

پیش بینی بازی CS:GO

علیرضا کتانی یوسف زارع

گروه 17

استاد قدیری

ترم بهار سال 1401

فهرست

[**قسمت اول : فاز شناخت و حل مسئله** 3](#_Toc105442204)

[معرفی حوزه مسئله 3](#_Toc105442205)

[مرور کارهای مرتبط 4](#_Toc105442206)

[ایده های تیم 4](#_Toc105442207)

[**قسمت دوم : فاز آماده سازی داده ها** 4](#_Toc105442208)

[تشخیص داده های پرت 4](#_Toc105442209)

[تبدیل و استاندارد سازی داده ها 6](#_Toc105442210)

[دسته بندی مجدد متغییر های دسته های 7](#_Toc105442211)

[سبد بندی متغییر های عددی 8](#_Toc105442212)

[**قسمت سوم : فاز EDA** 9](#_Toc105442213)

[بررسی روابط تک متغیره و تحلیل رفتار متغیرهای پیش بین و هدف 9](#_Toc105442214)

[قسمت بررسی روابط چند متغیره 12](#_Toc105442215)

[سبد بندی بر مبنای پیش بینی مقدار 13](#_Toc105442216)

[استخراج متغیرهای جدید بر اساس متغیرهای قدیمی 14](#_Toc105442217)

[**قسمت چهارم : فاز پیش مدل (Setup)** 16](#_Toc105442218)

[روش انجام Cross-validation 16](#_Toc105442219)

[بالانس داده ها 16](#_Toc105442220)

[تعیین baseline 16](#_Toc105442221)

[**قسمت پنجم : فاز مدل سازی** 17](#_Toc105442222)

[انتخاب و پیاده سازی الگوریتم ها 17](#_Toc105442223)

[اطمینان از عملکرد بهتر نسبت به baseline 17](#_Toc105442224)

[تنظیم بهینه و hyper-parameter ها 17](#_Toc105442225)

[**قسمت ششم : فاز ارزیابی** 18](#_Toc105442226)

[معرفی مجموعه معیار های ارزیابی و محاسبه آن ها 18](#_Toc105442227)

[برآورد هزینه های خطا 18](#_Toc105442228)

[گزارش برگشت به فاز های قبلی 18](#_Toc105442229)

[مشخص کردن بهترین مدل از بین مدل ها همراه با پارامتر تعیین شده 18](#_Toc105442230)

[تحلیل نقاط قوت و ضعف کار انجام شده و پیشنهاد بهبود 18](#_Toc105442231)

# **قسمت اول : فاز شناخت و حل مسئله**

## معرفی حوزه مسئله

* نوع فرایند: پیش بینی نتیجه بازی ، برد یا باخت گروه تروریست یا گروه کانتر
* نوع بازی: حمله دوطرفه گروه ها و کشتن تمام اعضای یک گروه ،بمب گذاری گروه تروریست و منفجر شدن آن، خنثی سازی گروه کانتر
* حجم دیتابیس :122 هزار داده و 96 اتربیوت
* چگونگی تحلیل: تحلیل برمبنای امتیازات ،سلامتی اعضا، نوع اسلحه های هر گروه، نوع تجهیزات حفاظتی

لینک دسترسی به صفحه دیتابیس مورد نظر :

<https://www.kaggle.com/datasets/christianlillelund/csgo-round-winner-classification>

لینک اطلاعات بیشتر در مورد بازی cs:go :

<https://en.wikipedia.org/wiki/Counter-Strike:_Global_Offensive>



تصویر 1 : cs-go

## مرور کارهای مرتبط

* پاک سازی دیتا (حذف چندین ستون بدون کاربرد، بررسی رکورد های missing)
* ترکیب ستون ها براساس دو معیار damage, type
* ترکیب ستون ها و کشیدن نمودار های مرتبط
* سبد بندی متغیرها با سه روش و هر کدام سه تعداد مختلف سبد
* بررسی روابط بین داده ای با متد های مختلف

## ایده های تیم

در این پروژه قصد داریم که به تعدادی سوال برای دستیابی به هدف اصلی مسئله پاسخ بدهیم که هر کدام از آن ها اطلاعاتی به ما میدهند :

* کدام ویژگی بیشترین ارتباط را با برنده تیم در هر دور دارد ؟
* آیا برخی از سلاح ها برای گروه ها مفید هست ؟
* پراکندگی مقادیر تایم، سلامتی و پول چگونه است ؟
* تحلیل فردی متناسب با قدرت بازیکن و تجهیزات
* تاثیر گروه مقابله بر نتیجه بازی

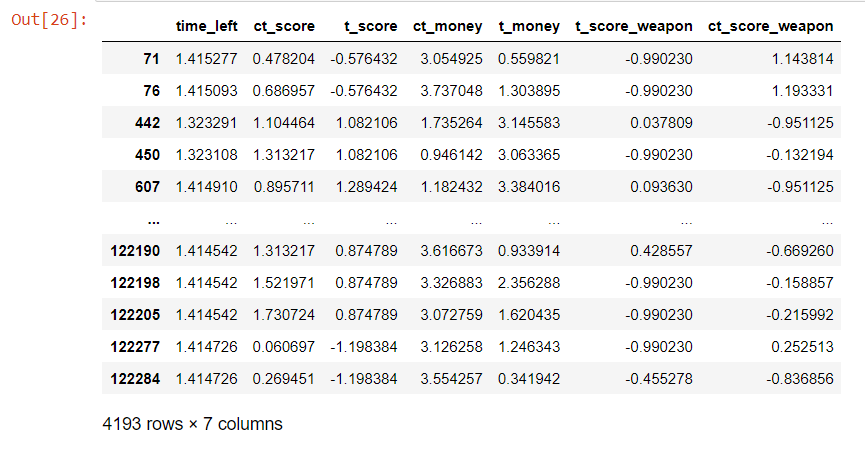
# **قسمت دوم : فاز آماده سازی داده ها**

## تشخیص داده های پرت

برای تشخیص داده های پرت در اینجا ما از سه روش استفاده کرده ایم که به ترتیب بیان میشود :

