Parametter

Các thông số: Trong RapidMiner, đây là những thiết lập các giá trị kiểm soát và ngưỡng mà một nhà điều hành sẽ sử dụng để thực hiện công việc của mình. Đây có thể là các tên thuộc tính và vai trò của hệ thống, hoặc các thuật toán khai thác dữ liệu mong muốn sử dụng.

Null: sự thiếu vắng dữ liệu. Giá trị là không được ghi lại, không rõ, hoặc không xác định.

Comma Separated Values ​​(CSV): Một định dạng văn bản phổ biến cho các bộ dữ liệu sử dụng dấu phẩy để phân cách các thuộc tính.

Trong phân tích bộ dữ liệu của mình, Richard đã phát hiện ra rằng hoạt động của khách hàng trong các lĩnh vực truyền thông kỹ thuật số và sách, và hoạt động của họ với các thiết bị điện tử bán trên trang web của công ty mình, dường như có rất nhiều điểm chung với khi một người mua một eReader.

eReader\_Adoption: thuộc tính này chỉ tồn tại trong tập dữ liệu huấn luyện.

+Những người mua trong vòng một tuần phát hành của sản phẩm :"Innovator. +Những người mua sau tuần đầu tiên nhưng trong những tuần thứ hai hoặc thứ ba :”Hội Quán sớm.

+Những người mua sau ba tuần nhưng trong hai tháng đầu năm là "Phần lớn sớm '. +Những người mua sau khi hai tháng đầu năm là "Đa số cuối. Thuộc tính này sẽ phục vụ như nhãn hiệu.

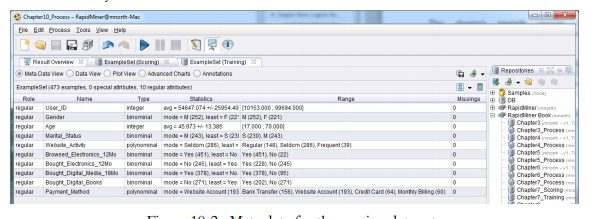
DỮ LIỆU CHUẨN BỊ ví dụ bao gồm hai bộ dữ liệu: Chapter10DataSet\_Training.csv

Chapter10DataSet\_Scoring.csv

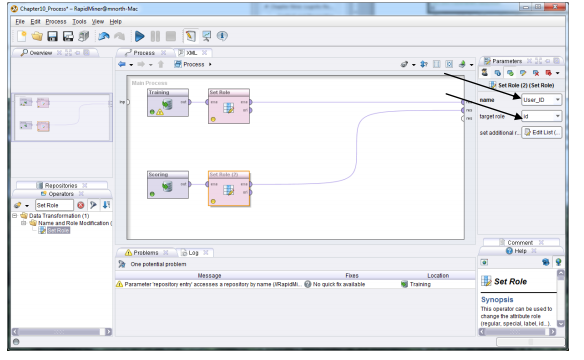
hoàn tất các bước sau đây:

1) import cả hai bộ dữ liệu vào kho lưu trữ RapidMiner. Bạn không cần phải lo lắng về các loại dữ liệu thuộc tính bởi vì các nhà điều hành Tree Quyết định có thể xử lý tất cả các loại dữ liệu. Lưu chúng trong kho lưu trữ với tên mô tả.

2) Kéo và thả cả hai bộ dữ liệu vào một cửa sổ trình chính mới.Chạy mô hình để kiểm tra dữ liệu và làm quen với các thuộc tính.

