REPORT ON ALGORITHMIC BIAS ASSESSMENT

Prepared By SigmaRed Technologies Inc.

Version 1.0

February 7, 2024

CONFIDENTIALITY NOTICE:

This report and its contents are restricted and intended solely for use by authorized officials of The Ministry of Foreign Affairs of the Kingdom of the Netherlands (NL MFA). Any unauthorized disclosure, copying, or distribution of the information herein is strictly prohibited. If you are not the intended recipient, please notify the sender immediately and delete this report from your system.

SigmaRed

TABLE OF CONTENTS

Abstract

Overview

1) High-quality, knowledge-intensive decision-making on short-stay visa applications is the foundation of excellent consular service. The impartial and efficient processing of Schengen short-stay visa ("Kort Verblijf Visum," KVV) applications from travelers is paramount to advancing the interests of economic diplomacy, family visits, and tourism. The early recognition of potential opportunities and risks of applicants plays a pivotal role in positioning the Ministry of Foreign Affairs at the forefront of executing its responsibilities within the Dutch migration system. This proactive approach not only upholds the integrity of MFA visa processes but also significantly contributes to enhancing the security of the Netherlands by preventing illegal migration.

2) The BAO* Classification Model, employed by the Netherlands Ministry of Foreign Affairs, supports the processing of visa applications as part of the 'Information-Support Decision: Short Stay Visa' (IOB/KVV) process. This BAO Classification Model methodically assesses the opportunities and risks associated with visa applications, categorizing them according to their opportunity and risk profile to facilitate more efficient decision-making processes for short-stay Schengen Visas.

3) The BAO Classification Model's unique role necessitates a tailored approach to evaluation. Traditional metrics are not applicable, and alternative methods, such as causal testing and localized bias assessments, are employed to assess the model's fairness and effectiveness. It is imperative to maintain clarity on the model's advisory capacity to avoid misconceptions about its influence on visa application outcomes.

Technical Bias Assessment & Report

1) As required by the Ministry of Foreign Affairs, SigmaRed Technologies has conducted statistical bias tests on the BAO algorithm used for the efficient processing of Schengen Visa applications. The assessment includes calculating various bias assessment metrics, a comparative analysis of rejection rates, an initial causal inference analysis and interpretation of the computed metrics, and documentation of gaps and recommendations. It was also investigated whether profiles influence only the processing time or also affect the decision-making process of officers.

2) This report has been prepared based on the above assessment activities. The report explains the methodology used for technical bias assessment, applicable bias metrics, their values and interpretations, limitations of using some other bias metrics, insights of data, identified gaps, and recommendations.

***BAO** – BAO stands for Buitenlandse Zaken Analyse Omgeving

3) The report concludes with an appendix providing additional details, including a comparison between the BAO Classification Model and traditional supervised classifiers, supplementary tables and graphs from the exploratory data analysis and bias evaluation, a compilation of metrics deemed inadmissible, references, and a glossary.

Our Conclusion

1) To evaluate bias within the BAO Classification Model, we performed statistical bias tests on protected attributes on the given dataset which was up until January 2024.

2) Our findings indicate that, when normalized for historical data, there is no disproportionate discrimination based on age, marital status, or gender. However, we observed that applications with a Yemeni nationality (0.121% of all applications) were found to have a higher presence in risk profiles relative to their rejection rates, suggesting bias.

3) The BAO Classification Model does not generate predictions but categorizes applications based on predefined criteria. As such, there is no concept of a "wrong grouping," which renders metrics like False Positive Rate and Family Wise Error Rate irrelevant. These metrics presuppose a binary outcome of right or wrong, which does not align with the BAO model's function.

4) Our analysis reveals a notable correlation between BAO Profile and visa application outcomes, even after accounting for application information like Age, Gender, Nationality, and marital status. However, due to the absence of specific data, the correlation may be attributed to unobserved factors and we are not able to conclude that this correlation imply causation.

Summary of our recommendations

1) As correlation does not equate to causation, we recommend further experimental studies to evaluate whether the BAO Response impacts a visa officer's decision.

2) We also recommend establishing a comprehensive monitoring system for the classification model to promptly detect any shifts in data patterns or performance and bias indicators.

Introduction

The Ministry of Foreign Affairs of the Kingdom of the Netherlands (NL MFA) performs a core task of processing and evaluating short-stay Schengen visa applications. In an effort to streamline this process, the NL MFA employs a rule-based classification algorithm, which is instrumental in identifying opportunity and risk profiles among applications towards efficient processing and facilitation of bonafide applications. This algorithmic approach is designed to optimize application processing times for individuals deemed low risk, thereby enabling decision-making officers to allocate more attention to the more intricate cases.

The BAO classification model categorizes visa applications into opportunity and risk profiles which are then converted into tracks (fast track, regular track, and intensive track), as part of BAO response. The evaluation process considers the application's information and, where applicable, the details pertaining to their host in conjunction with the broader migration system partners associated with the NL MFA, such as the Repatriation and Departure Service (DT&V), the Immigration and Naturalization Service (IND), the Royal Military and Border Police (KMAR), and the Social Affairs and Employment Inspectorate (ISZW).

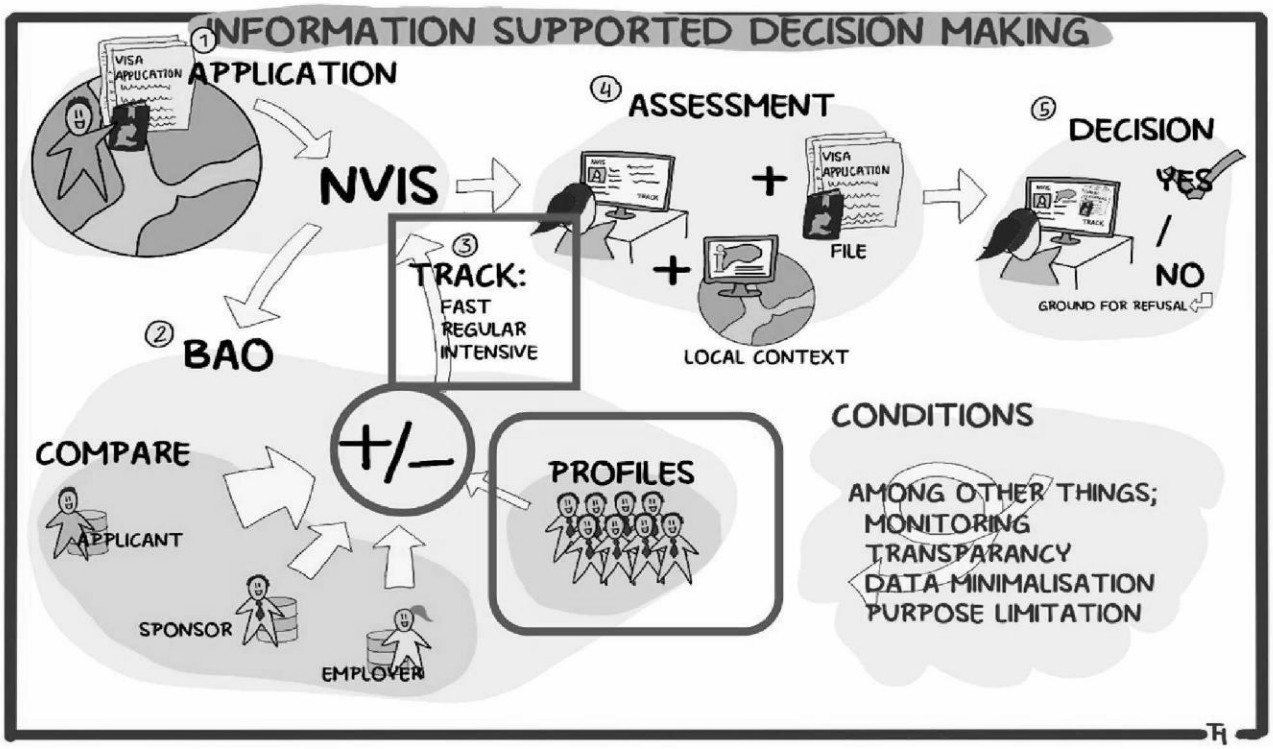
In adopting a classification model to enhance the efficiency of the Schengen Short-Stay visa application process, the NL MFA has implemented a series of management measures. This comprehensive set of measures underscores the principle that the classification model does not dictate the automatic approval or denial of visa applications and ensures the process's integrity. These measures are detailed further in the documentation made available by the NL MFA.

Among the management measures mandated by the NL MFA is the requirement for an Independent technical bias assessment. In response to this requirement, the NL MFA has engaged SigmaRed Technologies Inc (SigmaRed) to conduct an Algorithmic Bias Assessment of the BAO Classification Model. The specific parameters and objectives of this assessment are outlined in the subsequent Scope section of this report.

Scope

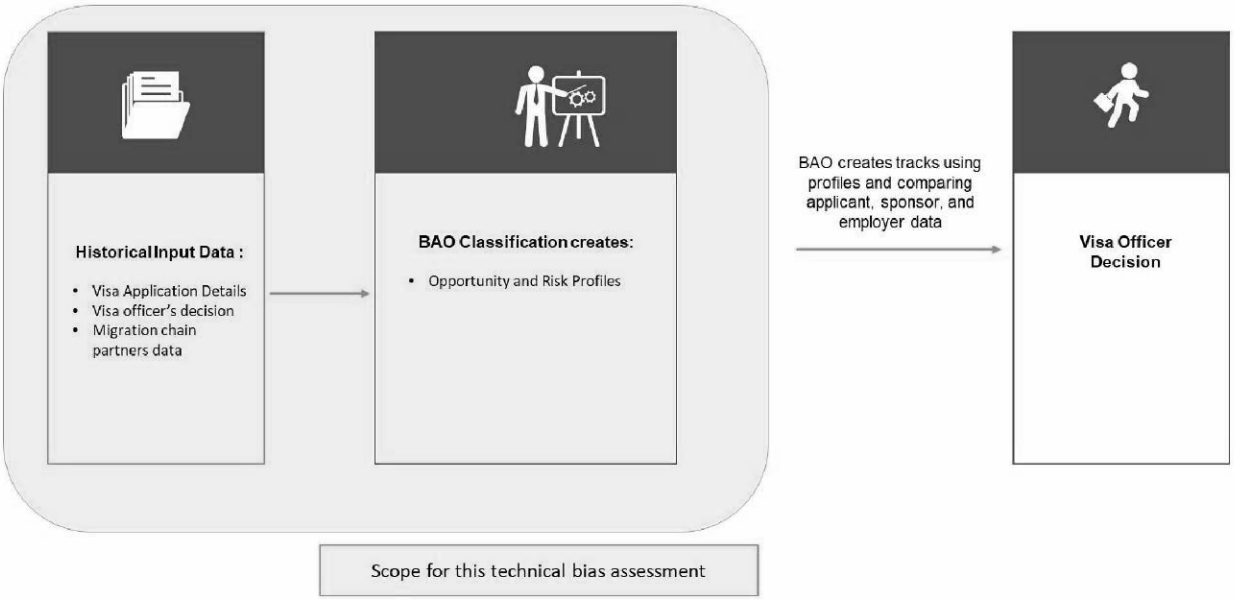
The scope is to perform a statistical bias test on the BAO algorithm applied during the assessment of Schengen Visa (BAO/IOB).

The following diagram shared by NL MFA gives an overview of the overall Information Supported Decision Making Process. As noticed in the diagram below the out of BAO response is the tracks (Fast, Regular, Intensive) and applications are mapped to one of these tracks and given to visa officers for further assessment and decision.



Source: NL MFA Team

The scope of this technical bias assessment is focused on the analysis of historical data and technical bias review of profile creation algorithm within BAO, as given in the scope diagram below. The details of in-scope and out-of-scope items are described in the following sub-section.



In Scope

1. Detailed technical bias assessment of the BAO Classification Model, including:
    a. Review of the model code and its underlying algorithm.
    b. Analysis of the applicable datasets:
        i. Historical Visa Applications (Input Data to the Classification Model).
        ii. Profiles generated by the model (Output Data from the Classification Model).
    c. Bias Analysis solely on Protected Attributes

2. Detailed Exploratory Data Analysis (EDA):
    a. Assessing the distribution of features in both input and output data.
    b. Calculating the percentage of individuals assigned to Opportunity and Risk profiles across various attributes used by the BAO Classification Model

        i. The main purpose of stay.
        ii. Place of application.
        iii. Nationality.
        iv. Gender.
        v. Marital status
        vi. Age group.
        vii. Occupation.
        viii. Visa application decision