* Zscore

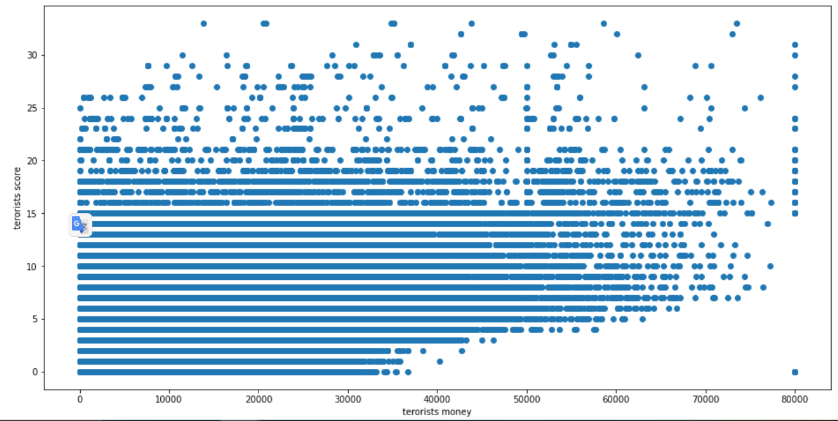
در این قسمت فرمول Zscore را برای 7 متغییر انجام داده و هر کدام از آن ها که مقدارش خارج بازه [-3,3] قرار دارد جزوه داده های پرت محاسبه میشود.

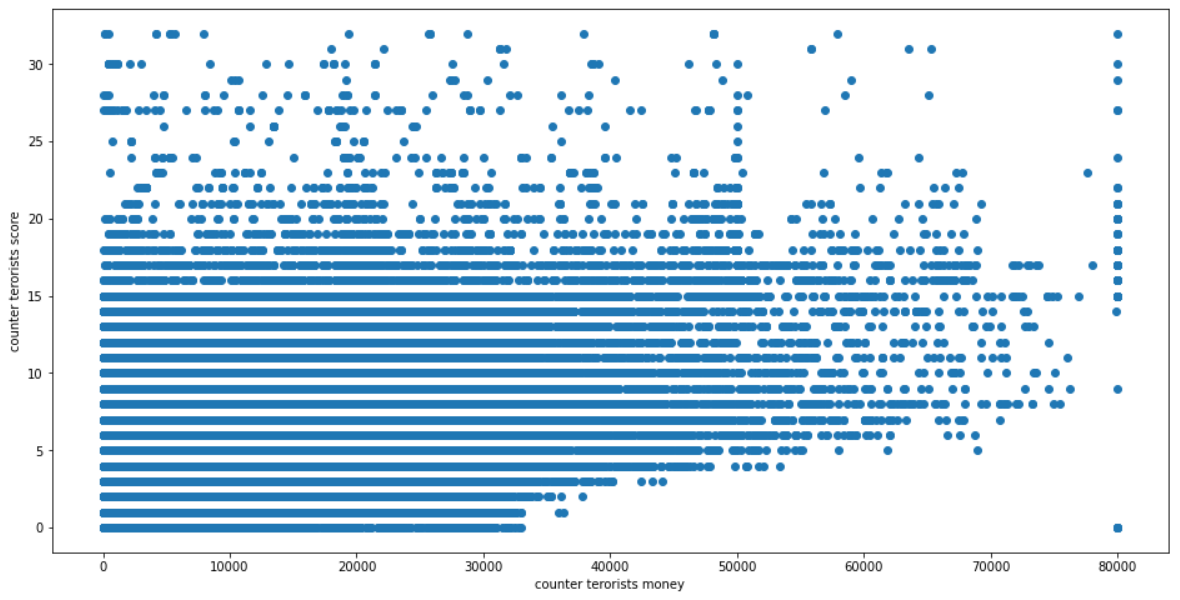


تصویر 2 : روش z-score بر روی دیتابیس

* Scatter plot

در این روش ما با استفاده از نمودار Scatter بر روی هر دو متغییر انتخاب شده میتوانیم داده های پرت بین این دوتا متغییر را تشخیص بدهیم ، نمودار های زیر برای تشخیص داده های پرت بین score , money میباشد که نمودار اول مربوط به ct و دومی مربوط به t یا همان تروریست می باشد.

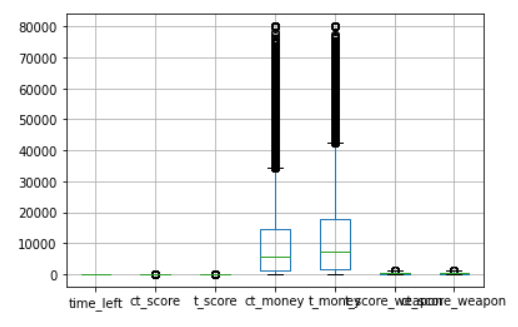




تصویر 3 : scatter plot بر روی اتریبیوت money / score

* box plot

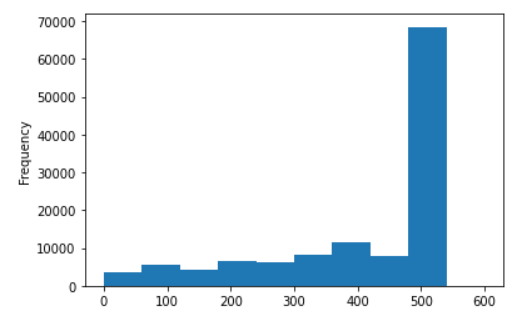
در این قسمت از نمودار box استفاده میشود که از اعدادی که از box خارج باشند جزوه داده های پرت محاسبه میشوند البته که در این جا به دلیل زیادی داده های پرت نمیتوان تشخیص کافی داد.



