My Data Literacy Project
(Replace this with your Project Title)

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

Abstract

As often mentioned in the news and in the daily life, people suggest that a higher

1. Introduction

As often mentioned in the news and in the daily life, people suggest that a higher expected salary influences the job aspirations of the youth. We are curious about the underlying truth of thgis stetement. We have three Datasets, one Datset shows the Expected Salaries for various industrial sectors, the second one shows the amount of students in the STEM sector in Tübingen from the Wintersemester 2005/2006 to the Wintersemester 2023/2024. The third dataset contains information about the amount of students, who finish high school in the certain year. We did a feature analysis of the three datasets to get a better understanding of the data but the main focus of our analysis is on predicting the amount of students in the STEM sector for the next years. We used different approaches like a Random Forest approach with SARIMA and a Multivariate Autoregression Model to investigate which approach works best for our purpose.

2. Data and Methods

We used three different pre collected datasets, our University data was collected by the Statistics Department from the University of Tuebingen. They used the immatriculation data from the students to create this dataset, which leads to a quite accurate data which is necessary for the usage in our models. // HERE WE NEED MORE INFORMATION ABOUT THE DATA FROM THE OTHER DATASETS//
The first step was to do some basic feature engineering on the datasets. We used various plots like line graphs, bar charts and pie charts to get a better understanding of the data and detect trends. After we finished the basic feauture engineering we started the main analysis of the Vector Autoregressive Model (VAR). 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.
We've chosen the vector autoregressive model, because we have multivariate data and want to get a deeper understanding, if one of the datasets affects the others, so that we are able to predict one time series with another. A vector autoregressive model doesn't predict a single time series analyses, but multiple. Each time series is a linear combination which consists of n lags and an error term. A lag is the data from the past which we use to estimate a new data point. We define the amount of lags in the hyperparameter p which defines the maxamount of lags which are used in the linear combination. This linear combination also uses an error term, because we can't guarantee that our model predicts the correct output.
One crucial step in the usage of those AR models, is that we have to ensure that our data is stationary. Data is stationary if those three attributes are fulfilled

1. The mean μ is constant
2. The standard deviation is constant
3. There is no seasonality in the data

// CHECK IF OUR DATA FULFILLS THOSE THREE ATTRIBUTES ///
After we ensured this, we did a PCAF to find statistically significant lags which we could use to find an optimal hyperparameter p of our vector autoregressive model. A PCAF only represents the direct influence of the past datapoint S_{n-i} for $i \in \{0, 1, ..., n\}$ to S_n. This gives us a graphical visualization with an error band of non statistical significant lags in // GIVE A RANGE //.
After we used this PCAF we tried the different amount of lags, which looked like they have a significant input on the current datapoint S_n. With the optimal lag we did a Granger

[*] Equal contribution [1] Matrikelnummer 12345678, first.last@student.uni-tuebingen.de, MSc Machine Learning [2] Matrikelnummer 12345678, first.last@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.

Causality Test to get an understanding if one timeseries really influences another. The Granger Causality Test, does a t-Test and F-Test checks if we can predict S_i with $i \in \{2, 3\}$ the time series S_j $i \neq j$. The Granger Causality Test internally injects lags from the time series S_J into the linear combination of S_i, then it does a t-Test and an F-test. As soon as one of those terms is verified by both tests, then the test result is true.

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

References