**Project Name: Customer Prediction**
**Team Name: Team TVS**
**Authors:  Tirth, Vishal and Sakshi**
**Date: 12/18/2023**

**Problem Statement**

The main challenge presented by Netrality was to identify potential customers from a prospect customer list, leveraging data-driven insights and machine learning techniques. Netrality provided three key CSV files: one containing the current customer list, another with the current billing data of current customers over various locations, and the third with the prospect customer list.
The primary objective was to use the provided data sets to develop a predictive model that could identify potential customers from the prospect list. This involved exploring and cleaning the data, understanding the relationships between different features, and applying machine learning algorithms to make predictions**.**

**Summary of Approach**

The approach to solving the business problem of identifying potential customers for Netrality involved a comprehensive data-driven approach, which included data exploration, feature engineering, and the application of various machine learning algorithms. Initially, we addressed missing values and inconsistencies in the provided CSV files and created categorical and numerical data frames for analysis. The team explored various directions for analysis, such as geographical locations, employee growth rates, revenue, and departmental budgets. We utilized correlation matrices to identify highly correlated features and guide the selection of relevant variables. Feature engineering was experimented with to create new variables that might capture essential information. While we had initially planned to use the billing data of current customers as the target variable to predict billing of prospective customers, we then decided against it as the billing data was not normalized, and there was no linear relation between the features and the billing. Therefore, we then decided to use revenue as our target variable, as it was highly correlated with most of the features.

**Summary of Results and Conclusion**

The final results of our approach to solving the business problem of identifying customers for Netrality indicate that the Random Forest Regressor performed the best among the various models tested. We chose this model based on its ability to predict revenue, which serves as a proxy for identifying valuable customers.

**Key points and decisions in our approach include:**

1. **Feature Selection:**
    - We selected features that demonstrated a high correlation with revenue. Specifically, we chose “Total Funding Amount”, “Number of Locations”, “and “Employees”.
    - Budgets were highly correlated as well, but we opted for a diverse set of features and excluded them in favor of capturing different aspects of customer potential.

2. **Model Selection:**

- The Random Forest Regressor was chosen as the final model due to its robust performance in predicting revenue. This model takes into account the interaction of multiple decision trees, providing a more accurate prediction.

3. **Model Performance:**
   - The Random Forest Regressor exhibited exceptional performance metrics on the test data:
     – RMSE (Root Mean Squared Error): 0.081
     – R2 (R-squared): 0.897
     – MAE (Mean Absolute Error): 0.050
   - These metrics reflect the model's accuracy, explaining a high percentage of the variance, and providing small errors in prediction.

4. **Common Companies Predicted:**
   - We predicted 300 potential customers using the Random Forest Regressor and an additional 300 potential customers through the CatBoost Regressor. Remarkably, there are 260 companies that overlap in both predictions.

In conclusion, the Random Forest Regressor, trained on carefully selected features, stands out as a reliable tool for identifying potential customers. The overlap of predicted companies from different models adds another layer of validation, reinforcing the potential of the identified companies as valuable prospects for Netrality.

**Table of Prospective Customers to Target**
*Company IDs of Potential Customers identified*

| | | | | |
|---|---|---|---|---|
| 155353090 | 369938550 | 459073254 | 5851944 | 60720958 |
| 459963342 | 68863214 | 130203765 | 3444162 | 141738322 |
| 3834943 | 13830837 | 2508566 | 48187827 | 41320983 |
| 74203899 | 39469842 | 93323946 | 16220332 | 58804259 |
| 6275089 | 22315545 | 43897815 | 412002344 | 10256729 |
| 55727016 | 103841907 | 60310227 | 8590337 | 31342638 |
| 20775334 | 24182874 | 18633856 | 18579882 | 2245058 |
| 345275896 | 430983 | 16859276 | 65536388 | 63243962 |
| 9438760 | 26968154 | 30048670 | 462169976 | 21403020 |
| 297664519 | 39600454 | 12227700 | 688376 | 67421650 |
| 13207636 | 104333869 | 100221071 | 106138232 | 266727 |