تصویر 4 : box plot بر روی دیتابیس

## تبدیل و استاندارد سازی داده ها

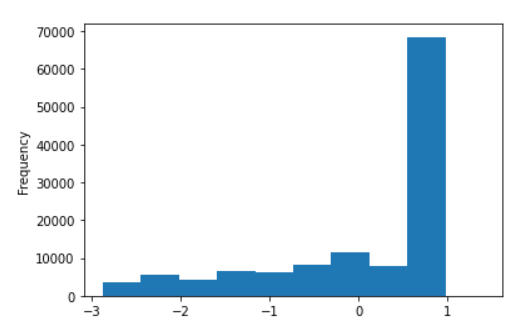
در این جا ما برای استاندارد سازی از دو روش بر روی داده ی t\_health استفاده کرده ایم.



تصویر 5 : t\_health plot

* Zscore

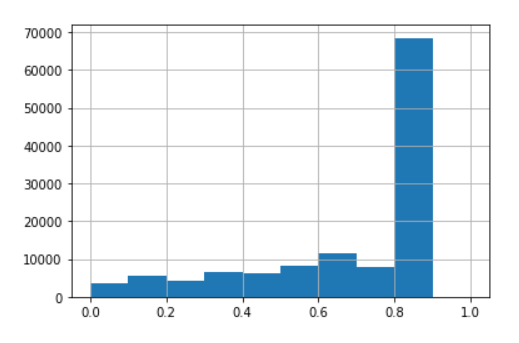
نمودار بالا که مربوط به t\_health می باشد با استفاده از این روش به نمودار زیر تبدیل میشود.



تصویر 6 : z-score t\_health plot

* Min-Max

در این قسمت نیز با استفاده از روش Min-Max که روش سریع تری نسبت به بالا میباشد بر روی داده ی t\_health به نمودار زیر دست پیدا کرده ایم.



تصویر 7 : Min-Max t\_health plot

با توجه به نمودار های به دست آمده به نظر می رسد که روش Min-Max با بازه هایی که مشخص کرده است برای این متغییر مناسب می باشد.

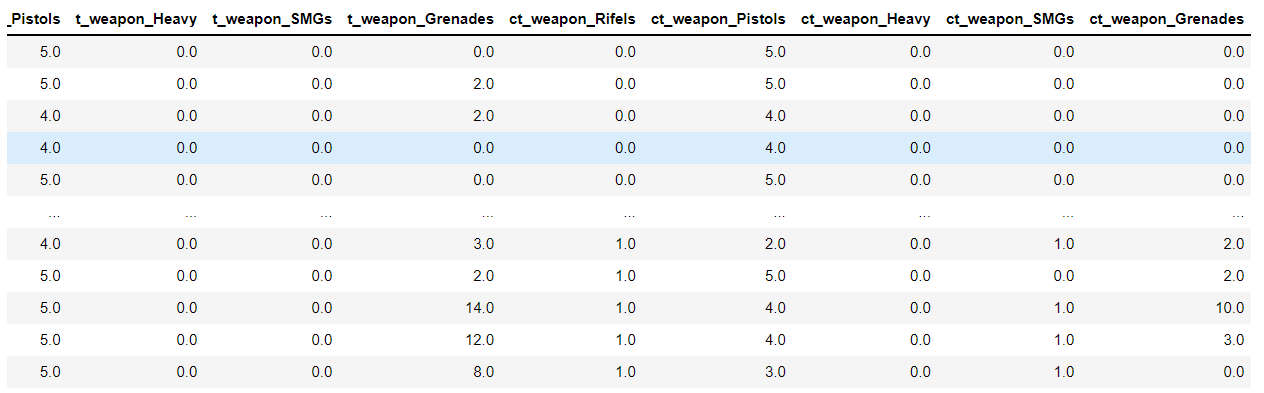
## دسته بندی مجدد متغییر های دسته های

با توجه به دیتابیس متوجه میشویم که داده های مربوط به اسلحه زیاد بوده و حدود 68 ستون را در اختیار دارد برای کاهش ستون های اسلحه از روش دسته بندی برای اسلحه ها استفاده میکنیم به صورتی که اسحله ها را با در نظر گرفتن گروه های زیر 5 ستون ها تقسیم میکنیم.



تصویر 8 : دسته بندی های Weapons

حال با استفاده از این تقسیم بندی 68 ستون را به 10 ستون تبدیل کرده ایم که 5 ستون آن مربوط به تروریست و 5 ستون دیگر مربوط به گروه مقابله با تروریست می باشد ، بخشی از ستون های تقسیم بندی شده :



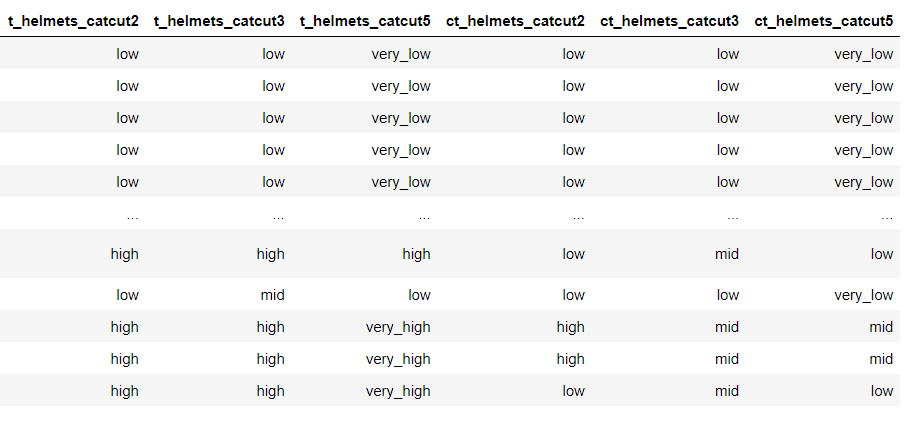
تصویر 9 : جدول دسته بندی های Weapons

## سبد بندی متغییر های عددی

سبد بندی یا همان binning را با دو روش cut و qcut بر روی متغییر های عددی انجام داده ایم

