# **DSP305x REPORT**

Đề bài: “We will build a predictive model that predicts the three fastest horses (in any order) in a race. Your customer requires criteria for all races from 2018 to now with the true positive horses at least 49.3% (i.e. if you have 600 fastest horses in 200 races, you must predict 294 horses correctly)”.

# **I. Exploratory Data Analysis**

## 1. Business Understanding:

- Hiểu về hệ thống đua ngựa ở Nhật Bản: Đua ngựa ở Nhật Bản là một môn thể thao cưỡi ngựa phổ biến, với hơn 21.000 cuộc đua ngựa được tổ chức mỗi năm. Có ba loại đua diễn ra ở Nhật Bản – đua đường phẳng, đua có chướng ngại vật và đua Ban'ei (còn được gọi là đua nháp).

Ở Nhật Bản, đua ngựa được tổ chức bởi một hiệp hội có tên là JRA( Japan Racing Association) và NAR( National Association of Racing). JRA chịu trách nhiệm tổ chức các sự kiện đua ngựa tại mười trường đua lớn ở các khu vực đô thị, trong khi NAR chịu trách nhiệm tổ chức các sự kiện đua ngựa địa phương trên khắp Nhật Bản. Hệ thống quản lý đua ngựa này chỉ có ở Nhật Bản.

Các cuộc đua cá cược hàng đầu của Nhật Bản diễn ra vào mùa xuân, mùa thu và mùa đông; cuộc đua hàng đầu là Japan Cup.

Các giải đấu chính có thể kể tên như: Tháng 2 có giải February Stakes (sân đất dài khoảng 1600m), tháng 3 có Fukuyru Stakes (sân đất 1800m), tháng 4 có Oka Sho (sân cỏ 1600m), tháng 5 có Tokyo Yushun( Turf 2400m), tháng 11 có giải đấu nổi tiếng là Japan Cup (sân cỏ 2400m).

Các trường đua lớn có thể kể đến như: Tokyo Racecourse( Japan Cup được tổ chức ở đây), Morioka Racecourse, Nakayama Racecourse.

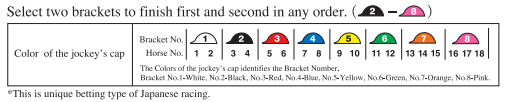
Các loại ngựa: Sẽ có các cuộc đua chỉ cho ngựa đực, ngựa cái hoặc thiến, sau đó có thể giải đấu sẽ xét về tuổi tác như dưới 3 tuổi và trên 3 tuổi..vv

Đối với người chơi sẽ có các cách cá cược như:

. Win: Cược con ngựa nào sẽ về nhất. Ví dụ con ngựa có số hiệu [3].

. Place: Cược bộ các con ngựa về nhất, nhì, ba như ( [3][14][16]). Nếu chỉ có khoảng 5-7 ngựa đua, thì sẽ chỉ được chọn bộ 2 ngựa đua về nhất và về nhì ([3][14]). Không quan trọng thứ tự.

. Bracket quinella: Cược 2 cặp về nhất nhì, không quan trọng thứ tự (cặp số hiệu 2 và 8)



Bracket quinella

. Quinella: Cược 2 con về nhất nhỉ, không quan trọng thứ tự ([3][14])

. Quinella place: Cược cặp (1st, 2nd) hoặc (1st, 3rd) or (2nd, 3rd) , không quan trọng thứ tự (3-16 or 3-14 or 14-16)

. Trio: Cược 3 con về 1st, 2nd, 3rd,, không quan trọng thứ tự.

. Exacta: Cược 2 con về 1st, 2nd, theo thứ tự.

. Trifecta: Cược 3 con về 1st, 2nd, 3rd, theo thứ tự.

- Phân tích các yếu tố ảnh hưởng đến kết quả trận đấu như:

. Đường đua phù hợp: Mỗi con ngựa có thể sẽ có 1 loại đường đua đặc biệt với nó hơn những loại đường đua khác. Ví dụ như khi đua với đường đua đất sẽ đua tốt hơn là đua đường cỏ.

. Khoảng cách: Có những con ngựa phù hợp chạy bền và giữ phong độ trong khoảng cách xa hoặc có những con phù hợp chạy nhanh trong khoảng cách ngắn( nhưng khi chạy dài thì khó chiến thắng hơn).

. Người đào tạo( trainer): Khả năng đào tạo tốt hỗ trợ cho người cưỡi ngựa có thể sử dụng hiểu quả khả năng của con ngựa.

. Người cưỡi ngựa: Khả năng cưỡi ngựa, có thể xử lý được các loại đường đua như đường đất, phẳng, có chướng ngại vật…

. Tình trạng sức khỏe của ngựa: Nếu ở trận đấu trước ngựa có chấn thương, sẽ có thể ảnh hưởng đến phong độ ở các trận sau mặc dù các thành tích trước đó rất tốt.

. Tốc độ của ngựa trong các trận đua trước: Phân tích tốc độ trung bình của con ngựa trong cả trận đua cũng là 1 yếu tốt quan trọng để dự đoạn được con ngựa này có khả năng đua thắng cao không.

. Giống ngựa: Đây cũng là 1 yếu tốt rất quan trọng, thường ta sẽ xem xét về nguồn gốc, giống thuần chủng hay lai..

. Cân nặng: Cân nặng của con ngựa, cân nặng của người cưỡi và đai mà con ngựa mang theo trong cuộc đua. Trong trường hợp con ngựa chạy rất nhanh nhưng do cân nặng mang thêm sẽ có thể ảnh hưởng đến kết quả của cả trận đua.

. Yếu tố thời tiết: Mưa, nắng, ẩm. Sẽ có các con ngựa chẳng hạn chạy ngày nắng nóng rất tốt nhưng khi chạy ở môi trường mưa, ẩm thường có kết quả không tốt.

