**Project Name: - Predicting Potential Prospect Customers**

**Team Name: - The FLASH**

**Authors: Sai Pranay Tummala and Varun Manjunath**

**Date: - 05/13/2024**

**Problem Statement**

The main Business problem presented by *“Netrality Data Centers”* was to identify and acquire potential customers from the Prospect Customers list. We have applied and utilized various Machine Learning techniques to perform the task for our Project. Four essential CSV files were provided by Netrality viz., ‘All Time Billing’, ‘Last Month Billing’, ‘Current Customers’ and ‘Prospect Customers’. The Prospect Customers list is where we need to acquire the potential Customer ID’s.

The main aim for us is to create a model which could predict and recognize all the possible clients from the Prospect Customers List with the help of Billing histories.

**Summary of Approach**

The main aim given to us by the Netrality Data Center was to predict Potential Customers who will be the ideal ones replacing the Current Customers. To tackle the Business Problem, we used various Applied Machine Learning Techniques such as the Exploratory Data Analysis (EDA), Feature Engineering, Several Models of Regressions, various Clustering methods and various uses of Data Visualization methods to identify the patterns. In the first, we identified and addressed the null values/missing values from the csv files given. We explored the Data Sets in order to understand which label/column can be our features such as ‘Revenue’, ‘Estimated Budgets’, ‘Number of Budgets’, ‘Funding Amounts’ etc. In the initial steps we targeted our Billing and features consisting different labels from Current Customer and also the billings. Later we understood that we have to have ‘Target’ as our Billing and Features must not be the Billing (All Time or Last Month) completely. When we were driving ahead with the actual requirement needed, we weren’t able to obtain satisfactory metric scores from the models which were created to predict the Potential Customers. But we used an effective way of predicting by using the Clustering Methods (K-Means, Birch, Spectral Clustering). Which helped us in obtaining the Potential Company IDs which could be replaced and perform better in the near future.

**Summary of Results and Conclusion**

* The gist of our results Is that we were able to obtain the Potential Company IDs, by using “Clustering Methods”, which seemed to be one of the most effective ways of Predictions to be established.
* With evaluations at and each and every clusterID helped us understand the bonds established between the Prospect and Current Customers file.
* Hence, we can confidently say that we were successful in obtaining the Potential CompanyID’s from the Propsect Customers Data File.

**Table of Prospective Customers to Target**

*Customer IDs that are identified from Potential Customers csv file using Clustering Method.*

|  |  |  |
| --- | --- | --- |
| 7703659 | 155353090 | 67421650 |
| 2441797 | 253034667 | 24576142 |
| 159662254 | 37962377 | 5128546 |
| 104333869 | 5619763 | 25185178 |
| 24231957 | 38683398 | 43897815 |

**Details of the Modeling and Process Approach**

The Methodology through which we have analyzed and understood the Data involved various factors such as “Exploratory Data Analysis”, “Feature Engineering”, “Data Visualization”, “Creating and Selecting Models” and finally “Predicting the Potential Prospect Customers”.

**(I) Exploratory Data Analysis (EDA): -**

* In order to use the data effectively, we went through several steps of analyzing the data by clearing out the null/0’s/NaN values from the Data Frames.
* We identified the data types within the Data Frames and converted a few of the categorical value columns into numerical value columns, in order to fit the data effectively. Few of them are: - ‘Primary Industry’, ‘Revenue Range’, ‘Employee Range’, ‘Ownership Type’, ‘Business Model’, ‘Company Country’, etc.
* We also created different sets of Data Frames between ‘Billing’ and ‘Current Customers’ by merging them based on the ‘Company IDs’.

**(II) Feature Engineering(\*): -**

* Our Analysis led us to create different new features for the created model or methods, for obtaining better results.
* We created new features within the (i) “Current Customers” & “Prospects Customers” and (ii) New features were created between “Billings(All\_Time & Last\_Month)” and “Current Customers”.

**(i). New features created within Current and Prospect Customers: -**

(1) Total Estimated Budgets of all Departments: - Where we sum up all the Estimated Budgets across each department.

(2) Average Revenue Per Place: - Where we used the columns Revenue and Number of locations and performed mathematical operation.

(3) Average Employees Per Place: - In this we used the columns Employees and Number of locations and performed mathematical operations.

(4) Total Number of Employees across all Location: - We used the Employees and Number of Locations Columns and performed mathematical operation.

(5) Fund Change: - We used the Total Funding and Recent Funding Columns and performed mathematical operations.

(6) Average Salary: - We used the Revenue and Employees Columns and performed mathematical operation.

(7) Employees Growth Rate: - We used the Past 1 and 2 years of Growth Rate Columns and performed mathematical operation.

