A Work Project, presented as part of the requirements for the Award of a Master's Degree in Business Analytics from the Nova School of Business and Economics.

Creating a Product to Segment Donors and Predict Donor Churn

AI Ethics in NGO's: Implications of Biased or Unfair Machine Learning in the Non-Profit Sector

Ana Raquel de Oliveira Bilro 43329

Work project carried out under the supervision of: Qiwei Han 17-12-2021

Abstract

The social sector is far from thriving nevertheless with the help of emerging computation technologies advancements can be done, compensating the scarcity of resources. With these systems, organizations can classify their donors and target those who prosper more return. Nonetheless, the use of these models should consider bias and fairness as central questions. In a sector focused on social good, the ethical challenges of machine learning are key. This thesis focused on a case study, done in Portugal 2021, but aims to take the conversation on bias and fairness further, addressing the technological challenges and limitations of machine learning tools.

Keywords – Ethics, Fairness, Bias, Machine Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

I. Group Research: Creating a Product to Segment Donors and Predict Donor Churn

1. NGO Presentation

This project uses AMI's data to create models capable of being generalized to other NGOs. AMI is a private Portuguese organization whose mission is to provide humanitarian aid and to promote human development (AMI: Vision, Mission and Values 2021). Although AMI has supported over 10 000 people in Portugal through 15 social facilities, the organization is globally oriented, having carried intervention activities in 82 countries. As per mentioned by the NGO's representatives, it is composed of a small technical team, most collaborators work full-time and have non-technical background and little experience in computer science.

2. Context and Problem Definition

AMI has around 70 000 donors - 93% and 7% companies - and a vast majority of Portuguese contributors (99%). Although the foundation relies on monetary donations to maintain its activity, there is a notorious absence of recurrence: donors are unlikely to donate more than once. The lack of regular donations, and consequential financial instability, challenges AMI's viability. To increase monetary contributions, the NGO should grow the number of active donors - those who not only registered, but also contributed monetarily -, the number of donations and the amount donated per individual. As a result, the following project scope was defined:

- Create a product to segment donors: group donors based on similarities. Consequently, the possibility of predicting fall-backs will be improved by virtue of a better

understanding of each donors willingness to donate and his/her characteristics. Furthermore, it will allow to identify the donors who are more likely to make regular donations and target them. Ultimately, this will improve the ability to retain donors and increment donations.

- Develop a model to predict donor churn: The model will render, for each donor, their probability of churn (the risk of not contributing to AMI again). This metric will be important to understand the willingness to donate, prevent churn, and allow more stability in the management of ongoing activities.
- Create an interactive dashboard for data visualization: This tool will help visualize data intuitivel, providing easiness of interpretation. So, AMI's managers will be able to extract lessons and make data-driven decisions.

3. Data and Methodology

3.1. Data Description

The information was extracted in September 2020 from the NGO's own Customer Relationship Manager (CRM), where inputs are made by all business areas and the outputs are used by the departments of Communication, Image and Sensibilization, and Marketing. It contains details on the individuals and companies registered to AMI since 1990, as well as information about the donations, including their value and type (monetary, volunteering, goods, or services). Following General Data Protection Regulation (GDPR) rules, all the data delivered was anonymized and, therefore, could be used with no restrictions. It is recorded in six excel files containing only active donors.

Before an in-depth review of these records, it is relevant to establish two important distinctions. First, AMI can receive both monetary and non-monetary contributions. The first will generate a monetary flow, therefore including sales, donations, and funding. For the latter, the value of

the offer is evaluated by AMI considering the time granted by the donor, in case of volunteer work, or the value of the good or service provided. Additionally, we should distinguish potential contributions from the ones that are finalized. AMI may contact previous donors to assess their interest in participating in new initiatives. If the donor does not respond, it will be classified as "potential".

3.2. Exploratory Data Analysis

With the intent of gaining a better understanding of the data provided, an Exploratory Data Analysis (EDA) was performed. The main purpose is to understand the data prior to making any assumptions that might affect future decision making. It focuses on finding relevant patterns, correlations, outliers, and mistakes. To simplify interpretation, visualizations were created using the matplotlib and seaborn libraries in Python. The inferred conclusions from the current analysis are useful to support the decisions on the procedures to fill missing data. The general approach to dealing with missing values started by defining a threshold of 70%. Following GDPR rules, AMI can only contact individuals who gave their consent - hence, records whose consent date is not Null. Consequently, AMI only has permission to approach 0.17% of their single donors, blocking the implementation of targeted communication. Since this is a legal requirement, it cannot be deleted.

