EN3100 - Project Report Examination Form.

Version 2

*When completed, upload this **whole** document onto Learning Central as directed – by Thursday 7th May.*

Guidelines for the completion of the EN3100 Project Report Examination Form.

1. You will have received this Form, for your completion on **Monday of Week 12, 4th May by 4pm.**
2. Answer all questions in the spaces indicated. You can refer to your dissertation, including figures and tables to help present your answers.
3. You must use Calibri size 11 font. **You must not change the margins, line spacing, or delete any text already in this document, such as these notes or the original questions**. Your answers must not exceed the page limit given in each section. Anything exceeding the page limit will not be considered – hence your answers need be precise and concise. The length of each answer can vary, provided the overall page limit specified in each section is not exceeded.
4. You have 72 hours to complete and return the form.
5. The completed form MUST be uploaded onto Learning Central, via Turnitin, by **Thursday 7th May at 4pm.** An Assignment "Submission of Project Report Examination Form" has been set up on Learning Central, under EN3100.

Plagiarism

Students are reminded that they are expected to produce their own original work. Forms are being returned via Turnitin, and disciplinary action may be taken against students with high similarity scores, other than with their own previously submitted reports.

Extenuating Circumstances

Students are reminded that they are working from home and are expected to be available to complete this work. If you feel you have a genuine reason which will prevent you from meeting the deadline, other than the general issue of the ongoing pandemic, you must submit an Extenuating Circumstances

Section A: To be completed by the Examiner and/or Supervisor

Student Name *Cameron Stumpf*	Student Number *C1673094*
Student Email address StumpfCA@cardiff.ac.uk	
Project Title *Machine Learning in Financial Engineering*	
Supervisor Name *Kensuke YOKOI*	Examiner Name *Rossi Setchi*

Section B: General Questions

These questions are the same for every student. This section, including your answers, **must not exceed one page for all four answers**.

Question 1: State the main deliverables of your project.
This project seeks to deliver a practical comparison between traditional decision tree machine learning methods and more advanced neural network machine learning methods in regression forecasting applications; forecasts are based on a standardised academic stock market dataset.

Question 2: Explain the new techniques that you have learnt and applied during your project.
In undertaking this project, I have learned and made use of a wide range of python libraries in and around data science, data visualisation, and machine learning; more specifically, I have made use of the sci-kit learn library, and Facebook's PyTorch machine learning library, both used globally in data science research, in finance, and in business. *Due to the nature of the PyTorch machine learning library, I was required to learn a great deal about neural network architectures, once again gaining a detailed and comprehensive grasp of not only it's operation, but it's potential in business and government applications as the prevalence of big data in the modern world grows.*

Question 3: Describe your greatest achievement during the project which enabled you to complete the objectives.
My greatest achievement in the project was the point at which my neural network model produced behaviour forecasts of the NASDAQ-100 that outperformed comparable experiments with decision tree methods. *This point marked a significant milestone in the project as it made evident the viability of neural network machine learning methods when viewed as opposed to decision tree methods that had been proven superior in previous work. This milestone additionally served as a 'base-camp' from which further neural network experiments were designed and evaluated against one another.*

Question 4: Explain any technical difficulties that limited your progress. This should include any impact the coronavirus pandemic has had, for example, inability to access equipment or computer facilities, but **NOT** any impact already reported via Extenuating Circumstances.
It was my intention to not only produce regression forecasts, but to additionally undertake a series of classification experiments, predicting whether the stock value would have increased or decreased between two points in time (the difference between which being the forecast range) as a 1/0 return argument. *These experiments would have required significantly more computing power than I had available as it relied on significantly more complex neural network architectures than the regression analysis to produce accurate results. I was in communication with Cardiff University's ARCCA supercomputer lab organising the use of one of their servers at the time of the COVID-19 shutdown; I was thus forced to abandon these experiments and instead focus on just the regression analysis.*

Section C: Specific Questions (*Agreed by the Supervisor and Examiner*)

These questions relate to your specific project/report. This section, including your answers, **must not exceed three pages for all six answers** (and original questions).

Question 5: Compare Neural Nets and decision trees in terms of the transparency of their decisions. Is explainability important in financial forecasting?
The results that Neural Network and decision tree methods produced were ultimately of the same form, both methods would return a continuous value that forms an estimate for the future value of the NASDAQ-100 stock index. *Explainability is of significant importance in financial forecasting as ultimately these forecasts serve only to advise and direct a financial analyst, not to replace one. The current limitations of machine learning must be carefully considered in how a machine learning application is designed, as we are still a significant way away from artificially recreating human intuition or "gut-feeling", which continues to play a vast role in stock market trading.* *This was the original intention with comparing regression with classification results for the same forecasting task. Regression results are a far more valuable indicator than classification results for a stock trader as they give a broader insight into a stock's behaviour. Whilst classification problems may return a greater accuracy, they are not anywhere near as descriptive as regression; an automated trading tool could only operate off of a classification machine learning model.* *Comparison between the two methods would have offered interesting insight into the trade-off between explainability and accuracy in results.*

