Capstone Batch 4 - client briefing

Email from your head of engineering

Dear new employee
Welcome to the team, and congratulations on passing the grueling interview process! We're glad to have you as the newest member of our team here at *Awkward Problem Solutions™.*

As I'm sure you know from the interview process, we are a consultancy that tackles the hard data science problems no-one else will touch. We already have your first assignment. Don't let us down.
o
The United Kingdom Department of Police has hired us to help solve some delicate problems. They have a stop and search policy for every police department, that's supposed to ensure that officers stop people and cars when there is probable cause. There have been accusations in the press that the police tend to stop and search certain minorities at a higher rate than others.

Another sensitive area is related to the question of asking suspects to remove articles of clothing for search, in which the accusations are that women of certain age groups and ethnicities are searched more than others. I do not need to stress that these accusations are extremely serious, and can lead to the dismissal of officers.

They have hired us to investigate whether the data proves any of these claims. If we are successful in investigating, they want us to create and host an API endpoint for authorizing searches. This will be integrated into a police system that the officer will need to use, ensuring uniformity of decisions. You will be responsible for setting this system up, and running it for a year.

I have attached the email from Dr. Anabelle Wilson, the head of IT from the Department of Police. Dr. Wilson will be your primary point of contact during this project. You must ensure that you disambiguate all requirements, as I believe the project may still be under-scoped in some aspects.

As you may imagine, this project is a huge deal for our consultancy, we're counting on you to deliver on this one. Naturally if you have any questions you can ask me, but I'm counting on you to lead and execute from start to finish.

Yours,
Henry
VP of Data Science
Awkward Problem Solutions™.

Email from Dr. Annabelle Wilson

Dear Henry

I'm happy to hear that we'll be working together again. As you know we have been having some unfortunate coverage in the press regarding our stop and search operations. I'm unaware of how credible these claims are, and want you as an external party to investigate and present an objective report.

Additionally we want your company to design and implement a system that our officers will need to use to approve the stopping of a person/car. It will be integrated into our own internal approval system, so all we need is a REST Api that our own code can call.

Regarding data, we've been collecting information on the search operations since 2017, with the following:

- `Age range`
- `Date`
- `Gender`
- `Latitude`
- `Legislation`
- `Longitude`
- `Object of search`
- `Officer-defined ethnicity`
- `Outcome`
- `Outcome linked to object of search`
- `Part of a policing operation`
- `Removal of more than just outer clothing`
- `Self-defined ethnicity`
- `Type`
- `Station`

The **training dataset** has approximately 660,000 observations, I expect this should be enough.

Regarding the analysis, we want you to search for evidence that any of our stations may be discriminating on gender, ethnicity or age regarding who they chose to stop, and also who they ask for any clothing to be removed.

As you may imagine these are very delicate topics, and we want to ensure that we treat both the public and our officers with respect. We therefore need you to propose what metrics would suggest that anything unacceptable is being observed.

Naturally there may be trends, so if something seems to have already been corrected it is less relevant than if it still occurs. Some training and changes of policy may impact the data over time.

Our expectation from experience is that some populations exhibit higher levels of delinquency than others, but would expect that the success rate of the searches would not vary significantly between populations.

The report will be read both by me and by members of our policy committee, please use this structure, and ensure that each section is written for the relevant audience.

After the analysis is completed and your report has been understood we will run a proof of concept with your company for the use of your API for a few of our police stations.

The main objective with this API is that searches are performed only when there is more than 10% likelihood that the search will be successful. We will evaluate this per station and per search objective, as we believe that some stations over search on certain search types.

We also expect that your API should be able to level the discovery rate without significantly diminishing our overall ability to detect offences. More explicitly, we want to level the discovery rate between ethnicities for every station and for every search objective.

The data from the year 2020 will not be in the training set, and we will call it as if it is live with data from that year. We won't be sending you any requests from searches we didn't perform, as naturally we wouldn't know the outcome.

From your last email I understand you've hired a specialist to deal with this project. I'm very much looking forward to working with them, and am of course available to answer their questions as they occur.

Best regards
Dr. Annabelle Wilson
Head of IT
Department of Police, Her Majesty's Government, UK

Email from Dr. Annabelle Wilson (Jan 25th)

Dear <your name>

Thank you for your thoughtful email with your questions, I'm happy to see you have already got a solid understanding of our data and challenges. I've attempted to answer the main questions, but will be going over your remaining questions over the course of the next few days. Please find my answers below.

What should be considered a successful search?