(i) Analysis at Donor Level

AMI provided data on 62 032 single donors and 3 164 companies, made up of their characteristics and demographics. However, from a total of 31 variables regarding information on individuals, as per the established threshold, only 14 variables were considered usable. In reference to the companies, we preserved 10 out of the original 14 features. Thus, decisive characteristics that could help improve segmentation (individual's age or company’s sector)

were disregarded. The amount of missing data is one of the biggest obstacles to the project - addressed again in section IV - and must be considered throughout all decisions. There is a balanced distribution of donors by gender (51.8% men, against 48.1% women). We can also assume that AMI's appeal is mainly national: 99.75% of single contributors are Portuguese and only one company is foreign (Spanish). On average, an individual will donate 130€ to AMI in their donor journey, and a company 3 652€[1]. The total value offered by a company varies significantly, ranging from 1€ to 4 103 803.93 €. Most single donors (91.8%) are not registered as "AMI Friends", a loyalty program that encourages frequent donations, and 3.5% of companies partake in this initiative. AMI allocated every individual that did not have a known registration date to 1990, ergo the peak in enrolments in that year. Registration sources were grouped into 13 categories for a simpler analysis of its impact.

(ii) Analysis at Donation Level Potential Contributions

AMI initiated the recording of potential contributions once the NGO transitioned to CRM, in 2018. Since CRM implied a higher granularity and maintenance of information, data quality improved. However, pattern recognition is hampered by the low number of observations - there was only one full year of data (2019) at the time of the analysis. Most contributions are set as "Lost", most likely because AMI makes untargeted contacts, automatically creating leads that do not generate response. This enhances the value of the present project: targeted decisions on segments help trigger actions by the donors.

Finalized Contributions

[1] These values refer to the sum of all contributions ever made by a donor - single and company, respectively.

These records were either imported from AMI's old database, or already had a receipt associated, indicating the transaction was completed. These contemplate 170 750 logs from individuals between January 2004 and September 2020. As for companies, the team had access to information updated in February 2021 to improve model performance. It portrayed a total of 16 979 company contributions starting in 2004. Less than 50% of the individuals in AMI's database contributed at least once during the time frame mentioned. The maximum number of contributions made by a donor was 399. On average, each person donated 6 times and each company 4 times. However, most are one-time donors (61% of single donors and 57% of companies). The number of contributions per year has been tendentiously decreasing since 2011. GDPR rules restricted proactive contacts to subscribers to raise awareness and propose contributions, aggravating this trend. On average, a person will assist AMI 125 days after the previous contribution. Over 90% of donors had not made any endowment in the past 1.5 years, at the time of the analysis, coherent with the fact that most of them are one-time donors.

Over 99% of contributions from individuals (60% for companies) refer to donations -voluntary transfers of money. The average value of received from individuals is of 54€. If we consider only non-monetary contributions (goods and services, volunteer work, visibility, and recycling), this comes up to 718€. As for monetary, the average value collected is 50€. The average value of a giving made by a company is 1 137€ and has been increasing since 2018. Monetary ones are lower (932€) than non-monetary (1 489€). The number of contributions made by a donor and their value have a correlation of ~0.04, indicating independence.

3.3. Methodology

To achieve the goals propositioned in subsection 4, we will resort to ML algorithms. Their application requires an understanding of the data used, as well as its cleaning. The decision on

the best models requires the testing of multiple algorithms and the comparison of their outputs. The final verdict will be based on predefined measures - suitable to each algorithm - and important criteria, according to the project scope.

3.4. Pipeline and Data Preparation

(i) Data Cleaning

Once data ingestion was finalized and information from the different sources was joined, it was prepared for the model. Consolidation and cleaning of the data involved the deletion of outliers, including errors - for instance, donors registered in 1930 (prior to the NGO's foundation) - and numerical variables whose z-score[2] was higher than five. Missing values were filled according to context inferred internally and, when necessary, an external data source, the Iberian Balance Sheet Analysis System (SABI)[3] was used to get firmographics on the companies. We transformed the data types to meet the model's requirements and performed feature engineering, including upstream categorization (grouping values into wider categories) and the creation of additional variables (e.g., days since donors' last contribution). Finally, we resorted to the "scikit-learn" library to encode categorical variables (one-hot-encoding) and standardise numerical features (standard scaler).

(ii) Feature Selection

After feature engineering, 494 variables integrated the final table. Feature selection followed a multidimensional approach that combined analytical decisions - deleting attributes based on

[2] Numerical measurement that represents the distance of a value to the group's mean.

[3] SABI is a database with information on companies from Portugal and Spain.

the lack of data quality or high levels of correlation - and functional evaluation, based on AMI's experience. It is proposed as functional because it considers a practical interpretation of the problem, centring on an intuitive view of what characteristics should be key in a segmentation. At last, we tested different combinations of variables and selected the one that provided the best result.

