MSc. Applied Data Science thesis

Universiteit Utrecht

Prediction of master student influx: Faculty of science

Supervisor
Prof. Dr. Gerard Barkema

Second supervisor
Dr. Albert Gatt

Student/Author
Thomas Mosterd, 6280838

1 Abstract - Individual

The goal of this thesis is to accurately predict the student intake for the masters Nanomaterial Science, Computing Science and Theoretical Physics at the Utrecht University. The data used to made this prediction is a dataset containing detailed information about the applications in the last 2 years, and a dataset containing the number of applications since 2010. Utrecht University is interested in these predictions as they have to form classes, assign teachers and select fitting class rooms. The University can not afford to wait until the final number of enrolments is known as that is too late. The model used to make this prediction is the Sarimax model, a time series forecasting model. Using the 2021 data as a test set, the model parameters were tuned to ensure the prediction for 2022 is as accurate as possible. Two separate predictions were made. The first one is based on the number of applications on June 5th. This date is chosen as it is the first data point after the application deadline. For the other prediction, another Sarimax model was used to predict the number of applications in June, based on the data in May. Then the predicted applications were used by the model. This allows for predictions to be made a month earlier, which could be beneficial for the University. The predictions made by the June model appear to be the most accurate, and while the exact number could differ, there is a high probability it will fall in the confidence interval. Predictions made by the May model appear slightly less accurate, but still accurate enough to be used if the extra month is needed.

2 Preface - Shared

This thesis contains a shared research part and therefore also a shared written part. The goal of this research project from Utrecht University was to predict the number of master students that will start any of the following 12 masters from the faculty of science;

- Artificial Intelligence (AINM)
- Nanomaterials Science (NASC)
- Computing Science (COSC)
- Experimental Physics (EXPH)
- Game and Media Technology (GMTE)
- History and Philosophy of Science (HPHS)
- Business Informatics (MBIM)
- Climate Physics (CLPH)
- Mathematical Sciences (MSCM)
- Theoretical Physics (THPH)
- Human-Computer Interaction (HCIM)
- Applied Data Science (ADSM)

The 12 masters have been divided over four students so that each student will predict the influx for three of the masters:

- Wannes Coppelmans (MSCM, HPHS, CLPH)
- Thomas Mosterd (NASC, COSC, THPH)
- Sander Vonk (AINM, EXPH, MBIM)
- Tommy Wirken (HCIM, GMTE, ADSM)

In this thesis, there is an indication for each section or subsection whether the part is written together (shared) or individually (individual).

3 Introduction

3.1 Motivation and context - Shared

During summer holiday, students are relaxing and releasing stress from the past school year. Meanwhile, universities are very busy preparing for the upcoming year. Classes have to be formed, teachers have to be assigned to classes, classes have to be assigned to rooms and each year it is a surprise how many students will show up in September.
Because the scheduling of rooms and especially staff needs to be done well in advance, it is usually not possible to do this when the number of students that will start a master is already certain. The number of students that apply only gives an indication of the actual number of starting students. Students can be rejected by the university or can cancel their application at any time. Therefore, the exact number of students that should be counted on is usually known only a few days or weeks before the start of the programme.
The aim of this research is to make a prediction, at the latest at the start of June, of the number of starting students in September for each of the selected masters from the faculty of science department at the University of Utrecht. To make this prediction, information like the number of applications and enrolled students of previous years is taken into account.

3.2 Literature overview - Shared

The most similar is research performed by Tilburg university [1] focusing on both bachelor and master applicants and basing admission rates on several factors. Including the attendance of students at events organised by the university and the distance between the residence of the student and the university. It was found that the attendance of these events is a successful way of predicting the intake of students.
Most research in this space is based in the USA. Which is harder to compare as the financial challenge of entering a university is very different to the situation in The Netherlands. Looking at an example of these papers [2] it can be seen that the data used is much more detailed than datasets in this paper. These data points include ethnicity, income of the parents, education of parents and achieved grades in the previous education. These data points allow a more extensive research into relevant factors for admission, but it would not be possible for Utrecht University to supply this kind of data as this goes against the privacy laws within the EU.

3.2.1 Influx prediction

While many universities have the problem of not knowing well in advance how many students will start a programme, no published literature on the prediction of student influx seems to be available. Yet, making predictions for the future is very common practice and so there are many algorithms that might be able to make predictions on the number of enrolling students for the upcoming year. The models that make future predictions based on “historical” data are called time series forecasting models.

