به‌نام خداوند بخشنده و مهربان



**دانشگاه اصفهان**

**دانشکده مهندسی کامپیوتر**

**گروه نرم ‌افزار**

**پروژه دوم**

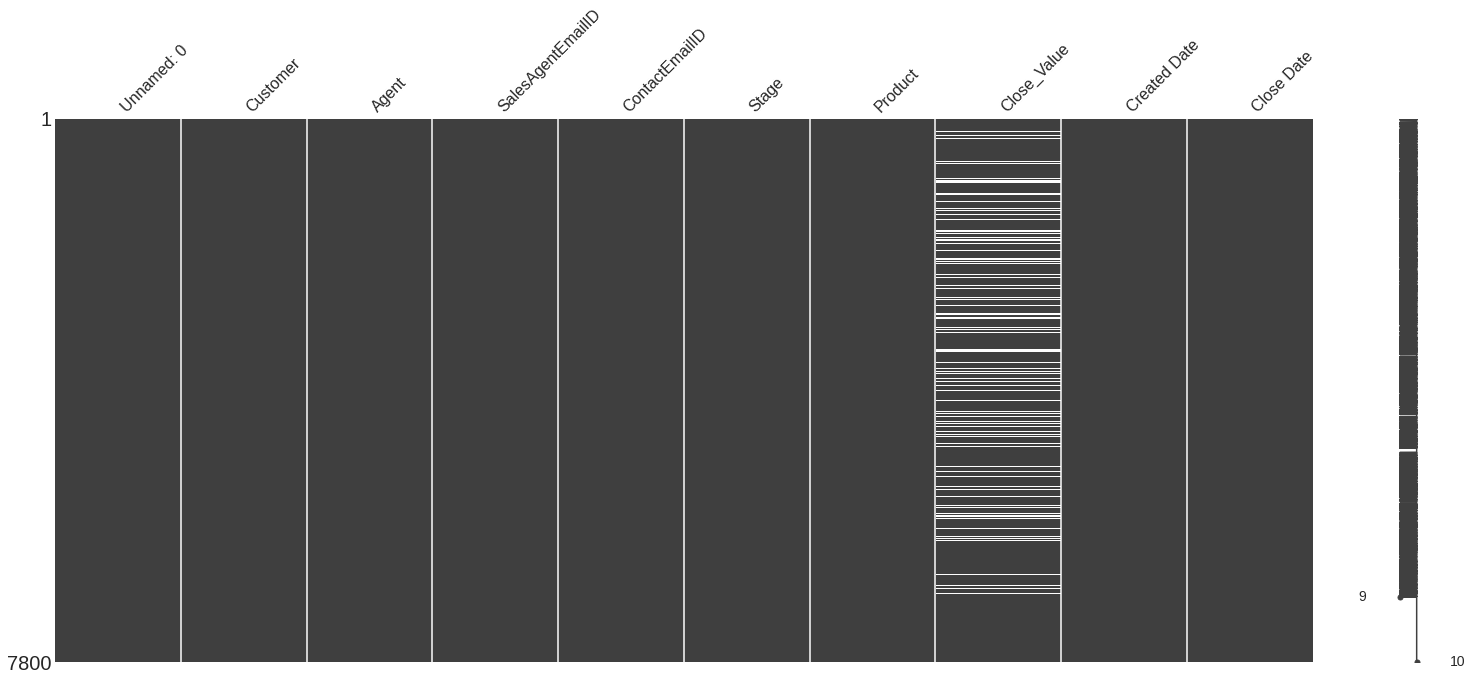
**طبقه بندی**

**استاد: دکتر رضا رمضانی**

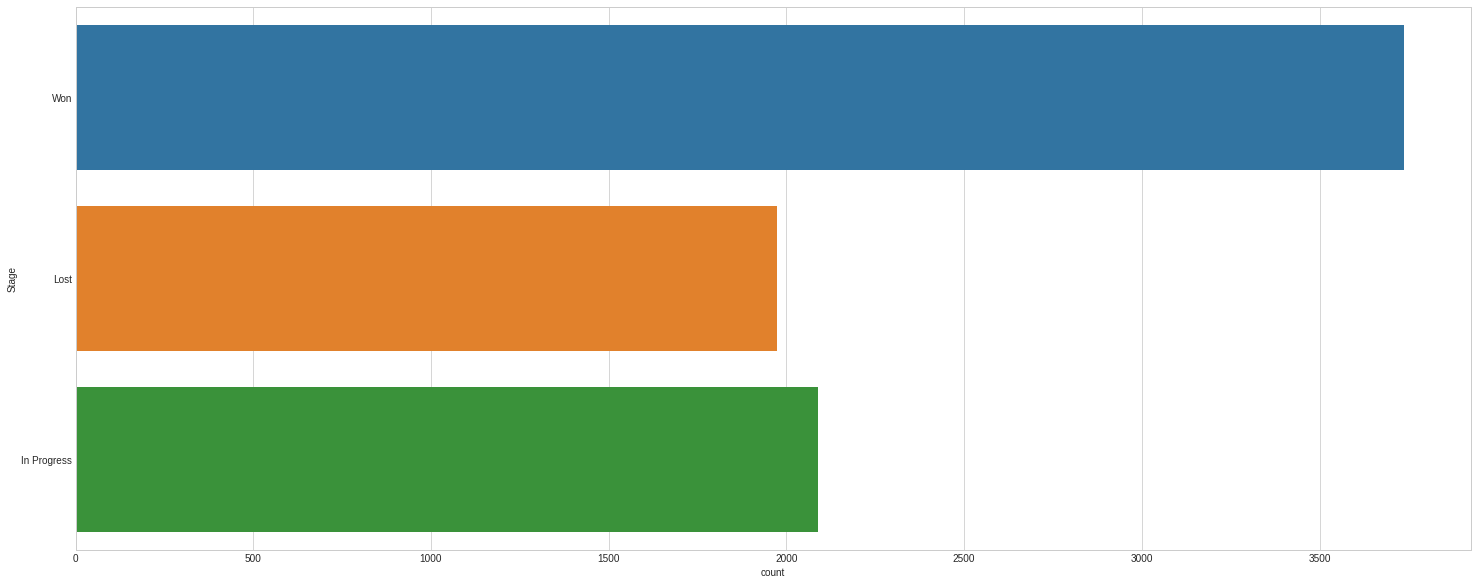
**ارائه دهندگان:**

**امیر سرتیپی 993614019**

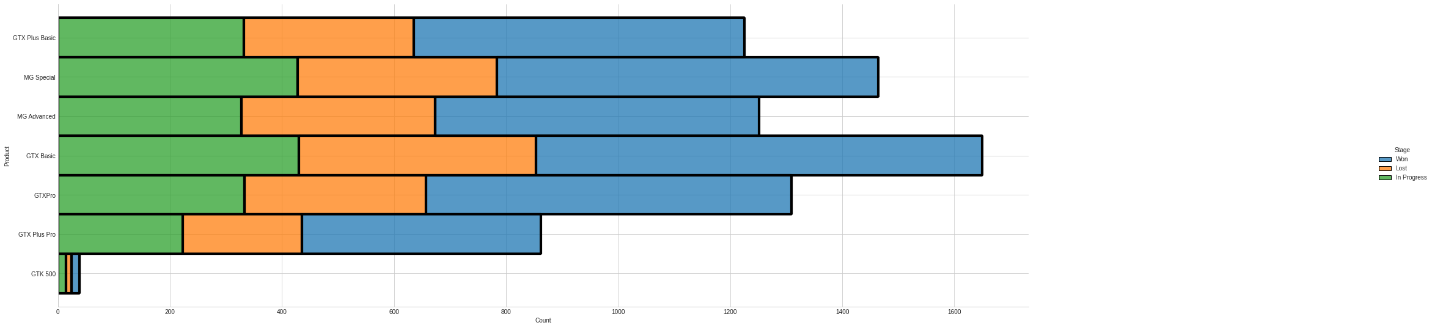
ابتدا یک نگاهی به دیتاست می‌کنیم تا missing value هارا شناسایی کنیم. که همانظور که مشاهده می‌شود در ویژگی close value تعداد زیادی missing value داریم.



یک شمای کلی نیز از تعداد لیبل‌هایی که داریم می‌بینیم. در مجموع نمودارهای زیادی برای تحلیل بهتر داده‌ها ترسیم شده است.



در این شکل تنوع محصولات و نسبت 3 لیبل قابل مشاهده است.



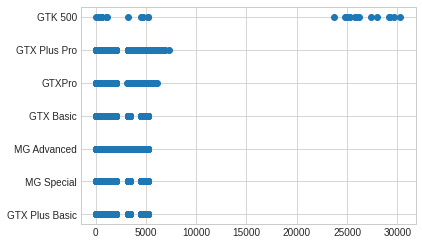
## ستون‌ها

از داده‌های اینتراکش به دو شکل استفاده شد. در ابتدا ویژگی نوع ارتباط به عنوان ویژگی در دیتاست اصلی در نظر گرفته شد. و همچنین به کمک مدل پریترین شده‌ی flair که برای انالیز احساست استفاده می‌شود برای هر معامله تعداد اینتراکشن‌هایی که بین تاریخ create\_date و cloes\_date بود عملیات sentiment انجام شد. در خروجی اعداد بین -1 و 1 به دست امد که هرچه اعداد نزدیک تر به -1 باشد نگرش منفی تری وجود دارد و احتمال برد کمتر است و بر عکس هرچه به 1 نزدیک‌تر باشد احتمال برد بیشتر است. حال جمع همه‌ی سنتیمنت ها مشخص کننده‌ی عدد sentiment هست که در دیتا فریم مشاهده می‌شود.

