به‌نام خداوند بخشنده و مهربان

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

گروه نرم افزار

پروژه دوم

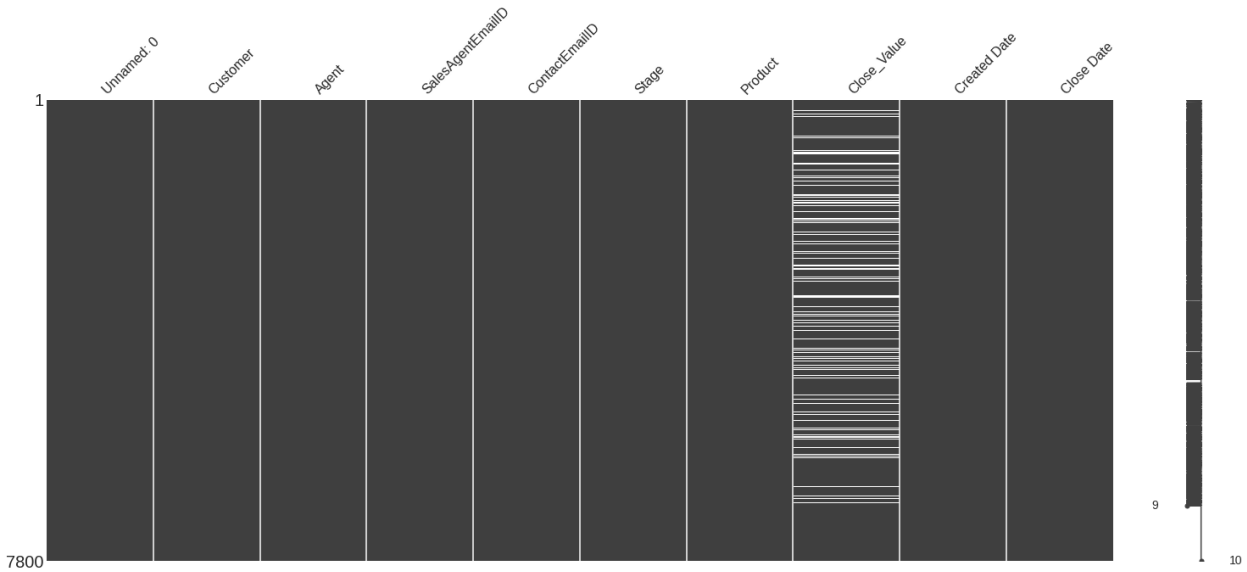
طبقه بندی

استاد: دکتر رضا رمضانی

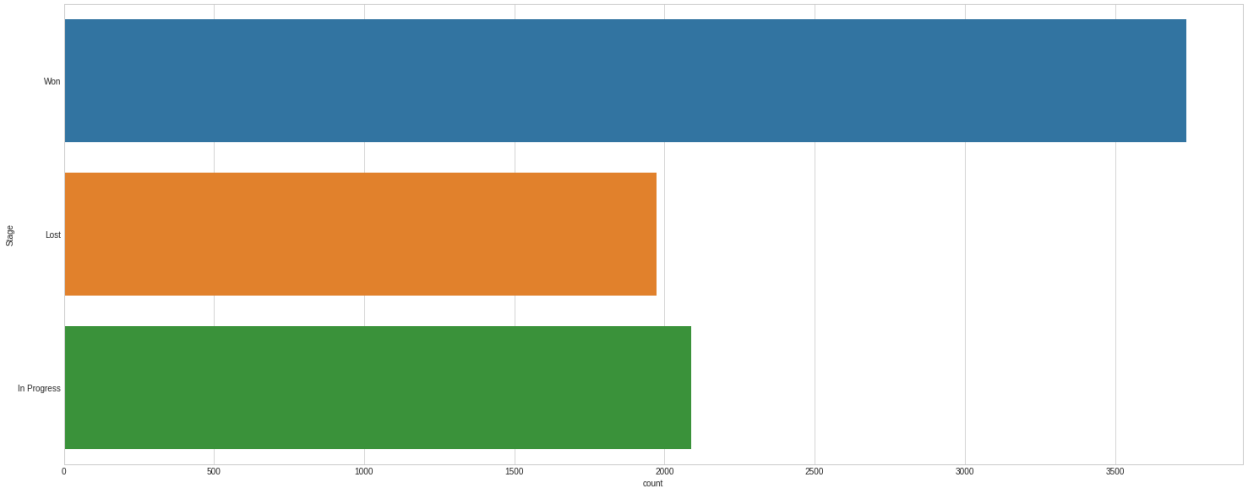
ارائه دهندگان:

