S&P_TR Prediction and Simulation

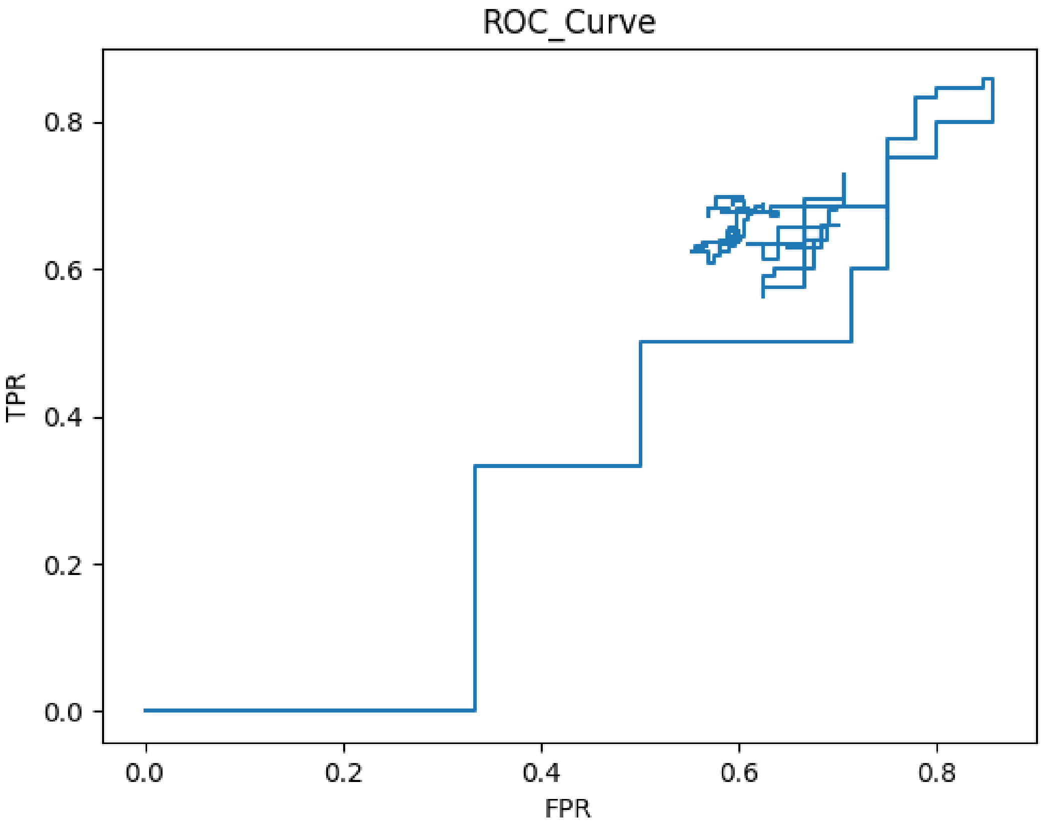
Task1:

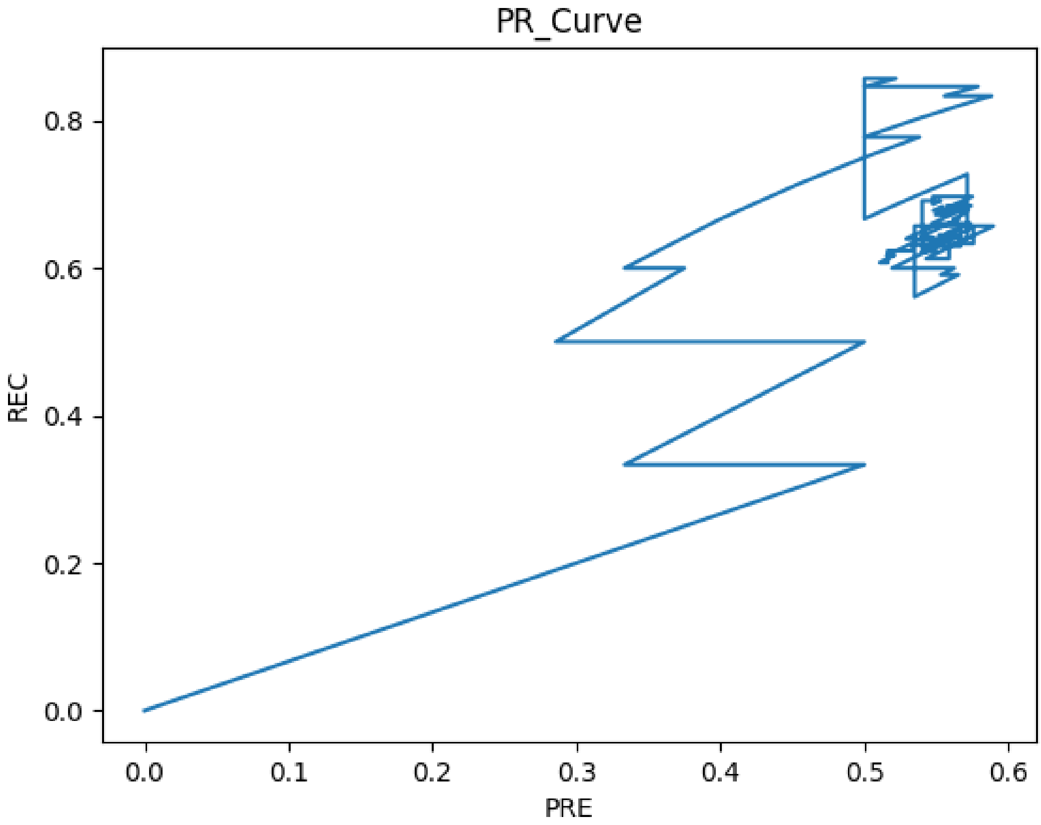
Here a model based on LSTM is built to predict the next 250 points of the close price. The input size is set as 2*250 and the output size is 250.
The dataset contains around 5k years' points, so nearly 20 years. The first 16 years are used as the training data's feature and the 2rd-17th years' data as the training targets. The 18th and 19th years' data as test feature and the last year's data as test targets. The model has trained 30 epoch.
Accuracy, precision, recall and f1 score are calculated as the table below shows. Since positive and negative share the same importance in our prediction goal, accuracy should be paid more attention to, which is 0.548 which is a little bit more than 0.5.