برای تاریخ‌ها به جای close date و created date مدت زمانی که یک معامله باز بود در نظر گرفته شده و به عنوان duration در نظر گرفته می‌شود.

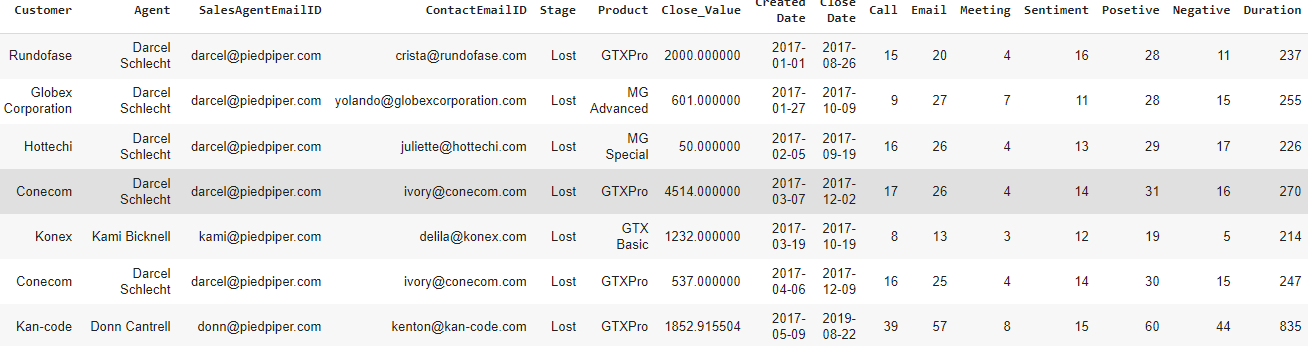
از دیتاست interaction نوع هر interaction نیز در دیتاست اصلی اورده شده است. که برای هر اینتراکشن برای هر معامله 3بار email ، 2 بار call و 1 بار meeting صورت گرفته است. هرکدوم از این ویژگی‌ها نیز به عنوان یک ستون در نظر گرفته شده‌اند.

در قیمت‌های برای یک کالای خاص داده‌ی نویزی وجود دارد که قیمت‌های بسیار بالا هستندن که ‌انها را حذف می‌کنیم.



خروجی دیتاست به شکل زیر می‌باشد.

همچنین تعداد sentiment های مثبت و منفی نیز در ستون‌ها بیان شده است.



## نتایج حاصل از الگوریتم

الگوریتم‌های گفته شده در صورت پروژه بر روی دیتاست امتحان شد و DT دقت بهتری داشت. نتایج به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion matrix | Accuracy | F1 score | **Algo** |
|  | 0.5210084033613446 | 0.6109215017064846 | KNN |
|  | 0.6911764705882353 | 0.8051259390190014 | SVM |
|  | 0.7436974789915967 | 0.8225024248302618 | MLP |
|  | 0.9592050209205021 | 0.9691699604743081 | DT |
|  | 0.4992997198879552 | 0.5517241379310345 | NB |
|  | 0.6624649859943977 | 0.7969671440606572 | RF |

برای label زدن InProgress ها از روش Active Learning استفاده شد که دقت و معیار F1 Measure به شرح زیر می‌باشد.

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion matrix | Accuracy | F1 score | **Algo** |
|  | 0.9673252279635258 | 0.9652945924132366 | DT |

## Hyperparameter

پس از انتخاب DT به عنوان الگوریتم اصلی کانفیگ زیر به GridSearch پاس داده شد تا تمامی موارد نیاز را بررسی کند. کانفیگ به شکل زیر می‌باشد.

params = [{'criterion':['gini', 'entropy'],

           'max\_features': ['auto', 'sqrt', 'log2', None],

           'random\_state':[0]},

          ]

خروجی الگوریتم بهترین پارامتر‌ها به شکل زیر معرفی کرده است.

