Hello everyone, I am Tommy Cheung from the March 2021 Data Science stream. Today I will be presenting my capstone project, which is called Monitus. It is a prediction tool about the financial performance of a company based on different source of data.

In this presentation, I will be covering the business value, the basic idea of the project, the data handling process, how I develop my model, how I deploy my app, and lastly the future plans and conclusion.

So why would I start a project on financial statement? When I do my research, I found there is little machine learning study on this particular matter. So, I, as an accountant before, try to apply prior knowledge in my working experience. Also, my prior NLP project has the similar area of study, so this capstone project is an extension from that. Finally, the tool can be used as one of my portfolio to demonstrate the skills I learnt from this course for my career development.

So what is Monitus? In terms of business value, I want to deploy an app which allows people to easily access financial statement information and prediction result in just 1 simple app. Therefore, for investors, they can easily obtain latest F/S in just one-click. And for their investment appraisal purpose, the growth figures predicted can be one of their referencing parameters. And the media can use it as a search engine for the latest company information, and use the bankruptcy prediction engine to search for potential companies that can go bankrupt any soon. Finally, for others like machine learning students, they can explore the different ML models used in the prediction as educational purpose.

So why is Monitus? I chose this name because it is a easy to remember name, and also it is a Latin word that represent advisory and remainder, which echoes to the main theme of this project.

Now straight to the main goal of the project. (5 goals)

First, where to obtain the dataset? For legal documents, I downloaded it using API from SEC of the US government since our major study target is US listed companies. Next the F/S info is requested from financial modeling prep dot.com. Last, for training purpose of bankruptcy problem, I found a Kaggle dataset that fulfills the requirement.

After getting the data, most of my time will be preprocessing data. This process is tedious and challenging because I have to select the best set of data to input to prediction model. The financial statements contains many unnecessary or redundant columns which must be deleted first before training the model, otherwise it will lower the overall accuracy. Besides, the text data from the document needs to be transformed into array before doing any training as well. So let’s look at the Taiwan bankruptcy dataset first.

The dataset already provides a classification between bankrupt and non-bankrupt companies (using 0/1) vs many financial ratios. First I have to remove redundant columns and dropped columns with nulls columns. Also, although most of the figures falls in range between -1 to 1, some figures gets in range of over 10^9, therefore a standardscaler is used to normalize the dataset. Another problem you can see is the data set contains most non-bankrupt company, which makes sense because bankrupt companies are rare. Therefore, I used SMOTE to evenly distribute the dataset.

So, regarding the columns chosen, it should have little correlation with other columns, the ratio has to be useful and can be matched with the US dataset for prediction. Also, many columns have null figures, which make sense in financial world. However, since some of the ratio requires division, dividing null figures may create trouble during modelling. Therefore, I have to either remove those columns or rows respectively.

Here is the heatmap before and after selecting the right columns. As you can see, the prior graph shows many columns with high correlation between each other, which are bad for prediction.

Now the model part. I chose classification models including (…) and use gridsearch CV for hyperparameter tuning, and you can see the result here. Overall the accuracy score is very high except I intentionally train using the dataset without SMOTE, and you can see the difference because the original dataset is so imbalanced. After consideration, I have chosen random forest as the model to be output using pickle to save the model for later.

Next is the earnings call dataset. Some of the work has been done before in my NLP project, but I have enhanced the model by using Countvectorizer and Tfidfvectorizer to make a vocabulary list for prediction in my app.

Here is the result from prediction. As you can see the dividend growth has the best performance on prediction so my model will focus on predicting this parameter.

Finally the investment appraisal part. This is an explorative work as it is very complicated to train figures with no ordinary trend and no pre-trained model found from any online database. I attemped to use ratios and cashflow figures to predict growth. Since Monitus is an ongoing project, I will only deploy this prediction on the app until it has a high enough accuracy.

So I have to do the filtering process like before, and you can see the result here.

I have defined the growth figure using different types of classification, for this one, -1 rate means growth decreased by over 100% over the year, and will be classified as 1. 2 will be rate lower than 1.7%, which is the US GDP growth rate I used as a benchmark. And over 100% growth will be 4 and remaining will be 3. You can see the distribution of the growth ratios I chose here, which is OK for the model.

I chose the normal classification models including (…) and also added a neural network using MLP Classifer. And you can see the result here, the average accuracy scores are all below 50%, which are unsatisfactory. I will further improve it in the future.

So now, let’s see how the app goes!