Netrality Data Analysis Report

Applied Machine Learning - DSC 681
Prof. Jared Mroz

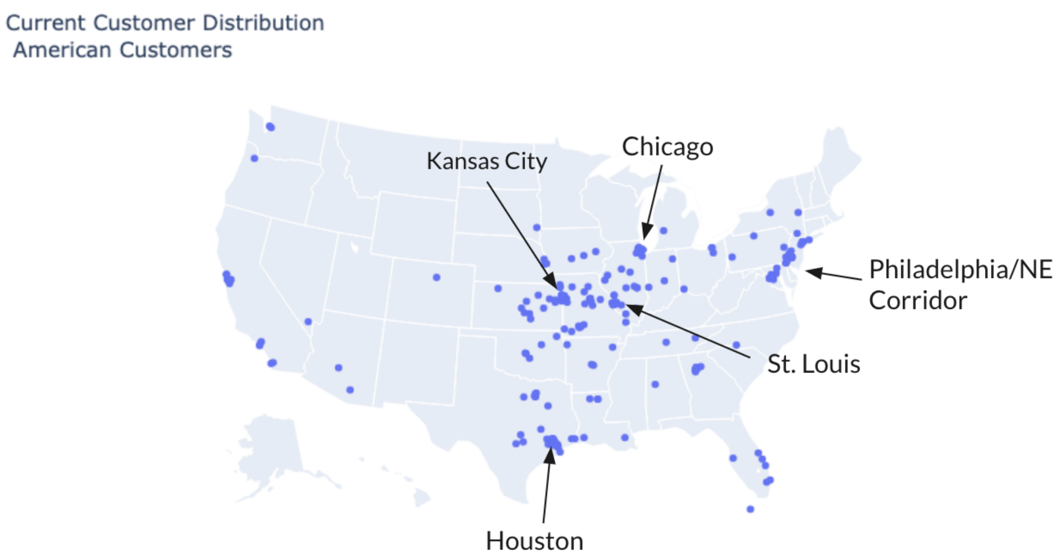
Team JYS
Joe Gallagher, Yashasvi Bhati, Shivam Bhagat
December 18, 2023

Problem Statement

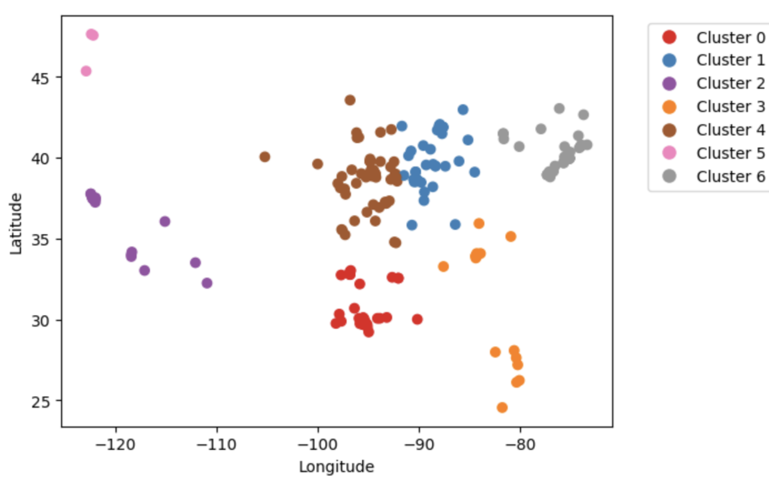
1) Can we predict which prospective customers are most likely to convert, based on their similarity to Netrality's current customers?
2) Of those prospective customers, can we predict the monthly revenue they would generate for Netrality?

