**Data Audit Report**

Prepared by

Jason Noble

In support of

Predictive model to identify employees who are thinking of leaving Fortune Corp

Requested by

Fortune Corp, SVP of Human Resources

September 30, 2024

**Introduction**

The analytics team has been asked by the Fortune Corp Senior VP of HR is seeking to understand why employees voluntarily leave the company. We will first need to confirm that enough data has been collected so that a predictive model of employee voluntary attrition can be built and tested.

It is the analytics team’s understanding that the target variable segments for this model are:

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

The goal is to build a predictive model of employee attrition in order to use the model to find current employees who might be thinking of leaving, so proactive steps can be taken to retain them.

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:
   * Numeric - data fields that are continuous numeric data.
   * Categorical - data fields with distinct levels or values which represent categories; can be a number or a label, nominal or ordinal.
   * Character - data fields whose values are characters and are not otherwise classified as categorical.
   * Date - data fields that are identified as calendar dates.
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 Fortune Corp HR Senior VP, as listed in Table 1.

Table 1. Datafiles Received

|  |  |  |  |
| --- | --- | --- | --- |
| Filename | Type | # Of Records | File Contents |
| Fortune\_acct | SAS | 4,867 | Payroll and performance data |
| Fortune\_attrition | SAS | 262 | Employees who left between 2015-2017 |
| Fortune\_hr | SAS | 4,867 | Background employee data |
| Fortune\_survey | SAS | 1,470 | Data collected from the employee survey |
| Fortune\_credit | csv | 4,867 | FICO score (SVP thinks this might be predictive) |

**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 (5): Employee\_no DailyRate HourlyRate  MonthlyIncome PercentSalaryHike  Categorical Fields (4): PerformanceRating StockOptionLevel  Department OverTime  Character Fields (1): SSN  Date Fields (0):  ***Records:*** 4,867  ***Columns:*** 10 |
| ***Notes:*** The data field Employee\_no appears to be a field we could utilize for our ID and merge.  Will need to remove dashes in SSN and make numeric to merge with file #5 |

Table 2. Fortune\_ACCT – Numeric Data

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

Table 3. Fortune\_ACCT – Categorical Data

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

|  |  |
| --- | --- |
|  |  |
|  |
|  |  |
|  |

|  |  |
| --- | --- |
|  |  |
|  |
|  |  |
|  |

Table 4. Fortune\_ACCT – Character Data

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

|  |
| --- |
| ***Datafile #2:*** Fortune\_attrition  ***File Analytic Contents:***  Numeric Fields (1): Employee\_no  Categorical Fields (0):  Character Fields (0):  Date Fields (1): depart\_dt  ***Records:*** 262  ***Columns:*** 2 |
| ***Notes:*** The data field Employee\_no appears to be a field to ID and merge. |

Table 5. Fortune\_attrition – Numeric Data

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

Table 6. Fortune\_attrition – Date Data

|  |  |
| --- | --- |
| *Analytic calendrical fields* |  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #3:*** Fortune\_HR  ***File Analytic Contents:***  Numeric Fields (1): Employee\_no  Categorical Fields (4): Education EducationField Gender birth\_state  Character Fields (1): first\_namE  Date Fields (2): Birth\_dt hire\_dt  ***Records:*** 4,867  ***Columns:*** 8 |
| ***Notes:*** The data field Employee\_no appears to be a field we could utilize to ID and merge. |

Table 7. Fortune\_HR - Numeric Data

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

Table 8. Fortune\_HR - Categorical Data

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

Table 9. Fortune\_HR - Character Data

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

Table 10. Fortune\_HR – Date Data

|  |  |
| --- | --- |
| *Analytic calendrical fields* |  |
|  |  |
|  |
|  |  |
|  |

|  |
| --- |
| ***Datafile #4:*** Fortune\_survey  ***File Analytic Contents:***  Numeric Fields (8): Employee\_no DistanceFromHome NumCompaniesWorked TotalWorkingYears TrainingTimesLastYear YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  Categorical Fields (8): BusinessTravel EnvironmentSatisfaction JobInvolvement JobLevel  JobSatisfaction RelationshipSatisfaction MaritalStatus WorkLifeBalance  Character Fields (0):  Date Fields (0):  ***Records:*** 1,470  ***Columns:*** 16 |
| ***Notes:*** The data field Employee\_no appears to be a field we could utilize to ID and merge. |

Table 11. Fortune\_survey - Numeric Data

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

Table 12. Fortune\_survey - Categorical Data

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

|  |
| --- |
| ***Datafile #5:*** Fortune\_credit  ***File Analytic Contents:***  Numeric Fields (2): FICO\_SCR SSN  Categorical Fields (0):  Character Fields (0):  Date Fields (0):  ***Records:*** 4,867  ***Columns:*** 2 |
| ***Notes:*** The data field SSN appears to be a field we could utilize to merge with Fortune\_acct, file #1.  Will need to make values integers, lose the decimals, in order to merge. |

Table 13. Fortune\_credit - Numeric Data

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

**Modeling Sample**

|  |  |
| --- | --- |
| **Segment** | **Count** |
| **Event --Employees that voluntarily left and took survey** | 262 |
| DEPART\_DT |  |
| **Non-Event -- Current Employees that took survey** | 1,233 |
| BUSINESSTRAVEL but no DEPART\_DT |  |
| **Total (target) sample** | 1,495 |
|  |  |
| **Total records in datafile** | 4,892 |

Sample sizes in the target segments appear to be of sufficient size to support the predictive model. Even though the event segment is on the low side with 262 observations, the ratio of non-events to events is about 5:1.

**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 has 92 missing values, clarify how to handle
* DailyRate: confirm 10.2 is 102
* Attrition Employee\_no has 25 duplicates, confirm only 237 employees left in time frame
* Gender has 359 values marked N/A, clarify how to handle
* Birth\_state has 648 missing values, clarify how to handle
* MARITALSTATUS has 100 missing values, clarify how to handle

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

* Employee\_no appears to have embedded info with a range of 2,316 to 999,908 for 5,000 people, explain any meaning in the employee numbers
* There are only 237 distinct Employee\_no values in the attrition file
* Define PerformanceRating categories
* Define StockOptionLevel categories
* Define ENVIRONMENTSATISFACTION categories
* Define JOBINVOLVEMENT categories
* Define JOBLEVEL categories
* Define JOBSATISFACTION categories
* Define RELATIONSHIPSATISFACTION categories
* Define WORKLIFEBALANCE categories