Report on the Flourishing Data Set

Introduction to Machine Learning

A.I. Booster

Joaquin ARIAS

13-10-2024

INTRODUCTION

This report aims to analyze and, if possible, draw conclusions based on the dataset named Flourishing. This data focus on measuring various aspects related to the well being of employees both in their professional and personal ambits of life. The data has been collected from 248 individuals and will be helpful to study their work and life overall satisfaction and which variables influence in their prosperity.

The data is composed of many variables but the main ones and the ones who will help us measure the dimensions of the satisfaction and well being of the 248 employees are the **pro_quant** (understood as professional flourishing), **priv_quant** (private life flourishing) and **positivity**.

Flow variable is also understood as a measure of the overall mind state of individuals and their ability to thrive under difficult situations or accomplish complex tasks.

In this dataset, we will consider quantitative variables age, pro_quant, priv_quant and flow. And our categorical variables are going to be sex, famstatus, education, pro_cat, priv_cat and positivity. The ID variable is a number that represent each individual and is given to every participant.

IMPORTING AND PREPARING OUR DATA SET

1. The first step is always importing all the libraries and important packages we are going to use through out the proyect. In this case and at the beginning it was the following packages we imported:
    a. Pandas
    b. Numpy
    c. Math
    d. Matplotlib
    e. Seaborn
    f. Os
    g. Unittest
    h. Warnings
2. Our second step was to ensure that the working directory is correct so that our notebook does not use absolute paths. We completed this task using OS and changing the directory.
3. Then we import the complete dataset and transform it into a data frame. To complete this task we use the pandas function read excel. We only needed to specify the file name because it doesn't have special separators or encoding. To

get the ID out of the way and to give it a use we transformed it to the index and got the following table:

ID	age	education	sex	famstatus	pro_cat	pro_quant	priv_cat	priv_quant	positivity	flow
344	72	6	2	2	3	82	3	83	4.142857	44
317	45	6	2	2	3	81	3	83	4.000000	44
444	43	5	1	1	3	79	3	78	5.000000	47
270	60	2	2	3	3	79	3	82	4.666667	42
52	49	2	2	2	3	77	3	73	2.285714	33

4. Education, sex, fam status, pro cat, priv cat and positivity all have a number that represents a specific category. To be able to visualize the categories and not only numbers we transform them into the specific labe creating dictionaries that contain the equivalence and using the function map to replace them.
5. In the special case of positivitiy, we want to maintain the float number but also have the category. So we create a function that classifies them. This is the final result of our table:

ID	age	education	sex	famstatus	pro_cat	pro_quant	priv_cat	priv_quant	positivity	flow	pos_cat
344	72	Has completed 5 years after high school degree...	Female	Separate/Divorced	Flourishing	82	Flourishing	83	4.142857	44	emotionally flourishing
317	45	Has completed 5 years after high school degree...	Female	Separate/Divorced	Flourishing	81	Flourishing	83	4.000000	44	emotionally flourishing
444	43	Has completed up to 4 years after high school ...	Male	Single	Flourishing	79	Flourishing	78	5.000000	47	emotionally flourishing
270	60	Has completed only a high school degree	Female	Couple	Flourishing	79	Flourishing	82	4.666667	42	emotionally flourishing
52	49	Has completed only a high school degree	Female	Separate/Divorced	Flourishing	77	Flourishing	73	2.285714	33	moderately emootionally healthy

UNIVARIATE ANALYSIS

Summary of Categorical Variables:

- **Sex**: The majority of the participants are female. Making up about 61% of the interviewed individuals. Males only sum to 39%.
- **Family status**: Most of the participants have a partner while single and divorced have almost the same amount near 15%.
- **Education**: A large portion of our population is well educated. Almost half of it has completed more or 5 years after secondary education.
- **Professional Flourishing Category**: The majority of people are considered mentally healthy. Its good to see that 6% more of people are flourishing than languishing.
- **Private Flourishing Category**: In this case, most participants are moderately mentally healthy too. But a bigger portion of them are flourishin. And only 6% are languishing. This could suggest that most people in our dataset have a relatively stable private life.

