University of St.Gallen

School of Management, Economics, Law, Social Sciences, International Affairs and Computer Science

Clickstream Prediction and Analysis

Term Paper

Workshop Fundamentals of Data Science

Johannes Binswanger

Thomas Aeschbacher

05.11.2021

Alex Sebastiani (19-615-095)

Sacha Roduit (20-618-252)

Alex-Paolo Marchese (18-614-776)

Sophie Lejay (18-614-081)

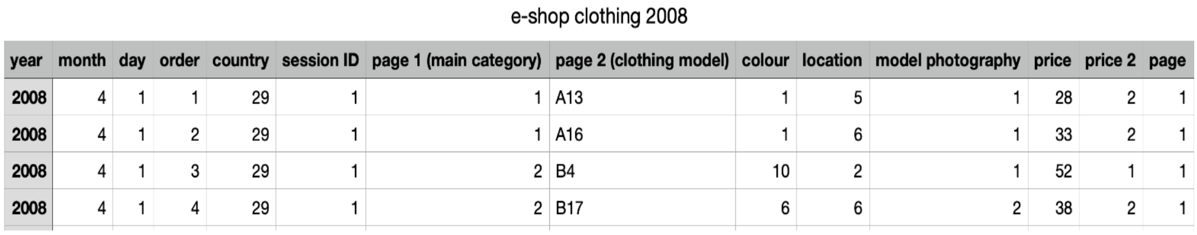
Contents

Table of Figures

1. Description of the Dataset

Our aim is to predict the number of clicks for an online store offering clothing for pregnant women based on different features. We assume the number of clicks to be positively correlated to the quantity of orders. Hence, our analysis could help online clothing stores determine whether and to what extent certain features could be used successfully for predictive purposes. In order to optimize our results, we seek to experiment with various machine learning models, namely linear, polynomial and logistic regressions, neural networks as well as random forests. Furthermore, cross validation is used to assess the accuracy of a model and mitigate overfitting. Our project represents both a regression and classification problem as we attempt to predict both the exact number of clicks and whether this number is below or above a certain threshold.

The features are composed of four numerical variables and seven categorical variables. The former consists of the day, month, price and page number within the e-store website, while the latter includes the country of origin of the IP address, product category, Session ID, colour of the product, location on the page, model photography (i.e. front or side view) and a variable informing whether the price of a particular product is higher than the average price for the entire product category (i.e. 'price 2' in the table). Moreover, the number of clicks corresponds to the column name 'orders' in the dataset. Since all information was collected over the same year (2008), the feature 'year' was disregarded. The dataset includes over 165'400 rows and does not contain any missing values.



e-shop clothing 2008

year	month	day	order	country	session ID	page 1 (main category)	page 2 (clothing model)	colour	location	model photography	price	price 2	page
2008	4	1	1	29	1	1	A13	1	5	1	28	2	1
2008	4	1	2	29	1	1	A16	1	6	1	33	2	1
2008	4	1	3	29	1	2	B4	10	2	1	52	1	1
2008	4	1	4	29	1	2	B17	6	6	2	38	2	1

Figure 1: Dataset Sample

2. Data Cleaning, Grouping and Readjustments

As previously mentioned, the dataset came with big advantages. It contains over 160’000 rows with no NA values, as can be verified in the pre-processing folder called 0. analysing original dataset. Having done so, the group quickly realized that it was not as insightful as it appeared at first glance. It began when we all shared the accuracy rates of our different machine learning algorithm types, as none of them was able to deliver an accuracy of over 70%.

In a first step, we used all the features possible to make a prediction, but this led to very poor results. We thus started questioning ourselves on whether we fitted the algorithms in the wrong way. After having verified that this was not the case, we started to discuss which variables could be left out due to lack of or even inverse correlation with our target column “page visits” called orders in the original dataset. Given that our dataset covers entries from five months in one single year (2008), we hence decided to scrap the information related to the date.

Furthermore, because a few customers saw up to 195 products in one session, we concluded that we could not treat them as if the page visits came by the same number of individual visitors. Therefore, we aggregated the data in order to have one row per individual visitor. For visitors with more than one search, we kept the common information (the country) and kept the mean of the cost as well as the most frequent value of the other relevant columns (category, colour, model photography and page) among all the searches per session.