| | | | | |
|---|---|---|---|---|
| 17402544 | 5358630 | 10814149 | 2901487 | 345283492 |
| 3523141 | 24461754 | 8110814 | 38126010 | 51711289 |
| 24576142 | 88376327 | 23776193 | 15712435 | 27820656 |
| 239305146 | 4506176 | 54823666 | 1720519 | 62014529 |
| 27722128 | 66505148 | 39584392 | 116424706 | 37934049 |
| 9751686 | 128123007 | 15877691 | 84099764 | 14155984 |
| 30245334 | 94568709 | 2441797 | 24904409 | 13043336 |
| 128860355 | 34687140 | 21692686 | 19513364 | 30347697 |
| 7258303 | 14946173 | 196821345 | 122634324 | 52447678 |
| 164856312 | 7897402 | 70336972 | 7834830 | 4128773 |
| 24431896 | 355780297 | 91406858 | 157713259 | 72238328 |
| 56049732 | 3539473 | 36487399 | 254869582 | 3573275 |
| 1507503 | 39977791 | 1114181 | 106676542 | 138102457 |
| 94519784 | 50040045 | 11127417 | 56199764 | 30196506 |
| 33018022 | 95650875 | 168509596 | 29818882 | 49302654 |
| 47942451 | 524857 | 118145647 | 9604546 | 90883103 |
| 34537887 | 9012358 | 59557563 | 47730929 | 51386296 |
| 169231161 | 13140007 | 75679349 | 45662682 | 76416262 |
| 356818961 | 9634147 | 34390962 | 32899972 | 128565250 |
| 15125590 | 10190854 | 32199830 | 28268742 | 345620672 |
| 15794314 | 45992365 | 22856817 | 234914216 | 56953751 |
| 31693712 | 41058369 | 345458296 | 17662709 | 19812244 |
| 15896733 | 7783252 | 225738822 | 12337715 | 371833565 |
| 38650584 | 36739880 | 29868189 | 358896585 | 114263816 |
| 358707387 | 344486428 | 71581827 | 24231957 | 374265599 |

| | | | | |
|---|---|---|---|---|
| 348727434 | 40603148 | 13699410 | 58213077 | 54742651 |
| 81789264 | 12913103 | 344472790 | 66421453 | 141951307 |
| 36848621 | 94652898 | 347071113 | 21545051 | 296960124 |
| 8850023 | 51315878 | 12272288 | 34570613 | 4280349 |
| 62342734 | 136036427 | 47582115 | 41058643 | 144933765 |
| 31802379 | 32489569 | 104104167 | 358636815 | 148046227 |
| 11260615 | 129729226 | 14516709 | 19071074 | 14713364 |
| 20827376 | 41323685 | 56526980 | 348500421 | 559692221 |
| 35881054 | 566128558 | 154239344 | 350915803 | 343397512 |
| 37847515 | 16466078 | 24854701 | 20070834 | 369350661 |
| 50495653 | 2540471 | 19890428 | 1804856 | 40403053 |
| 9867169 | 48547873 | 23675590 | 38027519 | 23545246 |
| 24544229 | 99107151 | 188678563 | 9487723 | 44501455 |
| 63031304 | 136118787 | 297468076 | 44534402 | 54821316 |
| 356414046 | 15191640 | 17815664 | 54412460 | 51200156 |
| 80970479 | 1461214 | 5619763 | 4901834 | 136872493 |