امیر سرتیپی ۹۹۳۶۱۴۰۱۹

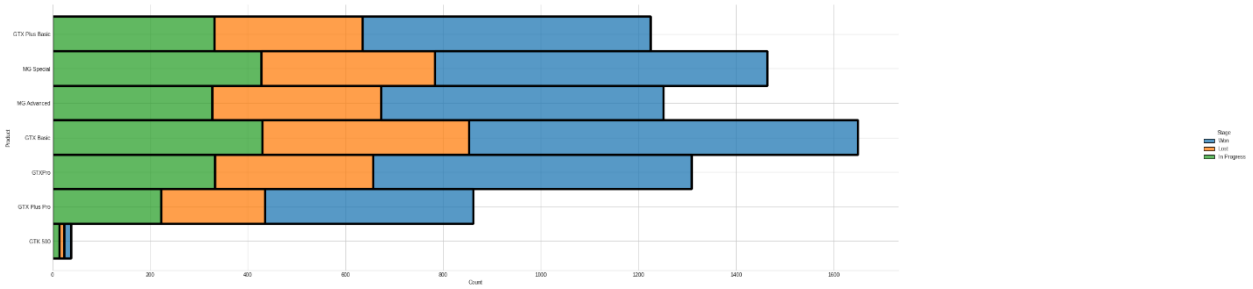
ابتدا یک نگاهی به دیتاست می‌کنیم تا missing value هارا شناسایی کنیم. که همانظور که مشاهده می‌شود در ویژگی close value تعداد زیادی missing valueداریم.



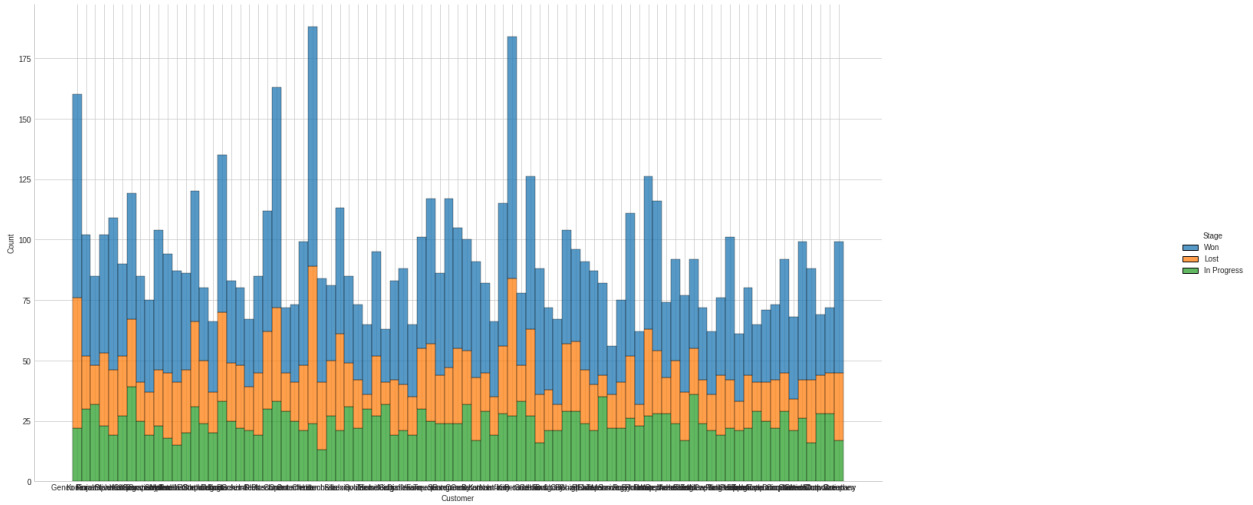
یک شمای کلی نیز از تعداد لیبل‌هایی که داریم می‌بینیم. در مجموع نمودارهای زیادی برای تحلیل بهتر داده‌ها ترسیم شده است.



