ML Midterm Project Report

Alvaro Guijarro May
226883@students.hertie-school.org

Niklas Pawelzik
221791@students.hertie-school.org

Justus v. Samson-Himmelstjerna
222918@students.hertie-school.org

Abstract

Women in the European Continent face different policies depending on their country of residence regarding abortion procedures. Some countries have a more flexible approach to this issue and offer a safer environment to the women involved, but unfortunately this is not the case everywhere [1]. In the case of German States, distance to neighboring countries, abortion policies in neighboring countries, fraction of population with migrant background, and foreign population in the area could be some predictors to identify future foreign or local patients for abortion procedures. In order to generate a Logistic Regression Model for Binary Classification, a dataset with the characteristics needed to solve this question had to be created from scratch. For the year 2021, in Germany there were approximately 100,000 registered abortion procedures performed throughout the 16 federal states, with less than 1,000 registered to be from foreign women. Correcting for this data imbalance, a prediction model with a balanced acccuracy score of 0.845 was achieved.

1. Proposed Method

[1] The data available to us contains information on abortions per year per state. It distinguishes 'foreign' abortion-seekers (without further information on the distribution of countries of origin), which enables us to create a binary classification model between 'foreign' and 'local'. Independent variables can be used to enable a Logistic Regression Model to estimate the probability that an event belongs to a specific class[2]. In the given context, a Logistic Regression Model is therefore a fitting tool to approach to solve our research question: Can independent variables containing information on population composition and neighbouring countries be used by our model in order to predict labels of 'foreign' or 'domestic' to abortions in the states of Germany? The independent variables we have decided to evaluate for each German State with this Logistic Regression Model are the following:

1. *Proximity of neighboring countries (geographical distance)* / Restrictiveness of respective countries (using abortion atlas scores)
2. *Foreign population per state (share of overall population)* / Restrictiveness of respective countries of origin (using abortion atlas scores). Also interesting for network effects (foreigners going for their abortion to states where they have relatives/social network; medical and social professionals speaking the abortion seeker´s native language)
3. *Population with a migrant background per state.* / Restrictiveness of respective countries of origin (using abortion atlas scores). Interesting for network effects (foreigners going for their abortion to states where they have relatives/social network; medical and social professionals speaking the abortion seeker´s native language)

The binary dependent variable we are trying to predict is the type of abortion seeker (foreigner / german citizen) that have a procedure performed on them. In order to predict this variable, we developed the following Linear Regression Model:

$$ln(\frac{p}{1-p})) = \beta_0 + \sum_{i}^{26} \beta_{ia} X_{ia} + \sum_{i}^{62} \beta_{ib} X_{ib} + \sum_{i}^{62} \beta_{ic} X_{ic} + \epsilon$$

The *a* group of variables involves the fraction of population of every German State with a migrant background (13) and the abortion atlas policy index (13) of European countries. Only migrant groups from 13 European countries surpass the statistically significant size to be included in the micro census data, i.e. smaller communities with migrant background could not be inlcuded. The *b* group of

[1] Here's a link to Alvaro's GitHub account https://github.com/Alvaroguijarro97/ML_Group_Project, Justus's https://github.com/jvsamson and Niklas's \unskip\protect\penalty\@M\vrulewidth\z@height\z@depth\dpff

variables involves the fraction of foreign population residing in each German State (31) and the abortion atlas policy index of the European country the group came from. The *c* group of variables involve the average distance betweet the German states' capital cities and the primary cities of each European country belonging to group *b* (31), as well as the abortion atlas policy index of of those countries (31). Every German State will have a different ranking for the information obtained for variable group *a*, *b*, and *c*, so taking this into consideration the order in which the variables appear for each group is sorted in ascending order, e.i: $X_{1a} < X_{2a} < X_{3a} < ... < X_{3n}$.

