A Prediction and Analysis of the interplay between Students and Salary Prospects in the STEM Field

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

Abstract

Our research focus is to provide valuable insights into how future salaries affect the decision-making of students when it comes to selecting their major at the University of Tubingen. For this purpose we explore the dynamic relationship between student enrollment, expected salaries, and graduation rates. We are exploiting the Vector Autoregressive model to demonstrate a reliable forecasts for future STEM student numbers, underscoring the significant influence of projected income on these predictions. Furthermore, by manipulating the salary variable in our analysis, we observe a corresponding increase or decrease in the number of students. This implies a sensitivity in enrollment trends to changes in expected salary, highlighting the potential impact of financial considerations on academic choices.

1. Introduction

In recent years, a growing number of high school graduates have chosen the path of higher education. Science, technology, engineering, and mathematics, known as STEM, make up a large proportion of all students, although this academic career is known for being one of the most demanding. This paper aims to explore the potential influence of financial incentives, uncovering hidden relations.

The analysis is built on three underlying datasets. The first dataset captures the historical enrollment figures of students at the University of Tübingen. It contains the total number of students for each semester, for all study programs ever offered. This dataset is maintained and published by the Statistics Bureau of the University of Tübingen. The University of Tübingen makes a good candidate to discover potential structure in the number of enrollments because STEM programs are strongly represented here. To quantify the salary expectations, a dataset from the Deutsches Statistisches Bundesamt is used. It contains the salary trend for all major industrial sectors in Germany. This dataset does not adjust for inflation, therefore an additional dataset from the Deutsches Statistisches Bundesamt is used to account for that (Section 2.1). It contains the yearly inflation rate in Germany. The number of students is also limited by a further parameter. Namely the number of people eligible for an academic career, hence all high school graduates. The Statistisches Landesamt Baden-Württemberg provides data about high school graduates in Baden-Württemberg for every year. We will not take other states into account, under the assumption that the number of students leaving Baden-Württemberg, to study elsewhere, cancels out with others moving to the University of Tübingen.

Our central contribution is a Vector autoregressive model (VAR) discussed in Section 2.2 that is built using just mentioned data. After processing (Section 2.1) the data, the model is optimized by different methods, see Section 2.3. This ensures that the relevant structure in the data is captured by the model. After verifying that the VAR model functions as expected and testing the model on a test set, Figure 1, we can predict all three input time series. As we are only interested in the number of students enrolled and the development of the salary, our research is focused on these two instances. This allows us to conclude in Section 2.4 that salary expectations have a significant impact on the amount of STEM students. Further, we can model different salary development scenarios, and see how the students would likely react (Section 3).

2. Data and Methods

2.1. Data Processing

Each data set used in the analysis had to be processed, to work for our analysis. The university data is only available in PDF format to download. To process the data, it had to

[*]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 5993286, paul.kaifler@student.uni-tuebingen.de, Bsc Computer Science [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.