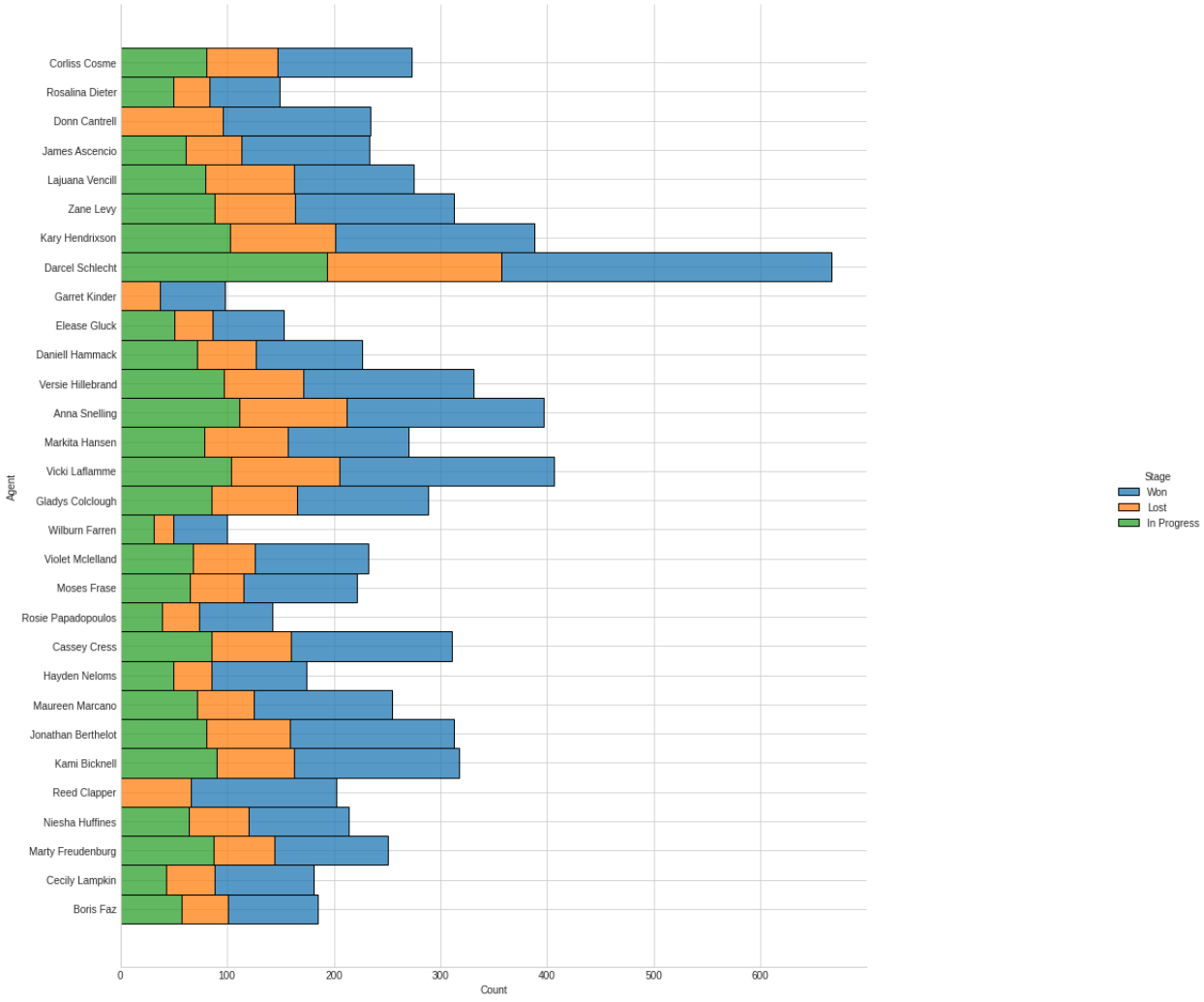
در این شکل تنوع محصولات و نسبت ۳ لیبل قابل مشاهده است.