4. Model Application and Evaluation

4.1. Segmentation Model

(i) Algorithms attempted

Since there were are no previously defined clusters, we can consider there is no target variable for the segmentation. As such, we tested several unsupervised learning models[4]. The first attempt used DB-Scan, a density-based clustering algorithm (Scikit-learn 2021). For this algorithm, a minimum number of donors is defined for each cluster. Similar donors are grouped into the same cluster, and those that are in regions with a density below the minimum threshold are marked as outliers. However, even with a low number of necessary neighbours for each cluster, most donors were classified as outliers, which is hardly interpretable. K-Means is a fast, widely used method that minimises the inertia (maximizes the cohesion within the same group), by partitioning the observations to the clusters with the nearest mean (Scikit-learn 2021). The algorithm was experimented with multiple combinations of variables and the elbow method (Soppin, Ramachandra and Chandrashekar 2021) was applied to derive the optimal number of clusters for each trial. PCA was applied before K-Means, for a two-dimensional representation of the model, facilitating the comparison of the performance of the different sets of attributes (Brownlee 2020).

[4] Algorithms that find patterns and train on data that does not possess a label associated (Jordan e Mitchell 2015).

(ii) Model Evaluation

The best option should be heterogeneous inter-cluster, whilst coherent intra-cluster, so that it is easy to characterize a donor based on the group he/she is inserted in. Separability and interpretability are key when evaluating the model. To ensure these, two-sample t-tests and one-way ANOVA were employed at each attempt. The former evaluates if feature values were different from one segment to another, by examining whether the differences are statistically significant (P-value< 0.05) (Webster 2013). The latter tests the null hypothesis that two or more groups have the same population mean. The silhouette-score - a similarity score where "the best value is 1 and the worst value is -1" (Scikit-learn 2021) - was also considered when comparing the distinguishability of the outcomes, i.e., whether the object resembles its own cluster and not the others. Finally, a homogenous size of the clusters was used as cohesion indicator.

(iii) Model Decision

The best model focusesmostlyon interpretability, to be user-friendly for the client.It outperformed the others for all metrics, but the silhouette-score. Standardizing the data can make separability more difficult, which could explain this lower score. The final model used 4 different clusters and a set of features[5] and characteristics heterogeneous enough to embark efficient marketing actions. Cluster 0 ended up with approximately 7k donors, 1 with 5k, 2 with 24k and 4 with 31k. Cluster 0 was found the persona most important to retain, as they donate more than once and a relatively high value in their lifetime as donors.

[5] Supervised machine learning retrieves an output from an input based on pair examples.

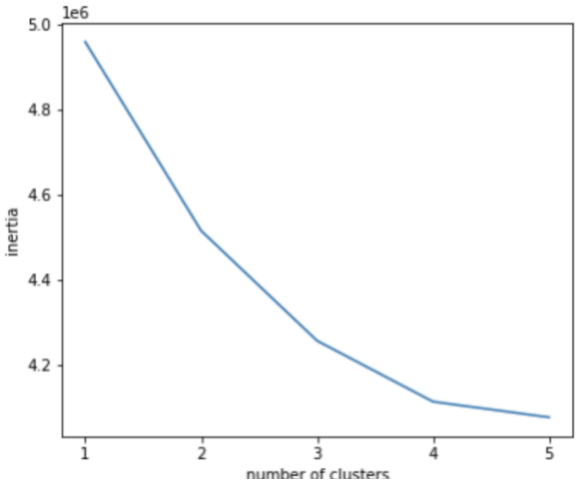


Figure 1 - Elbow method with selected features.

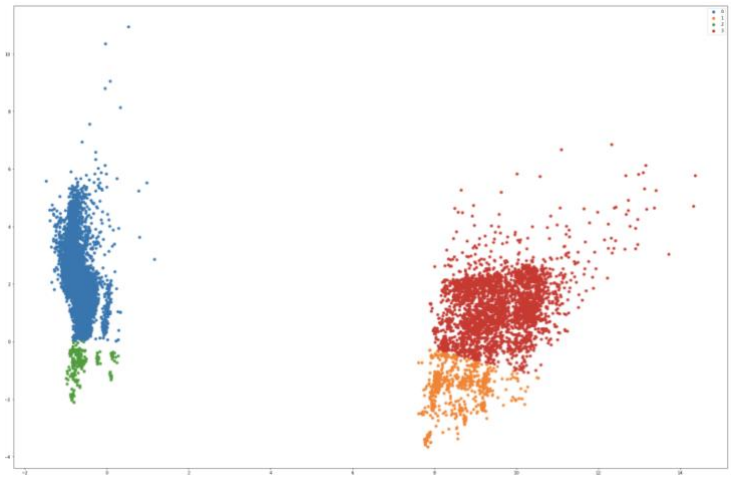


Figure 2 - Clusters from K-Means application (k=4) with selected features.