(8) Financial Aspect: - We used the Revenue, Employees, Total Funding Amount and Total Est. Budgets of all the Departments Columns and performed mathematical operations.

(9) Total Financial Aspect: - We used the Revenue, Employees, Total Funding Amount and Total Est. Budgets of all the Departments Columns by summing up all of them.

(10) Total Amount: - We used the Total Funding Amount and Total Est. Budgets of all the Departments Columns and performed mathematical operations.

(11) Profit\_Current\_Customers: - We used the Revenue and Total Est. Budgets of all the Departments Columns and performed mathematical operation.

**(ii). New features created between Billings and Current Customers: -**

1. Total Income Generated by a company based on Last Month/All Time Billing: - We used the Revenue and Total Billing Columns and performed mathematical operation.
2. Efficiency based on Last Month/All Time Billing: - We used the Total Income Generated by a company and Total Est. Budgets of all the Departments Columns and performed mathematical operation.
3. Income generation per employee based on Last Month/All Time Billing: - We used the Revenue and Total Billing Columns and performed mathematical operation.
4. Total Revenue Acquired per employee basing on Last Month/All Time Billing: - We used the Efficiency Based on Billing and Employee Columns and performed mathematical operation.
5. Total Revenue Acquired per location basing on Last Month/All Time Billing: - We used the Employees and Total Billing Columns and performed mathematical operation.
6. Customer Contribution on overall Billing based on Last Month/All Time Billing: - We used the Revenue and Total Billing Columns and performed mathematical operation.
7. Revenue Ratio of the Current Customer (Last Month/All Time Billing): - We used the Revenue and Total Billing Columns and performed mathematical operation.
8. Revenue Ratio of the Current Customer per Employee (Last Month/All Time Billing): - We used the Revenue and Employee Columns and performed mathematical operation.

**(III) Data Visualization: -**

* To determine which Industry is drawing the highest amount of Revenue, we established plots between “Primary Industries” and “Revenue” for Current Customers and Prospect Customers Data File.
* To determine which state is topping the Revenue, we established plots between “Company State” and “Revenue” for Current Customers and Prospect Customers Data File.
* Below are the plots: -

**Current Customers: -**

A graph with different colored bars

Description automatically generated A graph of different colored lines

Description automatically generated

**Prospect Customers:-**

A graph of different colored bars

Description automatically generated A graph of different colored lines

Description automatically generated A graph of different colored lines

Description automatically generated

**(IV) Creating and Selecting the Models:-**

* After rigorous analysis and trials for selecting the features and targets, we selected Features and Targets from the default available ones and from the newly created ones (Feature Engineering\*).

**(i). Selected Features:-** “Revenue (in 000s USD)”, “Recent Funding Amount (in 000s USD)”, “Number of Locations”, “Est. IT Department Budget (in 000s USD)”, “Est. HR Department Budget (in 000s USD) and “Total Financial Aspect”.

**(ii).** **Selected Target:-** “Total Revenue Acquired per Employee basing on Last Month”.

* We then developed different models of Regression such as the “Random Forest Regressor”, “XGB Regressor”, “K Neighbors Regressor”, “Extra Tree Regressor", “Gradient Boosting Regressor”, “LGBM Regressor", “Cat Boost Regressor".
* Where we performed rigorous tunning measures for the models to give us the standard metric scores.
* We tried testing our model run for 100 times in order to check the performance of it.
* But unfortunately, no such model gave us the desired results. Out of these models we selected “Random Forest Regressor” as a Tentative Best Model, since its metrics was on the positive side after multiple runs when compared to the other ones.

**(🡪). Clustering Methods:-**

* In order to obtain/acquire the Potential Company IDs, we tried predicting it by using “Clustering Methods”, which is one of the most effective ways of Predictions to be established.
* In this “Clustering Method”, we used only the “Current Customers” and “Prospect Customers” csv files, in which we created a ‘Binary’ column for both the csv files.
* ‘Binary’ – 0’s for ‘Current Customers’(Blue Bubbles) and ‘Binary’ – 1’s for ‘Prospect Customers’(Orange Bubbles).
* Since, clustering method can take only the input data, we have given only the features for the clustering models – **‘K-Means’, ‘Birch’ and ‘Spectral Clustering’.**
* The selected features are:- 'Revenue (in 000s USD)', 'Est. Marketing Department Budget (in 000s USD)', 'Est. Finance Department Budget (in 000s USD)', 'Est. IT Department Budget (in 000s USD)', 'Est. HR Department Budget (in 000s USD)', 'Number of Locations', 'Recent Funding Amount (in 000s USD)’, ‘Total Amount’\*, ‘Total Financial Aspect’\*, ‘Fund Change’\* and ‘Average Revenue Per Place’\*. (where :- ‘\*’ = New Features created via Feature Engineering)
* Then we evaluated the set of clusters, by picking up each and every cluster ID in order to predict and identify the strong related bonds which would be clustered between the “Prospect and Current Customers”.
* The following table shows more about the Prediction of Company ID’s from the Prospect Customers Data File.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Clustering Method** | **Number of Clusters (n\_clusters)** | **Evaluation at which Cluster ID** | **Predicted Prospect CompanyIDs** | **Current Customers IDs** | **Features Selected** |
| Spectral Clustering | 20 | 11th | 7703659 | 368816466 | 1st Group |
| K-Means Clustering | 7 | 4th | 155353090 | 1475361, 2555990 | 1st Group |
| Birch Clustering\* | 5 | 2nd | 67421650, 2441797, 253034667, 24576142, 5128546, 104333869, 5619763, 25185178 | 3194954, 32394811, 27181208 | 2nd Group |
| Birch Clustering\* | 7 | 1st | 24231957, 38683398, 159662254, 43897815, 37962377 | 28117851 | 2nd Group |