3.2.2 Time series forecasting

Time series forecasting models use data from the past, such as stock prices, sales of products or unemployment rates and based on these data, a prediction of the stock price, number of sales or unemployment at some point in the future is made. There are many algorithms that can make such predictions based on time series data. There are different types of forecasting models; for example regression models, decompositional models, moving-average models, exponential models and deep learning models. All of these models have a different approach to predicting future data.
(Linear) regression tries to fit a line as closely as possible to all data points [3]. When this line is extended beyond the known dates, a prediction for future values can be seen. This rather simple model will fit a straight line to the data while an exponential, parabolic or different function may be more appropriate. There are many of such more complex models and some of the most important ones are described in the book Forecasting: Principles and Practice [4]. Firstly, exponential (smoothing) models are commonly used to solve time forecasting problems. They are able to follow different patterns in data and mirror these in their predictions but they are normally not well suited to follow seasonal trends unless they are extended. Decompositional models are better suited to pick up on seasonal trends. They decompose a seen trend in data in a seasonal and a cyclic (for example yearly) component and use this decomposition to make future predictions. Moving-average models, which are treated extensively by Hyndman and Athanasopoulos [4], are also very commonly used for time series forecasting and can also be extended to support seasonality. These models aim to describe auto-correlation (similarity between observations as a function of the time lag between them) in data.
Arguably the most complex method to solve time forecasting problems is using deep learning, or more specifically neural networks. The operation of these networks is based on the structure of animal brains and while they are able to solve very complex problems, a lot of data is usually required to train these models. Finding out how a neural network comes to a certain result using its input is extremely difficult [5].
Out of the different time forecasting methods that have been discussed, moving-average models are the most promising solution for solving this time series forecasting problem. While a relatively simple moving-average model like AutoRegressive Integrated Moving Average (ARIMA) might not make very accurate predictions, this model can be extended to pick up on seasonal data and can also include external variables. A Seasonal Autoregressive Integrated Moving Average with eXogenous regressors model (SARIMAX) model should theoretically be able to mirror a pattern seen in data, taking into account the time of year and external data. In this case, SARIMAX might make a prediction of the number of enrolments in September based on the development of student applications and the time of year while also taking the number of enrolments of previous years into account [6].

3.3 Research question - Individual

What number of students can be expected to start their masters Nanomaterial Science, Computing Science and Theoretical Physics at the Utrecht University

in September 2022.

4 Data - Shared

4.1 Selected data exploration results

4.1.1 Master bachelor herkomst

The file "master bachelor herkomst", contains information on the number of applications per bachelor from students for masters and pre-masters that are being or have been offered from 2011 until 2022 at the Utrecht University. The file contains what bachelors the students who signed up for a (pre-)master have followed or are still following. In many cases, the bachelor is not present in this file since only UU bachelors are stored so students from different universities are counted towards the unknown bachelor student total. For the purpose of this research, the pre-master programmes were removed from the data. After further cleaning of the data by merging studies in the data that changed names, 12 master programmes were left. For each of these programmes, the number of students that applied for a masters could be found by summing up the count of students whose bachelors were known combined with the number of students with an unknown bachelor. Besides these data on the origin of students who applied for master programmes, there is another data sheet where the number of enrolled students for masters can be found per bachelor.
The application data is updated yearly at the start of June, since the deadline for applying for a master programme is June first. The number of enrolled students is added after the first of September each year. These data contain the number of applications from 2011 until 2022. The total number of enrolled students at the start of each master is known for every year from 2011 up until 2021.

4.1.2 Ati trend

Another dataset, called ati trend, is a dataset that stores student applications and tracks the status of the applications. The data starts in April 2017 with two updates per year and from October 2019 onwards there is an update roughly every week (50 updates per period from the 28th of November until the 13th of November). An important date in these data is the first update in June (June fifth), since this is the first update after the application deadline for masters that start in September. After that, the number of applications is set and only the status of applications can change. Applications can have the following statuses; rejected by university, cancelled by student, condition: pre-master, waiting for complete file, request for review, (conditionally) admitted – acceptance unknown, (conditionally) admitted – accepted, enrolled. These statuses are ordered from least to most likely that a student will enrol in a programme and are tracked from the moment a student applies for a master programme. Since students can apply for multiple masters, the number of applications a student has made is also stored. Another important date is the 28th of November, as all ongoing applications with September as their starting date are cancelled so the number of applications drops by a lot.
These data contain both a yearly and a seasonal trend. The number of applications grows from December until June each year and generally most applications are made in May.
There are 406.622 rows and 39 columns in the data (up until the most recent

update on June fifth, 2022) that store the information as we mentioned before. For each row there is the date that indicates when the row was updated. There is also an indicator for the gender and for the highest followed and completed education of a student. Information regarding the programme that a student applied for like the name, Osiris key and the starting date of the master are also stored. The data does not include personal information with which a student can be identified like a student id, residence, or the exact education a student has had.

