DATA LIFECYCLE MANAGEMENT

ASSIGNMENT 2 REPORT
DATA QUALITY ISSUES

AKANKSHA SHRIMAL - MT20055

ANJALI - MT20082

Assumption : The code file is in .ipynb format, and data is loaded at runtime from google drive.

Introduction

Given dataset - UCI Poker hand dataset. It has 11 attributes and 1025010 instances. Thus it is an 11 dimensional dataset with 25010 samples for training and 1000000 instances for testing. Each record/instance is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes (suit and rank), for a total of 10 predictive attributes. There is one Class attribute that describes the "Poker Hand". The order of cards is important, which is why there are 480 possible Royal Flush hands as compared to 4.

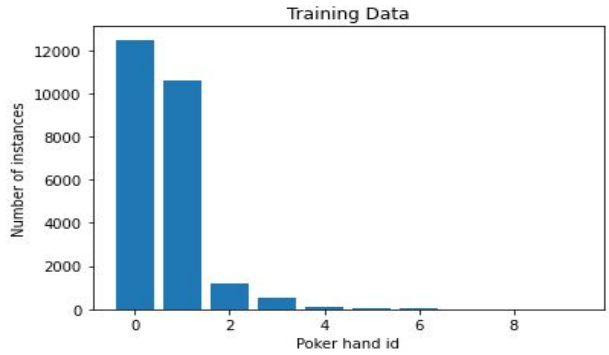
VARIOUS DATA QUALITY ISSUES IN DATASET

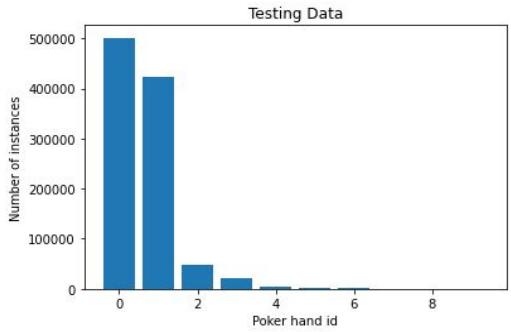
Class Imbalance	**Yes**
Missing Data	No
Outliers	No
Duplicate data	**Yes**
Skewed data	No
Data Homogeneity	No
Class Overlapping	**Yes**

VARIOUS DATA QUALITY ISSUE ANALYSIS OVER GIVEN DATASET

Class Imbalance

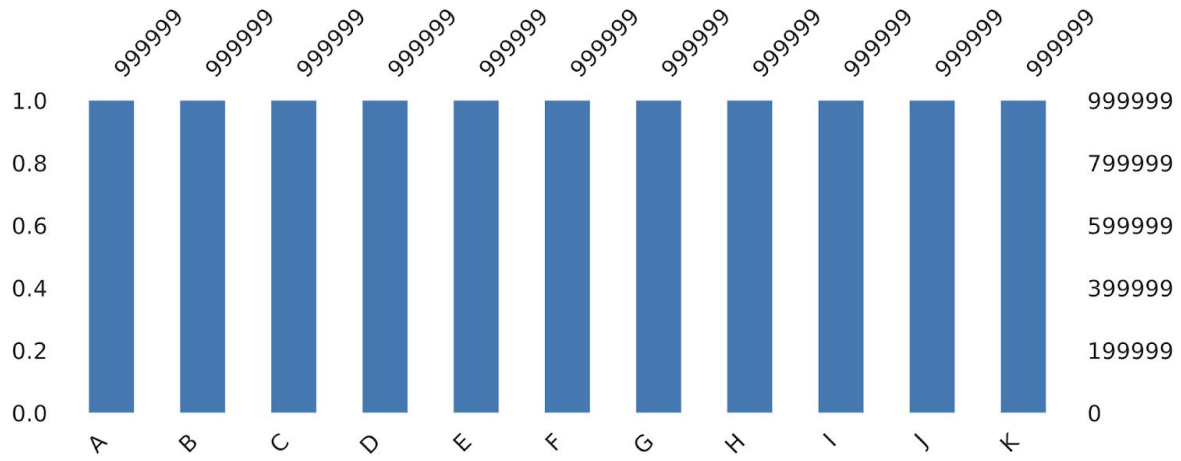
The given dataset is extremely imbalanced. Imbalanced data is very challenging for machine learning algorithms used for classification as they work in a way where they assume that equal data is available for all the classes.