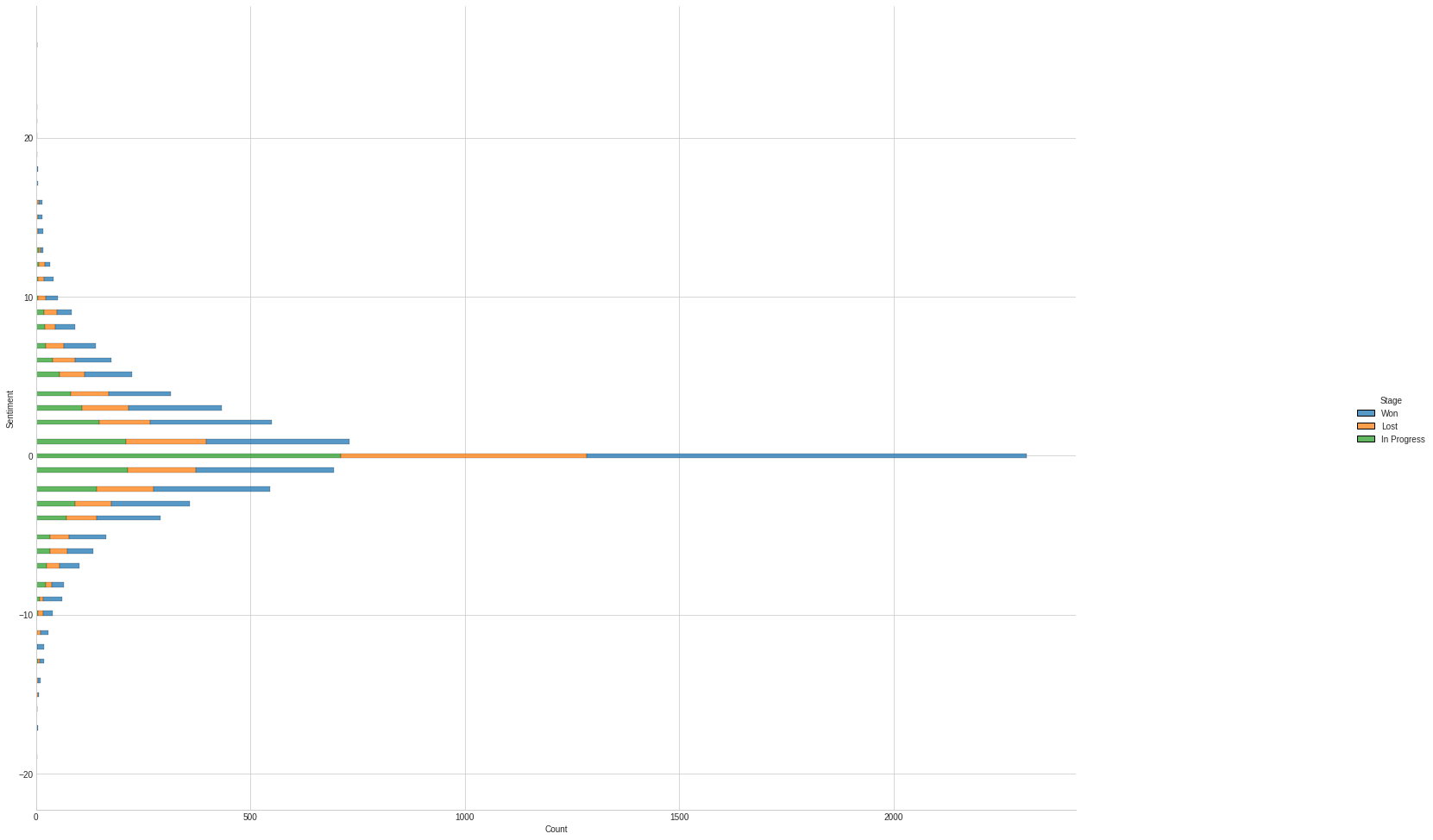
{'criterion': 'entropy', 'max\_features': None, 'random\_state': 0})

## آموزش مدل

برای آموزش مدل ابتدا کل داده‌های لیبیل دار مدل رو اموزش دادیم و InProgress هارو حذف کردیم که دقت مدل بر اساس 10-fold-cross-validation به 90 رسید.

پس از آن داده‌های InProgress را با این مدل label زدیم و مدل را دوباره ترین کردیم که دقت مدل به عدد 85 رسید. این مدل در اصطلاح robust تر خواهد بود.

براساس نمودار زیر و سنتیمت نسبت به Label ها می‌توان نتجیه گرفت سنتیمت تاثیری بر won یا loss نداشته است چون معاملاتی وجود دارد که عدد sentiment بالایی دارد اما منجر به باخت شده است و بر عکس.



موارد موردنیاز برای پروژه در آدرس زیر در دسترس می‌باشد.

<https://github.com/amirsartipi13/DM-P2-Classifications>