* cut

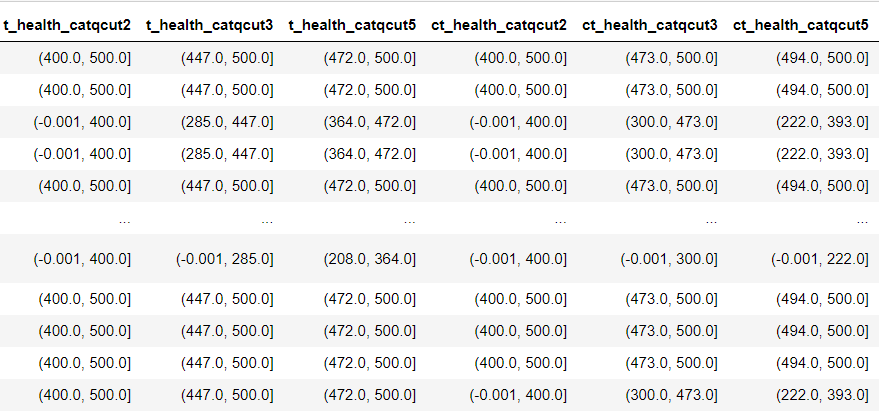
متغییر های عددی را از دیتابیس جدا کرده و سپس بر روی آن ها سه بار روش cut را انجام می دهیم که هر بار به ترتیب دسته های 2 تایی، 3 تایی ، 5 تایی تقسیم کرده ایم ، بخشی از ستون های به دست آمده :



تصویر 10 : دیتابیس بعد از سبد بندی cut

* qcut

در این قسمت نیز همان تقسیم بندی های روش cut را انجام میدهیم فقط در اینجا با استفاده از روش qcut نتیجه حاصل میشود ، که به صورت بازه ای در ستون ها قرار میگیرد.

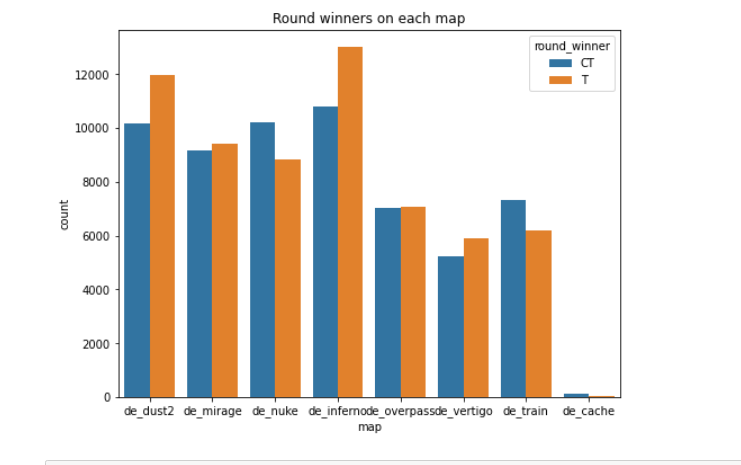


تصویر 11 : دیتابیس بعد از سبد بندی qcut

# **قسمت سوم : فاز EDA**

## بررسی روابط تک متغیره و تحلیل رفتار متغیرهای پیش بین و هدف

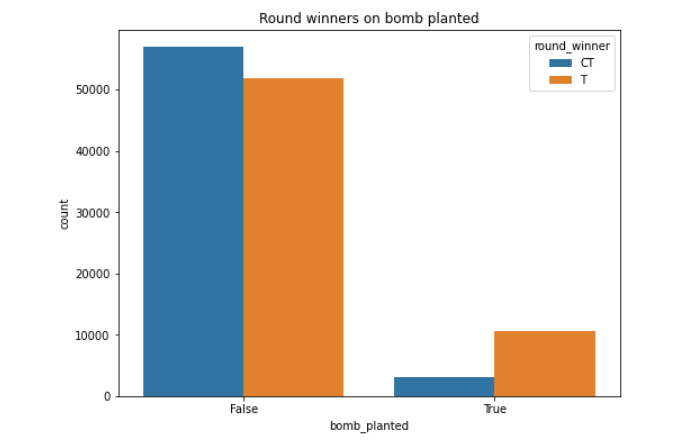
در ابتدا به کشیدن نمودار های نوع بار پرداختیم .از آنجا که برخی از نمودارهای نیاز به نرمال سازی داده ها داشتند به وسیله کراس تب این کار انجام شده است.



تصویر 12 : نمودار توزیع برنده ها در هر map

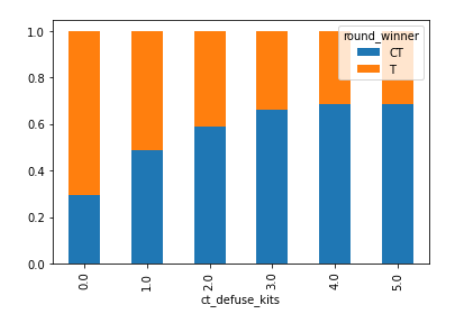
برای مثال در نمودار بالا مشاهده میشود که در چه زمین های بازی چه گروه های موفق تر عمل میکنند که میتواند تا حد زیادی به مکان قرار گیری هر گروه مرتبط باشد.

یا مثلا در نمودار زیر مشاهده میکنید که قرار گرفتن یا نگرفتن بمب تا حد زیادی موجب تغییر نتیجه بازی خواهد شد به گونه ای که درصورتی فعال سازی تقریبا شانس گروه کانتر به کمتر از نصف رسیده است .پس بهتر است روی گروه خنثی سازی و مهارت های حمله جمعی به مکان بمب گذاری کار شود.