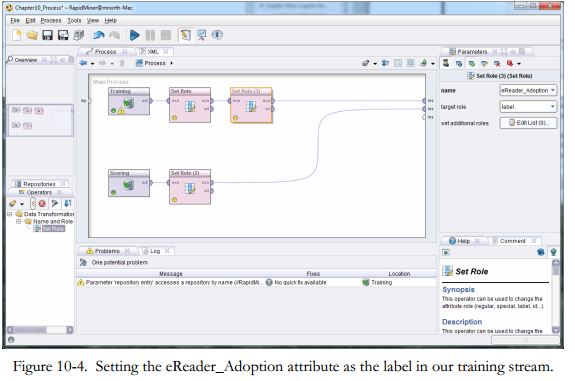


3) User\_id không tham gia trong việc gán nhãn lớp hay nói cách khác nó là thuộc tính không tiên đoán, nó là một thuộc tính sinh ngẫu nhiên để nhận biết giữa các KH. Có 2 cách xử lý thuộc tính không tiên đoán: loại bỏ thuộc tính bằng cách chọn Select Attributes operator. Hoặc cả 2 bảng dữ liệu trên phải được cài đặt như sau: trong khu vực thông số ở phía bên tay phải của màn hình, thiết lập vai trò của các thuộc tính user\_id 'id'. Điều này sẽ để lại các thuộc tính trong các tập dữ liệu trong mô hình, nhưng nó sẽ không xem xét các thuộc tính như một yếu tố dự báo cho thuộc tính nhãn.

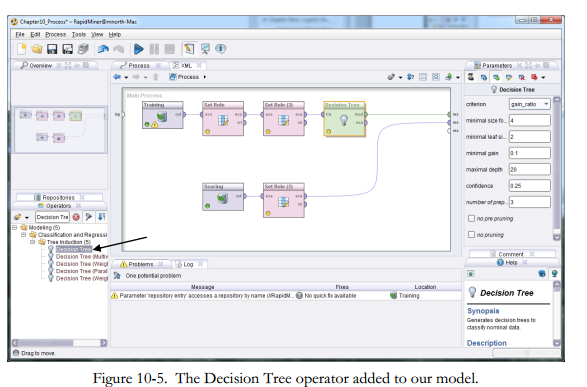


4). Vai trò: làm cho mỗi bản ghi dễ dàng kết nối trở lại với từng người sau đó, có thể dễ dàng biết ai để liên hệ trong giai đoạn triển khai các dự án khai thác dữ liệu.

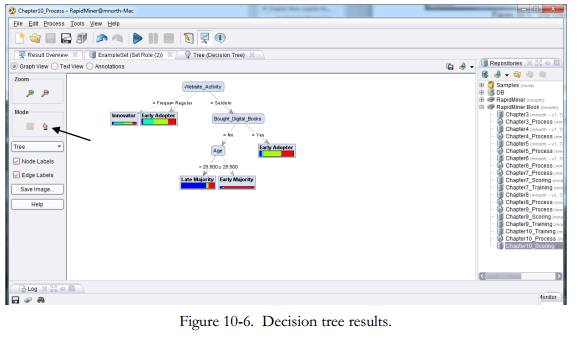
Trước khi thêm một toán tử cho cây quyết định, vẫn cần thêm một bước chuẩn bị dữ liệu. Với hi vọng tìm ra một luồng tập huấn để cung cấp một thuộc tính nhãn. Trong ví dụ này chúng tôi muốn dự đoán adopter thế hệ tiếp theo khách hàng eReader của nhóm Richard có thể có.Vì vậy, nhãn hiệu của chúng tôi sẽ là e Reader\_Adoption.



5). Tiếp theo, tìm kiếm trong tab Operators “Decision Tree”. Chọn toán tử cho câu quyết định cơ bản và thêm nó vào tập huấn luyện.



6) Chạy mô hình và chuyển sang tab “Decision Tree” trong quan điểm kết quả. +> cây sơ bộ:



7). Trong hình 10-6, nút là hình bầu dục màu xám, đó là các thuộc tính phucjvuj cho việc dự đoán thuộc tính nhãn hiệu. Lá là những điểm cuối có nhiều màu, cho chúng ta thấy sự phân bố các loại danh mục thuộc tính nhãn hiệu theo các chi nhán của cây đến điểm lá.

