**CS 519 Applied Machine Learning I Project**

**Team Members: Md Ishtiaq Ahmed, Israel Gonzalez**

Task List:

1. brief description of the dataset
2. include a data dictionary
3. summary dataset (no of rows, columns, how many null values, unique values,)
4. discard the null columns, single value columns
5. organize columns in different category
6. doing pca
7. look out for outliers using plots
8. delete the outliers
9. select the features
10. create a column for prediction/churn

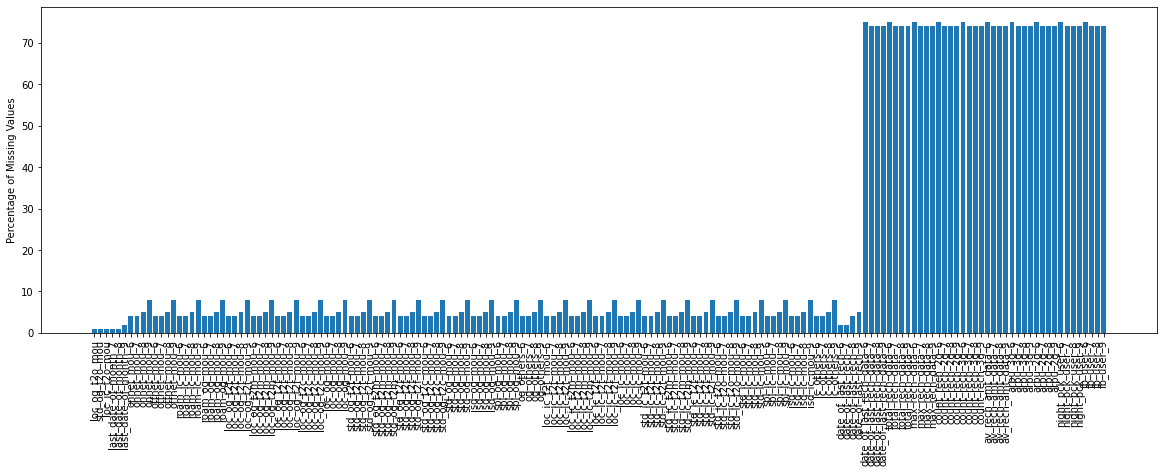
The churn of a telecom customer can depend on many things, and it is challenging to predict customer churn. That is why we were looking for a telecom customer dataset with a large number of instances along with a good amount of features. Finally, we found a dataset from <https://data.world/kishoresjv/telecomchurn/workspace/file?filename=telecom_churn_data.csv>

**This dataset contains 99,999 instances and 225 features.** Below is a brief description of the dataset:

* Each row of this dataset represents one unique customer
* All the features are telecom customer attributes related to what services they are using, spending on different services, talk time, data usage, recharge amount, data of last usage, date of recharge, types of data pack and many others
* Every attribute data is for four month
* One additional column is added to indicate churn. Customer who did not generate any revenue in the month of September fall under churned customer
* A detail data dictionary is provided in a separate excel sheet.

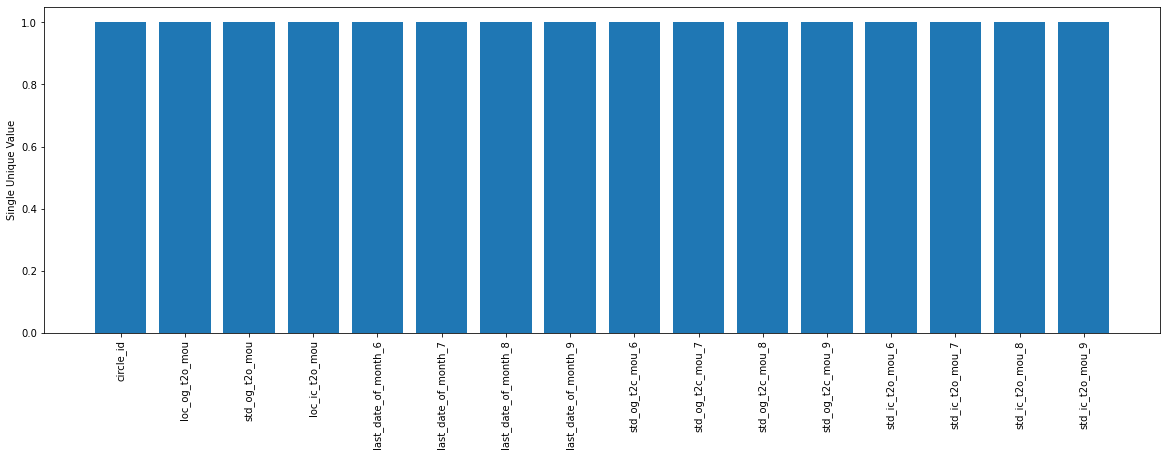
**Null value Analysis:**

From the below figure, we found that there are 40 features where more than 70% value is null. These columns are deleted since they will not add significant value.



**Unique Value Analysis:**

There are a total of 16 features where only a single unique value is present. These features are deleted from the dataset.



**Filling up the NaN Values:**

There are many standard way to fill up the null values. We have used mean value of each numeric column to fill up null values. After performing the action below is the current status of percentage of null value > 0:

A picture containing logo

Description automatically generated

So, there are still some null values present in the date column. Since date values cannot be replaced with mean, we will replace them with zero.