LISBON DATA SCIENCE ACADEMY

Stop and Search Operations in UK Policing

Model Deployment Analysis

Prepared for:
The United Kingdom Department of Police

Prepared by:
Francisco Oliveira
Data Scientist
Data Science Department

Table Of Contents

1. Summary

The United Kingdom Department of Police has requested an investigation into accusations of discrimination in their stop and search policy. Our task was to analyse the data to **determine if there is evidence of discrimination based on gender, ethnicity, and age**. Additionally, we created and hosted an API endpoint for authorising searches, which was integrated into their police system for uniformity of decisions. The main objective of the API is to ensure that searches are performed only when there is a **more than 10% likelihood of success** and to **level the discovery rate without significantly diminishing the overall ability to detect offences**. A proof of concept was run with our company for the use of the API in a few police stations.

We are pleased to report that the test run of the API for authorising searches went smoothly without any problems. During the test run, there were a total of 4200 successful API requests made.

Table number 1.1

Business requirements	Technical requirements
Search for evidence of discrimination on gender and ethnicity across stations.	Precision score between different ethnicities/genders should not vary 5% in each station.
Verify if women of certain age groups and ethnicities are being asked more often to remove outer clothing.	Precision score when clothes are removed on women of different ethnicities/age-ranges should not vary 5% in each station.
Searches are performed only when there is a more than 10% likelihood of success.	Initiate a search only if the predicted probability of success is greater than or equal to 10%.
Level the discovery rate without significantly diminishing the overall ability to detect offences.	Maximise the recall rate to stay above 90% to detect the majority of offences, while ensuring that the level of discovery rate remains consistent across different population sub-groups. Specifically, aim to limit the discrepancy in precision scores between sub-groups to no more than 5% at each station and ensure that the discrepancy between stations, measured as the average precision score per station, does not exceed 10%.
Create and deploy a REST API.	Ensure the REST API remains operational and processes information as requested without any disruptions or errors.

2. Results Analysis

2.1 Model Performance

The observed differences in the values presented in Table 2.1 indicate variations between the expected and observed outcomes. These differences provide valuable insights into the performance of the model and highlight areas that require attention and improvement.

The F1 score, which represents the harmonic mean of precision and recall, shows a slight increase in the observed value compared to the expected value. This indicates that the model is performing moderately well in terms of balancing precision and recall.

Looking at precision, we observe a marginal increase in the observed value. This suggests that the model is making fewer false positive predictions, thereby improving its ability to accurately identify positive cases.

On the other hand, there is a slight decrease in recall, indicating that the model is missing some positive cases compared to what was expected. This highlights the need to enhance the model's sensitivity to capture a higher proportion of true positive cases.

Analysing the average differences across different subgroups, such as gender, ethnicity, and age range, varying levels of disparities between the expected and observed outcomes are revealed. These differences emphasise the presence of biases or imbalances in the model's predictions across demographic groups, necessitating further investigation and refinement.

Overall, the model's performance, as evaluated based on the known outcomes, falls short of the desired level. Although there are some improvements in certain metrics, such as precision and reduction of bias across genders, there are still significant biases across ethnicities and age ranges. These findings highlight the need for further fine-tuning and improvement of the model to ensure fairness, accuracy, and robustness in its predictions.

Table number 2.1

	Expected	Observed
F1 Score	0.35	0.38
Precision	0.22	0.24
Recall	0.87	0.85
Average difference (Gender))	0.14	0.10
Average difference (Ethnicity)	0.17	0.37
Average difference (Age Range)	0.24	0.22

2.2 Success on requirements

In this section, we will assess the success of our model in meeting the requirements outlined in Report #1. Specifically, we will evaluate how the new population and our model performed compared to what was initially reported and expected. By analysing key metrics and comparing them to the expected values, we can gain insights into the model's behaviour and identify areas for improvement.

One of the primary requirements was to achieve a recall rate of over 90% to capture the majority of offences. However, upon evaluating the model's performance, we observed a slight decrease from the expected value of 87% to the observed value of 85%, still falling short of the expected 90%. This suggests that the model missed some positive cases, indicating the need for enhanced sensitivity to detect a higher proportion of true positive cases.

Another important requirement was to limit the discrepancy in precision scores between sub-groups to no more than 5% at each station, with an average difference not exceeding 10 percentage points between stations. Upon analysing the results (Figure 1 and Figure 2), we found that we were close to achieving a maximum difference of 5% between ethnicities and genders in terms of success rates.

However, when examining the difference found that we were close to achieving a maximum difference of 5% between ethnicities and genders in terms of success rates. For both subgroups combined (as depicted in Figure 3) and in each station (Figure 4), we fell short of meeting that requirement.

When evaluating the discrepancy between stations (as also shown in Figure 4), which was measured as the average precision score per station, we found that the observed variations exceeded the specified threshold of 10 percentage points.

Some level of discrepancy was expected as seen in our expected results Table 2.1 where our Averages differences all surpassed the 10% requirement. We saw some small improvements in Age and Gender, even having Gender pass the requirements with a change of 14% to 10%. On the other hand when comparing the observed precision scores for different ethnicities to the expected values, we observed a substantial difference. The average difference in precision scores between ethnicities was 37%, exceeding the specified threshold of 10% and the expected value of 17%. This indicates significant biases or imbalances in the model's predictions across different ethnicities, which directly contributed to the poor precision performance observed. Addressing these biases is crucial to ensure fairness and accuracy in our model's predictions and to meet the specified requirements.

In Table 2.1, we observed a small increase in precision, with the expected value of 22% rising to an observed value of 24%. Similarly, the F1 score showed a slight improvement, increasing from the expected value of 35% to an observed value of 38%. While these improvements are encouraging, they are still not sufficient to meet the desired level of precision and recall balance.

The disparities in the observed data regarding the success rates and precision scores across different sub-groups, especially in ethnicity, highlight the necessity of taking immediate action to rectify these issues. By doing so, we can ensure that our model provides

reliable and consistent predictions across all population sub-groups in all stations, mitigating any potential biases or unfairness. It is crucial to investigate the root causes of the disparities in precision scores and develop strategies to recalibrate the model's performance to align with the desired objectives.

Given these findings, it is evident that the model's performance falls short of meeting the desired level outlined in the requirements. The disparities and biases in success rates and precision scores across different sub-groups suggest the presence of systemic issues that need to be addressed to ensure fairness, accuracy, and consistency in the model's predictions.

On a positive note, the REST API performed exceptionally well, remaining operational and processing information as requested without any disruptions or errors. This highlights the robustness and reliability of the API in handling the provided data.

In conclusion, while the recall rate and precision scores did not fully meet the specified requirements nor had the expected values, the analysis provides insights into the reasons behind these shortcomings, primarily related to biases and imbalances across demographic groups. The findings underscore the need for further fine-tuning, investigation, and refinement of the model to address these issues and achieve the desired objectives of maximising recall, maintaining consistent discovery rates, and limiting the discrepancy in precision scores between sub-groups. Additionally, the operational stability and reliability of the REST API have been successfully maintained, ensuring smooth processing of information.

2.3 Population Analysis

We conducted a comparative analysis between the training data and the new population data. By examining both the overall data and specific stations/months, we aim to identify and explain the observed differences. This analysis provides valuable insights into the demographic and behavioural changes within the new population.

1. Analysis of Overall Data:

We examined various aspects, including gender, age range, officer-defined ethnicity, legislation, and object of search.

