Netrality Prospect Company Recommendation Project

Team HDR: Hasibul Jishan, Daniel De Las Heras, and Ryan Pittman

Monday, December 18, 2023

Problem Statement

Our client, Netrality Data Centers, is looking to expand its business by targeting new companies to acquire as customers. They provided us with a dataframe of current customers, a dataframe of billing coefficients associated with those customers at each Netrality location, and a dataframe of prospect companies they could potentially target in the future. The dataset of current customers consisted of 329 rows, each representing statistics associated with companies that Netrality currently does business with. The columns in this table reflected various quantitative and categorical features of those companies such as their revenue, employee growth rates, industry, etc. The dataset of prospect companies consisted mostly of the same columns, but held 1929 rows representing companies that Netrality is not currently doing business with. Lastly, the current billing dataset held the standardized billing coefficients of the 329 current companies at each billing location. These payment metrics were calculated twice; once of lifetime coefficients and another of last month's billing payment coefficients. Our goal was to assess the prospect companies based on the criteria of Netrality's current customers using machine learning models to ultimately identify a consolidated list of companies that would be most advantageous for Netrality to gear their marketing campaigns toward and acquire as clients. It takes time and money to acquire customers. We wanted to determine which of the 1929 prospect companies were actually worth seeking business with in order to optimize profitability.

Summary of Approach

After taking all the necessary steps to clean, standardize, and preprocess the data, we ran a multitude of supervised and unsupervised machine learning models to see which one would provide the most accurate subset of recommended companies from the larger prospect list. Such machine learning models included logistic regression, k-NN, XGB classifier, etc. We started off by creating two binary columns called "label" and "label2" which represent whether each company reported total billing coefficients above average or average and below. Companies reporting above average billing payments were assigned a 1 while all other companies were given a 0. The details on the differences between "label" and "label2" and how they were formulated are discussed later in this paper. Our logic was to use these columns as the response variables in our machine learning models since Netrality would want to target companies that improve their revenues the most. Unfortunately, after running these models and continuously trying to improve them, we came to the conclusion that due to the skewness and nonlinear nature of the datasets, we could not improve our models' accuracy scores past 65%. However, we realized that since these models were run independently of one another, we could take the set difference of all the companies each one recommends and use this set as our most accurate recommended list. The law of independent events in statistics proves that the false positivity rate of our final recommended list drops significantly when taking this set difference.

Summary of Results and Conclusion

As aforementioned, set differences were used to minimize the false positivity rate of the recommended lists produced by each of our individual models since they had relatively low accuracy scores. Each model was run twice with 1,000 iterations; once with using "label" as the response variable and another time with using "label2" as the response variable. The set difference was taken between all the models using the same response variable. This produced two recommendation lists. We then decided to take the set difference between these two lists one more time in order to optimize our accuracy and consolidate the number of recommendations made. As aforementioned, the reasoning behind using two separate response variables will be later discussed. However, it is important to note that the "label2" variable accounts for company size (which appeared to be an unintended confounder in our models) while the "label" variable does not. Overall, we are 99.93% confident that the 298 companies in our final recommended list accurately reflect customers that fit the criteria of current customers who make above average billing payments. The table of our final recommendations is depicted below.

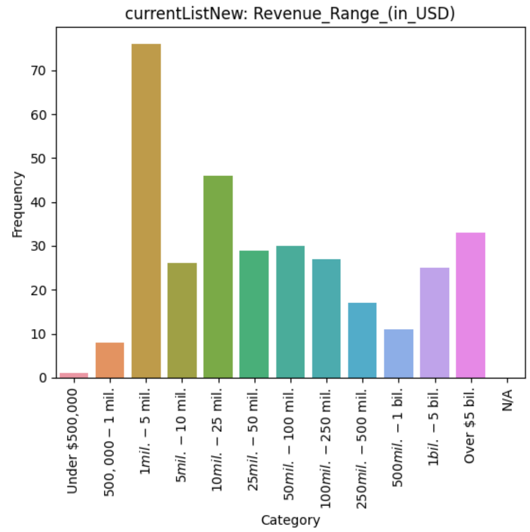
Table of Prospective Customers to Target