4.2 Data preparation for analysis including motivation (integration, missing data analysis, etc.)

4.2.1 Master bachelor herkomst

To prepare these data to be input for a model, a few steps had to be taken. Firstly, pre-masters had to be removed from the data since this research is about masters only. After cleaning the data it had to be restructured due to the layout. The applications and the enrolments for each of the masters from all the bachelors had to be summed up in order to get the total number of applications and enrolments per master program.

4.2.2 Ati trend

In this dataset, pre-masters were also removed. Then the rows with a start data not in September (some masters also start in February) were removed.
For this analysis, only the date (DATUM), the number of applications (AANM_AANTAL), and the programme (BK_PROGRAMMA) were retrieved from the dataset. After replacing the NA values with zero and setting the correct format for the date column, the dataset was split into twelve different datasets based on the master programmes. The data of each master programme follows its own pattern. Therefore, it is necessary to separate the masters to create different prediction models.
The data for most masters (except Applied Data Science and Human-Computer Interaction) begins in June 2016, but between June 2016 and November 2019, the data was updated twice a year. Beginning in November 2019, data was updated weekly. To make a forecast in a time series, the data must be consistent. An example of this inconsistency can be seen in Figure 1. Therefore, the data prior to November 2019 was removed so that the data is consistent on a weekly basis.

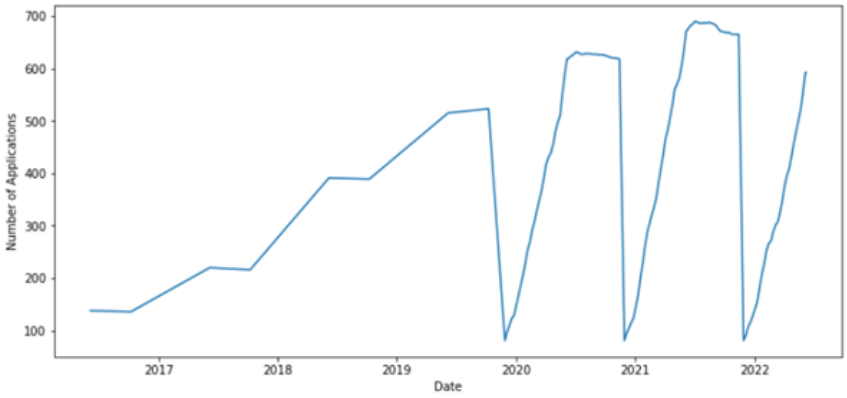

Figure 1: Number of applications for Artificial Intelligence (example), before removing the data before November 2019.

The final step in preparing the data is to create a column called WeeksTillDeadline. This column uses the date column to count how many weeks remain until the registration deadline. It was found that these data work well as a predictor in the SARIMAX models to predict the number of applications. For one of the masters (History and Philosophy of Science), the year also proved to be a useful variable to predict the number of applications. The year was therefore retrieved from the date column, and added to the dataset for this particular master.

4.3 Ethical and legal considerations of the data

Both of the datasets are fully anonymised and students cannot be identified with these data. The data is not publicly available and will not be shared with third parties.

5 Methods- Shared

5.1 Translation of the research question to a data science question

To predict the influx of master students, past data has to be fed to a model that can then make a prediction based on the data that is provided. The two datasets require different approaches since the "master bachelor herkomst" file has 11 years of data for the enrolments and 12 years for the applications, with yearly updates. On the other hand, the "ati trend" dataset is much more complex and has weekly updates, but only for the past nearly three years. So while these data are much more extensive and has more data points, fluctuations in the final number of applications and enrolments have a much bigger effect on the prediction than for the data that spans almost twelve years.
To come to a good prediction, the steadiness of the data that contains the number of applications and enrolments of the past 11 years was combined with the knowledge there is on the development of the number of applications and their statuses for the current year. A model was developed that could predict the number of enrolled students based on the number of applications in June and another model was built that could predict the total number of applications there would be in June based on the data up until May. Combining these two methods resulted in a model that was able to make a substantiated prediction of the number of enrolments in 2022.