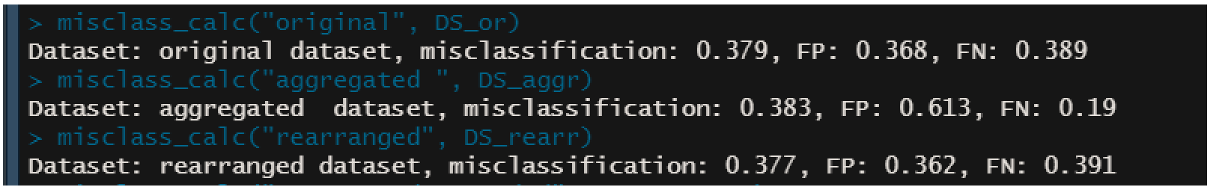
Despite this, the results did not seem to significantly improve. Given that we had already invested a considerable amount of time into it, we really strived to get a good prediction through the dataset. This led us to create a third version of the dataset. As the goal of our algorithms is to predict the amount of page visits given a product and some features related to it, we decided to group all rows containing the same configurations of colour, price and category and specify their count.

In the following sections, we are describing how we individually made use of the three types of datasets to train and test our algorithms. We filtered or manipulated the datasets according to our respective needs. Some of us also used a fourth dataset, which originated from the third dataset. It is an extended version, i.e., it takes more columns into account, which led to more unique rows with their respective count.

3. Logistic Regression

The script concerning the approach through the logistic regression can be found in the respective folder. It takes the data from three different datasets contained in the Data folder (the original one and the two elaborated versions). After transforming the amount of page visits binarily into 1s (higher than the median value, to signal a high amount) and 0s (for low visits), we fitted the GLM model to calculate the relationship between this value and the other ones available.

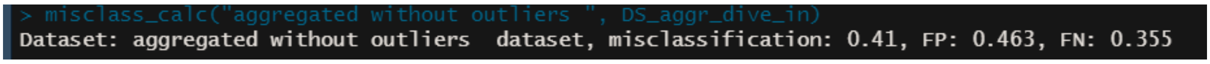
The following steps consisted in making predictions and assessing the percentage of misclassification, false positives (FP) as well as false negatives (FN). This was done for all three datasets. By outputting the three metrics of % of misclassification, FP and FN, we were able to compare the three datasets (as can be seen in the image below).



```
> misclass_calc("original", DS_or)
Dataset: original dataset, misclassification: 0.379, FP: 0.368, FN: 0.389
> misclass_calc("aggregated ", DS_aggr)
Dataset: aggregated  dataset, misclassification: 0.383, FP: 0.613, FN: 0.19
> misclass_calc("rearranged", DS_rearr)
Dataset: rearranged dataset, misclassification: 0.377, FP: 0.362, FN: 0.391
```

Figure 2: Metrics for the three datasets

Dataset two stood out, as it showed a high level of FP and a low one of FN. This could be explained by the fact that the median in that dataset was 4 and that the values of clicks per individual user session could be quite high in some cases. This could be seen by plotting a histogram. We removed the outliers (page visits with value <= 15). It helped in balancing out the relation between FPs and FNs, but caused the misclassification value to get higher.



```
> misclass_calc("aggregated without outliers ", DS_aggr_dive_in)
Dataset: aggregated without outliers  dataset, misclassification: 0.41, FP: 0.463, FN: 0.355
```

Figure 3: Metrics for dataset two after having removed the outilers

The comparison between the metrics of the three datasets showed that the last one was the most significant one, even though it did not make a big difference overall. For proper comparison with other models, we used the caret package to create a model that would run a k=10 fold cross validation. It led to the same metrics, confirming that the best accuracy received through logistic regression by using the selected dataset is 62.3% (1-0.377).

4. Neural Network

The high error in the linear and logistic regression models led us to believe that our f might be more complicated than initially assumed. We therefore decided to train a neural network as it usually helps yield better results. To achieve this, the package ‘neuralnet’ was used with one hidden layer containing three neurons (see diagram below). In addition to the value of the constants for the hidden and output layer, the optimal weights are calculated so as to optimize the result. Categorical variables may be integrated into a neural network through one-hot encoding. Nonetheless, we chose to only use numerical variables as features for this model for time efficiency purposes as the code already took a significant amount of time to run due to the high number of rows.