4.2 Churn Prediction Model

(i) Churn Definition

For exiting prediction, the team used a supervised mode[6] where the target variable should

[6] Included demographics (gender and region); firmographics (field and legal form); consent; information on the value, frequency, and recency of the contributions.

indicate whether the donor is churned. Since AMI did not have a definition of churn prior to the project, the team defined, together with the NGO, a list of rules for this classification. First, only people who contributed with a monetary donation at least once should be eligible for this evaluation. Then,

- If the donor only donated once, then the time since last contribution was taken into account. If the number of days since that donation was higher than the average number of days between contributions for all donors, the donor was classified as churned.
- If the donor contributed multiple times, the team compared the last contribution to the donor's usual frequency. If the number of days since the last time the donor contributed was higher than the average number of days between contributions for that donor, he/she was classified as churned.

(ii) Pipeline for Classification Model

Most of AMI's givers are one-time donors and, consequently, have churned. As such, our data is imbalanced. We employed synthetic minority over-sampling technique (SMOTE) (Imbalanced-Learning 2021) to generate the same number of observations for the minority class (not churned) as for the majority class(churned), so the model can learn better. Synthetic points are generated from the potential points that underly the pairs of actual data, by assuming they would follow a similar distribution, i.e., anything between 2 points from the minority class is likely to be from that class too. The resulting samples were used to train multiple classification models. The team used grid search to decide on the optimal values for a set of hyperparameters including the penalty to apply (L1 or L2), to avoid overtraining (Géron 2017) (Scikit-learn 2021).

(iii) Model Evaluation

The model is designed to identify new donors who are at risk of churn. The data was split into a training set and a test to evaluate the capability of generalizing the model to donors not used to train it. Accuracy and precision were chosen as the most relevant metrics to indicate the quality of the predictions. The first returns the percentage of correctly predicted points out of the total points, the second the percentage of points predicted as positive (churned) that are correct (True Positives). If a person is at risk of churn but is not predicted, AMI will apply to them undifferentiated actions. If they are wrongly predicted as churned, AMI will use part of their resources for targeted actions and increase their costs, thus False Positive cost is higher.

(iv) Model Decision

Since the results for the baseline model (Logistic Regression) and the more complex ones (XG-Boost [11]) were not significantly different, the team opted to use the Logistic Regression. This still allowed to get a precision of 99.9% and accuracy of 99.73% for the test set, while performing faster and implying lower costs than the other options. We should consider that the results of the model are highly influenced by the fact that the dataset usable for the churn prediction is very small. For the test set, a total of 5 285 donors were predicted as churned, versus 5 213 not churned (only monetary contributors considered). As AMI considered that the probability of churn would be more useful for their business decisions than a binary classification, the final output was changed to meet those needs.

(ii) Model Results

The results of both algorithms with the inclusion of additional socio-economic data were analogous to the previous attempts. Regarding segmentation, on the one hand, the frequency

of donations and their average value varied from cluster to cluster, resembling the results of the former model. The first cluster (0) included the donors who contributed more often (16 times on average) and the highest value (disregarding companies). On the other hand, the silhouette score rounded 0.65 and the clusters were virtually undistinguishable regarding their demographic characterization. For instance, when geographical features were used, Vale do Tejo was the most popular region (~30% of records) in all segments.

II. Individual Research: Fairness and Bias

1. Introduction

The present thesis was carried out as the final proof for the achievement of the Masters in Business Analytics diploma by NOVA School of Business and Economics. The project was made under the supervision of Professor Qiwei Han, in collaboration with the Portuguese non-profit organization Assistência Médica Internacional (AMI), alongside the advisory of faculty members and external consulting from Accenture.

Although there have been many advances on technological tools that allow targeted actions, they are not often used for the social sector, with most campaigns running ad hoc. We can see that in other sectors, such as the private one, accuracy in campaigns is prioritized (Bodenberg & Roberts, 1990; Gönül & Shi, 1998). This is often due to the discrepancy of data in both sectors, notwithstanding as data collection increases in Non-Governmental Organizations (NGO) it becomes possible to process that data and extract conclusions. The first and foremost purpose of the present work is to show how machine learning can be applied to automate donor segmentation in the social sector in a manner that is useful for the organizations to increase their financial revenue by contributing to targeted marketing actions, and secondly to reflect upon how the ethics of these systems with special regard to fairness and bias.