5.2 Motivated selection of method(s) for analysis

The problem of predicting the number of enrolling students for a master programme is a time series forecasting problem, since the number of enrolled students in previous years is known and the goal is to extend the trend that is seen in the data in such a way that an accurate prediction for September 2022 is made. Because it is desirable to have an accurate prediction as early as possible, multiple methods have been tested and used.
Because the total number of applicants is known after the application deadline, which is June first, a prediction of the number of enrolling students can be made using the ratio between the number of applications and enrolments in previous years. However, this ratio changes quite a bit between different masters and over the years. Therefore, it is also important to look at this problem as a time series forecasting problem to pick up on trends that may be happening. Furthermore, it is desirable to get a prediction before June so the university can for example already start looking for more staff members if a relatively high number of students is expected. To do so, the weekly development of applications is also taken into account and time series forecasting is again used here to extrapolate in order to predict the number of applications in June.
Out of the different time series forecasting models and other methods, SARIMAX appeared to give the best results so this model was further developed.

5.2.1 Predicting the number of enrolments based on the applications and trend

Several models were tested to make predictions based on both datasets that have been described. Moving-average models are commonly used for time series analysis and ARIMA models are the most popular type [4]. Multiple models were compared, including linear regression, Vector AutoRegression (VAR), Simple Exponential Smoothing (SES), ARIMA and SARIMAX. Out of these models, SARIMAX seemed the most promising as it gave the most accurate predictions before fine-tuning the models. Therefore this model was chosen and further developed. Time series analysis is frequently done using SARIMAX and it was unsurprising that this model gave the best result early on [6].

5.2.2 Predicting the number of applications in June based on the trend up until May first 2022

Again, for solving this problem, SARIMAX showed the most promising results when testing multiple models. Again, this model was chosen and fine tuned in order to predict the number of applicants after the deadline on June first, based on the ati trend data up until May 2022. The number of applications from the weeks before are important in this model to follow a trend. The predictor that has been used is called "WeeksTillDeadline". After getting inaccurate predictions for the master HPHS, it was found that the year is also an important predictor for this particular master. Therefore, the year is added in the prediction model for this master.

5.3 Motivated settings for selected method(s)

The SARIMAX model needs a number of input parameters to work properly. These parameters are p, d, q, and t, all of which have different functions. The p is the order for the auto-regressive component, d for the integrated component and q for the moving-average component. t is the parameter controlling the deterministic trend. There are several approaches to choosing values for these parameters. One approach is to look at the (partial) autocorrelation graph and perform a Dickey Fuller test to find the best values for the model. The disadvantage of this approach is that human vision is required to choose the optimal model, which results in the code for the predictions not being reproducible (especially for each master individually).
An approach that can be performed automatically is to test every possible combination in parameters and choose the combination with the lowest error. All models were trained and tested based on data up to the previous year. The model with the lowest error was then selected to train the data up to 2022, after which the predictions for the current year were made. To calculate this error, the "mean absolute percentage error"(MAPE) between the predictions and the actual values is taken. Figure 2 shows the combination of parameters tested for the model and Figure 3 shows how the MAPE is calculated.

```python
def arima_configs():
    models = list()
    # define config lists
    p_params = [0, 1, 2]
    d_params = [0, 1, 2]
    q_params = [0, 1]
    t_params = ['n','c','t','ct']

    for p in p_params:
        for d in d_params:
            for q in q_params:
                for t in t_params:
                    cfg = [(p,d,q), t]
                    models.append(cfg)
    return models
```

Figure 2: Creating all possible combinations of parameters for SARIMAX within the given ranges. These parameters are p, d, q, and t, all of which have different functions. The p is the order for the auto-regressive component, d for the integrated component and q for the moving-average component. t is the parameter controlling the deterministic trend. The 'c' indicates a constant term, 't' indicates a linear trend in time, and 'ct' includes both. 'n' means no trend term.

```python
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
```

Figure 3: Used for calculating the mean absolute percentage error(MAPE) to measure the performance of the SARIMAX model. MAPE is the sum of all errors divided by the sum of true y (or forecast).