\**Ngoài ra sẽ có các yếu tố thống kê dựa vào các trận đấu trước để ta có xem xét kết quả trận đấu như Tổng số tiền thưởng mà con ngựa đã kiếm được cho đến hiện tại, Tổng số lần chiến thắng mà con ngựa đã tham gia, Tổng số trận tham gia trước đây của đấu sĩ( kinh nghiệm)..*

- Mục đích dự án: Như yêu cầu của dự án, ta sẽ tạo một Machine learning model để dự đoán tất cả các trận đua từ 2018 đến hiện tại với True positive( tỷ lệ dự đoán ngựa thắng) ít nhất đạt được 49.3%( ví dụ có 600 con ngựa đua tốt nhất trong 200 trận, ta phải dự đoán được chính xác 294 con ngựa về trong top 3).

- Tiêu chí hoàn thành: Hiểu được các kiến thức nghiệp vụ về đua ngựa, tìm ra được các Insights tốt, xử lý được raw data( missing values, data type, duplicated data, merging..), hiểu được các yếu tố và sự ảnh hưởng của chúng tới kết quả trận đua.

Và cuối cùng, hoàn thành báo cáo( report) đáp ứng các tiêu chí của dự án.

## 2. Data Understanding:

*\*Phần code tham khảo file EDA.ipynb*

2.1 Descriptive Statistics and Univariate Analysis.

Về data, ta sẽ có 9 bảng được mô tả như sau:

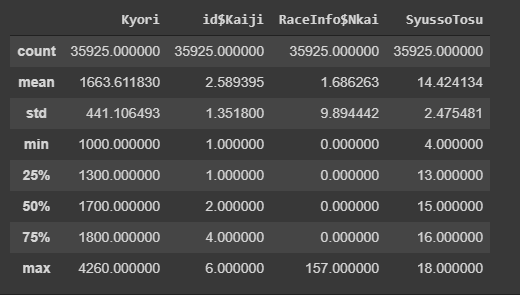
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **English Name** | **Table Name** | **Full Name** | **No. Rows x columns** | **Identities** |
| 1. Race details | RA | RACE( tất cả thông tin các cuộc đua) | 35925 x 115 | Year, JyoCD, Kaiji, Nichiji, RaceNum |
| 2. Horse race information | SE | UMA\_RACE( thông tin con ngựa trong cuộc đua) | 525663 x 78 | Year, JyoCD, Kaiji, Nichiji, RaceNum, KettoNum |
| 3. Racehorse master | UM | UMA( thông tin con ngựa) | 178842 x 238 | KettoNum |
| 4. Jockey master | KS | KISYU( người cưỡi ngựa) | 1742 x 632 | KisyuCode |
| 5. Master of training | CH | CHOKYO( trainer) | 5142 x 587 | ChokyosiCode |
| 6. Producer master | BR | BREEDER( nơi huấn luyện) | 17042 x 32 | BreederCode |
| 7. Owner master | BN | BANUSI( người sở hữu ngựa) | 8997 x 32 | BanusiCode |
| 8. Breeding horse master | HN | HANSYOKU( giống ngựa) | 144166 x 24 | HansyoKunum |
| 9. Weather condition | WE | WEATHER( thời tiết) | 18 x 24 | TenkoCD |

Ở đây chúng ta sẽ có 2 bảng chính là RA và SE.

Bảng SE có field KakuteiJyuni là target, tức vị trị về đích của ngựa mà ta cần dự đoán.

**Descriptive Statistics**

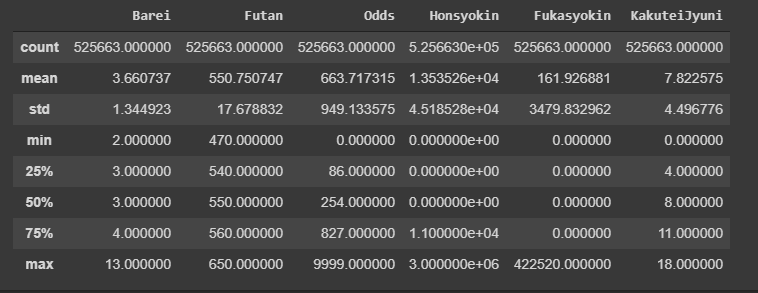
- Bảng RA:



RA Descriptive Statistic

* Kyori( khoảng cách của cuộc đua): Trải dài từ 1000m đến 4260m, trung bình khoảng 1663m.
* Kaiji( Số lần tổ chức cuộc đua hay giải đua, tại 1 trường đua, trong 1 năm): Có thể chỉ có duy nhất 1 giải hoặc lên tới 6 giải đua.
* Nkai( tổng số giải thưởng ở các trận đua trước của giải đua nếu là giải đua nhiều năm, không có null values): Rất ít khi có các giải đua nhiều năm, nếu có có thể lên tới 157 các giải từ các năm trước của giải.
* SyussoTosu( số lượng ngựa thực sự tham gia trận đua, không bao gồm các con ngựa đã đăng ký nhưng hủy tham gia hoặc bị loại trừ vì lý do nào đó): Có từ ít nhất là 4 con tham gia đến cao nhất là 18, trung bình có 15 con tham gia.
* Không có cột kiểu numerical nào có missing data.