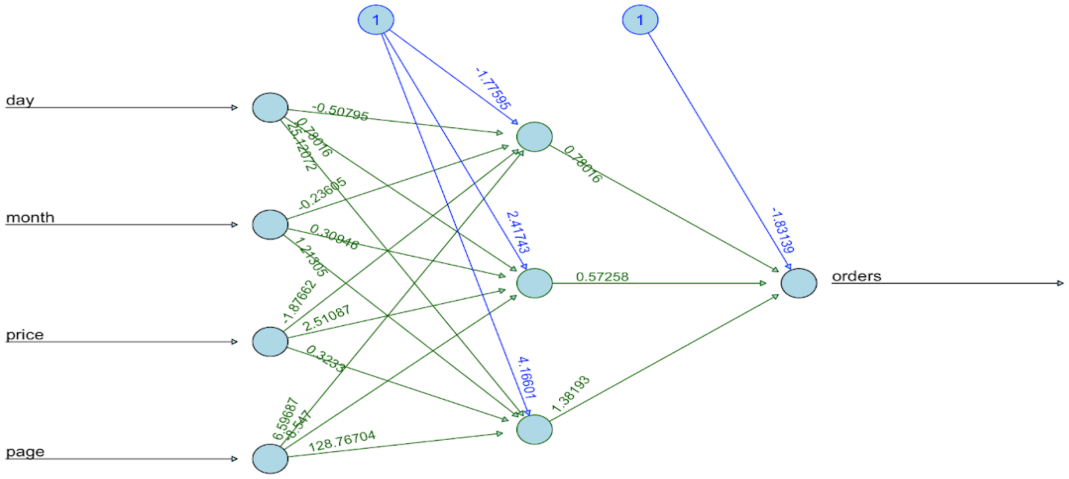


Figure 4: Neural Network Diagram

For the regression problem, the mean absolute error (MAE) was found to be around 5.5 which seemed quite promising at first glance. However, the graph plotting real versus predicted values strongly suggested that the chosen features are uncorrelated to our target variable and thus cannot serve as effective predictive tools (see diagram below). As for the classification problem, the neural network led to equally poor outcomes as it displayed a MAE of around 0.4.

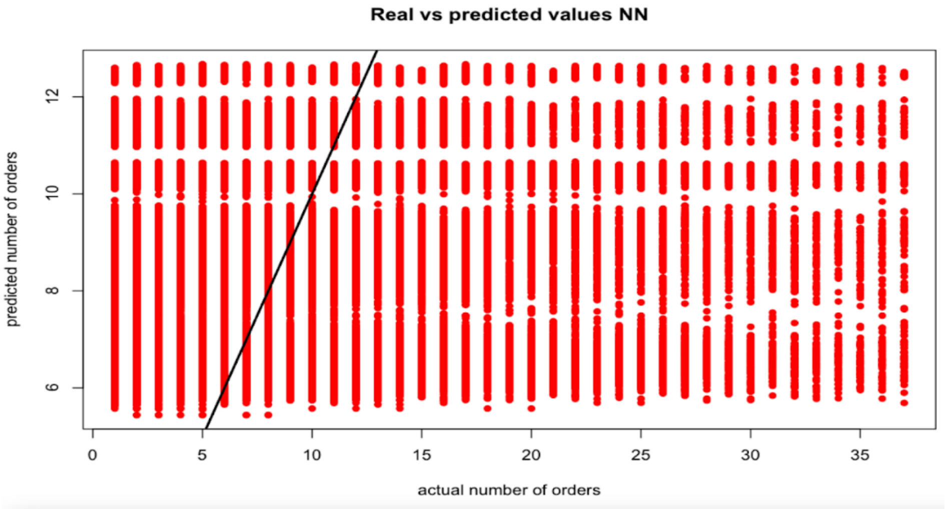


Figure 5: Validity of Neural Network

5. Random Forest

The Random Forest is generally quite a strong predictive model. It has the great advantage over traditional regression trees to be less dependent on the data, which means reducing the variance and is therefore much more useful when applying to new dataset. After splitting the data in a test and train part, the Random Forest builds many regression trees based on a sample of the train data with only a few features selected and takes the mean of the prediction of all these trees to give in the end a prediction with much lower variance.

It has some value that can be tuned, such as the number of features taken for each sample of growing trees. If all features are taken (bagging), the model will then always take the most correlated value and might miss a better combination.

For our prediction’s goal, the Random Forest is quite interesting since it can handle categorical variables, without turning them in a numerical value. It was used for a regression purpose. Unfortunately, the result is also there very disappointing. With an R-squared of about 40% and a Mean Absolute Error of 6.75 with a mean of the clicks of 9.78, the model is, therefore, not useful to make a good prediction. Even with tuning, the MAE and the RSME reduced slightly, but it was still not enough to use the model in a real case scenario.