Summary of Approach

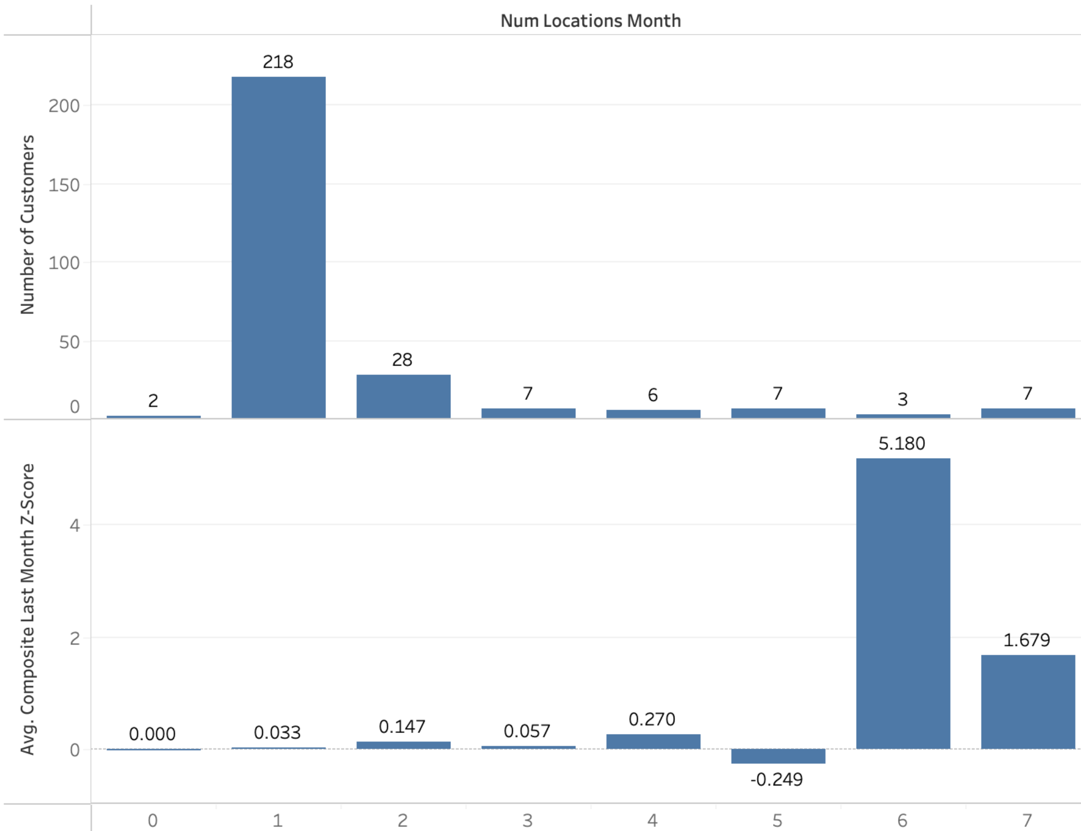
JYS leveraged the geodata of each of Netrality's customers and prospects to look for specific regions where Netrality Sales should focus their efforts. The location of prospective customers in relation to Netrality's Data Center locations is a feature that JYS strongly feels can be used to select strong candidates that Netrality can convert into valuable customers.



We used the zipcode of each customer or prospect to assign each a latitude and longitude. Using these coordinates, we mathematically clustered their locations. These clusters, mirroring regions within the continental United States, were used as additional features in our modeling efforts to predict the monthly revenue that Netrality would expect to see from individual customers.



As we attempted to model potential revenue, it became apparent to us that there are other indicators of a customer's lifetime value that would be important to predict. Using joined current customer and current billing data as a base, we assigned each customer a value for the number of data centers that they were present in during the last billing cycle. Again using geographic and financial data, we attempted to predict the number of Netrality locations that a customer would place themselves in.