Summary of Quantitative Variables:

- **Age:** The average age of participants is around 42 years and we have a wide age range.
- **Flow:** The mean is 33, this indicates that participants expecience a moderate level of flow, understanding these as a level of deep focus, inmersion and ability to accomplish tasks.
- **Positivity:** There's much spread of values in this variable. These translates the subjective nature of measuring positivity.

CONCLUSIONS:

- As we have said before, this dataset is primarily focused on the well being of the individuals in their individual and work life, measuring their flow and positivity.
- This includes demographic variables such as age, gender and education.

BIVARIATE ANALYSIS

Both Categorical

Contingency Tables, Relationship with Chi-Square Test and Strength with Cramer's V:

- **Professional Flourishing vs Positivity Category:**
  - **Chi-square Statistic:** 93.43
  - **P-value:** 5.88 e-18
  - **Cramer's V:** 0.434
  - **Interpretation:** The low p-value indicates that there is a highly significant association between these two variables and the Cramer's V suggest a moderate to strong association. This is the strongest relation found in the bivariate analysis and it suggests that there is an important link between how a person thrives professionally and their emotional well being and positivity on life.
  - **Insight:** Since professional well being contributes to positivity, focusing on work life balance and integration will ensure that professional achievements do not come at the cost of emotional well being.
    When the company struggles with low performing employees, implementing career coaching focused on positivity would help improve in their performance.
- **Education vs Positivity Category:**
  - **Chi-square Statistic:** 26.87
  - **P-value:** 0.0298
  - **Cramer's V:** 0.19
  - **Interpretation:** The p-value indicates that there is a significant association between education and positivity if we define alpha 0.05. The association is weak or moderate between the two variables, so, the different levels of education are likely associated with positivity.
  - **Insight:** Giving employees the opportunity to enroll in educational programs will boost their positivity and this will lead to better performance in their professional and private life.
- **Sex vs Family Status:**
  - **Chi-square Statistic:** 8.26
  - **P-value:** 0.0161
  - **Cramer's V:** 0.18
  - **Interpretation:** The p-value indicates that there is a significant association between education and positivity if we define alpha 0.05. The association is weak or moderate between the two variables, so, there might be a relation between being a male or female and your family status.

CONCLUSIONS:

- The relationship between professional flourish and positivity is the most significant and strong. This confirms that there is a connection between professional success and having a positive outlook on life.
- Education vs. positivity and sex vs. female status are significant but their strength is much weaker.
- The remaining comparisons have no significant relationship.
- **Both Categorical**

Both Quantitative

Correlation Table, Relationship with Correlation Test and R Value:

- **Professionally Flourishing vs Privately Flourishing:**
  - **R-value:** 0.6357
  - **P-value:** 1.77 e-29
  - **Cramer's V:** 0.434
  - **Interpretation:** Strong positive relationship between these two variables. Meaning that those who flourish personally are also likely to flourish professionally. This is the most significant relationship, due to both the strength of the correlation and the extremely low p-value.
  - **Insight:** Career satisfaction programs that address both personal and professional goals and how they connect can benefit the employees well being and their work performance.
- **Professionally Flourishing vs Flow:**
  - **R-value:** 0.5362
  - **P-value:** 7.28 e-20
  - **Interpretation:** This suggest that the higher is someone professional flourishing, the higher their flow.
  - **Insight:** Creating and fomenting work enviroments that lead to flow states, such as minimizing distractions, allowing autonomy and setting clear objectives can help employees thrive professionally.
- **Privately Flourishing vs Flow:**
  - **R-value:** 0.4049
  - **P-value:** 1.92 e-11
  - **Interpretation:** This suggest that the higher is someone privately flourishing, the higher their flow. But it must be taken into account that this relationship is weaker than the professional flourishing and the flow.
  - **Insight:** Incorporating personal development and mindfulness programs and practices can hel individuals vecome more present and engaged in their private lives, which may lead to longer periods of flow.
- **Age vs Flow:**
  - **R-value:** 0.3680
  - **P-value:** 2.27 e-09

    - **Interpretation:** This indicates that as age increseas, flow tends to increase as well. This is significant but weaker than the associations explained above.
- **Age vs Professional Flourishing:**
    - **R-value:** 0.2296
    - **P-value:** 0.00027
    - **Interpretation:** This indicates that as age increases, the individual will experience a slight flourish in their professional life.
- **Age vs Privately Flourishing:**
    - **R-value:** 0.1516
    - **P-value:** 0.0169
    - **Interpretation:** There is a very weak but significant relationship between age and privately flourishing. As the individual gets older, they might experience a slight increase in their private life flourish.
    - **Insight:** Companies can benefit from age specific programs that help spread older individuals strengths, such as mentoring roles.