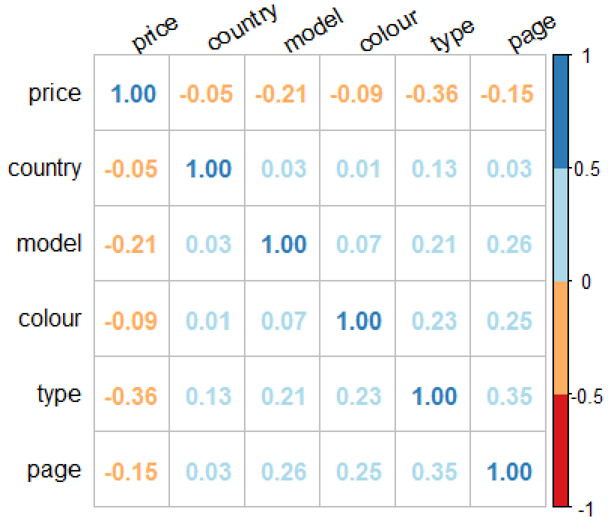
Using our aggregated dataset decreases the result of the MAE reaching 5.28 but the mean of the clicks also decreases to 7.88. The prediction thus improves slightly but not enough to have a useful model. Concerning our rearranged dataset, the results were even worse. Therefore, we did not tune both random forests.

In order to consolidate the result of the Random Forest, we tried a boosted tree. Since it also uses regression trees as a basis, the only difference is that it considers the previous results of the regression trees to build a new one. The result was comparable with the RF result and therefore confirms the first result. No tuning was done for this part as it was clear that the model would not be useful even with an improvement of the parameter.

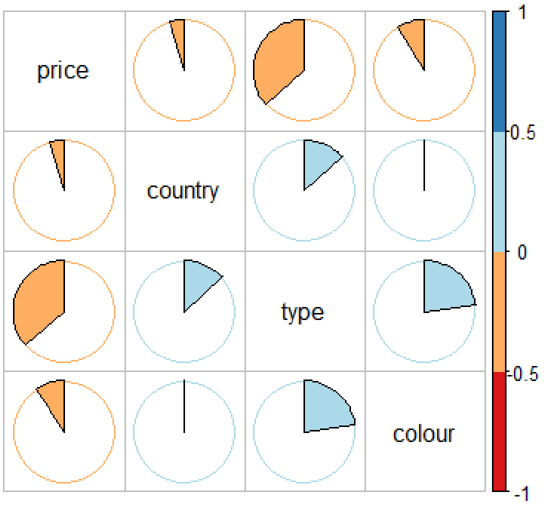
6. Linear and Polynomial Regression

The main goal of these two models was to find a regression capable of predicting mainly two features. These were the type of products[1] and the price. Alas, the results obtained were very poor with the best model obtaining an R2 value of 20% and an RMSE value of 11,21. As can be seen in Figures 6 & 7, the two features were selected since they have the strongest correlation among all features.

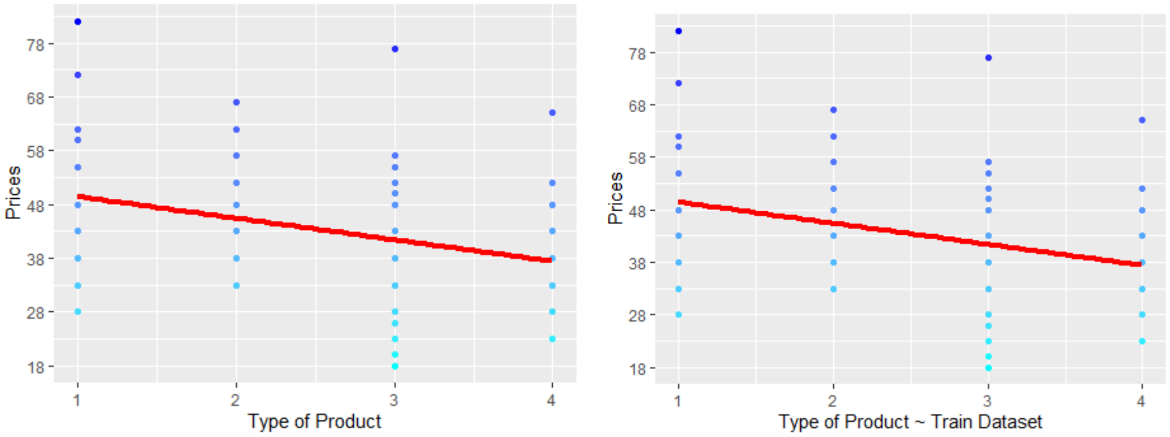
Through the linear regression, visible in Figures 8 & 9, which show respectively the dataset used for the regressions and the training dataset obtained from the previously mentioned dataset. It can be seen a decreasing trend within the four product types, a trend due to the fact that the fourth product category are the various sales made on the website. Moving instead on Figures 10 & 11, or the polynomial regression of degree, respectively, two and three, such a trend becomes far more evident. Unfortunately, as already said, the best model, i.e., the third-degree polynomial regression, has obtained an R2 value of 20 % and a value of RMSE around 11,21.



