**Scenario Overview**

Financial compliance regulations require organisations to surveil all channels of communication used by their regulated employees. Due to the volume of data, and the number of rules which may trigger an alert, the resources required to review all the alerts is very high. This leads to offshore outsourcing of initial reviews to groups which are not secure and not accurate with their review. This inaccuracy can lead to false negatives which can lead to regulatory backlash and massive fines for the organisation. The goal of machine learning in this case would be to reduce the number of:

* False positives – fewer false positives reduce resources wasted on reviewing unnecessary alerts, possibly making an offshore review team obsolete, thus improving data security.
* False Negatives – these must be minimised at all costs due to the tightening of regulatory investigations world-wide.

**1) What data would you use?**

Our current product is, fundamentally, an archive and search tool with a wide range of supported data sources. The communications data required is already provided by our clients due to their legal obligations to capture and store all their data (depending on the country, for 7+ years). The surveillance product which would utilise machine learning would use this comms data as one of the data sources.

The other source of data would be provided by their existing surveillance workflows which again, they are legally obligated to follow. This would provide us metrics on which alerts are real and which are false alarms.

The rules and scenarios which trigger an alert must be provided as a starting point by an industry expert. These already exist in the current workflow. These are, as standard, simple lexicon searches with proximity constraints on word placement; e.g. ‘buy’ must be followed by ‘Tesla’ with a maximum distance between them of 5 words. One scenario is made up of many of the criteria such as in this example. Each of these criteria get scored, but currently the scores are static and based entirely on the ‘best-guess’ of industry experts.

**2) What are your key input and output variables?**

The input variables in this scenario are the words in the communications and their proximities. The model could be extended to include other factors such as the time of day, normal topics of conversation, tone of voice (in audio comms), HR data like company hierarchy, etc.

The output variable is score given to the object. The score mentioned in the previous section is currently calculated based on the criteria in the scenario/rule which are satisfied by a given communication. This score is then used to determine what type of alert is triggered (mild, medium, severe).

**3) What type of machine learning problem is this?**

This is a non-parametric, supervised, prediction model.

Non-parametric: There is no natural or obvious assumption to be made about the shape of the function given the number of possible combinations of words in messages or criteria in each rule. The volume of communications data and alerts generated by even a small company should be enough to approach this as from a non-parametric perspective.

Supervised: The organisation is required to review any alerts by law as proof they are doing their due diligence in ensuring their staff are acting according to the latest regulations. These alert review outcomes (false positive, valid but not serious, valid and serious, etc.) are stored by the platform. These indicate what the score assigned to the communication should have been and can be used to train the model, making this a supervised learning problem.

Prediction: This was the trickiest to decide as in theory it could be considered either. To contradict what I have said above, it could be a classification problem as the communications could be classified into the different types of alerts the company deals with. However, due to the business requirement of flexibility to change the boundaries of alert categories at will when process or regulations change, a score which can then be input into another system is more appropriate. A continuous score is more suited to a prediction problem than a classification problem.

**4) What steps would you take to solve this problem through machine learning?**

As discussed in the lectures, there are 10 steps in the machine learning process. I will recap on these here.

1. Define the purpose of the ML project – this has been discussed in the scenario overview.
2. Obtain the data set for the analysis – agreements would need to be made with clients to be able to use their private and confidential data sets to be able to train the model. Once these agreements have been signed off, the data is already being collected and stored in the existing product.
3. Explore, clean, and pre-process the data – The communications dataset is cleaned at the point of collection to ensure that all the communications of a particular trader are accounted for, and a full set of records is stored. The communications / input variable data is already processed by the existing product as it is required for search and retrieval. The output variable / alert review outcome data will need to be cleaned as there is no guarantee that the alert review is accurate or complete. It may well be that on a Friday at 4:30pm the review agent decides to click false alarm on every record to be able to go home and start their weekend. This is one example of a case which would need to be identified and handled, possibly by exclusion or by assigning the original score by the existing (non-ML supported) scoring model.
4. Dimension reduction and feature engineering – Communications can often be split up into multiple sections. There can be a header, a body, metadata, footer, disclaimer, etc. It may be found upon investigation that some of these are irrelevant to the alert scoring. For example, it is unlikely that the footer of an email will hold anything worthwhile to review, so all of the words in the footer are excluded from the input.
5. Determine the ML task at hand – This has been determined in question 3 to be a non-parametric, supervised, prediction model.
6. Partition the data – Due to the data being stored entirely in SQL databases, it is very easy to randomly partition the data into identical database structures containing the training, testing and validation sets.
7. Choose the ML techniques – Some initial testing would need to be carried out before I could decide on an ML technique, but I would start with a decision tree to build a regression model. Based on the results of this, I would carry on with the model produced or carry out further research on what other options are available.
8. Use the ML technique – Covered slightly in step 7, the model would need to be trained using a range of hyperparameters to find the (seemingly) optimal model. This will require many models to be trained and evaluated.
9. Interpret the results – Of all the models produced, each will need to be compared to its peers and the best selected.
10. Deploy the ML technique – A result of step 2 would likely be that the clients providing their data for training would want access to the model produced. This would require pushing the resulting model into production globally. This model will be used by existing and future clients to help with identifying real alerts in communications and helping the clients to stay ahead of the regulators.

**5) What might cause missing data in your data set? Which approach outlined in the lecture materials do you think would be most suitable for dealing with missing data, and why?**

There are two side from which data could be missing: input and output variables.

Input variable data could be missing due to a problem with comms collection. Although a massive problem and a logistical nightmare, this is sometimes unavoidable. Another cause of this could be malicious actors who do not want to be surveiled and talk either in person or on unrecorded channels. Once again, this provides an incomplete set of records as only half of a conversation may then be recorded. If the data cannot be recovered, it cannot be included in the surveillance and hence will be excluded from the model. However if the data is partially available, this can still be surveilled, reviewed and therefore included in the dataset for the model.\*\*

The major potential for missing points in the output variable data is down to human error. If the alert reviewers don’t properly review every alert, the unreviewed alerts are missing outcome datapoints. Using an algorithm to fill in the most likely outcome and including this datapoint in the model is possible and suitable. The algorithm which generates the initial alert score before human review should give a decent “best guess” of what score to fill in.

\*\*Note: malicious or purposeful circumvention of surveillance is included in some rule criteria, which will trigger an alert when regulated users try to take a conversation “off line”.