**Data Audit Report**

Prepared by

Charlene Wolthers

In support of

Predictive model of employee voluntary attrition

Requested by

The SVP of Human Resources at Fortune Corp

March 31, 2024

**Introduction**

The analytics team is tasked with developing a predictive model aimed at improving the understanding of why employees voluntarily leave the company. To facilitate this request, the team has received several datasets containing information on current donors and past employees.

The qualifications for the target sample are those who have taken the survey. This sample will be broken up into two segments:

1. Employees who voluntarily attritioned.
2. Employees who are still with the company.

These datasets are provided to aid in constructing a that a predictive model of employee voluntary attrition. The objective is to use this model to find current employees who might be thinking of leaving, so proactive steps can be taken to retain them. There is no data-dictionary for these datafiles. The project's endpoint involves consolidating these files into a single, comprehensive dataset for analysis.

The purpose of this data audit is to ensure the following:

1. The analytics team has received all datafiles intended for this project.
2. The analytics team understands the content, layout, and format of these files.
3. The data in these files are of sufficient integrity and quantity to support the model development.

This data audit consists of 4 sections:

1. Datafile Summary: A list and description of all datafiles received.
2. Datafile Detail: For each datafile, tables showing all data fields received, their values, summary statistics, and distributions. Data fields are categorized into one of 4 types of analytical variables:
   * Categorical - data fields with distinct levels or values which represent categories; can be a number or a label, nominal or ordinal.
   * Date - data fields that are identified as calendar dates.
   * Numeric - data fields that are continuous numeric data.
   * Character - data fields whose values are characters and are not otherwise classified as categorical.
3. Modeling Sample – After merging all supplied datafiles, a determination is made as to whether there is adequate sample size for each target sample to support model development.
4. Questions – The auditing process will uncover data integrity issues. This section lists what the analytics team has found in this regard. This section also poses specific questions on data field definitions, field coding, and interpretation, answers to which will facilitate the team’s model development effort.

**Datafile Summary**

The analytics team has received 5 datafiles from The SVP of Human Resources as listed in Table 1.

Table 1. Datafiles Received

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | File Type | # Of Records | File Contents |
| fortune\_acct | SAS | 4,867 | Payroll data |
| fortune\_attrition | SAS | 262 | Employees who left the company over the 2015-2017 period |
| fortune\_credit | CSV | 4,867 | FICO score |
| fortune\_hr | SAS | 4,867 | Background employee data |
| fortune\_survey | SAS | 1,470 | Data collected from the employee survey |

**Datafile Detail**

Each datafile contains the analytic data fields as shown in the following tables. Note that the data fields have been classified based on their potential analytical usage.

|  |
| --- |
| ***Datafile #1:*** fortune\_acct  ***File Analytic Contents:***  Numeric Fields (4): DailyRate HourlyRate MonthlyIncome PercentSalaryHike  Categorical Fields (4): Department OverTime PerformanceRating StockOptionLevel  Character Fields (2): employee\_no ssn  Date Fields (0):  ***Records:*** 4,867  ***Columns:*** 10 |
| ***Notes:*** The data field employee\_no appears to be a row id or index field. |

Table 1. fortune\_acct - Numeric Data

|  |  |
| --- | --- |
| *Analytic numeric data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 2. fortune\_acct - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |

Table 3. fortune\_acct - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #2:*** fortune\_attrition  ***File Analytic Contents:***  Numerical (0):  Categorical (0):  Character (1): employee\_no  Date (1): depart\_dt  ***Records:*** 262  ***Columns:*** 2  ***Notes:*** The data field employee\_no appears to be a row id or index field. |

Table 4. fortune\_attrition - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |

Table 5. fortune\_attrition - Date Data

|  |  |
| --- | --- |
| *Analytic date data fields* |  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #3:*** fortune\_credit  ***File Analytic Contents:***  Numerical (1): fico\_scr  Categorical (0):  Character (1): ssn  Date (0):  ***Records:*** 4,867  ***Columns:*** 2  ***Notes:*** The data field ssn appears to be a row id or index field. |

