Risk Forecasting with Machine Learning (ML)

1 – Introduction

The goal of this assignment is to forecast risk over a specified time frame using different models, both standard and Machine Learning (ML) models, and compare these models. The models selected for this project were GARCH, Decision Trees, Random Forests, Gradient-Boosted Decision Trees, and Neural Networks. The chosen risk forecast for the models is to predict the forward looking 5-day volatility, i.e. next working week's volatility. The volatility calculation is consistent across all models and uses the standard deviation over a 5-day window, annualised by multiplying by the square root of 252 (typical number of trading days in a year).

This project trains and tests each model based on an 80:20 training-test split for each chosen stock, JPMorgan (ticker: JPM), Apple (ticker: AAPL), and Tesla (ticker: TSLA), with a start date of one day after TSLA's IPO date: 30-June-2010. This split provides a good trade-off between training the model well and assessing its generalisation capability on unseen data. These 3 stocks were specifically chosen due to their reflection of different industries and growth levels, with JPM being a finance firm, AAPL being a more stable technology blue-chip, and TSLA being a more volatile growth stock.

__Success Metrics for Each Model:__

1. **Volatility directional classification %**: predicting whether the next 5d period will have a higher or lower volatility than the last
2. **ROC curves & AUC scores (0-1):** this is more robust than the first since it considers all classification thresholds and is better in the case of dealing with imbalanced datasets where one direction dominates. A high AUC score informs us the model is performing well at distinguishing between positive and negative values, in this case it is predicting whether the next 5d period will increase in volatility or not
3. **Estimate accuracy (Mean Absolute Error & Root Mean Square Error):** concrete measures of volatility prediction error magnitude to gauge the precision of models (the lower the better)

2 – Data Collection, Preparation, & Compound Returns Data

__Data Highlights:__

- **Adjusted Prices:** divided raw price data by the CFACPR to account for stock splits
- **Log Volume:** used log volume to handle the right-skew of trading volumes better
- **Log Returns:** applied a logarithm to simple returns to turn them into compound returns
- **Training-testing Data Split:** split all data into 80% training data (commencing on 30-June-2010) and 20% testing data (commencing on 04-April-2021)
- **5-day Volatility Windows:** this time period was chosen as it reflects a trading week and allowed for better model performance due to the sheer unpredictability and lack of linearity in financial markets which made 1-day predictions ineffective
- **Consistent Model Features:** used the same features across all models (lagged returns, rolling volatilities, return patterns, market features, and feature interactions)

The plot in Figure 1 shows the more stable, lower returns of finance firm JPM, the strong blue-chip technology returns of AAPL, and the more volatile growth stock returns of TSLA.

