Lead Scoring Case Study

Abstract

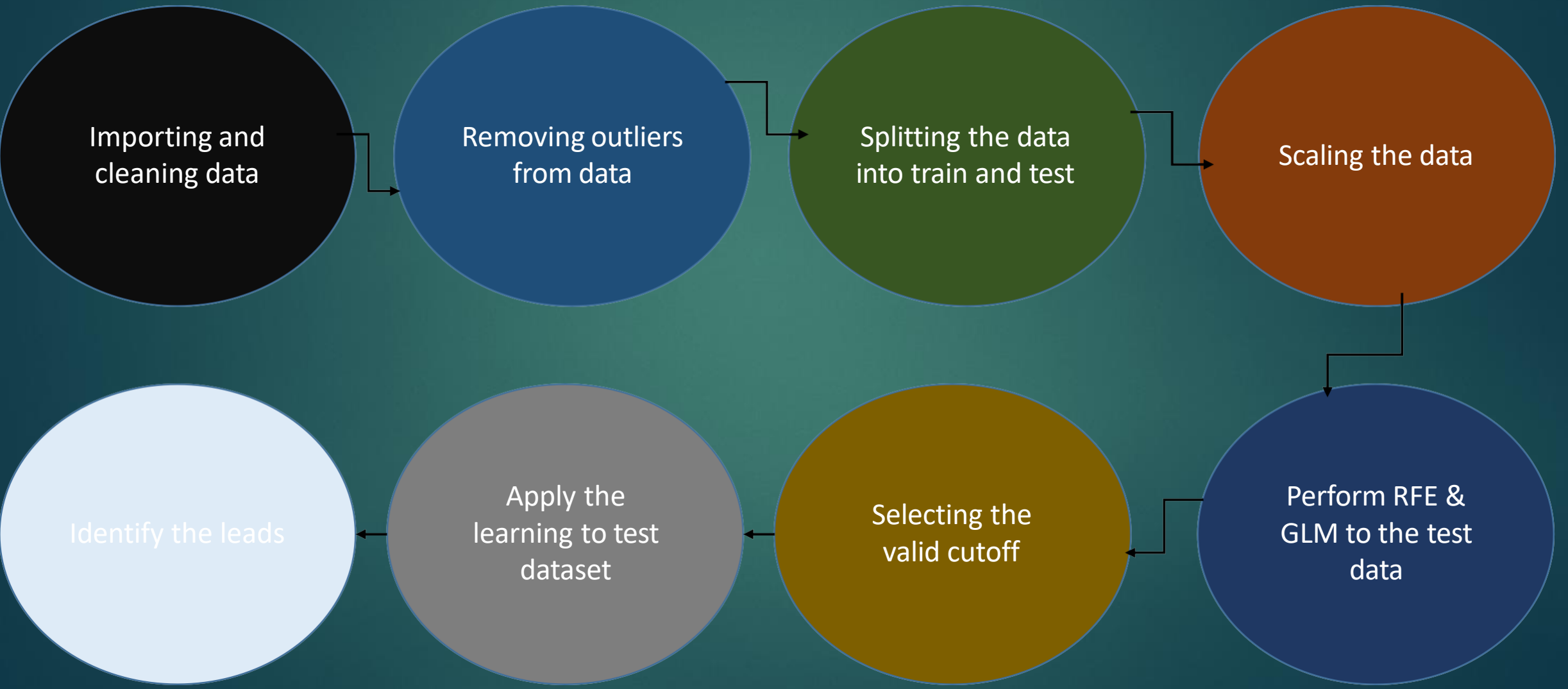
X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. Although X Education gets a lot of leads, its lead conversion rate is very poor.

For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted

Goals:

- As and analyst we need to find the ‘Hot Leads', Our target is to increase the lead conversion to 80%.
- *Building a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads*

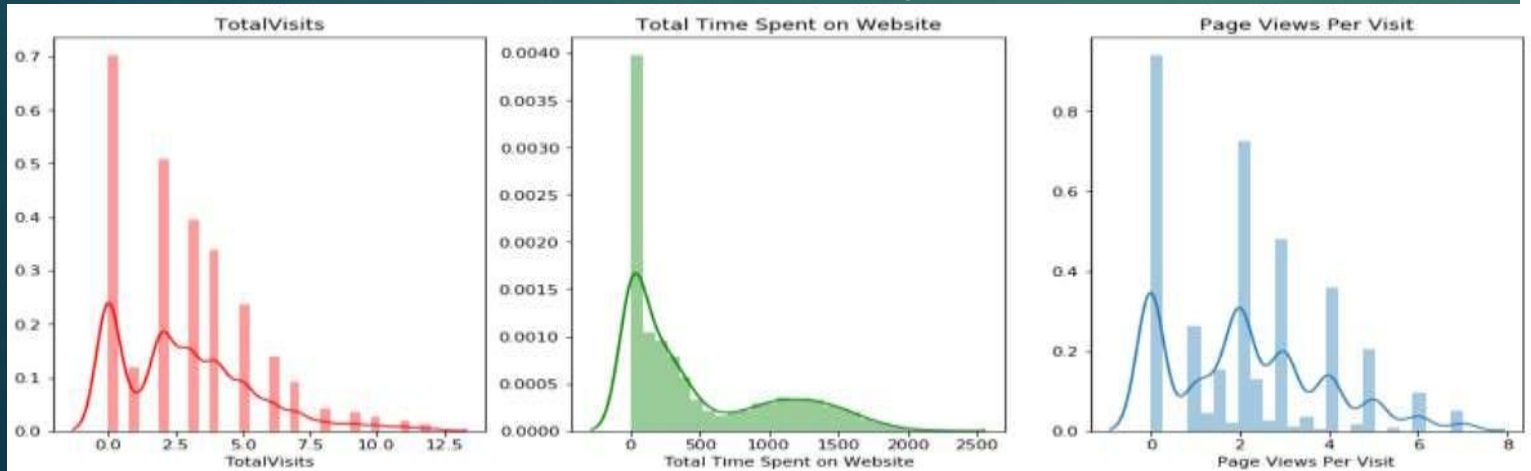
Problem solving methodology

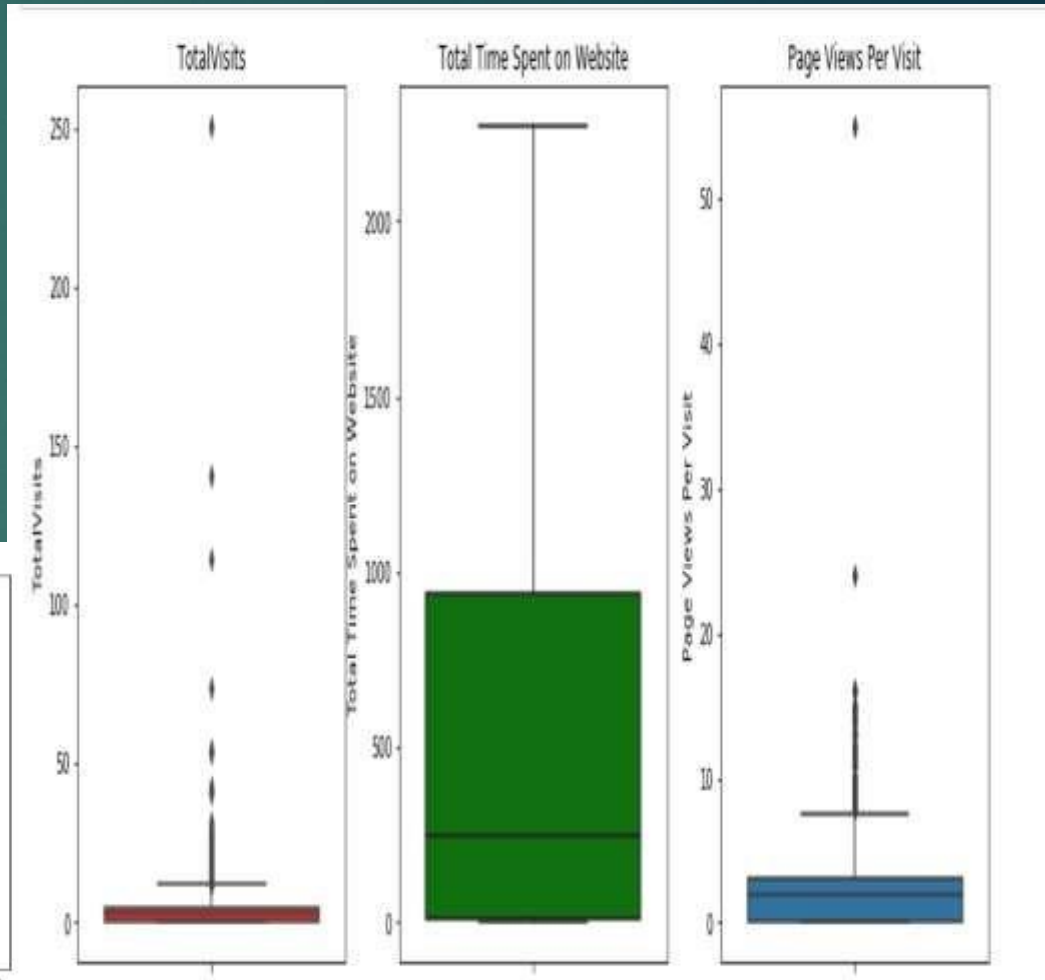


Outlier Removal & Dummy columns

As we can see our data is heavily affected by the outliers so in the first step we have removed the outliers from the dataset then our primary goal is to deal with the categorical feature in this case we are using one hot encoding to create the dummy columns so that we can pass this created columns in our model.

After removing the outliers we can see our data is not normally distributed which will be best for analysis.





Scaling and splitting

Now first we have divided our data into train and test where train will be 70% and test will be 30%.
After train test split we will perform scaling on out dataset as we already know scaling should be done after the train test split .The test set should be untouched we will apply our learning from the trainset to the test data.

We have applied MinMaxScaler our data will lie between 0 and 1 , we can see in the below graph that our Minimum value is 0 and max value is 1

	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Search	A free copy of Mastering The Interview	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Origin_Quick Add Form	...	Last Notable Activity_Form Submitted on Website
count	6190.000000	6190.000000	6190.000000	6190.000000	6190.000000	6190.000000	6190.000000	6190.000000	6190.000000	6190.0	...	6190.0
mean	0.078514	0.241263	0.212865	0.306033	0.000969	0.305654	0.526171	0.082068	0.005493	0.0	...	0.0
std	0.269000	0.211889	0.240765	0.250302	0.031121	0.460721	0.499355	0.274490	0.073915	0.0	...	0.0
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0
25%	0.000000	0.000000	0.003081	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	...	0.0
50%	0.000000	0.250000	0.107835	0.285714	0.000000	0.000000	1.000000	0.000000	0.000000	0.0	...	0.0
75%	0.000000	0.333333	0.408451	0.428571	0.000000	1.000000	1.000000	0.000000	0.000000	0.0	...	0.0
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.0	...	0.0

RFE,GLM & VIF

After 7th iteration we can see we are getting descent P-values and our VIF values are also in control.
Where you can easily identify the top 5 variable
which can explain the regression model

	Feature	VIF
11	What is your current occupation_Unemployed	4.99
1	TotalVisits	3.41
2	Total Time Spent on Website	2.07
4	Lead Source_Olark Chat	2.07
0	Do Not Email	1.83
7	Last Activity_Email Bounced	1.78
3	Lead Origin_Lead Add Form	1.44
9	Last Activity_Olark Chat Conversation	1.42
12	Last Notable Activity_SMS Sent	1.41
5	Lead Source_Welingak Website	1.26
6	Last Activity_Converted to Lead	1.12
10	What is your current occupation_Student	1.11
8	Last Activity_Form Submitted on Website	1.02
13	Last Notable Activity_Unreachable	1.01

Dep. Variable:	Converted	**No. Observations:**	6190
Model:	GLM	**Df Residuals:**	6175
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2541.8
Date:	Sun, 09 Jun 2019	**Deviance:**	5083.6
Time:	11:48:49	**Pearson chi2:**	7.16e+03
No. Iterations:	7	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1841	0.209	0.880	0.379	-0.226	0.594
Do Not Email	-1.3550	0.193	-7.030	0.000	-1.733	-0.977
TotalVisits	0.8394	0.213	3.934	0.000	0.421	1.258
Total Time Spent on Website	4.5473	0.169	26.969	0.000	4.217	4.878
Lead Origin_Lead Add Form	4.0255	0.215	18.708	0.000	3.604	4.447
Lead Source_Olark Chat	1.4240	0.125	11.397	0.000	1.179	1.669
Lead Source_Welingak Website	2.8117	1.029	2.734	0.006	0.796	4.828
Last Activity_Converted to Lead	-1.0544	0.219	-4.814	0.000	-1.484	-0.625
Last Activity_Email Bounced	-0.7405	0.361	-2.049	0.040	-1.449	-0.032
Last Activity_Form Submitted on Website	-0.8435	0.359	-2.352	0.019	-1.546	-0.141
Last Activity_Olark Chat Conversation	-1.5099	0.167	-9.057	0.000	-1.837	-1.183

Validation Metrics

The most important part is to check all the validation metrics which can tell you the performance of your linear model.
Here we can see the accuracy is pretty good but we are not interested about those people who are not converted our model ha identified all the zeros but we are more interested in identifying the people who are actually a hot leads.
Our confusion Metrix is sowing that our model has misidentified 693 people and marked then as not a hot lead.

Metrics

- Accuracy
- Confusion Metrix
- Sensitivity
- Specificity
- False Positive Rate
- Precision
- Recall

Predicted	not converted	converted
Actual		
not_converted	3358	437
converted	693	1702

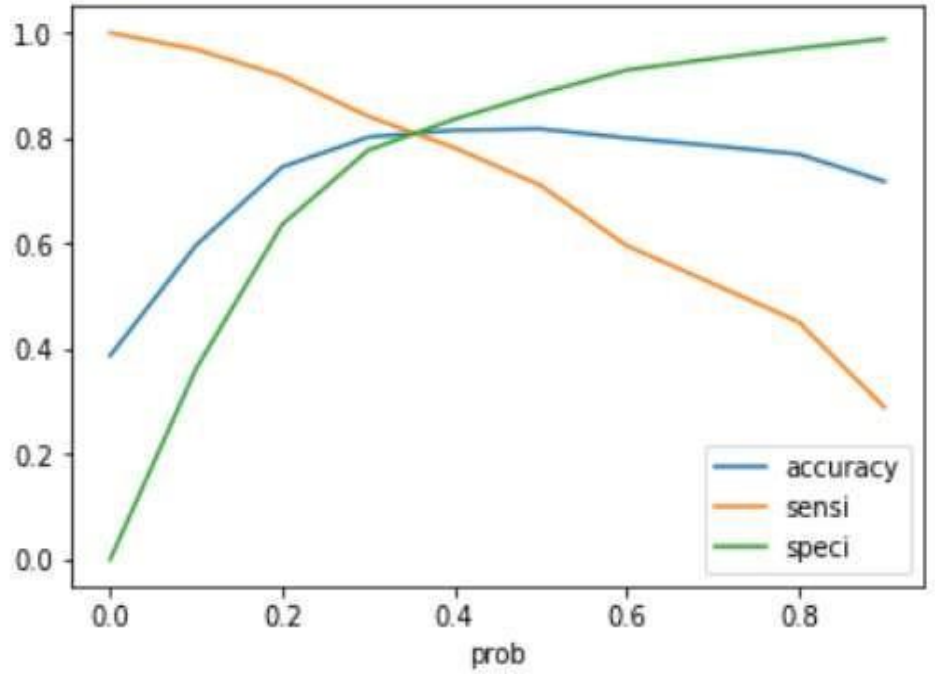
Accuracy	0.817447496
Sensitivity	0.710647182
Specificity	0.884848485
TPR	0.710647182
FPR	0.115151515

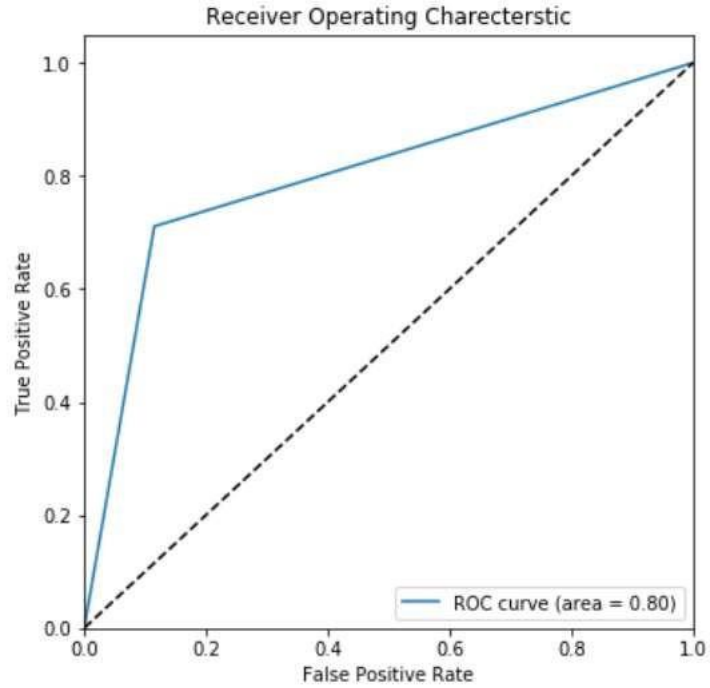
ROC, Cut-off , Sensitivity & Specificity trade-off

Below ROC curve is doing pretty good for our model and AUC is equal to .80,we can see that the with the decrease of FPR the TPR is increasing.

Where as we can see a tradeoff between sensitivity and specificity we can easily identify the cutoff, For our model we have selected 0.3 as our cutoff.

	prob	accuracy	sensi	speci
0.0	0.0	0.386914	1.000000	0.000000
0.1	0.1	0.596931	0.968267	0.362582
0.2	0.2	0.745234	0.918163	0.636100
0.3	0.3	0.801939	0.841336	0.777075
0.4	0.4	0.814863	0.781628	0.835837
0.5	0.5	0.817447	0.710647	0.884848
0.6	0.6	0.800323	0.596242	0.929117
0.7	0.7	0.785622	0.524008	0.950725
0.8	0.8	0.769628	0.451357	0.970487
0.9	0.9	0.717609	0.288935	0.988142





Prediction on Test Data

After applying all the learnings on the test data we can see though our accuracy has decreased but our main target is to increase the sensitivity .Our model is doing quite a good job to identifying a 82% of the hot leads.

Predicted	not converted	converted
Actual		
not_converted	1271	385
converted	172	826

Accuracy	0.790128109
Sensitivity	0.827655311
Specificity	0.606103958
Precision	0.682080925
Recall	0.827655311

Conclusion

For generating the lead score we are multiplying the percentage which we got from the logistic model with 100 , Below we can see the scores for our top 5 leads

Out[101]:

	ID	Converted	Convert_Prob	Final_Prediction	Lead Score
1767	7219	1	0.999719	1	99.97
761	3478	1	0.999493	1	99.95
414	8074	1	0.999454	1	99.95
232	6383	1	0.999465	1	99.95
2040	7579	1	0.999019	1	99.90