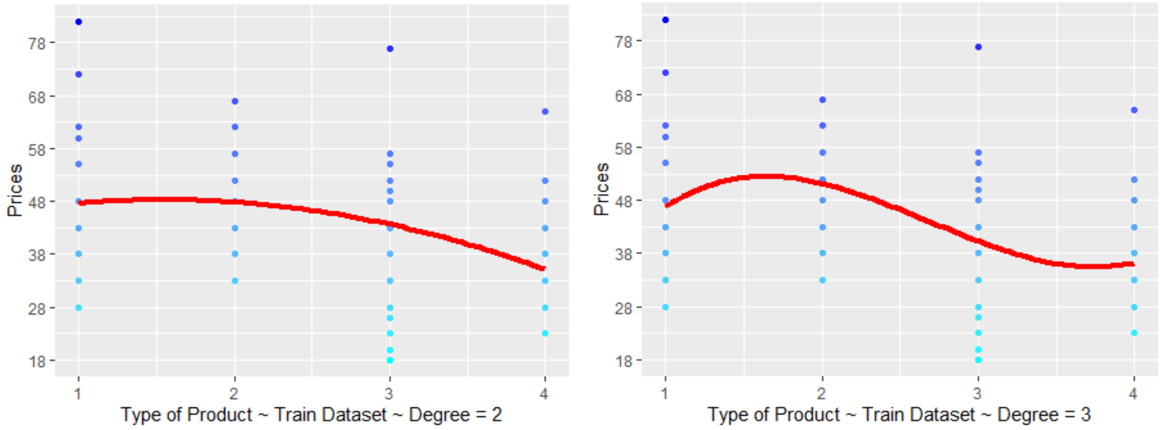
	price	country	model	colour	type	page
price	1.00	-0.05	-0.21	-0.09	-0.36	-0.15
country	-0.05	1.00	0.03	0.01	0.13	0.03
model	-0.21	0.03	1.00	0.07	0.21	0.26
colour	-0.09	0.01	0.07	1.00	0.23	0.25
type	-0.36	0.13	0.21	0.23	1.00	0.35
page	-0.15	0.03	0.26	0.25	0.35	1.00



Figures 6 & 7: Correlation Plots



Figures 8 & 9: Linear Regression ~ Type of Products and Prices



Figures 10 & 11: Polynomial Regression of Second and Third Grade ~ Type of Products and Prices

Such a uselessness of these models can be seen in Figure 12 which indicates the intervals of predictions of the “best” model. As it can be seen, the intervals of the Data Test are very large, and therefore such a model as already said has a very low chance to make any good predictions. In other words, it is essentially unusable in any business strategy approaches.

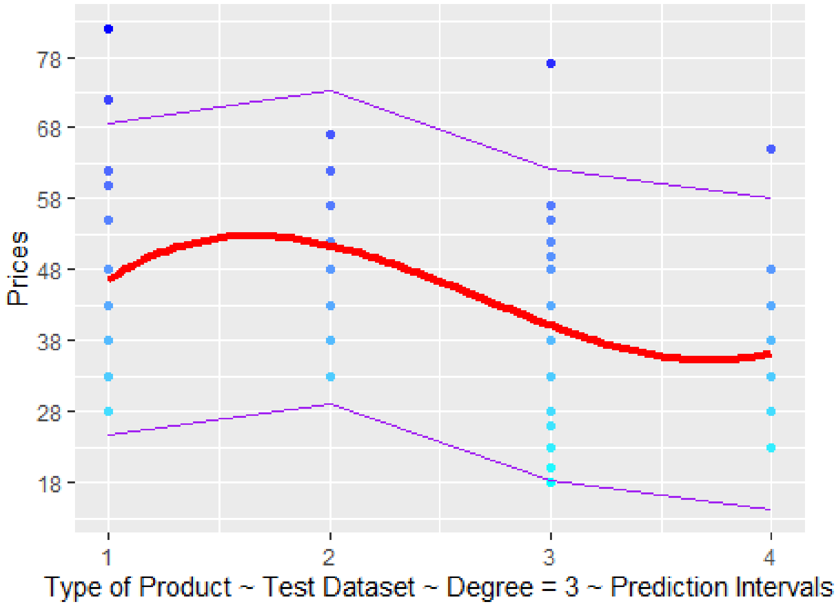


Figure 12: Polynomial Regression of Third Grade with Prediction Intervals ~ Type of Products and Prices

