Answers

We will have to keep in mind the given [Guidelines](#) for creating the report and holding all the references from our research.

Summary of Case Study

- Lets make few things clear:
- Data: User Feedback Data
- Before answering the how questions, lets get on the same page by answering the 'why' questions.
- Why analyse the performance: Data helps identify problems, learn what is working and have active reception of user's important issues.
- Why we need data Science with this system and data??
- Because: Because feedback links are available on all the gov.uk pages, the advent of spams became inevitable leading to added noise in the data. [Data in Government Blog]([How we are using machine learning to detect GOV.UK feedback spam - Data in government (blog.gov.uk)](#))

  > (their quantity is radical enough (12% of total responses in 2022) to not ignore them while making decisions, they dilute the sensitivity of other genuine issues causing security problems, response delays and missing the real concerns)

- Solution (for excluding these spams out of pool of meaningful feebacks): Filter them. For filtering we first need to establish a way to classifiy each feedback as spam or not spam. Earlier this required manual efforts which resulted in wastage of unncessasy time and money of civil services.

HOW ??:

- If this classification can be done with machine learning and data science, it becomes convinent to process as well as possible to automate the whole system to get clean data.

- Machine learning is a technique where the data learns meaninful insights from the data itself. In our case, civil services chooses to use supervised machine learning, a techique where we have a label to each input instance stating which target our inputs corresponds to. In this case, we have a classification problem using 2 discrete classes: spam and nort spam (ham). Civil services created a training dataset where

each feedback entry was mapped to one of the label. This data curation process although takes time but once done, can be used n-number of times until what we expect is not different from what our data represents.

- The machine learning model learns to improve its performance based on the curated data we just talked about. Once its done with its training, civil servants assessed the working of the model on a unseen dataset (test dataset) and compare the model's predictions with assigned labels. Importantly, we can use these assessment probability scores to demand a high level of confidence in our model's predictions, reducing mislabelling of legitimate feeback. Some ML algorithms can produce statistics on feature importance which tells us which characteristics played how much role in getting the respective predictions.

- In general, Problems suited for ML can be effectively solved with hueristic based methods with less time and data so in this case, as a best practice, they first tried "rule-based" spam detection method on the new data pipeline [Inside Gov.uk] ([Working with user feedback during COVID-19 - Inside GOV.UK (blog.gov.uk)](#)) as one of these heuritic methods to test their performance. Their experimentation resulted that rule-based approach on their data struggle with spam that had subtle combination of indicators, hence they move forward with machine learning while knowing the advantageous areas to focus on.

- Data pipelines to solve different problems different data sources.:
  - Read the Link above.

- Using agile principles to rapidly delivering a working solution.

- Used [Machine learning Canvas]([Machine Learning Canvas — OWNML](#)) to quickly define the scope of their problem

- Using [PyCaret Python Package](#) to automate the training of different classifiers and choosing the best out of them to manually fine tune based project requirements.

- Chose Random Forest Classifier, an ensemble model to get their final predictions

- Canvas helped quantify the complexity of their dataset and the need to standardise the versions used across them. Data version control [DVC](#)is used to achieve such data and model standardisation. → Ensures consistency and reproducibility of results

- For performance analyses: During training we used a loss function, during inference we use confusion matrix for classification tasks to determine how many predictions were wrong or write and in which way. They also used confusion matrix to visualize the occurance of false positives, where real feedback was identified as spam. Further analysing the performance, they ranked individual feature importance to assess the features used in the model for their role and level of impact on model's classification predictions.

- Using all the tools mentioned allowed them to deploy working solutions at pace with streamlined planning and setted priorities. They used rule-based approach to compliment agile development and decide if the current solution can be improved or changed for the good in the next cycle.
- Careful monitoring of the iteration process to tackle spams that have been catered to fool the system and cause "model drift". Enchancement in dataset's size and complexity is a necessary part in strengthening the process. With the incremental complexity and size of the data, they introduced the use of 2 additional performance metrics: precision and recall and determine a classification threshold that strikes an optimum balance between the two.
- Use of open source development is an integral part of their approach as it provides a refined collaboration across teams, test novel techniques and speed up the processing time.