Subject: Data Quality Issues and Optimization Process

Dear Jack Sparrow,

I hope you are doing well.

I want to share some thoughts and insights regarding the data analysis done on users, receipts, and brand datasets. We must take quick action regarding the data quality issues I found.

1. **Missing values:** Based on my analysis, many values exist in the datasets. The missing values in the users dataset are less, which can be quickly addressed. But the missing values in the receipts and brands datasets are higher comparatively.

    - **Missing values in the users dataset:** There are 62 missing values in the lastLogin feature, followed by 56 and 48 missing values in the state and signUpSource features, respectively.

    - **Missing values in the brands dataset:** Various features in the data contain more than 50% missing values. The most missing values are observed in the categoryCode and topBrand features, with 56% and 52%, respectively. Also, there are a few missing values in category and topBrand features, with 13% and 20%, respectively.

    - **Missing values in the receipts dataset:** The receipts data has the most features containing missing values compared to the other two datasets. The highest number of missing values are observed in the pointsAwardedDate feature with 52%. It is followed by bonusPointsEarned and bonusPointsEarnedReason with 51% each. There are missing values in other features too.

2. **Duplicated Rows:** A considerable amount of data is duplicated in the users dataset. There are 283 duplicated rows, meaning there are only 212 original values.

3. **Outliers:** I observed some outliers in the purchasedItemCount feature of the receipts dataset.

- I found these issues by doing in-depth exploratory data analysis. To resolve these issues, I would require some things to be addressed:

    1. Missing Values: Some features have more than 30% missing values. Any conclusions drawn from such data can be wildly inaccurate about the actual trend in the market. Please let me know the source of this missing data so I can gather more information to address this issue.

2. Duplicated Rows: Duplicated rows imply that the user records are repeated multiple times. This can cause skewness in the analysis. I need your guidance on how to go forward on this issue. Any information on the cause of the duplication can also be beneficial.

3. Outliers: I observed some outliers in the receipts data. I decided to consult with you regarding the outlier. I would appreciate your insights into whether these outliers are valid or not. This would help me in choosing the right approach to treating the outliers.

- To optimize the data assets we are trying to create, I would like to get any new or additional datasets that would enable me to reach more appropriate conclusions. It would give better trends and insights about the market. Also, it would be beneficial if I knew the target customers and business cases so that I work towards optimizing the assets accordingly.

- Also, I have some concerns regarding our performance and scaling issues that we might face during production. As the size of the dataset increases, it would take much work to handle the size of the data. It would be much more complicated to scale the data seamlessly. These issues can be addressed by using data storage efficiently. Also, using the cloud-based storage available in the market would be great. This would ensure that we can accommodate the performance and scaling demands. Regular optimization tests and comprehensive monitoring can help keep these issues in check.

I believe addressing these issues will allow us to move forward correctly. Please let me know if you need more information regarding these issues. I look forward to discussing this in detail.

Regards,
Hemanth Reddy