In terms of gender, we observed slight changes. The male population decreased by 0.68%, while the female population increased by 1.11%. Notably, there was a substantial increase of 264.10% in the "Other" category. (figure 5.5)

Regarding age range, the 10-17 and 18-24 groups experienced decreases of 9.65% and 14.86%, respectively. In contrast, the over 34 age group saw an increase of 24.31%, and the under 10 age group had a significant increase of 155.80%.(figure 5.6)

The analysis of officer-defined ethnicity revealed a decrease in the White population (-4.86%) and increases in the Black (+20.75%), Asian (+10.77%), Other (+12.73%), and Mixed (+76.47%) ethnic groups.(figure 5.7)

Regarding legislation, there were notable changes. Notably, the use of the Criminal Justice Act 1988 (section 139B) increased by 349.41%, while the Criminal Justice and Public Order Act 1994 (section 60) decreased by 61.33%. The usage of the Firearms Act 1968 (section 47) also saw a significant decrease of 67.23%.(figure 5.8)

In terms of the object of search, there were significant changes in several categories. Notably, there was a decrease in searches related to psychoactive substances (-98.04%) and firearms (-66.64%), while searches for game or poaching equipment increased by 79.47%.(figure 5.9)

2. Analysis of Specific Stations and Months:

In addition to the overall analysis, we conducted a focused analysis considering specific stations and months present in both datasets.

Within these specific stations and months, we observed slight changes in gender distribution, with a 0.10% increase in males and a 0.24% decrease in females. The "Other" category saw a notable decrease of 8.75% (figure 5.10).

The age range analysis showed a 5.72% increase in the 10-17 group, a decrease of 13.15% in the 18-24 group, a decrease of 3.26% in the 25-34 group, and an increase of 17.08% in the over 34 group. The under 10 group saw a 4.29% increase (figure 5.11).

In terms of officer-defined ethnicity, there was a decrease in the Asian population (-10.31%) and increases in the Black (+5.25%), Mixed (+2.99%), Other (+24.04%), and White (+0.15%) ethnic groups (figure 5.12).

There were noticeable changes in regard to legislation. For instance, the use of the Criminal Justice Act 1988 (section 139B) increased by 14.34%, while the Criminal Justice

and Public Order Act 1994 (section 60) saw a significant increase of 224.44%. The Firearms Act 1968 (section 47) usage decreased by 44.46% (figure 5.13).

Regarding the object of search, there were noteworthy variations, such as a 14.56% increase in searches related to articles for use in theft, a 65.91% increase in searches related to game or poaching equipment, and an 82.50% increase in searches related to psychoactive substances (figure 5.14).

Overall, the observed changes in the data indicate some differences between the train data and the new population.

3. Deployment Issues

3.1 Re-deployment

Although our previous efforts did not result in complete success in meeting the requirements outlined in Report #1, we remain determined to address the observed limitations and strive towards achieving the desired outcomes. The analysis of the new population, along with the overall data, has provided valuable insights into the changes that have occurred. Building upon these insights, we are now preparing for a redeployment of the model with a renewed focus on improving its performance.

One of our primary objectives during redeployment is to achieve a recall rate of over 90% in both the training and new population datasets. We recognize that a high recall rate is crucial for effectively detecting the majority of offences. To accomplish this, we have developed a custom function that calculates the optimal threshold for different models. This function ensures that the selected threshold achieves a recall rate of at least 90% while also maximising precision.

Furthermore, redeployment offers us the opportunity to implement ongoing monitoring and evaluation mechanisms to track the model's performance over time. Through rigorous evaluation processes and continuous analysis, we can swiftly identify any disparities or shortcomings that arise and take corrective measures promptly. This iterative approach ensures that the model evolves alongside the changing demographics and behaviours of the population, promoting fairness and equality in our law enforcement practices.

Our focus during redeployment is to address the observed limitations and strive towards achieving the desired outcomes. We are committed to obtaining a recall rate of over 90% in both the training and new population datasets. By utilising our custom function to determine optimal thresholds and continuously monitoring the model's performance, we aim to maximise its precision while minimising bias and discrimination. The table 5.1 in the annexes will provide a comprehensive overview of the model scores, allowing us to select the most promising models that, in this case, were the models containing sensitive information.

Those models were then tested on the new population as seen in table 5.2. Initially, we considered using the LGBMClassifier; however, extensive RAM usage was observed when receiving API requests, leading to app crashes. Additionally, the Random Forest classifier was excluded due to its tendency to generate searches for everyone. To address these issues, we opted for the same model used in Report #1. This model was further enhanced by incorporating sensitive features, resulting in improved performance with the new threshold for predicted probability. Notably, this modified model exhibited better efficiency in terms of RAM usage while handling API requests in the application.

The features considered for the new model are :

Categorical

- Legislation
- Object of search
- Part of a policing operation
- Gender
- Age Range
- Officer-defined ethnicity
- Station

Numerical

- Hour
- Day of the Month
- Day of the Week
- Month

To make a final prediction, if the predicted probability is equal to or higher than 35%, the observation is predicted as True, and if it is lower, it is predicted as False.

The API was also changed so now it returns 405 if there is any of the new information Gender, Age Range, Officer-defined ethnicity and Station is missing.

3.2 Unexpected problems

During the deployment of the model, we did not encounter any unexpected problems. The process went smoothly without any crashes or unexpected data. However, we did face some challenges during the testing phase, which we will discuss further.

One issue we encountered during testing was the compatibility of the scikit-learn version. It is essential to ensure that the scikit-learn version used during training the model is compatible with the version installed in the deployment environment. In cases where the scikit-learn version mismatch occurred, it caused compatibility errors, and we had to resolve them by either updating the scikit-learn version or modifying the model to be compatible with the installed version.

During the testing phase, we faced a specific challenge related to excessive RAM usage and API crashes. As mentioned in section 3.1, initially, we used a certain model that resulted in significant RAM consumption, causing the API to crash. However, after switching to a different model and closely monitoring the RAM metrics, we were able to identify that the issue was primarily caused by the specific model choice. This experience highlighted the importance of carefully selecting and evaluating models to ensure efficient resource utilisation and the stability of the API.

Dealing with null values was another aspect that required attention during the deployment. In real-world scenarios, it is common to encounter missing or null values in the input data. To handle null values, we implemented appropriate strategies, such as imputation techniques like mean imputation, median imputation, or filling missing values with a specific marker value. It was crucial to ensure that the deployment environment and the API were equipped to handle and process null values appropriately without causing any errors or unexpected behaviour.

Overall, while the deployment itself went smoothly, the testing phase presented some challenges related to scikit-learn version compatibility, integrating custom transformers into the API, and handling null values. These challenges required careful attention and appropriate solutions to ensure the model's performance and functionality were not compromised.

3.3 Learnings and Future Improvements

Throughout the analysis and deployment of the model, we have gained valuable technical insights that guide us towards making improvements and achieving the best possible outcomes.

One key technical learning from the model redeployment was the development of a custom function to calculate optimal thresholds. This function ensured a recall rate of over 90% while maximising precision. By fine-tuning the threshold, we struck the right balance between recall and precision, improving the model's effectiveness in identifying offences while minimising false positives.

However, there are still areas for further improvement. Some key suggestions include:

- Addressing limitations: Thoroughly analysing and addressing the limitations observed during the initial deployment is crucial. This could involve refining the model's architecture, exploring different algorithms, or incorporating advanced techniques like feature selection to improve performance and reduce bias.

- Model refinement: Fine-tuning the model's hyperparameters and exploring additional algorithms can help optimise its performance. During the deployment, we attempted to incorporate a location feature by retrieving latitude and longitude information using an API. However, we encountered performance issues and slower processing times. To enhance efficiency, alternative methods or optimizations should be explored to include location-based information without compromising system performance.

- Adapting to external factors: The model should be regularly updated and retrained to account for changes in legislation, societal dynamics, or demographic shifts. Monitoring external factors and incorporating them into the model's retraining process will ensure its continued effectiveness and fairness.

- Bias detection and mitigation: Developing techniques to detect and mitigate potential biases within the model is important. This could involve conducting fairness assessments, analysing the impact of different features on the model's predictions, and implementing mitigation strategies to ensure equitable outcomes across gender, ethnicity, and age groups.

By addressing these improvements, the model's performance can be enhanced, ensuring higher recall rates, improved precision, and reduced discrimination. Regular evaluation and iteration will allow for continuous optimization and alignment with evolving societal and operational requirements.