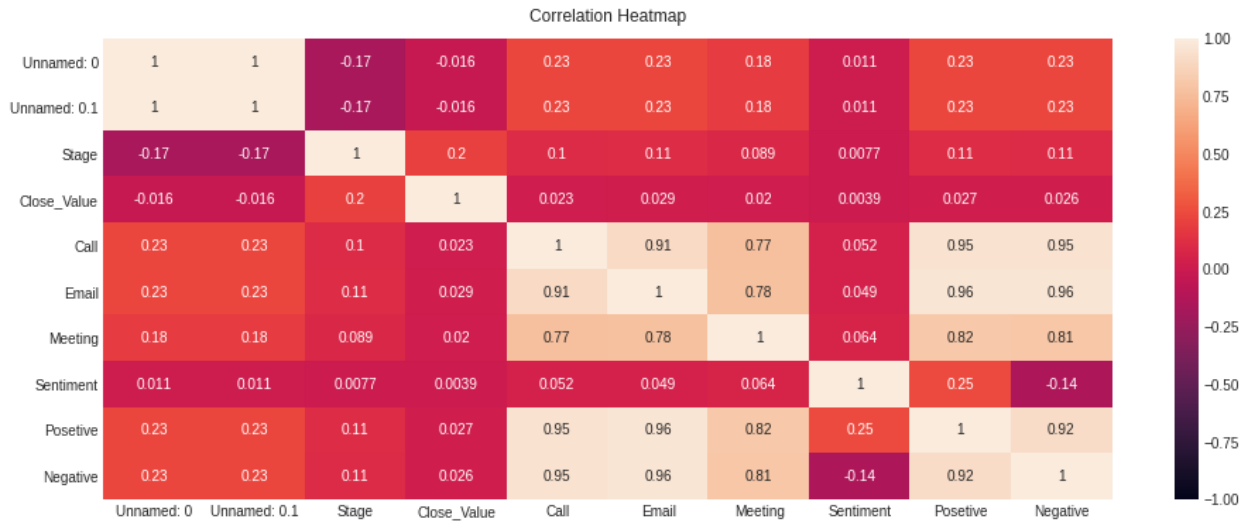


در نمودار زیر تنوع مشتری‌ها به همراه تعداد برد و باخت و اینپراگرس رو مشاهده می‌کنید.

