rate, β_2, changed from around 6 to around -10 when we added the intercept parameter β_1. Also note how correlated the two βs are. The fact that there is not a regular circle and that we see more of a line indicates this. Moving on to the MCMC analysis and the logistic regression plot of this 2 parameter model :

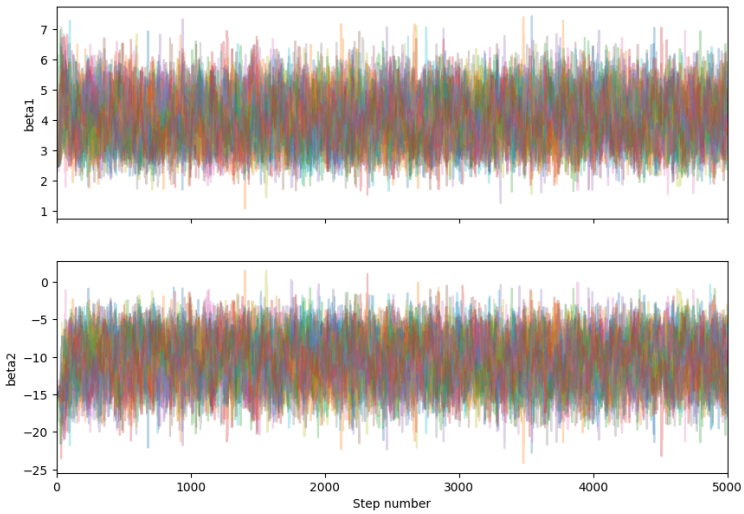


Figure 5: Trace Plots of β_1 and β_2

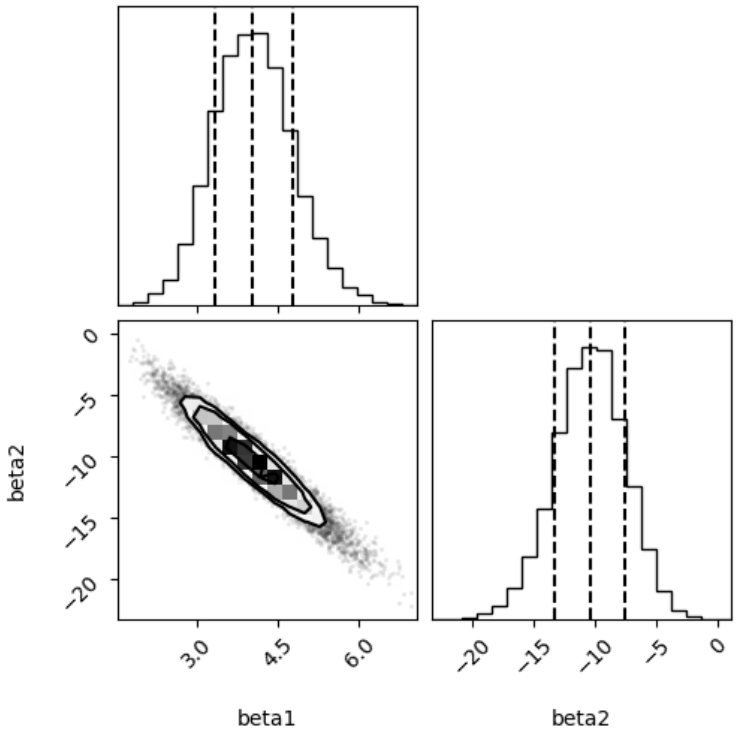


Figure 6: Contour Plots for β_1 and β_2

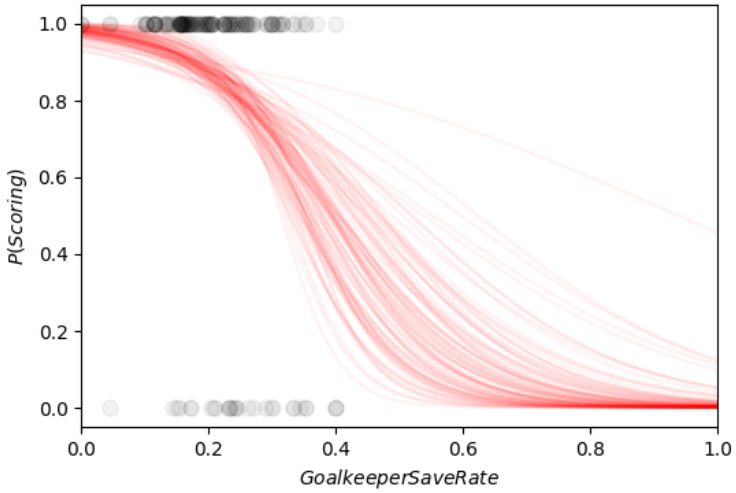


Figure 7: Logistic regression plot of our data fitted to two parameters; P(Scoring) against Goalkeeper save rate