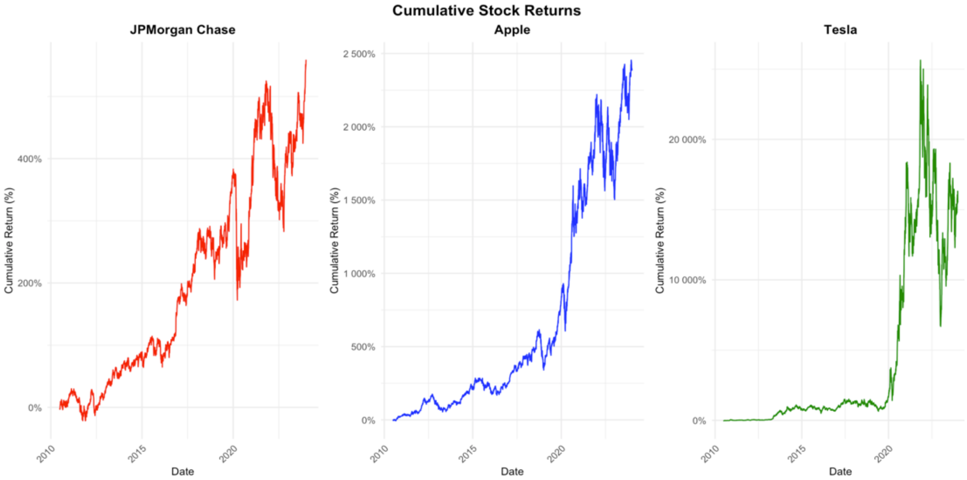


Figure 1: Compound Returns for JPM, AAPL, TSLA Over Time

3 – Summary Statistics for JPM, AAPL, and TSLA

	Mean	Std Dev	Skewness	Kurtosis	Min	Max
JPM	0.00071	0.01754	0.23	13.2	-0.15	0.18
AAPL	0.00110	0.01768	-0.07	8.36	-0.13	0.12
TSLA	0.00212	0.03585	0.32	7.92	-0.21	0.24

Looking at the three stocks it is clear to see that TSLA is the most volatile stock with the highest risk but also the highest mean return, which is to be expected but could reduce the accuracy of our model estimates later.

4 – Typical GARCH (1, 1) Model

A 5-day rolling volatility prediction was chosen due to the poor performance of models on predicting the next day's volatility, as shown by the data from a 1-day GARCH on TSLA where they are inferior to even random chance (50%). In comparison, the 5-day GARCH volatility predictions were much more accurate and had a better directional accuracy. These 5-day GARCH models will be used as a reference to see the improvement from our ML models.

	RMSE	MAE	Directional Accuracy
TSLA 1d Volatility GARCH	0.4051	0.3375	28.13%
JPM 5d Volatility GARCH	0.1121	0.0912	68.3%
AAPL 5d Volatility GARCH	0.1127	0.0896	66.22%
TSLA 5d Volatility GARCH	0.2295	0.1860	68.90%

From the table we can see that the GARCH performed well with reasonable RMSE/MAE values, and quite high directional accuracy. As expected, the volatility of TSLA's stock led to much higher error values, likely due to its more volatile nature.

5 – Decision Trees

Decision trees perform best for binary directional accuracy (upwards or downwards volatility) as they split data into clear decision boundaries and optimise for classification as opposed to continuous predictions. However, their constant outputs struggle to capture smooth relationships thus they perform only reasonably in regression prediction.

	RMSE	MAE	Directional Accuracy
JPM 5d Volatility D.T	0.1036	0.0775	71.30%
AAPL 5d Volatility D.T	0.1113	0.0801	68.55%
TSLA 5d Volatility D.T	0.2237	0.1724	70.23%

The Decision Trees performed slightly better than our GARCH regarding their directional accuracy (c.70%) but performed a lot better in their prediction errors with lower RMSE/MAE.

6 – Random Forests

Random forests are essentially an ensemble of multiple decision trees that combine their outputs to improve their volatility prediction accuracy and reduce overfitting. This model is more robust and generalisable as random forests train each decision tree on a random sample data subset and random features subset.

	RMSE	MAE	Directional Accuracy
JPM 5d Volatility R.F	0.0976	0.0752	71.76%
AAPL 5d Volatility R.F	0.1039	0.0763	65.8%
TSLA 5d Volatility R.F	0.2105	0.1634	65.5%

I would argue the Random Forest ensemble performed better than the Decision Trees. It had a slightly worse directional accuracy compared to our Decision Trees but had improved prediction accuracy (lower RMSE/MAE) compared to our GARCH and Decision Trees.

7 - Gradient-Boosted Decision Trees

Our gradient boosted decision trees build on decision trees by sequentially training a series of trees, where each tree corrects the errors of the previous ones, optimising performance through gradient-based minimisation of a loss function. Here the XGBoost package has been used.

	RMSE	MAE	Directional Accuracy
JPM 5d Volatility G.B.M	0.1008	0.0772	70.53%
AAPL 5d Volatility G.B.M	0.1042	0.0772	68.55%
TSLA 5d Volatility G.B.M	0.2223	0.1739	70.84%

Our Gradient-Boosted Decision Trees scored well regarding classification with high directional accuracy and had the second-best RMSE/MAE values out of all the models.