Summary of Results and Conclusions

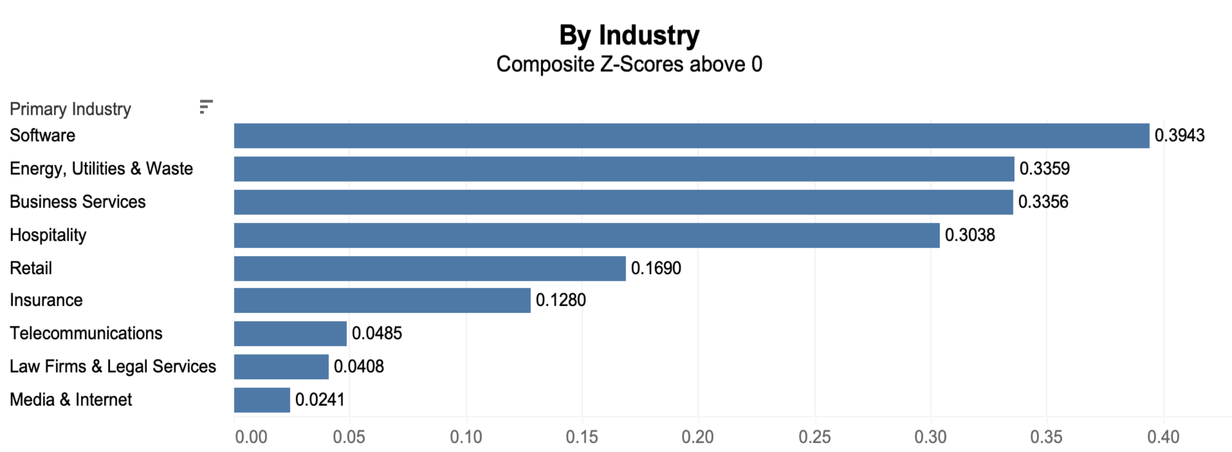
Target prospects headquartered in Indiana

Based on our analysis of the locations of each of Netrality's current customers, JYS determined that Netrality should seek out prospective customers based in Indiana. When plotting each customer's location, we'd noticed that Netrality's customers are primarily located in the same metropolitan areas as Netrality data centers. **This is true for all Netrality data centers except for Netrality's Indy Telcom Center.**

- Texas (Houston Data Center): 61 Customers
- Missouri (St. Louis and Kansas City Data Centers): 56 Customers
- Pennsylvania (Philadelphia Data Center): 22 Customers
- Illinois (Chicago Data Center): 22 Customers
- Kansas (Kansas City Data Centers): 22 Customers
- Indiana (Indianapolis Data Center): **3 Customers**

60% of Netrality's customers are located in Texas, Missouri, Pennsylvania, Illinois, or Kansas. All have identifiable clusters of customers, presumably providing a revenue stream at each of the data center locations that they surround. This is a trend that Netrality should seek to continue at its Indy Telcom Center.

While there are only 3 active customers located in Indiana, Netrality has 57 prospective customers located in the Hoosier State. But of these 57, who are the best to target? JYS analyzed the average total last month revenue Z-Score for all customers by primary industry. In doing so, we were able to discern which industry had customers which generated above average revenue for Netrality.



When we filter down the 57 Indiana prospects to those industries which contain higher that average paying customers, we are left with the following 31 prospects to target:

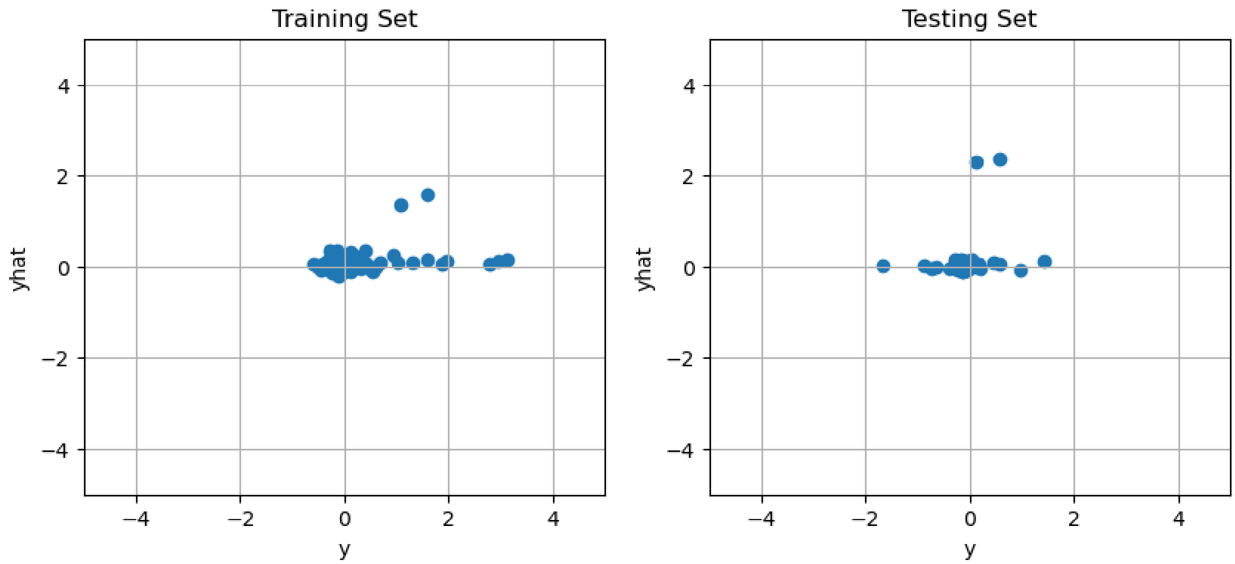
Table of Prospective Customers to Target