- Bảng SE:

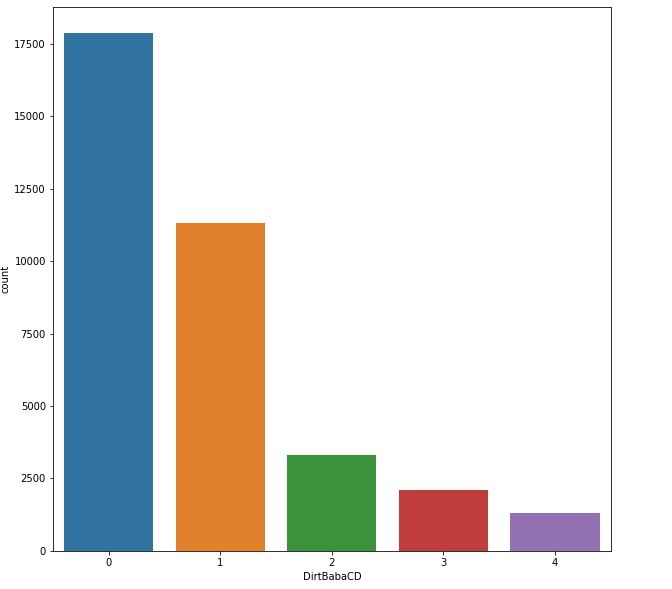


SE Descriptive Statistic

* Barei( tuổi ngựa): Thấp nhất là 2 tuổi, cao nhất là 13 tuổi, trung các con ngựa tham gia ở độ 3 đến 4 tuổi.
* Futan(khối lượng đảm nhiệm(đai + người cưỡi), đơn vị là 0.1 kg): Trung bình khoảng 55 kg.
* Odds( tỷ lệ các cược cho thắng cuộc): Có thể là 0 hoặc có thể lên tới gần 10000 trong khi trung bình chưa tới 1000, đây có thể là dữ liệu outlier, hoặc bị sai trong quá trình nhập liệu hoặc là giá trị mặc định không được áp dụng.
* Honsyokin( tiền thưởng thu được của trận đua): Số tiền thưởng tối đa có thể thu được là 3.000.000 Yên.
* Fukasyokin( tiền thưởng thêm từ trận đua): Cao nhất là 422520 yên.
* KakuteiJyuni( trình tự vệ đích): 18 là vị trí cuối cùng có thể có ở các trận đua này.
* Không có cột kiểu numerical nào có missing data.

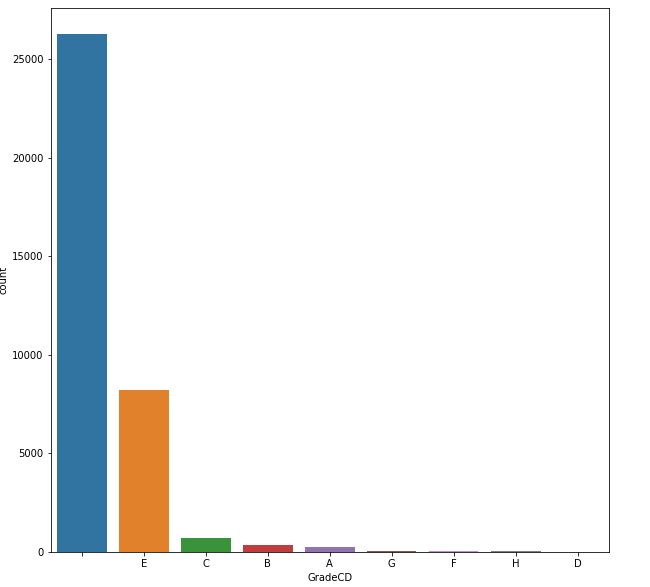
**Univariate Analysis**

- DirtBabaCD( thuộc bảng RA, categorical type): Code tình trạng sân đất.



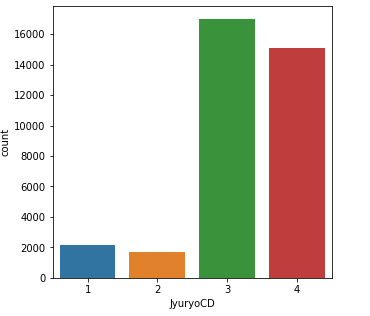
Tập trung chủ yếu ở giá trị 0( dữ liệu chưa được set, lý do hầu như là từ cuộc đua đia phương hoặc cuộc đua ở hải ngoại) và 1( đạt tiêu chuẩn).

- GradeCD( RA, categorical type): Cho biết đây có phải là giải đua đặc biệt hay không.



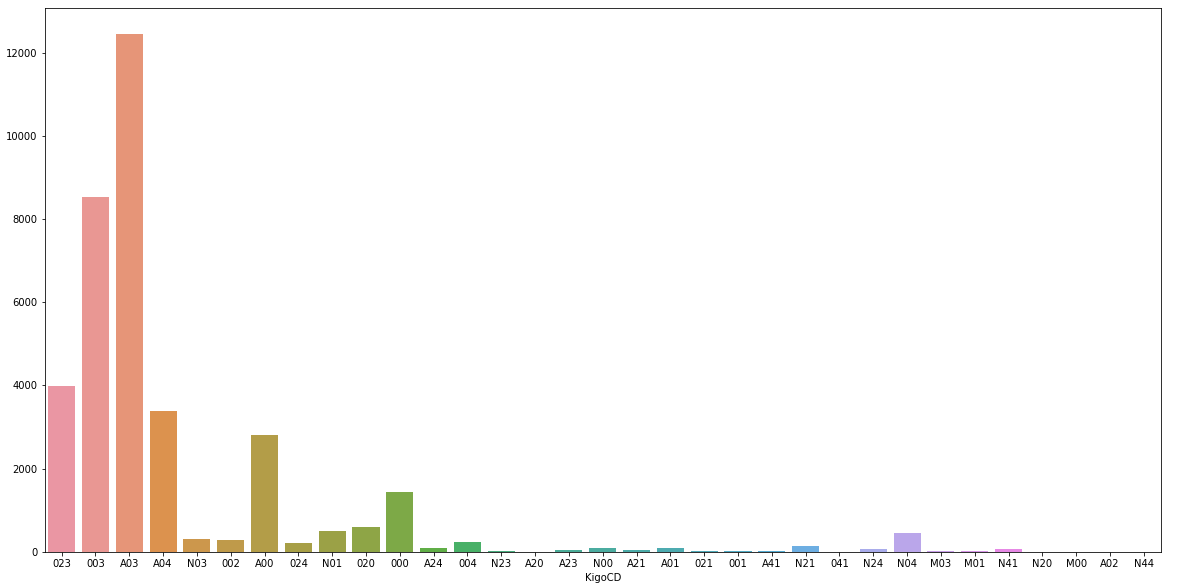
Nhiều nhất là giá trị trống( empty value, dữ liệu chưa được set, lý do hầu như là từ cuộc đua đia phương hoặc cuộc đua ở hải ngoại), tiếp đến ta có E( các cuộc đua đặc biệt ngoài giải thưởng nặng).