A search is considered successful if the outcome is positive, **and** is related to the search. The positive outcomes are:

- `Local resolution`
- `Community resolution`
- `Offender given drugs possession warning`
- `Khat or Cannabis warning`
- `Caution (simple or conditional)`
- `Offender given penalty notice`
- `Arrest`
- `Penalty Notice for Disorder`
- `Suspected psychoactive substances seized - No further action`
- `Summons / charged by post`
- `Article found - Detailed outcome unavailable`
- `Offender cautioned`
- `Suspect arrested`
- `Suspect summonsed to court`

The flag `"Outcome linked to object of search"` determines whether the search was relevant for this outcome.

Do we use the self identified ethnicity or police defined ethnicity?

You should use police defined ethnicity, as it is the only data the officer has at the moment when they make the request to search. In your analysis we expect you to check whether these generally match the self-definition, or if there is any significant difference.

What do we do when we receive unknown ethnicities at runtime? If a new ethnicity that we have not seen in the training dataset arrives after going live, what do we do? Reject prediction? Fill with 'other'?

I must admit this is an option I wish we hadn't added to the police application, but it seems that officers have indeed used "other" in some circumstances. We trust your best judgement on this one.

The Metropolitan station (which is very large and has lots of data in the training set) has the features Outcome linked to object of search and Removal of outer clothing without any data (always missing). Is this a known problem, and if so how should we proceed?

Thank you for bringing this to our attention, I've contacted the administration at the Metropolitan and asked that they fix their data entry. Please include this in your report, and do not use the Metropolitan station's data for training your models. They will not be in the test set.

Sometimes the Outcome linked to object of search column has no data, is this acceptable?

If there was a search and the outcome has not been written please consider it to be False. We've found that officers tend to write the outcomes when they find something, but forget to go back to the application and fill in a None when they don't find anything.

Should Removal of more than just outer clothing be filled with False?

Yes, except when it's just a vehicle search, in which case it makes no sense and should be kept as NaN. Otherwise it's considered a data imputation error.

Best regards
Dr. Annabelle Wilson
Head of IT
Department of Police, Her Majesty's Government, UK

Email from Dr. Annabelle Wilson (Jan 27th)

Dear <your name>

Please find the answers to more of your questions below.

For each of the columns in the provided dataset, would it be possible to provide a definition and a list of all possible values?

> While we completely understand the validity of this request I'm afraid that has not been conducted by our IT department, and so we are counting on you to give us your best interpretation of the data as it currently exists.

There seems to be some significant change in the data over time, both in volume and changes in the features, is this expected?

> Yes, this is expected due to changes both in reporting practices and also due to the change in crime over time with the economy, etc.

How should we deal with mismatches between 'self-defined ethnicity' and 'officer-defined ethnicity'?

> For the purposes of model training and of analysis of potential discrimination please use the officer reported values. We are however interested in knowing how the officer defined ethnicity differs from the self-identification, so please add that to your report.

How should we deal with missing data in clothes removal searches? Is it expected that so much is missing?

> We've asked officers to mark that field as False when they didn't do a clothes search, but we presume it's safe to assume that sometimes they will not have marked it, and it will mean they didn't search.

There are records where 'Outcome' is 'A no further action disposal' but 'Outcome linked to object of search' is True.

> Thank you for bringing this to our attention, it suggests that there are data entry issues. Please detail this in your report so that we can take action with the stations where this is occurring.

In the briefing we were told that "the success rate of the searches should not vary significantly between populations". Is there (A) a defined target for this variation? Also, (B) how should we define population in this case? What would be (C) an acceptable difference between police station search rates?

> This is a very interesting question. We are not data scientists here at the IT department, so please tell us in your report if our expectation is unreasonable. We would hope that (A) there would not be a discrepancy of more than 5 percentage points between population sub-groups, which would be defined as (B) a (station, ethnicity, gender) tuple, and that the discrepancy between stations (average per station) would not be larger than 10 percentage points. We have no way to tell whether our expectations are realistic, of course, but trust that

you can guide us in this process. Please note that we are only concerned about age when deciding about clothes removal. For now we are not interested in ages when deciding whether or not to conduct a search.

How should we measure discrimination? As an imbalance in search success rate between groups or as an imbalance in the relative search rate between groups?

Our current priority is in making sure no population is over-searched, which we are defining as having equal success rates. We know that due to correlation with economic status different groups will not have the same search rate, which is acceptable for now.

How should we deal with the gender "Other" regarding discrimination?

Please exclude it from the analysis, unless you find some very strange pattern that is worth reporting. Regarding using it in modelling please use your best judgement.

Our preliminary analysis found evidence that searches are concentrated on given geographical locations. Is this acceptable or can it be considered a form of discrimination?

We don't necessarily find that unexpected, as our stop operations are generally conducted in specific roads or certain buildings (for instance around nightclubs). Please let us know in your report if you find this may introduce any biases.

When we receive the ground truth for the predictions we made?