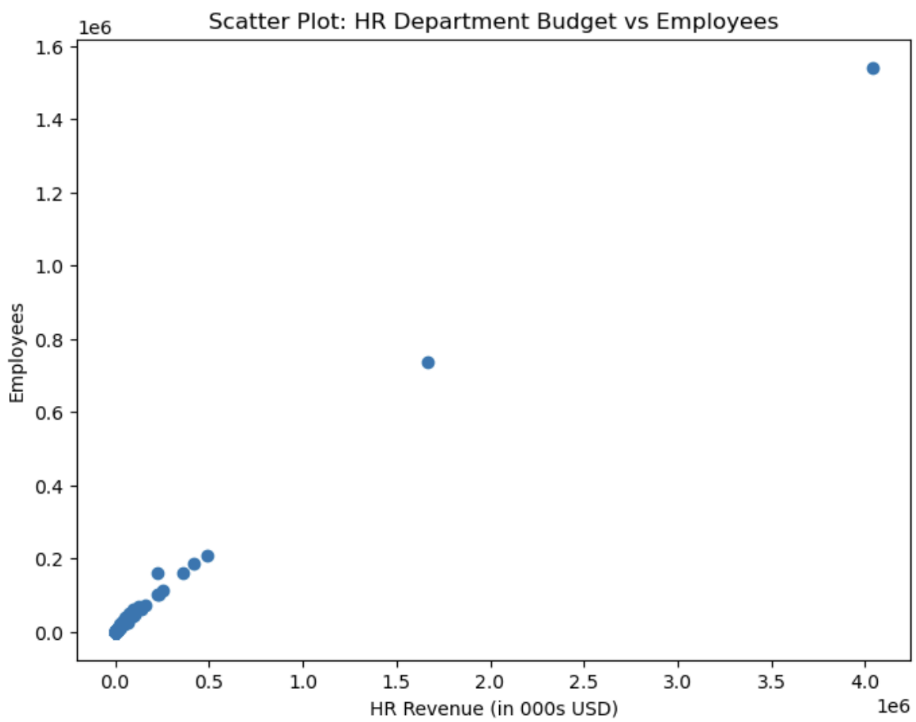
**Details of the Modeling and Process Approach**