4. Learnings and Future Improvements

Throughout our analysis and deployment of the model, we have gained valuable insights that can guide us towards making improvements and ensuring the best possible outcome for our clients. In this section, we will discuss known areas where we can implement clear improvements and explore tasks that can shape our future approach to the problem.

One potential improvement is the utilisation of national census data. By incorporating this data, we can gain a more comprehensive understanding of search rates and address potential oversearch. Analysing search rates based on location demographics will help identify any disparities or biases that may exist. This approach goes beyond success rates alone, providing a nuanced assessment of law enforcement practices across different areas.

Continuously retraining the model with new data is crucial to ensure its relevance and accuracy. By incorporating the latest population information, the model can capture evolving patterns and behaviours, maximising its effectiveness. Regular updates and retraining maintain the model's adaptability to changing landscapes, ensuring it remains accurate and reliable.

Analysing the new data for instances of oversearch is an important exploratory task. Thoroughly examining patterns and trends in the data can provide insights into potential issues. By conducting in-depth analyses on specific stations or regions where oversearch is observed, targeted feedback and guidance can be provided to mitigate the problem.

Furthermore, exploring additional exploratory tasks and rethinking the problem from different angles can lead to innovative solutions. For example, conducting sentiment analysis on public discourse related to stop and search policies can provide a broader understanding of public perceptions and concerns. This information can be used to shape the model's development, making it more responsive to societal needs and fostering public trust.

Incorporating feedback loops with officers in the field can also be beneficial. Regular feedback from officers regarding their experiences with the API and the search authorization process can help identify areas for improvement, address any usability issues, and ensure that the system is user-friendly and intuitive.

Additionally, considering the ethical implications of the model and its deployment is crucial. Continuous evaluation of the model for bias, fairness, and potential discriminatory outcomes is necessary. Implementing mechanisms to monitor and mitigate any unintended biases or unfairness will contribute to a more equitable and just system.

By implementing these improvements and pursuing further exploration of the data, we can enhance the model's performance and contribute to a fair and unbiased law enforcement system. It is essential to maintain a proactive and data-driven approach, constantly seeking ways to improve our methods, and collaborating with relevant stakeholders to ensure the best outcomes for our clients and the communities we serve.