Question 6: Can dimensionality reduction compromise the accuracy of the forecasting? Is there an optimal number of dimensions for a given dataset?
Too much dimensionality reduction can absolutely compromise the accuracy of forecasting, a lack of dimensionality reduction, however, can too similarly compromise a model's accuracy. *The point of equilibrium must be found experimentally. Finding the correct balance point between feature importance and feature variety is a critical task in model optimisation and, when correctly done, has a vast positive impact upon model's performance.* *The optimal number does exist; however, it may vary depending on the requirements of the machine learning model. For the same dataset, a machine learning practitioner may wish to forecast results in a time of relative instability, whilst another may wish to predict major disturbances, the former would perhaps benefit from high dimensional analysis, and lower dimensionality for the latter.*

Question 7: What methods were used to verify the correctness of the code, which was written?
A great deal of debugging and correctness evaluation came down to logic checks. In performing a standardised experiment multiple times under identical conditions, systematic errors could easily be identified and fixed in the code. *Errors in model underfitting or overfitting could easily be solved by modifying the hyperparameters or data conditions of the dataset, however, errors in the coding itself could only be identified by inspection, either with systematic errors in repeated experiments, or in forecast values that were*

"too good to be true".

If forecasts were too far from the true values then errors in neural network forwards or backwards propagation processes may be occurring, each being easily inspected and rectified. Errors in the data pipeline would result in similar errors, but required strong coding logic ability to fault-find and fix, especially as PyTorch's TorchTensor class object behaves in an incredibly unintuitive way to someone from an Engineering background (I may just be speaking for myself).

Results that were unusually accurate could be the result of incorrect data preparation, perhaps the model was being trained on the testing dataset (indicated by higher training loss than validation loss), this error occurred several times and was the cause of great distress.

As a final check, asking colleagues to review my code allowed for flagging of simple mistakes that I would have otherwise been blind to.

Question 8: Is data fitting important in financial forecasting and why?

I believe that data fitting can be important in financial forecasting, however, as of late it is becoming less and less important in favour of higher dimensional machine learning problems. True, neural network machine learning models are capable of curve fitting; however, they are equally capable of fitting to datasets significantly more complex than simple 2 dimensional problems, allowing for more interesting relations to be considered in forecast development.

Question 9: How are major disruptions in the markets such as pandemics taken into considerations in forecasting?

Major disruptions and shifts in the financial market are actually something that my forecasting model was very ineffective at detecting. Models trained specifically to detect large shifts may be designed, however they would require a very detailed, specially curated dataset with a large variety of features that focussed specifically on highly disruptive events (natural disasters, pandemics, market crashes, etc.). Even if trained on such a dataset, it would be difficult to predict the impact of a pandemic such as the coronavirus when a pandemic of this global scale has never occurred in the modern world.

Machine learning models such as mine are trained on relatively uneventful data with the purpose of making forecasts in uneventful times, these models respond poorly to market shifts as evident when my trained models were applied to the test dataset. Midway through my designated test dataset there occurs a significant upward shift of the market, none of the machine learning models trained were able to effectively respond to this shift resulting a large disparity between predicted and true behaviour of the NASDAQ-100 index.

My dataset consisted entirely of "lagging-indictors", meaning that the features upon which the forecasts were based were purely reactive to disruptive events. Had "leading-indicators" such as GDP, unemployment, or sentiment analysis of news outlets been used in addition to the main dataset, the models may have learned to respond more effectively to market disruptions.

There is great scope to develop machine learning models trained specifically to detect these large scale disruptive events; this would be an excellent subject for further work, especially in engineering as detection of market shifts so closely resembles failure detection in engineered components, but this is vastly beyond the scope of my project.

Question 10: How often should models be retrained to maintain their high degree of accuracy?
From my perspective, there is no set time frame against which a model's lifespan can be arbitrarily assessed. In a true-to-life application of machine learning methods in the financial sector, there may be a vast number of models operated parallel to one another, each used in an ensemble of other indicators to form a financial analyst's assessment of a subject. *To ensure accuracy, a model must be retrained after any major market shift occurs, however, to an analyst it may be greatly beneficial to retain an outdated model to gauge the impact of a financial event based on previously anticipated marked behaviour. Forecast error, in and of itself becomes, a metric for which a market can be evaluated.* *The major market shift that occurred mid-way through my test dataset would have been an ideal time to retrain the model, however, if no such market shift occurred then the model's effective lifespan could have been double or triple what it was.* *An important consideration in retraining a model after a market disruption must come from available data. As evident from my neural network experiments when the model is starved of data, you must wait until enough data has been collected under the new market conditions before NN methods can be utilised. Until then, more flexible methods must be used such as the decision tree methods that were evaluated during this project.*