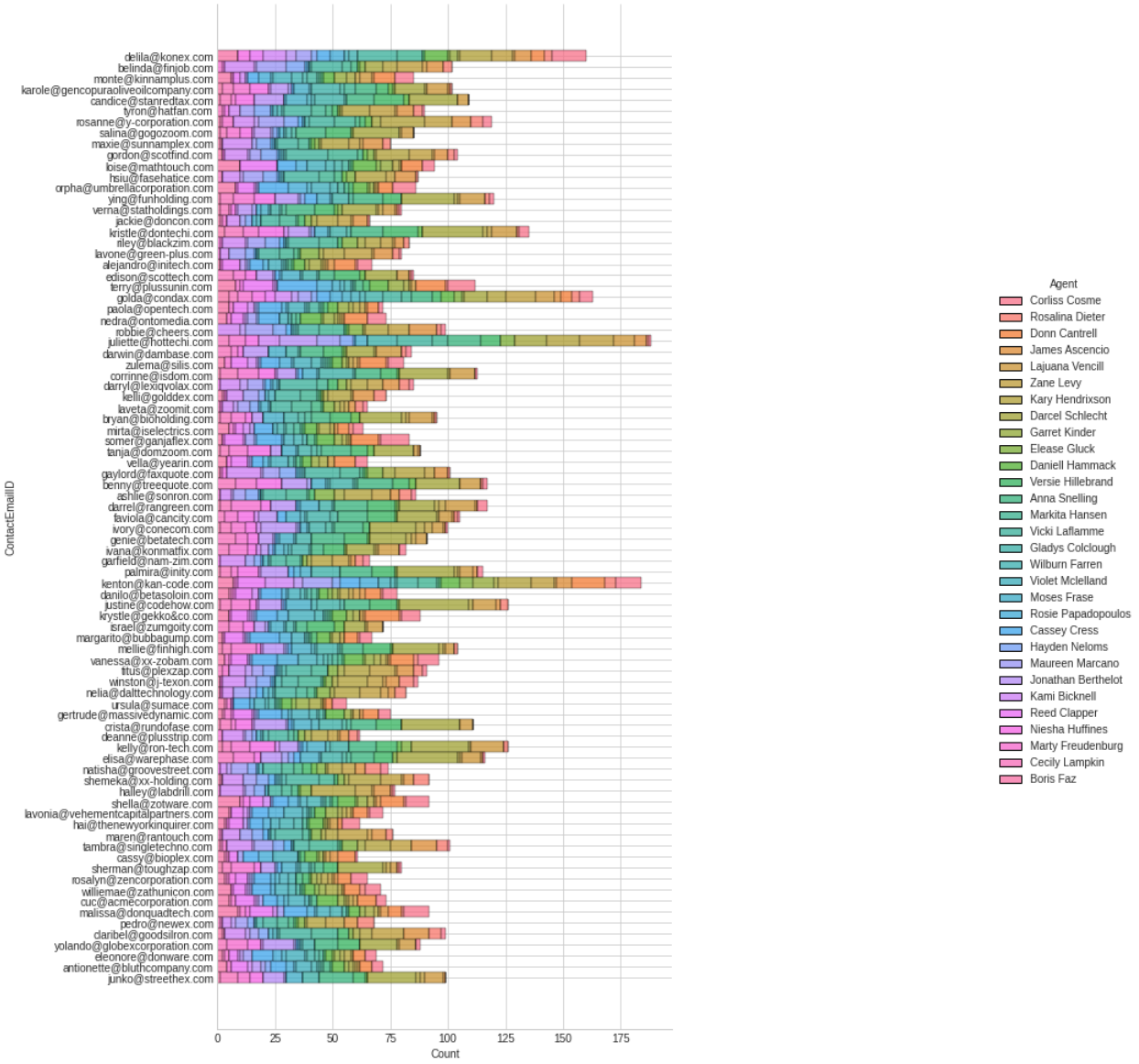


در نمودار زیر عملکر هر agent مشاهده می‌شود که چه تعداد باخت و برد و اینپراگرس دارند.


Corliss Cosme
Rosalina Dieter
Donn Cantrell
James Ascencio
Lajuana Vencill
Zane Levy
Kary Hendrixson
Darcel Schlecht
Garret Kinder
Elease Gluck
Daniell Hammack
Versie Hillebrand
Anna Snelling
Markita Hansen
Vicki Laflamme
Gladys Colclough
Wilburn Farren
Violet Mclelland
Moses Frase
Rosie Papadopoulos
Cassey Cress
Hayden Neloms
Maureen Marcano
Jonathan Berthelot
Kami Bicknell
Reed Clapper
Niesha Huffines
Marty Freudenburg
Cecily Lampkin
Boris Faz
Agent
0
100
200
300
400
500
600
Count
Stage
Won
Lost
In Progress


Correlation Heatmap
Unnamed: 0 1 1 -0.17 -0.016 0.23 0.23 0.18 0.011 0.23 0.23
Unnamed: 0.1 1 1 -0.17 -0.016 0.23 0.23 0.18 0.011 0.23 0.23
Stage -0.17 -0.17 1 0.2 0.1 0.11 0.089 0.0077 0.11 0.11
Close_Value -0.016 -0.016 0.2 1 0.023 0.029 0.02 0.0039 0.027 0.026
Call 0.23 0.23 0.1 0.023 1 0.91 0.77 0.052 0.95 0.95
Email 0.23 0.23 0.11 0.029 0.91 1 0.78 0.049 0.96 0.96
Meeting 0.18 0.18 0.089 0.02 0.77 0.78 1 0.064 0.82 0.81
Sentiment 0.011 0.011 0.0077 0.0039 0.052 0.049 0.064 1 0.25 -0.14
Posetive 0.23 0.23 0.11 0.027 0.95 0.96 0.82 0.25 1 0.92
Negative 0.23 0.23 0.11 0.026 0.95 0.96 0.81 -0.14 0.92 1
Unnamed: 0
Unnamed: 0.1
Stage
Close_Value
Call
Email
Meeting
Sentiment
Posetive
Negative
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

در نمودار زیر تنوع مشتری‌های برای هر agent هست که نشان می‌دهد تقریبا هر customer با بیشتر agent ها در ارتباط بوده است.



آنالیز احساسات بر روی اینتراکشن‌ها

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 -۱ و ۱ به دست امد که هرچه اعداد نزدیک تر به -۱ باشد نگرش منفی تری وجود دارد و احتمال برد کمتر است و بر عکس هرچه به ۱ نزدیک‌تر باشد احتمال برد بیشتر است. حال جمع همه‌ی سنتیمنت ها مشخص کننده‌ی عدد sentiment هست که در دیتا فریم مشاهده می‌شود.