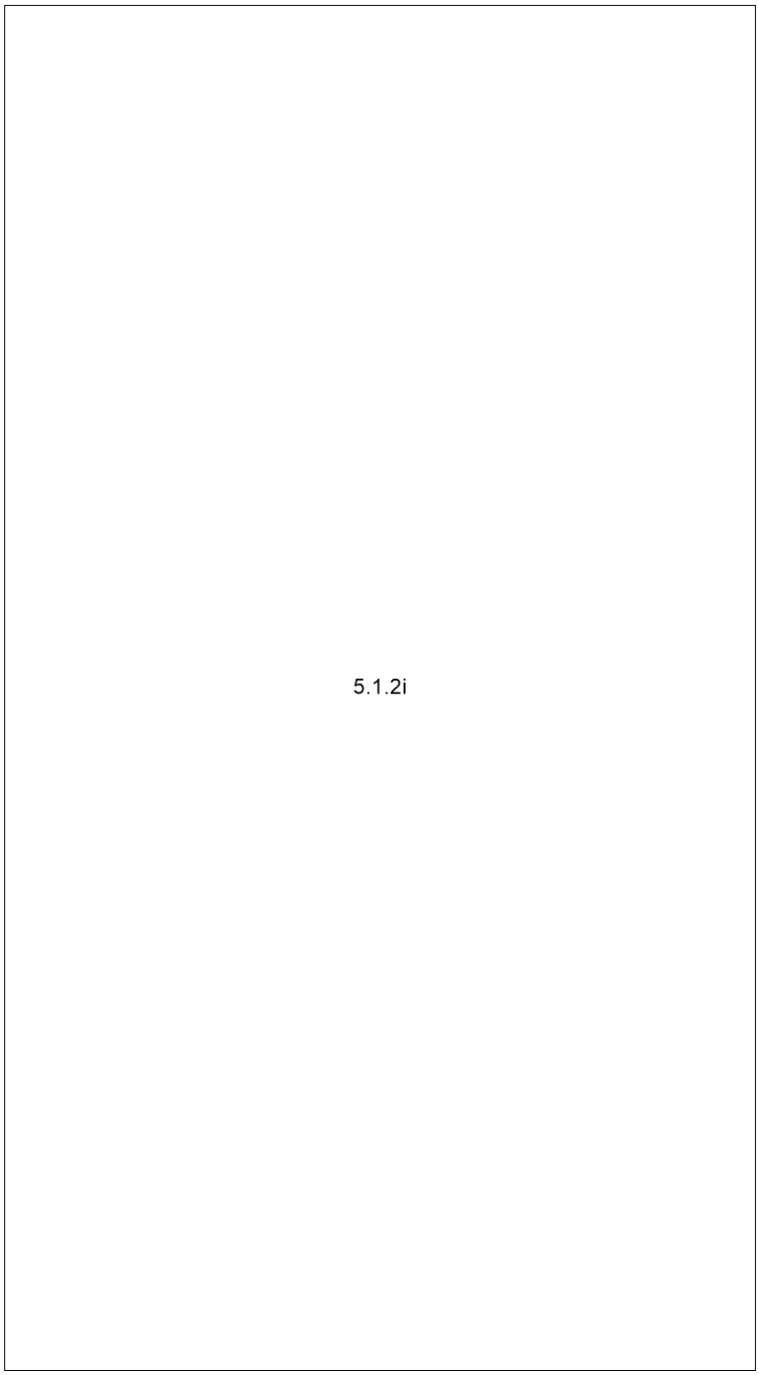
3. Bias Testing:

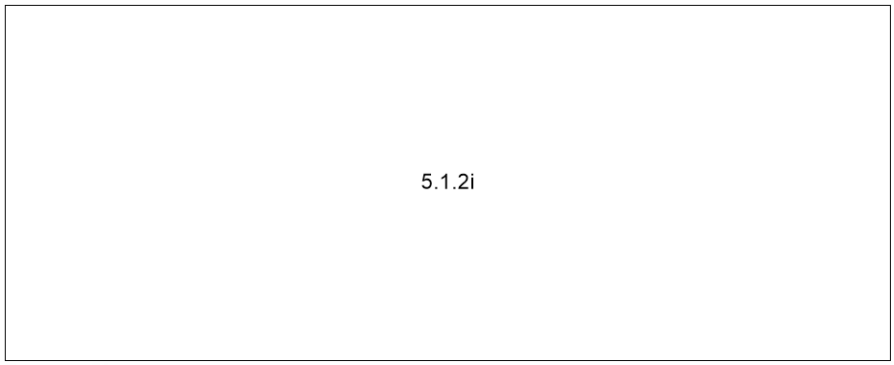
    a. Identify, develop, and implement techniques for bias evaluation, considering historical group-wise performance.
    b. Analysis of bias metrics derived from the bias evaluation techniques applied.
    c. Determination of infeasible bias evaluation strategies.

Dataset Features

Data for the classification model is collected based on historical data of the individual application files and information on potential "hits" from the NL MFA's migration chain partners. Components of the BAO, in addition to the profiles, include migration chain partner sources and the information from the application itself. All these components are weighed and subsequently lead to the final assignment of a track to a visa application by means of the BAO (fast, regular, intensive).

Dataset given to us for Bias analysis, contains the following features:





Out of Scope:

The following are out of the scope of this technical bias assessment:

1. Evaluation of bias attributed solely to the decision-making officer's discretion: The assessment does not cover biases attributed solely to the decision-making officer's discretion, considering personal biases beyond the scope of the algorithm.
2. Process Biases: The scope is only technical bias assessment and does not include other functional or process biases.
3. Legal reviews: Legal review concerning the algorithm's compliance with relevant laws and regulations.
4. Mitigation: Development and implementation of technical mitigation measures for any identified biases.

Description of the Current Classification Model

The current classification model employed by the BAO is a rule-based algorithm designed to assist decision-making officers by providing them with structured information. This model is not a predictive tool but rather a system that groups visa applications into predefined profiles based on specific features. This model then becomes a part of the BAO Response which outputs three tracks: Fast, Intensive and Regular that designate the intensity of the case to be handled. Hence, the model does not directly take or propose visa application decisions.

Below is a detailed description of the classification model and its operational framework.

Model Creation

BAO's classification algorithm follows the rules established under the foreign ministry's guidelines. The model does not learn from past cases, as per its design, but is built on a set of fixed rules and does not evolve with new data. The BAO model is used to create profiles that represent groups of applications sharing common characteristics. When a new application matches the profile's characteristics, it is classified as an opportunity or risk profile. This model then becomes a part of the BAO Response which outputs three tracks: Fast, Intensive and Regular for visa officer consideration when the application is processed by visa officers.

Data Collection Sources

Data for the classification model is collected based on historical data of the individual application files and information on potential "hits" from the NL MFA's migration chain partners. The model does not utilize a learning algorithm and does not adapt based on historical case outcomes. Instead, it operates within the confines of the established guidelines.

Components of the BAO, in addition to the profiles, include migration chain partner sources and the information from the application itself. All these components are weighed and subsequently lead to the final assignment of a track to a visa application by means of the BAO (fast, regular, intensive).

Existing Safeguards

The classification model is designed with several safeguards to enable fairness and compliance. Notably, the BAO does not determine the outcome of a visa application. Any decision to reject an application must cite grounds for refusal as mentioned in the EU Visa code, ensuring that the classification model's output is only utilized to determine the amount of time spent on a particular application. Additional safeguards are present in the NL MFA's documentation of the BAO Classification Model.

Model Classification

The model categorizes applications into risk and opportunity groups based on hit rates and rejection rates. These groups are defined as follows:

1. Opportunity Group: Applications with a refusal percentage lower than 5% and a hit rate lower than 0.25%.
2. Opportunity Group 2: Applications with a refusal percentage lower than 2.5% and a hit rate between 0.25% and 0.5%.
3. Risk Group: Applications with a hit rate higher than 5%.
4. Risk Group 2: Applications with a hit rate between 1% and 5% and a refusal percentage higher than 16%.

In practice, the two Opportunity Groups are consolidated into a single category, as are the two Risk Groups. This simplification streamlines the decision-making process while maintaining the integrity of the risk assessment of the BAO Classification Model.

One point of concern, however, is that the Classification Model utilizes rejection rates to perform the grouping. However, if the decision-making officer bases their decision on the BAO Response itself, it could lead to a negative feedback loop by causing intensive track applications to get rejected more, hence being classified into intensive track applications again. To investigate this further, we have conducted a Causal Inference Test, which is described in detail later in the document.

Further information about the comparison between BAO Classification Model and traditional supervised classifier models can be found in the Appendix.

SigmaRed

Approach to the Evaluation of the BAO Classification Model

Introduction

This section of the algorithmic assessment report focuses on the evaluation metrics applicable to the BAO Classification Model. The model's role in decision-making is to assist officers without replacing their judgment. Consequently, traditional metrics that depend on true and predicted values are unsuitable for assessing the model's performance as explained below.

The BAO Classification Model does not generate predictions but categorizes applications based on predefined criteria. As such, there is no concept of a "wrong grouping," which renders metrics like False Positive Rate and Family Wise Error Rate irrelevant. These metrics presuppose a binary outcome of right or wrong, which does not align with the model's function.

Inadmissible Metrics

Due to the nature of the BAO Classification Model and its integration into the decision-making process, several conventional evaluation metrics are inadmissible. These metrics typically require a comparison between predicted and actual outcomes, which is not applicable in the current scenario where the model's output is not a direct prediction of visa decisions.

The following metrics, along with their formulas, are deemed inadmissible for evaluating the BAO Classification Model. Appendix A provides a detailed list of other inapplicable metrics.

SL. No	Metric Name	Metrics Overview	Why this metric is not admissible
1	Accuracy	In the context of tests or models, accuracy measures how well they can correctly identify or predict outcomes. It is a way to assess how reliable and trustworthy a model is in providing the right results. For example, let's say you have a medical test that is designed to detect a certain disease. If the test has an accuracy of 90%, it means that out of 100 people tested, it will correctly identify 90 people who have the disease and correctly identify 90 people who do not have the disease.	The BAO classification model is not a "learning algorithm" which does not predict the outcome. It only maps the applications into respective profiles and is not an automated decision-making system. Hence, the accuracy metric is not applicable.
2	False Positive	False positives refer to the number of incorrect positive identifications in a given situation. It means that a model wrongly identifies something as positive when it is actually negative. For example, if a medical test incorrectly identifies 10 out of 100 healthy people as having a certain disease, those 10 cases would be considered false positives.	To consider this metric, one needs to know the actual decisions against the predicted decision of the BAO classification model. As mentioned earlier, the BAO classification model does not" predict" any decisions and is not an automated decision-making system. It only maps the application into respective profiles, and hence this metric is not applicable.

3	False Positive Rate	This metric represents out of all actual negative points how many points are falsely predicted as positive. For example, let's say we have a medical test for a certain disease. If the false positive rate of the test is 5%, it means that out of 100 healthy individuals who take the test, 5 of them would receive a positive result even though they don't have the disease. Hence, the false positive rate is 5%.	Same as the above given for false positive rate
4	True Positive	True positives refer to the number of correct positive identifications in a given situation. It means that a model accurately identifies something as positive when it is indeed positive. For example, if a medical test correctly identifies 10 out of 100 individuals who actually have a certain disease, those 10 cases would be considered true positives.	Same as the above given for false positive rate
5	True Positive Rate	This metric represents, out of all actual positive points, how many points are correctly predicted as positive. For example, if a medical test correctly identifies 90 out of 100 people with a certain disease as having the disease, those 90 cases would be considered true positives, and the true positive rate is 90%	Same as the above given for false positive rate
6	F1 Score	The F1 score is based on the harmonic mean of Precision and Recall. These two metrics are based on True Positives, False Positives, and False negatives, as given below.	Same as the above given for false positive rate
7	Precision	Precision is the ratio of true positives to the sum of true positives and false positives, where true positives (TP) are instances correctly predicted as positive, and false positives (FP) are instances incorrectly predicted as positive by the model.	Same as the above given for false positive rate
8	Recall	Recall is the ratio of true positives to the sum of true positives and false negatives, where true positives (TP) are instances correctly predicted as positive, and false negatives (FN) are instances incorrectly predicted as negative by the model.	Same as the above given for false positive rate

On the same lines as given above, the following bias metrics are not applicable as well. These are described in detail in Appendix A.

a. Equal Opportunity Difference
b. Equalized Odds
c. False Discovery Rate
d. False Discovery Rate Difference
e. False Discovery Rate Ratio
f. False Omission Rate
g. False Omission Rate Difference
h. False Omission Rate Ratio
i. False Positive Rate Ratio
j. False Negative Rate Ratio
k. Average Odds Difference
l. Error Rate Difference

Infeasibility of Evaluating Bias by Solely Comparing Rejection Rate

Assessing algorithmic bias by comparing visa rejection rates before and after the implementation of profiles generated by the BAO Classification Model was considered. However, there are complexities and limitations associated with this approach, particularly that we can't attribute the changes in rejection rate before and after a profile was created to the BAO Classification Model or other externalities.

The BAO Classification Model is designed to assist visa officers by providing additional information through tracks (fast, regular, intensive), but does not replace their decision-making process. Consequently, evaluating the model's bias by comparing visa rejection rates against the tracks is not straightforward. The decision officer may not be convinced with the application details and may reject or ask for an interview for further review. This may also be due to various other variables that could influence rejection rates, including but not limited to changes in application volumes, shifts in geopolitical contexts, and alterations in immigration policies, which are not feasible to analyze as part of technical bias assessment.

Our preliminary analysis confirmed that the BAO Classification Model's scores generally correlate with visa officers' decisions. However, for example, if the rejection rate for a particular group was 3% before the creation of their profile but increased to 6% after the creation of the profile, one 'can't attribute that increase solely because of the presence of a risk profile. There could be many external confounders (such as the global political landscape, lack of funds, lack of evidence to prove the purpose of the visit, lack of ties to the home country, etc., which are not accounted for in the BAO) that can increase rejection rates and hence causes these results.

As such, we do not recommend using rejection rate comparisons as a standalone metric for evaluating the model's bias due to the complexities above and the risk of drawing erroneous conclusions.

Conclusion

The BAO Classification Model's unique role necessitates a tailored approach to evaluation. Traditional metrics are not applicable, and alternative methods, such as causal testing and localized bias assessments, are employed to assess the model's fairness and effectiveness. It is imperative to maintain clarity on the model's advisory capacity to avoid misconceptions about its influence on visa application outcomes.

Exploratory Data Analysis (EDA)

Overview

We have performed exploratory data analysis of the data provided to us. The relevant EDA analysis conducted is explained below, and further EDA details are included in the Appendix.

Data Preprocessing Steps

As given in the document shared by NL MFA titled "Lifecycle of NL BAO/IOB profiles" and based on discussions with the NL MFA team, the following selection criteria are considered for data preprocessing. Also, it is mentioned in the factsheet document that "for an applicant to be included in a profile, the applicant must be 18 or over and must have had to apply for a short-stay visa. The profiles are drawn up on the basis of at least 200 visa applications and several characteristics".

