EMPLOYEE ATTRITION

A DATA MINING CASE STUDY

JULY 20, 2017

AUTHOR: SOHIL SHAH
NEW YORK UNIVERSITY

SOHIL SHAH - EMPLOYEE ATTRITION

Value People - Value Resources!!
Employee Attrition (IBM Data)

Business Crux/ Overview of the Problem Statement:

One of the most important resource for successful functioning of any organization or company is the People resource. Hence, losing the right people from the company can be a huge setback. Thus, understanding the factors or reasons for attrition makes it, all the more, necessary for a company or organization. Our client is from the pharmaceutical industry in USA which faces around 14 – 15 % attrition rates on an average. The data we have also validates this fact and we have around 237/1470 = 16.12% which is quite alarming. It is very important for us to thus, identify the key parameters or reasons for the employees leaving the company (voluntarily), in order to, retain them if they are a part of the good performing bracket and hence, Attrition is very important target variable.

Research & Analysis:

The hyperlinked figures are numbers from research websites linked already.

- True positive: **Attrition** people correctly identified as **Attrition**
  TP → Cost of Replacement + Training costs + Performance Hit for 5 months of New recruit by 50 %
  $ 4000 (average) + $ 1200 + 2.5*6503 (average of Monthly Salary → Data) = **- $ 21,457.5**

  Negative as it is cost incurred to the company but it can be reduced/ optimized as we have taken a correct prediction.

- True Negative: **Not Attrition** people correctly identified as **Not Attrition**
  TN → 0 [As the person is still there and we had predicted he will be here it won't cost the company]

- False positive: **Not Attrition** people incorrectly identified as **Attrition**
  FP → Incentive cost + salary hikes + stock options
  15.2% [Average of Percent Salary Hike from data] * (6503*12) [Average annual salary] = **- $ 11861.5**

- False Negative: **Attrition** people correctly identified as **Not Attrition**
  FN → Cost of Replacement + Training costs + Performance Hit for 5 months of New recruit by 50 %
  $ 4000 (average) + $ 1200 + 2.5*6503 (average of Monthly Salary → Data) = **- $ 21,457.5**

SOHIL SHAH - EMPLOYEE ATTRITION

Data Steps:

1. Recode Attrition column 1→ Yes to Attrition; 0 → No to Attrition
2. Converted all the numerical but categorical variables into Categorical in Azure

Model Parameters:

Metric	Logistic Regression	Boosted Decision Trees	Decision Forests	Neural Network	Bayes Point Machine
FP	8	8	4	17	6
FN	48	55	55	39	48
Overall Error	0.127	0.143	0.134	0.127	0.122
Sensitivity	0.342	0.247	0.247	0.466	0.342
F1	0.472	0.364	0.379	**0.548**	0.481
AUC	0.836	0.771	0.807	0.794	0.837

Parameter & Model Selection:

Based on above calculations and parameter values, we see that the client would be impacted maximum by False Negative and to some extent by False positives. We also see that True positives is also incurring similar costs but we can reduce it as we would be predicting the correct attrition outcome. False Positive and False Negatives are sudden to the company or due to wrong promotion/ hikes/ stock – options. Hence, we would like to focus on reducing both **False Negative** and **False Positive** values. The key metric for the model performance evaluation in the current business scenario is **F1 value.**

As we can compare the values above, the best model is **Neural Network** model with **F1 value = 0.472**

Quality of Models

The quality of models is not upto the mark. If we consider the following table calculation as shown below we get to see that even the best model is costing us $ 1 million plus. Also here, we have disregarded the True Positive cost as it is not a sudden expense.