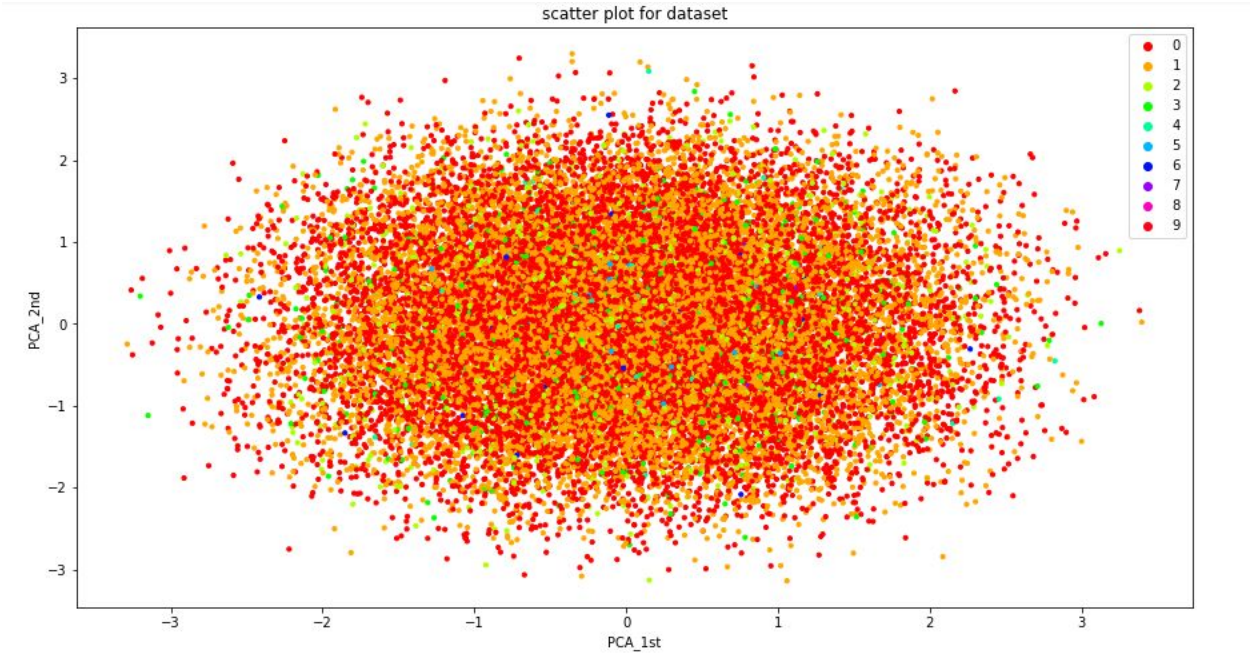
Missing Data

The dataset contains no missing values. It was mentioned in the description and verified through code too (using both sklearn and pandas profiling tool)



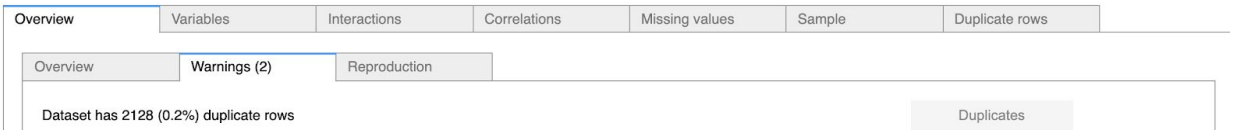
Outliers in data

There are no outliers present in the dataset as indicated by the plot given below:



Duplicate Data

According to the report generated by the pandas profiling tool, the dataset has 2128 duplicate rows. Multiple duplicate records can not only increase computation and storage,but also produce incorrect insights so it must be removed.



Skewed Data

Skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. It refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. The given dataset does not contain skewed data.(using pandas profiling tool)

Completeness

It is the fraction of non-null values in the dataset. Using pydeequ, for each attribute it is verified that whether each attribute follows completeness or not. It is found that all the attributes follow completeness. A value of 1 for completeness indicates no missing values.(Here each attribute is renamed from A to K)

```
Column 'E'
       completeness: 1.0
       approximate number of distinct values: 4
       datatype: Integral
Column 'J'
       completeness: 1.0
       approximate number of distinct values: 13
       datatype: Integral
Column 'F'
       completeness: 1.0
       approximate number of distinct values: 13
       datatype: Integral
Column 'A'
       completeness: 1.0
       approximate number of distinct values: 4
       datatype: Integral
Column 'I'
       completeness: 1.0
       approximate number of distinct values: 4
       datatype: Integral
Column 'G'
       completeness: 1.0
       approximate number of distinct values: 4
       datatype: Integral
Column 'B'
       completeness: 1.0
       approximate number of distinct values: 13
       datatype: Integral
Column 'C'
       completeness: 1.0
       approximate number of distinct values: 4
       datatype: Integral
Column 'H'
       completeness: 1.0
       approximate number of distinct values: 13
       datatype: Integral
Column 'K'
       completeness: 1.0
       approximate number of distinct values: 10
       datatype: Integral
```

```
Column 'D'
       completeness: 1.0
       approximate number of distinct values: 13
       datatype: Integral
```

Data Homogeneity

Using the profile (histogram) generated from pydeequ, it was inferred that all the features are integral and homogeneous. So there are no problems related to non uniform data format.

Class Overlapping

There is a class overlapping problem found in this dataset along with class imbalance. Here is the tsne(2 features) scatter plot to visualise the separation between the classes. It could be seen below.