9225195	*32237127*	*44550034*	*77906104*
24986484	*39455353*	*66831197*	*77906104*
27444162	*40208956*	*68111514*	*85590595*
95125692	*112164926*	*343499768*	*346400019*
352040245	*353579035*	*354546833*	*354601525*
357080017	*358612313*	*365525965*	*372338553*
397659115	*403915494*	*500997527*	*507377945*
547323117			

Details of the Modeling and Process Approach

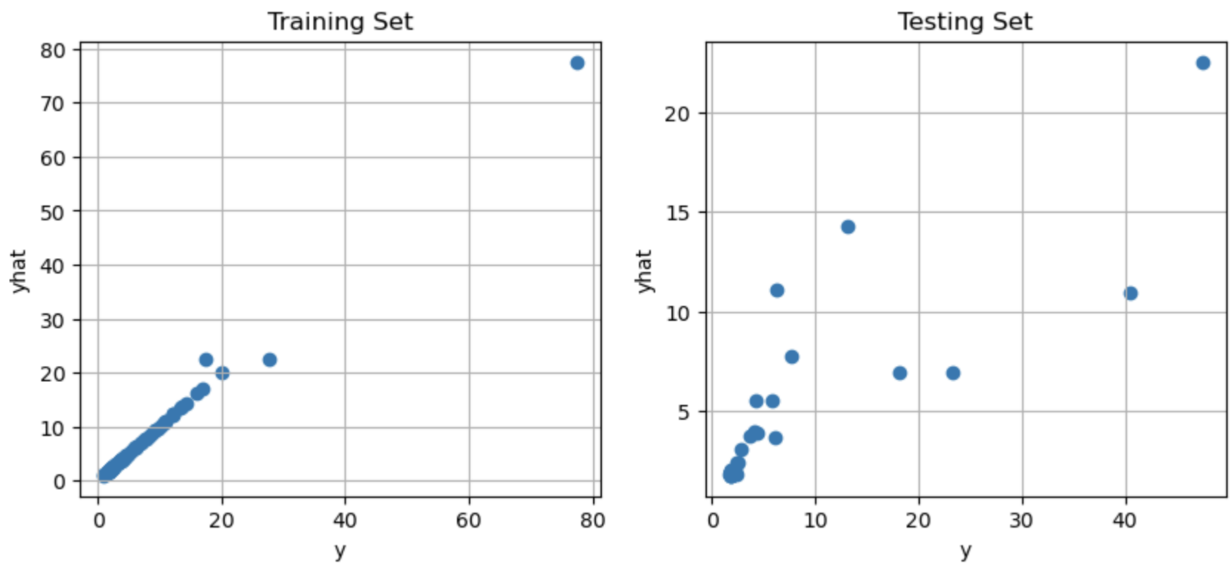
While JYS was able to discern a list of priority prospects based on our exploratory data analysis, we were still left to model predictions of a customer's monthly payment toward Netrality, and the number of data center locations they would place themselves in.