SOHIL SHAH - EMPLOYEE ATTRITION

Model Type	FP	FN	Total Cost
Logistic Regression	8	48	$1,124,851.78
Boosted Decision Trees	8	55	$1,275,054.28
Decision Forests	4	55	$1,227,608.39
Neural Networks	17	39	$1,038,487.52
Bayes Point Machine	6	48	$1,101,128.83

Cost due to FP	$11,861.47
Cost due to FN	$21,457.50

Work Space:

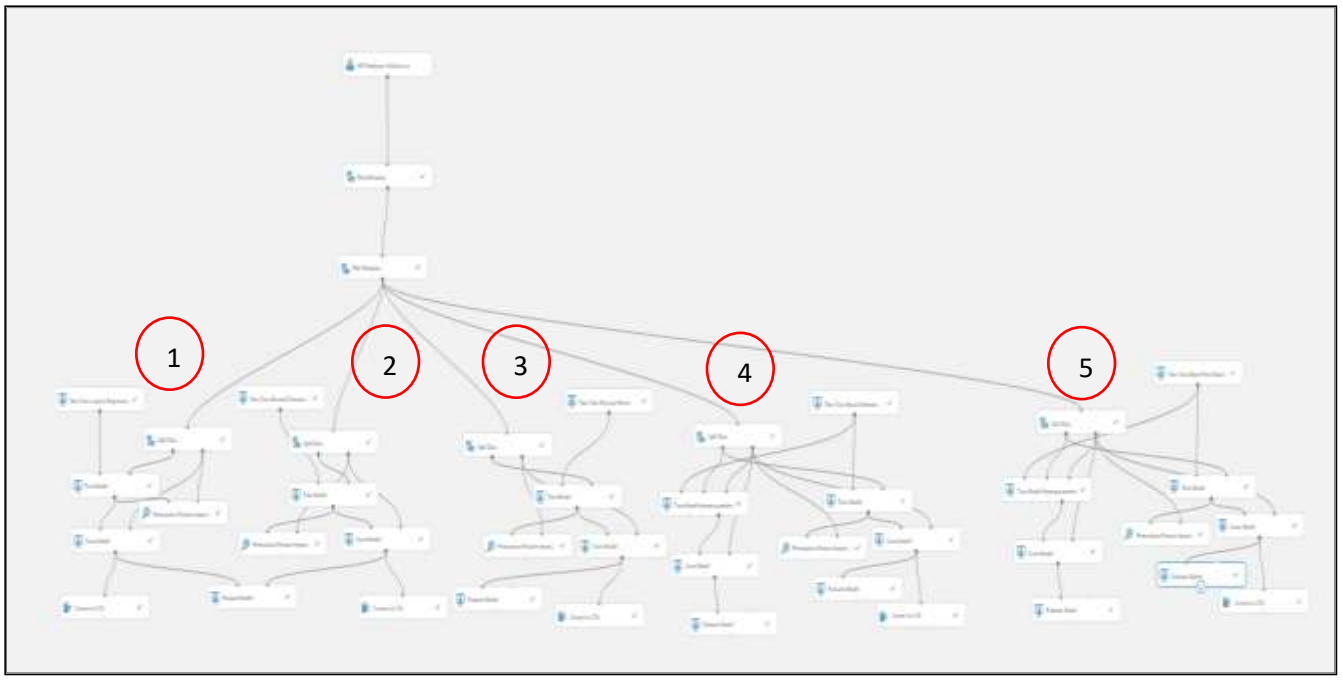


Figure 1: (1) CART, (2) Logistic Regression, (3) Boosted Trees, (3) Decision Forest (4) Neural Networks (5) Bayes Point Machine