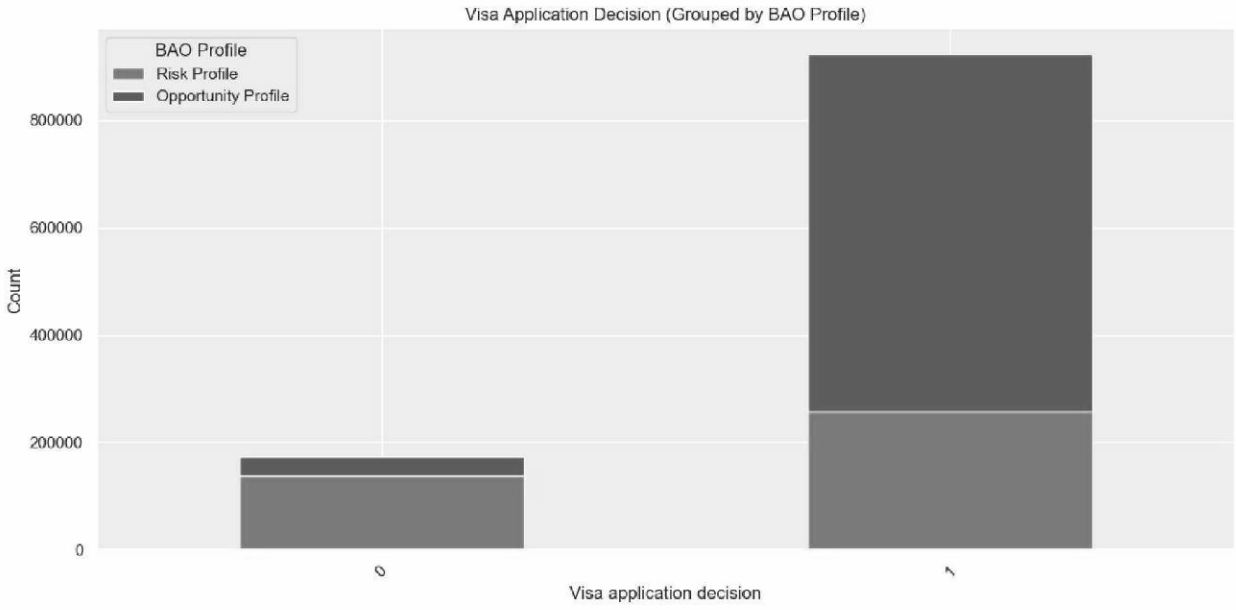
1. Only keep rows where there is an associated BAO Profile. Since our bias assessment approach calculates bias in the BAO classification model, we are not able to assess bias when no profile is assigned.
2. Only keep rows where the Visa Application Type is "C" meaning only short-stay visas are considered in our analysis.
3. Only keep rows where the visa application destination was the Netherlands.
4. Removing rows where Applicant Gender is "Onbekend", as there's only one application in that category.
5. Only keep rows where the application is above 18 years of age as mentioned by NL MFA team.
6. Grouping the VTBG (conditional acceptance) decision into the positive group as there are only small number of applications with VTBG values, and they are grouped into acceptance.
7. Only keep rows where there is a Visa Application Decision. We remove applications with no decision as it's required to calculate any of the causal effect or bias evaluation.*

* Only 0.5% (5,779) of all visa applications after following steps 1 through 5 contain a null visa application decision

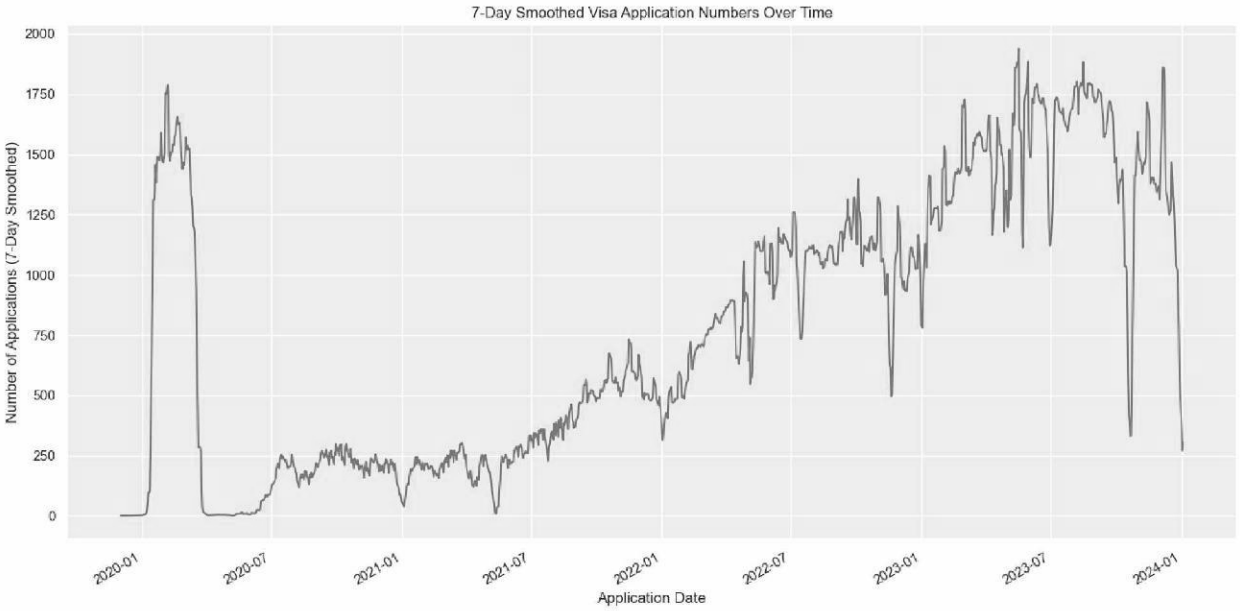
EDA Detailed Analysis and Interpretation:

EDA Detailed Graphs

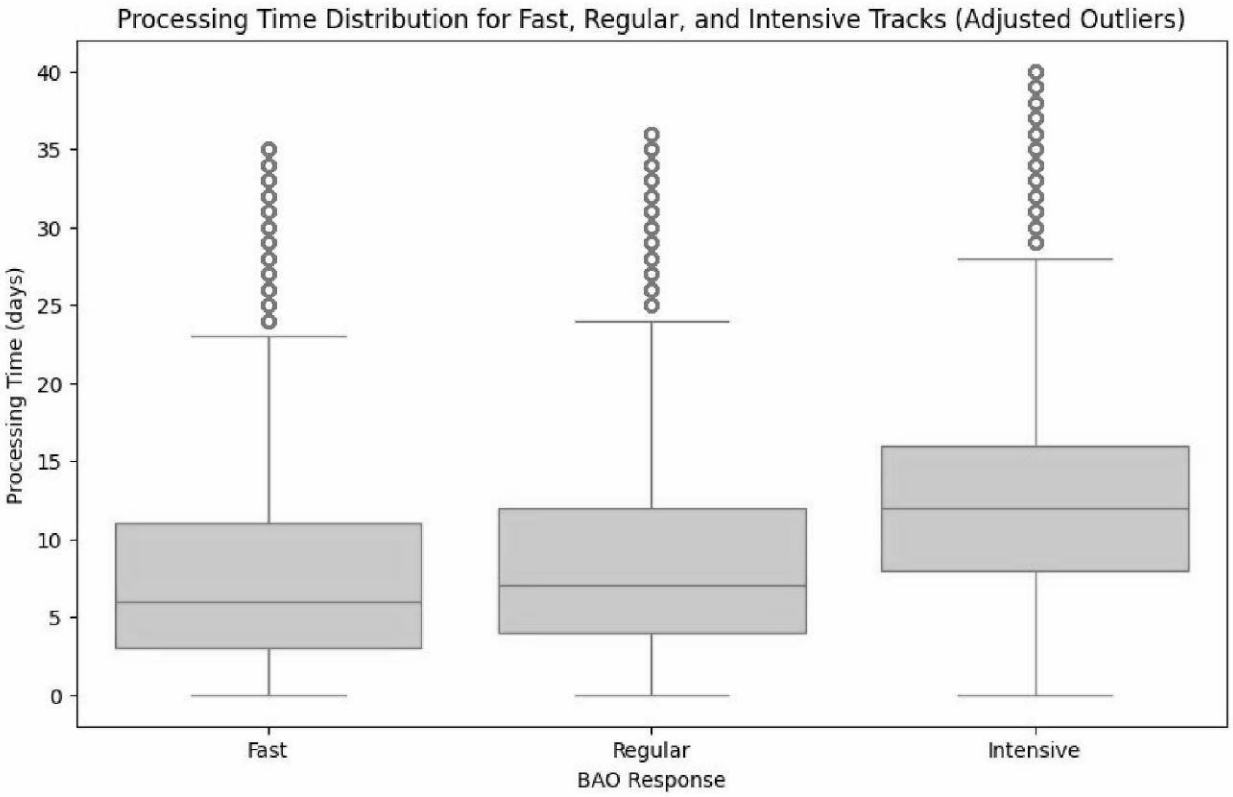
Visa_application_decision: There are 3 different visa application decision possibilities with 'Positive' being the most common (affecting 923,894 applications).



We observe a high number of positive BAO Profiles being associated with a positive visa decision (label 1) and vice versa, however, this statement 'doesn't prove causality as the reason for the positive or negative visa decision can also depend on other factors in the application. Furthermore, this graph reinforces that just because an application is placed into a Risk Profile, it 'doesn't mean that their application will get rejected and vice versa.

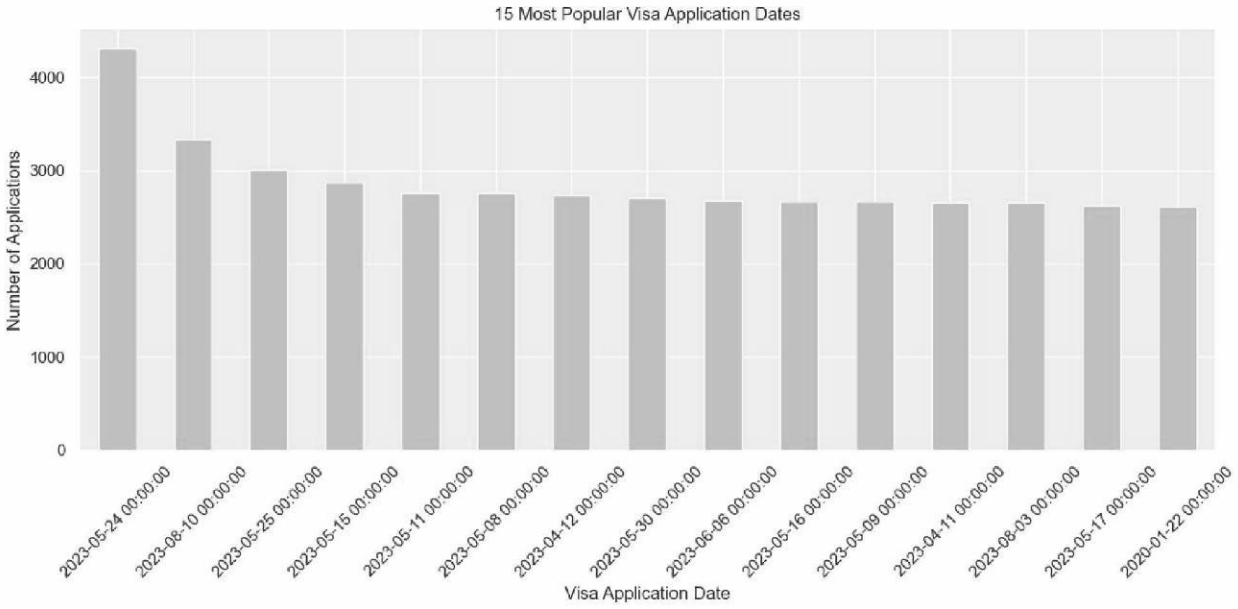


The graph above displays the trend in the number of visa applications over time. You can observe fluctuations and potential patterns or seasonality in the volume of applications. The main reason for this noticeable fluctuation is COVID-19, which has substantially reduced international travel. Various other factors, such as holiday seasons, changes in visa regulations, or other global events, could also have an impact on international travel.



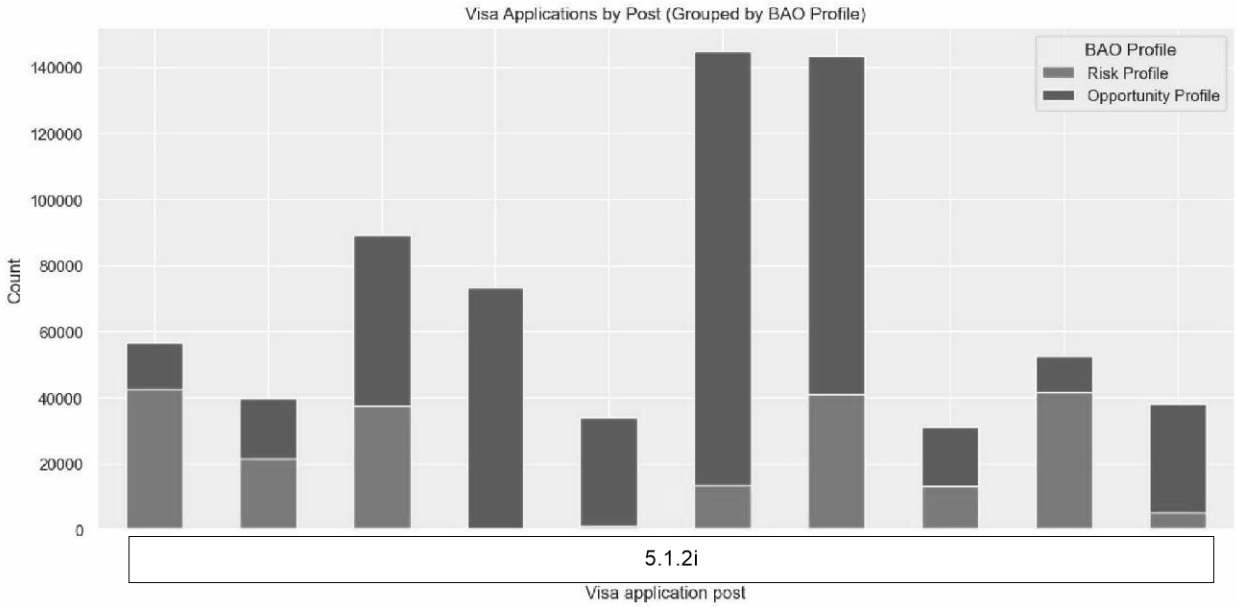
An important point regarding Fast, Regular, and Intensive Track applications is the difference in processing time. As we observe in this Box Plot, the median number of days to process an Intensive Track application is a few days more than a Fast Track application (13 days vs 6 days). The median number of days to process a regular track application is 7 days. Most of the regular track applications (~99.5%) don't have a BAO profile associated with them. Hence, though we don't consider applications without BAO profile for bias analysis, we have considered those application here, to include regular track as well in the graph above.

