### **Part 4 - Document the problems**

| **data issue discovered** | **Why is this a problem?** | **How did you fix it?** | **Why is this fix appropriate?** |
| --- | --- | --- | --- |
| The data was a combination of different unnecessary data types which were not required for training models. | Models can only be trained via numerical data. | By only selecting columns having numerical data and neglecting else. | This will allow us to train our model. |
| There we a lot of columns present in the data set provided, which was difficult for us to observe properly all at a time. | Unable to observe the dataset properly might lead to inconvenience. | By performing transpose of the dataframe and making the columns vertical. | Without this fix, we might face inconvenience while analyzing the dataset. |
| Due to the large size of the dataset, it was physically tiring to point out NULL values from data sets. | The models cannot handle missing values. | By applying isnull() function on the dataframe to check if there are any null values in each column. | This will result in appropriate training of the model. |
| Distribution of data. | Values which are very different from others in a column (outliers) may lead to several problems such as inaccurate model performance, etc. | By performing EDA (exploratory data analysis) on the dataset and selecting data which lies within a desired range. | This will lead to better model performance. |