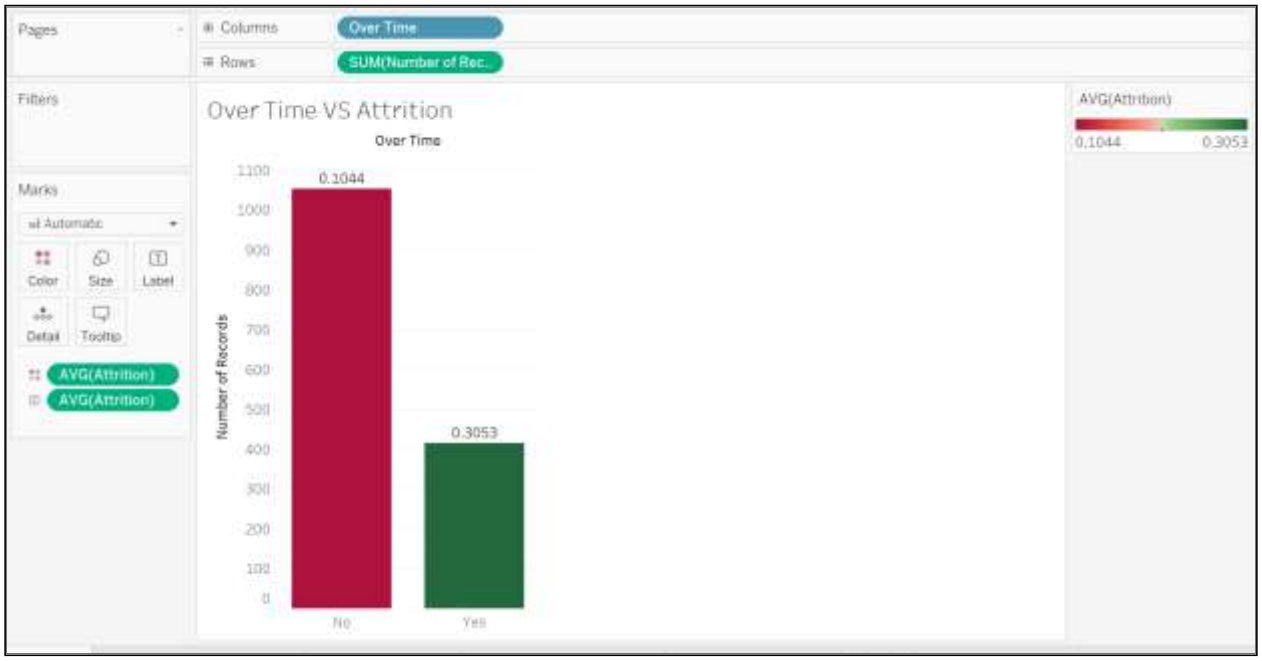
Best Predictors Analysis

From the Neural Network model, our best 5 predictors are:

1. Over Time
2. Stock Option Level
3. Job Role
4. Job Satisfaction
5. Department

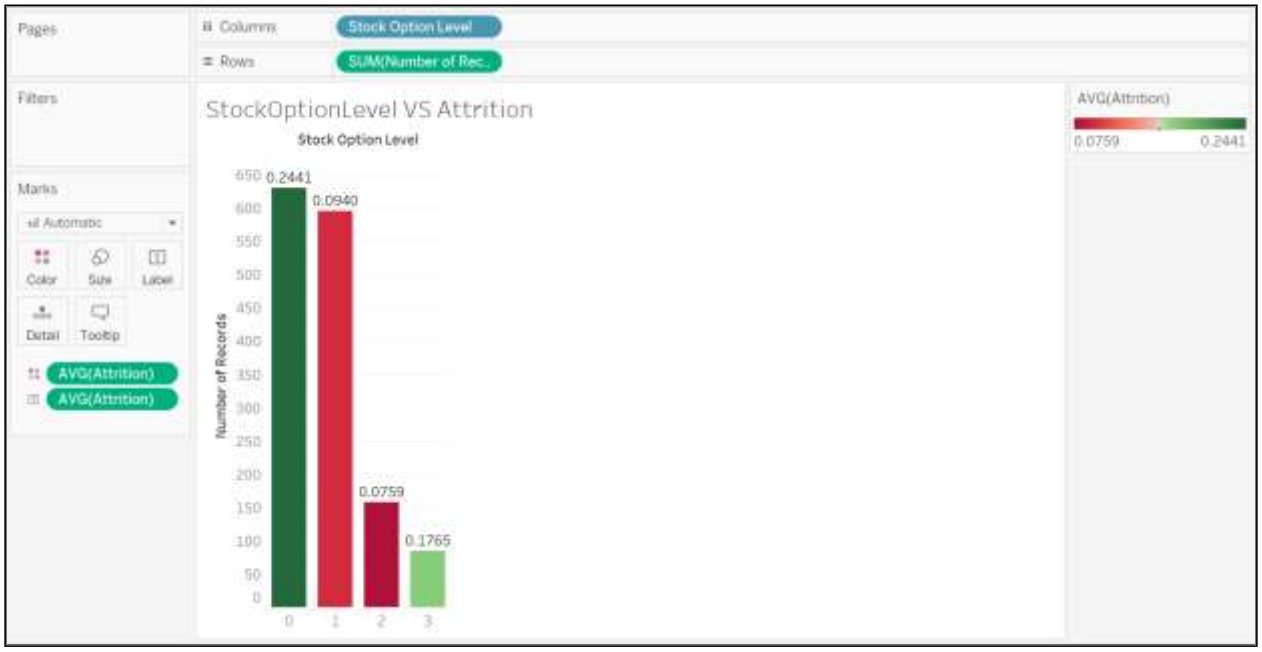
SOHIL SHAH - EMPLOYEE ATTRITION

1. **Impact of Over Time on Attrition**



As we can see above, those people who are doing **overtime** are having **higher attrition** rate of around **30.5 %**

2. **Impact of Stock Option Level on Attrition**



As we can see above, those people who don't have any stocks are more in numbers are also having **higher attrition** rate of around **24.5 %**

SOHIL SHAH - EMPLOYEE ATTRITION

3. Impact of Stock Option Level on Attrition

As we ca

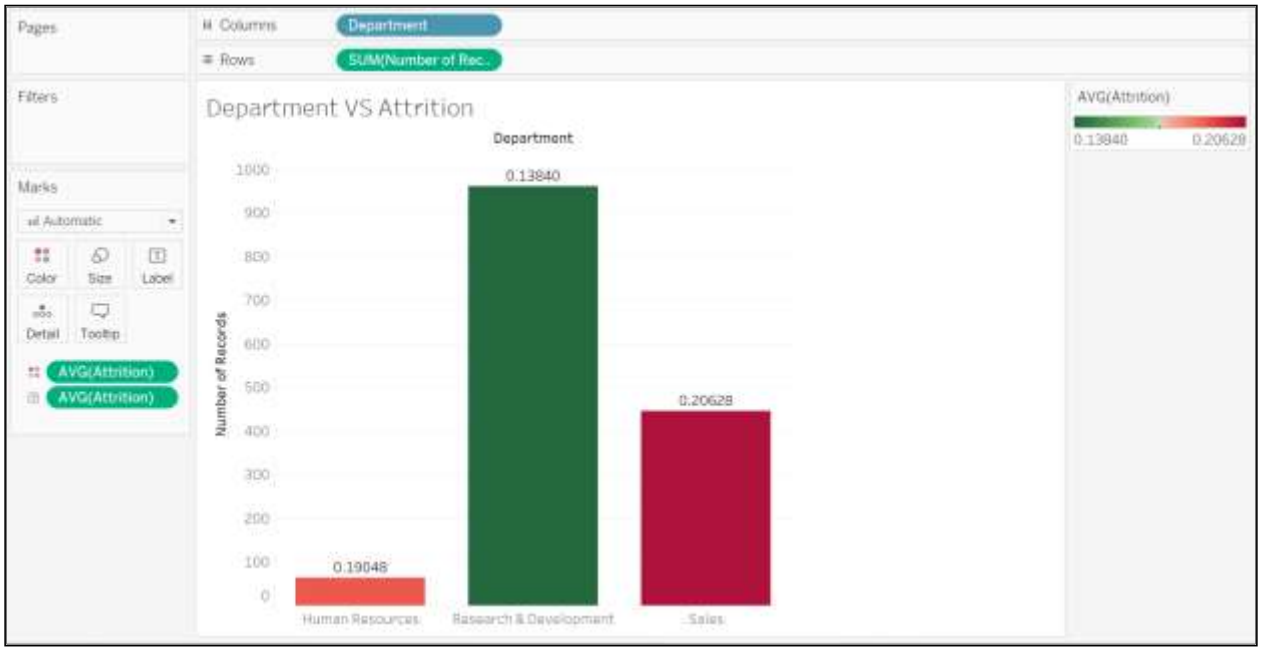
]n see above, employees working as Sales Representative are having **highest attrition** rate of around **39.8 %**

4. Impact of Job Satisfaction on Attrition

As we can see above, employees having **NO Job Satisfaction** are having **highest attrition** rate of around **22.84 %**

SOHIL SHAH - EMPLOYEE ATTRITION

5. **Impact of Department on Attrition**



As we can see above, employees from SALES department are having **highest attrition** rate of around **20.62 %.**