This plot already makes much more sense now! It agrees with our prior knowledge of how penalties work but this is not the best we can do. We'll step it up once again by adding an additional parameter but this time we will couple it with an additional predictor variable, the goal difference between the kicker's team and the opposing goalkeeper's team right before the penalty is taken. We're also going to jump straight the MCMC

analysis this time to get right to the crux of our investigation. It would also take too long to completely explore the three dimensional parameter space to get a complete posterior distribution.

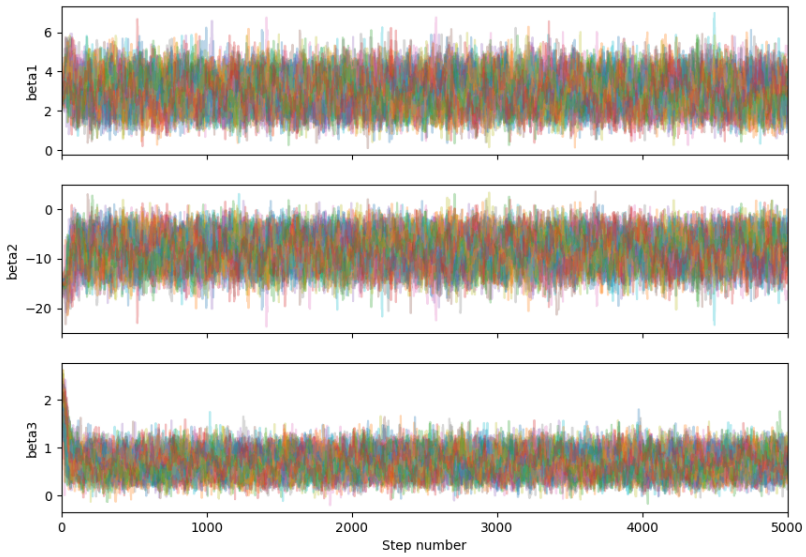


Figure 8: Trace Plots of β_1, β_2 and β_3

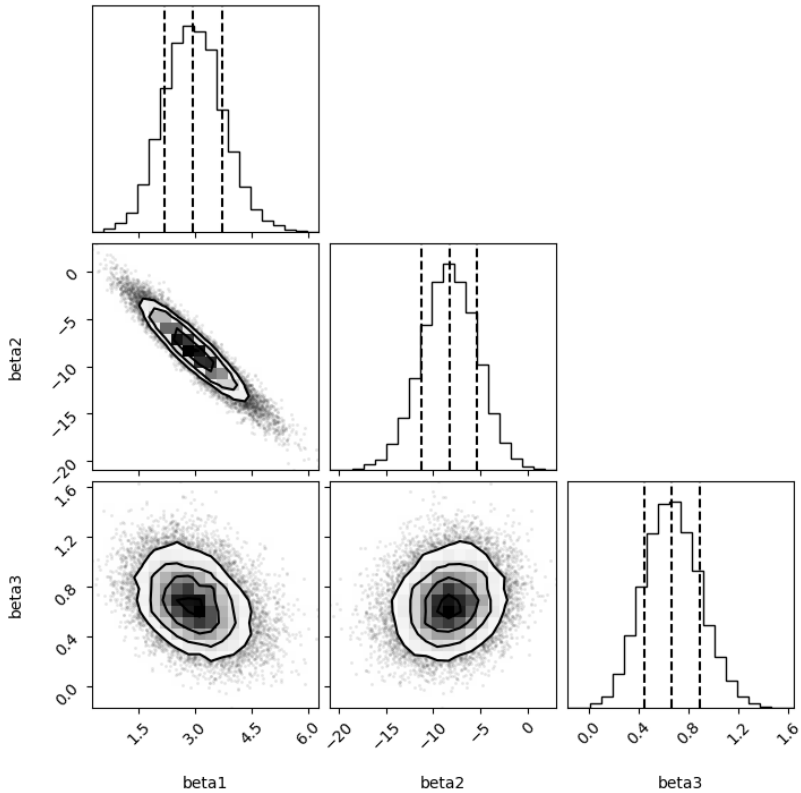


Figure 9: Contour Plots for β_1, β_2 and β_3

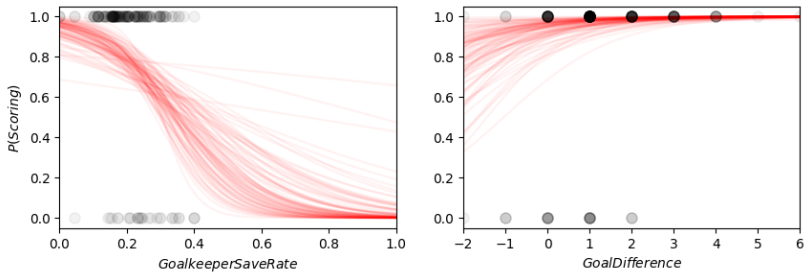


Figure 10: Logistic regression plots with $P(\text{Scoring})$ against Goalkeeper save rate and Goal difference

Once again we can observe that our parameter values changed when a new one was added, although this time the change was not very significant. One reason could be because of random fluctuations in the MCMC algorithm, but it also could be that the new predictor variable does not correlate with the other two β values as much as they correlated with eachother. Notice also how the β_3 value does not strongly influence the chances of scoring as much as our first predictor variable, β_2, does. This explains why β_3 is relatively small compared to the other parameters as can be seen in Figure 9. However, when the game has already been won, and the opposing side is demoralised, there is a greater chance that the penalty is scored. This is because the β_3 value is positive but small, meaning that a higher goal difference leads to a greater chance of scoring. Again, this lines up with our intuition.

We also tried to analyze our data using a 3D interactive grid and even though it might be harder to see the trend, it still is visible when looking at how pale the red colour gets depending on the region observed.