تصویر 13 : نمودار قرار گرفتن بمب در بازی

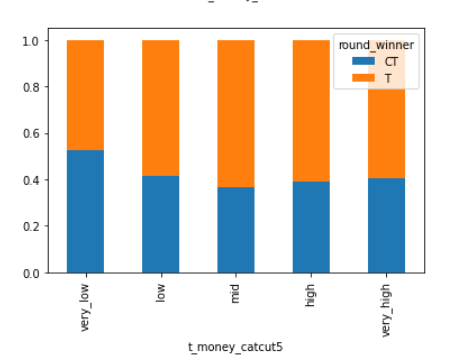
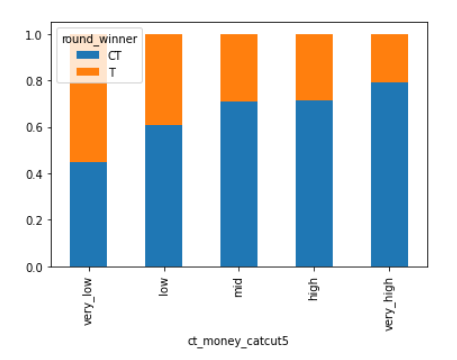
در شکل زیر نمودار تعداد افرادی (از گروه کانتر) که کیت خنثی سازی در بازی دارند و تاثیر آنان در برنده شدن بازی مشاهده میشود .



تصویر 14 : نمودار تعداد افراد دارای کیت خنثی

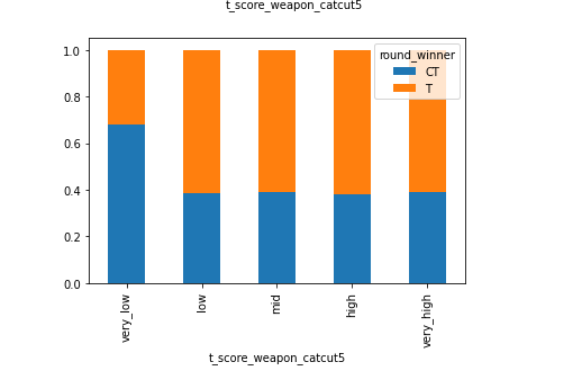
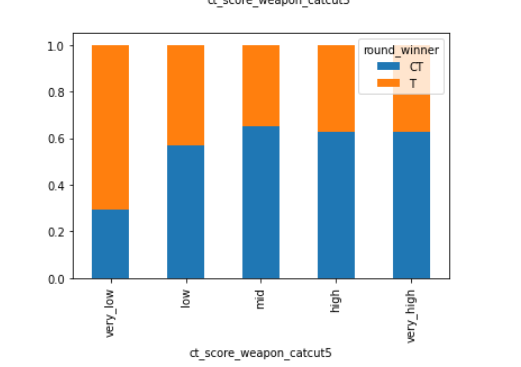
مشاهده میشود که هرچه تعداد کیت بالاتر میرود شانس برنده شدن کانتر بیشتر میشود .کمی قبل تر دیدیم که کارگذاشتن بمب به شدت از احتمال موفقیت کانتر میکاهد پس یکی از ابتدایی کارها برای اصلاح ان میتواند این باشد که گروه کانتر مجبور باشند تعداد کیت خنثی سازی بیشتر از 3 تا بردارند.

نمودار زیر دسته بندی پول بر اساس میزان برنده شدن در بازی را نشان میدهد همانطور که مشاهده میشود با بیشتر شدن پول در هر دسته احتمال موفقیت ان دسته مشابه ها بالا میرود. ولی به طور کلی این میزان افزایش برای دسته کانتر بسیار بیشتر از دسته تروریست هست و گروه کانتر رفتار بهتری را نشان میدهند.



تصویر 15 : نمودار دسته بندی پول بر برنده شدن هر گروه

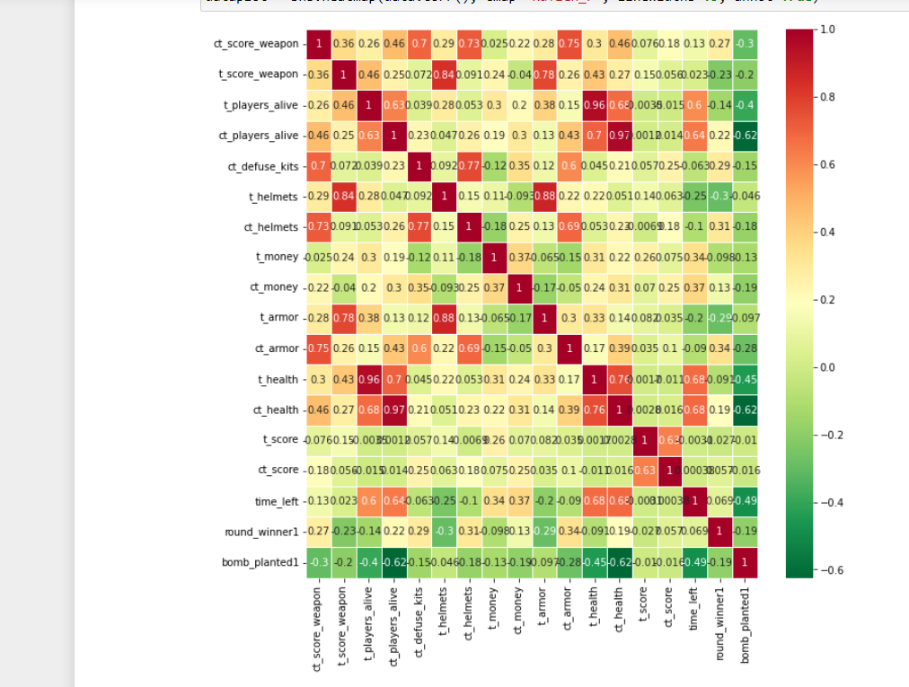
دو نمودار زیر تاثیر برداشتن اسلحه با اسیب بیشتر را نشان میدهد. متوجه خواهیم شد که به طور تقریبا یکسان هر چه میزان اسیب پذیری جمعی گروه بالاتر باشد متعاقبا احتمال برد نیز بیشتر خواهد بود.