2. Literature Review and Motivation

The present work was carried in Portugal year of 2021, it is thereupon of foremost importance to have a characterization of the social socioeconomic paradigm of the country in contemplation of understanding some of the key aspects than lead to the conclusions of the project. Because there is no consensus on the definition of a non-governmental organization, the statistical findings differ from study to study, hence the following figures refer to different populations. According to Statics Portugal, there are around 27 985 Social Entities[7] in Portugal, with an average of 5,2 workers per organization (Alexandra Esteves 2015), a low number considering the significative existence of units with a big dimension (i.e., Holy houses of mercy). The latest Social Economy Sector Survey Statistics Portugal[8] (2019), shows that the organizational structure are usually 1 or 2 hierarchical levels and most of them have reached seniority (20 years old or more). Top managers, as in the person who occupies the highest hierarchical level, are mostly elected (79,9%) and only slightly more than a third (32,9%) consider themselves as "moderately autonomous" in the use of information technologies. With regards to employees in the sector, almost a 1/3 receives the national minimum wage. Concerning the operations, almost half (45,8%) of NGO's stated not to use key performance indicators for evaluation / monitorization of their activities. Close to 93% of these entities did not use methods of measuring social impact (2018). To get a grasp of the Portuguese situation it is also important to mention the European and global paradigm to be able to contextualize the social sector of Portugal. We can see that the United States of America is highly more

[7] "Social Economy (SE) entities active in 2018, headquartered in Portugal, (...) including 5 large families: Cooperatives, Mutual associations, Holy houses of mercy, Foundations and Associations with altruistic goals (AAG)."

[8] "Statistical operation within the scope of the National Statistical System (NSS), which emerged following the realization, in 2017, of the Management Practices Survey (MPS) of non-financial companies."

founded by private donors, in contrast to Portugal. This is an especially important point to mention, considering the need that NGO's have of donations to sustain their activity.

Being the scarcity of resources one of the most mentioned issues of Portuguese social sector, the use of such is of high interest. Most organizations are understaffed and in a disproportional ration to their donors. This means contacting donors should be done in the most efficient manner cost wise avoiding wasteful expenditure. We have seen how the use of Artificial Intelligence (AI) can help segment donors making it possible to profile them, ergo improving retention and acquisition rates ((Aaker 2009), (Aquino 2002), (Sargeant 2011)).

Notwithstanding the use of Artificial Intelligence should not be done without consideration to bias and fairness of the Machine Learning (ML) system. It is relevant to consider the structure of the NGOs, who are the providers of the data to produce these models, when evaluating their outcome. In fact, a study, drawn up by the Catholic University of Portugal for the Calouste Gulkenkian Foundation (Franco 2015), finds that most NGOs are led by highly educated volunteer middle-aged white men, and the statutory boards are not very open to external participation and scrutiny. On the other hand, most paid workers are female who often lack training to perform their activity.

The publication of works in the ethic field concerning machine learning and algorithms is ever growing with the increased use of such technology, it is however important to mention that this is a new field and that the concept of bias and fairness are relative concepts, and their definition is highly dependent of the context they are applied to. In sooth, we have been seeing multiple cases of how faulty models can be harmful for the population such as in the American penal system (Angwin 2016), as the Gender Shades project (Gebru 2018) on facial recognition systems or Google's machine translation (Kuczmarski 2018), evidencing the importance of this

measures as a last step in the evaluation of any ML model, including the ones proposed in the present project.

3. Methods for Bias and Fairness Evaluation

For the present project, following on the conclusions from the data analysis, we selected the variables within the dataset that could potentially lead to unfair or biased outcomes. To make the evaluations we resorted to existing libraries and focused on the Aequitas Toolkit.

Fairness is widely referred to as the ability of a model to be accountable on the treatment and impact (Dino Pedreshi 2008). In this work the fairness on impact was measured recurring to the use of measures based on parity. The model is fair when the proportion of predicted elements as positive of a certain label is the same inter-group (Pedro Saleiro 2019).

3.1. Potential discrimination within our data and context

For our dataset, there were five demographic characteristics to observe regarding their influence on the model regarding bias and fairness. These variables are “gender” (e.g., male), “region” (e.g., Lisbon), “education level” (e.g., masters, “job title” (e.g., Dr.) and “age” (e.g., 50). Unfortunately, due to high number of missing values the assessment for these parameters was only possible for two variables (“gender” and “region”), thus forward the study will focus on those two. This means that the full analysis of the model is compromised by the lack of data quality.

3.2. Aequitas Toolkit

The Aequitas Toolkit, created by the Centre for Data Science and Public Policy of the University of Chicago, is an open-source bias and fairness audit toolkit made public in 2018. It enables to test models for the two metrics in study: fairness and bias.