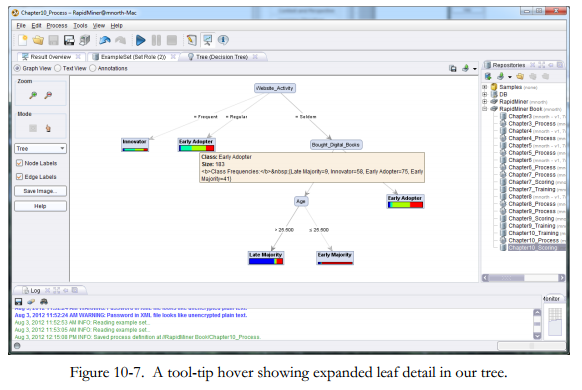
Chúng ta có thể nhìn thấy trong cây này Website\_Activity là yếu tố dự báo tốt nhất có hay không một khách hàng sẽ mua eReader mới của công ty . Nếu hoạt động của người frequent or regular, chúng ta thấy rằng họ có thể sẽ là một Innovator hoặc Early Adopter,tương ứng . Tuy nhiên, nếu họ ít khi sử dụng trang web ,thì mua sách kỹ thuật số trở thành dự đoán tốt nhất tiếp theo của thể loại thông qua đọc sách điện tử của họ . Nếu họ đã không mua sách kỹ thuật qua các trang web trước đó , thì Age là một thuộc tính tiên đoán tạo thành một nút , với folks trẻ áp dụng sớm hơn những người lớn tuổi . Điều này được nhìn thấy trên các nhánh

cho hai lá đến từ các nút Tuổi trong hình 10-6. Những người ít khi sử dụng

trang web của công ty , chưa bao giờ mua sách kỹ thuật số trên trang web, và lớn tuổi hơn 25 ½ có nhiều khả năng đất trong các thể loại Late Majority, trong khi những người có cùng một cấu hình nhưng dưới 25 ½ được dự đoán Early Majority .

Trong ví dụ này, bạn có thể xem cách bạn đọc các nút , lá và nhãn chi nhánh khi bạn di chuyển qua các cây .

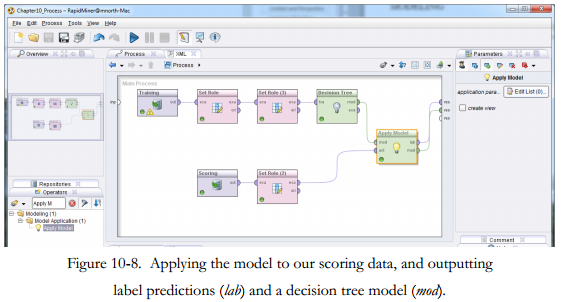
Trước khi trở về quan điểm thiết kế , dành vài phút để thử một số trong những công cụ ở bên trái bên tay phải màn hình . Kính lúp có thể giúp bạn nhìn thấy cây của bạn tốt hơn , lan rộng ra hoặc nén các nút và lá để tăng cường khả năng đọc hoặc để xem chi tiết của một cây lớn cùng một lúc. Ngoài ra, hãy thử sử dụng biểu tượng ' hand ' dưới chế độ (xem mũi tên trên hình 10-6 ) . Điều này cho phép bạn bấm và giữ trên từng lá hoặc các nút và kéo chúng xung quanh để nâng cao khả năng đọc cây của bạn . Cuối cùng , hãy thử lơ lửng chuột qua một trong những lá trong cây . Trong hình 10-7 , chúng tôi thấy một tool -tip hộp di chuột cho thấy chi tiết của lá này . Mặc dù dữ liệu huấn luyện của chúng tôi sẽ dự đoán rằng người dùng "regular" trang web Early Adopters, mô hình không phải là 100 % dựa trên dự đoán đó. Trong di chuột , chúng ta đọc thấy trong tập dữ liệu huấn luyện , 9 người phù hợp với hồ sơ này là Late Adopters, 58 are Innovators, 75 are Early Adopters and 41 are Early Majority . Khi chúng tôi đến giai đoạn đánh giá , chúng ta sẽ thấy không chắc chắn trong dữ liệu của chúng tôi sẽ chuyển thành tỷ lệ phần trăm conf, tương tự như những gì chúng ta đã thấy trong Chương 9 với hồi quy logistic .