تصویر 16 : نمودار تاثیر گذاری اسلحه بر برنده شدن هر گروه

## قسمت بررسی روابط چند متغیره

در این قسمت از ابزار correlation های نوع پیرسون ، کندال و اسپیرمن استفاده کردیم . همچنین برای مشاهده تصویری از ابزار هیت مپ با دو نوع corr خود پایتون و همچنین اسپیرمن بهره بردیم.



تصویر 17 : Correlation Plot

شاید عجیب ترین نوع ارتباط بین میزان اسیب اسلحه ها با کلاه خود که ابزار حفاظتی است .

ارتباط دیگر رابطه منفی میزان زمان زندگی نوع کانتر با بمب گذاری است که دلیل ان کشته شدن همه افراد بازی با بمب است. نکته دیگر بیشتر بودن تاثیر بمب گذاری بر افراد پلیس است که دلیل ان تلاش تیم کانتر برای خنثی سازی بمب حین کارگذاری ان است.

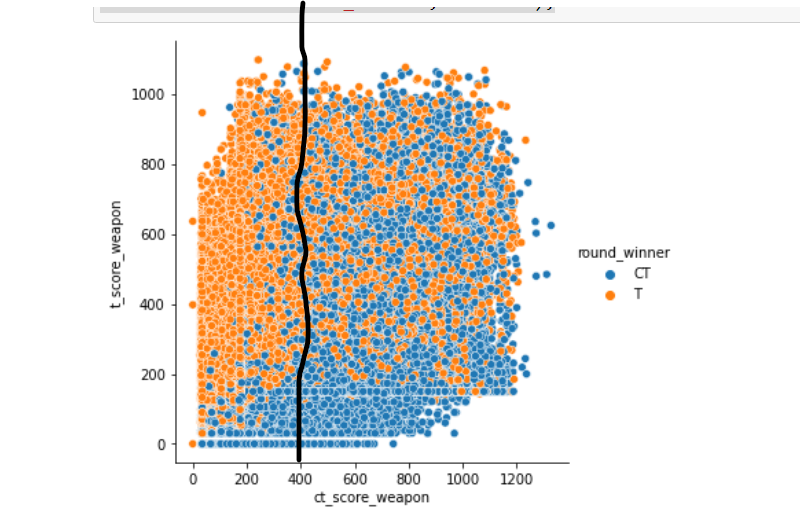
همچنین میزان سلامتی هر دو گروه با زمان کل بازی ارتباط مثبت دارد که دلیل ان واضح است. ارتباط جالب دیگر رابطه قوی کلاه خود با زره حفاظتی است در کانتر کلاه خود عمدتا با زره به فروش میرسد و همچنین هر دو در یک قسمت فروش قرار دارند پس برداشته شدن هر دو طبیعی است.

ارتباط دیگر تاثیر منفی پول و کیت خنثی سازی است. که یعنی پول زیادی برای خریدن ان لازم است. از قبل میدانیم برای بهتر شدن وضعیت بمب گذاری برای گروه کانتر نیاز به کیت های خنثی سازی بیشتری داریم (از سه تا بیشتر ). و حال اینجا میبینیم که قیمت این نوع کیت بسیار گران است پس باید برای عملکرد بهتر قیمت ان را کاهش بدیم.

همچنین هرچه اسلحه میزان اسیب بیشتری داشته باشد مشاهده میشود که سلامتی تا حد خوبی بالاتر میماند با توجه به رابطه 0.5 در نمودار.

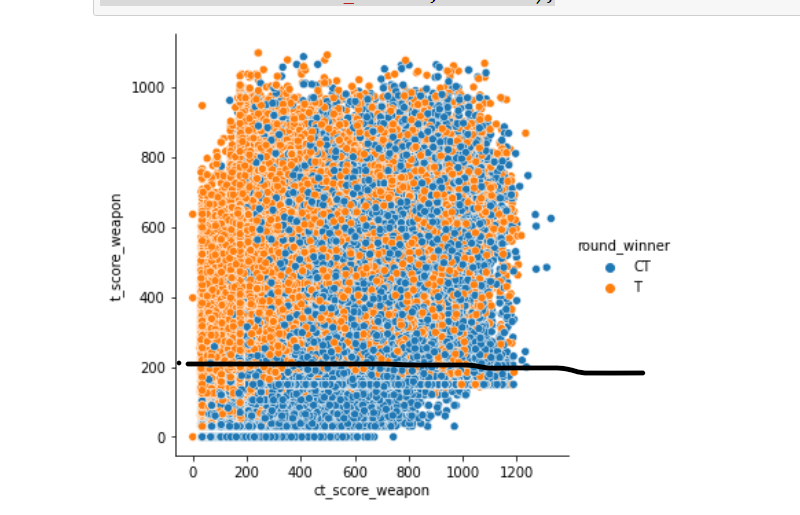
## سبد بندی بر مبنای پیش بینی مقدار

در ابتدای کار به نمودار سه متغیره اسیب اسلحه کانتر و تروریست و همچنین برنده شدن هر طرف میپردازیم.



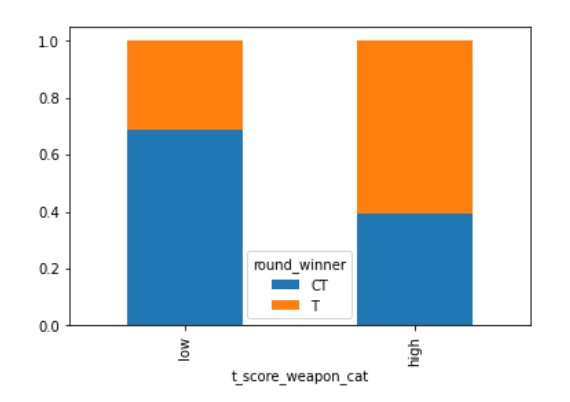
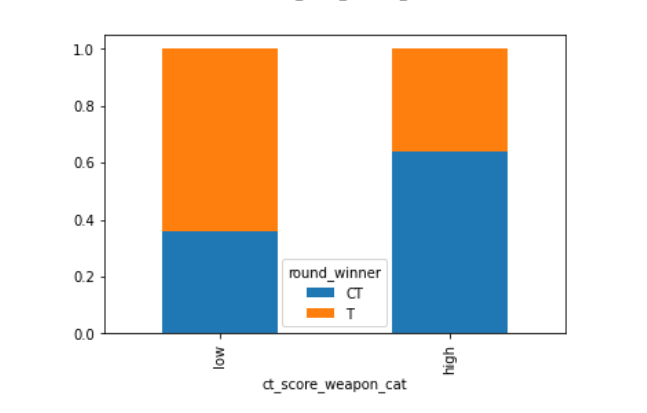
تصویر 18 : نمودار دمیج اسحله هر دو گروه بر اساس برنده بازی

