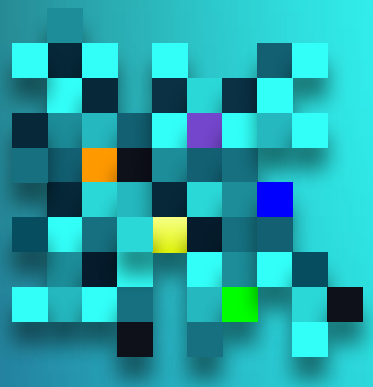
loxz digital



The Loxz Digital MLR Pearson Correlations Heatmap Methodology

2022 Pearson Methodology Report

By Yiming Zhang
Lead Data Scientist, Loxz Digital

Q1 2022
Methodology Report

loxz digital

Abstract

In this methodology paper, we will dive into how to generate rich insights from a survey respondent to compare their specific score with a unique user group that represents the respondent's profile. This short report will attempt to introduce the definition of Pearson Correlation, how to implement it with a heatmap and how to add unique filters to generate high-value insights for your Machine Learning exploration and experiments. Also, the limitations of the current version and future considerations will be discussed.

Motivation

The MLR score that our Machine Learning Diagnostic Assessment represents is a reliable indicator of the ML maturity of an entity in academic institutions (i.e. a student), or an organization compelled to offer and procure machine learning services or build ML models in-house (companies that build ML solutions). Our goal is to provide feedback on how those question features are correlated for a certain group of respondents or survey takers. For example, a freshman student might be interested in how his strengths or weaknesses are compared to his peers or a company in the healthcare industry would like to know how the aggregate scores of this industry are correlated. For this purpose, we will provide a client with a filtered heatmap that visualizes the correlation of the features for a group that represents him. The filters could be the grade of a student, the industry of a company, or the number of data scientists that an organization employs. Filters provide a way to focus on certain characteristics that are important for the respondent.

Pearson Correlation and Correlation Matrix

Correlation is a term used to represent the statistical measure of the linear relationship between two variables. In this case they are the questions IDs. It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find a correlation between all of these variables and store them using the appropriate data structure, the matrix data structure is used. Such a matrix is called a **correlation matrix.**

The Pearson correlation coefficient r between two variables x and y can be calculated using the following formula. $\bar{x}$ is the mean value of x and $\bar{y}$ is the mean value of y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The value of the correlation coefficient can take any value from -1 to 1.

- If the value is 1, it is said to be a positive correlation between two variables. This means that when one variable increases, the other variable also increases.
- If the value is -1, it is said to be a negative correlation between the two variables. This means that when one variable increases, the other variable decreases.
- If the value is 0, there is no correlation between the two variables. This means that the variables change in a random manner with respect to each other.

A correlation matrix denotes the correlation coefficients between variables at the same time. It is computed by calculating the Pearson Coefficients between any two features in the filtered survey data.

loxz digital

What does the Loxz Digital ML Survey Heatmap represent or explain?

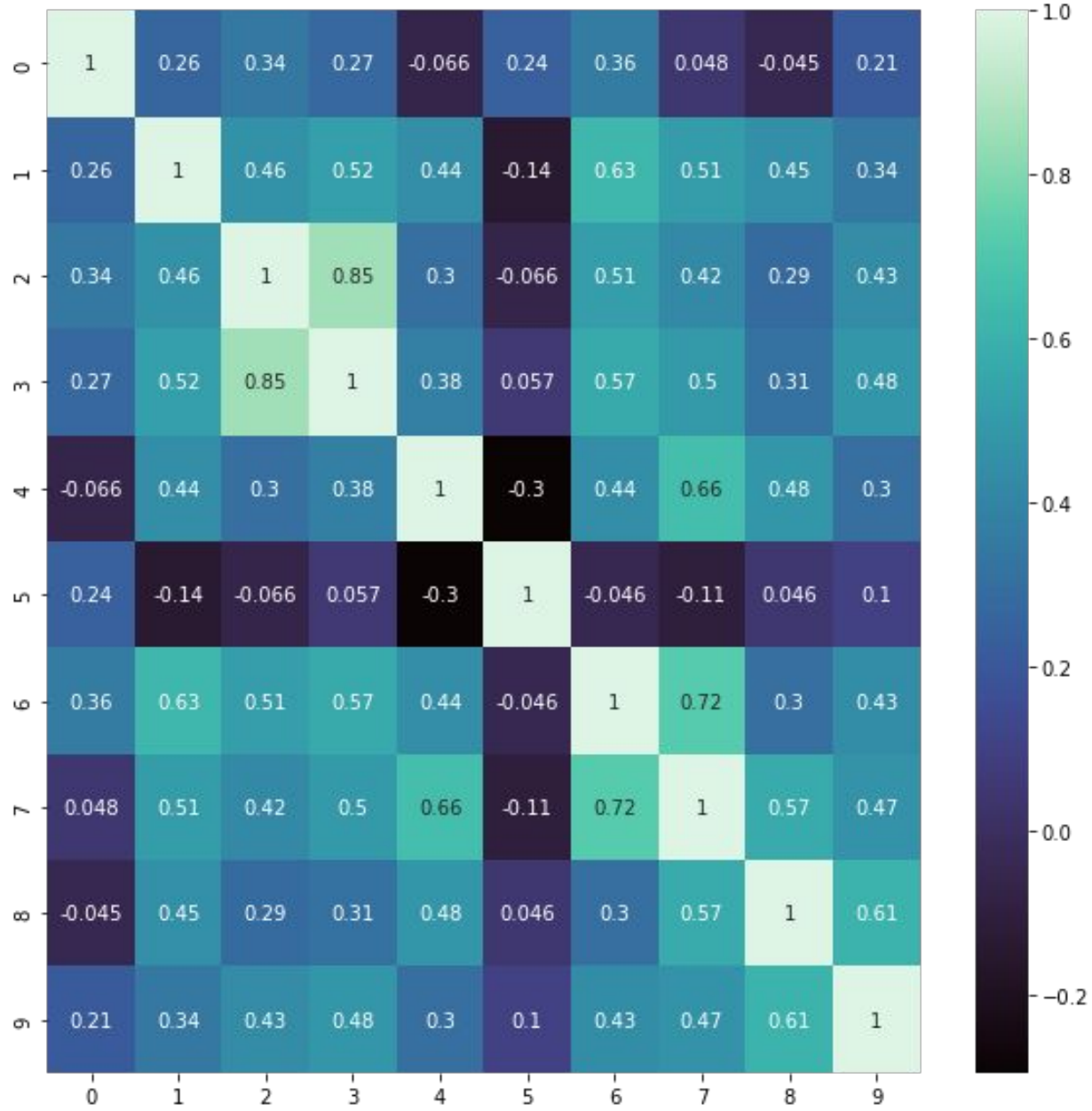
In order to help the audience or respondents realize how their ML Survey score are correlated in each question, we collected a user-question matrix M, where each entry M_{ij} represents the score that user i got by answering question j.