There will be a moment before Feb 12th when we will send you the results of the predictions, so that there is a chance to correct if needed.

What does it mean for the "Part of a policing operation" feature to be missing?

Some searches are conducted as part of a police operation, and others may be conducted opportunistically. For instance if there is a car crash and the driver seems under the influence of drugs the officer may conduct an ad-hoc search.

What should we do regarding missing data in the other features? (see questions above)

Some questions about missing data have already been answered, but for the other features please use your expertise to decide how to deal with missing data, and justify it in your report.

In the training data set there are objects which were never searched as part of a policing operation Should we discard these searches from the analysis ?

Please tell me if your expertise suggests otherwise, but it would seem that that data could be relevant, as not all searches happen in a policing operation.

Could you please disambiguate what "Object of search'" means in this dataset?

You can think of it as synonymous to "objective of the search.'

The data suggests that there are subjects with ages under 10

Thank you for bringing this to our attention. Please detail this in your report so that we can take any necessary action.

Could you clarify what it means to conduct a search only when there is more than 10% likelihood that the search will be successful?

Ideally we want a system that gives a good probability of the search being successful, and if the probability is lower than 10% then the search should not happen. Ideally that will reduce the cases where the officer should "obviously" not search. Please let us know what the best way to measure this is in your report, as we suppose you must have some smart way to turn this data into probabilities.

Is there a minimum number of operations to consider a police station?

Actually that may be worth considering at a population sub-group level, or (station, ethnicity, gender) tuples. Please consider that if they are smaller than 30 people then we have no significance, or suggest an alternative metric based on your expertise. Again, please note that we are only concerned about age when deciding about clothes removal.

Given that we are giving "Go / No go" answers to officers who would otherwise do the search, it is inevitable that the discovery rate will go down, even if the search success rate increases. What would be a minimum Search Success Rate for the app? And what would be a minimum detection rate?

We are hoping this pilot project will give us an idea of what is possible, and we are of course aware that this will slightly lower our discovery rate. I've been able to convince my own leadership team at the ministry to let the IT department run this experiment, and we are counting on you to help us show them that this improving success rate is doable without sacrificing too much discovery rate.

There seems to be a lot of discrepancy between stations. Is this something we should attempt to correct?

Ideally yes, we're hoping that having a unified policy will in part correct differences between stations. Having said that we're more focused in minimizing differences between sub-populations (see definition above), so please let us know in your report what a good trade-off looks like.

Is it equally important to detect grave offenses (e.g. discovery of weapons) and the ones that would have only resulted in a warning? And should we treat all legislation categories as equal?

That is a very good point, but for this proof-of-concept please consider all offenses equal in value for the sake of simplicity.

In order to assess if the data samples are representative of the group demographics for the total population, can we use the data from 2011 England and Wales census?

This is an interesting idea, but for the moment it seems like overkill. If we can level the success rates we would already have a good improvement. Please see the question above about imbalances.

The data changes over time, as was mentioned in the briefing. Are there any specific dates when such changes are expected to have occurred? (changes in IT systems or reporting practices)

Again this is a fair question, but we at HQ don't actually have visibility over the individual reporting systems at the different stations. If there are any large changes on particular dates we would be very interested in learning about it in your report.

Can you confirm that each observation represents a search performed on some individual?

The searches can be individual, vehicle, or individual and vehicle. For the purposes of our proof of concept, we can assume that if any search of an individual in the vehicle was successful then it would be reported as such.

How important is explainability in that modelling process? In other words, is it more important to be able to explain the reasons behind why a car/person is selected to be searched, or is it more important to maximise the amount of contraband found whatever the reasons?

For the current proof of concept we are mostly interested in knowing what is possible, so as long as you are able to debug your model and understand what is going wrong if needed please assume you will not have to show up the internal workings. If the proof of concept is successful and we move to production that will probably have to change, thank you for bringing this to my attention.

Is "search objective" the same meaning as "search type"?

Search objective shows up as "search object", and search type as "Type" in the dataset.

The available Latitude Longitude, is related with the 'station' column?'

There will probably be some correlation, but we'd expect that it was the location where the officer was then they conducted the search. Please let us know in your report if this is not what the data suggests.

How should we interpret "'A no further action disposal'"?

Interpret it as "either proven innocent, or insufficient evidence to proceed".

What is the current stop and search criteria?

It has varied over time and between jurisdiction. We are attempting to unify it across our stations, which is where hopefully this project will help us.

Would it be acceptable to approve searches only when officer and subject of search have the same ethnicity?

> There are no plans for anything of this nature at this time, I believe that would go against UK law.

Could you provide us data on the gender of the officer? This could improve a definition of the policy on non gender discrimination for cloth removal.