1. **visa application_date**:



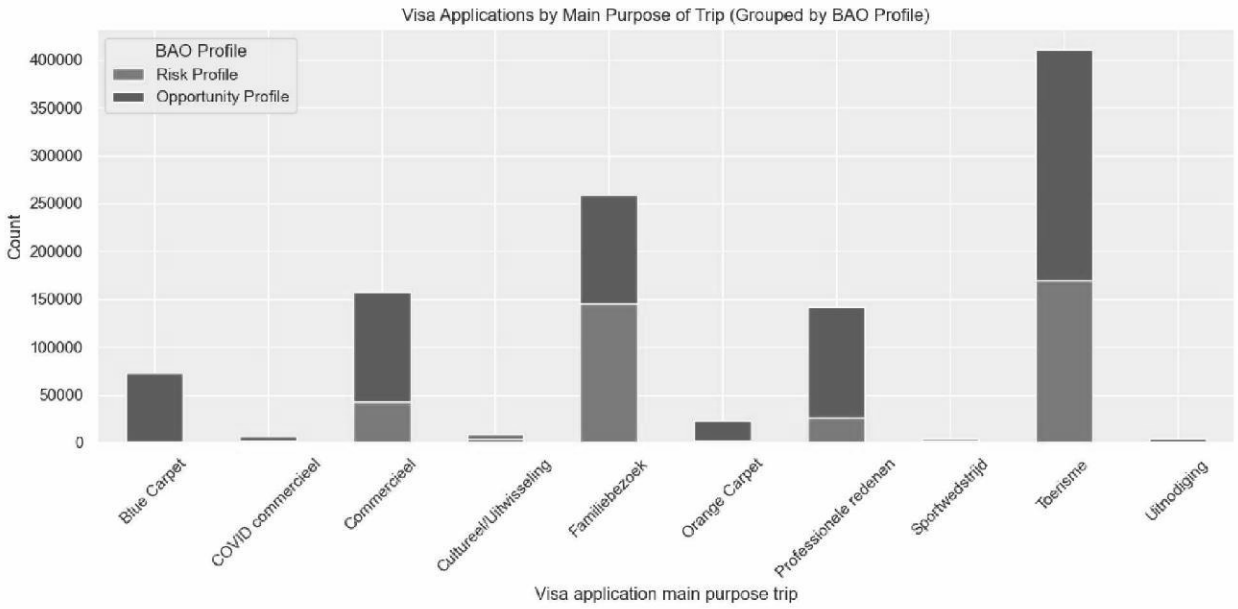
The above chart depicts the highest number of visa applications on any given data over five years. The top 15 values are considered above. Applications were made on 1,330 different dates.

2. **visa application_post**:



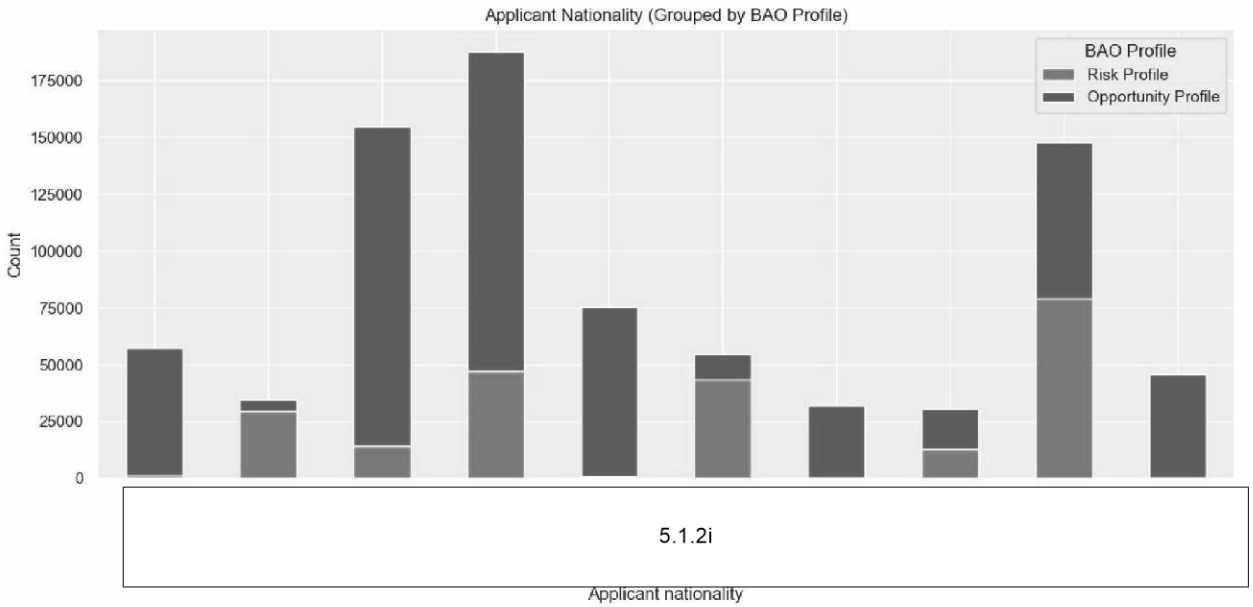
The above chart portrays, for the top 10 application posts, the number of applications attributed an Opportunity profile and the number of applications attributed a Risk profile for each post. There are 89 different posts or offices where applications were submitted, with 5.1.2i being the most frequent (144,671 applications).

3. **visa application_main purpose_trip**:



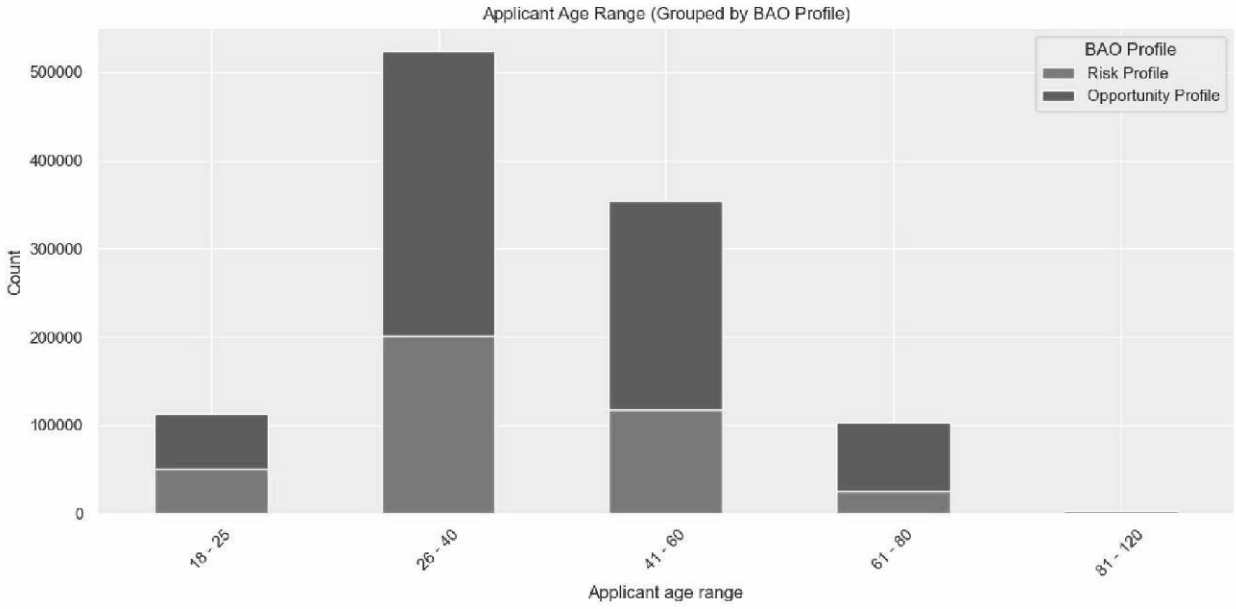
There are 20 different stated main purposes of trips, with 'Toerisme' (Tourism) being the most common (410,724 applications). The above chart portrays, for the top 10 main_purpose_trip of the applications, the number of applications that are attributed Opportunity profiles and the number of applications that are attributed Risk profiles for each main_purpose_trip.

4. **application_nationality**:

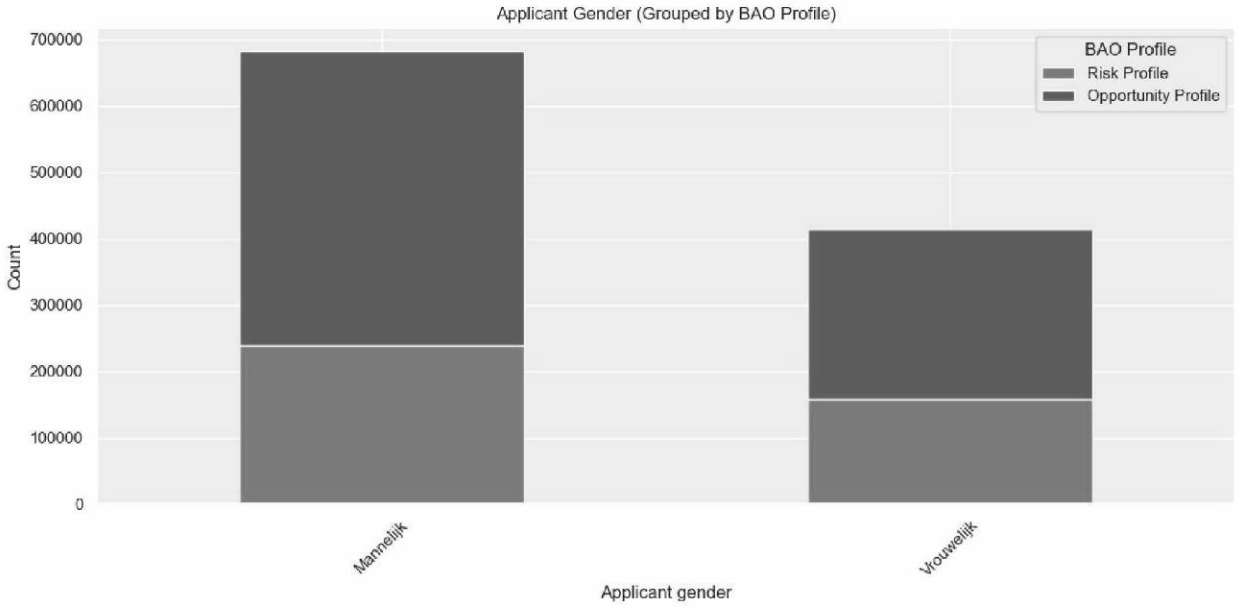


Applications come from 138 different nationalities, with 5.1.2i Nationality being the most frequent (187,889 applications). The above chart portrays, for the top 10 application_nationality of the applications, the number of applications attributed an Opportunity Track profile and the number of applications attributed a Risk profile for each application_nationality.

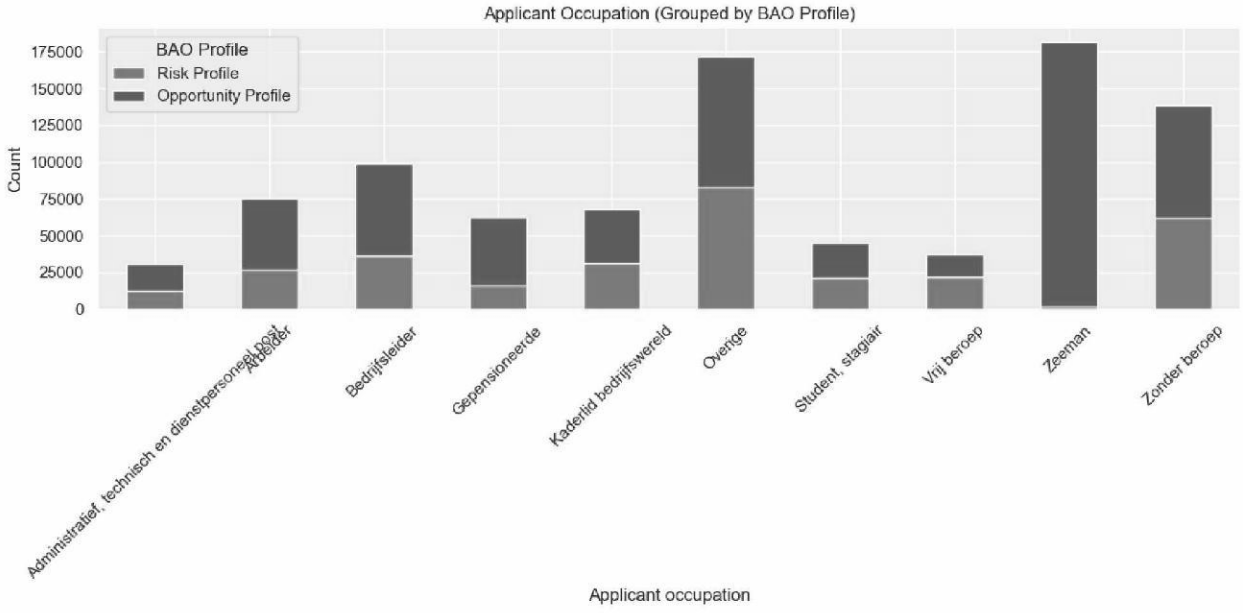
5. **application_age_range**: There are 5 different age ranges, with '26 - 40' being the most common range (523,369 applications).



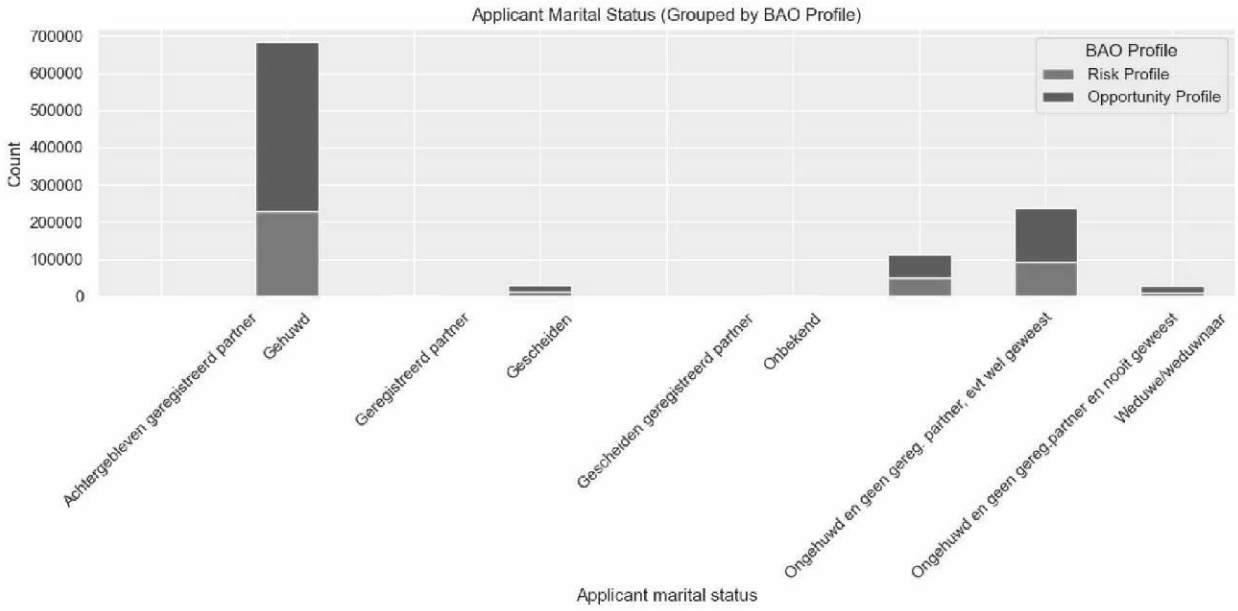
6. **application_gender**: There are two genders listed, with 'Mannelijk' (Male) being the more common gender (682,603 applications).



7. **application_occupation**: Applications have listed 36 different occupations, with 'Zeeman' (Seaman) being the most frequent (181,719 applications).



8. **application_marital_status:** There are 9 different marital statuses, with 'Gehuwd' (Married) being the most common (682,569 applications).



Causal Inference

Overview

In the context of the algorithmic assessment report, the causal inference section was designed to estimate the influence of the 'BAO Profile' on visa application decisions. This analysis can provide additional insight into whether the tracks only causes additional visa processing time or if there might be a causal impact of tracks on the final decision. To achieve this, we implemented relevant techniques to evaluate if there is causality between the BAO response and visa officers' decisions after controlling for potential confounders (in this case, information that a visa officer makes a decision with that the BAO Classification Model also has access to such as Nationality, Gender, Age Range, Marital Status, Post, Purpose of Trip, and Occupation).

Our analysis revealed that the presence of an Opportunity BAO Profile is likely to correlate with a positive outcome in visa application decisions (and vice versa).

It is important to note that while the BAO tracks appear to be a significant factor, visa officers consider a multitude of elements when making their decisions. Therefore, the results of this causal inference should be interpreted as one piece of the broader decision-making process rather than a definitive indicator of causality within the model. Furthermore, as correlation 'doesn't imply causation, an experiment, as expanded upon in the Recommendation Section, is suggested to confirm the results.

Detailed Analysis

Inverse Probability Weighting Analysis

Our approach:

1. Algorithm Choice: XG Boost (eXtreme Gradient Boosting) was selected for its popularity and efficacy in handling large datasets, leveraging a boosting technique to combine predictions from weak learners, typically decision trees (this decision tree, however, is entirely different from the BAO Classification Model).

2. Estimation of Propensity Scores: The XG Boost algorithm was employed to estimate propensity scores, representing the likelihood of receiving the treatment 'BAO Profile' based on observed covariates. Optimal hyperparameters were selected through experimentation to ensure model performance. The rationale behind using the XGBoost algorithm for this purpose is further explained in the Appendix.

3. Calculation of Weights: Weights were computed as the inverse of the estimated propensity scores using the formula:

$$Weight = 1 / Propensity\ Score$$

4. The process of applying weights promotes a balanced distribution of covariates between treated and untreated groups:

    - Application of Weights: Calculated weights were applied to each observation in the dataset by multiplying the outcome variable and other relevant variables.

    - Analysis of Weighted Data: Analysis was conducted on the weighted data to derive the Average Treatment Effect (ATE) and Mean Squared Error (MSE).

$$ATE = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$$

In our approach to evaluate the causal effect of the BAO response on the visa application decision, we chose the Visa Application Decision as the dependent variable (outcome) and the BAO Response as our treatment. We would have to achieve a near-zero ATE to observe no causal effect.

Observed Results and Interpretation:

The causal inference analysis conducted using the Inverse Probability Weighting (IPW) approach with the XGBoost algorithm yielded insightful results regarding the impact of the 'BAO Profile' on visa application decisions. The Average Treatment Effect (ATE) calculated from the analysis was 0.868, which indicates that a Positive BAO Profile (as indicated by the Opportunity Profile) is associated with an increase in the probability of a positive visa application decision by ~87 percentage points after accounting for the confounders using Inverse Propensity Weighting. This ATE value quantifies the average increase in the likelihood of a positive outcome attributable to the Opportunity Profile across the population studied. The robustness of the XGBoost algorithm in handling the underlying complexities of the dataset was crucial in deriving this estimate, ensuring that the observed effect is not confounded by the distribution of other covariates.

In addition to the ATE, the Receiver Operating Characteristic Curve (ROC AUC) score of 0.98 was obtained, which reflects the model's strong ability to discriminate between those who received a positive visa decision and those who did not, based on the 'BAO Profile.' Furthermore, this suggests that the propensity scores are going to be relatively accurate determinants of ATE. The results from this causal analysis provide a nuanced understanding of the 'BAO Profile's' influence and underscore the importance of considering a range of variables when interpreting the algorithm's decisions.

It should also be noted that other Causal Inference Approaches exist, such as Two-Stage Least Squares (2SLS). However, they rely on having a valid instrument that is correlated with the treatment variable but not directly correlated with the outcome except through the treatment.

Two-Stage Least Squares (2SLS) is a method that helps deal with a problem called endogeneity, where variables might be connected in a way that confuses our results. In simple terms, 2SLS works in two steps. First, it predicts the values of a variable using another variable that doesn't cause errors. Then, in the second step, these predicted values are used to get accurate results in the main analysis.

However, 2SLS only works well if we have the right kind of predictor, called an 'instrument.' An instrument has to be unrelated to errors and only affect our main variable through its impact on another variable. If we can't find a good instrument, 2SLS-based results might be less trustworthy.

For the above reason, 2SLS isn't an applicable approach due to the lack of instruments.

Inter-Temporal Bias Analysis

Overview:

In our assessment report, we have employed a comparative analysis of user profiles across two distinct periods to evaluate algorithmic bias, focusing on the relationship between risk profile percentages and rejection rates. The assumption is that fairness is reflected by a proportional relationship between these two metrics. We analyzed visa applications and their profiles in 2022 and 2023, calculating disparate impact ratios for groups based on Nationality, Marital Status, and Gender to identify any significant shifts that could indicate bias. This approach enables us to evaluate the inputs to the BAO Classification Model (rejection rate) with its outputs (in this case, we're observing Risk Profile Percentages). Furthermore, as this is a bias analysis, we are limiting our scope to protected attributes.

Our findings indicate that there has been no disproportionate discrimination across the evaluated groups other than for the Yemeni Nationality, where there is a Normalized Disparate Impact Ratio of 0.52. The ideal scenario would show a direct proportionality between changes in rejection rates and risk profile percentages, which would suggest an absence of bias.

Given that the BAO Classification Model 'doesn't play a causal role in the visa officer's decision, this result indicates that it's more likely for a decision-making officer to put additional time into evaluating this application. However, it would still have to cite a ground for refusal as mentioned in the EU Visa code. As such, we recommend continuously monitoring the BAO Classification Model and its data for any signs of drift (a change in the distribution of data coming in), as well as bias (routinely checking that the risk profile percentage doesn't exceed the rejection rate for any particular group).

Detailed Analysis

The Inter-temporal bias approach is designed to identify and measure algorithmic bias by comparing the disparate impact ratio across different groups and timestamps. The disparate impact ratio is calculated as the ratio of the mean outcomes (the percentage of applications for a particular subgroup in a risk profile) for two groups. The algorithm used for this approach also normalizes this ratio by dividing it by the ratio of the mean of a normalizing attribute (the model's inputs – rejection rate) for the two groups.

The algorithm used for this inter-temporal bias uses bootstrapping to estimate the confidence interval of the normalized ratio. Bootstrapping is a statistical method that involves generating multiple samples from the original data and calculating the statistics for each sample. The confidence interval is then estimated from the distribution of these sample statistics.

This algorithm provides a robust and statistically sound method for identifying and quantifying bias in data. It is particularly useful for analyzing data with multiple protected attributes and allows for the comparison of bias between different groups within each attribute.

There are other parity metrics available, including Statistical Parity Difference, however, in experimentation, we have observed them to closely follow the Disparate Impact Ratio in its results and hence haven't included them.

In this scenario, the risk profile percentage for men applying from Asia is determined to be 20% in the year 2022. This means that, based on various factors, 20% of applications by men from Asia are categorized as having certain risk factors that require closer evaluation in 2022. Let us assume that the rejection rate for men applying from Asia is determined to be 2% in that same time period. The ratio between both gives us 10.

If we repeat this analysis and observe the risk profile rate and rejection rate to be 22% and 2.2%, respectively, we observe a proportional change between the risk profile rate and rejection rate (as the ratio is still 10). However, if the rejection rate is 2.2% but the risk profile rate went to 30%, that indicates a significant disproportional outcome (with the ratio now being 13.6).

We utilize the risk profile rate here to compare the differences in negative outcomes controlling for the model's inputs (rejection rate). The same analysis could be performed with opportunity profile rate as well, however, due to the binary nature of the profiles, these techniques would yield the same result.

In simple terms, if the rejection rate matches the risk profile percentage, and this holds true over time, it shows that the algorithm is fair. It's not treating Asian men unfairly because the rejection rate matches the identified risk profile percentage in both time periods considered. However, if there is a disproportionate outcome, it signifies that the BAO Classification Model is amplifying the difference in rejection rate.

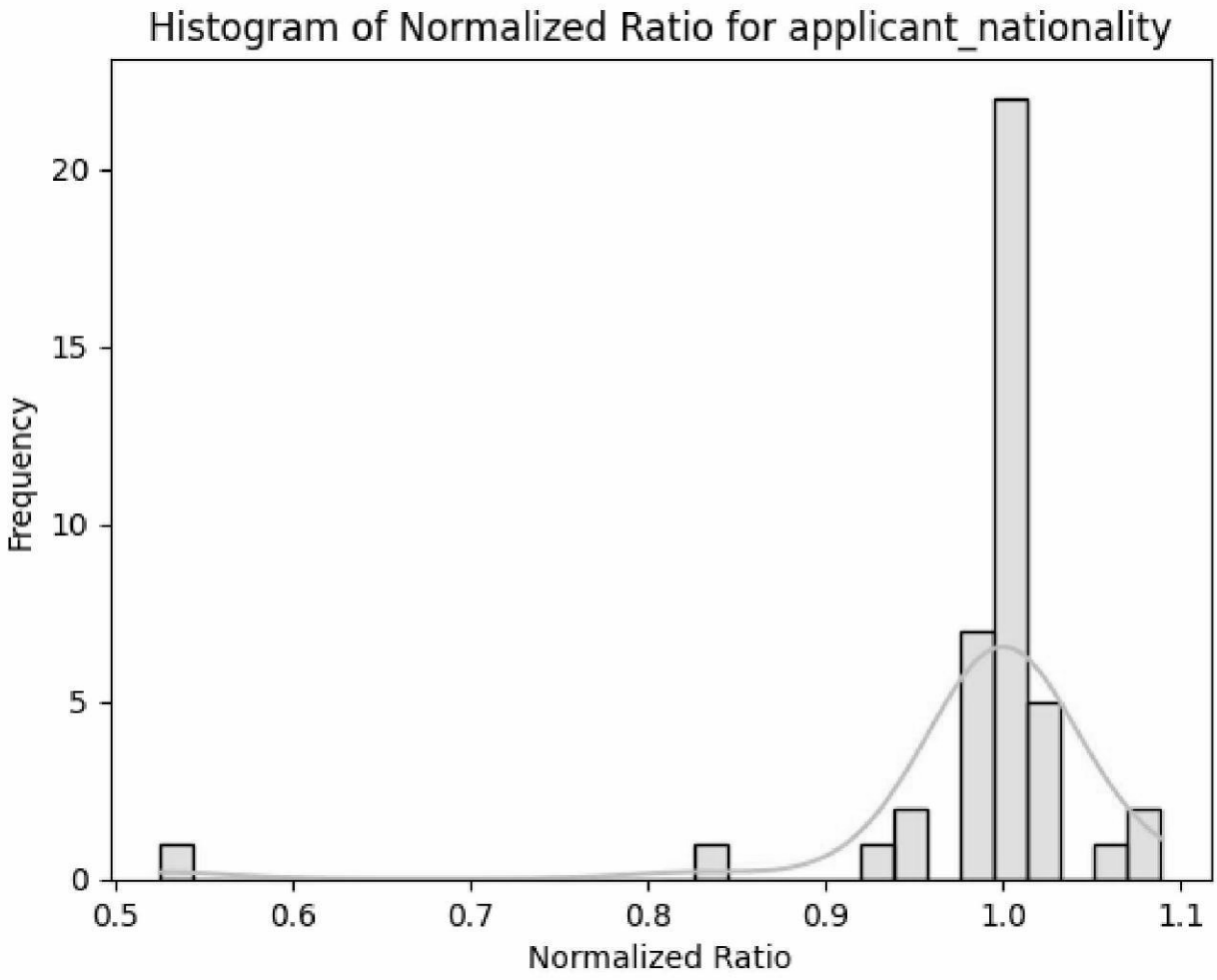
In practice, though, it's almost impossible to achieve a perfectly corresponding ratio, hence the acceptable threshold for this metric hovers between 0.8 to 1.2, as suggested by the State of California Guidelines on Employee Selection Procedures in October 1972.[2]

Observed Results and Interpretation

- The overall mean disparate impact ratio for all groups is approximately 0.960, indicating that there isn't bias on average across subgroups even without normalizing with the rejection rate.
- The overall mean normalized ratio is approximately 0.97, which is within the acceptable thresholds between 0.8 to 1.2, and indicates no bias on average after normalizing with the rejection rate.
- One group has been identified whose normalized ratios are outside the acceptable threshold of 0.8 to 1.2, indicating potential concerns regarding disparity:
    1. Yemeni group ('Jeminitische') with a normalized ratio of 0.525

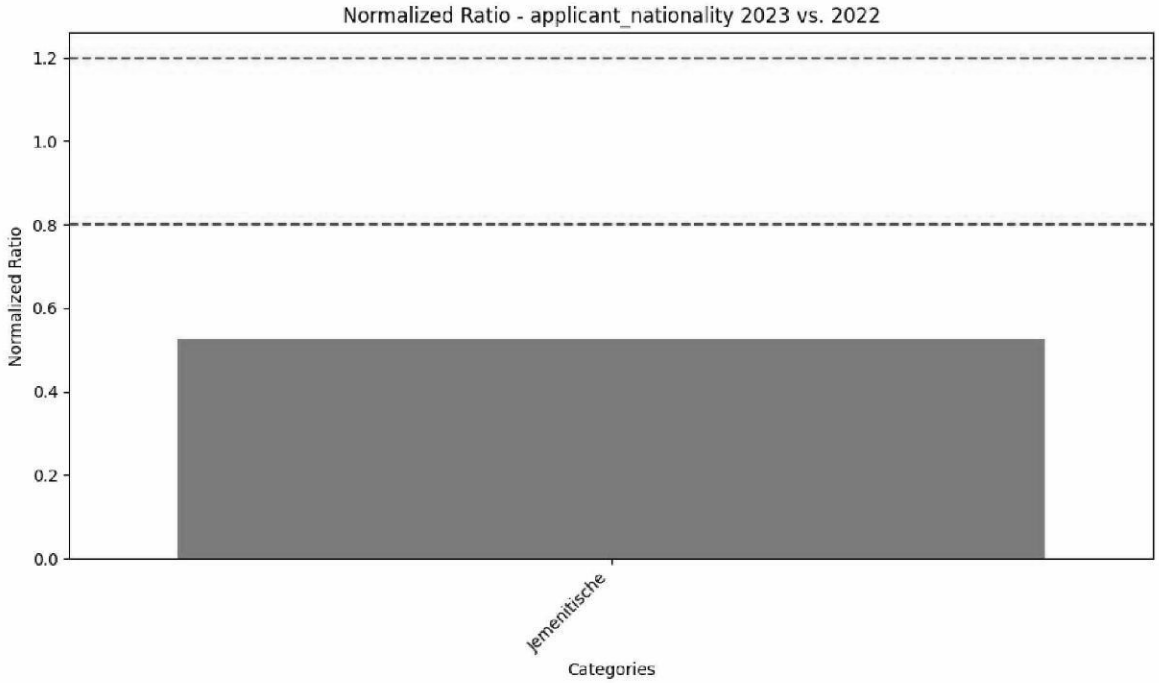
Reference:

[2]https://spb.ca.gov/content/laws/selection_manual_appendixd.pdf



The histogram shows the distribution of normalized ratios across all attributes combined. The blue dashed lines mark the threshold values, indicating where the acceptable range lies. The majority of the normalized ratios are concentrated around 1, but there is a noticeable spread, indicating variability across different groups and potential disparities. As observed, there is only one group that is outside of the threshold of 0.8 to 1.2, and that is the Yemeni (" Jemenitische") Nationality.

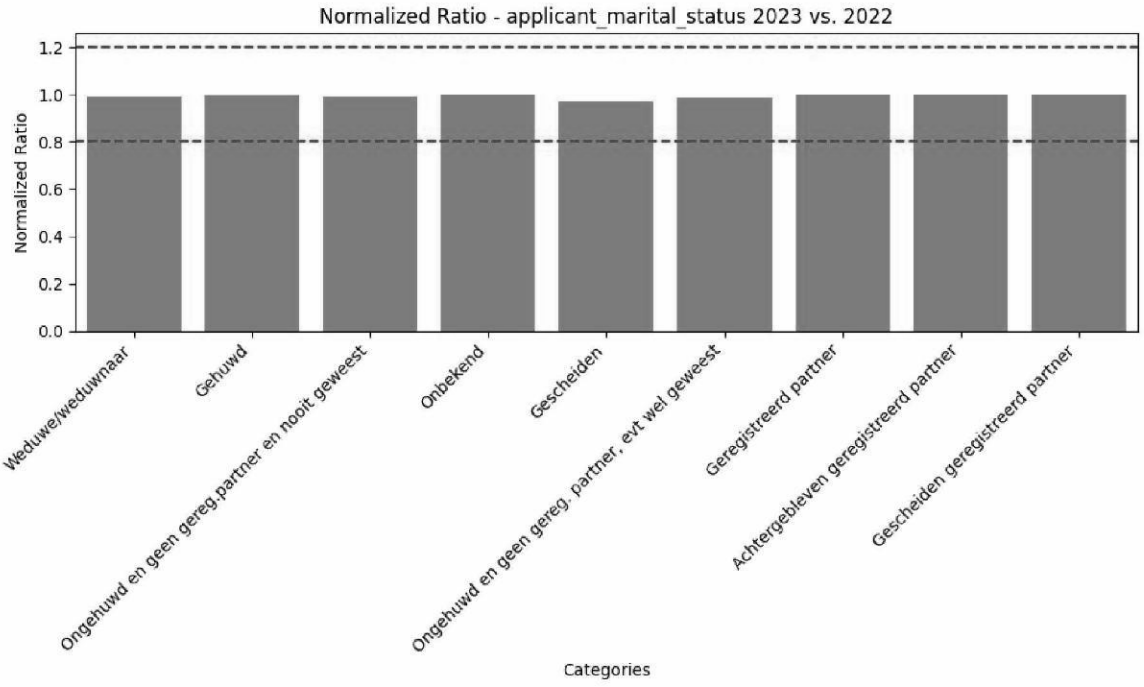
1. Nationality



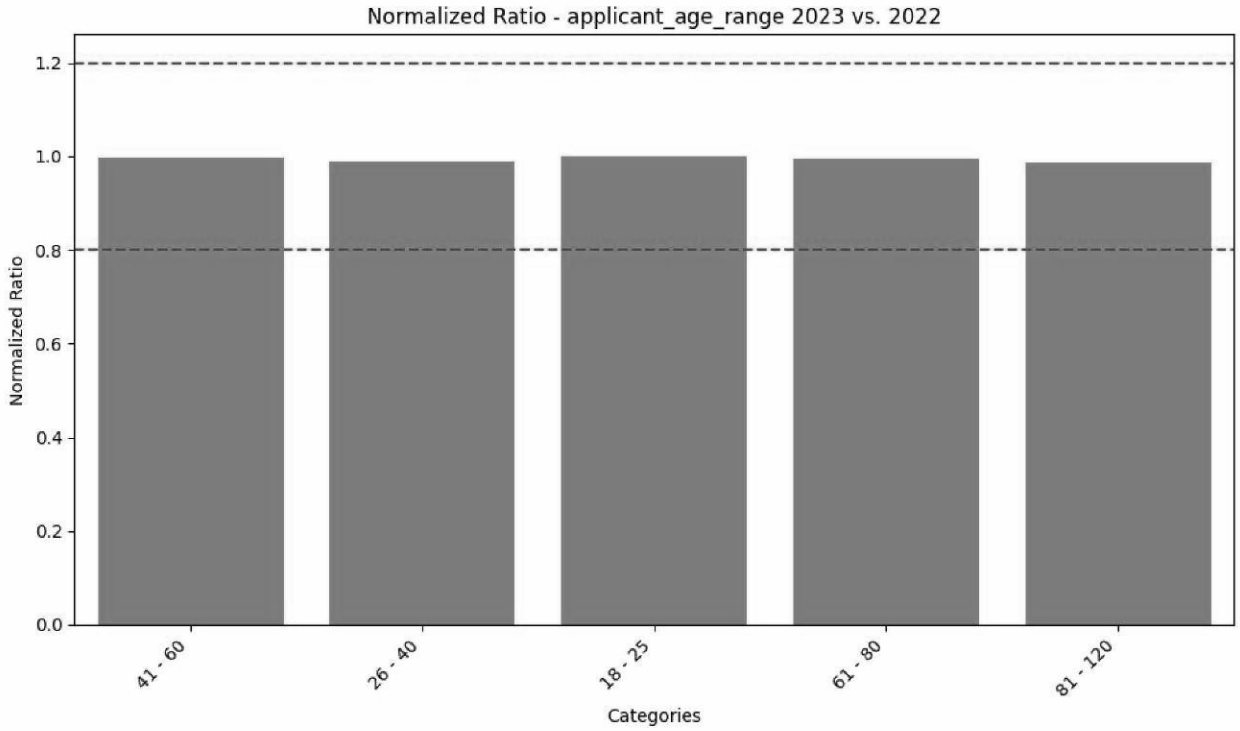
Some summary statistics for the "Jemenitische" population is given below (Applications whose Visa Application Decision is Null isn't considered as there isn't a decision made yet).

In this graph, we can observe that the nationality "Jemenititsche" is below the acceptable threshold, indicating a potential disproportionate amplification of bias by the model. It is important to note that this particularly refers to the BAO Classification Model labeling groups as a risk profile, but it doesn't indicate that these Nationality applicants are going to get their visa rejected.

2. Marital Status



This graph represents the Normalized Ratios for different Marital Statuses. As observed, all values fall within the threshold, so there is no disproportionate discrimination observed.



This graph represents the Normalized Ratios for different Age Ranges. As observed, all values fall within the threshold, so there is no disproportionate discrimination observed.

Number of Groups with Bias when not filtering off at least 200

It is also important to note that when identifying groups that may have bias in them, we have filtered out those who have 200 applications or less (for example, if a particular nationality has less than 200 applications, they wouldn't be considered for our bias analysis as the profile doesn't utilize the nationality in it's creation). We utilize this guardrail as the same policy is adopted by the BAO Algorithm when creating the profiles. With this technique we have observed a significant number of potentially biased Nationalities and Age Groups be discarded as they 'didn't meet the minimum number to establish a sound statistical claim. This indicates that the guardrail set by the NL MFA has been effective at preventing false statistical claims (which can happen if the sample size is too low).

Inter-Group Bias Analysis

Overview

The bias evaluation of the BAO Classification Model was conducted to detect and assess potential inter-group bias by examining the relationship between risk profile percentages and rejection rates across different demographic groups. The analysis was performed by comparing two groups differentiated by a single variable, such as Nationality, gender, age range, or marital status. The algorithm calculated the ratio of risk profile percentages to rejection rates, providing a metric to identify any disproportionate bias. While no disproportionate discrimination was found across gender, marital status, or age group variables, a bias was detected against the Yemeni Nationality, indicating a concern within the model that necessitates further scrutiny.

In the causal inference analysis, the focus was to determine whether the BAO Response had a direct effect on the decisions made by officers. If no causal link was found, the identified bias would primarily contribute to longer processing times rather than directly influencing decision outcomes. The results of this analysis are crucial as they help to differentiate between processing inefficiencies and actual decision-making biases, which have distinct implications for addressing the issues within the system.

The observed results from the bias evaluation revealed significant variations in the Disparate Impact Ratio, particularly concerning Nationality. The Normalized Ratios, which account for historical performance, showed less variation and were closer to the ideal value of 1, suggesting a reduced disparate impact after normalization. However, the Yemeni Nationality stood out with a consistent bias that exceeded the acceptable threshold by 14.92%. Furthermore, an analysis across combinations of nationalities and genders were calculated, and the only disproportionate bias observed was between Yemeni men and Yemeni women, with Yemeni women being biased against (and an observable Normalized Ratio of 1.33, which is 10.8% above the threshold)

As mentioned in the Inter-Temporal Bias Analysis Section: Given that the BAO Classification Model 'doesn't play a causal role in the visa ' 'officer's decision, this result indicates that it's more likely for a decision-making officer to put additional time into evaluating this application, but would still have to follow proper NL MFA Rules regarding citing particular Visa Codes when making their decision. As such we recommend continuously monitoring the BAO Classification Model and its data for any signs of drift (a change in the distribution of data coming in), as well as bias (routinely checking that the risk profile percentage doesn't exceed rejection rate for any particular group).

Detailed Analysis:

The Inter-group bias analysis algorithm works by comparing the risk profile percentages and rejection rates between two groups that differ by only one variable. This could be Nationality, gender, age range, or marital status. The algorithm queries for the aggregate positive risk profile percentage for both groups and uses the BAO Response column, which takes into account these variables, including hit rate and rejection rate. The presence of bias is evaluated by calculating the ratio between risk profile percentages and rejection rate. This approach allows for a detailed comparison across different groups and helps in identifying any potential bias. It is important to note that the algorithm can also be applied to perform multi-group analysis, where bias is evaluated across a combination of variables, such as Nationality and gender. However, to ensure that the identified bias is localized and attributable to a specific variable, the algorithm only changes one variable at a time.