در شکل بالا مشخص است که اگر میزان اسیب اسلحه های گروه کانتر از 400 کمتر باشد به طور مشهودی احتمال موفقیت گروه تروریست بسیار بالاتر خواهد بود پس یک دسته بندی میتواند این گونه باش .همچنین از زاویه دیگر داریم:



تصویر 19 : نمودار دمیج اسحله هر دو گروه بر اساس برنده بازی

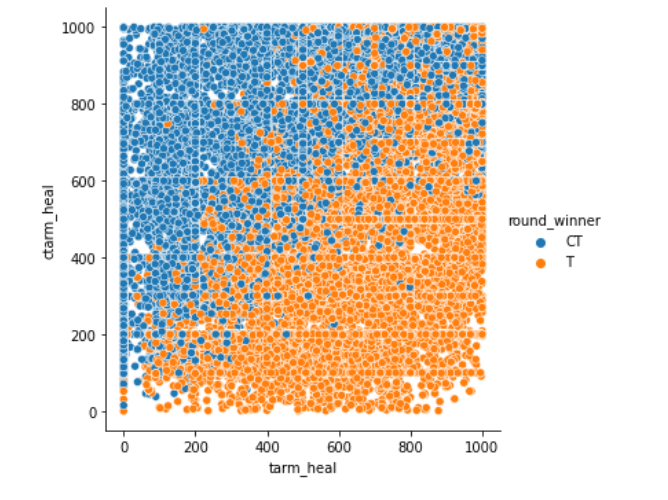
همانطور که در شکل بالا مشخص شده اگر میزان اسیب پذیری اسلحه های گروه تروریست از مقدار 200 کمتر شود به طور مشهودی احتمال موفقیت گروه کانتر بالاتر میرود پس یک دسته بندی هم به این شکل داریم. حال اگر نمودار های متناظر را رسم کنیم نتیجه دلخواه را خواهیم گرفت. نمودار سمت چپ برای کانتر و سمت راست برای تروریست است و آنچه پیش بینی کرده بودیم به واقعیت پیوسته است.



تصویر 20 : نمودار دسته بندی دمیج اسلحه هر گروه بر اساس برنده بازی

## استخراج متغیرهای جدید بر اساس متغیرهای قدیمی

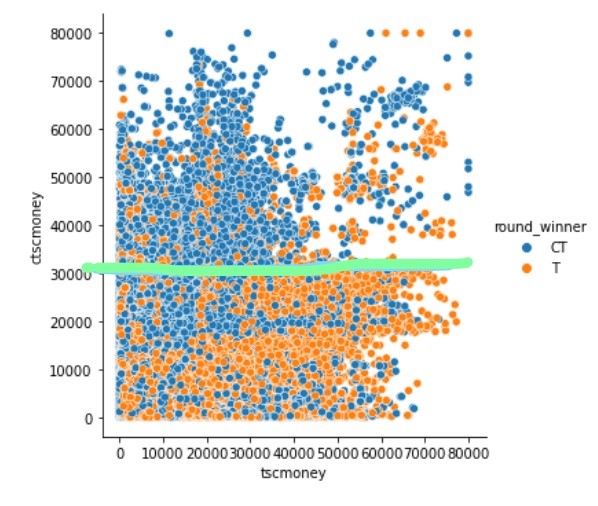
در اولین گام از این مرحله ساخت دو اتربیوت میزان آسیب اسلحه ها به جای حدودا 20 اسلحه برای هر طرف است که در نهایت منجر به ساخت 2 ستون به جای 40 ستون شده است. در روش دیگر به ترکیب این اسلحه ها به گونه ای که در خود بازی قرار گرفته اند پرداخته ایم به گونه ای که حدود 5 دسته جدید ایجاد شده است. سپس متغیر ctarm\_healرا پدید آوردیم که ترکیب دو متغیر سلامتی و زره است که نمودار ان به صورت زیر در می آید.

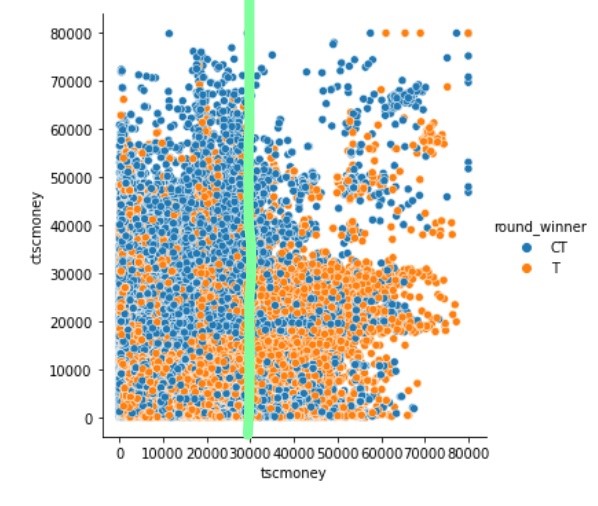


تصویر 21 : نمودار ctarm\_heal / tarm\_heal بر اساس برنده بازی

همانطور که مشاهده میشود پیدا کردن الگو کمی سخت است. هر دو طرف به صورت یکسان پیش روی کرده اند و یک نمودار y=xرا به عنوان مرز خود اختیار کرده اند.