2. Experiments

Data used:

1. **Abortions by foreign and local population in German States for the year 2021.** The Federal Statistical Office of Germany published a data set containg information on abortions performed annually at the federal level. This data is subdivided into the individual annual quarters. Furthermore, information on the origin of the patient is listed for the individual procedures, whereby a distinction is made between the various federal states and the category "foreigners". The dataset contains information for all 16 federal states with an annual number of about 100,000 abortions performed in Germany from 1996 up to 2021. We have wrangeld that data in such a way, that we could easly acces for each feaderal state within Germany the fraction of foreign abortions in comparison to the abortions performed overall within Germany. [1]

2. **Abortion Index Score of European Countries.** This Index is an initiative by the European Parliamentary Forum for Sexual and Reproductive Rights (EPF) and International Planned Parenthood Federation European Network (IPPF EN). It is an in-depth analysis of abortion policies across Europe, which scores 53 European countries and territories in accordance with their legal frameworks in reference to their access to safe abortion care.

   The questions and structures for the Atlas were designed by a group of experts in sexual and reproductive health and rights. They came up with an overall score on a scale from 0-100 composed of four sections with several sub-categories: "access", "legal status of abortion care", "clinical care and service delivery", "information and on-line information". Since this data set is only published as an PDF-Document, we needed to extract the data points. Next we filtered the data for rows and columns that matter in the given context and created a datafram that could be compared to our other data.[2]

3. **Foreign Population Distribution of German States.** The statistics on foreigners of the Federal Statistical Office contain different representations of the regional distribution of the foreign population. Among other things, they list the federal states, sex, years of age and country groupings/citizenship. With the help of this database we hope to identify driving factors for higher shares of abortions performed on foreigners, checking for variables such as composition of foreign population or proximity to neighboring countries with stricter abortion policies. [3]

4. **German Population with migrant background in German States.** The Federal Office for Migration and Refugees developed in 2020 a microsensus[4] to approximately identify the different migrant groups German citizens have as background, derived from the country of origin of their respondents or parents. For each of the federal states in Germany, this dataset contains the number of citizens that have a migrant background from one of the EU 27 countries, and only the most significant groups were kept for further research, since some migrant population groups were so small that their share in the overall population of the state was negligible.

5. **Geographical distance of German States to European Countries.** Containing information for City Name, Longitude, Latitude, Country, State, Region, City Status, Population Density, Population, and Id of more than 4 million unique cities around de world, we created a database with the information needed to calculate the distances between the 16 German States and the different countries in the European continent. With this database we calculated the average distance between the German State's Capitals and all of the cities available in the dataset from each one of the European countries, in order to determine the distance between them.[5]

6. **Final combined dataframe.** In order to create the dataset needed to run the classification model, the transformed versions of foreign population, migrant background, distance to european countries, and abortion procedures, all organized by the 16 federal states, were combined and cleaned. This was done by creating a naming convention that could be merged with every single dataset. We were than able to unpivot all dataframes from wide to long format, making each observation one state/country combination, instead of each observation a state with several countries as variables per single observation. We than created place holder values within a new column "Country Score" in every dataset, that allowed us to insert for each observation the corresponding abortion score for the correct

country. Then we sorted those in such a way that they would lists for each state the respective foreign population, as well as migrant population, replacing the place holder country names with value from the abortion atlas. In the end we where able to combine all these datasets we the geographical data we had compiled for each federal state in relation to 31 sorrinding countries. The resulting dataset contained a data point for each one of the abortion procedures performed in 2021 on the different German states, segregated by an identifyer variable that differentiates between procedures performed in foreigners or locals.

After compiling the dataset needed to run the model, we identified that we were going to be working with an unbalanced dataset:

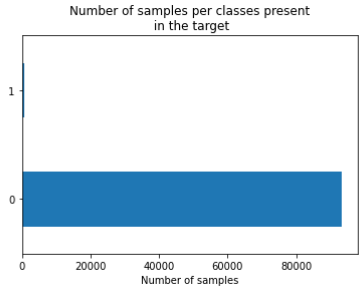

Figure 1. Dataset Distribution of Abortions performed on locals (0) and foreigners (1) in Germany