Initially, we had tried to use linear regression models, which produced poor results as seen below in the plots of a Ridge regression model used to predict a customer’s monthly spending Z-Score:




```
Training Metrics:
R squared: 0.12561163154594546
Mean Absolute Error: 0.22940300299864175
Mean Squared Error: 0.23046540107049046
Root Mean Squared Error: 0.4800681212812307

Testing Metrics:
R squared: -0.948727951161622
Mean Absolute Error: 0.2607930616308785
Mean Squared Error: 0.21936177303839013
Root Mean Squared Error: 0.46836072960741504
```

After moving on to regression tree models, we saw models with greater accuracy, although they did not reach the skillful model threshold of an R-Squared value of 0.7. The below plots are the training and testing results of a Decision Tree regressor to predict monthly spending Z-Score. The features we selected as predictors were the regional clusters labels we assigned from the geodata, the label of the data center location where the customer had the highest Z-Score, the ratio of a customers IT Budget to their revenue, and the total number of data center locations they were in.



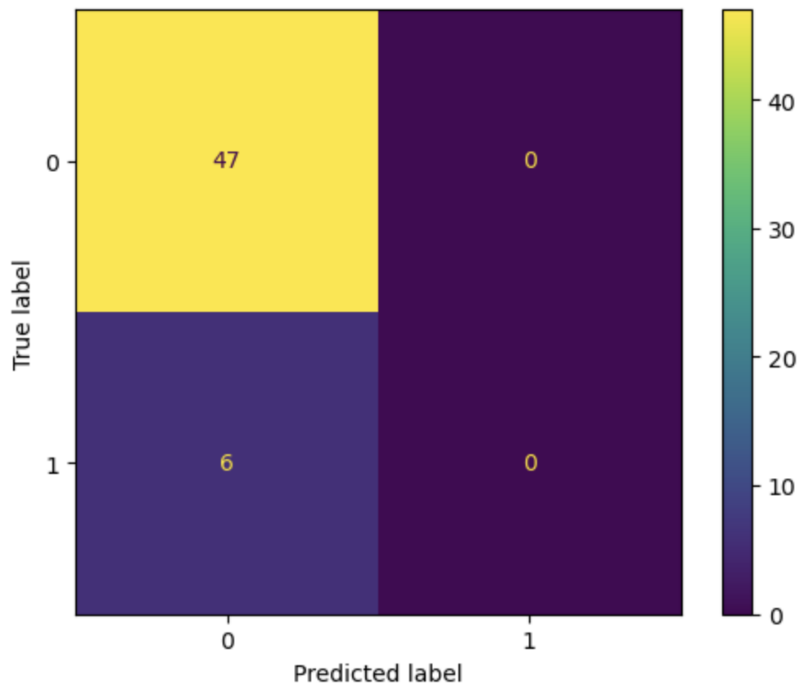

```
TRAINING METRICS:
R squared: 0.993532780616302
Mean Absolute Error: 0.04683427615315316
Mean Squared Error: 0.24303014647342586
Root Mean Squared Error: 0.492980878405467

TESTING METRICS:
R squared: 0.5242092301118387
Mean Absolute Error: 1.7527113622857147
Mean Squared Error: 34.41412513991037
Root Mean Squared Error: 5.866355354043119
```

While this model proved to be more accurate, there was still improvement to be desired, and we were experiencing issues with large amounts of error in the model predictions despite its higher accuracy.
We finally moved on to using the Random Forest regression model, which produced our best model to date.

We later discovered that our model's accuracy was being manipulated by the inadvertent inclusion of the 'num_locations_month' field, denoting the number of Netrality locations a customer was in during the last billing cycle, into our labels that we were attempting to predict.

After correcting this we were again met with a poor model, with negative R-squared values, and again had to rethink what we wanted to predict.

We decided to attempt a logistic regression model that would predict if a customer's Monthly Composite Spending Z-Score was positive or negative. Positive would indicate that the customer spent more than the average customer on Netrality's services, and negative would indicate that they spent less than average.




```
Accuracy: 0.8867924528301887
Precision: 0.0
Recall: 0.0
```

This attempt merely feigned accuracy, as the number of customers who had a negative Z-Score (indicated by the 0 label) far outnumbered those who had a positive Z-Score (indicated by the 1 label).

Ultimately, none of our attempts to model the income of current customers based on their financial and location data proved successful. We believe that more data is needed to account for the possible differences in payment structures and service plans that could vary from customer to customer.

Because of our failure to accurately model the generated revenue of Netrality customers, our suggestions of preferred prospects are based entirely on the exploratory data analysis that we performed throughout the project.

