All just Greedy for Money? - An Prediction and Analysis of the interplay between Graduate Statistics and Salary Prospects in the STEM Field

Abdallah Abdul-Latif [*1] **Lisa-Maria Fritsch** [*2] **Paul Kaifler** [*3] **Maximilian Schnitt** [*4]

Abstract

This paper examines whether the expected salary plays a role in which subject graduates choose to study. We have limited ourselves to the STEM (science, technology, engineering and maths) field. We analyse the influence of the three parameters: number of Graduates, of Students and Expected Salary, have on each other. Therefore we use the Vector autoregressive model and a PACF. This showed that the expected salary has an or has no? influence on how many school leavers decide to study in the STEM field. We were also able to make the prediction that the expected salary and the number of students in the next two years will be similar to the current situation, but slightly reduced. If the salary is increased or decreased in the analysis, it can be seen that the number of students goes up or down accordingly.(or else?)

1. Introduction

The trend in recent years shows that more and more pupils are taking their high-school diploma and then going to university. Many assume that one of the reasons for this is better salary prospects. We are interested in whether this is actually true. To investigate this, we used three data sets. The first data set is from the Statistics Bureau of the University of Tübingen. This can be used to determine how many students there are in which semester and in which subject in Tübingen. We limit ourselves to the STEM field because this area promises a high and growing salary. We also use the data from Tübingen because the STEM field is well represented here. This university can therefore be used as a representative. However, this can lead to a bias in our prediction. We obtained our second data set from the Statistisches Landesamt Baden-Württemberg. It contains the number of high-school graduates from Baden-Württemberg for the past years. We restrict ourselves to Baden-Württemberg, as it can be assumed that most students in Tübingen come from Baden-Württemberg. As this does not apply to all students, there is also a bias. The last data set we need was obtained from the Deutsches Statistisches Bundesamt. It contains the salaries of different industrial sectors in Germany. Here we have used the sectors that are relevant for the STEM field. We extracted the relevant data from the three datasets and analysed them with the vector autoregressive model and a PACF (Section 2). This results in a model that can be used to predict how many students there will be in the STEM field in the next two years and how the expected salary will develop. We can also analyse the interplay of the three parameters and predict what will happen if we modify the salary. (Section 3)

2. Data and Methods

In our research we used three datasets, which are divided as follows: The first dataset is from the Statistics Bureau of theUniversity of Tübingen., which contains the total amount of students from the winter semester 2005/2006 to the winter semester 2023/2024. The data was collected from the enrollments of students, which registered for a study in Tübingen. The second dataset is from the Statistisches Landesamt Baden-Württemberg which contains data about the high-school graduates from 1970 to 2023. Our third and last dataset is from the Deutsches Statistisches Bundesamt and contains the salaries of the large industrial sectors in Germany from 2005 to 2021. One main problem with the mapping between university subjects and actual jobs is that there exists no perfect one-to-one mapping. We also removed the inflation from the salaries to get the real wage, because the inflation could cause a false pattern detection in our autoregressive model. To reduce the amount of bias introduced in this step, we also used the Inflation data from the Deutsches Statistisches Bundesamt over the years.

[*]Equal contribution [1]Matrikelnummer 5977981, abdallah.abdul-latif@student.uni-tuebingen.de, MSc Computer Science [2]Matrikelnummer 4189024, lisa-maria.fritsch@student.uni-tuebingen.de, MSc Computer Science [3]Matrikelnummer 12345678, first.last@student.uni-tuebingen.de, MSc Media Informatics [4]Matrikelnummer 6040570, maximilian.schnitt@student.uni-tuebingen.de, BSc Computer Science.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023. Copyright 2023 by the author(s).

To get the most of this model, we wrote a parser, which parses text data generated from a PDF. This parser basically uses regular expressions to extract the amount of students per years divided into the semesters and created a new dataset with the usage of pandas.

Our first steps involved some feauture engineering on the three different datasets to get a better understanding of the data and also detect patterns in the data like trends and distributions over time. After we finished the basic feauture engineering step we started the main analysis of the Vector Autoregressive Model (VAR). We've chosen the vector autoregressive model, because we have multivariate data and we assumes that there is a correlation between the given time series. A VAR model don't predict a single time series but one for each input series. Those modles have one hyperparameter p which define the amount of previous data points which should be represented in the output series. Each prediction is a linear combination with the form $S_n = C_1 \cdot S_{n-1} + C_2 \cdots + C_i \cdot S_{n-i} + \epsilon$ for $i \in \{1, ..., |lags|\}$. In our case we have three incoming time series which also leads to three outgoing time series. Our outgoing prediction for the data point n based on lags is represented as

$$\begin{bmatrix} S_{1,n} \\ S_{2,n} \\ S_{3,n} \end{bmatrix} = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,n-1} \\ C_{2,1} & C_{2,2} & \dots & C_{2,n-1} \\ C_{3,1} & C_{3,2} & \dots & C_{3,n-1} \end{bmatrix} + \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,n-1} \\ S_{2,1} & S_{2,2} & \dots & S_{2,n-1} \\ S_{3,1} & S_{3,2} & \dots & S_{3,n-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

Each time series $S_n \; \forall i \in \{1, 2, 3\}$ describes the n-th data point for the input time series i. The term ϵ is added to approximate an error, which is induced by the model.
We need to assure that our model is stationary, which means that the the stochastic process doesn't change over time. WHY?. Our data is stationary if we are able to assure this for all our time series data inputs:

I The mean μ is constant

II The standard deviation is constant

III There is no seasonality in the data

After we ensured, that our data is stationary, we used a PACF plot to find lags which could be statistically significant. The main reason why we choose PACF in contrast to ACF is that we only want to look at direct influences from S_{n-1} to S_n and not from S_{n-3} to S_n. // CONCLUSION FOR PACF //

After we found the optimal lag configuration of our model, we used a Granger Causality Test, to statistically prove that the salary time series really influences the student decision. The main idea of Granger Causality is that we are able to predict the data point $S_{i,n}$ directly with one or multiple other

Figure 1. PACF Plot for our VAR Model

time serieses. The Granger Causality Test uses exactly this idea, by inserting data points $S* := S_{j,m}$ for $j \in i : j \neq i$ and $m < n$. After we inserted a couple of terms we do a t-Test and an F-Test on this new $S*$ and as soon as both tests return statistical significance for the same S* we can assure that the three time series are Granger Causal.

3. Results

In this section outline your results. At this point, you are just stating the outcome of your analysis. You can highlight important aspects ("we observe a significantly higher value of x over y"), but leave interpretation and opinion to the next section. This section absoultely *has* to include at least two figures.

4. Discussion & Conclusion

Use this section to briefly summarize the entire text. Highlight limitations and problems, but also make clear statements where they are possible and supported by the analysis.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

Notes

Your entire report has a **hard page limit of 4 pages** excluding references. (I.e. any pages beyond page 4 must only contain references). Appendices are *not* possible. But you can put additional material, like interactive visualizations or videos, on a githunb repo (use links in your pdf to refer to them). Each report has to contain **at least three plots or visualizations**, and **cite at least two references**. More details about how to prepare the report, inclucing how to produce plots, cite correctly, and how to ideally structure

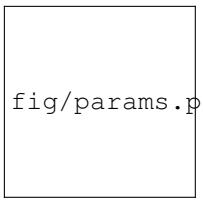

Figure 2. This is a figure caption. Write this, if we end up using that fig ;-)

your github repo, will be discussed in the lecture, where a rubric for the evaluation will also be provided.