- JyuryoCD( RA, categorical type): Code cách tính trọng lượng của ngựa.



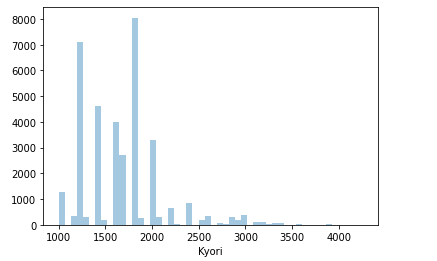
Chủ yếu ta có cách tính 3( cân nặng theo tuổi) và 4( cân nặng đặc biệt).

- KigoCD( RA, categorical type): Code chủng loại cuộc đua ngựa( Cuộc đua chỉ toàn ngựa đực, cuộc đua chỉ toàn ngựa cái, ngựa thiến hay cuộc đua hỗn tạp.v.v.).



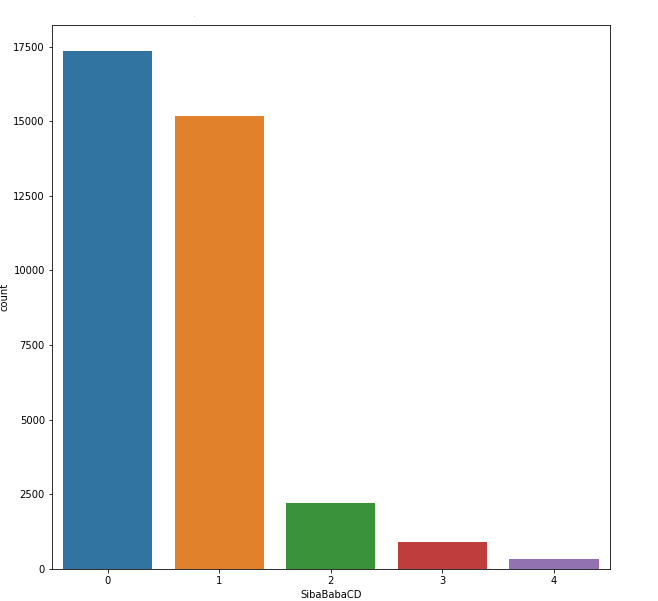
Tập trung chủ yếu là 023( F&M DES), 003( DES), A03( MIX DES), A04( MIX SD), A00( MIX), 000( dữ liệu chưa được set, lý do hầu như là từ cuộc đua đia phương hoặc cuộc đua ở hải ngoại). Chiếm phần lớn đều là các cuộc đua hỗn tạp gồm ngựa đực, cái, thiến...

- Kyori( RA, quantitive data): Khoảng cách cuộc đua.



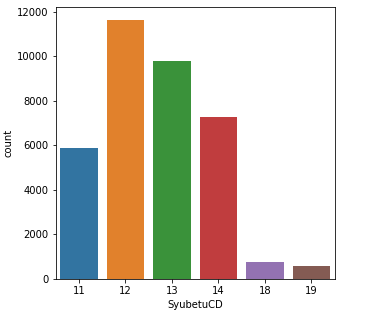
Tập trung chủ yếu trong khoảng 1000~2000m.

- SibaBabaCD( RA, categorical type): Code tình trạng sân cỏ.



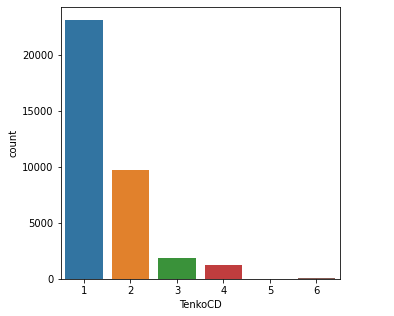
Các cuộc đua hầu như đạt tiêu chuẩn với giá trị 1( standard) ngoài ra còn có 2( Good), 3( Yielding), 4( Soft). Còn lại có 0( chưa được set value).

- SyubetuCD( RA, categorical type): Code cho biết cuộc đua này dành cho loại ngựa bao nhiêu tuổi.



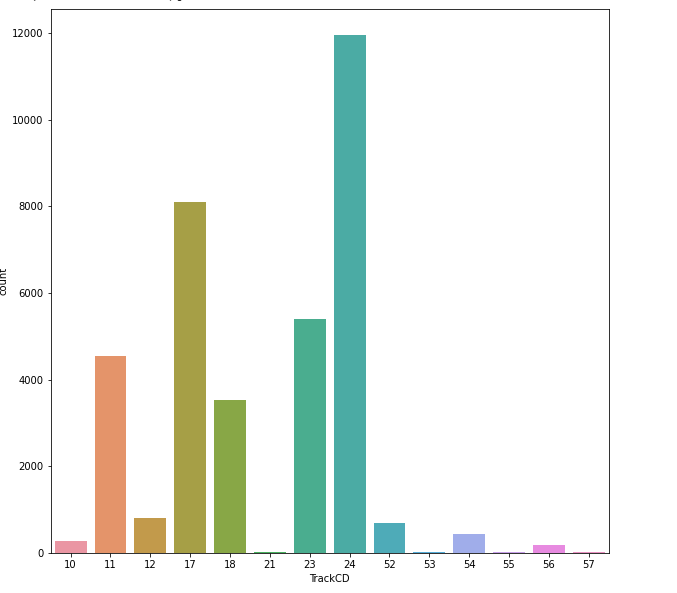
Không quá chênh lệch giữa 11( 2 tuổi), 12( 3 tuổi), 13( từ 3 tuổi trở lên), 14( từ 4 tuổi trở lên). Còn lại rất ít cuộc đua 18( từ 3 tuổi trở lên, STEEPLE-CHASE), 19( từ 4 tuổi, STEEPLE-CHASE).