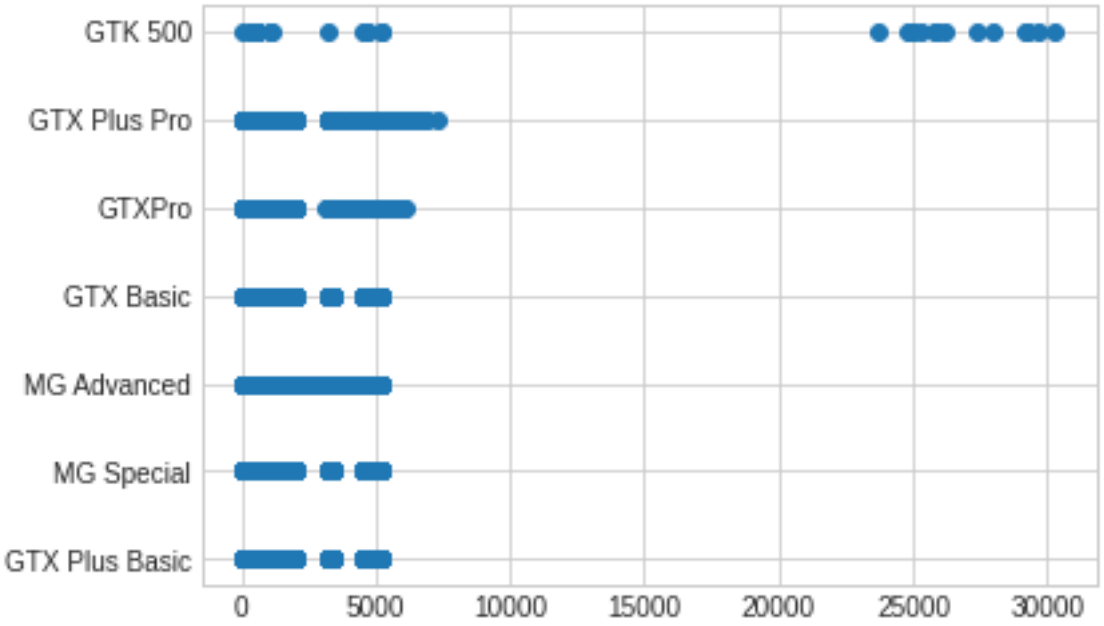
ستون‌های دیتاست

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 ، ۲ بار call و ۱ بار meeting صورت گرفته است. هرکدوم از این ویژگی‌ها نیز به عنوان یک ستون در نظر گرفته شده‌اند.

حذف داده‌های نویزی

در قیمت‌های برای کالای GTK 500 داده‌ی نویزی وجود دارد. با توجه به این که بیشتر قیمت‌ها کمتر از ۱۰۰۰۰ است، پس اعداد بالاتر برای این داده نویز به حساب می‌ایند که باید حذف شوند.