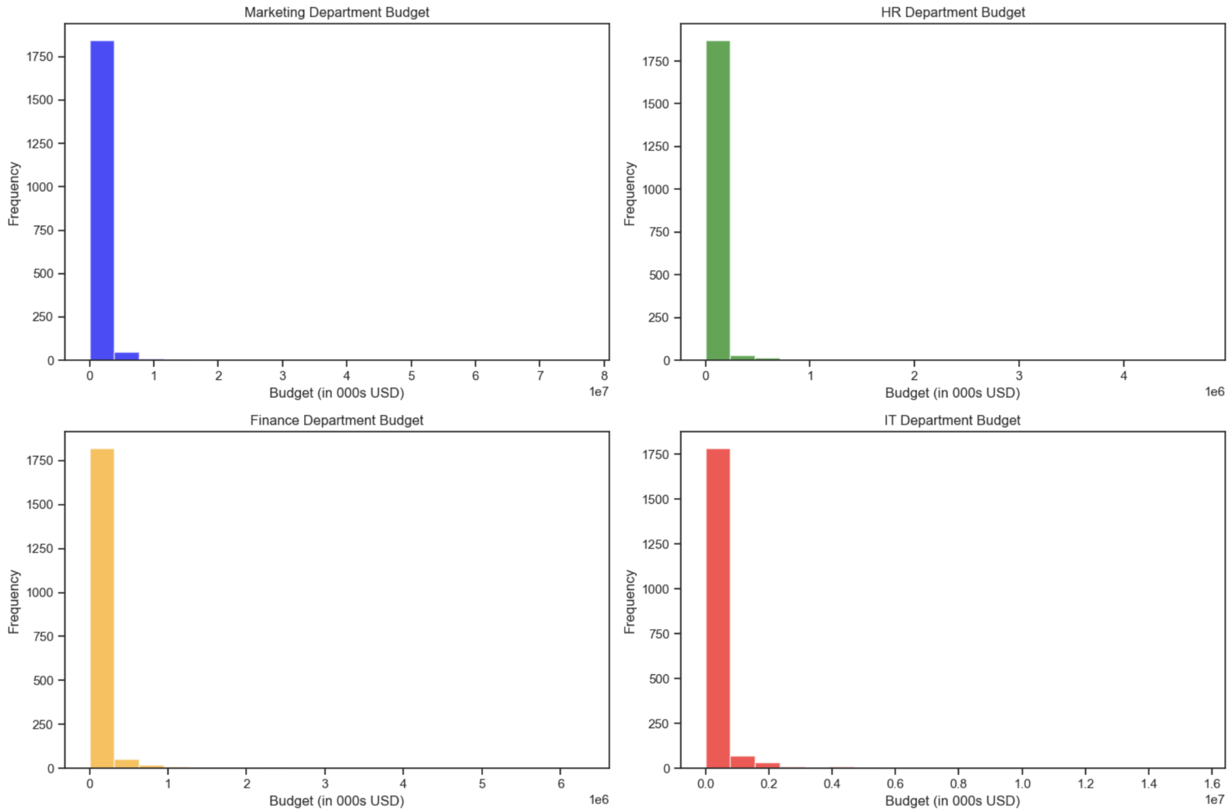
**Exploratory Data Analysis (EDA):**

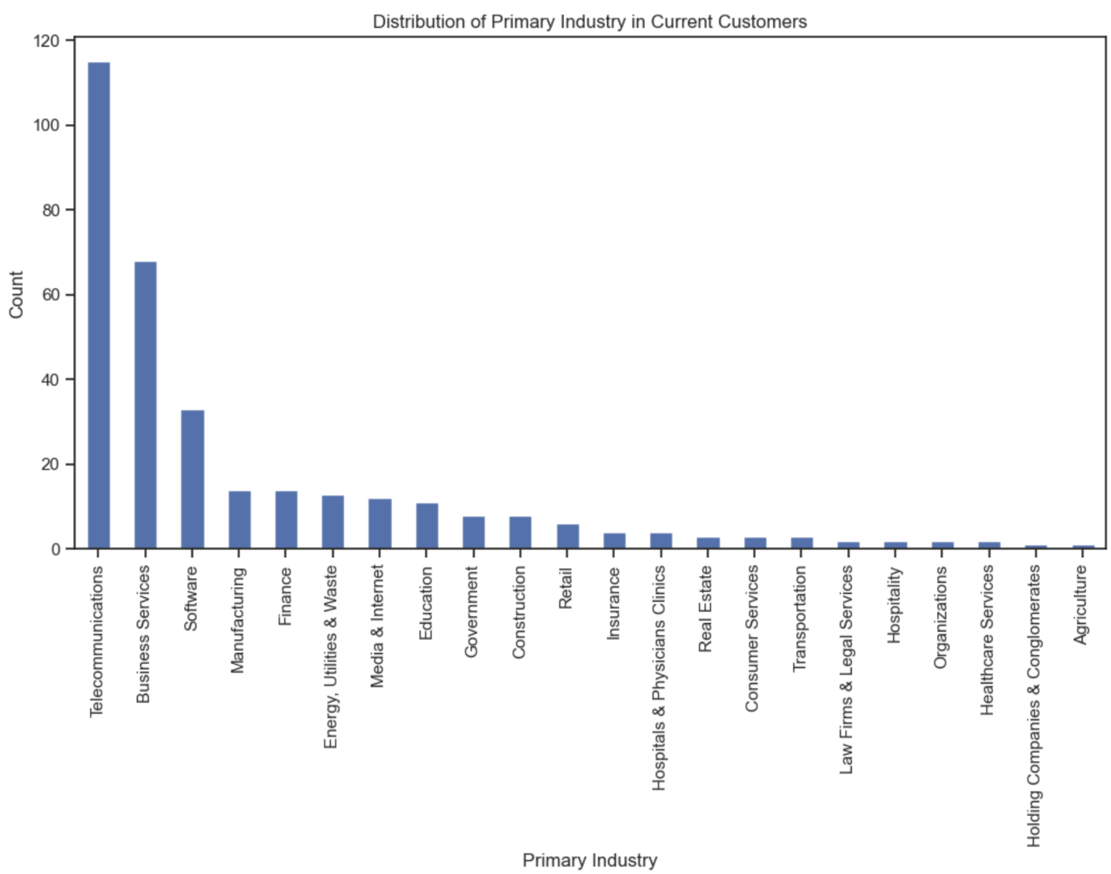
1. **Data Loading and Segregating:**
    - We started by loading the CSV files and cleaning it by removing nulls. We then segregated the data into categorical and numeric dataframes. The billing csv file was already segregated into Last Month Total and Lifetime Total so we created two separate dataframes for that by summing up the billing of various locations.

