Robert L. Prattico
robertprattico@gmail.com
(514) 260-4247
Monday, October 30, 2023

Cincinnati Red Baseball Analytics Trainee

Second Round Assessment Write-Up

Initial Solution Ideas

When I first reviewed the dataset, a variety of approaches to solve the use case crossed my mind. It was clear that this problem entailed building a binary classification model that can yield the probability of a pitch belonging to a the “dew point unaffected” or “dew point affected” class. Given that there was no clear target variable, I first thought to use an unsupervised machine learning algorithm to group the pitches into distinct clusters. My initial thought was that a model like K-Means could be trained to distinguish the pitches that were and were not affected by dew point without the need for a clear label indicating so. I soon realized that an unsupervised learning model would not be able to calculate the probability of the effect of dew point. Furthermore, we could not tell for sure if the clusters it built were definitely and solely related to the effect of dew point.

I identified the PLATE_X and PLATE_Z features as being the most important in determining the presence of a dew point effect; they relate to the final destination of the ball plotted on a Cartesian plane after factoring in all other elements of its trajectory. However, it wouldn’t be possible to build a classification model with two target variables. I would either have to train two separate models (one for the x-axis, one for the y-axis), or find a way to combine the two features into one.

I gave considerable thought to the first option. I considered converting PLATE_X and PLATE_Z from Cartesian coordinates to polar coordinates, but that would not have solved the issue of having two target variables. Instead, I built two linear regression models: one for the horizontal placement of the ball above home plate (PLATE_X) and another for its vertical placement above home plate (PLATE_Z). I fed the data into the algorithms to see what horizontal and vertical positions it predicted. I then compared the predicted values for PLATE_X and PLATE_Z to their actual values, essentially comparing where the ball was predicted to have been placed on the Cartesian plane above home plate to where it actually was placed. I used hypothesis testing to estimate the probability that dew point affected. The null hypothesis was that dew point did not affect the pitch, while the alternate hypothesis did. I calculated the average and standard deviation of horizontal and vertical positions for every pitch type for both left and right-hand throws. This would allow me to calculate a Z-score that was appropriate for that specific type of pitch. That Z-score would then lead to a p-value that represented the probability the pitch was affected by dew point.

While this approach initially seemed reasonable to me, I realized that it had one major flaw: it was based on the notion that the predicted value of PLATE_X and PLATE_Z was more correct than their actual values. Fundamentally, this assumption was wrong. So, I scraped this method and went back to the drawing board.

Final Solution

Having tried different approaches that maintained the PLATE_X and PLATE_Z, I explored a solution that based on the combination of these two features. I knew that a logistic regression model would be the best algorithm to yield probabilities of a classes, I just needed to find a way constructed those classes from the two target variables. I devised the "Is Outlier?" class to accomplish just that.

The "Is Outlier?" labels a pitch as having been affected by dew point based on whether or not it lies outside of the typical placement of a pitch. An important aspect of the "Is Outlier?" class is that it takes into the qualitative aspects of the pitches that are not reflected in the data. Different types of pitches will land in different ranges of the plane, so to deduce outliers for all the pitches together would erroneous. Similarly, different pitchers would have different abilities and style of throwing each pitch, so the identification of outliers needed to consider this qualitative element as well. Once the dataset was segmented by pitcher and pitch type, I then highlighted the outlier by calculating the interquartile range for both PLATE_X and PLATE_Z in each segment. I identified the pitches that lied outside the ranges by assigning 1 in the "Is Outlier?" class. At this point, I took the assumption that, since all other possible variable elements of the pitch were accounted for, these outliers occurred because of unconventional dew point levels. I may have missed dew point-affected pitches that lied within the interquartile ranges, but unfortunately there was no spot them with the data provided.

Having constructed a target variable in the "Is Outlier?" class, I was finally able to move to training the logistic regression model. There were very few instances of outliers, so I had to remedy the issue of under sampling in the dataset to avoid any bias in the model. I dummified categorical variables, standardized all values to be on the same measurement scale, and performed hyperparameter tuning to find the optimal configuration for the logistic regression model. I could have also performed a feature selection to remove irrelevant predictors, but the model performed strongly without it. In the end, my logistic regression model had a very good accuracy score of 96%. Finally, I applied the model to predict probabilities of a 1 in the "Is Outlier?" field to obtain my final submission.

While the model I built was strong, there are still a couple of flaws that could have impacted on its performance. For one, the methodology I applied to engineer the "Is Outlier?" class may have neglected some key intricacies in the game of baseball, the external conditions the pitch was thrown in, or in the minute details of the ball's trajectory. These intricacies, not the effect of dew point, may have been the reason why the pitches were identified as "outliers". Furthermore, the approach I took to resolve the issue of under sampling could have misled the model; training a model on such a high quantity of synthetic data can make it inadequate when applied to real data.

All in all, I greatly enjoyed completing this project. Being able to apply my skills and knowledge of data science and storytelling to solve a real-world use case in sports analytics was a great experience. I thank you for the opportunity, for your time, and for your consideration. I look forward to receiving your feedback.

Regards,

Robert L. Prattico