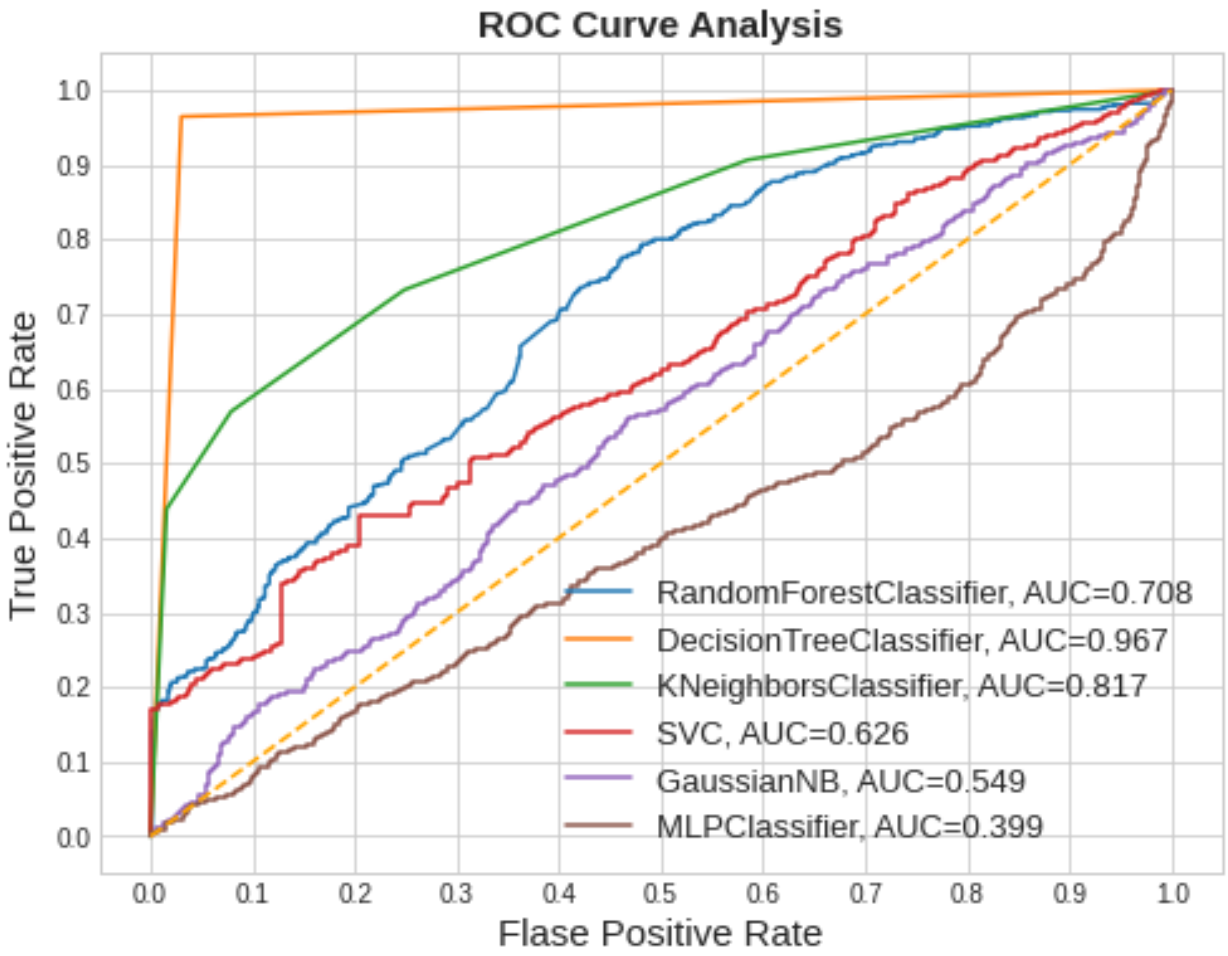


خروجی دیتاست به شکل زیر می‌باشد.

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

Customer	Agent	SalesAgentEmailID	ContactEmailID	Stage	Product	Close_Value	Created Date	Close Date	Call	Email	Meeting	Sentiment	Posetive	Negative	Duration
Rundofase	Darcel Schlecht	darcel@piedpiper.com	crista@rundofase.com	Lost	GTXPro	2000.000000	2017-01-01	2017-08-26	15	20	4	16	28	11	237
Globex Corporation	Darcel Schlecht	darcel@piedpiper.com	yolando@globexcorporation.com	Lost	MG Advanced	601.000000	2017-01-27	2017-10-09	9	27	7	11	28	15	255
Hottechi	Darcel Schlecht	darcel@piedpiper.com	juliette@hottechi.com	Lost	MG Special	50.000000	2017-02-05	2017-09-19	16	26	4	13	29	17	226
Conecom	Darcel Schlecht	darcel@piedpiper.com	ivory@conecom.com	Lost	GTXPro	4514.000000	2017-03-07	2017-12-02	17	26	4	14	31	16	270
Konex	Kami Bicknell	kami@piedpiper.com	delila@konex.com	Lost	GTX Basic	1232.000000	2017-03-19	2017-10-19	8	13	3	12	19	5	214
Conecom	Darcel Schlecht	darcel@piedpiper.com	ivory@conecom.com	Lost	GTXPro	537.000000	2017-04-06	2017-12-09	16	25	4	14	30	15	247
Kan-code	Donn Cantrell	donn@piedpiper.com	kenton@kan-code.com	Lost	GTXPro	1852.915504	2017-05-09	2019-08-22	39	57	8	15	60	44	835

ROC



با توجه به نمودار ROC و سطح زیر نمودار می‌توان متوجه شد که DT بسیار عملکرد بهتری را دارا می‌باشد.

