Problem Statement

Indeed routes leads to the sales team based on the potential quality of each lead. The system predicts the potential value of each lead and assigns out the leads from highest to lowest probability until each sales representative has a full sales book.

The dataset shows leads that were assigned and never assigned to a sales rep. The purpose of this exercise is to estimate the incremental impact the sales representatives had on revenue. In other words, how much more did these leads spend because there was sales intervention?

Exploratory Data Analysis

Column Definitions

Unnamed: 0

- dtype = int64
- Assumption = After analyzing the column 'Unnamed: 0', it seems that it is exactly same as index of the rows so it wouldn't be of much use to perform further analysis. I feel that we can drop this column from the dataset, as any correlation of 'unnamed' column (if exists) with other columns would be a spurious correlation. Therefore, dropping this column for reducing noise in the data.

advertiser_id

- dtype = int64
- Mostly, IDs are always unique but still let's check for any existing duplicate advertiser IDs
- After checking whether there is more than 1 advertiser id in the column 'advertiser_id', it was found that each value is unqiue.
- Since, 77891 is same as the total number of rows in our dataset, we can infer that all the IDs are unique in the 'advertiser_id' column This would be good to uniquely identify an advertisement but this won't of much use for further analysis. Therefore, dropping this column for reducing noise in the data.

assigned

- dtype = int64
- Assumption = Treating this as a categorical variable as this shows whether leads were assigned or not assigned to the sales rep.

date_assignment_starts

- dtype = object
- Assumption = This is the date and time when the advertising assignment starts

date_assignment_ends

- dtype = object
- Assumption = This is the date and time when the advertising assignment ends

first_revenue_date

- dtype = object

- Assumption = This is the date when first revenue was collected.
    - Would take its difference in days with the date_created to use it as an input variable after analyzing its correlation.

date_created

- dtype = object
- Assumption = This is the date when advertisement got created

age

- dtype = int64
- Assumption = Age of the advertisement. This also contains negative values, which seems ambiguous. Negative value of age might signify the age of a lead before advertisement was. I'd treat this as a variable which might contribute towards predicting the revenue and ignore the ambiguity of age being negative.

assign_days

- dtype = int64
- Assumption = Number of days assigned to a lead. This also contains negative values, which seems ambiguous. Negative value of days might signfy the days of a lead before the advertisement. I'd treat this as a variable which might be contributing towards predicting the revenue and ignore the ambiguity of number of assigned days being negative.

revenue

- dtype = float64
- Assumption = Revenue generated by a lead

The further analysis and visualizations can be viewed in this code file.

Ambiguities

- It has been observed that revenue column has 74551 missing values out of 77891 total rows, which means that approximately 95.7% values are missing in the revenue column.

Possible approaches for missing data in a column:

1. We can drop all the rows with missing values.
    a. Advantage:
        i. Complete removal of data with missing values results in robust and highly accurate model
        ii. Deleting a particular row or a column with no specific information is better, since it does not have a high weightage

    b. Disadvantage:
        i. Loss of information and data Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset and in our case, it is more than 95%

2. Replacing missing values with Mean/Median/Mode:
    a. Advantage
        i. This is a better approach when the data size is small but in our case the data is not very small It can prevent data loss which results in removal of the rows and columns.

    b. Disadvantage:
        i. Imputing the approximations and we add variance and bias to the data works poorly compared to other multiple-imputations method.
3. Predicting The Missing Values:

We can predict the nulls with the help of a machine learning regression algorithm. But since applying regression isn't producing good and accurate results, so I am not using this method.

    a. Advantage:
        i. Imputing the missing variable is an improvement as long as the bias from the same is smaller than the omitted variable bias Yields unbiased estimates of the model parameters
    b. Disadvantage:
        i. Bias also arises when an incomplete conditioning set is used for a categorical variable Considered only as a proxy for the true values
4. Using Algorithms like KNN
    a. Advantage:
        i. Does not require creation of a predictive model for each attribute with missing data in the dataset Correlation of the data is neglected
    b. Disadvantage:
        i. Is a very time-consuming process and it can be critical in data mining where large databases are being extracted Choice of distance functions can be Euclidean, Manhattan etc. which do not even yield a robust result, so not using this approach as well

- It has been observed that assign days column has values for leads which were not assigned. This seems a bit confusing as data variable definitions are not clear.
- Age has some negative values under unassigned categories and I have masked those values with 0 as negative age does not makes sense until a reason is specified.

Questions

Question 1

How many leads are represented in this dataset? Describe both the assigned and unassigned populations. What is the average revenue of each group?

Question 2

What are the most important metrics to consider when answering the problem statement? Why?

Question 3

Analyze any existing relationship between account age and revenue.

Question 4

What is the incremental value of assigning a lead to the sales team?

Question 5 (Bonus Question)

Investigate the data however you wish and discuss any interesting insights you can find in the data. Don't feel pressured to spend hours on this.

Answers to these are provided in this code file.

Conclusion

The data had a lot of missing values and ambiguities as written above. Descriptive analysis has been done and there was very less correlation of the feature variables with the target variable revenue. Data isn't much reliable to proceed with the predictive and prescriptive analysis. Applying any prescriptive analysis would lead to underfitting. Tried applying various regression techniques but the data isn't producing promising predictions. One of the regression techniques can be viewed in this code file.