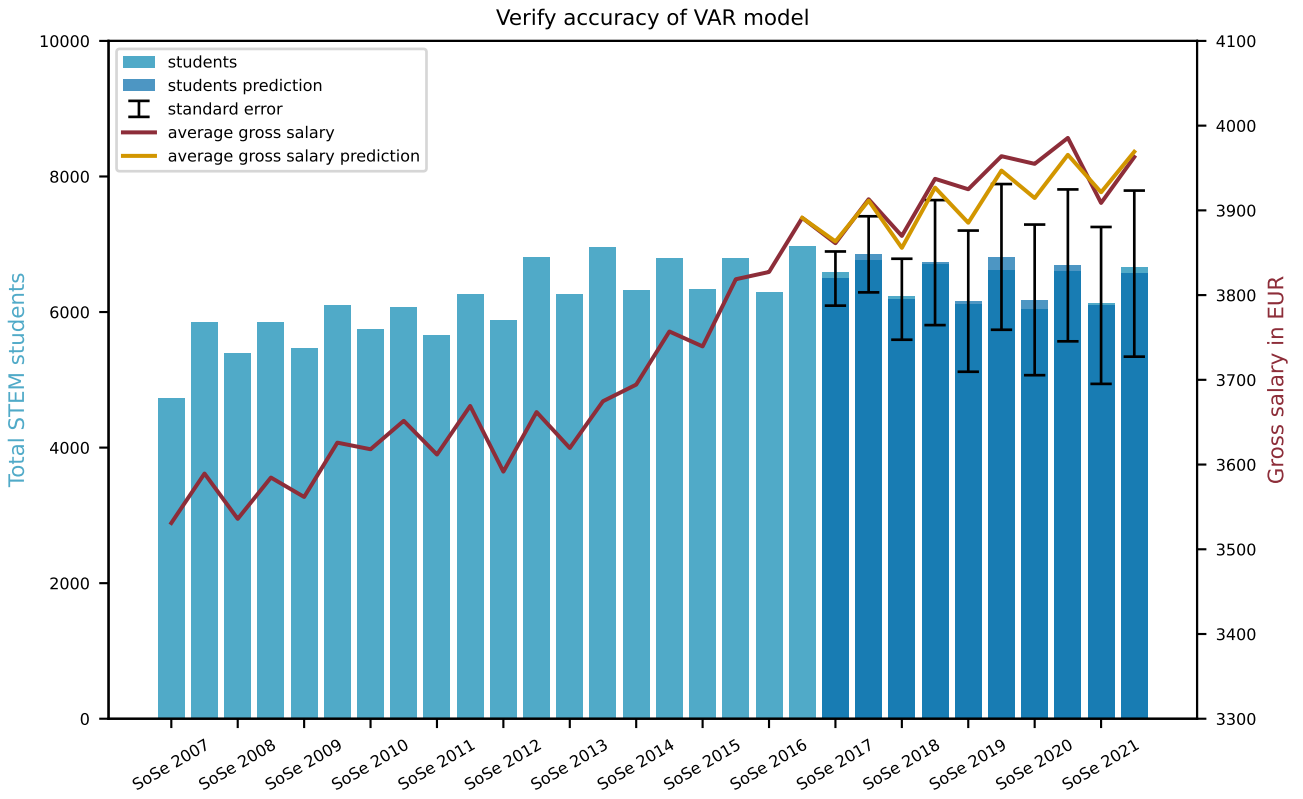

Figure 1. VAR model Prediction with optimal lag value of 5 on the test data set. Standard error with confidence interval value $\alpha = 0.05$.

be transformed into a more versatile format, using two steps. First, it has been transformed into CSV, which resulted in a corrupted file. A custom-written parser was then used, to restore the original data structure. In the year 2012, two years of high school graduates finished, due to the transition from G9 to G8 in Baden-Württemberg. This outlier in the data had to be flattened out to ensure the VAR model could be fitted correctly. The model uses the mean square error as a loss function, which makes outliers influence the data disproportionately. Half the number of students from the year 2012 has been spread over the next four years to flatten the data. The Deutsches Statistisches Bundesamt from which the average gross salary is obtained, does not account for inflation in its data set. Since we need to maintain comparability between past years, the salary has been adjusted for a yearly cumulative inflation rate.

2.2. Model creation

A vector autoregressive model is used to analyze relationships between multiple time series. The model creates a linear combination for each prediction which can be expressed in a matrix-vector notation.

$$\begin{bmatrix} S_{1,n} \\ S_{2,n} \\ \vdots \\ S_{m,n} \end{bmatrix} = \begin{bmatrix} C_{1,1} & C_{1,2} & \dots & C_{1,n-1} \\ C_{2,1} & C_{2,2} & \dots & C_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m,1} & C_{m,2} & \dots & C_{m,n-1} \end{bmatrix} \cdot \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,n-1} \\ S_{2,1} & S_{2,2} & \dots & S_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ S_{m,1} & S_{m,2} & \dots & S_{m,n-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The ϵ vector approximates an error, which is induced by the model. One main benefit of the VAR model is that it also assumes that $S_{i,j}$ is also influenced by $S_{k,l}$ $i, k \in \{1, \dots, m\} : i \neq k$ and $j, l \in \{1, \dots, n-1\} : j \neq l$. As mentioned in (Kotzé, 2023) a VAR model has one major advantage, over other possible approaches, which is its simplicity. Therefore its results remain interpretable. This model only uses one hyperparameter which describes the amount of past data that is used to make a prediction. This parameter is called the lag. The vector autoregressive model heavily depends on the stationarity assumption which ensures that the stochastical process stays the same over time. In this context, this is assured by the following three axioms: The mean μ is constant, the standard deviation is constant and there is no seasonality in the data.

Since we are only interested in predicting two timesteps into the future, which is equivalent to one year, we can safely assume that our data is stationary.