5.3.1 Predicting the number of enrolments based on the applications and trend

To train the SARIMAX model, different subsets of master programmes were created since some of the masters showed similar patterns in for example their ratio between applications and enrolments or in their popularity over time. The subsets that have been used to train the model that generates predictions for the master programmes, is shown in Figure 4.

Masters	Trained on data of the masters											
AINM	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
NASC	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
COSC	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
EXPH	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
GMTE	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
HPHS	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
MBIM	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
CLPH	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
MSCM	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
THPH	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
HCIM	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM
ADSM	AINM	NASC	COSC	EXPH	GMTE	HPHS	MBIM	CLPH	MSCM	THPH	HCIM	ADSM

Figure 4: This table shows the data selection for training SARIMAX and making predictions. It indicates for each master what data was used to train the prediction model. For each master, the number of past applications and enrolments from 2011 until 2022 are used. For example, to train a model to predict the number of enrolments for AINM in 2022, the data for AINM, EXPH, GMTE, MBIM, THPH and HCIM on the number of applications and enrolments from 2011 onward have been used.

To predict the number of enrolments, only the yearly number of enrolments and applications is available. This means that these data does not contain seasonality and it was therefore not included in the model.
The external variables that were added to the model to predict the number of incoming master students are the name of the master programmes and the number of applications of the previous 11 years including (the prediction of) the number of applications in 2022. Because there is a positive correlation between the number of applications and enrolments, it is important to take the number of applicants into account. While the ratio between applications and enrolments differs between years and masters, this relation can still be picked up by the model and used for making a prediction. Adding both external variables improved the predictions of the model significantly.
To test the accuracy of the predictions of the model, the model has been run with the data for 2021 instead of 2022. This way, the predictions could be compared to the real number of enrolments for each of the master programmes. The predictions for the number of enrolments in 2021 compared to the real number of enrolments can be seen in Figure 5.

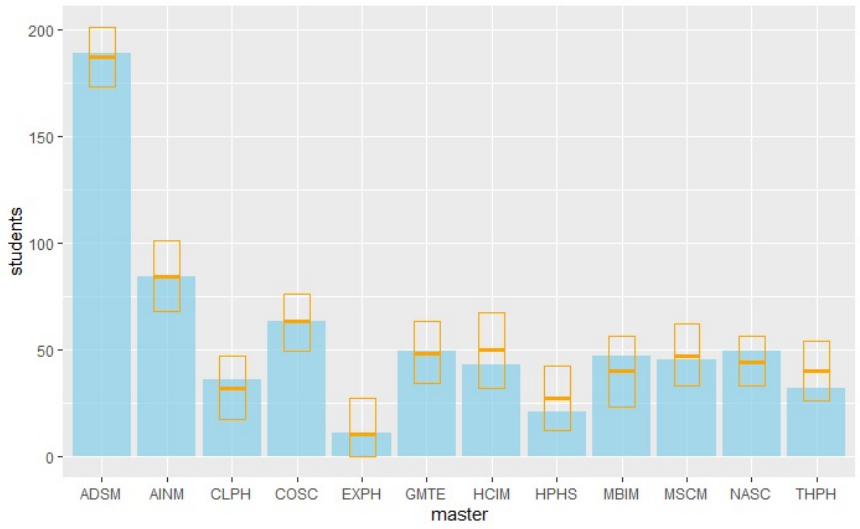

Figure 5: This figure shows the number of enrolled students for 12 master programmes in 2021. The bars represent the number of students enrolled at the start of a master programme. The yellow rectangles represent the range of the prediction that has been made for 2021. The lowest yellow horizontal line shows the minimum number of expected enrolments, the top yellow line shows the maximum expected number of students and the middle yellow horizontal line indicates the predicted number.

5.3.2 Predicting the number of applications in June based on the trend up until May first 2022

The ati trend data does contain seasonality. However, including this increased the complexity of the model while not leading to a significantly better model so it was decided to leave seasonality out of the model.
The most important predictor in this model is the WeeksTillDeadline column. This column turns out to be significant in predicting the number of applications in June because it starts when the first applications are coming in and ends on the deadline. You could say that this predictor creates a new index.
Another variable that was tested to act as a predictor was the year. This variable was found to be significant in predicting the number of applications only for the History and Philosophy of Science master. Therefore, the variable is used as a predictor only for this master.

