**Data Audit Report**

Prepared by

Mounica Madireddy

Kornkanok Somkul

In support of

Predictive Model of Employee Voluntarily Turnover

Requested by

The Human Resources Department

June 3, 2018

**Introduction**

Finding the main determining attribute which mainly affects the target sample from the selected qualifications merged from a vast dataset is considered to be the main objective of this Audit Report.

The analytical team has been asked by the SVP of Human Resources to build a predictive model to identify current employees who might be thinking to leave the company. The deliverable will be a scored file.

The target sample qualifications provided by the HR department are as follows:

* Former employees who voluntarily attritioned
* who took the survey

The non-target sample qualifications provided by the HR department are as follows:

* Current employees
* who took the survey.

It is essential that the analytical team has a full understanding of the quality and quantity of data provided to it in support of the analytical request.

Hence, the purpose of this data audit is to ensure that:

• all data received by the analytical team for the project are consistent with the team’s understanding of the requested analytical deliverable;

• that the team is reading and interpreting these data correctly;

• that the team has received all data intended to be supplied;

• that the data are functionally usable for modeling purposes.

The data audit is broken into four main sections:

1. Dataset Summary – A list of all datasets received.
2. Dataset Detail – For each dataset, tables showing all data variables received. It is important that this section be reviewed to ensure that the analytical team has all the data sent, the data are being read correctly and the data have reasonable values.
3. Modeling Sample – Based on the requestor’s sample requirements, a determination is necessary as to whether adequate sample is available to support modeling.
4. Questions – Specific questions that the analytical team needs answered to ensure that the team fully understands the data and that the data can support the requested analytical deliverable.

**Dataset Summary**

Fortune Corp, a maker of specialized laboratory equipment for the pharmaceutical industry, began business in June 1980. Priding itself on employee job satisfaction, the company is seeking to understand why employees voluntarily leave the company.

The analytical team has received 5 data files from the IT department as listed in Table 1. The team also received a data document which defines the data fields shown below.

*Table 1. Data Files Received*

|  |  |  |
| --- | --- | --- |
| File Name | File Type | File Contents |
| FORTUNE \_CREDIT | CSV | Employees’ FICO Score |
| FORTUNE \_ACCT | SAS | Payroll data |
| FORTUNE \_ATTRITION | SAS | Employee attrition information during 2015-2017 |
| FORTUNE \_ HR | SAS | Employees’ background |
| FORTUNE \_SURVEY | SAS | Employee Survey |

**Dataset Detail**

Each dataset contains the data fields as shown in the following tables.

Dataset #1: FORTUNE \_CREDIT

File Contents: This dataset contains 2 variables with 4867 observations. Out of which, 2 attributes are of numerical datatype. These attributes are summarised in table 2 .

*Table 2. Numerical Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **Count** | **Missing** | **Mean** | **Median** | **Min** | **Max** |
| SSN | 4867 | 0 | 100553379 | 555327345 | 550735837 | 999995259 |
| FICO\_SCR | 4867 | 0 | 675 | 727.1853298 | 726 | 820 |

Dataset #2: FORTUNE\_ACCT

File Contents: This dataset contains 10 variables with 4867 observations. Out of which, 5 attributes are of numerical datatype ,1 attribute is of character datatype and 4 attributes are of categorical data These attributes are summarised in table 3, table 4 and table 5.

*Table 3. Numerical Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **Count** | **Missing** | **Mean** | **Median** | **Min** | **Max** |
| DAILY RATE | 4775 | 92 | 10.2 | 801.4532356 | 798 | 1499 |
| HOURLY RATE | 4867 | 0 | 30 | 65.8463119 | 66 | 100 |
| MONTHLY INCOME | 4775 | 92 | 1009 | 6609.52 | 4908 | 199999 |
| PERCENT SALARY HIKE | 4867 | 0 | 11 | 15.2202589 | 14 | 25 |
| EMPLOYEE \_NO | 4867 | 0 | 2316 | 500918.04 | 497846 | 999908 |

*Table 4. Character Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Length of Character Field** | | | | | |
| **Field Name** | **Count** | **Missing** | **Min** | **Mean** | **Median** | **Max** |
| SSN | 4867 | 0 | 12 | 12 | 12 | 12 |