2.3. Choosing lag

To find the optimal hyperparameter we used two established methods, the partial autocorrelation function and information-based criteria as described in (Peter J. Brockwell, 2016). Particularly the AIC and BIC. The PACF only

considers the impact of the previous data point S_{n-1} on S_n to build a $(1-\alpha)$ confidence interval. We choose $\alpha = 5\%$ and get a 95% Confidence Interval. The optimal value according to PACF is 3.

	students	salary	graduates
Prediciton 1	0.704%	0.871%	2.658%
Prediciton 2	2.786%	0.301%	0.286%

Table 1. Difference in percent between the true value and predicted value for students, salary, and graduates. Lag set to 3.

The percentage differences between the true values and the predicted ones, especially for the students, are quite high, see Table 1. To get better results the AIC and BIC are used to find a better lag for the hyperparameter p. The abbreviation AIC stands for Akaike Information Criterion and it tries to find an optimum between the complexity of the model and fitting the data. This should prevent overfitting. BIC is short for Bayesian Information Criterion and penalizes more complex models. Therefore BIC leads to less complicated models than the ones from AIC, which sets the two apart. Both BIC and AIC return 6 as an optimal lag value for our data set. Since we are focusing on predicting students, low errors in that domain are more important to us. Minimizing these errors within the range from 3 to 6 leads to an optimal lag value of 5.

	students	salary	graduates
Prediciton 1	0.484%	0.419%	0.111%
Prediciton 2	1.346%	0.306%	0.939%

Table 2. Difference in percent between the true value and predicted value for students, salary, and graduates. Lag set to 5.

The percentage errors in Table 2 show that there is a significant improvement in the prediction of student numbers compared to the lag order of 3 from Table 1.

2.4. Test for causality

To ensure that the model does not try to find structure in the data, where there is none, we applied the Granger causality test. Granger causal implies that one time series is the result of another shifted time series. In this case, we try to find out if e.g. the salary time series is just a shifted version of the student's or the graduate's time series. The test verifies this by using series S_i and trying to insert data points from S_j or S_k $i \neq j \neq k$ to predict the outcome of the model. A t-test and an F-test are performed, if both tests return statistical significance, one time series can be constructed by another. The Granger causality test reports back no statistical significance for our data set, which means that predicting is a nontrivial operation.

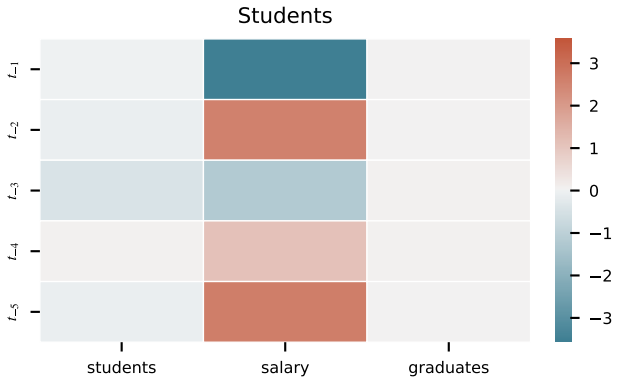

Figure 2. Coefficients of VAR model to predict students time series for t_0. t_{-1} to t_{-5} are the values of each time series respectively at the lagged time step, which is equal to one semester. t_{-1} is the most recent value.

2.5. Fitted model

With the correct lag chosen, the model fits the data by minimizing the mean square error. The returned coefficients describe the impact of a specific lagged value from one time series on predicting the next value. Figure 2 visualizes all coefficients relevant for predicting the number of students in the next semester. Notably, salary is the most relevant information for the prediction. All five coefficient values for each time step are non-zero, indicating a strong correlation. Graduates, on the other hand, are not as relevant, recognizable by values close to zero.

The suggested lag value of 5, earlier obtained in Section 2.3 is retrievable as well. The absolute value of the coefficient for predicting students on t_{-5} is the third largest, adding a lot of information to the model.

To further understand the data underlying correlations, one can look at the correlation matrix of residuals in Table 3. This matrix describes the degree to which the model's errors, called residuals, are related to each other. High correlations between errors can indicate a not fully captured underlying structure. Values close to zero, like -0.115 between salary and graduates, indicate that the errors in predicting the salary variable are not related to the errors in predicting the graduates variable.

	students	salary	graduates
students	1.000	-0.115	0.111
salary	-0.115	1.000000	-0.902
graduates	0.111	-0.902	1.000

Table 3. Correlation matrix of residuals using 5 lags.

