1. Explain the difference between Recurrent Neural Network (RNN) and Long / Short Term memory (LSTM) in layman terms.

Ans: LSTM neural networks are an improvement of Recurrent Neural Networks because they allow the model to retain useful information for a longer period of time without affecting the model’s output parameters in the fitting process. While theoretically RNNs and LSTMs should be able to handle the same kind of tasks, the process of fitting an RNN on output values of its previous iterations leads to the problem of exploding or vanishing gradients, usually the latter which LSTM are adept at handling.

The reason vanishing gradients occur is because as the number of layers within the RNN increase, the amount of reduction achieved in the loss function diminishes, leading to an eventual stagnation in the fitting process, loss of information over many layers and nonoptimal solutions. LSTMs solve this by storing parts of the information in a separate component/cell of the neural network deemed to be useful later on in-place of using all of the output from a previous input. This stored information can then be utilized in determining the output at a later sequential step of the model without having to feed it in the fitting process at every step.

1. You are given a list of quantitative factors that are commonly used to spot/analyze companies. Please explain in detail which machine learning algorithm would you choose and how would you design and train the model to predict returns?
   1. Identify the datatypes used for all the variables (datetime/str/number) and ensure they’re appropriate
   2. Identify null values within the dataset and try to fill them using suitable metrics such as interpolation or lagged values/slices/grouped averages
   3. Plot the distribution of the label and features to identify potential outliers and check whether they’re reasonably justified
   4. Split the dataset into the training and test sets
   5. Utilise and evaluate a series of algorithms including Linear Regression, Decision Tree Regression, Random Forest, CatBoost etc.
   6. Evaluation metrics for the regression algorithms would include the R2 metric and a plot of the predicted vs actual label values
   7. Utilise the best-performing algorithms, identify the most important/contributing features and check whether utilizing them alone improves the linear regression model at all
   8. Check for the possibility that any of the algorithms used fill in for each other’s drawbacks such as sensitivity to outliers, multicollinearity in the case of many features etc.
      1. If possible, try to use some reasonable combination of the algorithms for the final model
2. Referring to the excel attached below, please write a simple code to train the given quarterly data.

* Column E is your target **Y output for the specific start and end period.**
* **Column F and beyond are your factors as at the start period.**

**Using the model built, predict the top 5 companies with the highest returns for the perio**d 31 Dec 20 to 31 Mar 21.



Submit your answers in a python code.