Index	CompanyID					
0	2441797	355780297	313524325	116245188	23284479	5784162
1	155353090	49958972	19939690	5851944	129728745	13370133
2	19513364	72238328	37606830	357686568	22922051	355769145
3	39600454	15614775	56199764	56674605	25428225	94765130
4	136118787	3573275	158679592	6278960	348727434	81264919
5	345283492	154239344	40861703	6065973	40157265	1967792
6	5358630	128860355	37847515	35671385	24871911	69371375
7	8110814	10190854	103858861	39305645	353601857	10852424
8	1461214	34818453	29818882	346356055	346089987	9224425
9	9867169	16859276	87820582	62392222	68863214	66864182
10	8590337	1720519	31802379	153011246	65143293	31038037
11	297468076	88376327	54742651	47989454	115147644	371449923
12	3444162	26968154	345620672	72074644	125075912	55953757
13	15712435	93176798	9487723	1524090	10339502	48466099
14	13207636	94568709	16237253	91358593	21203264	19677250
15	7258303	11126273	346921123	346960592	80692712	31474409
16	239305146	116424706	15564820	48616172	6324534	346340982
17	12227700	13140007	23776193	33030147	41615409	48466094
18	60310227	4506176	41976469	353760348	15224727	82562409
19	27722128	168509596	13631525	20317318	13009984	356330880
20	234914216	4280349	61688814	352324273	257884545	100827459
21	1114181	41323685	17402544	14680168	7261816	50955929
22	412002344	144933765	37323650	42747577	63750701	10340178
23	30245334	29344353	2654475	34604820	42536961	347738793
24	14155984	70175811	15860110	4753140	5542350	347078310
25	11127417	9634147	481290639	34065457	61506294	9044145
26	103841907	7834830	2540471	36848621	54167280	344439078
27	62014529	135545775	112951694	131989912	118899830	100090286

28	15794314	164856312	65750983	144245275	119734544	31118573
29	58804259	84099764	19170452	23857792	29693197	89130720
30	2953966	39977791	15319076	20349415	371769443	113672248
31	14516709	348205593	59884285	21007086	36671558	86688809
32	31342638	106676542	4606512	82331778	432282252	26378341
33	1507503	27148954	40026602	17192825	26935736	78120278
34	57705757	22856817	19385483	49187454	400149291	60015931
35	66421453	3834943	43927242	113295822	37732439	39455061
36	14946173	20037711	7209161	26527592	36739880	132055687
37	9751686	22516014	33166475	4409572	11274549	351250618
38	41058369	175292976	124371024	157713259	353609645	157155360
39	24576142	46697605	353608190	115654741	24150322	92521051
40	1804856	195493909	91504888	80871947	28935398	14451207
41	12913103	12272288	345638177	17815664	95568461	77448833
42	56526980	148046227	3183391	114475590	27673265	187957333
43	90883103	51200156	11438187	98599255	347050066	353887720
44	266727	344472790	48301568	41123861	61443556	100811291
45	18633856	22807075	8110529	111641624	230703538	365525965
46	5619763	11240319	30196506	356180489	10605954	89915648
47	24182874	129729226	29231431	13374268	17631006	355627427
48	15877691	100221071	168640828	44828366	347400265	
49	33280815	89440993	13910366	150291170	3557694	

Details of the Modeling and Process Approach

We began our project by conducting exploratory data analysis (EDA) on the three datasets provided by Netrality. We printed the descriptive statistics for the quantitative columns, created various graphs like pie charts and bar charts for the categorical columns, and even constructed a correlation heat map to test for multicollinearity throughout the data. From there we were able to draw some important conclusions. For starters, the correlation matrix allowed us to identify the relationships between certain variables. In our initial models, we avoided incorporating variables with high correlation coefficients based on these findings. Our EDA also allowed us to acknowledge how skewed and nonlinear the datasets were. As shown on the right, this was evident in the histograms we constructed from our quantitative columns along with their respective kurtosis scores from the descriptive statistics.

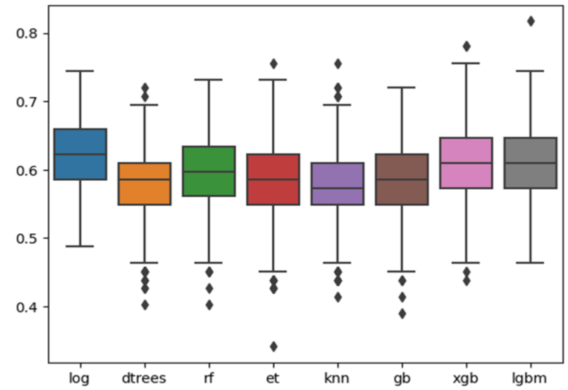



```
Founded_Year                                    8.081917
Revenue_(in_000s_USD)                          86.217632
Est_Marketing_Department_Budget_(in_000s_USD) 130.992717
Est_Finance_Department_Budget_(in_000s_USD)    73.310792
Est_IT_Department_Budget_(in_000s_USD)         72.644307
Est_HR_Department_Budget_(in_000s_USD)        225.733661
Employees                                     201.276473
Past_2_Year_Employee_Growth_Rate               39.860218
Alexa_Rank                                      0.070918
Total_Funding_Amount_(in_000s_USD)             89.942488
Recent_Funding_Amount_(in_000s_USD)           147.241166
Number_of_Locations                            50.052211
dtype: float64
```

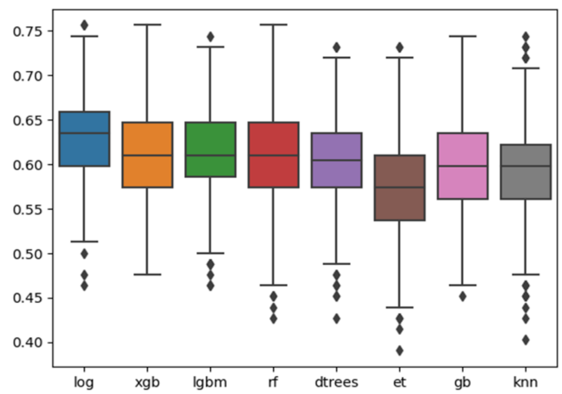
Identifying this skewness led us to our next procedure; preprocessing the data. Our first step was to merge the billing and current customer datasets so that we could pair each company ID to its billing coefficients across each billing location. Then, even though the billing coefficients were already standardized by Netrality, we decided to log transform the data to account for skewness. Our next step in preprocessing was to sum the lifetime billing coefficients for each company and make a new column in the dataset with the totals. We chose to use the lifetime billing coefficients over last month's billing coefficients since we believed that the lifetime billing data better encompasses the billing history of each company at their billing locations and that last month's data could simply be an outlier. We then created a binary "label" column which assigned a 1 to companies with above average total lifetime billing sums and a 0 to all other companies that made average or below average billing payments. We also created another column called “label2” which took company size into consideration by dividing the lifetime billing coefficient sums by the number of Netrality locations each current company occupies. Companies with an average lifetime billing sum higher than the median for any given location were assigned a 1 in the “label2” column while all other companies received a 0. Accounting for company size in such a manner was important since we noticed that our models were skewing their recommendations toward larger companies that had the revenue to do more business with Netrality. These two label columns would be the response variables in machine learning algorithms we ran thereafter. This is because Netrality, like all other companies, wants to target customers that would optimally strengthen their top line. Finally, we ended our data preprocessing measures by filling N/As with appropriate values and removing outliers. We also altered some of the features in our dataset to enhance their influence in our machine learning models. For example, we changed the “founded year” column to reflect the age of the company in years by subtracting each companies' founded year from 2023.

Once our preprocessing steps were completed, we began running various models and evaluating their performance. We started with unsupervised machine learning models such as k-means clustering to help us identify groups that exist among the current companies based on the features of our dataset. Afterall, if we could identify the features in common among all current companies with a 1 in their “label” column, we could query the prospect list for companies with similar features. Once we realized that those groups exist among the current customers, we decided to pursue various supervised machine learning models for our final model.

A 25%/75% train/test split was used to fit all of our supervised machine learning models and each one was run twice with 1,000 iterations using either “label” or “label2” as the response variable. We began by running multiple linear regression models on our dataset to identify which companies to recommend and the projected billing coefficients they should have if Netrality were to acquire them as clients. However, due to the nonlinear nature of the dataset, the R-squared and accuracy scores of these models were very poor; even after optimizing feature selection based on our correlation heat maps and log transforming the data to account for skewness. Therefore, we decided to try a variety of other classification and regression models to see if we could improve our accuracy scores. Those models included: logistic regression, k-NN, decision trees, XGB classifier, LGBM classifier, extra trees, gradient boost classifier, and random forests. While the accuracy scores of these models were a significant improvement from the multiple linear regression models, they still failed to surpass 65%. These individual accuracy scores are depicted in the boxplots shown to the right. We quickly came to the conclusion that no matter how much additional preprocessing we performed, the accuracy scores of our individual models would not continue to increase. Therefore, we leveraged the law of independent events in statistics to minimize the false positivity rate of our final recommendation list. This was done by running all of the aforementioned models using "label" as the response variable and taking the set difference of all the companies each model recommended. The same was performed using "label2" for the response variable as well. Then, one final set difference was taken between those two recommended lists. All of the individual models that were incorporated in our "best model" are shown in the table below along with their performance statistics.



"label"



"label2"

Model Metrics:

	Dataset	Iterations	Accuracy	Precision	Recall	ROC
log	Binary Label 1: Above Average Billing	1000	0.6195243902439020	0.6463322699041460	0.6134214818202210	0.6224625710406190
dtrees	Binary Label 1: Above Average Billing	1000	0.579060975609756	0.6278835559870410	0.5965723386136260	0.5842922582514270
rf	Binary Label 1: Above Average Billing	1000	0.598109756097561	0.6333298705961670	0.5850409861296030	0.6024469774281960
et	Binary Label 1: Above Average Billing	1000	0.5808902439024390	0.597896806561957	0.6158254179198480	0.5812164058831300
knn	Binary Label 1: Above Average Billing	1000	0.578670731707317	0.6635935659046890	0.4022878871576310	0.5886196181893490
gb	Binary Label 1: Above Average Billing	1000	0.5816707317073170	0.5967617012026230	0.7225390417698560	0.5862241636680160
xgb	Binary Label 1: Above Average Billing	1000	0.6107439024390240	0.6261822895975850	0.6449590262029140	0.6111729210905140
lgbm	Binary Label 1: Above Average Billing	1000	0.6144146341463420	0.6296356498155190	0.6478615410582600	0.6148036011882580

	Dataset	Iterations	Accuracy	Precision	Recall	ROC
log	Binary Label 2: Above average billing per location	1000	0.6282073170731710	0.6454686132474380	0.5865442665168870	0.6307927218332030
xgb	Binary Label 2: Above average billing per location	1000	0.609719512195122	0.6121429265049470	0.6137670291986840	0.6114428921794250
lgbm	Binary Label 2: Above average billing per location	1000	0.6128292682926830	0.6162420949826670	0.6131596046922980	0.614596385350464
rf	Binary Label 2: Above average billing per location	1000	0.6092926829268290	0.6517314246566310	0.5292497077224650	0.6124123968709700
dtrees	Binary Label 2: Above average billing per location	1000	0.6028414634146340	0.6746082517038540	0.455908862630144	0.6047355173560370
et	Binary Label 2: Above average billing per location	1000	0.574219512195122	0.627285593749522	0.3758552440105350	0.5756882727761510
gb	Binary Label 2: Above average billing per location	1000	0.5982317073170730	0.6352977483853110	0.5537485466063450	0.6077586035960300
knn	Binary Label 2: Above average billing per location	1000	0.592890243902439	0.668547370945701	0.3766347282166030 0	0.5944837854740500