AXA exercise answers – Nishal Chandarana

**Question 1**

Data insights

* The dataset has over 400,000 rows and overall the outcome variable of interest – history of mental health problems - has 30% of rows with “Yes” and 70% with “No”
* Among the variables in the dataset, employment appears to have the one of the largest impact on the mental illness variable. When employed the mental illness rate is approximately 26%, while for unemployed near to 39%.
* Only having high school or associate degree education increase the mental illness rate approximately 5 points (28% -> 33%)
* We see that with income increase there is a large decrease in mental\_illness rate, falling from 38% for the lowest incomes to just 21%. This is likely linked to employment and education.
* Good sleep patterns and healthy diets are linked to a lower mental illness rate, this is intuitive based on a hypothesis that the body and mind’s health are linked
* None of history of substance use, chronic medical conditions, or family history show a strong link to the history of mental illness outcome, somewhat surprising.

**Question 2**

*Part 1 – variable exclusion/inclusion*

Several of the variables in the dataset have plausible causal links to mental illness:

* Age - getting older may make us more or less susceptible to mental illness, or make it more likely that it has been experienced at some point
* Marital status - having a partner may reduce likelihood of mental illness
* Employment - keeping working and productive may be beneficial for mental health
* Income - higher income may increase life satisfaction
* Alcohol consumption - high alcohol use may be a cause or coping mechanism for poor mental health
* Dietary habits - keeping the body healthy with good food may contribute to good mental health
* Sleeping patterns - may be a cause or result of poor mental health
* Physical activity - as with dietary habits, increased physical activity may improve mental health
* Family history - there may be a genetic element or home environment that causes more mental health to be inherited
* Substance abuse history - may be the cause or result of poor mental health

The variable "Name" does not have any plausible link to the outcome and so has not been used

*Part 2 – Model of choice*

I have used a Random Forests Classifier algorithm. It is suitable for this task as we are trying to predict a binary outcome (history of mental illness). Random forests is a flexible algorithm that can handle multiple data types - we have a combination of numerical and categorical variables here. It is capable of finding non linear relationships and interactions in data, which makes it suited to this complex problem of mental health where there are likely many contributing factors and non linear relationships.

*Part 3 – Model performance*

We see the model achieves 56% accuracy on the test set which is moderate. We can see the precision is quite low at 30% for the label=True subset, suggesting there are a lot of false positives - cases where the model predicts a problem in mental health history when there is not one. Meanwhile the precision is 70% for No cases – not necessarily impressive as we could achieve a 70% precision on “No” cases by always predicting “No”. The area under the ROC curve is 60% which is again moderate performance in classification.

*Part 4 – Model biases*

Looking at the feature importance of the model, it appears to highly value three variables which are intrinsically linked: income, employment and education level. We would want to better study the relationship between these variables as they likely have a high correlation. Including all of them in the model may not be optimal - the model is as such biased by these variables.

With more time, it would be beneficial to look at if there is a bias in where the model makes incorrect predictions, this may give some insight into how it can be improved.

**Question 3**

*Limitations*

The Random Forests algorithm isn’t fully interpretable so we cannot understand in detail how exactly the “important” variables are leading to the predictions that it gives. This makes it difficult to make causal inferences. We could either study the explainability of the model with a technique like SHAP or instead explore other machine learning methods like logistic regression that would give a clearer link between a variable’s value and it’s impact on a model prediction.

*Improving model performance*

The Random Forests algorithm has a lot of different parameters so further tuning of them could improve performance. We could remove some of the features that the model does not find useful and as such speed up the model training process and make it easier to scan the model parameter space. We may also want to engineer new features that the model could use to improve performance.

There also needs to be some analysis of an optimal threshold to use for deciding label predictions. We will want to consider the purpose of the model to decide if we want to prioritise high recall of depression cases, or if instead we want to be precise when we do predict depression, avoiding false positives.