Almost 99% of the abortion procedures performed in German States were for local patients, so in order to be able to identify if we could explain the remaining 1%, a balancing of the dataset had to be performed. With the help of class weights, the skewed distribution of the classes for this dataset can be modifyed in the logistic regression model so the training model can penalize the misclassification made towards the minority class (foreign patient) by setting a higher class weight and also reduce the weight for the majority class (local patient). [4]

Evaluation method: For this Logistic Regression model, we are using *Accuracy score* and *Confusion Matrix* to evaluate our results[3]. Accuracy is used in order to compute how many times the classifier was right (predicted and actual value are the same) and divide it by the number of samples in the training set.

$$Accuracy = \frac{(TruePositive + TrueNegative)}{TotalSamples}$$

The Confusion Matrix is used in order to visualize the results of the prediction model, here we can observe the True Positives, False Positives, False negatives, and True Negatives our model has produced with its current configuration. Visualizing these results will help identify

Experimental details: In order to ensure that within our logistic regression model, all parameter estimates had converged properly, we increased the number of iterations allowed. We did this for both of our implemented logistic regression modells. However, for the unweighted logistic regression model we only needed to set the maximum interations equal to 1000, while for the weighted model, we opted 1500 allowed iterations. For further modeling it might be interesting to check out the iterations history and to check the loss values during the training time.

Since our dataset contains numeric variable values that are different in scale within our X values we decided to perform a standardization to have a common scale while building our machine learning model. Due to the fact that the most popular techniques for scaling numerical data prior to modeling is the StandardScaler we decided to start with that for our base line model. Because Machine learning models learn from combining input variables to an output variable, differences in the scales across input variables may increase the difficulty of the problem being modeled. For now we have only used the default configuration of the StandardScaler, wich centers as well as scales the values in each column. Next we could explore how our results would differ if we used a scaling transformer on our dataset.

Results: The performances of our weighted Logistic Regression Model can be currently compared against two alternatives: An unweighted Logistic Regression Model and a Dummy Model. Given the extremely unbalanced classes, the Dummy Model and the unweighted Logistic Regression Model turn out to perform identically: the Confusion Matrix for the unweighted model shows that the classifier assigned the ”foreign” label to none of the estimated cases, it effectively acts as a Dummy Model.

This explains the extremely high unbalanced accuracy score of 0.994: the model´s explanatory power is equal to the share of abortions performed on Germans. This imbalance is also reflected in the high values of the Dummy model in the Precision-Recall Curve. Lastly, the limited use of this model is reflected in its balanced accuracy score of 0.5: When it is taken into account how unbalanced the classes are, the unweighted model has no explanatory power.

In contrast, the Logistic Regression Model that includes the computanionally generated weights assigns both labels, as seen in the confusion matrix. As expected, abortions seeked by Germans are the biggest group in prediction and actual label. More than 40.000 abortions were correctly predicted as performed on Germans, 6.501 abortions per-

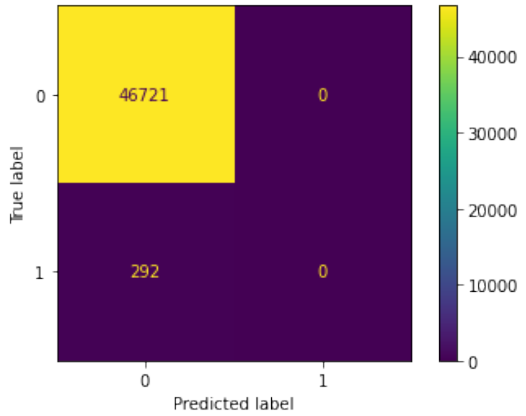

Figure 2. Confusion Matrix of the unweighted Logistic Regression Model with Abortions performed on locals (0) and foreigners (1) in Germany

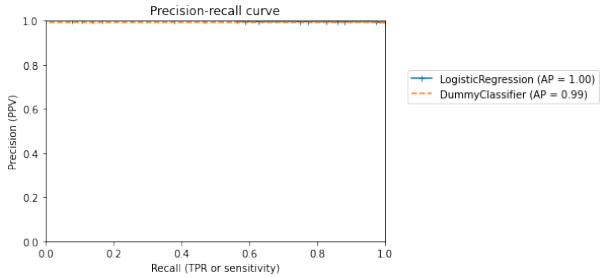

Figure 3. Precision-Recall Curve of the Logistic Regression Model and the Dummy Model

formed on Germans were labeled with the foreign predictor. Of the 292 abortions performed on foreigners, 242 were correclty predicted. While the performance on the Precision-recall curve is only slightly better than for the Dummy model (the maximum score of 1 limits the room for improvement drastically), the strength of these results is reflected in an accuracy score of 0.861 and a balanced accuracy of 0.845: a greatly improved score in comparison to the unweighted model.