Table 7. fortune\_credit - Numerical Data

|  |  |
| --- | --- |
| *Analytic numerical data fields* |  |
|  |  |
|  |

Table 8. fortune\_credit - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #4:*** fortune\_hr  ***File Analytic Contents:***  Numerical (0):  Categorical (4): Education EductaionField Gender birth\_state  Character (2): employee\_no first\_name  Date (2): birth\_dt hire\_dt  ***Records:*** 4,867  ***Columns:*** 8  ***Notes:*** The data field employee\_no appears to be a row id or index field. |

Table 10. fortune\_hr - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 11. fortune\_hr - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |
|  |  |
|  |

Table 12. fortune\_hr - Date Data

|  |  |
| --- | --- |
| *Analytic date data fields* |  |
|  |  |
|  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #5:*** fortune\_survey  ***File Analytic Contents:***  Numerical (7): DistanceFromHome TotalWorkingYears NumCompaniesWorked TrainingTimesLastYear YearsInCurrentRole YearsSinceLastPromotion  YearsWithCurrManager  Categorical (8): BusinessTravel EnvironmentSatisfaction JobInvolvement JobLevel JobSatisfaction MaritalStatus RelationshipSatisfaction WorkLifeBalance  Character (1): employee\_no  Date (0):  ***Records:*** 1,470  ***Columns:*** 16  ***Notes:*** The data field employee\_no appears to be a row id or index field. |

*Table 13. fortune\_survey - Numerical Data*

|  |  |
| --- | --- |
| *Analytic numerical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 14. fortune\_survey - Categorical Data

|  |  |
| --- | --- |
| *Analytic categorical data fields* |  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |
|  |  |
|  |

Table 15. fortune\_survey - Character Data

|  |  |
| --- | --- |
| *Analytic character data fields* |  |
|  |  |
|  |

**Modeling Sample**

|  |  |
| --- | --- |
| **Segment** | **Count** |
| **Available event (yes) sample** | 262 |
|  |  |
| **Available non-event (no) sample** | 1,208 |
|  |  |
| **Total (target) sample** | 1,470 |
|  |  |
| **Total records in datafile** | 4,867 |

Sample sizes in the target segments appear to be of sufficient size to support the predictive model.

**Questions**

1. Does the above information appear to be correct? Specifically:
   * Does the analytics 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. Here is a list of the data integrity issues the analytics team uncovered. Please review:

* DailyRate missing in 92 cases (1.9%).
* DailyRate max value is 1,499…accurate? 10.2 min value (all other values are integers)…accurate?
* MonthlyIncome missing in 92 cases (1.9%).
* MonthlyIncome max value is 199,999…accurate?
* Employee\_no variable in the fortune\_attrition datafile may contain possible duplicate records. There was a frequency count of 2 for several employee\_no.
* Gender missing in 359 cases (7.4%) – all categorized under N/A.
* Birth\_state missing in 648 cases (13.3%).
* MaritalStatus missing in 100 cases (6.8%).
* Distribution is heavily right-skewed for MonthlyIncome, YearsSinceLastPromotion, and YearsWithCurrManager.
* Distribution is heavily left-skewed for hire\_dt.
* Bimodal distribution for YearsInCurrentRole.

1. The following are specific questions the analytics team has about the data. Please review:

* Is DailyRate an average calculation? Does it include overtime pay?
* PercentSalaryHike was calculated over what period of time?
* What do values for PerformanceRating mean? Same question for Education, EnvironmentSatisfaction, JobLevel, JobInvolvement, RelationshipSatisfaction, WorkLifeBalance, StockOptionLevel, and JobSatisfaction.
* Are the values for the variable, DistanceFromHome, measured in miles?
* Please provide clarification of definition for the following variables: TrainingTimesLastYear, NumCompaniesWorked, and TotalWorkingYears.