We then compute the correlation matrix along all columns of the user-question matrix, so we will have a correlation matrix of the scoring of each question. We then visualize the correlation matrix using a heatmap.

If the correlation value at entry M_{ij} of the heatmap is positive (a warmer hue) then it means there is a positive linear correlation between the scoring of question i and question j. On the other side, a negative value (a colder hue) means there is a negative correlation between the scoring of question i and question j. The larger the magnitude of the value is, the stronger the correlation is, and no correlation with the value being 0. We provide some examples, later in this report.

Q1 2022
Methodology Report

loxz digital



loxz digital

Adding filters to the heatmap

As explained previously, upon submission of the ML Survey we will provide filtered heatmaps as a part of the insights. The filters are mainly created in two ways: filtering by user group or filtering by question group (subcategories). We can also customize the filters to desired attributes.

For the user group filtering, we take a subset of the rows in the user-question matrix as we filter users by their attributes. For the question group filtering, we look for a subset of the columns in the user-question matrix according to the subcategories. In both the MLR assessment for the academic industry and the organizational MLR, we have 5 different sub-categorical scores: For students, these sub-categories include:

- Data Preparation
- Modeling
- Career Trajectory
- ML Aptitude
- Business Value

For organizations, the sub-categories in addition to the overall MLR score are:

- Data Preparation
- Model Development
- Model Deployment
- Model Monitoring
- Business Value

Constraints and Limitations

The current design of heatmap has two constraints and limitations. One, when filtering for a specific group of users, for example, the size of the Data Science/Machine Learning team of a company or a specific research area for a student, the available observations are limited in quantity at this moment. Calculating the Pearson Correlation with insufficient data will not give a proper representation of how the scores are correlated for a certain group thus the filtering of the heatmap would need enough data to operate. That is why we will introduce the Pearson Correlation Heatmap at the end of Q2 2022. Second, our future generation of the ML Survey will include more features than just numerical scores. Thus, the current numerical-based Pearson Correlation will not apply to other types of features.

Future Considerations

There are mainly three areas for improvement. First, with our steady monthly respondent growth, a user-clustering method will be implemented for the industry version of the ML Survey in a quarter and two quarters for the student version. This will help us build better filtering by exploring potential user groups. Second, with the evolution of the future version of Survey, we will also include a feature importance plot for a user group for better explainability. Third, we will add a comparison-on-click functionality that once the user hits a grid at position (i, j) on the heatmap, a small pop-up with more information on his scores of question i and question j and the score distribution in his group.

loxz digital

Conclusion

We plot a Pearson Correlation heatmap to visualize how those question features are correlated for a certain group or sub-group that represents the survey respondent. User groups are partitioned by some categorical features such as the grade of a student in the academic version or the company size in the industry version. The heatmap will accentuate a strong correlation between features with a lighter hue with warm indicating positive correlation and cold indicating negative correlation so the reader will get instant feedback on how his result compares to his peers.

We strongly believe that the Pearson correlation coefficient Heatmap, will evolve but more importantly build strategic value for the organization and academic institutions that host the survey and where data in aggregate can be accessed to tweak academic curriculums for one and training purposes for organizations. Further, we believe that with the sub-scoring clustering in place, we believe that human resource organizations will find tremendous value in time savings when accessing the sub-scores of certain candidates.

Furthermore, it's conducive to point out that when some candidates have a positive correlation, (say seniors) and others have a negative correlation, (say freshmen) the academic institution can decipher these scores and prepare students better for a much steeper career trajectory, or determine if some ML resources are becoming obsolete. Finally for organizations, the deep insights provided in a Pearson Correlation Heatmap, will genuinely uncover positive or negative correlations around sub-categories. For example, some industries like Software and Technology may find a positive correlation among questions related to Model Monitoring while other industries such as Energy may find a negative correlation in such a sub-category. It's exciting to think about such insights our Pearson Correlation Heatmap will provide and the value it can bring to both your academic institution for students as well as your organizational development as the Machine Learning industry evolves.