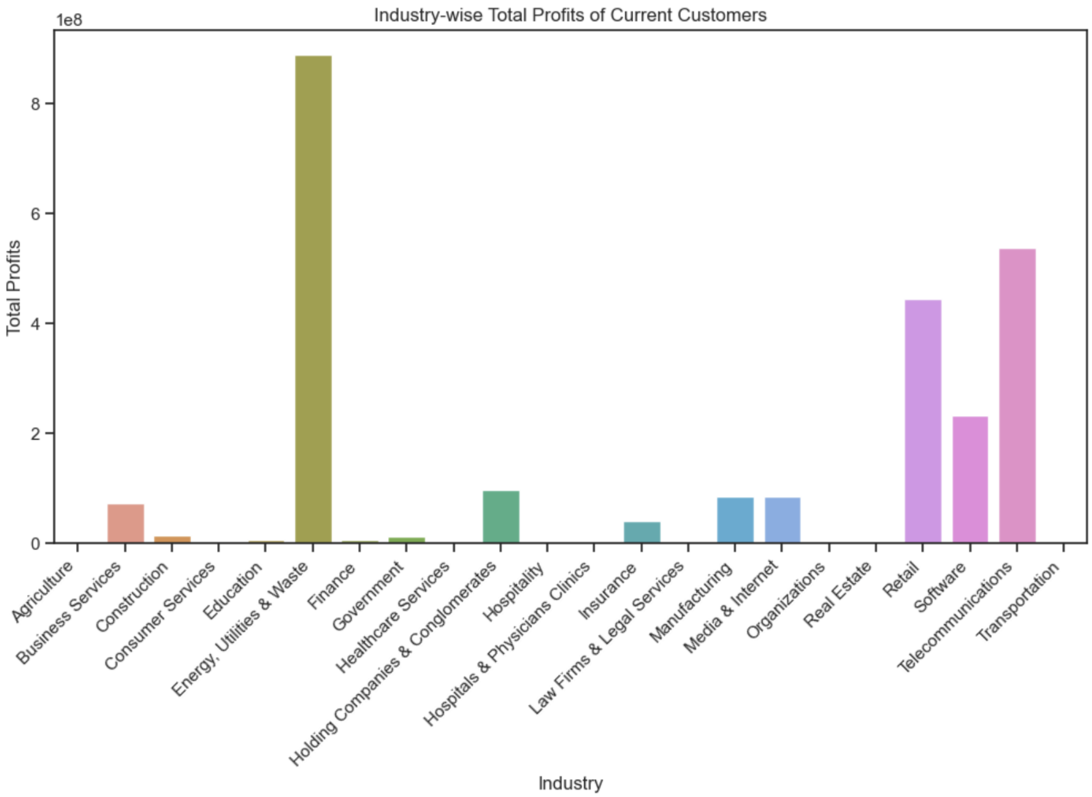
2. **Data Visualization:**
    - Explored the relationship between the HR department budget and the number of employees by creating a scatter plot. . We also examined the scatter plot for employee growth rate and revenue to understand potential correlations.
    - Created Histograms for the budgets of various departments (Marketing, HR,Finance, IT) to visualize their distributions.

- Constructed a correlation matrix to quantify and visualize the relationships between features in the dataset. This helped in identifying which features are correlated with each other
- Used a bar plot to display the distribution of companies across primary industries. This provided insights into the dominant industries within the dataset.