در سمت دیگر به ترکیب متغیر پول و امتیاز پرداخته ایم همانطور که مشاهده میشود هر گروه که این متغیر در آنها بزرگ تر از 30000 باشد تقریبا درصد خوبی از برنده بازی را تشکیل میدهد. متغیر جدید نام tscmoney(c) دارد.





تصویر 22 : نمودار ctscmonry / tscmoneyبر اساس برنده بازی

# **قسمت چهارم : فاز پیش مدل (Setup)**

## روش انجام Cross-validation

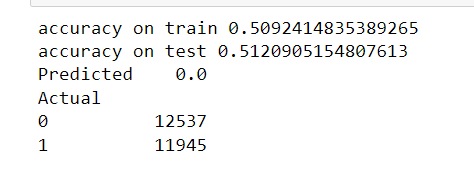
به منظور آماده سازی داده ها برای ورودی الگوریتم، به دو قسمت تقسیم کرده ایم. ستون خروجی ما یعنی round winner را y نامگذاری کرده و تمام ستون های دیگر را X قرار داده ایم. حال باید از دیتاست موجود بخشی را برای داده های تست جداسازی انجام دهیم. برای اینکار مقاله هایی خوانده شده تا درصد انتخاب تست از داده مشخص شود که این عدد 20% می باشد. با استفاده از تابع train\_test\_split داده ی تست را از Data موجود جدا میکنیم. در ادامه با تابع shape تعداد سطر و ستون های test , train را نمایش داده ایم.

## بالانس داده ها

یکی از مزایای خوبی که داده ی ما دارد بالانس بودن آن می باشد یعنی در حدود 51 درصد مواقع گروه تروریست و در 49 درصد باقی مانده گروه پلیس برنده بازی شده اند. و نیازی به بالانس کردن دیتابیس نیست.

## تعیین baseline

برای تعیین baseline از الگوریتم dummy Regressor با استراتژی میانگین استفاده کرده ایم به طوری که میانگین برندگان بازی را محاسبه میکند و تمام مقادیر پیش بینی داده های تست را همان مقدار میانگین می دهد. با توجه به این که مقادیر برد گروه تروریست بیشتر می باشد، میانگین ما برابر گروه تروریست است برای همین برنده تمام بازی ها را تروریست اعلام میکند و دقت این الگوریتم به دلیل بالانس بودن داده ها در حدود 51% است.



تصویر 23 : مقدار Baseline

# **قسمت پنجم : فاز مدل سازی**

## انتخاب و پیاده سازی الگوریتم ها

* Logistic Regression

دقت این الگوریتم در پیش بینی برنده بازی در حدود 74% است که بهتر از baseline عمل کرده اما هنوز هم جای بهبود در دقت با تغییر پارامتر ها وجود دارد.

* SVM

الگوریتم SVM عملکردی همچون الگوریتم Logistic Regression داشته و دقتی در حدود 75% را به ما می دهد.

* Naïve Bayes

برعکس دو الگوریتم بالا ، الگوریتم Naïve Bayes نسبت به دو الگوریتم بالا عملکرد ضعیف تری دارد و دقت 72% در پیش بینی برنده بازی دارد و به دلیل دقت کم جزوه الگوریتم های مناسب دیتاست ما قرار نمی گیرد.

* Neural Network

الگوریتم Neural Network ، با دقت 76% از تمام الگوریتم بالا بهتر بوده ولی هنوز نیز به دقت مورد نظر نرسیده است.

* XGboost

دقت الگوریتم XGboost ، 79% می باشد و بالا تر از Neural Network قرار میگیرد.

* Adaboost

این الگوریتم مانند XGboost خوب عمل نکرده و دقت کم 74% در پیش بینی برنده بازی دارد. در نتیجه مناسب داده ما نمی باشد.

* Decision Tree

الگوریتم Decision Tree مانند XGboost دقت 79% دارد و جزوه الگوریتم با بیشترین دقت قرار می گیرد.

* Random Forest

در نهایت الگوریتم Random Forest با اختلاف از بقیه الگوریتم ها، دقت 85% دارد که از بیشترین دقت الگوریتم های قبل 6 درصد بیشتر می باشد. دقت این الگوریتم در داده های train نیز قابل ملاحظه هست و حدود 99% می باشد.

## اطمینان از عملکرد بهتر نسبت به baseline

## تنظیم بهینه و hyper-parameter ها

برای بهینه تر کردن الگوریتم های بیان شده، از تابع RandomizedSearchCV استفاده شده است به صورتی که یک آرایه از پارامتر ها ، الگوریتم انتخاب شده ، داده های train را به تابع می دهیم تا هم بهترین دقت و هم بهترین پارامتر ها را خروجی بدهد. سپس با قرار دادن پارامتر های خروجی تابع در داخل الگوریتم دوباره آن را اجرا میکنیم و دقت به دست آمده از آن را با مقدار قبلی مقایسه می کنیم. در اکثر مواقع دقت بیشتر از مقدار قبل می شود که همین نشان دهنده بهینه شدن الگوریتم است، اما ممکن است برعکس آن نیز اتفاق بیفتد که باید چندین بار تابع RandomizedSearchCV را اجرا کرد تا بهترین پارامتر خروجی به دست آید. در کد ما سه الگوریتم را به صورت رندوم انتخاب کرده و این کار را برای آن ها انجام داده ایم تا اینکه مثلا برای الگوریتم Random Forest با قرار دادن بهترین پارامتر ها تا دقت 87% نیز توانسته ایم به دست بیاوریم.

# **قسمت ششم : فاز ارزیابی**

## معرفی مجموعه معیار های ارزیابی و محاسبه آن ها

* TPT
* TPCT

## برآورد هزینه های خطا

* MAE
* MSE
* RMSE
* R2

## گزارش برگشت به فاز های قبلی

## مشخص کردن بهترین مدل از بین مدل ها همراه با پارامتر تعیین شده

## تحلیل نقاط قوت و ضعف کار انجام شده و پیشنهاد بهبود