6 Results - Individual

The confidence interval used is a 95% confidence interval. For Computing Science, the predicted intake for 2022 using the June model is 69, with a confidence interval of 57-80. This is a slight increase over the previous years, 54 and 63, which can be explained by the increased number of applications. Using the in May predicted applications, the intake is expected to be 67. The prediction for Nanomaterial Science in June is 42, with the confidence interval being 28-57. the prediction is slightly lower than last year(49), but still higher than two years ago (35). These fluctuations are in line with the changes in applications. The prediction using the May model is 47. Which is still in line with the changes in applications. Lastly, the prediction for Theoretical Physics is 32, with a confidence interval of 21-44. This is the exact same as the year before, which is expected as there is a very small change in applications. The prediction in May is 34. The prediction of the model, based on the data in June, and the data of the previous two years is visualised in figure 6. The prediction of the model in May compared with June model is visualised in figure 7.

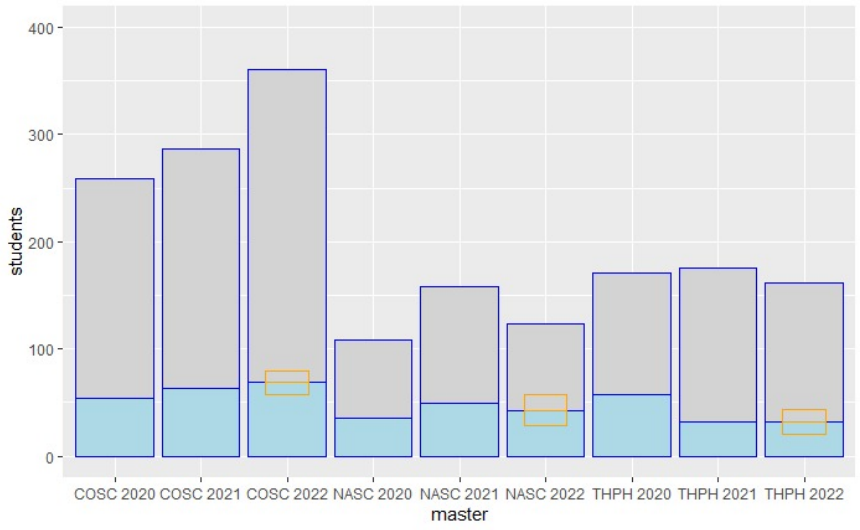

Figure 6: Graph showing the intake (blue) and applications (grey) for the masters. The blue bars for 2022 are the prediction, with the confidence interval being visualised with the orange bars.

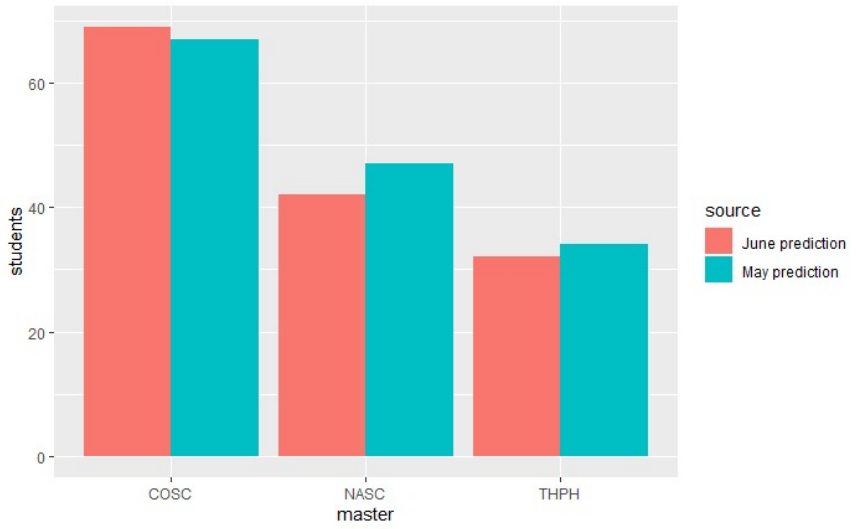

Figure 7: Comparing the results of the prediction using the actual intake numbers from June 2022 with the prediction based on the predicted intake numbers based on the data from May 2022.