- TenkoCD( RA, categorical type): Code thời tiết.



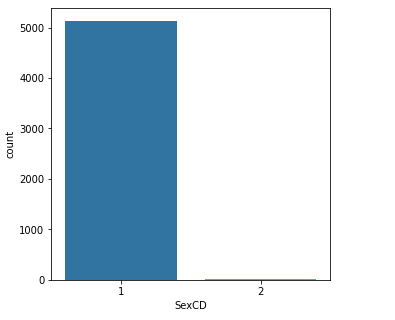
Chủ yếu thời tiết được đánh giá ổn là 1( Fine), bên cạnh đó có 2( Cloudy) là nhiều mây, còn lại 3( Rainy), 4( Drizzle), 6( Light Snow) là không nhiều.

- TrackCD( RA, categorical type): Code loại sân, sân cỏ, sân đất, chướng ngại vật.



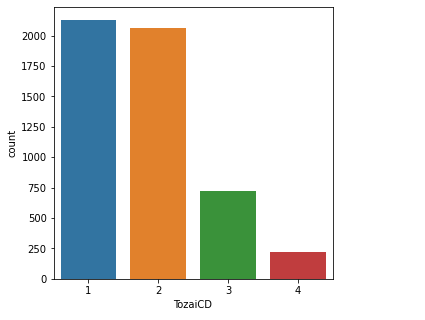
Tập trung chủ yếu với 11( đường cỏ phẳng, ngược chiều kim đồng hồ), 17( đường cỏ phẳng, theo chiều kim đồng hồ), 18( đường cỏ phẳng, theo chiều kim đồng hồ, bên ngoài), 23( đường đất phẳng, ngược chiều kim đồng hồ), 24( đường đất phẳng, theo chiều kim đồng hồ).

- SexCD( CH, categorical type): Code giới tính của trainer.



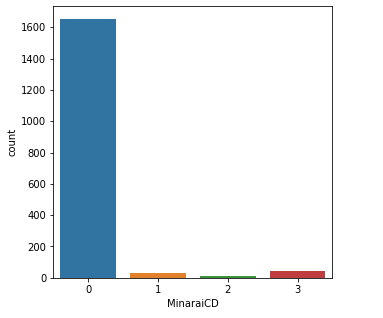
Chủ yếu là nam với giá trị 1( Male).

- TozaiCD( CH, categorical type): Code xuất thân của trainer.



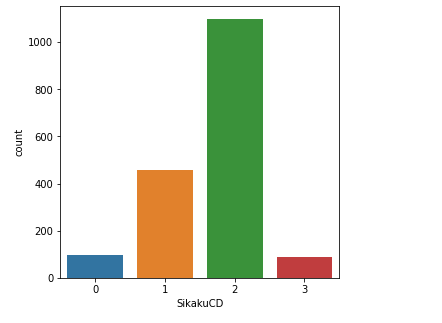
Tập trung chủ yếu ở các nơi như 1( Trung tâm huấn luyện Miura), 2( Trung tâm huấn luyện Ritto).

- MinaraiCD( KS, categorical type): Code thành tích của người cưỡi ngựa.



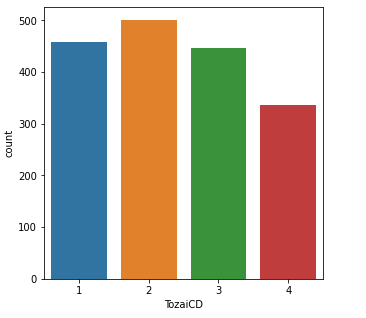
Chủ yếu là chưa được set 0, thường thì rất ít đấu sĩ/ người cưỡi ngựa có các thành tích tốt.

- SikakuCD( KS, categorical type): Code khả năng chạy đường đất, đường phẳng, chướng ngại vật.



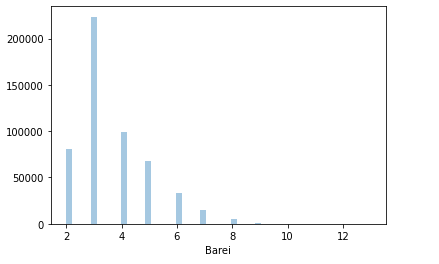
Chủ yếu là khả năng chạy các đường đua bằng phẳng như 1( Flatness), 2( Flat ground).

- TozaiCD( KS, categorical type): Code xuất thân của đấu sĩ.



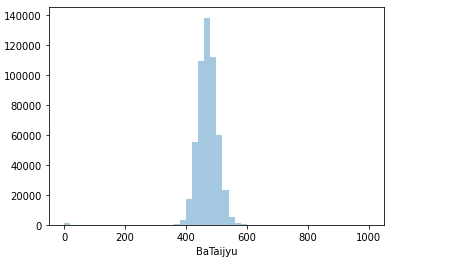
Hầu như trải đều ở các khu vực như 1( Trung tâm huấn luyện Miura), 2( Trung tâm huấn luyện Ritto), 3( từ các khu vực trong nước). Riêng tập trung ít nhất ở 4( từ nước ngoài).

- Barei( SE, quantitive data): Tuổi ngựa.



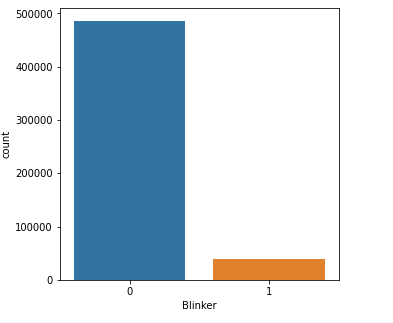
Chiếm phần lớn là những con ngựa 3 tuổi, sau đó ta có 2, 4 và 6 tuổi.

- BaTaijyu( SE, quantitative data, from 002Kg to 998Kg is valid.): Khối lượng ngựa.