*Table 5. Categorical Data*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Frequency | | | |
| **Field Name** | **Missing** | **Value** | **Count** | **Percent** |
| DEPARTMENT | 0 | Human Resources | 222 | 4.56 |
| Research & D | 83 | 1.71 |
| Research & Development | 3065 | 62.98 |
| Sales | 1497 | 30.76 |
| OVERTIME | 0 | No | 3497 | 71.85 |
| Yes | 1370 | 28.15 |
| PERFORMANCERATING | 0 | 3 | 4117 | 84.59 |
| 4 | 750 | 15.41 |
| STOCKOPTIONLEVEL | 0 | 0 | 2154 | 44.26 |
| 1 | 1920 | 39.45 |
| 2 | 507 | 10.42 |
| 3 | 286 | 5.88 |

Dataset #3: FORTUNE\_ATTRITION

File Contents: This dataset contains 262 observations and 2 variables. Out of which, 1 attribute is of numerical datatype and the other attribute is of date datatype. These attributes are summarised in table 6 and table 7.

*Table 6. Numerical Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Field Name | Count | Missing | Mean | Median | Min | Max |
| EMPLOYEE\_NO | 262 | 0 | 4043 | 523493.02 | 523913 | 997607 |

*Table 7. Date Data*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | Min Freq | | | Max Freq | | |
| Field Name | Count | Missing | Oldest | Most Recent | Year | | Count | Year | | Count |
| DEPART\_DT | 262 | 0 | 3/1/2015 | 12/31/2017 | 2016 | | 72 | 2017 | | 98 |

Dataset #4: FORTUNE\_HR

File Contents: This dataset contains 4867 observations and 8 variables. Out of which, 1 attribute is of numerical datatype, 1 attribute is of character datatype, 4 attributes are of categorical datatype and 2 attributes are of date datatype. These attributes are summarised in table 8, table 9 table 10 and table 11.

*Table 8. Numerical Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Field Name | Count | Missing | Mean | Median | Min | Max |
| EMPLOYEE\_NO | 4867 | 0 | 2316 | 500918.04 | 497846 | 999908 |

*Table 9. Character Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Length of Character Field | | | | | |
| Field Name | Count | Missing | Min | Mean | Median | Max |
| FIRST\_NAME | 4867 | 0 | 12 | 12 | 12 | 12 |

*Table 10. Categorical Data*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Frequency | | |
| FIELD NAME | Missing | Value | Count | Percent |
| EDUCATION | 0 | 1 | 565 | 11.61 |
|  | 2 | 916 | 18.82 |
|  | 3 | 1881 | 38.65 |
|  | 4 | 1332 | 27.37 |
|  | 5 | 173 | 3.55 |
| EDUCATIONFIELD | 0 | Human Resources | 94 | 1.93 |
| LS | 463 | 9.51 |
| Life Sciences | 1532 | 31.48 |
| Marketing | 447 | 9.18 |
| Medical | 1524 | 31.31 |
| Mkt | 105 | 2.16 |
| Other | 260 | 5.34 |
| Tech | 91 | 1.87 |
| Technical Degree | 351 | 7.21 |
| GENDER | 0 | Female | 1774 | 36.45 |
| Male | 2734 | 56.17 |
| N/A | 359 | 7.38 |
| BIRTH\_STATE | 648 | AK | 99 | 2.35 |
| AL | 96 | 2.28 |
| AR | 74 | 1.75 |
| AZ | 72 | 1.71 |
| CA | 82 | 1.94 |
| CO | 73 | 1.73 |
| CT | 79 | 1.87 |
| DC | 102 | 2.42 |
| DE | 89 | 2.11 |
| FL | 94 | 2.23 |
| GA | 72 | 1.71 |
| HI | 59 | 1.4 |
| IA | 94 | 2.23 |
| ID | 103 | 2.44 |
| IL | 70 | 1.66 |
| IN | 107 | 2.54 |
| KS | 111 | 2.63 |
| KY | 93 | 2.2 |
| LA | 106 | 2.51 |
| MA | 99 | 2.35 |
| MD | 114 | 2.7 |
| ME | 94 | 2.23 |
| MI | 78 | 1.85 |
| MN | 84 | 1.99 |
| MO | 76 | 1.8 |
| MS | 89 | 2.11 |
| MT | 100 | 2.37 |
| NC | 97 | 2.3 |
| ND | 50 | 1.19 |
| NE | 80 | 1.9 |
| NH | 74 | 1.75 |
| NJ | 107 | 2.54 |
| NM | 106 | 2.51 |
| NV | 145 | 3.44 |
| NY | 100 | 2.37 |
| OH | 106 | 2.51 |
| OK | 90 | 2.13 |
| OR | 90 | 2.13 |
| PA | 137 | 3.25 |
| RI | 67 | 1.59 |
| SC | 58 | 1.37 |
| SD | 90 | 2.13 |
| TN | 108 | 2.56 |
| TX | 94 | 2.23 |