7 Conclusion and discussion

7.1 Answering the data science question - Shared

Using a SARIMAX model turned out to be a satisfying approach for solving the time series forecasting problem of predicting master student influx based on the total number of applications in June. The confidence interval that was produced by the model was rather broad in some cases, especially for the masters with a relatively low number of applications. This can be explained since the outcome of a single student's application has relatively more impact on the final number of enrolments for smaller masters compared to larger ones.
SARIMAX also turned out to be a satisfying solution to predict the total number of applications in June based on the available data in May since the predicted number of applications was very close to the real number of applications. This was done to let Utrecht University predict the total number of enrolments in September earlier in the year, in May instead of June. If the enrolment prediction model is used in June it will give slightly more accurate predictions but the predictions it can make in May should be close enough to the ones made in June to be of use to Utrecht University.

7.2 Answering the research question - Individual

As can be seen in figure 6, the expected student intake for 2022 is not very different from 2021. This is mostly due to the number of applications not changing much from the previous years. So there is no reason to expect a very different result. Even though it is impossible to know the result at the writing of this thesis, it is expected that the actual intake will be close to the predicted number, or at least within the confidence interval. A different way to predict the intake includes the previously described model to predict the applications in May, instead of using the actual number when it is known. Using these predicted applications as the basis for the model means the prediction can be made in May instead of June, which could be desirable for the University. In figure 7, it is shown that while the prediction is somewhat different. The result is small enough for the prediction to still be usable. In conclusion, the model can be used to accurately predict the student intake as early as May. It can also be used after the deadline, with increased accuracy to compensate for the extra waiting period.

7.3 Describing implications for the proper domain setting - Shared

Although this study will not be made public, the prediction models used in this study may be useful for other (Dutch) universities. General predictors are used in the prediction models. The columns 'WeeksTillDeadline' and 'Year' are used to predict the number of applications in June, where the number of applications in June and the master programme are used for the final prediction. These variables can easily be translated into variables of other interested parties within the same domain.
There are a few things to keep in mind when using this research for other domain settings. This research focuses on master's degree programmes at a university. Therefore, it is recommended that this research be used only for university

master's degree programmes.
Another important point to keep in mind when using this research is that it is focused on a Dutch university. It is likely that other countries have different rules for applying for a master's degree, which may affect the conversion between applications and enrolments. Also masters that use a numerus fixus are not compatible with the model.

7.4 Discuss ethical implications and consideration - Shared

The predictions that are being made for the student influx are made in order to help Utrecht University with business like making schedules and hiring staff. Inaccurate predictions could lead to a decreased quality in the schedules for students or the amount of available staff. Yet, the University has dealt with the problem of predicting student influx for a long time already and the predictions that have been made for 2022 are only there to support the estimate that the university has made. An inaccurate prediction will therefore likely have little consequences for students or staff at Utrecht University. Moreover, the predictions that have been made do not deviate very much from the number of students in previous years so based on the predictions made for 2022, no big changes will have to be made within the university.

7.5 Future research - Shared

When making predictions for the student influx of the discussed masters in upcoming years, there are some insights that have been gained that may be useful.
Firstly, when analysing the "master bachelor herkomst" data, a big difference in conversion rate between students with a UU bachelor and students without a UU bachelor could be seen. Students who completed a UU bachelor are far more likely to enrol in a UU master when applying for one compared to students who did not have a UU bachelor. This can be seen in Figure 8 and Figure 9. A reason for this might be that international students are less likely to enrol in a master at Utrecht University compared to Dutch students since there is a higher barrier for these international students such as finding housing space and not speaking the Dutch language. Because of this difference in conversion from application to enrolment, the decision was made to make a difference between UU and non-UU applications. A problem with this method that was encountered later on, is that many students graduate from their bachelor in July, August or even September. Until these students graduate they will be listed in the data as not having a UU bachelor, even if they are doing their bachelor at Utrecht University. A possible solution to this problem could be to predict a conversion rate for the UU bachelors of students that applied and taking this into account with making a prediction. However, due to a lack of time this was not included in the current model so it was decided to not make a difference between UU and non-UU bachelor student applications. It is likely that when doing this, the predictions will become more accurate.
As was already discussed in the data section, the ati trend dataset was very useful since it contained weekly updates on student applications but the problem with it was that it only had this update frequency from late 2019 onward. While this has influence on the stability of the pattern that is seen over the years, the

Covid-19 pandemic also could have had an influence. It is reasonable to assume that the Corona virus had impact on the data from 2019 until at least June 2022. During the last few years, there were many Covid-related restrictions at times such as negative travelling advice and this likely lead to many international students choosing not to go study abroad. It's also likely that this trend is visible in the data and therefore also impacts the accuracy of the prediction model that was used to predict the student influx for September 2022. While this might not be too problematic for the 2022 prediction, when making future predictions it might be important to take the impact of the Covid-19 pandemic into account. On the other hand, the ati trend dataset will keep expanding over the years and the patterns in this dataset will grow more stable because of this. Therefore, in the future, these data will likely be far more important compared to the "master bachelor herkomst" data. One can assume that future predictions of master student influx will get more accurate because with each week and year that passes, more data becomes available, which a model such as SARIMAX can be trained on.