5. Annexes

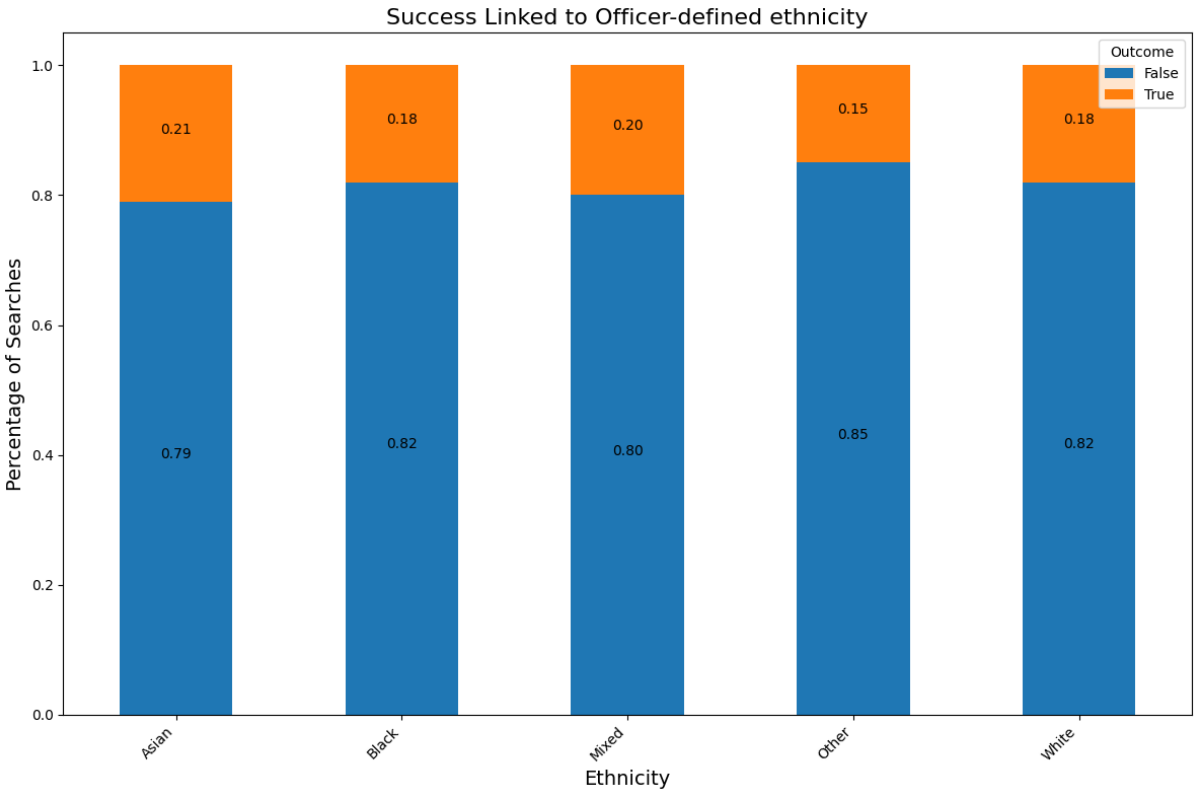

Figure 5.1

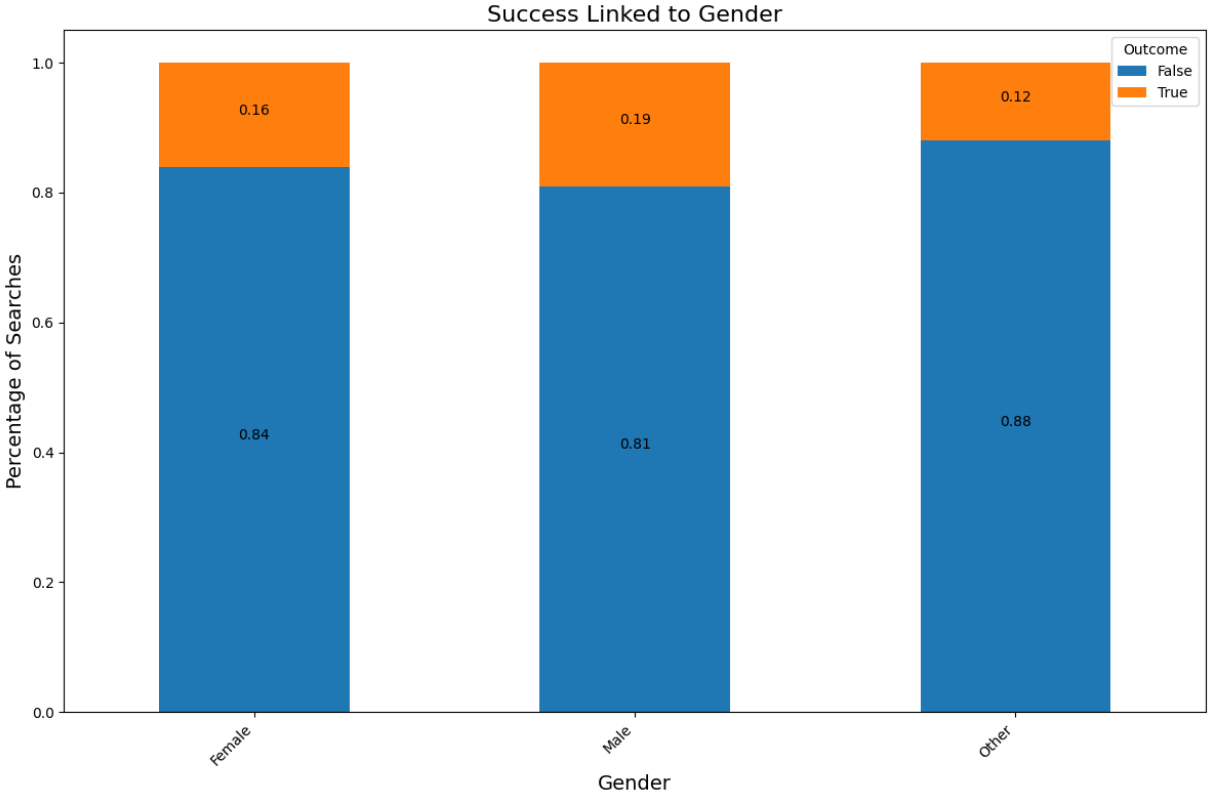

Figure 5.2

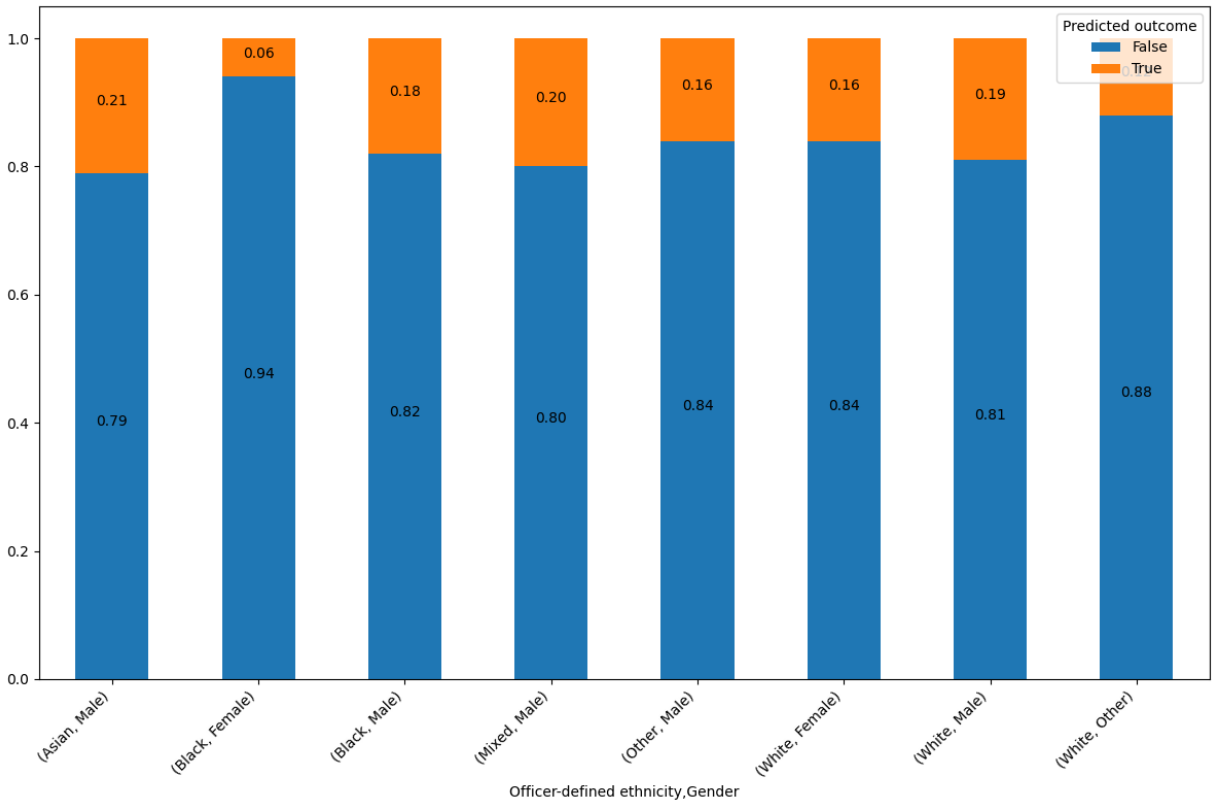

Figure 5.3

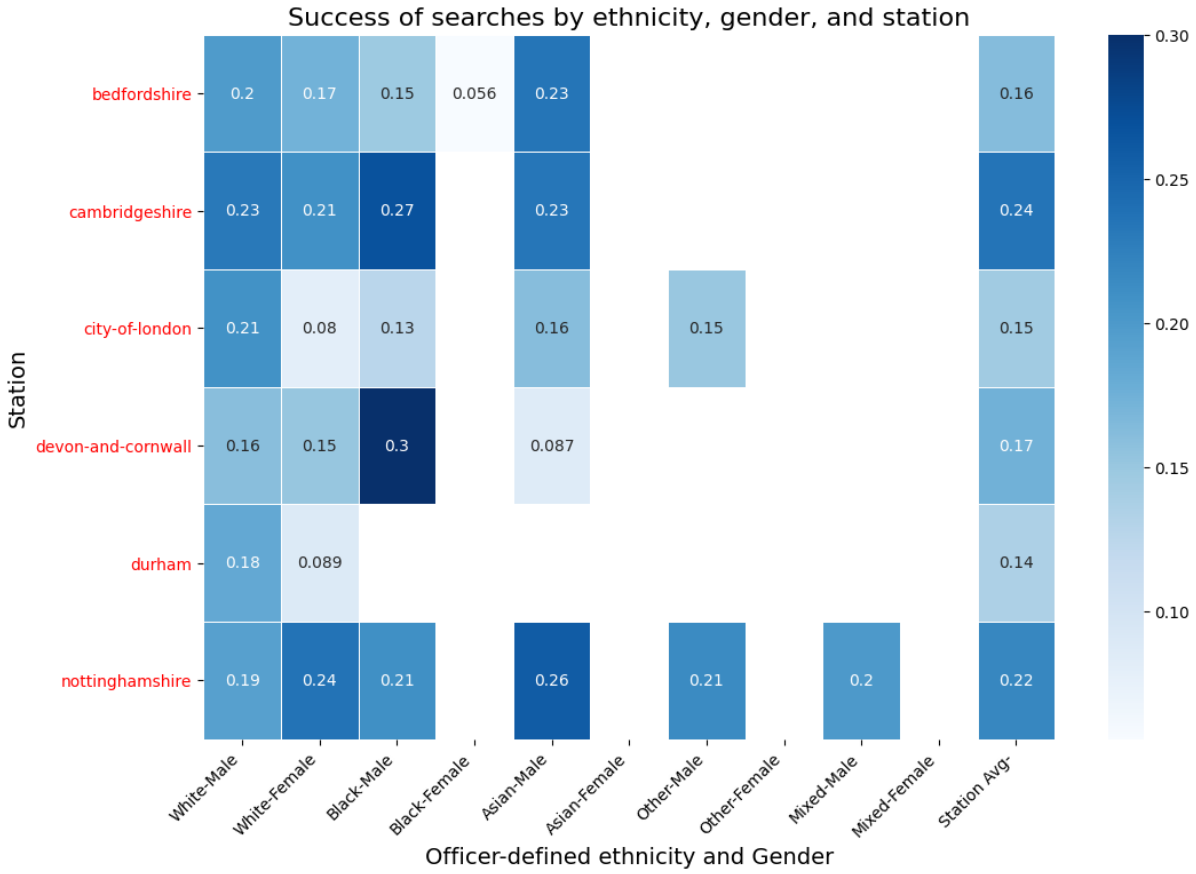

Figure 5.4

Figure 5.5

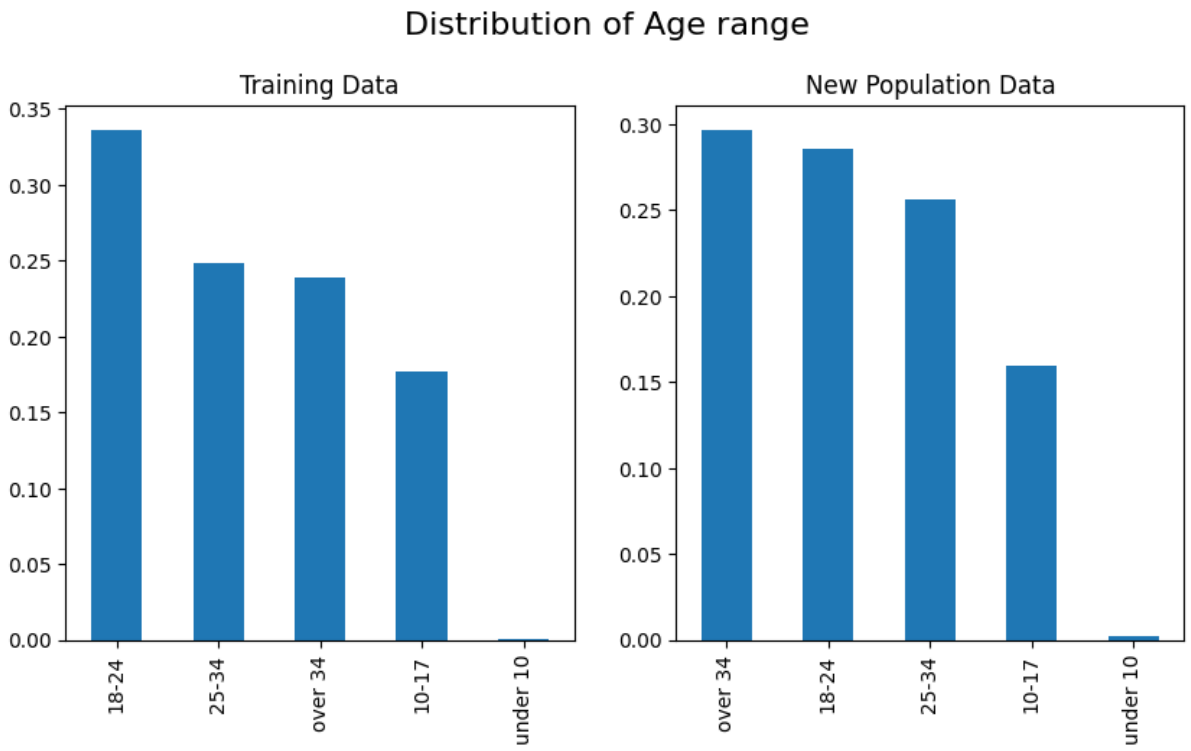

Figure 5.6

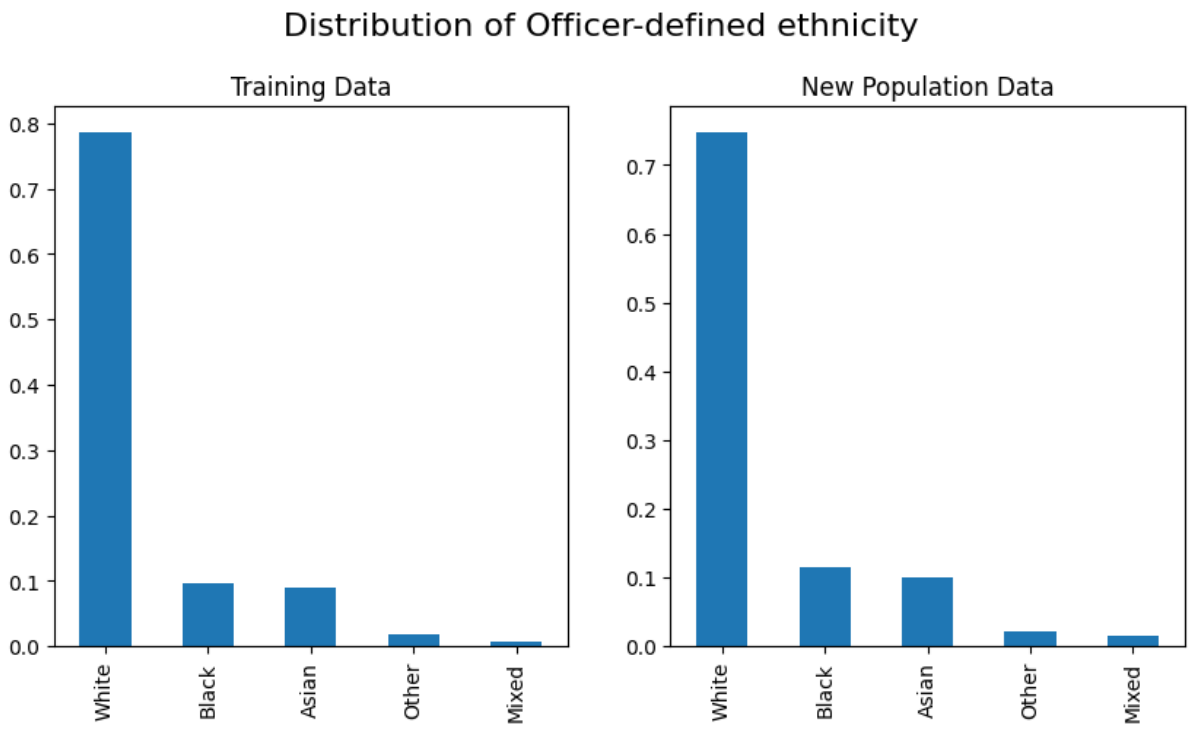
Distribution of Officer-defined ethnicity
Training Data
New Population Data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
White
Black
Asian
Other
Mixed

Figure 5.7

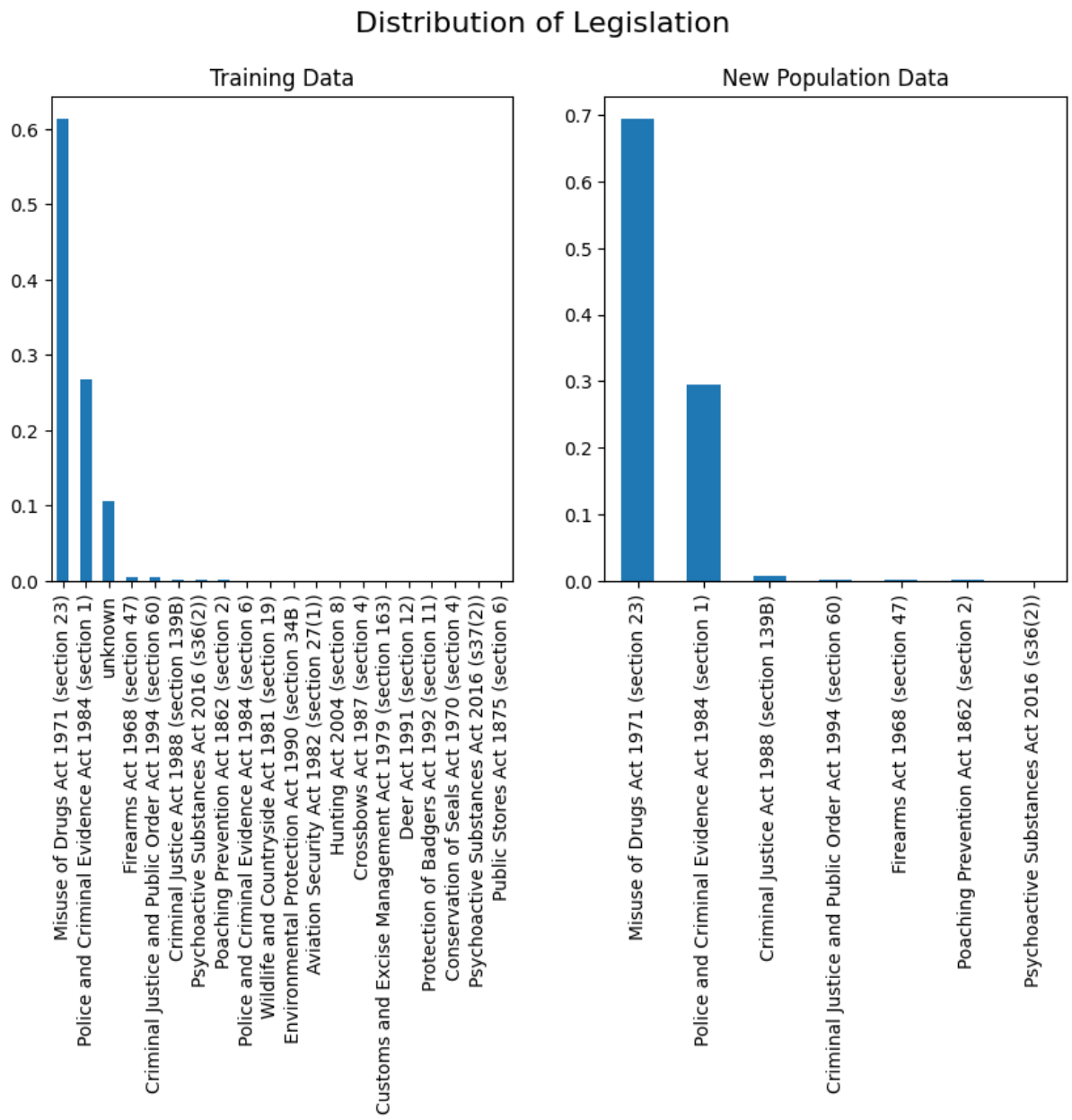
Distribution of Legislation
Training Data
New Population Data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Misuse of Drugs Act 1971 (section 23)
Police and Criminal Evidence Act 1984 (section 1)
unknown
Firearms Act 1968 (section 47)
Criminal Justice and Public Order Act 1994 (section 60)
Criminal Justice Act 1988 (section 139B)
Psychoactive Substances Act 2016 (s36(2))
Poaching Prevention Act 1862 (section 2)
Police and Criminal Evidence Act 1984 (section 6)
Wildlife and Countryside Act 1981 (section 19)
Environmental Protection Act 1990 (section 34B)
Aviation Security Act 1982 (section 27(1))
Hunting Act 2004 (section 8)
Crossbows Act 1987 (section 4)
Customs and Excise Management Act 1979 (section 163)
Deer Act 1991 (section 12)
Protection of Badgers Act 1992 (section 11)
Conservation of Seals Act 1970 (section 4)
Psychoactive Substances Act 2016 (s37(2))
Public Stores Act 1875 (section 6)
Misuse of Drugs Act 1971 (section 23)
Police and Criminal Evidence Act 1984 (section 1)
Criminal Justice Act 1988 (section 139B)
Criminal Justice and Public Order Act 1994 (section 60)
Firearms Act 1968 (section 47)
Poaching Prevention Act 1862 (section 2)
Psychoactive Substances Act 2016 (s36(2))