Classification Algorithms

Name	**DataSet**	**Iterations**	**Accuracy**	**Precision**	**Recall**	**Best**
Logistic Regression	*Combined Current Customers and Billing Data*	*100*	*.86*	*.0*	*.0*	*No*

Regression Algorithms

Name	**DataSet**	**Iterations**	**R2**	**MSE**	**MAE**	**Best**
Linear Regression	*Combined Current Customers and Billing Data*	*100*	*-0.9*	*.26*	*.21*	*No*
Decision Tree Regressor	*Combined Current Customers and Billing Data*	*100*	*0.52*	*34.4*	*1.75*	*No*
Random Forest Regressor	*Combined Current Customers and Billing Data*	*100*	*-1.19*	*0.05*	*0.15*	*No*
K-Nearest Neighbors	*Combined Current Customers and Billing Data*	*100*	*-0.08*	*0.03*	*0.1*	*No*
Extra Trees Regressor	*Combined Current Customers and Billing Data*	*100*	*-1.07*	*0.05*	*0.14*	*No*
Gradientboosting Regressor	*Combined Current Customers and Billing Data*	*100*	*-0.1*	*0.03*	*0.1*	*No*
XGBRegressor	*Combined Current Customers and Billing Data*	*100*	*-1.46*	*0.06*	*0.14*	*No*

Company ID	Company City	Primary Industry	Ownership Type	Business Model	Employees	Number of Locations	Est IT Department Budget (in 000s USD)	Revenue (in 000s USD)	Total Funding Amount (in 000s USD)
507377945	Indianapolis	Hospitality	Private	B2B	6,440	15	285,192	1,140,768	0
365525965	Indianapolis	Software	Private	B2B	1,886	11	20,493	553,880	0
352040245	Carmel	Business Services	Private	B2B	2,000	5	17,060	487,455	2,000
357080017	Indianapolis	Business Services	Private	B2B	3,100	11	14,315	409,028	0
354601525	Bloomington	Media & Internet	Private	B2B	1,600	2	10,005	357,356	2,000
112164926	Indianapolis	Media & Internet	Private	B2B	289	4	4,407	83,161	0
32237127	Indianapolis	Business Services	Public	B2B	408	11	3,925	112,171	0
77906104	Ellettsville	Telecommunications	Private	B2B	200	8	2,734	73,909	0
350690094	Evansville	Telecommunications	Private	B2B	274	32	2,079	56,194	0
39455353	Ligonier	Telecommunications	Private	B2B	150	14	2,050	55,432	1,000
95125692	Indianapolis	Software	Private	B2B	125	11	1,036	22,045	0
44550034	Indianapolis	Telecommunications	Private	B2B	74	2	1,016	27,473	13,000
85590595	Hebron	Telecommunications	Private	B2B	91	7	985	26,633	1,000
343499768	Fort Wayne	Software	Private	B2B	119	7	855	23,127	0
397659115	Indianapolis	Software	Private	B2B	100	5	777	21,006	72,288
353579035	Plainfield	Telecommunications	Private	B2B	35	2	265	7,186	0
27444162	New Lisbon	Telecommunications	Private	B2B	11	4	225	6,094	0
354546833	Fort Wayne	Telecommunications	Private	B2B	19	6	219	5,933	0
346400019	Indianapolis	Software	Private	B2B	10	6	160	4,344	1,100
372338553	Carmel	Software	Private	B2B	20	3	155	4,202	18,007
68111514	Westfield	Telecommunications	Private	Null	11	4	130	3,519	0
358612313	Peru	Telecommunications	Private	B2C	19	3	127	3,458	0
66831197	Noblesville	Telecommunications	Private	B2B	14	2	123	4,422	0
40208956	South Bend	Telecommunications	Private	B2C	13	2	115	4,138	0
403915494	Evansville	Telecommunications	Private	B2C	12	1	92	2,497	0
547323117	Carmel	Business Services	Private	B2C	6	1	72	2,069	0
500997527	Kokomo	Telecommunications	Private	Null	5	1	42	1,144	0
24986484	Indianapolis	Telecommunications	Private	B2B	6	2	42	1,157	19
9225195	Indianapolis	Business Services	Private	Null	4	1	25	539	0