The Aequitas Toolkit was used to audit the machine learning system created for the project. This system allows to "look for biased actions or outcomes that are based on false or skewed assumptions about various demographic groups", consequently facilitating informed and equitable decisions in the development of algorithmic decision-making models, such as the one of the present research project. The toolkit works with a Python library as a command line, where the user configures the bias metrics for chosen variables which Aequitas uses to generate a report. These reports allow data scientists to compare models, but also policy makers to understand if a certain system is biased before implementing it.

3.3. Selection of Metrics for Bias and Fairness Assessment

The machine learning model is one that produces a label for each donor, in this case the labels attributed can be "churned" or "not churn". In pursuance of selecting the best metric to evaluate fairness, we used the fairness tree as shown in figure 1.

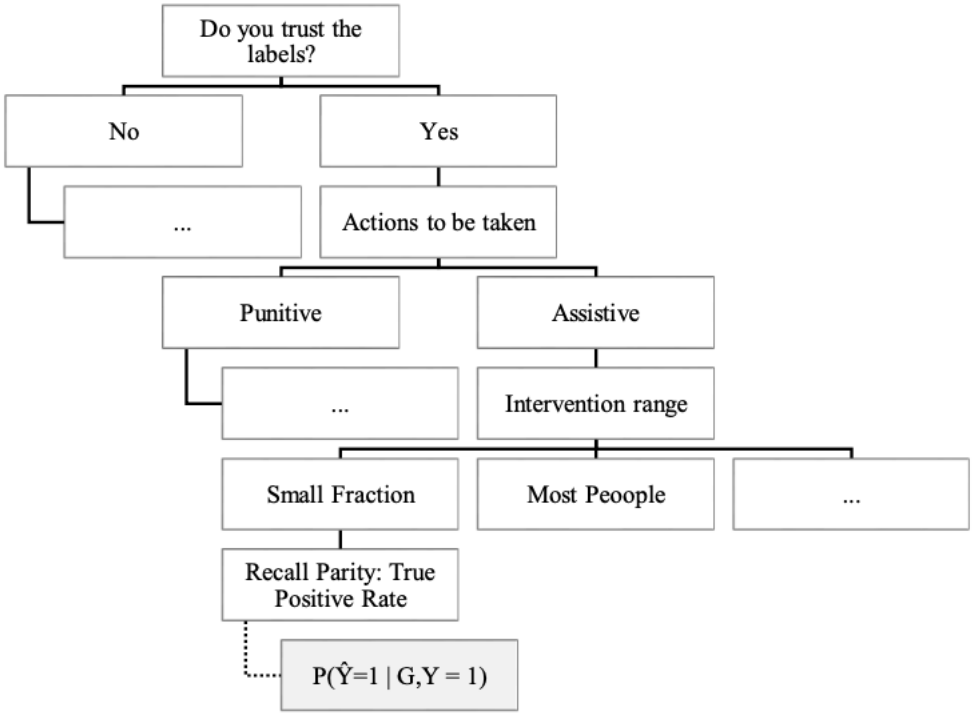


Figure 1 - Aequitas Fairness Tree Shortened

The first step is to define that we want our evaluation to be based on the errors of the system. A system is said to commit an error when a label (i.e., “churned” or “not churned”) is attributed to a certain individual but that does not correspond to the truth. Then, it is necessary to evaluate our trust on the labels. The churn definition is the one mentioned during the creation of the segmentation and churn prediction models. Given that the labelling systems was developed, in collaboration with the NGO, for this specific project they are to be trusted. Finally, we need to evaluate the nature of the intervention after the model outputs a result - assertive or punitive.

When we reach this point in the tree, we need to define whether or model is punitive or assistive. In our case, when a donor is classified as positive (likely to churn), AMI will allocate more resources to them to avoid that they churn effectively. In this sense, a positive