Comment on your quantitative results: The fact that our independent variables can in fact be used to build a classifier model matches our initial expectation. When inspecting the raw data on foreign and domestic abortions, it becomes obvious that the two states closest to the Polish border had by far the most foreign abortions, especially taking into account that both states have relatively small populations. These numbers are in line with the general public perception of "abortion travel" by Polish women being a phenomen especially in the German North-East. The high accuracy score of the weighted model confirmed this expectation and is therefore in line with our intuition. We were positively surprised however by the fact that such a

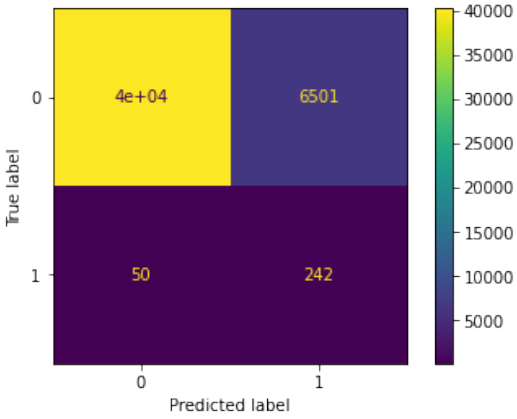

Figure 4. Confusion Matrix of the weighted Logistic Regression Model with Abortions performed on locals (0) and foreigners (1) in Germany

high share of the actual foreign abortions was correctly predicted by our model, especially given that all our variables operate on a state level only, due to the lack of local or individual data. The extreme difference between weighted and unweighted model does not come as a surprise given the extreme inbalance of the two classes, we anticipated that the unweighted model might behave like a dummy model as it actually does.

3. Future work

Overall, we are satisfied with the overall direction our project is taking. A promising point to continue the refinement of our model is to entangle how our independent variables interact and how we can improve model performance for instance by adding interaction terms and parameter modification. As described above, we expect the geographic variables (proximity to neighbouring countries) to be the strongest explanatory variables of our model. At the same time, we still see room for improvement in the operationalisation of the distance indicators. Currently, we take into account the distances between the German state capitals and the distance to the biggest countries in the neighbouring countries. As a first approximation, this seems to provide good results; however the explanatory power could probably be further improved by for instance including further cities per state or by factoring the cities population sizes.

Another approach to extract the full explanatory power of our data would be to test different scalers and additional models besides Logistic Regression Models. For example decision trees often perform better on imbalanced datasets like ours. This is due to the fact that their hierarchical structure allows them to learn signals from both classes.

References

[1] Europe abortion access project, 2020.

[2] A. Géron. *Hands-on Machine Learning with Scikit-Learn, Keras Tensor Flow*, chapter 14. O'Reilly Media, Inc.,, 2019.

[3] J. Jordan. Evaluating a machine learning model., July 2017.

[4] K. Singh. How to improve class imbalance using class weights in machine learning, October 2020.

Notes

[1] https://www-genesis.destatis.de/genesis/
/online?operation=table&code=23311-0006&
bypass=true&levelindex=1&levelid=1664961110565#
abreadcrumb

[2] https://www.epfweb.org/node/857

[3] https://www-genesis.destatis.de/genesis/
online#astructure

[4] https://www.bamf.de/DE/Themen/Forschung/
Veroeffentlichungen/Migrationsbericht2020/
PersonenMigrationshintergrund/
personenmigrationshintergrund-node.html

[5] https://simplemaps.com/data/world-cities