3. Results

It is not possible to find a lag value that gives the best results for all three time series. However one can set the optimal lag value such that it suits forecasting students best (Section 2.3). With the obtained insights into the relation between the number of enrolled students, salary expectations, and number of graduates, we are able to predict the upcoming four semesters, see Figure 3.

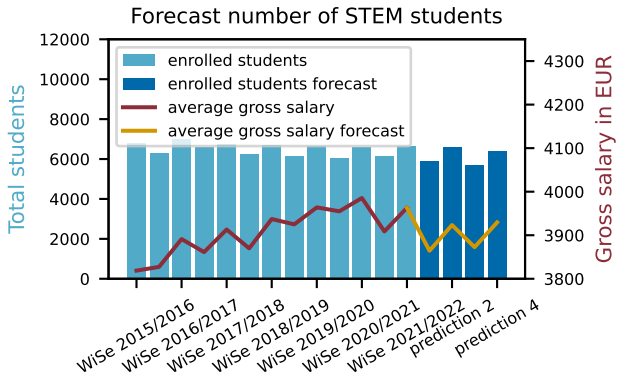

Figure 3. Predict the number of students enrolled in STEM courses.

It is also possible to construct different future scenarios. We generate a new data set with a 10% increase in salary expectations, Figure 4. This will be used as a basis to predict from, while the unchanged data set is used for training. The observed change in the prediction supports the obtained results in Section 2.5. There is a significant dependence between salary expectations and the number of students enrolled in STEM courses. As soon as the salary increases, there is also an upward trend visible in student numbers. The observed change is not massive which can be explained by the fact that the mass of students are students in higher semesters. The only direct major influence on the total number of enrolled students is made by people deciding to start this year. There won't be a lot of people changing subjects halfway through their academic careers. This also explains the huge spike in predicting the average gross salary. In the first few semesters, the amount of new students is not high enough, to capture the need for new people. Therefore the model suggests raising the salary even higher.

4. Discussion & Conclusion

In our research, we used a vector autoregressive model to predict the amount of students in the next semester. We used the salary data from the past 15 years and data about high school graduates in Baden-Württemberg as a prior. We optimized our Model to the best possible capacity, but there are strict limitations that can not be overcome. Vector autoregressive models are only capable of detecting linear

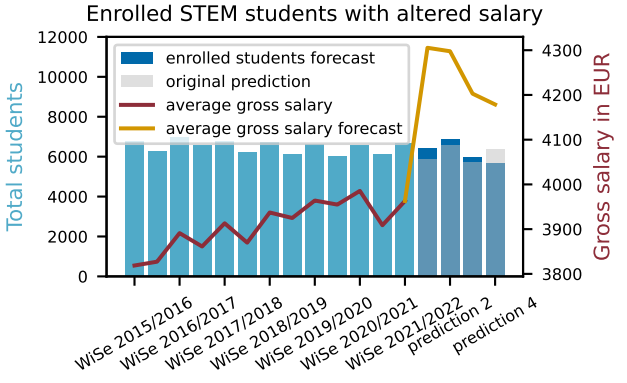

Figure 4. Predict the number of students enrolled in STEM courses using an increased salary by 10%.

patterns in the data. This occurs, because we only have linear combinations, based on values $S_{i,j}$ and coefficients to scale those values. If we wanted to improve our prediction we had to replace our model with other machine learning applications. A promising approach for this could be to implement a Random Forest algorithm. Our model on the other hand compensates for any potential lag in accuracy by being easily interpretable. Each parameter of the model influences exactly one part of the predictions and is therefore transparent in its process. Also worth mentioning is that our predictions are only based on three different inputs. There are certainly other variables impacting the number of students enrolled in courses at a university. Especially ones, that are not easily quantifiable and are subject to personal preferences. However, we are confident that our selected set of data captures the main aspects, as our model performs great on the provided test set.

Contribution Statement

Abdallah Abdul-Latif analyzed the optimal parameter lag and mapped courses to salary sectors. Lisa-Maria Fritsch was responsible for data pre-processing. Paul Kaifler fitted the model to the data and created visualizations. Maximilian Schnitt parsed the university data and analyzed optimal parameter lag. All authors jointly did the analyses of the data and wrote the text of the report. A repository containing the complete analyses and results can be found on GitHub.

References

Kotzé, K. Vector autoregression models. Technical report, School of Economic, 2023. URL https://kevin-kotze.gitlab.io/tsm/ts-10-note/.

Peter J. Brockwell, R. A. D. *Introduction to Time Series and Forecasting*. Springer, 2016.