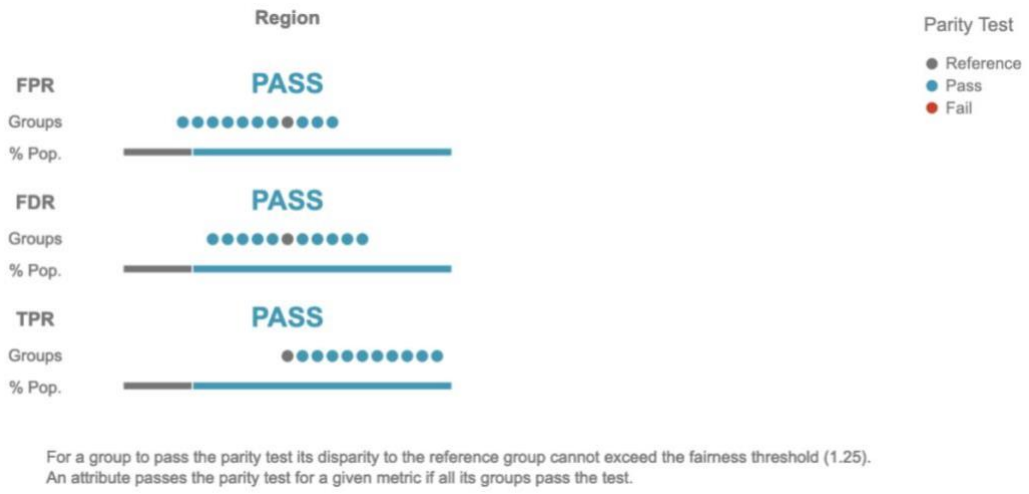
classification entails more support from the NGO and, for this reason, we should consider our model as assistive, which leads us to use the FNR to evaluate the system.

Not contacting a donor who is truly likely to churn (FN) impedes the possibility to change their mind and avoid them churning. Additionally, AMI can only take targeted actions for the donors that gave consent to be contacted, which is a very small part of the total donors. As a result, the costs of communicating with unnecessary donors (wrongly alarmed as likely to churned - FP) should not be relevant, confirming the importance of using the FNR.

1. **False Negative Rate** - $FNR = \frac{False\ Negative}{Labelled\ Positive}$, *meaning the ratio of false negative among a group of positively labeled.*

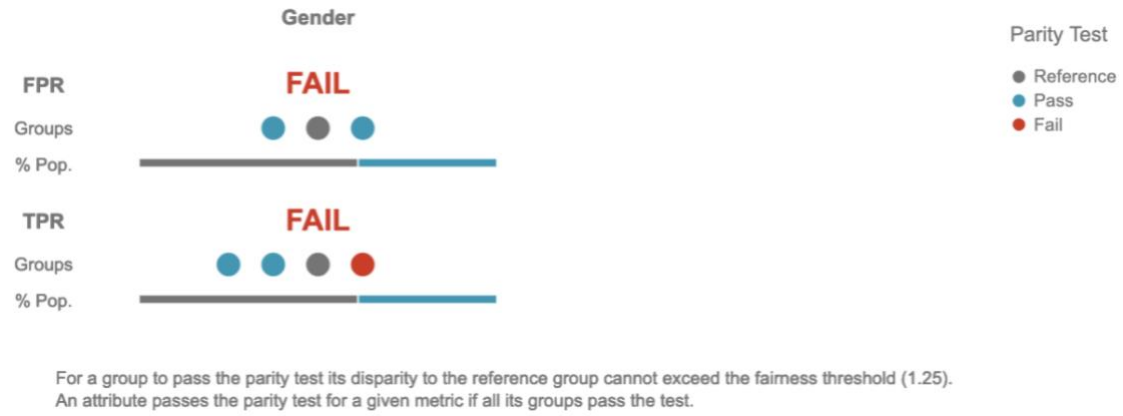
4. Results

As expected, the most heavily populated regions correspond to the ones with highest number ofdonorsandvice-versa. For the variable region the reference group to calculate bias measures used was Lisbon. The output results were as followed:



From the output of the toolkit, we can see that, for the variable "region" the model is not acting in an unfair form having had passed in the Parity Test for all three assertive metrics ("FPR", "FDR" and "TPR").

Regarding the variable "gender" the reference group used was male, and the output of the model were the following:



As we can see, the model, which accounts for three groups of the attribute "gender" (male, female, companies and other) failed on the parity test both for the False Positive Rate (FPR) and True Positive Rate (TPR). Nevertheless, TPR shows disparity ~1 for both female and companies, indicating fairness. This is probably happening because the group Other was the value used to fill missing values for the attribute "gender". Since the model seems to perform well for the cases where we have quality data, we will consider that it is being fair in its classification independently of gender.

5. Discussion

5.1. Limitations and their implications

After looking at the results of the project it is important to reflect upon the limitations that can influence the veracity of the conclusions reached at the end.

The biggest limitation of the project had to do with the data quality. The fact that there were a lot of missing values lead us to discard a lot of variables that could have had a strong impact on the model. Another important consideration is the filling of missing values for a lot of the attributes, this increases the risk of reaching false conclusion given that the analysis in being made relying on information with low quality. The method for the filling of the missing values chosen also has a big impact on the results.

Another important limitation to be consider is the "consent" variable, there is a very small poll of donors who allow AMI to contact them. This means that, following the rules of the General Data Protection Regulation (GDPR) in place in Europe, the organization will not be able to establish a relationship with most of the individuals registered in their database. Considering this even after having a profile per segment and the ability to tailor their actions towards such, the sample is composed by such low number of people that it might be difficult to assess the measures taken based on the characteristics of the clusters.

