Hi Nic,

You wrote:

*I think the result looks similar to the scikitlearn tools that I am using. I think that it indicates that there is a relationship between pictures and companies, but that may not be due to emotions hence we get a weak unclear but real correlation.*

You wrote that you got 62%. Was that overall percentage? I would have thought that random forest, which combines several strategies, would do better than binomial logistic regression

Your research is limited to emotions. In a formal way (which should be followed in your thesis), you have the following (suggested) research questions:

1. Is it possible to extract in a reliable way the emotional state of people from their self-provided and unstandardized pictures on Linkedin?
2. Are there statistical differences between the emotional states of software engineers working at Apple and a people not working at Apple?
3. Is the difference enough to design a classifier that can decide if between random working people if someone works at Apple or not?
4. Is this classification a predictor for the chance that somebody would be hired at Apple

Your work up to now:

1. Yes, definitely
2. Yes, they are not large but they are there. T-test/Anova show the difference both for Apple and non-Apple SW. And there were differences between Apple and the whole group of not-Apple. This is for mean scores. Maybe I should also run some tests for the distributions. I don’t think this one is sufficiently researched, but it is not an area that you are interested in ☹. I will try a little more when I have time.
3. This requires, at least in standard statistics, less noise in the data and better significance than 2). What I found (but you have to confirm) is that there are several ways to obtain a marginal classification. I tried binomial logistic regression with the following independent variables:
   1. Scores on emotions
   2. Scores on anger \* all other variables (mediators in statistical terms(
   3. Scores on variables and on Variable ^ 2 (variable times itself, quadratic). (not before reported to you)

All the results are comparable. The classifier is reasonable to see if people do NOT work at Apple. It is not very good to see if they work at Apple.

Please look at <https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php> under the heading classification. That is general, has nothing to do with regression, and is also relevant for you. Look at specificity, sensitivity and positive and negative predictive value. The negative predictive value is quite good in your work, the positive predictive value is not.

1. Not attempted, not a good research question given the tools that you have.

You wrote:

“it indicates that there is a relationship between pictures and companies”

I disagree for two reasons:

1. You cannot know. It has not been tested. It is possible, but there is nothing in your data that gives an indication. I found some influence from Anger. Maybe this points to dominance. Maybe dominant guys wear more ties in their pictures. Maybe wearing a tie in a picture says something. But this is not an indication, it is at most a suggestion for further research.
2. “Is there a relationship between features on pictures and companies?” is not a good research question, because it requires that an infinite number of features to be tested. Does it depend on the color of the shirt? If there is a painting in the background? How tidy or messy the background is? …… One needs to have a finite list of things to look for. In explorative research people start on a hunch. In order to maximize the chance to find something, usually people start with generating ideas from a reasoning starting with known data and known theories

The last thing is safer but more boring. Therefore, this is usually suggested by supervisors for Master’s and Ph.D. work. Higher chance to find something. People with a job might want to try the exploration.

I think it would be fair to suggest that with a weak correlation like you found, there is a chance for a more sensitive test if you take thousands of pictures of a person, a thing also called a video.

Please have a look at

https://hal.inria.fr/hal-01668386/document

(and not as good) <https://www.researchgate.net/profile/Semila_Fernandes/publication/325427595_Automation_in_Recruitment_A_New_Frontier/links/5e943eaba6fdcca78911f57a/Automation-in-Recruitment-A-New-Frontier.pdf>

Your work fits in the attempts to automate the HR recruitment process via AI. One can look at video (expensive/time consuming for applicant) and automated text analysis. You tried to find something that works from pictures.

I read somewhere but did not find it back that there are proprietary tools to analyze videos from applicants and tell if they should be hired. As they are proprietary and with secret sauce, you have no idea what the reliability, bias, etc. is.

Leslie

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.