نتایج حاصل از الگوریتم‌ها

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

Algo	F1 score	Accuracy	Confusion matrix
KNN	0.6109215017064846	0.5210084033613446	[[207, 275], [409, 537]])
SVM	0.8051259390190014	0.6911764705882353	[[76, 406], [35, 911]])
MLP	0.8225024248302618	0.7436974789915967	[[214, 268], [98, 848]])
DT	0.9691699604743081	0.9592050209205021	[304, 12], [27, 613]])
NB	0.5517241379310345	0.4992997198879552	[[273, 209], [506, 440]])
RF	0.7969671440606572	0.6624649859943977	[[0, 482], [0, 946]])

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

Algo	F1 score	Accuracy	Confusion matrix
DT	0.9652945924132366	0.9673252279635258	[[675, 21], [22, 598]]

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.

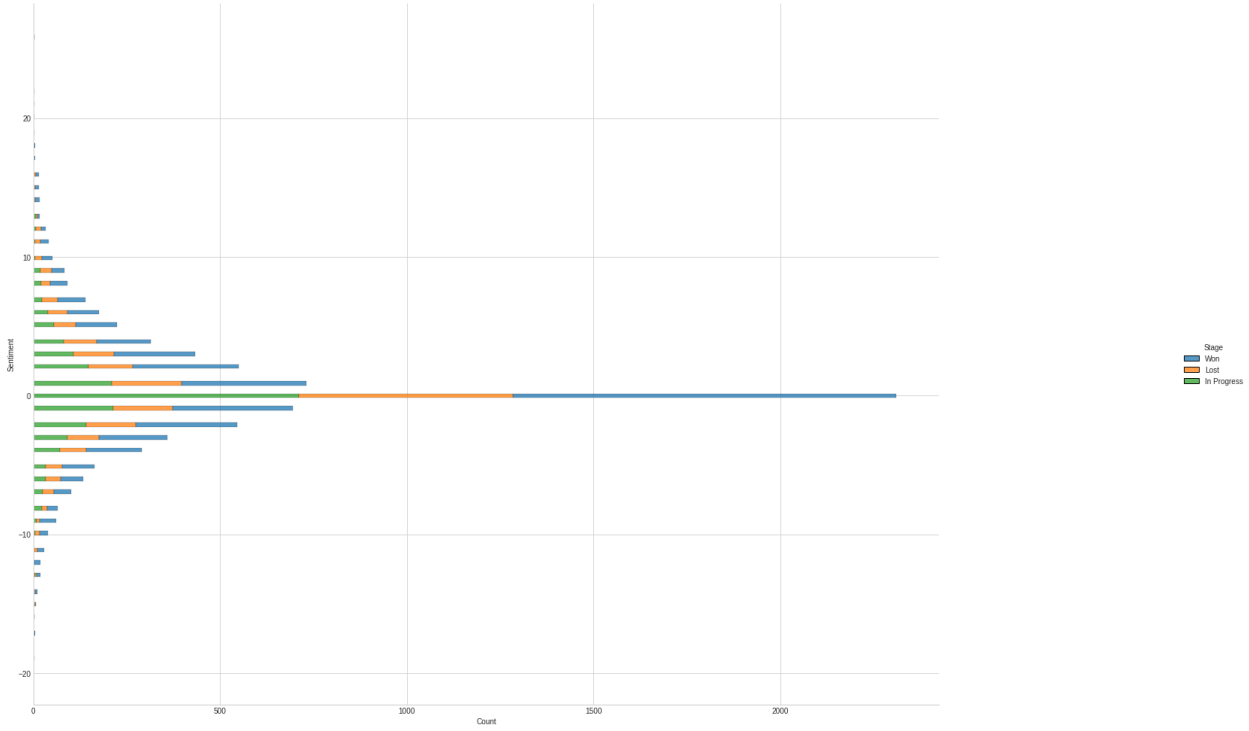
```
{'criterion': 'entropy', 'max_features': None, 'random_state': 0})
```

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 ۹۰ رسید.

پس از آن داده‌های InProgress را با این مدل label زدیم و مدل را دوباره ترین کردیم که دقت مدل به عدد ۸۵ 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