CONCLUSIONS:

- The strongest and most significant association is between professional flourishing and privately flourishing. This is a good indicator that highlights the importance of having a professional well being to achieve a private well being.
- Professionally flourishing and privately flourishin relate to flow in a positive wat too. This means that flourishing in any aspect relates to an increase in flow.
- Also, age and flourishing measures present a weak, bus existent relation. Indicating that as someone gets older, they can see a slight increase in their both private and professional well being.

Quantitative Vs. Categorical

Anova Table and Significance of the difference between groups:

In this analysis it is important to choose which outputs have interesting information to consider because there is a lot of noise.

- Positivity Category Vs. Private and Professional Flourishing
  - **F-Value:** 48.94, 43.38
  - **P-Value:** 8.46 e-25, 1.66 e-22
  - **Interpretation:** This relationship shows that positivity is strongly associated with private and professional life flourishing. Higher levels in positivity lead to a better personal and professional well-being.
  - **Insight:** This finding suggests that the company could promote positivity and that those activities could lead to an improvement in the private aspects of their employees such as satisfaction, selfsteem, relationships and overall happiness but also on the professional side as their career development and performance.
- Positivity Category Vs. Flow
  - **F-Value:** 20.36
  - **P-Value:** 8.24 e-12
  - **Interpretation:** This relation show that people who are positive most of the time are also more likely to engage deeply in their activities and get things done.
  - **Insight:** This confirms that the company should invest in positivity programs to boost their employees abilities to work under pressure and achieve great results. This will also lead to an increase in productivity.
- Education Vs. Flow
  - **F-Value:** 3.75
  - **P-Value:** 0.00275
  - **Interpretation:** This finding indicates that education level influences in the ability to experience flow. The greater the education, the more often the experience of flow.
  - **Insight:** Understanding this insight also provides important information for the company. If they invest in educational and training programs, making special emphasizes in continuous learning and skill building experiences, it will help individuals reach a flow status that will boost their productivity and work experiences.

CONCLUSIONS:

- Positivity is one of the most important variables to achieve private and professional flourish. Its important to invest in it to see results of well being in the employees.
- Positivity also helps boost flow performance. It is vital to achieve a good mind state that can help achieve work goals and objectives.

- Education is also important to get to a flow state, the more learning, the better the performance of the employees.

LINEAR REGRESSION

SIMPLE LINEAR REGRESSION – USING POSITIVITY TO PREDICT FLOW

Given the results obtained, we crafted a simple linear regression that can use positivity as the predictor (X) to predict the flow as the outcome (Y). The results explained below:

- **Coeffcient for Positivity:** 3.612
    - Interpretarion: For each additional unit of positivity the flow score is expected to increase by 3.612. As positivity increases so does the individuals flow.
- **R2 (Coefficient of Determination):** 0.193
    - Interpretation: 19.3% of the variability in flow score can be explained by the individuals positivity score. This shows us that 80.7% of the variability in flow is explained by other factors not included in this model.
- **Root Mean Squared Error (RMSE):** 5.915
    - Interpretation: The predictions made by the model deviate form the actual flow values in 5.915 units of flow. This can be considered a low value and a good fit.

Insight: Even thought increasing positivity can help improve flow, efforts that only focus on this variable might not be enough to get a significan improvement in flow experiences.

We also crafted two simple linear regressions trying to use age and professional score as predictors to determine flow, none of this regressions got good indicators as a result and we discarded both.

MULTIPLE LINEAR REGRESSION – USING POSITIVITY TO PREDICT FLOW

To start creating a multiple linear regression model we use as input all of the variables given, quantitative (age, pro_quant, priv_quant and flow) and categorical (sex, famstatus, education, pro_cat, priv_cat and positivity). The results of the model are the following:

- **R2 (Coefficient of Determination):** 0.423
    - Interpretation: 42.3% of the variability in the flow score can be explained by all the predictors given in the data set.
- **Adjusted R2:** 0.380
    - Interpretation: When the number of predictors is taken into account, only 38# of the variability is exp´lained by this model. This makes us notice that some predictors are not adding explanatory power.