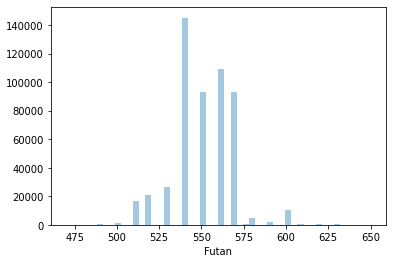
Tập trung chủ yếu từ khoảng 400 đến 600kg, đặc biệt nhiều với các con ngựa có khối lượng khoảng 450kg.

- Blinker( SE, categorical type): Code nịt che mắt của con ngựa.



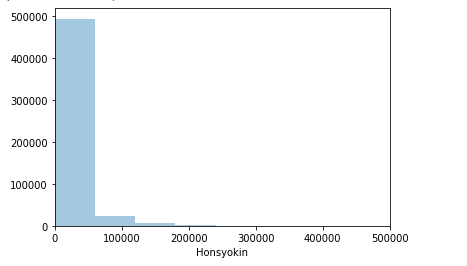
Chủ yếu là không sử dụng với giá trị 0, ít con ngựa sử dụng, tức với giá trị 1.

- Futan( SE, quantitative data, đơn vị 0.1 kg): Khối lượng đảm nhiệm của ngựa( người cưỡi + đai ngựa).



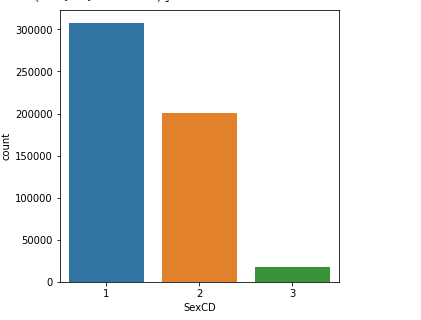
Khối lượng đảm nhiệm của ngựa nằm trong khoảng từ 50 – 60kg.

- Honsyokin( SE, quantitative data): Tiền thưởng thu được từ cuộc đua.



Phần thưởng có thể lên tới 200000 Yên.

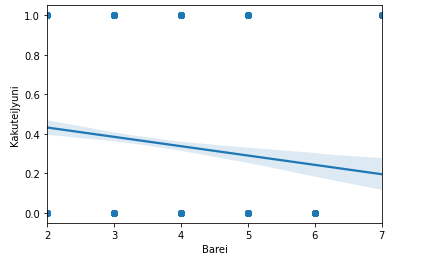
- SexCD (SE, categorical type): Giới tính con ngựa.



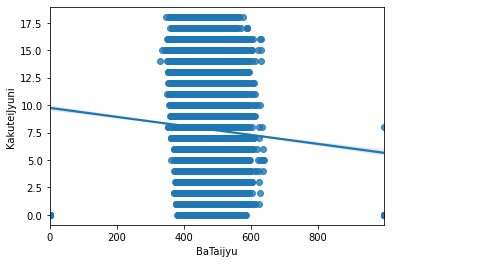
Nhiều nhất là 1( Male horse), sau đó đến 2( Female horse), ít nhất là 3( Castrated male horse, tức ngựa thiến).

2.2. Bivariate Analysis:

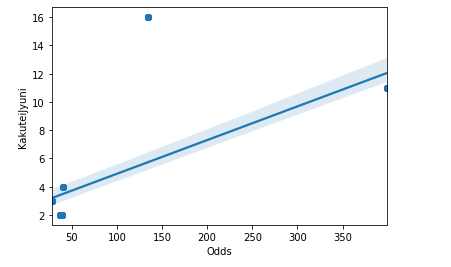
- Barei\_KakuteiJyuni: Từ cơ sở của loại đua rằng có các loại cuộc đua như là dành cho các con ngựa 2 tuổi, từ 3 tuổi trở lên, 4 tuổi trở lên. Xét cuộc đua từ 3 tuổi trở lên sẽ thấy được rằng ngựa càng lớn tuổi có khả năng về đích càng tốt.



- BaTaijyu\_KakuteiJyuni: Hình ảnh về mối quan hệ giữa BaTaijyu và KakuteiJyuni cho thấy rằng ngựa có cân nặng càng lớn, khả năng về đích càng được tốt..



- Odds\_KakuteiJyuni:Hình ảnh bên dưới cho thấy rằng, tỷ lệ cược thắng cuộc cho những con ngựa về sau sẽ cao hơn cho ngựa về trước, mà những tỷ lệ này được các nhà cái đặt trước các trận đua, điều đó chứng tỏ họ hoặc là có khả năng phân tích trận đua tốt hoặc hy hữu có thể là dàn xếp trận đấu. Yếu tố này cũng rất quan trọng khi xem xét về kết quả trận đấu.



Tương tự Odds ta cũng có Ninki( trình tự yêu thích các con ngựa được khán giả xếp hạng), nhưng dữ liệu Ninki có thể là trước hoặc có thể là sau trận đấu, đối với những trận đã diễn ra thường dữ liệu Ninki là sau trận đấu, nên nếu đưa vào model để train ở các bước sau sẽ là không chính xác.

Trên đây đều là những yếu tố chính có ảnh hưởng đến kết quả của cuộc đua, ngoài ra chúng ta cũng có thêm những features khác để việc phân tích được rõ ràng và chính xác hơn.

## Solution Proposal:

Về thuật toán phân loại, ta có thể cân nhắc 2 thuật toán nổi tiếng như KNN, LightGBM. Xem xét về yêu cầu dự án rằng chúng ta sẽ ưu tiên việc dự đoán True Positive cao tức ta sẽ có Precision metric cao hay Accuracy metric cao( đối với binary output, ta sẽ chuyển đổi output về dạng binary ở phần training). Vì vậy ta sẽ ưu tiên sử dụng LightGBM, bên cạnh đó việc load dữ liệu lớn với tốc độ nhanh cũng là một lợi thế của LightGBM.