**(1).** Where 1st Group ='Revenue (in 000s USD)', 'Est. Marketing Department Budget (in 000s USD)', 'Est. Finance Department Budget (in 000s USD)', 'Est. IT Department Budget (in 000s USD)', 'Est. HR Department Budget (in 000s USD)', 'Number of Locations', and 'Recent Funding Amount (in 000s USD).

(**2).** Where 2nd Group = 'Total Amount', 'Total Financial Aspect', 'Revenue (in 000s USD)', 'Est. Marketing Department Budget (in 000s USD)', 'Est. IT Department Budget (in 000s USD)', 'Fund Change', and 'Average Revenue Per Place'.

A grid of orange and blue dots

Description automatically generated

**(=>). Random Forest Regressor Performance:-**

* The performance of this model seemed to be on a better hand when compared to the other models created. The metrics of the Random Forest are as follows:- **“R2\_Score = 0.1322”, “MSE\_Score = 0.1857” and “MAE \_Score = 29589.47”.**
* Also, by using this model we predicted 10 Potential Company ID’s.

**Our findings via Regression models:**

**Table of Prospective Customers which can be Targeted (based on less percentage of metrics scores): -**

*Customer IDs that are identified from Potential Customers csv file using Random Forest Regressor.*

|  |  |
| --- | --- |
| 9650823 | 12963767 |
| 34423726 | 10076896 |
| 129761346 | 116015107 |
| 32902619 | 85395492 |
| 355675853 | 113934548 |

**Our metric scores for all the Regression Models which we have tried for: -**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **DataSet** | **Iterations** | **R2** | **MSE** | **MAE** | **Best** |
| *Random Forest Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *0.1322169415* | *0.1857524* | *29589.4741954* | *No* |
| *XGB Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *-0.48318* | *-0.483186* | *32172.060747242245* | *No* |
| *K Neighbors Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *0.07250091* | *0.147333* | *29688.791373* | *No* |
| *Extra Tree Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *-0.8778274* | *-0.877827* | *28525.3982022* | *No* |
| *Gradient Boosting Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *-11.950183* | *-11.95018* | *29040.4114209* | *No* |
| *LGBMRegressor* | *Combined Last Month, All Time and Current Customers* | *100* | *-0.2634464* | *-0.263446* | *31547.895756* | *No* |
| *CatBoost Regressor* | *Combined Last Month, All Time and Current Customers* | *100* | *0.15099623* | *0.19099623* | *28148.81804* | *No* |
| *Linear Regression* | *Current Customers and All Time Billing  (where target and feature had billings (which is not an ideal method to consider))* | *-* | *0.494677127* | *0.00040* | *0.00814* | *No* |
| *Ridge Regression* | *Current Customers and All Time Billing  (where target and feature had billings (which is not an ideal method to consider))* | *-* | *0.5302605* | *0.00037* | *0.0078* | *No* |
| *Lasso Regression* | *Current Customers and All Time Billing  (where target and feature had billings (which is not an ideal method to consider))* | *-* | *0.49468* | *0.00040* | *0.00814* | *No* |
| *Elastic Net Regression* | *Current Customers and All Time Billing  (where target and feature had billings (which is not an ideal method to consider))* | *-* | *0.4946* | *0.00040* | *0.00814* | *No* |

**Conclusion: -**

* To conclude, we can say that the approach and the solution for the problem given to us by the Netrality Data Center was solved by using “Clustering Methods”.
* By evaluating each and every clusterID for the clusters created, we were able to depict strong bonds between the Prospect and Current Customers Data File.
* On the whole, we can confidently say that we were successful in **“Predicting the Potential Prospect Customers”** at the end.