The only two conclusions which can be taken from such a graph is the fact that the second type of product, namely skirts, has a higher chance to have a higher price in comparison with the other two types of products, on the one hand. On the other hand, it can be seen that the third type of product, i.e., blouses, has the lowest prices, insight that can be seen from both the regression line and the dots plot.

However, in the process of finding "the best model", the group tried to use other features or to eliminate values to allow the model to become as reliable as possible. Unfortunately, as can be guessed, the third-degree polynomial regression model, obtained through the features type of products and prices was the "best" in that dataset.

[1] Type of product: 1-trousers; 2-skirts; 3-blouses; 4-sale

7. Further Outlooks

While our results were far from compelling, different methods do, however, exist which could have enhanced the validity and accuracy of our results. One of them could have been the use of the Prophet model, which adds a holidays effect to make the predictions better, this leads to a bad result in our case due to the short length of time in which the data were collected. Another method could be to cross-reference the data regarding births and future births over the period of the dataset and see how these influenced the trend of visits. Such collections of data could be obtained for the year 2008 now but at that time it was impossible to have the birth rate 9 months ahead. Another approach could concern the actual order of the products seen during the visit in the site, but such a value is unknown to us. We also tried to filter the target value to avoid extreme values and leave out customers with only a few clicks, but this does not improve our results significantly and makes models more difficult to implement in a real case. Having said that, it can be understood that the scarcity of the results of the various models could be improved, in order to have a more effective business strategy coming from such a dataset.

8. Business Application

All in all, our result does not provide the expected outcome. Furthermore, due to extremely poorly correlated data, it was not possible to achieve a useful prediction. If the company wanted to attain a more accurate prediction, they could look deeper at the time variable, for example, by collecting data over a period that would be long enough for them to successfully make use of the Prophet model. Recording the time spent on each page, the time passed between two clicks, or simply tracking in which part of the day the clicks occur could prove beneficial. It might be helpful to identify specific time zone with high traffic on the website and, thus, significantly improve clicks-prediction. We also encourage the company to precisely track where the clicks occur on the website. Finding whether the customer makes a click to quit, to look at an image, or to go to another page may also allow better features correlation and thus also improve the prediction. This dataset is not completely useless. From a descriptive statistics point of view, it provides some relevant information like where the customer is located and what is the most viewed product. However, it was not our goal to make descriptive statistics, and this was already done in a paper reporting about data mining applied on this dataset. But even for descriptive statistics the recording of the time and the position of the clicks that we suggested can lead to more meaningful conclusions, such as assessing the web traffic. The results could be applied to the business model and help the company in making better decisions for the future.

Declaration of Authorship

“We hereby declare

- that we have written this thesis without any help from others and without the use of documents and aids other than those stated above;
- that we have mentioned all the sources used and that we have cited them correctly according to established academic citation rules;
- that we have acquired any immaterial rights to materials we may have used such as images or graphs, or that we have produced such materials ourselves;
- that the topic or parts of it are not already the object of any work or examination of another course unless this has been explicitly agreed on with the faculty member in advance and is referred to in the thesis;
- that we will not pass on copies of this work to third parties or publish them without the University’s written consent if a direct connection can be established with the University of St. Gallen or its faculty members;
- that we are aware that our work can be electronically checked for plagiarism and that we hereby grant the University of St. Gallen copyright in accordance with the Examination Regulations in so far as this is required for administrative action;
- that we are aware that the University will prosecute any infringement of this declaration of authorship and, in particular, the employment of a ghostwriter, and that any such infringement may result in disciplinary and criminal consequences which may result in our expulsion from the University or our being stripped of our degree.

By submitting this academic term paper, we confirm through our conclusive action that we are submitting the Declaration of Authorship, that we have read and understood it, and that it is true.”

St. Gallen, 05 December 2021

Alex Sebastiani, Sacha Roduit, Alex-Paolo Marchese, Sophie Lejay