8 – Neural Networks (Avg. of 3 Iterations)

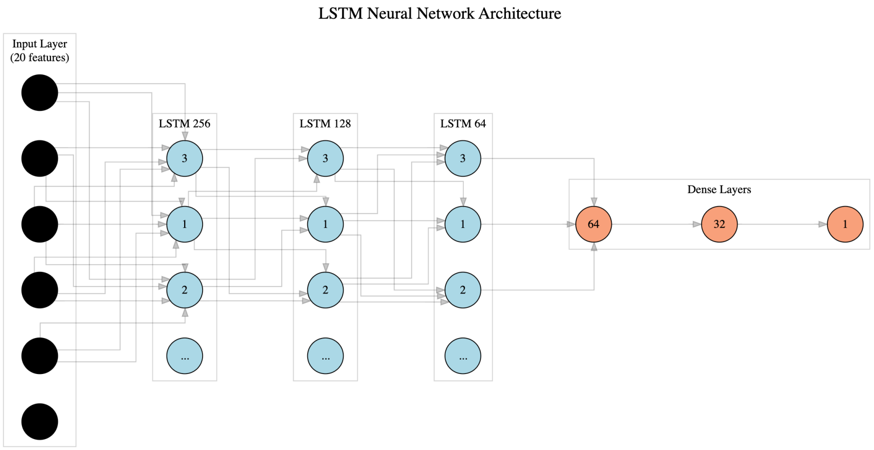
Neural networks leverage their ability to model complex, non-linear relationships in data. Our neural network is taken from the TensorFlow package, and each iteration was run with 300L epochs and patience level of 30L, with 3 runs being averaged to give our results below.

	RMSE	MAE	Directional Accuracy
JPM 5d Volatility N.N	0.136681	0.105735	49.30%
AAPL 5d Volatility N.N	0.133961	0.098634	48.50%
TSLA 5d Volatility N.N	0.330807	0.254626	47.06%

Our Neural Network models performed much worse in both classification and prediction accuracy compared to previous models. I took the average of three runs and saw varying directional accuracies with an outlier of 65% once, however due to the random nature and given that I do not have fixed weights and biases when I re-ran and averaged the outputs, the Neural Network achieved poor results overall.

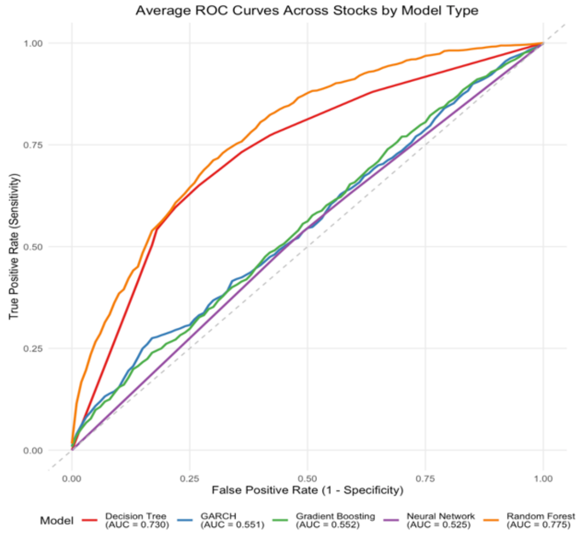
This result is likely due to Neural Networks being the most data-hungry of all our models and them typically performing the worst on tabular data. Having approximately 14 years of data, and then further splitting this into 5-day trading windows for our volatility calculations meant that our sample size was even smaller. This meant that our Neural Networks would try spot patterns in our noisy small financial dataset and thus had more autocorrelation and stationarity issues compared to our Random Forest models, as an example.