MÔ HÌNH

8) Trở về quan điểm thiết kế. tìm kiếm trong các tab Operator và Apply Model

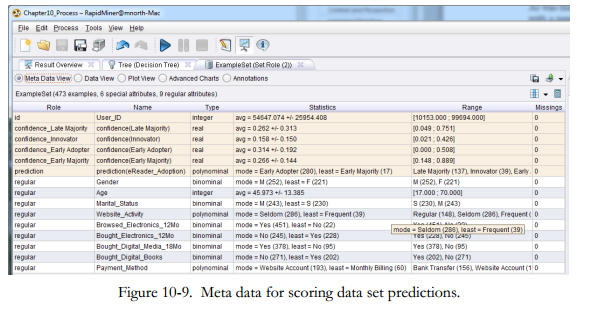
operator, đưa training and scoring streams cùng nhau. Đảm bảo rằng cả hai lab và mod cổng được kết nối với cổng độ phân giải để tạo ra kết quả đầu ra mong muốn.



9). Chạy mô hình. Bạn sẽ thấy kết quả quen thuộc cây vẫn giữ nguyên như trong hình10-6. Nhấp vào tab ExampleSet cạnh tab Tree. Cây của chúng tôi đã được

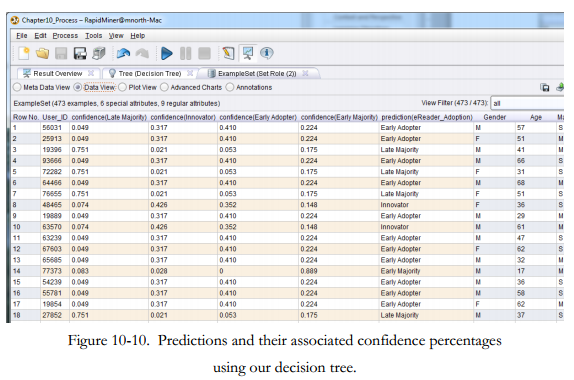
áp dụng cho dữ liệu score. Như là trường hợp với hồi quy logistic, thuộc tính conf

đã được tạo ra bởi RapidMiner, cùng với một thuộc tính dự đoán.

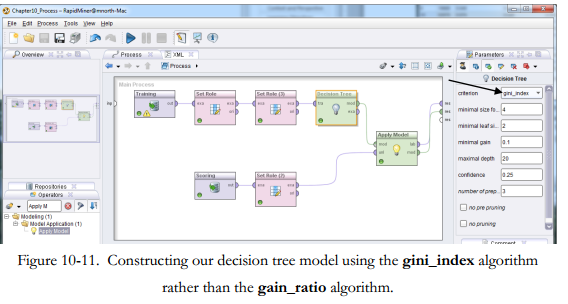


10). Chuyển sang Xem dữ liệu bằng cách sử dụng nút radio. Chúng ta thấy trong hình 10-10 dự đoán cho nhóm adoption của từng khách hàng, cùng với tỷ lệ phần trăm conf cho mỗi dự đoán.

