Task 4 Answer Video Transcript

Slide 1: Good Morning/Afternoon

Slide 2: Today I will be taking you through the first iteration of the home loans data science project.

This presentation will cover the Data Science Lifecycle which are the steps to any data science solution, the project overview, the process overview, a description of the data we worked with, the analysis conducted, the modelling and model evaluation process, and finally our recommendations based on our findings.

Slide 3: In attempting this problem we applied an iterative process called CRISP-DM or the Cross Industry Standard Procedure for Data Mining. This process entails that we understand the business and data, prepare the data for modelling, model based on the objectives of the business, and evaluate if the model performance is aligned with the objectives and if they are working towards having the model production listed. If not, we go back to business to gain a better understanding which will assist us in better understanding, therefore better preparing the data to be modelled such that the results are satisfactory.

Slide 4: In understanding the business, it is our understanding that the problem the business is faced with is a long and manual home loans application process that could take applicants up to 3 days in being notified on whether their application is successful or not. The business believes that this process can be reduced to a matter of seconds by automating it and using technologies such as Artificial Intelligence and Machine Learning.

Slide 4: We propose a solution where an applicant can apply for a home loan online, whether it be on his/her phone or PC. The application will consist of the applicant providing us with his or her information just like he or she would on any application and upon doing so, trigger a prediction from our machine learning model on the status of his or her application in a matter of seconds.

Slide 5: Now this model will be trained on historical data which consists of 614 records or historical home loan applications, with 422 being approved and 192 being declined. This data has 13 attributes, 5 of which are numerical and 8 categorical.

Slide 6: Some of our findings from our analysis of the data show that males have more applications and get approved relatively more than their counterparts. Another interesting finding is the correlation of the loan amount of the applicants and their respective incomes.

Slide 7: Having understood the business objective, and conducted an analysis we are comfortable proceeding to the modelling stage. This problem is a binary classification as we are predicting two classes - approval or declination of a loan. To get the best out of the model we will end up selecting we will need to prepare the data accordingly. This preparation includes how we handle missing data, scaling the features to a fixed range which makes the learning process ‘easier’ for some models - this is to ensure that the model does not give more weight to attributes with greater values without consideration of units, and the conversion of categorical attributes to numbers, for example, Female = 0 and Male = 1 as most models only understand numbers. It is important to note that Automated Machine Learning or AutoML requires no or very little data preparation. In our solution we used both AutoML and bespoke/traditional ML.

Furthermore, to ensure that we can evaluate our models fairly, we separate our data into two - the train set which will be the portion the models are fitted/trained on and the test set which will be the portion that the models are evaluated on. 80% was used to train and 20% for tests.

Slide 8: After training the selected model and using AutoML we obtained the following accuracies - 79% for AutoML and 77% for the selected model. It is important to note that accuracy is defined to be the sum of all the correct predictions made divided by all the predictions.

Slide 9: Based on the work conducted we believe that bespoke is better than AutoML even though it performed worse by 2% in our work because we know and understand exactly what went into the model and believe with more iterations, we can achieve better results. It also trains quicker and uses fewer resources.

With a few more iterations of CRISP-DM we believe we will be able to achieve results that are satisfactory and meet the objectives of the business, and finally see our solution integrated with the mobile and web app, and in the hands of our future customers, making the application process for them an easy and fast one!