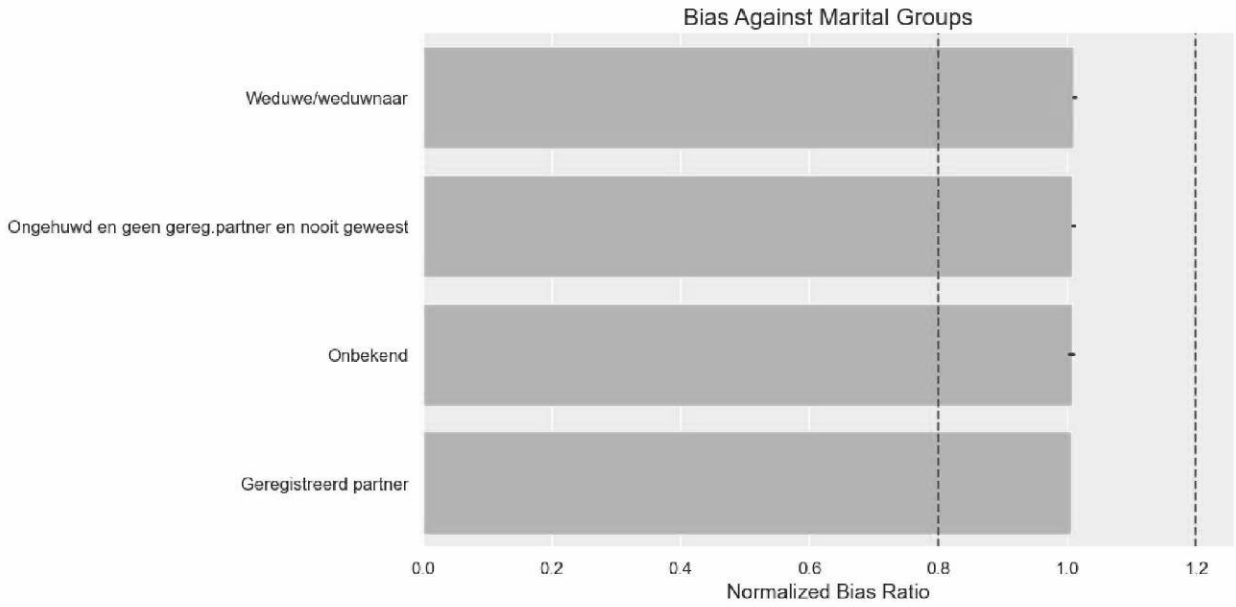
Observed Results and Interpretation:

The analysis of biases across the applicant's age range, gender, marital status, and Nationality, using both the Disparate Impact Ratio and Normalized Ratio, revealed distinct patterns of disparity. For the applicant's age range, gender, and marital status, the mean Disparate Impact Ratios were near parity, suggesting minimal biases within these categories. However, the nationality attribute exhibited a significantly higher mean Disparate Impact Ratio of 4.135, with a substantial variation indicated by a standard deviation of 7.848. This highlights the pronounced and variable impacts of bias on different nationalities. Conversely, when examining the Normalized Ratios, all attributes had means close to parity, with Nationality still showing a relatively higher degree of variation (std = 0.064), albeit less than observed in the Disparate Impact Ratio.

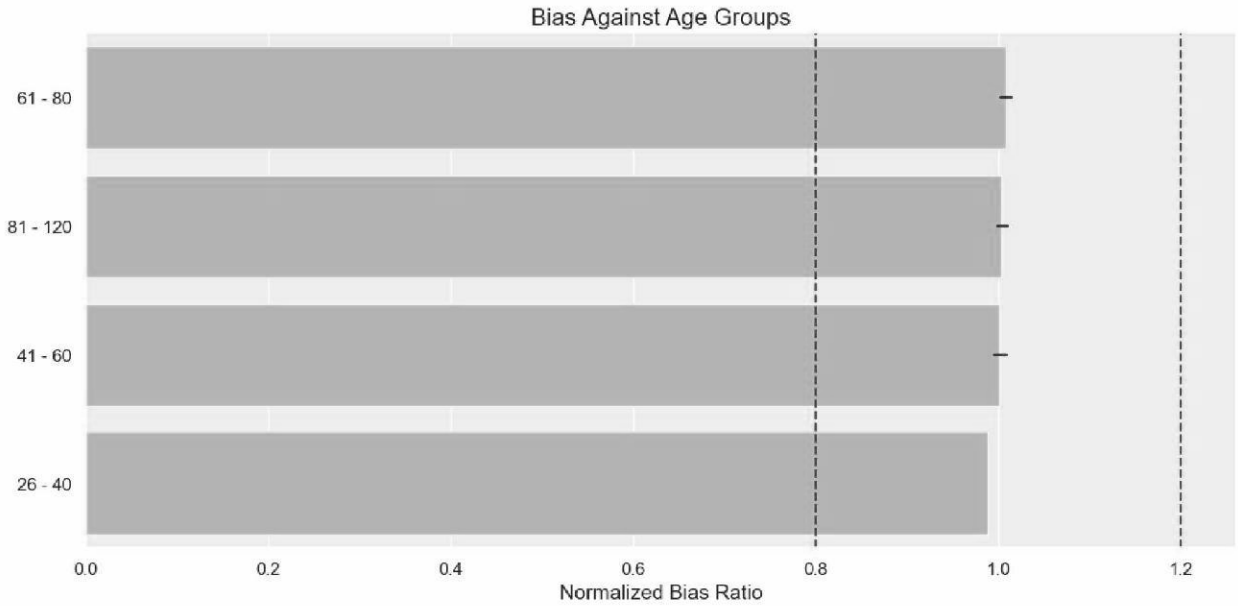
The contrast between the Disparate Impact Ratio and Normalized Ratio underscores the importance of normalization in bias analysis. While the Disparate Impact Ratio directly compares outcomes between groups, leading to potentially wide variations, especially pronounced in the nationality attribute, the Normalized Ratio provides a standardization that brings the means closer to parity and reduces the extent of variation observed. This normalization is crucial for a fair and comparable analysis across different attributes. Specifically, the significant variation and high mean in the Disparate Impact Ratio for nationalities point towards pronounced, varied biases against different national groups. However, normalization via the Normalized Ratio allows for these biases to be contextualized on a comparable scale, suggesting that while biases exist, they can be quantified and potentially addressed with targeted mitigation strategies.

Interpretation

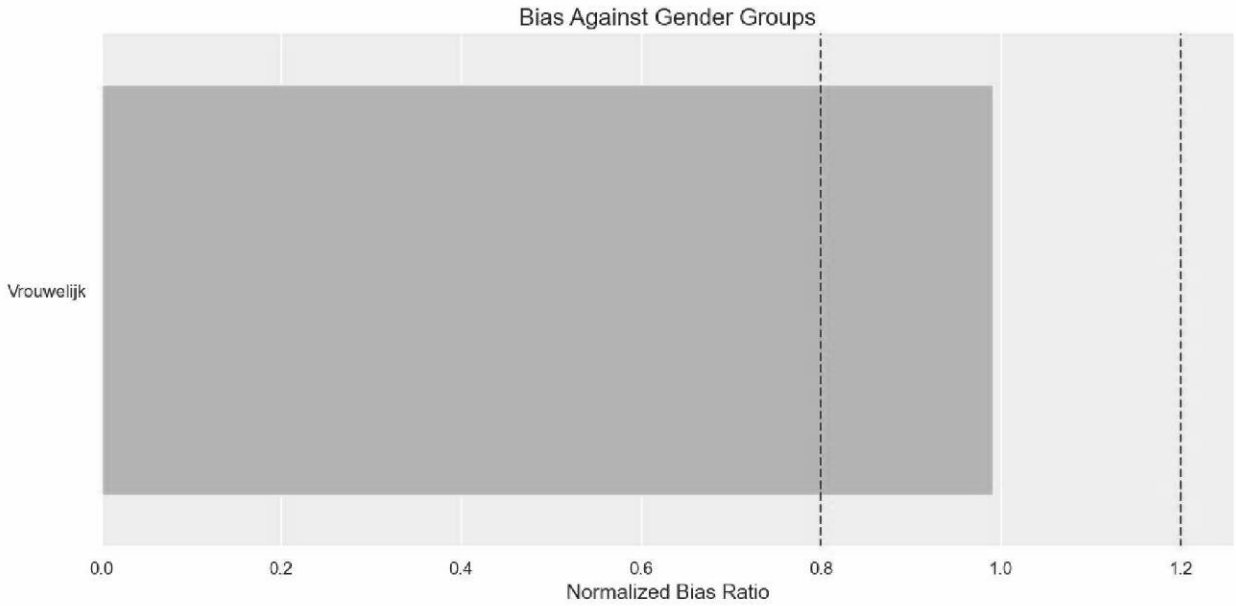
It is observed that in applications from the Nationality as Yemen ("Jemenitische ") , we noted disproportionate discrimination. In the case of Yemen, the bias threshold is exceeded by 14.92%. This indicates that the BAO Classification Model might be amplifying the risk profile presence of people from the Yemeni Nationality.



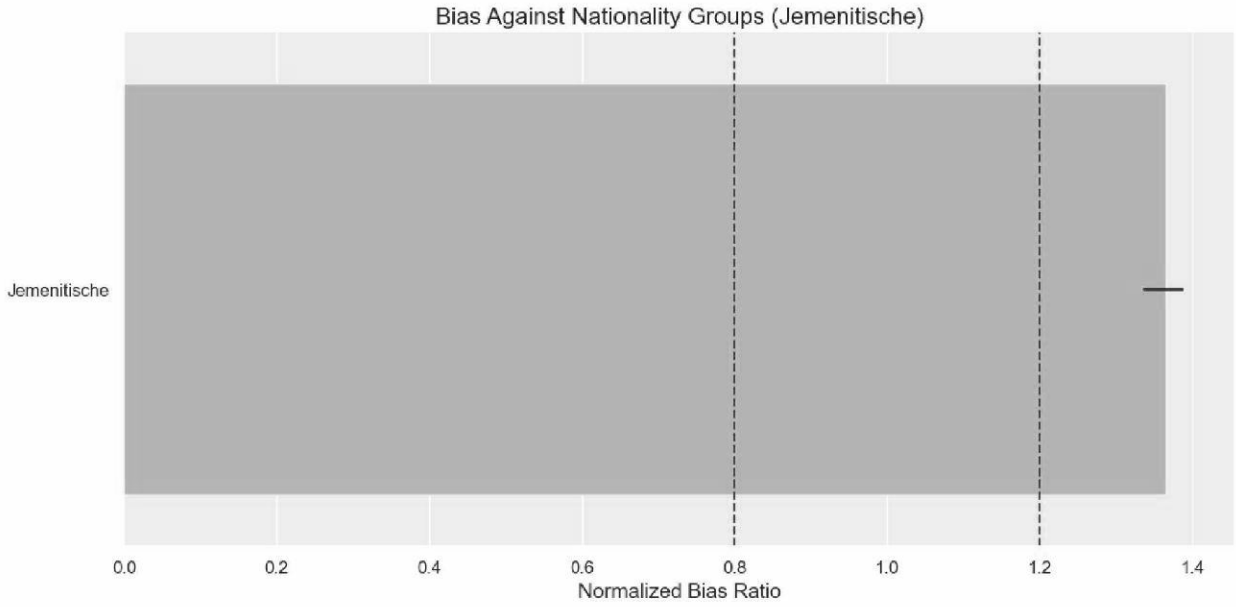
The above plot displays the Normalized Ratio for different Marital 'Status', and it can be observed that all sub-groups are within the acceptable Normalized Ratio thresholds, thus not raising any concerns for disproportionate discrimination.



The above plot displays the Normalized Ratio for different Age Groups and it can be observed that all sub-groups are within the acceptable Normalized Ratio thresholds, thus not raising any concerns for disproportionate discrimination.



The above plot displays the Normalized Ratio for different Genders. As this technique compares one category with all other categories and given that there are only two gender categories present in the dataset, only one bar is needed to represent the comparison between these groups. As the gender groups are in the acceptable threshold, so there is no disproportionate discrimination.



The above plot displays the Normalized Ratio for the Nationality attribute. Although there are plenty of nationalities, we have plotted the one that consistently exceeds the threshold of bias. As such we have plotted the Yemeni ("Jemenitische") Nationality, that we can observe has disproportionate discrimination against it.

Some insight into this method can be found here [3]

Reference:[3] https://www.sciencedirect.com/science/article/abs/pii/S2214804318302507

Recommendations

Given our in-depth bias analysis and causal inference tests, we provide the following recommendations.

Causal Inference and Bias Experimentation

Given the potential causal effect of BAO Response as well as the detection of bias in one nationality, we propose the implementation of a rigorous causal testing approach. To achieve this, we recommend the design and execution of an A/B experiment tailored to assess the model's impact on decision outcomes.

To test for causal effect and bias, visa applications that would traditionally be deemed as 'approvable' could intentionally be placed within a high-risk profile by the experiment design team. This intentional misalignment of risk profiles allows us to simulate a scenario where the model advises against approval, potentially leading to an increased likelihood of rejection.

By systematically analyzing the outcomes of this A/B experiment, we aim to discern whether the model inadvertently influences rejection rates. This experimental setup is crucial for uncovering any latent biases or unintended consequences that might arise during the model's advisory process.

The insights gained from this experiment will provide valuable empirical evidence regarding the model's robustness and fairness. It will serve as a proactive measure to ensure that the model does not inadvertently exacerbate biases in the decision-making process. Through such systematic testing, we can enhance the model's reliability, address potential challenges, and reinforce the integrity of the decision-making framework.

Continuous Model Monitoring

Given that there has been bias identified for one nationality along with noticeable drifts across variable distributions over time, we highly recommend performing continuous monitoring across both the visa application (input) and the generated profile (output) data. This evaluation would be done across three key domains: Model Performance, Data Drift, and Bias. By evaluating model performance rigorously it would ensure that rising numbers of potential misclassifications of the Classification Model are adequately identified. Data Drift closely links with this approach, as by evaluating the drift of data at a monthly, quarterly and yearly level, this would provide detailed insights that can be used both to update visa policies, as well as ensure that the classification model doesn't under/over utilize features.