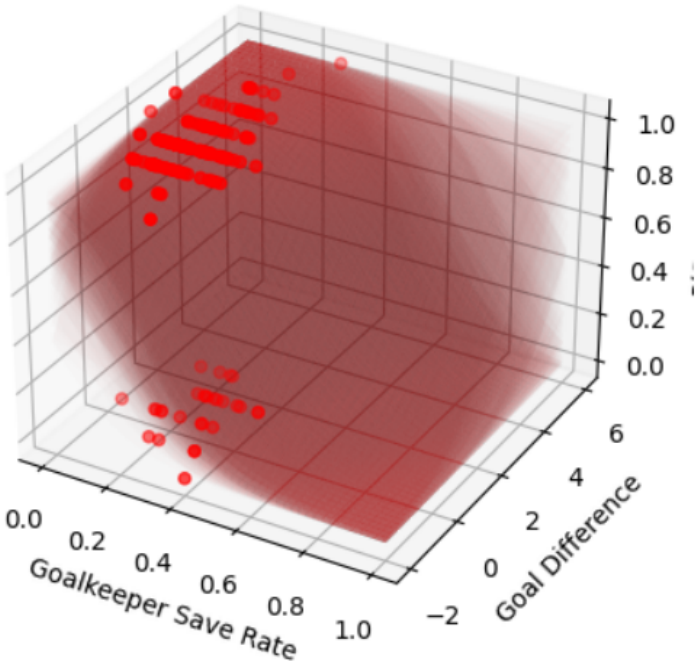


Figure 11: Logistic regression plots with $P(\text{Scoring})$ against Goalkeeper save rate and Goal difference

One additional piece of insight that might help us interpret this last grid plot is that the professional soccer player chosen to take our data from is none other than Cristiano Ronaldo, an excellent penalty taker, even among professionals. This suggests that some predictor variables (such as the Goal difference) which might have an incidence on the probability of scoring, do not have much of an effect on Ronaldo due to him being such an outlier.

4 Conclusions

Our data shows that the goalkeeper save rate is a high predictor for the probability of scoring a penalty, while the goal difference is a small predictor of scoring a penalty. The β_2 value associated with the goalkeeper save rate data is about -10, far from 0 compared to about 0.7 for β_3. Therefore we can say that the save rate data has a large effect on the logistic regression curve. It is also negative, implying that a greater goalkeeper save rate makes it less likely that a penalty will be scored; lining up with our intuition.

One area that we could have explored more in detail would be how the effect of those predictor variables and their associated parameter values change for different players. We would assume that for an average professional, the goal difference might have more of an impact on the probability of converting the penalty or not. Such deeper insight would have given us a more profound understanding of the mental aspect of penalty taking which is, we believe, very present in today's game.

References

[1] Wikipedia. *Logistic regression.* (accessed: 2023-04-17) (https://en.wikipedia.org/wiki/Logistic$_regression$). 2023.

5 Contributions

Minh and Michael contributed towards the data scrapping. Michael and Jordan and Minh all contributed to the coding and data analysis, and all three of us also contributed to the final paper written.