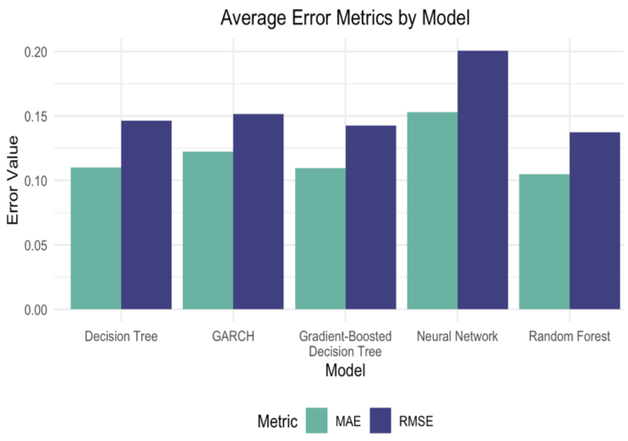
Figure 2: JPM LSTM Neural Network Architecture of an example run



Our LSTM neural network in Figure 2 progressively processes 20 input features through three stacked LSTM layers (256, 128, and 64 units) and three dense layers (64, 32, and 1 neurons) to predict JPMorgan's 5-day stock volatility.

9 – Model Success Comparisons with Classification ROC Curves & Regression RMSE/MAE Bar Graph





- **GARCH:** performs quite poorly in classification as seen by an AUC=0.551 had the second-worst prediction regression errors as seen by the bars in the right graph
- **Decision Tree:** performs second-best in classification with AUC=0.730 due to binary prediction nature of its leaf modelling, but performed moderate in error accuracy with the third-best RMSE & MAE values
- **Random Forest:** performs best in classification with AUC=0.775 and performs the best out of the models in actual regression prediction with the lowest RMSE & MAE values, as seen in the right graph
- **Gradient-Boosted Trees:** performs moderately in classification with AUC=0.552 but performs well with the second-best RMSE & MAE values
- **Neural Network:** performed the worst in classification with AUC=0.525 and in regression prediction with the worst RMSE & MAE values

10 – Conclusion & Analysis of Results

Conclusion:

In this assignment I have analysed 4 different Machine Learning models for risk forecasting, specifically for 5-day forward volatility prediction, and compared it to a standard GARCH model. Based on the success metrics from section 9, I believe that given the current data and standardised feature set, overall, the **Random Forests performed the best as they scored the highest on the classification risk predictions and had the highest regression accuracy** (with the lowest errors). Having done some further research, the AUC scores of our Random Forest models performed in line with other financial risk models with a typical range of 0.6-0.8. To summarise, the results in this project show that **ML models such as the Random Forest model perform better than standard volatility measures at predicting stock volatility.**

Regarding the performance metrics for different stocks (JPM, AAPL, TSLA), my hypothesis was consistent in so far as the fact that TSLA's volatility was the hardest to predict compared to the more stable JPM and AAPL stocks.

Surprises in Model Performances:

Originally, I had expected the more complex models to perform the best, specifically the Neural Networks and Gradient-Boosting Decision Trees. However, this was not the case. Examining the reasons behind this, I believe it stems from the data-intensive demands of these models, which are not well-supported by the limitations of 'useable' daily financial data. In addition, given the variance in their results, I believe averaging over 3 runs may be insufficient, however these models were too computationally intensive to perform more. Furthermore, the Neural Networks and Gradient-Boosting Decision Trees likely performed worse due to the nature of their pattern-seeking behaviour which was negatively affected by our noisy data and existing autocorrelations.

Recommendations:

I would recommend further analysis on the effectiveness of these financial risk forecasting models to be done with more data, less volatile stocks, and overall, more time spent running these models. Furthermore, I would suggest exploring the model's predictability effectiveness at different window lengths other than the 5-days used in this assignment. I would recommend an expanding-window for cross-validation to specifically see the model's performance as the data window increases, to test for overfitting. In addition, I believe more time should be spent on hyperparameter selection to improve the models. Finally, I would suggest supplementing stock data with more market indicators such as the volatility index (VIX), sentiment analysis through social media keyword analysis, and more detailed feature engineering where applicable.