DSC-681-01: Applied Machine Learning

Professor Mroz

Project reflection report

Submitted by Yashasvi Bhati

12/11/2023

Project Reflection Report - Netrality Data Centres

Introduction:

The primary goal of this project is to analyze how Netrality can identify an unrecognized pattern in its current customer base and use that information as a road map to generating more revenue from customers from the list of prospects. For the same, the class was provided with all the necessary datafiles by Netrality through Prof. Mroz which consisted of all the potential data that would play a significant role in identifying and predicting the next steps of Netrality in generating the leads.

Potential customer identification:

One of the best ways to identify which prospects can be turned into revenue making customers is to understand how they overlap with the current ones in various factors such as industry, location, service, budget, etc. Thus, team JYS's approach to getting kickstarter was to ensure that each team member works and spends enough time on each of the datasets to ensure and to identify already existing patterns. Through this analysis, there were certain key points discovered such as the use of Z Scores, proportion of US prospects being greater than in other

countries, etc. These key points not only helped the team understand Netrality's present position but also directed our analysis and findings. As we moved further into the weekly sprints, our analysis was driven by the findings in the past weeks.

The initial approach:

We performed EDA and noticed that there were plenty of NAN values. The team decided to convert them to 0's and not remove them from the dataset as it could possibly result in data becoming redundant and biased. The team focused and dedicated time in understanding the dataset and what we were looking at. We were also able to work on questions such as:

- The serving locations of Netrality
- Revenue per location last_month and lifetime
- Netrality do have fully credited customers
- No billing for recently acquired data centers like Pennsylvania

The datasets consisted heavily of numerical features than categorical ones and through our EDA, we were able to identify some of the key aspects like largest and least revenue made, primary and sub industries of current and prospective customers, concentrated locations of prospects, no.of employees highly correlating with the size of the company and their IT budgets, negative composite Z scores for Total lifetime, etc.

The approach with regards to methodology:

Since Netrality has physical data centers locations across the country and oceans, team JYS''s approach was to perform Geo-data clustering and try various machine learning algorithms and techniques on these geo-clusters to direct and lead our findings for this project. We believe that this can be a potential breakthrough to the problem statement as these geo-clusters will not only

plot on the axis but on the US map. The customer was divided into In-market and Out-of-market customers which resulted in the team targeting the In-market customers for our upcoming sprints. In-market customers are the ones who are located closer to the Netrality hubs and Out-of-market are located comparatively farther away from these hubs. Team used Geoclsuter and Plotly The findings from geo-data clustering with different algorithms is as follows:

- Distribution of current and american prospect customers
- Excluding outliers did not seem to be a good idea
- Identifying regional clusters
- Major customer hubs are centered around major metropolitan area
- 56 prospects are headquartered in Indiana
- Identifying futures hub of the prospects

Despite a few drawbacks of this approach such as not being successfully able to cluster multiple features, I personally believe that our method and approach to the problem turned out to be a decent approach as it makes it easier to follow along in real life too ! (Supporting statement: Rob flying to Indianapolis)

The data-centric way:

While talking about mathematically clustering of Netrality's current and prospects using K-means and fitting linear models to the training and testing data, team JYS did slow down on progress as the model was unsuccessful at making predictions on the customer's relative revenue for Netrality. However, the approach motivated us further to re-work on our linear models and apply regression tree analysis to our geographic and derived statistics.

Using the binary classification for segmenting the In-market and Out-of-market was again a good enough approach to be able to fit the data frame into clustering algorithms and the results were as expected. Moving further and when looking at the billing spends across each data center, there were a considerable amount of NAN values which we converted to 0 as we did not want the data to be compromised. This practice also lead to the team JYS experiencing numerous errors when running the regression tree analysis later on.

The IT and finance budgets of the prospects is something that Netrality should be concerned of and thus we created the variable 'Nerd_budget' which bridges the gap between prospects features and the current billing and spending patterns to identify and establish any similar patterns if they exist.

The training data turned out to be as good and with a Rsqaure of 0.99 however the testing data follows a set pattern and then fails with a Rsqaure of 0.52 with an MSE of 34.44. This did emphasize that our model had multiple errors which we wish to work upon in the upcoming final to be called a 'skillful model'.

Following week, we were able to represent more advanced regression models out of which Random forest proved to be the best with Rsqaure value of over 0.7. While running this model, we were predicting composite z-scores of sales revenue for customers across all Netrality customers. With over a 100 iteration, the training data did decent enough and our test metrics turned out to be Rsqaured of 0.71 which can be considered as a skillful model. To add on, this time our MSE was also as low as 13.79 compared to last. Team also played along with the features like IT and finance budget trying to predict the ratios and whether where they are located. Furthermore, the team also tried Feature Analysis to identify and see what the model decides the best features can be for this statement.

Henceforth, I do personally believe that this problem can be solved in a data centric way and it would just need a complete data with low errors and multiple runs of the various regression models.

Data availability:

Datasets provided by Netrality were enough when performing first hand procedures like EDA, PCA and data cleaning. However, as we delved deeper into the pool of composite Z-scores, not null values and zip codes for plotting the longitude and latitude, the team did realize that there could be a possibility of results turning out to be different than what they are now. Inclusion of all the Nan values might have bent the data into different directions and missing zip codes might have missed a potential budding prospects hub on the geo cluster. Team JYS was able to identity the following data missing:

- Use of composite z-scores in current billing
- CustomerID in current_billing is just 328 while the current_customers are 329 in total. Thus a discrepancy
- Few companies were fully credited and thus charged differently or not at all
- Not all the customers in the data list have city names and zip codes

Conclusion:

While the team felt that missing information was a great challenge as our approach was based on geo clustering and then using ‘km_labels’ to store all the geo-data, we do look forward to rising from the challenges and meeting the goals for this project.