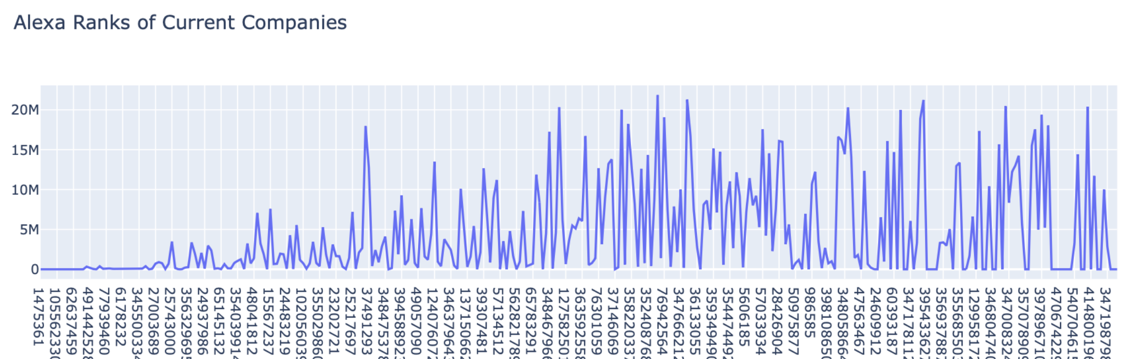



Marketing Department Budget
Frequency
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6 7 8
Budget (in 000s USD)
1e7
HR Department Budget
Frequency
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4
Budget (in 000s USD)
1e6
Finance Department Budget
Frequency
0
250
500
750
1000
1250
1500
1750
0 1 2 3 4 5 6
Budget (in 000s USD)
1e6
IT Department Budget
Frequency
0
250
500
750
1000
1250
1500
1750
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
Budget (in 000s USD)
1e7


Distribution of Primary Industry in Current Customers
Count
0
20
40
60
80
100
120
Telecommunications
Business Services
Software
Manufacturing
Finance
Energy, Utilities & Waste
Media & Internet
Education
Government
Construction
Retail
Insurance
Hospitals & Physicians Clinics
Real Estate
Consumer Services
Transportation
Law Firms & Legal Services
Hospitality
Organizations
Healthcare Services
Holding Companies & Conglomerates
Agriculture
Primary Industry



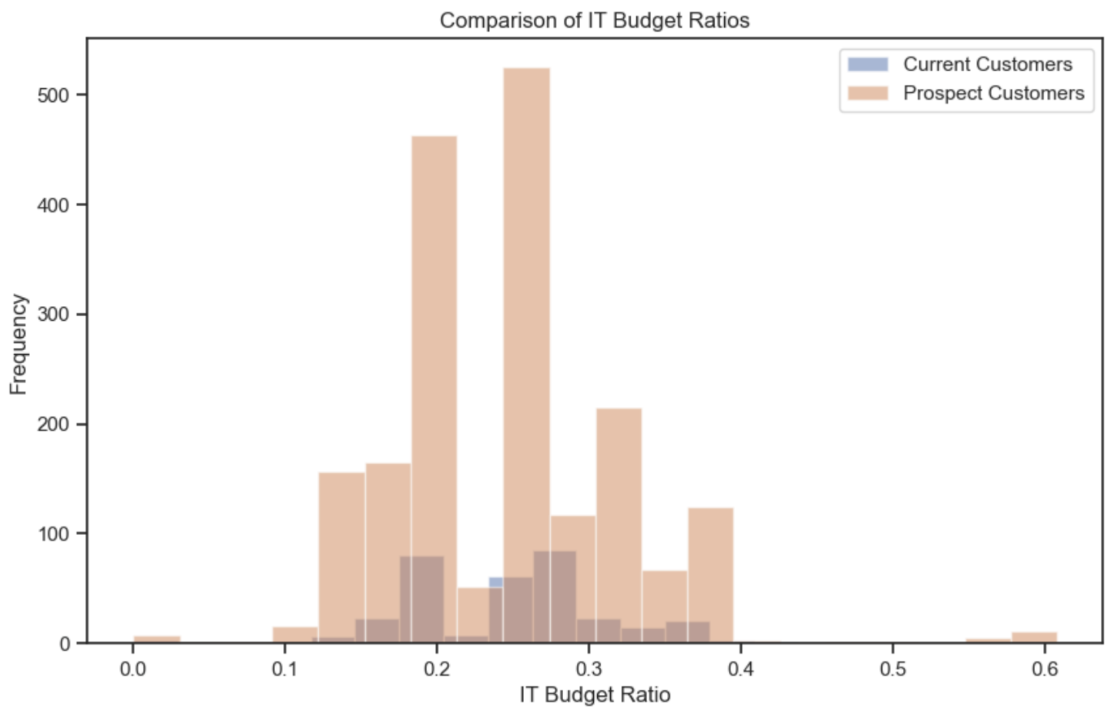

```
Highest Alexa Rank: 21841578.0 (Company ID: 350944369)
Lowest Alexa Rank: 0.0 (Company ID: 116244586)
```




**Feature Engineering:**

1. **Creation of New features:**
    - We summed up all the department budgets and created a new feature called as "Total Budget"
    - We calculated the Age of the Company.
    - We also calculated the Ratio of all the departments and created histograms to compare the ratio of current and prospective customers.