$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	0.548
$\text{Precision} = \frac{TP}{TP+FP}$	0.5679012345679012
$\text{Recall}=\frac{TP}{TP+FN}$	0.6814814814814815
$\text{f1 score}=\frac{2 \times prec \times rec}{prec + rec}$	0.6195286195286195

Also, ROC and PR curves are drawn below to show the performance of the model.

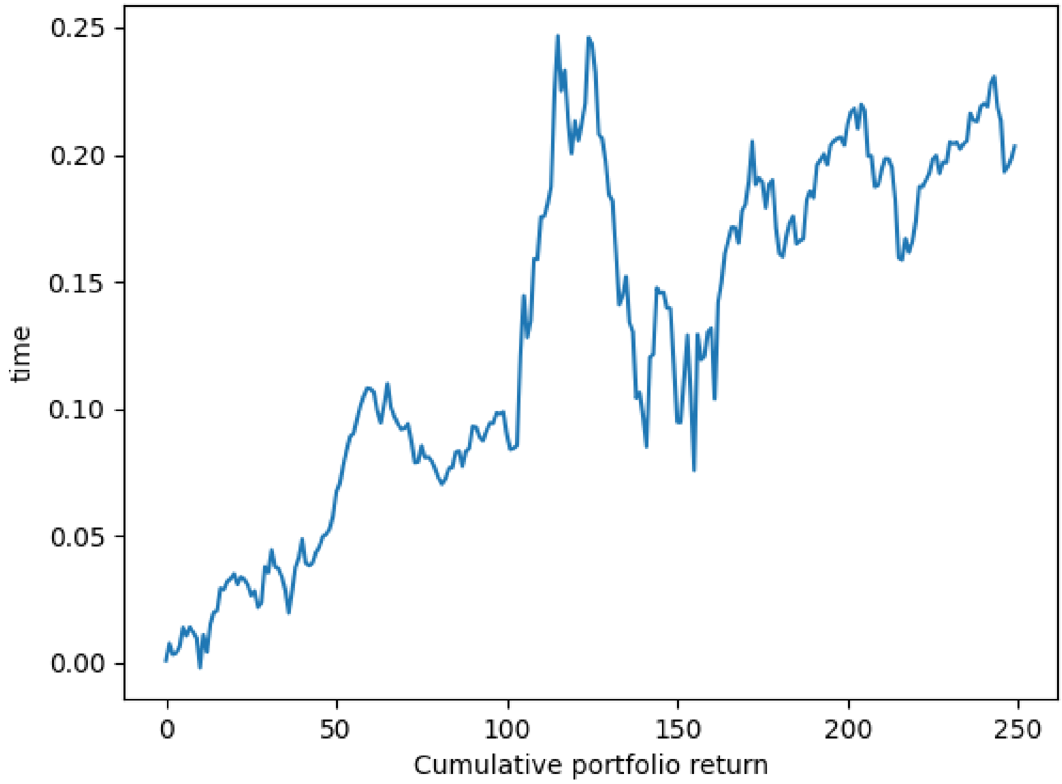
The output data and the rolling cumulative portfolio return are saved in the *result data* table in the shared drive folder.


ROC_Curve
TPR
0.0
0.2
0.4
0.6
0.8
FPR
0.0
0.2
0.4
0.6
0.8


PR_Curve
REC
0.0
0.2
0.4
0.6
0.8
PRE
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Task2:

Backtest is done by calculating the daily portfolio return and the rolling cumulative portfolio return. The curve of rollingCPR is drawn below.



Further Work:

The accuracy is only a little bit more than 0.5, in order to improve the model there are at least 2 methods as below:

The model sets the output size as 250, there is another option to set the output size as 1. When predicting the next year's data, the previous 2 years' data are used to predict the first data. When predicting the second data, the input data should be the previous 2 years' data except for the first entry and append the first prediction. The next 248 predictions analogy.

From the ROC and PR curve, overfitting might have occurred due to too many epochs. The performance of the model might be improved by reducing the number of epochs.