**Applications Data DQR**

Lin Zhu

4/4/2019

1. Data Description

Data name: Applications data

Number of fields: 10

Number of records: 1000000

Details variables: record, date, ssn, firstname, lastname, address, zip5, dob, homephone, fraud\_lable.

1. List of Info for Each Field

The table below listed the name of the field, the number of records, the percentage of the population, the number of unique values and the number of records with value zero in each field.

|  | Field name | # records | %populated | #unique values | # records with value zero |
| --- | --- | --- | --- | --- | --- |
| 1 | record | 1000000 | 100 | 1000000 | 0 |
| 2 | date | 1000000 | 100 | 365 | 0 |
| 3 | ssn | 1000000 | 100 | 835819 | 0 |
| 4 | firstname | 1000000 | 100 | 78136 | 0 |
| 5 | lastname | 1000000 | 100 | 177001 | 0 |
| 6 | address | 1000000 | 100 | 828774 | 0 |
| 7 | Zip5 | 1000000 | 100 | 26370 | 0 |
| 8 | dob | 1000000 | 100 | 42673 | 0 |
| 9 | homephone | 1000000 | 100 | 28244 | 0 |
| 10 | Fraud\_label | 1000000 | 100 | 2 | 985607 |

Categorical Fields: date, ssn, firstname, lastname, address, zip5, homephone, fraud\_lable.

Date/Time: date, dob

1. Description of each field

**record**:

Name: “record”

Field type: Categorical

Description: Number to track each record

All numbers in this field are unique

**Date**:

Name: “date”

Field type: Date/ Time

Description: the application date

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the date field. The most common value in this field is 20160816.

In order to get more insight, let’s take a look at the distribution of data in the month unit.

**A screenshot of a cell phone

Description automatically generated**

**ssn**:

Name: “ssn”

Field type: Categorical

Description: Social security number of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the ssn field. The most common value in this field is 999999999.

**First name**:

Name: “firstname”

Field type: Categorical

Description: first name of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the firtname field. The most common value in this field is EAMSTRMT.

**Last name**:

Name: “lastname”

Field type: Categorical

Description: last name of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the lastname field. The most common value in this field is ERJSAXA.

**Address**:

Name: “address”

Field type: Categorical

Description: address of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the address field. The most common value in this field is 123 MAIN ST.

**Zip code**:

Name: “zip5”

Field type: Categorical

Description: five first digits of zip code of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the zip5 field. The most common value in this field is 68138.

**Birthday**:

Name: “dob”

Field type: Date/ Time

Description: date of birth of the applicant

A screenshot of a cell phone

Description automatically generated

The bar chart displays the first 20 values in the dob field. The most common value in this field is 19070626.

In order to get more insight, let’s take a look at the distribution of data in the year unit.

A screenshot of a cell phone

Description automatically generated

From the plot, we find the value of dob is extremely high in 1907 which is 133986.

**Phone number**:

Name: “homephone”

Field type: Categorical

Description: phone number of the applicant

A screenshot of a cell phone

Description automatically generated

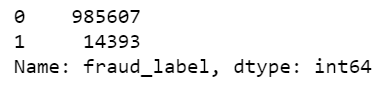
The bar chart displays the first 20 values in the homephone field. The most common value in this field is 9999999999.

**Fraud**:

Name: “Fraud”

Field type: Binary

Description: whether if the transaction is a fraud



0 represents the good records, while 1 means bad records which is a fraud. Most of the entries are labeled as not fraud.

No missing data.

A close up of a logo

Description automatically generated