- Extracted meaningful information from categorical features, and converted them into numerical representatives.




2. **Transformations:**
    - Applied log transformation on skewed numerical features to improve normality.
    - Scaled numerical features using techniques like Min-Max scaling to ensure consistent units.

**Model Selection and Training:**

1. **Target Variable:**
    - We initially selected 'Billing' as our target variable, but upon being unable to establish a significant relationship between 'Billing' and our features, we subsequently opted for 'Revenue' as the target variable for regression modeling.

2. **Feature Selection:**
    - Selected relevant features based on EDA, correlation analysis, and feature importances.
    - Utilized techniques like feature importances from Tree-based models.

3. **Model Training:**
    - Split the dataset into training and testing sets to evaluate model performance.
    - Trained multiple regression models, including Linear Regression, Lasso Regression, Decision Tree, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost, CatBoost, LightGBM, Support Vector Regression and Neural Networks.

4. **Model Evaluation:**

- Employed metrics such as Root Mean Squared Error (RMSE), R-Squared (R2), and Mean Absolute Error (MAE) to evaluate model performance.

**Prospect Customer Prediction**

1. **Prospect Data Preparation, Model Prediction and Threshold Selection:**
   - Processed the prospect customer data, ensuring it matches the format used for training.
   - Utilized the trained Random Forest Regressor to predict potential revenue for prospect customers.
   - Determined a threshold to classify potential customers based on predicted revenue.

**Common Companies Analysis**

1. **Comparison with Other Models and Evaluation:**
   - Employed CatBoost Regression to predict potential revenue for prospective customers.
   - Identified common companies between the predictions of Random Forest and CatBoost.
   - Evaluated the significance of common companies in terms of potential revenue.

**Conclusion**

1. **Best Model and Feature Importance**
   - Concluded that Random Forest Regressor is the best model based on the evaluation metrics. Highlighted the importance of features such as Total Funding Amount, Number of Locations, and Employees in predicting revenue.

| **Name** | **DataSet** | **Iterations** | **R2** | **MSE** | **MAE** | **Best** |
|---|---|---|---|---|---|---|
| *Linear Regression* | *Current Customer Dataset* | *1000* | *.248* | *2.640* | *1.9487* | *No* |
| *Lasso Regression* | *Current Customer Dataset* | *1000* | *.246* | *.2.644* | *1.956* | *No* |
| *Decision Tree Regressor* | *Current Customer Dataset* | *1000* | *.837* | *1.226* | *0.764* | *No* |
| *Random Forest Regressor* | *Current Customer Dataset* | *1000* | *0.898* | *0.96* | *0.591* | *Yes* |
| *Extra Trees Regressor* | *Current Customer Dataset* | *1000* | *0.903* | *0.946* | *0.647* | *No* |
| *AdaBoostRegressor* | *Current Customer Dataset* | *1000* | *0.898* | *0.969* | *0.648* | *No* |
| *Gradient Booster* | *Current Customer Dataset* | *1000* | *0.867* | *1.110* | *0.717* | *No* |
| *XGB Regressor* | *Current Customer Dataset* | *1000* | *0.879* | *1.055* | *0.707* | *No* |
| *CatBoost Regressor* | *Current Customer Dataset* | *1000* | *0.895* | *0.982* | *0.642* | *No* |
| *LGBM Regressor* | *Current Customer Dataset* | *1000* | *0.869* | *1.099* | *0.77* | *No* |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Support Vector Regressor* | *Current Customer Dataset* | *1000* | *0.108* | *2.877* | *2.164* | *No* |
| *Neural Networks* | *Current Customer Dataset* | *1000* | *0.320* | *2.511* | *NA* | *No* |