Dear XYZ,

I am Supriya, a data analyst with the Fetch Rewards team and would be working on the receipts, brands, and user datasets that we have in our database.

Since, I was analyzing the data in depth, I had some questions regarding the data sources, methodologies, and data collection methods with respect to the data. I also identified data quality issues with the current data we have and have suggested some ways to fix those issues, and scale and process the data as follows:

**Questions about data**

* What are the general business rules that need to be followed for any analysis?
* How has the data been collected? What are the different channels and sources?
* Is brand code the same for all categories of products in that brand? Is brand code different from product barcode?
* How does one define an ‘accepted’ receipt?
* Is it possible to document all information from a given receipt in a consistent format?

**Data quality issues discovery**

I followed the standard protocol to discover and assess data quality issues. Since data was available in JSON files I first converted them into a format compatible with the library I was using (pandas) so it was readable and accessible.

I explored the data through various inbuilt functions and methods in pandas. I assessed if the datatype of each column is relevant or not. Then I checked if there are any missing values in the data or even blank spaces. Next it was important to know the size of the data frame. This was done using ‘shape’ method.

The data entries were then assessed if multiple values occur in one row, or a single value occurs per row per column. For ease of analysis, linking databases and get more insight about the data, I tried to see if only one entry is there per row by manipulating the data.

**Potential resolution of data quality issues**

The data quality issues can be resolved by knowing the different data sources and channels the data is coming from and channels.

General business rules for example what receipt is considered acceptable etc need to be established before carrying out any analysis.

-Need to know the readable format for data processing

- Know the meaning of each data type and what it means for each metric in the data. So, in depth domain knowledge of the data and industry is essential.

- Find functions to convert to appropriate data types (formats) and fill in missing values using either imputation or dropping them based on the context of the data.

- Knowing functions to know how big the data is and if we have the data processing infrastructure in place to analyze this data.

- Convert non-atomic rows to atomic rows, to extract useful features and help in linking two databases

**Optimizing Data Assets**

To optimize data assets, it is imperative to know the business rules as well as the business problem or the objective of the analysis.

This would help in accurate feature selection, optimizing data cleaning, sub setting data accordingly and optimizing data linking.

**Performance and Scaling issues anticipated**

The given data is noisy with lots of irrelevant information. This can significantly impact machine learning models leading to poor prediction results as well as classification results.

It can also lead to overfitting of training data causing the model to perform poorly on the test/unknown data.

Scaling: Many firms' databases are being overwhelmed by the deluge of data they are facing due to exponential increases in the volume of data being produced and processed. A strategy known as "data scaling" has become critical for many organizations coping with ballooning datasets to manage, store, and process this overflow of data.

**Scaling solutions**

Scaling up: This sort of vertical scaling entails replacing your server with one that is faster and has more powerful resources (processors and memory). Because dedicated servers cannot be readily scaled, scaling up is usually a feature found in the cloud (since the move requires going to the data center to change the server manually and considerable downtime).

Scaling out: This horizontal scaling method entails the use of more servers for parallel processing. This is thought to be the ideal option for a real-time analytics project because you can build a proper infrastructure for your use case from the start and scale up as needed. In the long run, horizontal scalability tends to result in lower costs.

I hope you can answer my questions above and I hope my answers were useful.

Looking forward to working on this project.

Regards,

Supriya