## Data Preparation:

- Drop duplicated values: Ở các bảng RA, SE, UM, CH, KS với các keys lần lượt là ('id$JyoCD','head$MakeDate$Year', 'id$Kaiji', 'id$Nichiji','id$RaceNum'), ('id$JyoCD','head$MakeDate$Year', 'id$Kaiji', 'id$Nichiji','id$RaceNum', 'KettoNum'), ('KettoNum','ChokyosiCode') vì cùng 1 ngựa ở 2 thời điểm có thể có trainer khác nhau, (ChokyosiCode), (KisyuCode).

- Xử lý outliers: Futan( là khối lượng đảm nhiệm gồm người cưỡi và đai ngựa nhưng chỉ có giá trị 4 x 0.1kg là 0.4kg), BaTaijyu( khối lượng ngựa nhưng là 0kg hoặc lên đến 999kg), Odds( tỷ lệ cược 9999 trong khi giá trị trung bình chưa tới 700). Các outliers này có thể được chuyển sang giá trị mean của Series.

- Data split for training: Như đã nói ở mục đầu rằng ta sẽ có cột KakuteiJyuni thuộc bảng gốc ban đầu là SE làm label/ output cho dự đoán, vì yêu cầu của đề bài là chọn ra 3 con ngựa nhanh nhất của trận đua nên ta có thể quy output về dạng binary output, ngựa nằm trong top 3 về đích sẽ có giá trị là 1 và còn lại sẽ có giá trị là 0. Và các features sẽ là TenkoBaba$TenkoCD, JyokenInfo$SyubetuCD, JyokenInfo$KigoCD, Kyori, TrackCD, TenkoBaba$SibaBabaCD, TenkoBaba$DirtBabaCD, GradeCD, RaceInfo$Nkai, SexCD\_x Barei, Futan, Blinker, BaTaijyu, Odds, KeiroCD, UmaKigoCD, SanchiName, HinsyuCD, SexCD\_y, TozaiCD\_x, SikakuCD, TozaiCD\_y, MinaraiCD.

Bảng thông tin về các features liên quan:

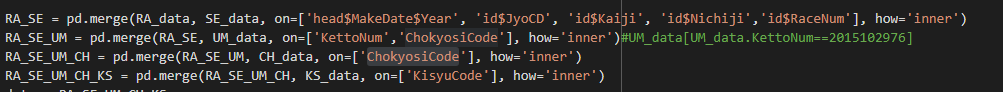
|  |  |
| --- | --- |
| **Tables** | **Features** |
| RA( thông tin tòan bộ race) | - TenkoCD( code weather). - SyubetuCD( race này dành cho ngựa bao nhiêu tuổi). - KigoCD( race code, race toàn ngựa đực, cái hay thiến) - Kyori( khoảng cách cuộc đua). - TrackCD( code loại sân). - SibaBabaCD( tình trạng sân cỏ). - DirtBabaCD( tình trạng sân đất). - GradeCD( loại cuộc đua là đua đường phẳng, đường đất hay giải thưởng nặng không tính điểm..). |
| SE( thông tin các trận đua) | - SexCD( giới tính ngựa). - Barei( tuổi ngựa). - Futan( khổi lượng đảm nhiệm đai + người cưỡi). - BaTaijyu( khối lượng ngựa). - Blinker( nịt che mắt của ngựa)**. -** Odds( tỷ lệ cược cho ngựa thắng cuộc). - Ninki( trình tự yêu thích do khán giả bầu). - UmaKigoCD( code loại ngựa). - Honsyokin( tiền thưởng thu được của race). |
| UM( thông tin con ngựa) | - SanchiName( gốc giống ngựa). - HinsyuCD( product code về giống thuần chủng or not). |
| CH( trainer) | - SexCD( giới tính trainer). - TozaiCD( code xuất thân của trainer). |
| KS( người cưỡi ngựa) | - SikakuCD( khả năng chạy đường đất, phẳng, chướng ngại vật). - SexCD( giới tính người cưỡi ngựa) - TozaiCD( code xuất thân). - MinaraiCD( thành tích). |
| \*SE( features không có sẵn và có thể được tạo mới từ bảng SE) | - win\_per\_count\_horse( tỷ lệ trận thắng của con ngựa ở các trận trước). - win\_per\_count\_kisyu( tỷ lệ trận thắng có sự tham gia của đấu sĩ này). |

*( tham khảo thêm file notebook ở mục Data Preprocessing)*

- Integrate data: Tạo mới column ‘DayMonthYear’ ở bảng SE từ bộ field (id$Year, id$MonthDay) từ đó tạo thêm các features mới hỗ trợ cải thiện model ở mục Model Improvement.

- Data type transformation: Bên cạnh việc chuyển các giá trị output về dạng binary, ta cũng sẽ chuyển một số field về dạng numerical là BaTaijyu( SE) và GradeCD( RA),KeiroCD( SE), TrackCD( RA)..sử dụng LabelEncoder của Scikit-learn. Như vậy mới có thể đưa tập data vào phục vụ training model.

- Data merging: Merge inner join, gồm 5 bảng gồm RA, SE, UM, CH, KS với các joining keys lần là (id$Year', 'id$JyoCD', 'id$Kaiji', 'id$Nichiji','id$RaceNum'), ('KettoNum','ChokyosiCode'), ('ChokyosiCode'), (KisyuCode).



*\*Chi tiết xem thêm ở file notebook*

# **II. Modelling & Evaluation**

## Selected Aglorithms:

Model được lựa chọn là LightGBM.

Lý do chọn model LightGBM: Là một gradient boosting framework sử dụng thuật toán tree-based. Có sự tương đồng với ensemble method giúp hỗ trợ tốt các dataset thực tế, vốn rất khó đạt được hết quả dự đoán/ phân loại cao. Bên cạnh đó việc xử lý nhanh lượng data lớn nhưng lại tốn ít bộ nhớ cũng là lý do tốt để chọn model này cho dự án.