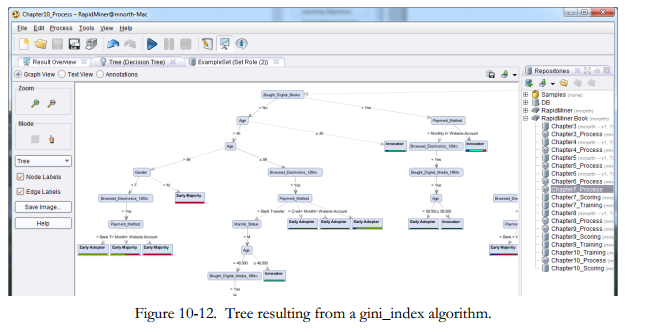
Không giống như các ví dụ hồi quy trong chương trước, có bốn thuộc tính conf, tương ứng với bốn giá trị có thể có trong nhãn (eReader\_Adoption). Chúng tôi giải thích những cách mà chúng ta đã làm với các mô hình khác mặc dù, tỷ lệ phần trăm lên đến 100%, và dự đoán là thể loại nào mang lại tỷ lệ phần trăm conf cao nhất. RapidMiner rất dễ (nhưng không phải 100%) tin rằng người 77373 (Row 14, Hình 10-10) là có được một thành viên của đại đa số ban đầu (88,9%). Mặc dù một số không chắc chắn, RapidMiner là hoàn toàn chắc chắn rằng người này sẽ không adopter đầu tiên (0%).



11). Chúng ta đã bắt đầu đánh giá kết quả mô. Chắc chắn một số thuộc tính khác cũng có thể là những tiên đoán trong tự nhiên. Hãy nhớ rằng CRISP-DM là theo chu kỳ trong tự nhiên, và trong một số mô hình kỹ thuật, đặc biệt là những người có ít dữ liệu có cấu trúc, một số thử nghiệm lỗi có thể tiết lộ những điều thú vị hơn trong dữ liệu. Chuyển về quan điểm thiết kế, nhấp chuột vào các nhà Operator Tree decision, và trong khu vực tham số, thay đổi thông số của tiêu chí ' criterion ' thành gini\_index ', như thể hiện trong hình 10-11.



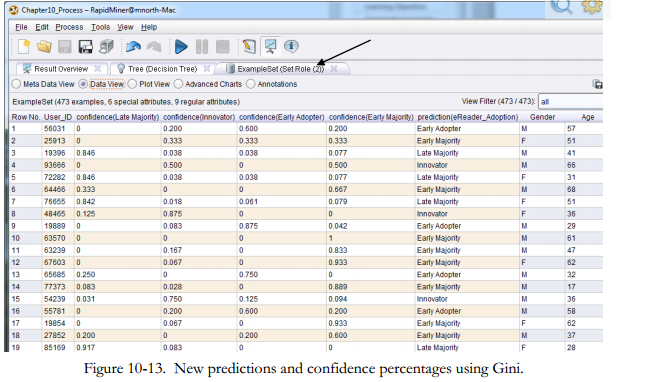
Chạy lại mô hình và nhìn:



Chúng ta thấy trong cây này có nhiều chi tiết hơn , granularity hơn trong việc sử dụng các thuật toán Gini như tham số cho cây quyết định. Chúng tôi cũng có thể sửa đổi các cây bằng cách quay lại xem thiết kế và thay đổi số lượng tối thiểu các mặt hàng để tạo thành một nút ( kích thước cho phân chia ) hoặc kích thước tối thiểu cho một chiếc lá . Thậm chí chấp nhận giá trị mặc định cho những thông số mặc dù, chúng ta có thể thấy rằng các thuật toán Gini đơn là nhiều nhạy cảm hơn là các thuật toán Gain Ratio trong việc xác định các nút lá . Mất vài phút để khám phá xung quanh mô hình cây mới này . Bạn sẽ thấy rằng nó là rộng lớn , và rằng bạn sẽ sử dụng cả hai công cụ Zoom và chế độ để xem tất cả . Bạn sẽ thấy rằng hầu hết

các thuộc tính dự đoán hiện đang được sử dụng , và độ chi tiết mà Richard có thể xác định khả năng loại áp dụng của từng khách hàng lớn hơn nhiều.

Chuyển sang tab ExampleSet trong Xem dữ liệu. Thay đổi thuật toán cơ bản, độ conf cũng được thay đổi.



