Hi Nic,

Here a quick overview of some SPSS outcomes. Do you have SPSS? Then I can send you some output files.

If you don’t have SPSS let me know and maybe I can do something.

Let me not send you output in detail now, then you can double check. Let me know if you have trouble.

I added a new variable “category”: 1 is Apple, 2 is SW engineer (not apple and not control) 3 = control (not software engineer and not apple).

First:

ANOVA with Post Hoc Tukey: the means between category 1 (Apple) and category 2 (SW) of anger and happiness are significantly different.

However, binomial regression between Apple and SW and same between Apple and Control does not find any solution.

T-test between Apple and not-Apple (SW+Control) finds statistically different means for happiness and anger. Anger has the best significance. It is quite high.

Binomial regression between Apple and not Apple works.

Forward regression (adding variables step by step):

Best initial guess if you don’t know anything (Assume everybody is NOT Apple = baseline to compare):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea,b** | | | | | |
|  | Observed | | Predicted | | |
|  | isAppleEmployee | | Percentage Correct |
|  | 0 | 1 |
| Step 0 | isAppleEmployee | 0 | 933 | 0 | 100.0 |
| 1 | 851 | 0 | .0 |
| Overall Percentage | |  |  | 52.3 |
| a. Constant is included in the model. | | | | | |
| b. The cut value is .500 | | | | | |

Overall percentage is the value to beat…

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Tablea** | | | | | |
|  | Observed | | Predicted | | |
|  | isAppleEmployee | | Percentage Correct |
|  | 0 | 1 |
| Step 1 | isAppleEmployee | 0 | 681 | 252 | 73.0 |
| 1 | 560 | 291 | 34.2 |
| Overall Percentage | |  |  | 54.5 |
| Step 2 | isAppleEmployee | 0 | 672 | 261 | 72.0 |
| 1 | 559 | 292 | 34.3 |
| Overall Percentage | |  |  | 54.0 |
| a. The cut value is .500 | | | | | |

The classification goes down for step 2, which is very unusual.

So prediction about who is NOT apple is quite good, but prediction about who IS apple is poor.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Summary** | | | |
| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
| 1 | 2460.499a | .005 | .007 |
| 2 | 2456.564a | .007 | .010 |
| a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001. | | | |

This is VERY poor. Consistent with overall percentage increase is only 2% or so (from 52.3 to 54%).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | | |
|  | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
| Lower | Upper |
| Step 1a | anger | .019 | .007 | 8.844 | 1 | .003 | 1.020 | 1.007 | 1.033 |
| Constant | -.499 | .145 | 11.844 | 1 | .001 | .607 |  |  |
| Step 2b | anger | .020 | .007 | 9.553 | 1 | .002 | 1.020 | 1.007 | 1.033 |
| fear | .041 | .021 | 3.923 | 1 | .048 | 1.042 | 1.000 | 1.085 |
| Constant | -.894 | .247 | 13.107 | 1 | .000 | .409 |  |  |
| a. Variable(s) entered on step 1: anger. | | | | | | | | | |
| b. Variable(s) entered on step 2: fear. | | | | | | | | | |

Fear is barely significant, better to look only at anger and to keep step 1 only.

Anger has high significance. Change for a fluke is only 3 in 1000.

Odds ratio (chance appl/chance not apple) = exp (B) is 1.02. Its CI is [1.007,1.033] So very marginal. Still this is enough to pick up the non-Apple ones quite nicely.

The effect size is so small that you have to have a large sample to see the difference. The separate samples SW and control are not large enough (probably, SPSS does not say why it cannot find a solution).

If you can get 62% instead of 54%, then that would be an incredible improvement.

Add as independents anger \* fear, anger \* happiness, etc. Those are moderators. https://en.wikipedia.org/wiki/Moderation\_(statistics)

Then Anger\* fear becomes a predictor with about the same quality of the prediction (forward regression), instead of Anger only.

There are also other solutions possible. When you do backwards, you can find a solution with similar quality and a lot of independents.

It seems like there are a number of possibilities for the predictors that have comparable quality of classification. There is probably not a solution that is a lot better, otherwise the system would have seen it.