Chọn lựa parameters cho model:

* learning\_rate: GBM làm việc bằng cách bắt đầu ước lượng các hệ số đầu tiên, sau đó cập nhật bằng các output của mỗi tree. learning\_rate này sẽ kiểm soát mức độ thay đổi của các hệ số. điển hình như 0.1, 0.001, 0.003. Ở đây tôi chọn 0.003 cho model đầu tiên.
* boosting\_type: gbdt, traditional Gradient Boosting Decision Tree.
* objective: Sử dụng binary để phù hợp với model là binary classification.
* metric: Lý do giống với objective, ta sử dụng ‘binary\_logloss’.
* max\_depth: Sử dụng parameter này để hạn chế độ sâu của tree.
* num\_leaves: Số lượng leaves trên tree.
* min\_data: 500, gán giá trị lớn để tránh việc tree phát triển quá sâu, có thể dẫn đến under-fitting. Gán mức hàng trăm đến hàng ngàn là đủ.

## Data Splitting

Chúng ta sẽ chia data gồm tất cả cuộc đua từ năm 2008 đến cuối 2017 cho việc training. Về dữ liệu test, ta dùng thông tin các trận đua năm 2018.

## Model training

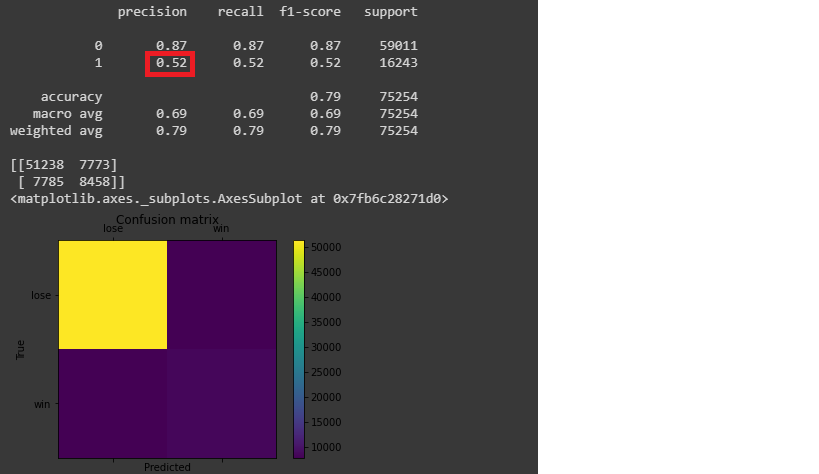
Tập train bao gồm 196877 records và 24 features.

Sử dụng method train() và các params đã nêu ở trên cho việc training.

## Model Evaluation

Implement Evaluation Measure, Assess the models according to the evaluation criteria

Sử dụng các metrics như precision, recall, f1-score cho bài toàn classification này, nhưng đề bài yêu cầu độ chính xác cào về True Positive nên ta sẽ lấy Precision làm metric chính để đánh giá.



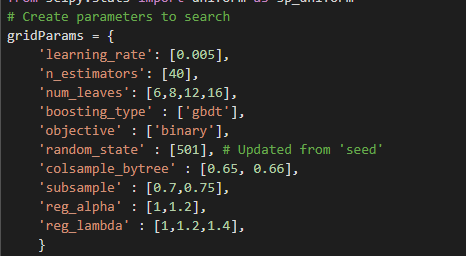
Kết quả model lần I

Như ở trên hình ảnh, kết quả Precision đạt 0.52( 52%) và cao hơn mục tiêu dự án là 49.3% với độ lệch 2.7%.

# **III. Model improvement**

## Tuning Hyper-parameters and Results

Sử dụng GridSearchCV, khoanh vùng các param search như sau:



Ta chọn ra được các params với best estimator:

{'boosting\_type': 'gbdt', 'colsample\_bytree': 0.65, 'learning\_rate': 0.005, 'n\_estimators': 40, 'num\_leaves': 16, 'objective': 'binary', 'random\_state': 501, 'reg\_alpha': 1, 'reg\_lambda': 1.4, 'subsample': 0.7}

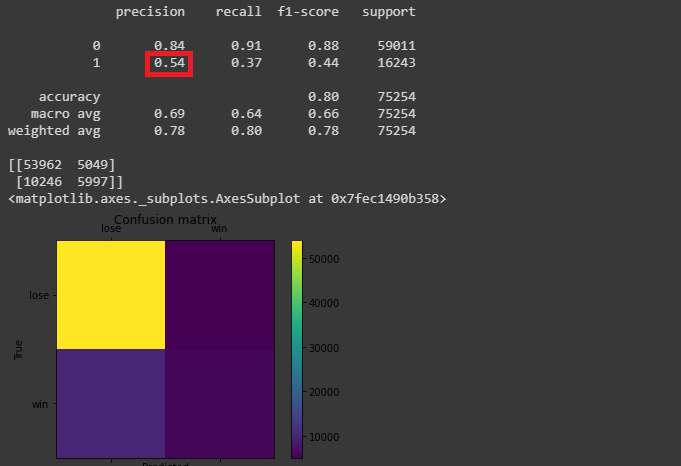
## Other model improvement technique & Results

- Dựa vào bảng SE ta cũng có thể tạo thêm các features mới như:

* win\_per\_count\_horse( tỷ lệ trận thắng của ngựa ở các trận trước).
* win\_per\_count\_kisyu( tỷ lệ trận thắng có sự tham gia của đấu sĩ ở các trận trước).
* Xử lý imbalanced data sử dụng kỹ thuật over-sampling( chi tiết xem thêm tại file notebook, phần Model Improvement).

## Model evaluation for improved model

Sau khi điều chỉnh lại model, ta thu được kết quả như sau:



🡪 Tăng thêm được 2% từ 52% lên 54%.

## Conclusion

Cuối cùng, sau khi hoàn thành các giai đoạn phân tích nghiệp vụ, xử lý dữ liệu, training model, ta đã hoàn thành model với mức chính xác True Positive ở mức 54%. Như vậy có thể nói hoàn thành mục tiêu ban đầu, đã dự đoán thành công top 3 những con ngựa xuất sắc nhất ở các trận đua năm 2018 với TP trên 49.3%.