*Table 11. Date Data*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | Min Freq | | Max Freq | |
| Field Name | Count | Missing | Oldest | Most Recent | Year | Count | Year | Count |
| BIRTH\_DT | 4597 | 270 | 6/12/1956 | 5/27/1999 | 1956 | 7 | 1982 | 283 |
| HIRE\_DT | 4867 | 0 | 10/10/1975 | 12/11/2017 | 1975 | 1 | 2015 | 521 |

Dataset #5: FORTUNE\_SURVEY

File Contents: This dataset contains 1470 observations and 16 variables. Out of which, 6 attributes are of numerical datatype and 8 attributes are of categorical datatype. These attributes are summarised in table 12 and table 13.

*Table 12. Numerical Data*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Field Name | Count | Missing | Mean | Median | Min | Max |
| DISTANCEFROMHOME | 1470 | 0 | 1 | 9.192517 | 7 | 29 |
| NUMCOMPANIESWORKED | 1470 | 0 | 0 | 2.6931973 | 2 | 9 |
| TOTALWORKINGYEARS | 1470 | 0 | 0 | 11.279592 | 10 | 40 |
| YEARSWITHCURRMANAGER | 1470 | 0 | 0 | 4.1231293 | 3 | 17 |
| EMPLOYEE\_NO | 1470 | 0 | 2583 | 510126.17 | 508447.5 | 999834 |
| TRAININGTIMESLASTYEAR | 1470 | 0 | 0 | 2.7993197 | 3 | 6 |

*Table 13. Categorical Data in dataset* FORTUNE\_ACCT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Frequency | | |
| Field Name | Missing | Value | Count | Percent |
| ENVIRONMENTSATISFACTION | 0 | 1 | 284 | 19.32 |
| 2 | 287 | 19.52 |
| 3 | 453 | 30.82 |
| 4 | 446 | 30.34 |
| JOBINVOLVEMENT | 0 | 1 | 83 | 5.65 |
| 2 | 375 | 25.51 |
| 3 | 868 | 59.05 |
| 4 | 144 | 9.8 |
| JOBLEVEL | 0 | 1 | 543 | 36.94 |
| 2 | 534 | 36.33 |
| 3 | 218 | 14.83 |
| 4 | 106 | 7.21 |
| 5 | 69 | 4.69 |
| JOBSATISFACTION | 0 | 1 | 289 | 19.66 |
| 2 | 280 | 19.05 |
| 3 | 442 | 30.07 |
| 4 | 459 | 31.22 |
| MARITALSTATUS | 100 | Divorced | 296 | 21.61 |
| Married | 635 | 46.35 |
| Single | 439 | 32.04 |
| RELATIONSHIPSATISFACTION | 0 | 1 | 276 | 18.78 |
| 2 | 303 | 20.61 |
| 3 | 459 | 31.22 |
| 4 | 432 | 29.39 |
| WORKLIFEBALANCE | 0 | 1 | 80 | 5.44 |
| 2 | 344 | 23.4 |
| 3 | 893 | 60.75 |
| 4 | 153 | 10.41 |
| BUSINESSTRAVEL | 0 | NON-TRAVEL | 150 | 10.2 |
| TRAVEL\_FREQUENTLY | 277 | 18.84 |
| TRAVEL\_RARELY | 1043 | 70.95 |

**Modelling Sample**

|  |  |
| --- | --- |
| **SEGMENT** | **COUNT** |
| Employees who voluntarily attritioned | 262 |
| And took the survey | 262 |
|  |  |
| **Available target sample** | **262** |
|  |  |
| Current employees | 4630 |
| And took the survey | 1233 |
|  |  |
| **Available non-target sample** | **1233** |
|  |  |

Note: the target sample size is on the low end of being acceptable.

**Questions**

1. Does the above information appear to be correct? Specifically:
   * Does the analytical team have all the data that was meant to be sent?
   * Is the team interpreting the data correctly?
   * Do the data appear to have reasonable values?
2. The following are specific questions the analytical team has about the data...
   * Are there additional former employee records available to increase the size of the target sample?
   * SSN is a character value in the FORTUNE\_ACCT dataset, while SSN in the FORTUNE\_CREDIT is numeric. So, the team has to convert SSN in the FORTUNE\_ACCT dataset to numeric…is this ok?
   * Are there any additional information indicates that how the survey have been collected? and the survey ranking is the highest number being the most positive, is that correct?