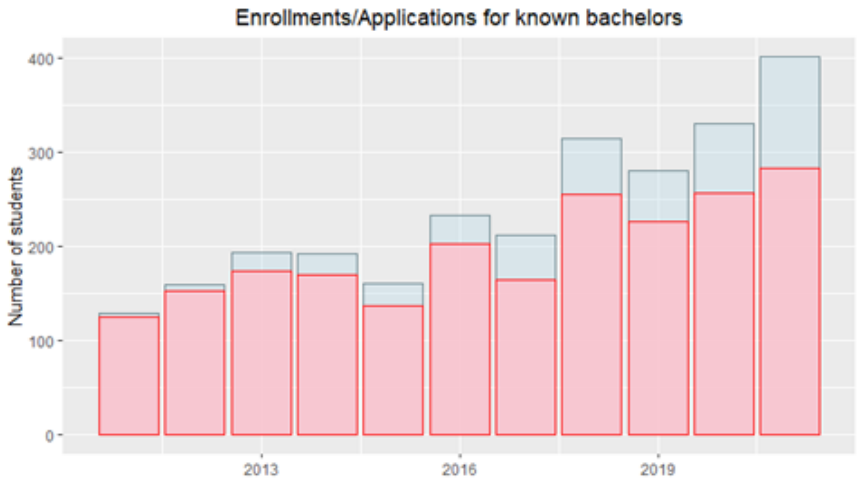

Figure 8: Blue represents the applications per year while red represents the actual enrolments per year.

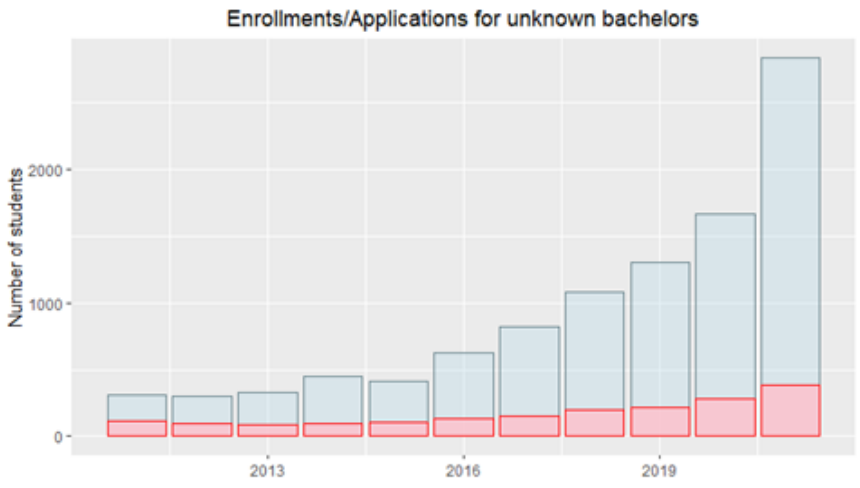

Figure 9: Blue represents the applications per year while red represents the actual enrolments per year.

8 Acknowledgements - Shared

We would like to thank Gerard Barkema for his experience and guidance during this research project and we would like to thank Marc Coemans for providing us with the data and his expertise.

References

[1] Yente Hamers. Predicting student enrollment. Msc thesis, Tilburg University, 2017.

[2] Ahmad Slim, Don R. Hush, Tushar Ojha, and Terry Babbitt. Predicting student enrollment based on student and college characteristics. In *International Conference on Educational Data Mining*, 2018.

[3] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, New York, 2013.

[4] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, Melbourne, 2018.

[5] Nikolaus Kriegeskorte and Tal Golan. Neural network models and deep learning. *Current Biology*, 29:231-236, 2019.

[6] Nari Arunraj, Diane Ahrens, and Michael Fernandes. Application of sarimax model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems*, 7:1-21, 2016.