HSBC Coding Challenge for AUS - 2019

- The Challenge : Building a Probability of Default (PD) model for Wholesale exposures
- Skills to be tested : Programming, Statistics (or AI/ML), Ability to pick up financial concepts
- Description : HSBC will provide a synthetic dataset with 10000 observations. The observations will include financial ratios of companies (which can be obtained from the financial statements) for multiple financial years and the corresponding performance over next 12 months. The performance variable is a 0/1 flag which would be 1 (bad customers) if the company defaulted on a obligation to the bank (i.e. they took a loan from the bank, but didn't repay on time), 0 otherwise (good customers). The target is to create a PD model with a score between 0 to 1000 which will discriminate bad (more prone to default) customers by allocating lower scores to them. The participants are free to choose any tool/programming language and any statistical/AI/ML technique.
- Performance metric : The model needs to be optimized for discrimination between good and bad customers. We usually use AUROC as a performance metric, the participants are free to use their own optimization criterion – however we will use AUROC for the final judgment. The model will also be run on a sample of 3000 observations for which the "default" flag will not be provided to the participants (this is to ensure that the models are not over fitting the provided data) – however to ensure homogeneity of population these 3000 observations will be generated in exact same way as the original 10000. 10 participants would be chosen based on the accuracy metrics, who will need to provide 1 page description of the method used, based on which HSBC will choose top three.
- Logistics: The dataset is provided in an excel file, along with an output template (Validation tab), which the participants will need to fill with the resultant model information. In the "score" column of the "validation" tab participants need to provide a model score (higher score should mean higher credit quality)

HSBC Coding Challenge for AUS - 2019

- The Dataset : This is a synthetic dataset which has been created using a summarized version of corporate financial statements available at HSBC. This leverages on the correlation structure and marginal empirical distribution of the financial ratios to generate the synthetic data. For further details please see the Iman-Conover method as described in https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/simulateMvMatrix.

- Data Quirks: Please note that given that nature of simulation which only takes into account the correlations and marginal distributions seen in the real world, some the individual observations might not seem plausible (for example you will see cases where quick ratio is more than current ratio – which is not possible by definition of these ratios). However, that should not pose a practical challenge in using this to come up with a model which can explain the “default” column using the supplied variables.

- Variables : 33 financial ratios (including some with absolute values e.g. sales). All absolute values are in ‘000 USDs. A brief definition has been provided for each ratio which you can use to test the conceptual soundness of the model (for example we do not expect customer riskiness and net sales to have a positive correlation).

- Performance metric : We will use the AUC measure as described in http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf.