Data Audit Report

Prepared by
Ryan Paw

In support of
Predictive model of employee voluntary attrition can be built and tested

Requested by
SVP of Human Resources

November 22, 2020

Introduction

The analytical team has been asked by SVP of Human Resources to build a predictive model of employee voluntary attrition can be built and tested.

The target sample qualifications, provided by SVP of Human Resources, are employees having taken the survey:

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

From this sample, the target segments for modeling will be:

1. Yes/event (1): Employees who voluntarily attritioned (left the company)
2. No/non-event (0): Employees who are still with the company

Assume the analysis is taking place in June 1, 2018.

The timeline for this project is as follows:

Milestone	Timeline
Data Audit/Aggregation	Week 1
Data Cleansing and Preparation	Weeks 2 - 3
Modeling Construction	Week 4
Scoring of Marketing File	Week 5
Marketing Campaign Commencement	Week 6

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. **Data File Summary** – A list and description of all data files received.
2. **Data File Detail** – For each data file, 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.

Data File Summary

The analytical team has received 5 data files from the SVP of Human Resources as listed in Table 1.

Table 1. Data Files Received

Filename	File Type	File Contents
fortune_credit	CSV	FICO score
fortune_acct	SAS	Payroll data
fortune_attrition	SAS	Employees who have left the company over the 2015-2017 period
fortune_hr	SAS	Background employee data
fortune_survey	SAS	Data collected from the employee survey

Data File Detail

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

Data File #1: fortune_credit
File Contents: fico_scr, ssn
Records: 4867

Table 2. Numeric Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
fico_scr	4867	0	675	727.1853298	726	820
ssn	4867	0	100553379	555327345	550735837	999995259

Data File #2: fortune_acct
File Contents: DailyRate, Department, HourlyRate, MonthlyRate, OverTime, PercentSalaryHike, PerformanceRating, StockOptionLevel, employee_no, ssn
Records: 4867

Table 3. Numeric Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
DailyRate	4775	92	10.2	801.4532356	798	1499
HourlyRate	4867	0	30	65.8463119	66	100
MonthlyIncome	4775	92	1009	6609.52	4908	199999
PercentSalaryHike	4867	0	11	15.2202589	14	25
employee_no	4867	0	2316	500918.04	497846	999908

Table 4. 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

Table 5. Character Data

	Length of Character Field					
Variable Name	**Count**	**Missing**	**Minimum**	**Mean**	**Median**	**Maximum**
ssn	4867	0	11	11	11	11

Data File #3: fortune_attrition
File Contents: depart_dt, employee_no
Records: 262

Table 6. Numeric Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
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	1/3/15	12/31/17	2016	72	2017	98

Data File #4: fortune_hr
File Contents: Education, EducationField, Gender, birth_dt, birth_state, employee_no, first_name, hire_dt
Records: 4867

Table 8. Numeric Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
employee_no	4867	0	2316	500918.04	497846	999908

Table 9. 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
		UT	91	2.16
		VT	120	2.84

Table 10. Character Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
first_name	4867	0	2	6.1588247	6	14

Table 11. Date Data

					Min Freq		Max Freq	
Field Name	Count	Missing	Oldest	Most Recent	Year	Count	Year	Count
birth_dt	4597	270	6/12/56	5/27/99	1956	7	1982	283
hire_dt	4867	0	10/10/75	12/11/17	1975	1	2015	521

Data File #5: fortune_survey
File Contents: BusinessTravel, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobSatisfaction, MaritalStatus, NumCompaniesWorked, RelationshipSatisfaction, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, employee_no
Records: 1470

Table 12. Numeric Data

Field Name	Count	Missing	Minimum	Mean	Median	Maximum
DistanceFromHome	1470	0	1	9.192517	7	29
NumCompaniesWorked	1470	0	0	2.6931973	2	9
TotalWorkingYears	1470	0	0	11.2795918	10	40
TrainingTimesLastYear	1470	0	0	2.7993197	3	6
YearsInCurrentRole	1470	0	0	4.2292517	3	18
YearsSinceLastPromotion	1470	0	0	2.1877551	1	15
YearsWithCurrManager	1470	0	0	4.1231293	3	17
employee_no	1470	0	2583	510126.17	508447.5	999834

Table 13. Categorical Data

		Frequency		
Field Name	**Missing**	**Value**	**Count**	**Percent**
BusinessTravel	0	Non-Travel	150	10.2
		Travel_Frequently	277	18.84
		Travel_Rarely	1043	70.95
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

Qualified Sample

Segment	Count
Employees who voluntarily attritioned (left the company)	262
Employees who took the survey	262
Available event (yes) sample	262
Employees who did not attrition (active employee)	4630
Employees who took the survey	1233
Available non-event (no) sample	1233
Total qualified (target) sample	1495
Total records in dataset	4892

Note – The target sample only includes employees that took the survey. 1495 out of the 4892 employees in this dataset completed the survey. If more employees took the survey, the qualified sample size would likely increase.

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. Here is a list of the data integrity issues the analytical team uncovered:
    - Birth_dt missing in 270 cases (5.87%)
    - DailyRate missing in 92 cases (1.89%)
    - MonthlyIncome missing in 92 cases (1.89%)
    - Birth_state missing in 648 cases (15.36%)
    - MaritalStatus missing in 100 cases (13.70%)

3. The following are specific questions the analytical team has about the data…
    - In the fortune_acct file, "Department" has 2 values that look similar to each other ("Research & D" and "Research & Development"). Are these files related? If so, do these department types need to be merged?
    - In the fortune_hr file, "hire_dt" includes dates before the company started in June 1, 1980: 10/10/75, 8/3/76, 10/6/79, 10/27/79, 12/7/79, 2/10/80.
    - In the fortune_acct file, "PerformanceRating" only has values 3 and 4. What does 3 and 4 represent? Is there supposed to be lower or higher values?
    - In fortune_hr, "EducationField" has 2 values that look similar to each other ("Life Sciences" and "LS"). Are these files related? If so, do these fields need to be merged?

- In fortune_hr, “EducationField” has 2 values that look similar to each other (“Tech” and “Technical Degree”) Are these files related? If so, do these fields need to be merged?
- In fortune_hr, “Gender” has a 359 “N/A” values. What does this represent? Do more gender categories need to be created (Transgender, non-binary, etc.)? Or is this a data entry error?
- Only 1470 out of the 4892 employees took the survey (30.20%). Why didn’t more employees take the survey?