**Learning Journal for Specialized Predictive Modelling and Forecasting Project**

Norvin Chandra A0195335Y

**Scope of Work**

Business understanding, building of predictive model, overall concept creation and incorporating Conjoint Analysis into the project

**Business Understanding**

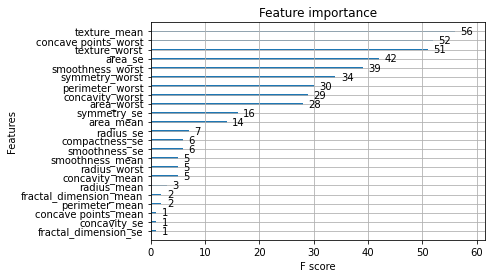
1. Our project deals with improving the predictive capability of traditional Fine Needle Aspiration given its low Sensitivity and Specificity as well as the need for repeated test to achieve acceptable diagnostic accuracy.
2. We did article search on Fine Needle Aspiration (FNA) method for biopsy. From articles written recently, we understand that FNA as an invasive method commonly carried out after a positive result from non-invasive probe actually has low sensitivity of approximately 64% for 1 sample. It improves to 91% for patients who underwent the procedures 3 times. The specificity is even lower at 56%. This is surprising for a commonly carried out procedure to have such a low predictive capability.
3. As a second level diagnostic procedure, FNA should ideally have high specificity (True Negative Rate) / low False Positives. It should have high confidence level in identifying patients who do not have cancer but were originally diagnosed positive in first line diagnostic (False Positives) as healthy persons to prevent unnecessary treatment. This is unlike first line model where the whole purpose is not to miss any cancer patient thus the focus on achieving highly sensitive model / diagnostic protocol.

**Building Prediction Model**

1. After confirming that there is no missing value, I proceed to scale the data using MinMax Scaler before training any model
2. For prediction model, I utilize scikit learn package in Python
3. Modelling attempts can be grouped into 3:
   1. Individual classification models
      1. K Neighbors Classifier (KNN)
      2. Neural Network
      3. Extreme Gradient Boosting (XG Boost)
   2. Stacked Classifier – using scikit learn StackingClassifier function
   3. Hyperparameter tuning and optimized model – utilize Bayesian approach implemented via Optuna package. Bayesian approach utilizes information from previous trial to steer the next trial to area with better chance of getting good result. This result in shorter tuning time and better hyperparameter compared against Grid Search.

**Implementation**

1. Among the models we tried, XGBoost with optimized hyperparameter gave the best combination of Sensitivity of 96.9% and Specificity of 97.2%. In terms of confusion matrix, there are only 3 False Positives and 2 False Negatives.
2. XGBoost allows us to identify importance of each of the features in the training data



The more an attribute / feature is used to make key decision with decision trees, the higher its relative importance.

1. Meanwhile we also have a Stacked Ensemble Classifier which produces Sensitivity of 87.3% and Specificity of 99.1%. This model produces only 1 False Positive but 8 False Negatives. In stacked ensemble classifier, it is not possible to interpret the importance each feature / attribute has in predicting the presence of malignancy.
2. The team had a few rounds of discussion on which model to choose as our final model given the need for high Specificity in second line of test.
3. We concluded that XGBoost has a more balanced Sensitivity and Specificity as well as the possibility to perform basic interpretation of the model through the Feature Importance score. Features with high importance score can then be consulted with medical expert in this area to confirm the role that feature play in diagnosing presence of malignancy. However, since medical field deals with human life, a non-black box model may be required to obtain approval from health authorities. In that case, Logistic Regression with PCA will be next best as it is not a black box model and has next best level of Sensitivity and Specificity.
4. After having a model that can assist traditional FNA, we initially had hard time incorporating what we have learned in this semester into this project. While reading through the lecture note, I got the idea to use Conjoint Analysis as a market positioning study before rolling out enhanced FNA as new service. Some of the team members think that the conjoint should be done while designing enhanced FNA. I proposed to use Conjoint Analysis at the end of project cycle once we have a prototype of enhanced FNA to study the positioning needed to successfully commercialize it. Conjoint Analysis is used to understand patients’ preference and to assist in designing the delivery of enhanced FNA in terms of anaesthesia requirement, price point to achieve wide acceptance and clinical as well as commercial success.

**Conclusion and Individual Lessons**

* Health Analytics has different challenges from our earlier project as in health analytics project we are dealing with the health and wellbeing of human.
* Instead of Accuracy, in health analytics the measure of Sensitivity and Specificity are more important.
* I learned new modelling technique such as Neural Network using Keras package and also stacking different classification model. In addition, I also learned Bayesian hyperparameter optimization using Optuna.
* In this semester I refreshed what I learned in 2nd semester on Predictive Analytics module in terms of model building and combined this with the Conjoint Analysis from the Specialized Predictive analytics that I learn in this semester
* I am grateful to my cooperative team mates and all lecturers for their effort in imparting their knowledge to me in this semester despite the challenges in this extraordinary time.