Finally, it should be taken into consideration that the fairness and bias definitions are not absolute concepts thereof neither are the results of their assessment. Different libraries can, because of this, render different outputs.

If the model automatically wrongfully predicts someone as "churned" just because they belong to a certain group, this will trigger a targeted action by the NGO. However, since this action in based on incorrect information this an inefficient resource allocation. The model will keep learning from that information and strengthen this biased decision. Additionally, e.g., if people of a certain region are always wrongfully predicted to be "not churned", AMI will probably stop investing resources regarding mass marketing in that area since that based on that analysis

it does not improve their financials. This howbeit, if a biased decision, will cause the NGO to lose potential donors.

5.2. Bias mitigation

Although the model is performing in a fair and unbiased manner, this evaluation should be done continuously. It is important to consider the implementation of measures that should be set in place to ensure the good maintenance of the system, especially the expected increase of

data dependency of organizations in the future. In the pre-processing phase, it is possible to add a parameter to the model to compensate the groups affected, ensure that groups are represented equally on training and test sets, remove the sensitive parameters, and perform resampling.

5.3. Donor segmentation model audit

As mentioned before this project is composed of two models: one of classification and another of segmentation. The bias and fairness audit were only performed on the classification part of the project because in unsupervised learning models there is no target variable, thus a confusion matrix is not calculated. The goal of the segmentation model is to group donors according to the characteristics. We want to take targeted actions according to their profile (including "gender", "region", etc.). If the model generates results that will lead to different actions for the different groups, the model is performing in accordance with its goal and should not be considered unfair.

6. Conclusions and Future Work

As shown previously, it is possible to determine that the social sector is still lagging when comparing to others in the use of recent technologies, in the computation field. Although the adoption of data systems is at an early stage the benefits of their use are clear, especially considering how automation and data processing can be an advantage for organizations with resource scarcity, a central problem in the sector. In terms of bias and fairness the topic is underdeveloped which takes part of a bigger problem of lack of formal systems in organizations. Undermining the importance of this issue represents more of a risk for the organizations themselves, who can lose donors and therefore donations as a result, than for the donors discriminated per se. According to our assessment, our model is not acting in an unfair nor biased form. Notwithstanding, the limitations of the NGOs present before and particularly

the lack of data can interfere with the conclusions, which might be misleading. As mentioned before, the methodology for filling missing values could be diversified to check the impact of those records. Given the subject nature of the concepts used through the work, experimenting libraries that have different definitions could be relevant as well to see how the assessment changes based upon them.

As a future work, it would be pertinent to study this metrics not only on other organizations of the social sector but also on other sectors and industries to compare how such topics are accounted for. For the non-profit sector the importance of data education seems to be ever growing. A lot of the project findings were compromised by the lack of data or lack of data quality. A correct data collection could have allowed for a much more extensive and truthful analysis.

Besides the barriers encountered during the project, the utmost significant conclusion was the realization of the magnitude of the impact that data can have on NGOs. Investing and integrating technology will clearly have a positive domino effect across society, yielding in surplus for all.

7. Works Cited

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.

Aaker, J. L., & Akutsu, S. 2009. Why do people give? The role of identityin giving. Journal of Consumer Psychology, 19(3).

Aquino, K., & Reed, A. 2002. The self-importance of moral identity. Journal of Personality and Social Psychology.

Sargeant, A., & Shang, J. 2011. Bequest giving: Revisiting donormotivation with dimensional qualitative research. Psychology &Marketing, 28(10).

Franco, Raquel Campos. 2015. Survey on the ngo sector in Portugal. Calouste Gulbenkian Foundation.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, Rayid Ghani. 2019. Aequitas: A Bias and Fairness Audit Toolkit. Center for Data Science and Public Policy University of Chicago.

Gebru, Joy Buolamwini and Timnit. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. Conference on Fairness, Accountability and Transparency.

Kuczmarski, James. 2018. Reducing gender bias in google translate. https://blog.google/products/translate/reducing-gender-bias-google-translate/.

Alexandra Esteves, Américo M. S. Carvalho Mendes, Ana Lourenço, Fernando Chau, Filipe Pinto, Francisca Guedes de Oliveira, Manuel Antunes da Cunha, Marisa Tavares, Raquel Campos Franco, Ricardo Gonçalves, Sara de Azevedo Garrido, Sofia Silva,Tommaso Ramus. 2015. Survey on the ngo sector in Portugal. Lisboa: Fundação Calouste Gulbenkian.