Figure 5.8

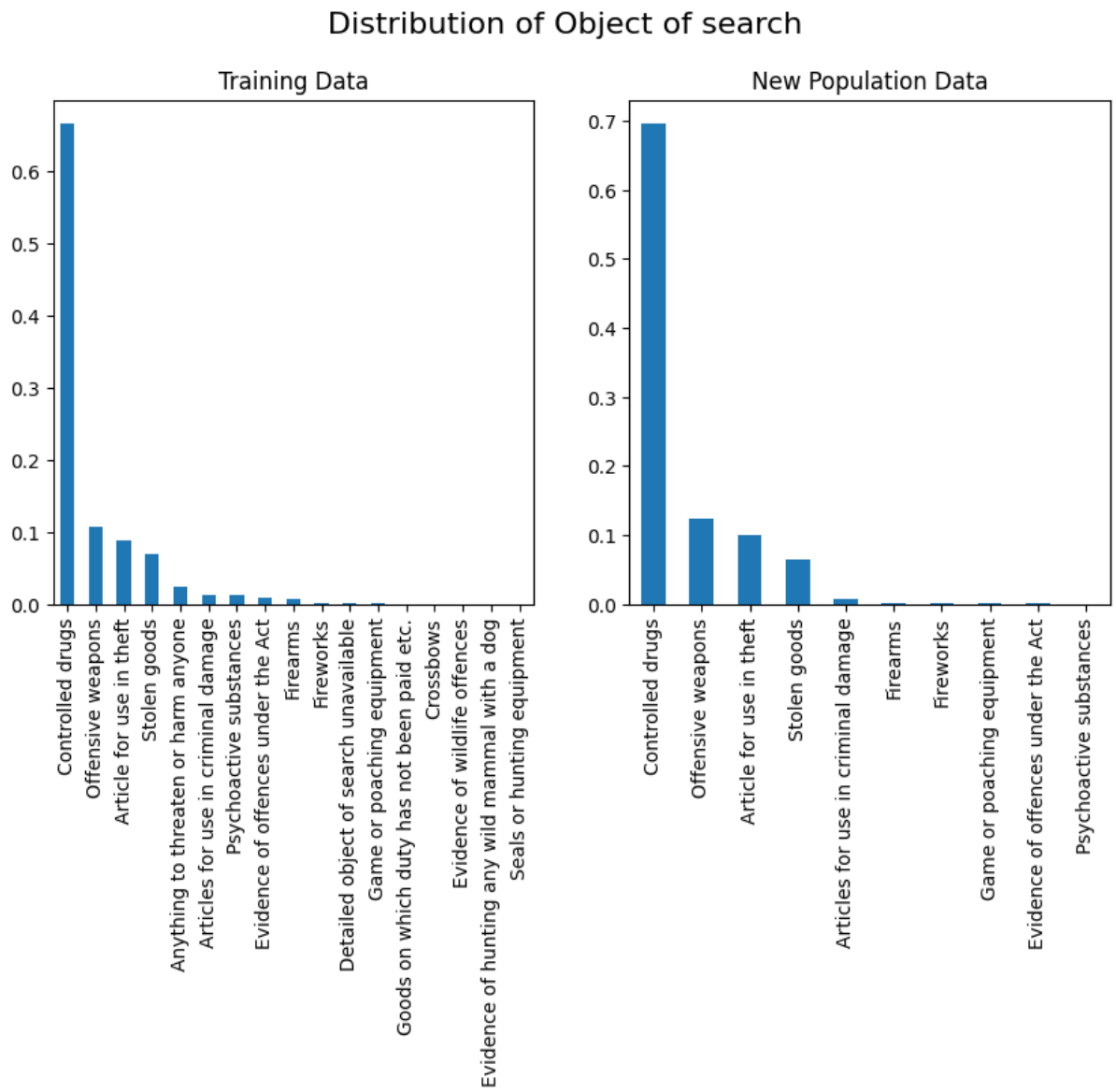
Distribution of Object of search
Training Data
New Population Data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Controlled drugs
Offensive weapons
Article for use in theft
Stolen goods
Anything to threaten or harm anyone
Articles for use in criminal damage
Psychoactive substances
Evidence of offences under the Act
Firearms
Fireworks
Detailed object of search unavailable
Game or poaching equipment
Goods on which duty has not been paid etc.
Crossbows
Evidence of wildlife offences
Evidence of hunting any wild mammal with a dog
Seals or hunting equipment

