DATA QUALITY REPORT FOR CREDIT CARDS AND CELL PHONES APPLICATION DATA

**Description**

**Dataset Name:** Credit Cards and Cell Phones Application Data

**Data Source:** DSO 562 project (synthetic data of real dataset)

**Time Period:** Jan. 2016 – Dec. 2016

**Number of Fields:** #10

**Number of Records:** 1,000,000

**Description:**

This is a synthetic dataset generated to mimic real application dataset. It includes 1 million application records from January 2016 to December 2016 with 10 fields. We aim to use supervised learning algorithms on this dataset to predict application/ identity fraud.

**Summary Table:**

**Numeric Fields:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| FIELD NAME | PERCENT POPULATED | MINIMUM VALUE | MAXIMUM VALUE | MEAN VALUE | STANDARD DEVIATION | PERCENT ZERO |
| **date** | 100 | 20160101 | 20161231 | NA | NA | 0 |
| **dob** | 100 | 19000101 | 20161031 | NA | NA | 0 |

**Categorical Fields:**

|  |  |  |  |
| --- | --- | --- | --- |
| FIELD NAME | PERCENT POPULATED | NUM OF UNIQUE VALUES | MOST COMMON VALUE |
| **record** | 100 | 1,000,000 | NA |
| **ssn** | 100 | 835,819 | 999999999 |
| **firstname** | 100 | 78,136 | EAMSTRMT |
| **lastname** | 100 | 177,001 | ERJSAXA |
| **address** | 100 | 828,774 | 123 MAIN ST |
| **zip5** | 100 | 26,370 | 68138 |
| **homephone** | 100 | 28,244 | 9999999999 |
| **fraud\_label** | 100 | 2 | 0 |

**Distribution of Fields:**

**Filed 1:**

**Name:** Record

**Description:** Unique identifier for each application. This field has 1 million unique integers.

**Field 2:**

**Name:** date

**Description:**

Date field represents the application date for each record and ranges from 2016-01-01 to 2016-12-31. No missing value or zero value appears in this field. In the histogram below, I divide the original dataset into 52 bins, which represent each 52 weeks in one year. Therefore, each bin represents the number of applications in each week.

**Graph:**

**Chart, histogram

Description automatically generated**

**Field 3:**

**Name:** SSN

**Description:**

SSN field represents the social security number used for each application. It is a 9-digit number issued to U.S. citizens, permanent residents and temporary (working) non-residents used to track individuals for taxation purposes. No missing value or zero value appears in this field. This filed has 835,819 unique values and the most common value is 999999999, which appears 16,935 times in this field. The graph below shows top 15 SSN number in this dataset.

**Graph:**

**Chart

Description automatically generated**

**Field 4:**

**Name:** firstname

**Description:**

Firstname field represents the first name of applicant for each application. No missing value or zero value appears in this field. This filed has 78,136 unique values and the most common value is EAMSRMT, which appears 12,658 times in this field. The graph below shows top 15 first name in this dataset.

**Graph:**

**Chart, bar chart

Description automatically generated**

**Field 5:**

**Name:** lastname

**Description:**

Lastname field represents the last name of applicant for each application. No missing value or zero value appears in this field. This filed has 177,001 unique values and the most common value is ERJSAXA, which appears 8,580 times in this field. The graph below shows top 15 last name in this dataset.

**Graph:**

**Chart, bar chart

Description automatically generated**

**Field 6:**

**Name:** address

**Description:**

Address field represents the address of applicant for each application. No missing value or zero value appears in this field. This filed has 828,774 unique values and the most common value is 123 MAIN ST, which appears 1,079 times in this field. However, the applications from this same address do not have same related information. The graph below shows top 15 address in this dataset.

**Graph:**

**Chart, histogram

Description automatically generated**

**Field 7:**

**Name:** zip5

**Description:**

Zip5 field represents the 5-digit zip code of applicant’s address for each application. No missing value or zero value appears in this field. This filed has 26,370 unique values and the most common value is 68138, which appears 823 times in this field. The graph below shows top 20 5-digit zip code in this dataset.

**Graph:**

**Chart, bar chart

Description automatically generated**

**Field 8:**

**Name:** dob

**Description:**

Dob field represents date of birth of applicant for each application. No missing value or zero value appears in this field. This filed ranges from 1900-01-01 to 2016-10-31 and has 42,673 unique values. The most common value is 1907-06-26, which appears 126,568 times in the dataset. The graph below shows the distribution of dob. This graph has 116 bins and each bin represent 1 year. We can find that the bin include year 1907 is higher than other bins.

**Graph:**

**Chart

Description automatically generated**

**Field 9:**

**Name:** homephone

**Description:**

Homephone field represents 10-digit home phone number of applicant for each application. No missing value or zero value appears in this field. This filed has 28,244 unique values. The most common value is 9999999999, which appears 78,512 times in the dataset. The graph below shows top 15 phone number in this dataset.

**Graph:**

**Chart, histogram

Description automatically generated**

**Field 10:**

**Name:** fraud\_label

**Description:**

Fraud\_label field represents if it is labeled as fraud activity for each application. No missing value or zero value appears in this field. This filed has 2 unique values. 0 represents non-fraud activity for this record and 1 represents fraud activity for this record. 0 and 1 occur 985,607 and 14,393 times correspondingly.

**Graph:**

**Chart

Description automatically generated**