Data Drift Importance Example: If a particular nationality has historically had very few visa overstays but have now had a significant increase in the number of visa overstays due to a foreign crisis. As the data is aggregated over 5 years, for every additional month there is only roughly a ~3.34% change in the data (one month is dropped and a new month is added). This would result in this Nationality still being classified as an Opportunity profile. However, given these recent developments it might be wiser to either not categorize this Nationality into a profile, or employ other measures to ensure a significant number of misclassifications don't occur. Continuous monitoring of every variable would strongly enable these insights to be identified earlier, especially by utilizing an alerts system when there has been a statistically significant distribution shift in the data.

It is also important to assess for drift in bias metrics over time. This will ensure that any potential bias arising is promptly identified and alerted according to its severity (i.e. how far out of the threshold its results are). Furthermore, constant documentation of bias evaluations will be significantly useful in building trust regarding the Classification Model with all its stakeholders.

To conclude, we strongly recommend monitoring the Classification Model across model performance, data drift and bias, as they would help pre-emptively catch any arising issues, enable dynamic and empirical evidence to potentially adjust model boundaries or variables used, and lastly identify rising bias in the Classification Model before it can increase out of proportion.

Appendix

A. Differences between the BAO Classification Model and Traditional Supervised Classifiers

The current classification model employed by the BAO differs significantly from traditional supervised classifier models in several key aspects:

1. Rule-Based vs. Predictive: The BAO's model is rule-based, meaning it operates on a set of predefined rules that are applied to sort applications into categories. In contrast, traditional supervised classifiers are predictive, using historical data to learn patterns and make predictions about new, unseen instances. The BAO's model categorizes applications into risk and opportunity groups based on predefined thresholds of hit rates and refusal percentages. Traditional supervised classifiers predict the likelihood of an outcome (e.g., risk of default) based on input features, without necessarily adhering to fixed thresholds.

2. Static vs. Dynamic Learning: The BAO's model does not learn or adapt from past decisions or outcomes (static). Traditional machine learning models, however, continuously update their parameters based on new data to improve prediction accuracy over time (dynamic).

3. Interpretability vs. Complexity: The BAO's model is highly interpretable because it follows a clear set of rules that can be easily understood and explained. Traditional classifiers, especially complex models like neural networks or ensemble methods, can act as "black boxes" with decision-making processes that are difficult to interpret.

4. Feature-Based Decision Making: Traditional classifiers often weigh the importance of various features differently and combine them in complex ways to make a prediction. The BAO's model, however, does not weigh features but uses them to check against predefined profiles, which are essentially sets of rules.

5. Outcome Influence: In the BAO's model, the algorithm's output is not the sole determinant of the final decision. It is one of several inputs used by officers who make the ultimate judgment thus impacting the intensity of the assessment. However, in many traditional classifiers, the model's output can be the primary determinant of the decision, especially in automated systems.

B. Additional Inadmissible Metrics

As mentioned earlier in this report, considering how the BAO Classification Model works and its role in decision-making, we can't use some of the standard model evaluation metrics. These metrics usually need a direct comparison between predicted and actual outcomes, but that doesn't apply here since the model doesn't directly predict visa decisions. In addition to the metrics given in the section inadmissible metrics section earlier, we can't use the following metrics to evaluate the BAO Classification Model. Hence, alternative methods that align with the model's advisory role and the decision-making context were employed in our analysis like causality analysis, inter-group analysis, and inter-temporal analysis.

To consider these metrics, one needs to know the actual decisions against the predicted decision of the BAO classification model. As mentioned earlier, the BAO classification model does not" predict" any decisions and is not an automated decision-making system. It only maps the application into respective profiles and hence these metrics are not applicable.

Sl. No	Metric Name	Metric Description
1	Equal Opportunity Difference	Equal Opportunity Difference is a measure used to check if a model treats different groups of people fairly. It looks at whether the model makes mistakes in predicting positive outcomes equally for all groups. If the Equal Opportunity Difference is small, it means the model is doing a better job of treating everyone fairly when it comes to positive predictions. It helps ensure that the chances of receiving a positive prediction are roughly the same for different subgroups, promoting fairness in the model's outcomes.
2	Equalized Odds (EO)	Equalized Odds is a fairness metric that aims to ensure equality in both the true positive rate (sensitivity) and false positive rate across different subgroups or demographic categories. In other words, it strives to make sure that the model performs equally well in correctly identifying positive instances (true positives) and in avoiding false positive predictions for all subgroups, promoting fairness in both aspects of the classification.
3	False Discovery Rate	The false discovery rate is the proportion of falsely predicted positive points out of all predicted positive points. It tells us how many of the predicted positive instances are actually false positives.
4	False Discovery rate difference	This metric looks at the difference in false discovery rates (FDR) between two groups: the unprivileged and the privileged. The goal is to have a value close to 0, which is considered ideal and fair. If the value is less than 0, it means there's a slightly higher benefit for the unprivileged group, and if it's greater than 0, there's a slightly higher benefit for the privileged group. Fairness is achieved when the metric falls within the range of -0.1 to 0.1
5	False discovery rate ratio	This metric, the false discovery rate ratio, compares the false discovery rates (FDR) between the unprivileged and privileged groups. It's calculated by taking the ratio of FDR for the unprivileged group to FDR for the privileged group. The ideal value is 1, indicating fairness. If the value is less than 1, it suggests a slightly higher benefit for the unprivileged group, while a value greater than 1 implies a slightly higher benefit for the privileged group
6	False Omission Rate	False omission rate is the proportion of falsely predicted negative points out of all actual negative points. It tells us how many of the actual negative instances are incorrectly predicted as positive
7	False omission rate difference	This metric, the false omission rate difference, looks at the gap in false omission rates between two groups: the unprivileged and the privileged. It's calculated by subtracting the false omission rate of the privileged group from that of the unprivileged group. An ideal value is 0, suggesting fairness. If the value is less than 0,

		it means there's a slightly higher benefit for the unprivileged group, while a value greater than 0 implies a slightly higher benefit for the privileged group. Fairness is considered achieved when the metric falls within the range of -0.1 to 0.1.
8	False omission rate ratio	The false omission rate ratio is indeed calculated by dividing the false omission rate of the unprivileged group by that of the privileged group. The interpretation of the metric aligns with the concept that an ideal value is 1, indicating fairness. A value less than 1 implies a slightly higher benefit for the unprivileged group, while a value greater than 1 suggests a slightly higher benefit for the privileged group.
9	False Positive Rate (Fall-out)	The ratio of false positives to the total number of actual negative cases is known as the False Positive Rate (FPR). It represents the proportion of negative instances that are incorrectly predicted as positive by a model
10	False positive rate ratio	The false positive rate ratio compares the false positive rates (FPR) between the unprivileged and privileged groups. It's calculated by dividing the FPR of the unprivileged group by the FPR of the privileged group. An ideal value is 1, indicating fairness. A value less than 1 suggests a slightly higher benefit for the unprivileged group, while a value greater than 1 suggests a slightly higher benefit for the privileged group.
11	False Negative Rate (Miss Rate)	The false negative rate is a metric that assesses the performance of a classification model. It specifically measures the proportion of instances that are actually positive but are incorrectly predicted as negative by the model. In other words, it represents the ratio of false negatives to the total number of actual positive cases. False negatives occur when the model fails to correctly identify instances that truly belong to the positive class, and the false negative rate provides insight into the extent of this misclassification.
12	False negative rate ratio	The false negative rate ratio compares how often the model incorrectly predicts negatives in the unprivileged group versus the privileged group. It is calculated by dividing the false negative rate (FNR) of the unprivileged group by the FNR of the privileged group. An ideal value is 1, suggesting fairness. A value less than 1 implies a slightly higher benefit for the unprivileged group, while a value greater than 1 suggests a slightly higher benefit for the privileged group.
13	Average Odds Difference	Average Odds Difference is a way to compare the average chances of an event happening between two groups. It's calculated by finding the average difference between the false positive rate (likelihood of a wrong positive prediction) and true positive rate (likelihood of a correct positive prediction) for unprivileged and privileged groups. The goal is an ideal value of 0, signifying fairness. If the value is less than 0, it suggests a slightly higher benefit for the privileged group, while a value greater than 0 suggests a slightly higher benefit for the unprivileged group. Fairness is considered within the range of -0.1 to 0.1.
14	Error Rate Difference	Error rate difference is a measure that looks at the gap in error rates between two groups, the unprivileged and the privileged. The goal is to have an ideal value of 0, indicating fairness. If the value is less than 0, it means there's a slightly higher benefit for the unprivileged group. Conversely, if the value is greater than 0, it suggests a slightly higher benefit for the privileged group
15	Specificity (True Negative Rate)	The true negative rate is the proportion of correct negative predictions (true negatives) among all actual negative cases
16	Negative Predictive Value	Negative Predicted Value (NPV) is a measure indicating how trustworthy a negative prediction or test result is. It is calculated as the ratio of true negatives to the total number of negative predictions. In a medical context, NPV helps assess the likelihood that a negative test result accurately indicates the absence of a particular condition. For instance, if a test has a high NPV of 95%, it suggests that 95 out of 100 people with a negative result are truly free of the condition. In contrast, a lower NPV, such as 70%, would imply that out of 100 people with a negative result, only 70 are genuinely free of the condition, and 30 might have the condition despite the negative result.
18	Statistical Parity Difference	Statistical Parity Difference measures the difference in predicted positive outcomes between two groups, like Group A and Group B. In simpler terms, it helps assess if there is fair and equal treatment across different groups. A value of 0 indicates equal outcomes, while a value less than 0 suggests a slightly higher benefit for the

		privileged group, and a value greater than 0 suggests a slightly higher benefit for the unprivileged group. It's a way to check for disparities in various outcomes, such as employment rates or loan approvals, between different groups.
19	Equal Opportunity Difference	Equal Opportunity Difference measures the difference in true positive rates between two groups. The true positive rate is the ratio of correct positive predictions to the total actual positive cases for each group. In the concept of equal opportunity, everyone should have the same chances for success, regardless of background. It aims to treat everyone fairly and eliminate discrimination based on factors like race or gender. An ideal Equal Opportunity Difference is 0, signifying equal treatment. A value less than 0 means a slightly higher benefit for the privileged group, while a value greater than 0 indicates a slightly higher benefit for the unprivileged group.

Conventional metrics used for evaluating predictive models rely on the presence of a binary or multi-class outcome, which allows for a direct comparison between the actual outcomes and the predictions made by a model. However, the BAO Classification Model operates differently in that it does not generate predictions that fit into these straightforward categories. Instead, it serves an advisory function, assisting in decision-making processes in a manner that does not align with the direct comparison approach of standard evaluation metrics.

Given this unique characteristic of the BAO Classification Model, it is clear that the traditional methods of model evaluation are not suitable for assessing its effectiveness. The model's value lies in its ability to inform and guide decisions rather than predict outcomes in a binary or multi-class format. Therefore, to accurately evaluate the performance of the BAO Classification Model, it is necessary to adopt alternative evaluation strategies. These strategies should be tailored to the model's advisory nature and the specific context in which it is used to make decisions.

In essence, the evaluation of the BAO Classification Model requires a shift away from conventional predictive accuracy metrics towards methods that can capture the model's impact on decision-making quality and outcomes. This might involve assessing the relevance and utility of the advice provided by the model, how it influences decision-making processes, and ultimately, how it affects the effectiveness of the decisions made. Such an approach ensures that the evaluation is aligned with the model's operational context and its role as a decision support tool.

C. Insight Into the Causal Inference Model

Why XG Boost:

The use of the XG Boost algorithm in the above scenario is just one choice, and it's not mandatory. Different algorithms could be used to estimate propensity scores. XG Boost is a popular choice due to its ability to handle complex relationships and high predictive performance.

- XG Boost can capture complex non-linear relationships between covariates and the propensity for treatment.
- XG Boost is robust to outliers and can handle missing data effectively.
- XG Boost provides feature importance scores, allowing to identify which covariates are more influential in predicting treatment assignment.
- XG Boost allows to fine-tune various hyperparameters to optimize model performance.

Hyperparameters:

When using XG Boost for propensity score estimation, we tuned the following hyperparameters and default values satisfies:

- Learning Rate (eta): It controls the contribution of each tree to the final prediction. Lower values make the model more robust but may require more trees.
- Number of Trees (n_estimators): The number of boosting rounds (trees) to build. A higher number may lead to better performance but could also increase the risk of overfitting.
- Max Depth (max_depth): The maximum depth of a tree. Deeper trees can capture more complex relationships but might lead to overfitting.
- Subsample and Colsample: These parameters control the sampling of the dataset during training. Subsample determines the proportion of training data to be used in each boosting round, while colsample determines the fraction of features to be randomly sampled for building each tree.
- Regularization Parameters (lambda and alpha): These control the L1 and L2 regularization terms and help prevent overfitting.

Causal Inference Model Glossary:

1. XG Boost (eXtreme Gradient Boosting): This is a type of machine learning algorithm, specifically a classification model, that is often used for its ability to handle large datasets. It works by combining the predictions from several simple models to create a final, more accurate prediction.

2. Propensity Scores: These are scores that represent the likelihood of receiving a certain treatment (in this case, 'BAO Profile') based on observed characteristics. In simpler terms, it's a score that predicts whether a certain condition applies to an individual based on their information.

3. Weights: In this context, weights are calculated as the inverse of the propensity scores. These weights are used to balance the dataset, ensuring that the treated and untreated groups are comparable. This helps to reduce bias in the analysis.

4. Average Treatment Effect (ATE): This is a measure of the average difference in outcomes between those who received the treatment (BAO Profile) and those who did not. In this case, an ATE of 0.868 suggests that, on average, having a BAO Profile is associated with the likelihood of a positive 'Visa Application Decision' by 86 percentage points.

5. Receiver Operating Characteristic Curve (ROC AUC): This is a measure of the overall performance of a binary classification model. It quantifies the ability of the model to discriminate between the positive and negative classes by considering the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across various classification thresholds.

6. Covariate Balance Checks: These checks are done to ensure that the treated and untreated groups are similar in terms of observed characteristics. This is important to make sure that any differences in outcomes are due to the treatment and not due to differences in characteristics between the groups.

7. Bootstrapping: This is a statistical method used to estimate the variability of a statistic, like the ATE, by resampling the data many times. The 95% confidence interval gives us a range in which we can be 95% confident that the true value of the statistic lies.