> Unfortunately we do not have any data on this.

Regarding the analysis of whether there is improper behavior regarding clothing to be removed, should all the protected classes be analysed?

> We’ve had issues in the press suggesting these are more prevalent with women of some ages, so please check across gender, age and ethnicity.

How much difference would be considered acceptable between stations?

> We don’t have a strong prior as to this, please report the hard numbers and give us your assessment of whether it seems significant enough to warrant further investigation.

Is our model expected to authorize removal of clothing, or just whether the officers can perform the search?

> No, the model will just be about search vs non-search, this part is purely an analysis.

There seem to be stations which aren’t reporting clothes removal information at all. How should we deal with this?

> Thank you for bringing this to our attention. Please detail this in your report so that we can take any necessary action.

Best regards
Dr. Annabelle Wilson
Head of IT
Department of Police, Her Majesty's Government, UK

Email from Dr. Annabelle Wilson (Jan 31st)

Dear <your name>

Please find the answers to more of your questions below.

What will be the API specification?

Your API should have 2 endpoints, **should_search/** and **search_result/**. The API is described below. As an alternative, you can also check our API documentation here.

1) **should_search/**
   This endpoint receives the following content:

```
{
 "observation_id": <string>,
 "Type": <string>,
 "Date": <string>,
 "Part of a policing operation": <boolean>,
 "Latitude": <float>,
 "Longitude": <float>,
 "Gender": <string>,
 "Age range": <string>,
 "Officer-defined ethnicity": <string>,
 "Legislation": <string>,
 "Object of search": <string>,
 "station": <string>
}
```

   If the observation ID already exists on the database, it should return an error message. Otherwise, it should predict an outcome for that observation (whether the officer should search or not):

```
  {"outcome": <boolean>}
```

2) **search_result/**
   This endpoint allows us to input the true outcome of a search, as follows:

```
{
   "observation_id": <string>,
   "outcome": <boolean>
}
```

   If the observation ID is not on the database, an error message should be displayed. If it is, the API should return an object with the observation ID, your predicted outcome and the true outcome given:

```
{
   "observation_id": <string>,
   "outcome": <boolean>,
   ~~"true_outcome": <boolean>~~ "predicted_outcome": <boolean>
}
```

Would it be possible to have a sample query in advance for testing purposes?

Please refer to the API specification link given in the previous answer: https://app.swaggerhub.com/apis/l6837/batch4-capstone/1.0.0.

Is there a requirement for response time from the API?

We expect it wouldn't take more than half a second or so to get a response after the server is up and running.

Will there be any stations in the test set which were absent in the training set?

The features in the test set will be subset of those in the training set. Having said that please make your app able to deal with new stations.

Regarding the features, are they going to contain the same information as the ones in the training set? For instance, could an "Object of Search" which has not been observed before show up in production?

There could be new categoricals in the features, so please take that into consideration when designing the API.

Is there any reason for some days to have much more stop & search than the majority of days? Any special events or specific operations?

This may be related to specific policing operations, but we do not have any insider information on this. Please mention it in your report.

During a policing operation are all vehicles/persons searched? Or is there still a selection by the officers on whom to stop?

Officers will request to search individuals and/or vehicles where they believe there is probable cause and that it is covered by legislation.

What is the meaning behind legislation? Could you provide some context for this?

The legislation field is the legislation under which the search was performed. It should in theory never be missing, but from what I understand it has been found to be missing in the dataset. Please let us know about this in your report.

May I ask how the results of these insights will be used? In case some of the reported issues are true, will there be decisions made on top of them?"

We can not disclose this information. Is there any reason why you feel you would need to know that? If you feel this prevents you from improving the quality of your solution please detail that in your report explaining in what way this would help you.

How would you like us to handle requests not containing racial, gender or age data? Should they be refused until all these fields are input or try to make a prediction based on other fields alone?

Your API should not refuse requests which are valid, but please let us know in your second report (after the model has been used) about any issues with the input data.

We have found that Station BTP has observations all over the country with no particular trend. Can you give us more detail on this station? Does it have a different role than the rest of the stations?

That would be the British Transport Police, who police the railways and light-rail services.

The column "Object of search" appears to have some overlapping items, such as "Offensive weapons" and "Firearms". Since these may pertain to legal terms, we would like clarification if we can agglomerate some of these under broader categories or if the distinction is strict (Firearms may refer to hunting gear, for example).

Generally we make this separation due to the prevalence of knife crime. Offensive weapons are generally knives and other non-firearm weapons. If you find that for the sake of statistical significance there is advantage in joining them we are not opposed to this, but ideally we'd like to keep them separate.

Best regards
Dr. Annabelle Wilson
Head of IT
Department of Police, Her Majesty's Government, UK