Figure 5.9

Figure 5.10

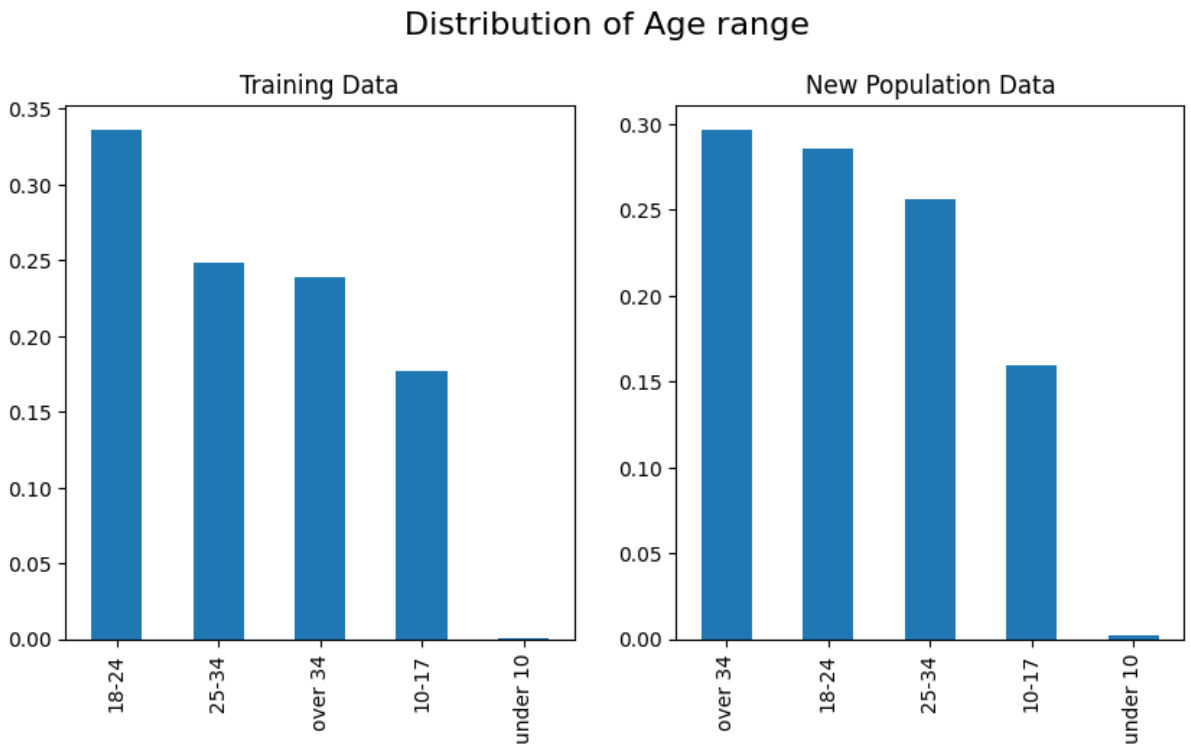

Figure 5.11

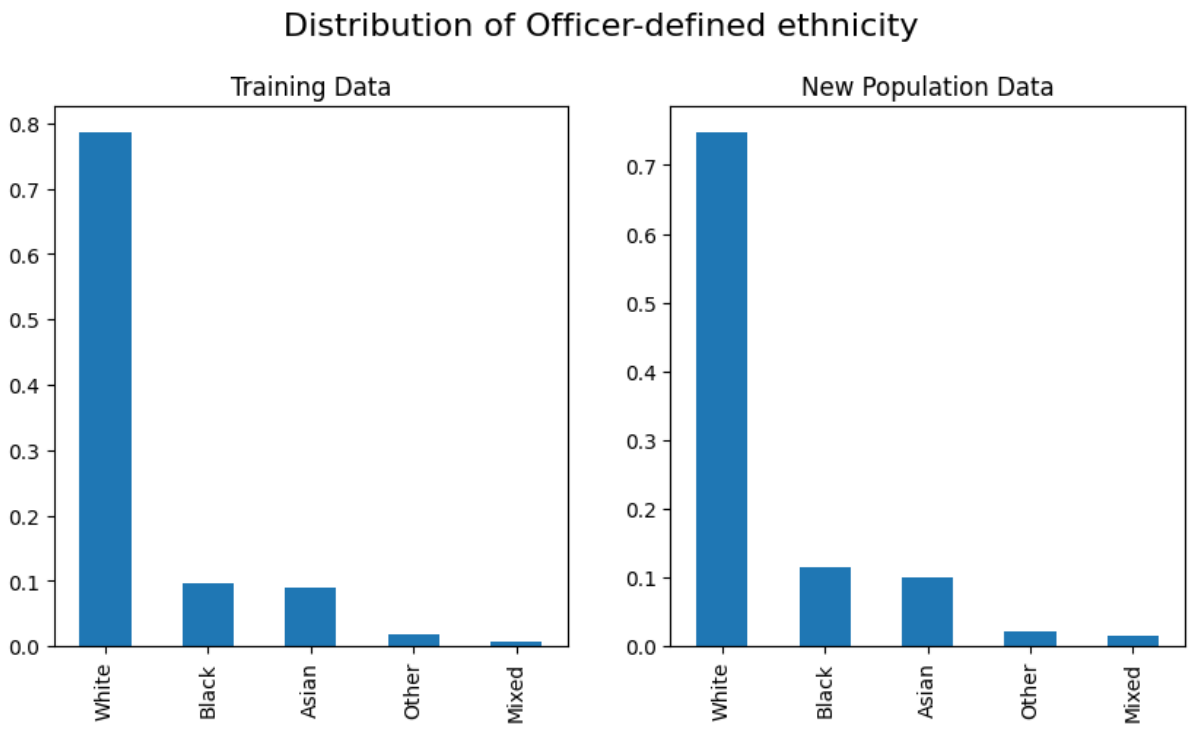
Distribution of Officer-defined ethnicity
Training Data
New Population Data
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
White
Black
Asian
Other
Mixed

Figure 5.12

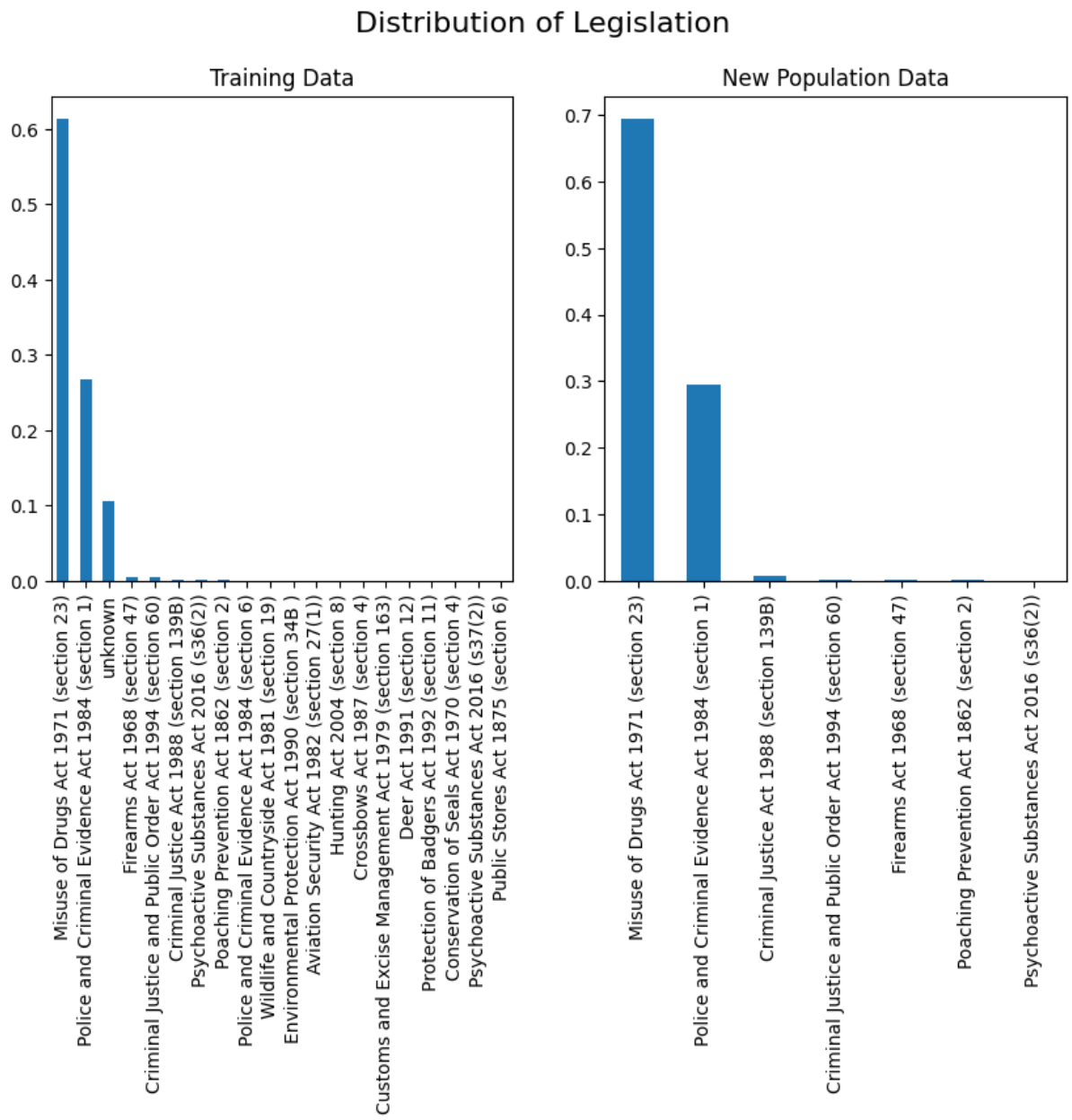
Distribution of Legislation
Training Data
New Population Data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Misuse of Drugs Act 1971 (section 23)
Police and Criminal Evidence Act 1984 (section 1)
unknown
Firearms Act 1968 (section 47)
Criminal Justice and Public Order Act 1994 (section 60)
Criminal Justice Act 1988 (section 139B)
Psychoactive Substances Act 2016 (s36(2))
Poaching Prevention Act 1862 (section 2)
Police and Criminal Evidence Act 1984 (section 6)
Wildlife and Countryside Act 1981 (section 19)
Environmental Protection Act 1990 (section 34B)
Aviation Security Act 1982 (section 27(1))
Hunting Act 2004 (section 8)
Crossbows Act 1987 (section 4)
Customs and Excise Management Act 1979 (section 163)
Deer Act 1991 (section 12)
Protection of Badgers Act 1992 (section 11)
Conservation of Seals Act 1970 (section 4)
Psychoactive Substances Act 2016 (s37(2))
Public Stores Act 1875 (section 6)
Misuse of Drugs Act 1971 (section 23)
Police and Criminal Evidence Act 1984 (section 1)
Criminal Justice Act 1988 (section 139B)
Criminal Justice and Public Order Act 1994 (section 60)
Firearms Act 1968 (section 47)
Poaching Prevention Act 1862 (section 2)
Psychoactive Substances Act 2016 (s36(2))