Lấy người trên hàng 1 (ID 56.031) là một ví dụ. Trong hình 10-10, người này đã được tính toán là có ít tỷ lệ phần trăm nhất bất kỳ một trong bốn loại adopter. Theo thuật toán Gain Ratio , chúng tôi đã 41% chắc chắn anh ấy sẽ early adopter, , nhưng gần 32% chắc chắn ông cũng có thể innovator. Nói cách khác, chúng tôi cảm thấy conf ông sẽ mua các eReader early, nhưng chúng tôi không chắc chắn cách ban đầu. Có lẽ đó là quan trọng đối với Richard, có thể không. Ông sẽ phải quyết định trong giai đoạn triển khai. Nhưng có lẽ sử dụng Gini, chúng tôi có thể giúp anh ta quyết định. Trong hình 10-13, cùng người đàn ông này đã được thể hiện để có một cơ hội 60% được early adopter và chỉ có 20% cơ hội để trở thành innovator.

Đôi khi mức độ tự tin trong một dự đoán thông qua một cây quyết định rất cao mà một thuật toán cơ bản nhạy cảm hơn sẽ không làm thay đổi giá trị dự đoán của một quan sát tại tất cả.

TÓM TẮT CHƯƠNG

Cây quyết định là mô hình tiên đoán tuyệt vời khi các thuộc tính mục tiêu là phân loại trong tự nhiên, và khi tập dữ liệu là các loại hỗn hợp. Mặc dù của chương này tập hợp dữ liệu không chứa bất kỳ ví dụ , cây quyết định là tốt hơn so với nhiều phương pháp tiếp cận thống kê có trụ sở tại xử lý các thuộc tính có giá trị bị mất hoặc không phù hợp mà không phải là cây xử lý ra quyết định sẽ làm việc xung quanh những dữ liệu đó và vẫn tạo ra kết quả có thể sử dụng .

Cây quyết định được làm bằng các nút lá ( nối với nhau bằng các mũi tên chi nhánh nhãn) , đại diện cho các thuộc tính dự báo tốt nhất trong một tập hợp dữ liệu. Các nút và lá dẫn đến tỷ lệ phần trăm sự tự tin dựa trên các thuộc tính thực tế trong tập dữ liệu huấn luyện , và sau đó có thể được áp dụng cho dữ liệu điểm cấu trúc tương tự để tạo ra các dự báo cho các quan sát điểm. Cây quyết định cho chúng tôi biết những gì được dự đoán , làm thế nào chúng ta có thể tự tin trong dự đoán , và làm thế nào chúng tôi đến dự báo . 'Như thế nào chúng tôi đến " phần sản lượng cây quyết định được thể hiện trong một giao diện đồ họa của cây .

Chương 11: MỤC TIÊU HỌC TẬP

Sau khi hoàn thành việc đọc và bài tập trong chương này, bạn sẽ có thể:

 Giải thích một mạng lưới thần kinh được, làm thế nào nó được sử dụng và những lợi ích của việc sử dụng nó.

 Nhận định dạng cần thiết cho dữ liệu để thực hiện thần kinh khai thác dữ liệu mạng.

 Xây dựng một mạng lưới dữ liệu khai thác mô hình thần kinh trong RapidMiner sử dụng tập huấn tập dữ liệu.

 Giải thích kết quả đầu ra của mô hình và áp dụng chúng vào một dữ liệu điểm thiết lập để triển khai mô hình.

Chúng tôi sẽ giúp Juan xây dựng một mạng lưới thần kinh , mà là một phương pháp khai thác dữ liệu có thể dự đoán loại hoặc phân loại theo cách tương tự như cây quyết định làm , nhưng mạng lưới thần kinh được tốt hơn trong việc tìm kiếm sức mạnh của kết nối giữa các thuộc tính.Các thuộc tính:

 Player\_Name : Đây là tên của người chơi. Trong giai đoạn chuẩn bị dữ liệu của chúng tôi , chúng tôi sẽ thiết lập vai trò của mình để 'id' , vì nó không phải là tiên đoán trong bất kỳ cách nào , nhưng quan trọng là phải giữ trong dữ liệu của chúng tôi thiết lập để cho Juan có thể nhanh chóng đưa ra khuyến nghị của mình mà không cần phải phù hợp với các dữ liệu về tên của các cầu thủ sau . (Lưu ý rằng các tên trong bộ dữ liệu của chương này được tạo ra bằng cách sử dụng máy phát điện tên ngẫu nhiên . Họ hoàn toàn ngẫu nhiên và bất kỳ tương tự cho những người thực sự là ngoài ý muốn và hoàn toàn conincidental . )

 Position\_ID : Đối với nhóm lượt Juan thể dục thể thao , có 12 vị trí có thể. Mỗi người đều biểu diễn như là một số nguyên từ 0 đến 11 trong các bộ dữ liệu .

 Shots : Đây là tổng số các bức ảnh, hoặc cơ hội ghi bàn mỗi người chơi mất trong mùa giải gần đây nhất của họ.

Make : Đây là lần số lượng các vận động viên ghi bàn khi chụp trong thời gian nhất mùa giải gần đây.

 Personal\_Points : Đây là số điểm các vận động viên cá nhân ghi trong nhất mùa giải gần đây.

 Total\_Points : Đây là tổng số điểm các vận động viên góp phần ghi trong

mùa giải gần đây nhất. Trong nhóm lượt Juan thể dục thể thao , số liệu này được ghi lại chỉ một vận động viên góp phần chấm điểm . Nói cách khác, mỗi lần một điểm vận động viên một quan điểm cá nhân , tổng số điểm tăng của họ từng người một, và mỗi khi một vận động viên góp phần một điểm đồng đội , tổng số điểm của họ tăng một là tốt.

 Assists : Đây là một số liệu thống kê cho thấy phòng thủ số lần các vận động viên đã giúp đỡ đội có được bóng từ các nhóm đối lập trong mùa giải gần đây nhất.

 Concessions : Đây là số lần chơi của vận động viên trực tiếp gây ra các opposing

đội thừa nhận một lợi thế tấn công trong mùa giải gần đây nhất.

 Blocks : Đây là số lần các vận động viên trực tiếp và độc lập bị chặn

nhóm đối lập bị bắn trong mùa gần đây nhất.

 Block\_Assists : Đây là số lần một vận động viên cộng tác với một đồng đội để

chặn bắn của đối phương trong mùa giải gần đây nhất. Nếu được ghi lại như một khối hỗ trợ , hai hoặc nhiều người chơi phải có được tham gia . Nếu chỉ có một cầu thủ bị chặn bắn, nó được ghi nhận như là một khối . Kể từ khi bề mặt chơi lớn và các cầu thủ trải ra, nó là nhiều hơn khả năng cho một vận động viên để ghi lại một khối hơn cho hai hoặc nhiều hơn để ghi lại khối hỗ trợ.

 Fouls : Đây là số lần , trong mùa giải gần đây nhất, các vận động viên cam kết sau pha phạm lỗi . Vì tắc nghẽn các nhóm khác cung cấp cho họ một lợi thế , thấp hơn này số lượng, tốt hơn hiệu suất của vận động viên cho đội bóng của mình.

 Years\_Pro : Trong tập dữ liệu huấn luyện , đây là số năm các vận động viên đã đóng tại trình độ chuyên môn . Trong tập dữ liệu điểm , đây là số năm kinh nghiệm

vận động viên , bao gồm cả năm như là một chuyên nghiệp nếu có , và nhiều năm trong tổ chức , cạnh tranhgiải đấu nghiệp dư.

 Career\_Shots : Điều này tương tự như các thuộc tính Shots , ngoại trừ đó là tích lũy cho toàn bộ sự nghiệp của vận động viên. Tất cả các thuộc tính sự nghiệp là một nỗ lực để đánh giá khả năng của người thực hiện nhất quán theo thời gian.