Key Predictors:

In the developed model, the following are the best predictors:

- **Emotionally Flourishing in Positivity:**
  Coefficient = 4.641
    - Interpretation: Individuals categorized as emotionally flourishing in the positivity variable experience a 4.641 unit increase in their flow variable.
- **Professional Flourishing:**
  Coefficient = 0.224
    - Interpretation: For every unit increase in the professional flourishing variable, the flow score increases by 0.224 units if we hold all the other factors constant.
      This aligns with the previous findings that work related well being is a driver of flow and performance.
- **Age:**
  Coefficient = 0.167
    - Interpretation: For every year that increases in the age of the individual, the flow score will increase by 0.167 units. This indicates that older people are better enabled to achieve flow state. We can assume that this is given due to a greater life experience and better understanding on work activities.
- **Male in Sex:**
  Coefficient = -1.512
    - Interpretation: Surprisingly, being a male is associated with a 1.512 unit decrease in flow score when being compared to females. This may be due to having an unbalanced set where the majority of participants are female.
- **Separate / Divorced in Fam Status:**
  Coefficient = -1.843

- Interpretation: Possibly due to the stability provided by being in a relationship or single, we see a 1.843 unit decrease in flow when people are separated.

Unexpected and Weaker Relationships:

- **Positivity:**
  Coefficient = -0.078
  - Interpretation: Surprising result given the previous findings in bivariate analysis. I suggests that when controlling the other variables described above, positivity may not have a strong impact on flow as we thought.
- **Private Flourishing:**
  Coefficient = 0.027
  - Interpretation: Each unit increase in private flourishing leads to a small effect increase in flow of 0.027. This is not a primary driver but has a positive effect.
- **Has completed up to 2 years after high scool in Education Levels:**
  Coefficient = 2.726
  - Interpretation: Individuals with a higher education have a bigger increase in their flow score. We can infer that studying helps to get a more structured approach to get tasks done.

A new model was generated excluding positivity and priv_quant. The model did not return better results and is not worth analyzing profoundly. This second model focuses a bit more on professional aspects and the demographic variables. If wanted by the company they can use either without sacrificing much explanatory power.

- **R2 (Coefficient of Determination):** 0.422
  - Interpretation: 42.3% of the variability in the flow score can be explained by all the predictors given in the data set.
- **Adjusted R2:** 0.384

CONCLUSION: Given that we do not obtain better results with the new model, we can conclude that both models show similar predictive capabilities and that the predictors given by the data set can predict a little bit more than 42% of the flow score.

LOGISTIC REGRESSION

We are going to work on a logistical regression model that has the aim to predict the positivity but transforming it to a binary value where 1 indicates a high positivity (above the median) and 0 indicates a low positivity. Our predictors are going to be quantitative (pro_quant, priv_quant, flow and age) and categorical (sex, famstatus, education and pro_cat as a dummy).

The model predicts positivity because it fits well and can be transformed into a binary variable. Flow is a continuous variable and is better handled with continuous outcomes as the linear regression model we made before.

Model Performance:

- **Accuracy:** 0.740
    - Interpretation: This indicates a good level of prediction power and suggests that the chosen predictors are useful for predicting between the two levels of positivity.
- **Sensitivity:** 0.760
    - Interpretation: This model correctly identifies 76% of individuals with high positivity. The model is effective when capturing true positive cases.
- **Specificity:** 0.720
    - Interpretation: Our model is slightly worse at recognizing individuals in the low positivity category.
- **Precision:** 0.731
    - Interpretation: The model is reliable 73% of the time.
- **F Score:** 0.745
    - Interpretation: When combining precision and recall, the precision of our model is almost 75%. This confirms that the model maintains a good balance between identifying true positives and avoiding false positives.

CONCLUSION:

This logistic regression model offers a valuable insight into how professional, private flourishing and demographic factors contribute to an individual's likelihood of having a high positivity.

This model could be used as a tool for well being assessment in a working place. This can help us provide support to those with low positivity and identify people with high positivity while giving both the support they need to help them flourish in their work place and to achieve work place.