Figure 5.13

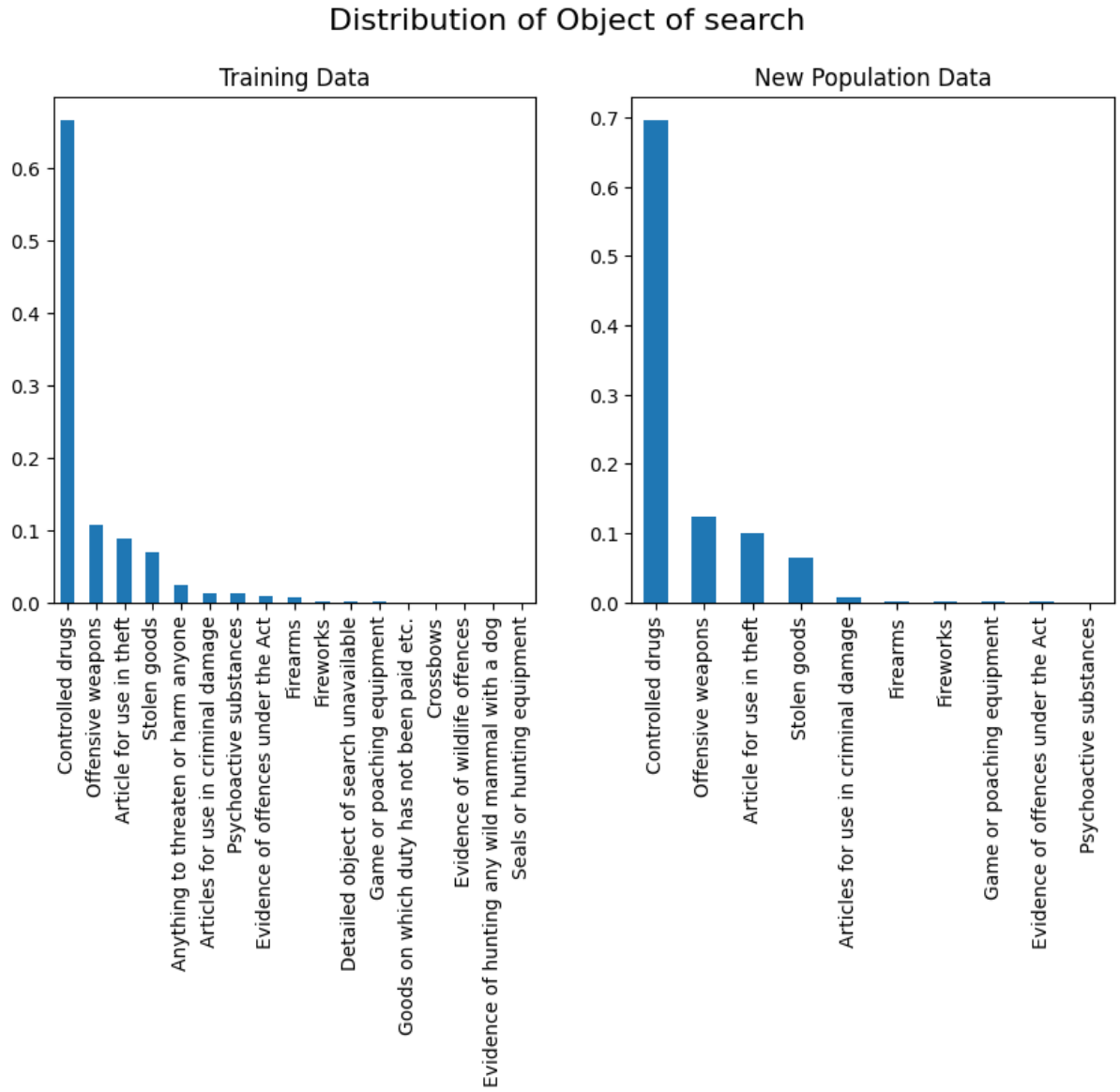
Distribution of Object of search
Training Data
New Population Data
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Controlled drugs
Offensive weapons
Article for use in theft
Stolen goods
Anything to threaten or harm anyone
Articles for use in criminal damage
Psychoactive substances
Evidence of offences under the Act
Firearms
Fireworks
Detailed object of search unavailable
Game or poaching equipment
Goods on which duty has not been paid etc.
Crossbows
Evidence of wildlife offences
Evidence of hunting any wild mammal with a dog
Seals or hunting equipment

Figure 5.14

Table number 5.1

	F1 score	Recall	Precision	Average difference (age, gender , ethnicity)
LogisticRegression	0.34	0.91	0.21	0.17
RandomForestClassifier	0.33	0.90	0.21	0.17
LGBMClassifier	0.35	0.90	0.22	0.19
LogisticRegression (with undersample)	0.34	0.91	0.21	0.17
RandomForestClassifier (with undersample)	0.33	0.90	0.21	0.17
LGBMClassifier (with undersample)	0.34	0.90	0.21	0.18
LogisticRegression (with sensitive info)	0.35	0.90	0.22	0.18
RandomForestClassifier (with sensitive info)	0.34	0.91	0.21	0.17
LGBMClassifier (with sensitive info)	0.36	0.91	0.22	0.19
LogisticRegression (with Latitude and Longitude)	0.35	0.85	0.22	0.19
LogisticRegression (with sensitive info and undersample)	0.35	0.91	0.22	0.19
RandomForestClassifier (with sensitive info and undersample)	0.34	0.90	0.21	0.17
LGBMClassifier (with sensitive info and undersample)	0.36	0.91	0.22	0.20

Table number 5.2

	F1 score	Recall	Precision	Average difference (age, gender , ethnicity)
LogisticRegression (with sensitive info)	0.38	0.96	0.23	0.17
RandomForestClassifier (with sensitive info)	0.36	1	0.22	0.16
LGBMClassifier (with sensitive info)	0.38	0.94	0.24	0.18