

**Faculty of Technology**

**Department of Computer Science**

**PROJECT PROPOSAL FORM**

|  |
| --- |
| PROM02  MSc Dissertation  Academic Year: 2022/23  Module Leader: Neil Eliot  Email: neil.eliot@sunderland.ac.uk |

|  |  |
| --- | --- |
| Student Name: |  |
| Student ID: |  |
| Programme: | MSc … |
| Mode: | FT/PT |
| Modules Completed | [Identify which modules you have completed] |

|  |  |
| --- | --- |
| Proposed Project Title: | Developing a predictive model for stock market movements using historical financial data and news articles. |
| Supervisor (if known and agreed): |  |

|  |
| --- |
| Aims (To be revised in the Planning Stage) |
| The aim of this project is: To analyze historical financial data and news articles to identify trends and patterns in the stock market;  To make accurate predictions about future stock market movements based on past data and news articles;  To develop a model that can continuously update predictions based on new data and news articles;  To provide insights into the drivers of stock market movements and identify potential risks and opportunities;  To improve investment decision-making by providing a more comprehensive and data-driven view of the stock market. |

|  |
| --- |
| Objectives (To be revised in the Planning Stage) |
| The objectives to achieve this aim are:  Data Collection: Acquire and clean a large dataset of historical financial data and news articles.  Feature Engineering: Extract meaningful features from the data to train the model.  Model Selection: Choose the appropriate machine learning algorithm that can accurately capture the relationship between the data and stock market movements.  Model Training: Train the model using the processed data and evaluate its performance using appropriate metrics.  Model Validation: Validate the model's predictions on an independent dataset to ensure its accuracy and robustness.  Model Deployment: Deploy the model in a real-world environment and integrate it with existing systems and processes.  Model Monitoring: Continuously monitor the model's performance and update it as needed to improve its accuracy over time. |

|  |
| --- |
| Skills from course (Include module codes) |
| The skill I will use from existing modules include:  Python Programming: A solid understanding of Python programming language is required to develop the predictive model.  Data Manipulation and Analysis: Knowledge of Pandas, Numpy and Matplotlib libraries to manipulate, clean and analyze the data.  Machine Learning Algorithms: Knowledge of scikit-learn library to implement various machine learning algorithms.  Natural Language Processing (NLP): Knowledge of NLP techniques and libraries such as NLTK and SpaCy to process and analyze news articles.  Model Validation: Knowledge of model evaluation techniques such as cross-validation and performance metrics such as accuracy, precision, recall, and F1 score.  Model Deployment: Knowledge of deployment techniques such as Flask or Django to deploy the model in a real-world environment.  Database Connectivity: Knowledge of SQL and databases such as SQLite or MySQL to store the data and model output. |

|  |
| --- |
| New skills I will need to develop |
| I will need to develop new skills in  Time Series Analysis: Understanding and implementing time series analysis techniques to analyze financial data and identify trends and patterns.  Sentiment Analysis: Developing the ability to analyze news articles and identify the sentiment behind them to better understand their impact on the stock market.  Ensemble Methods: Understanding and implementing ensemble methods such as Random Forest or Gradient Boosting to improve the model's accuracy.  Data Visualization: Developing the ability to visualize the data and results to gain insights and communicate findings effectively.  Cloud Computing: Developing the ability to deploy and scale the model using cloud computing platforms such as Linode, AWS or Google Cloud. |

|  |
| --- |
| Practical Element |
| The practical element will involve Data Collection, Feature Engineering, Model selection, Model validation, Model Deployment, Model Monitoring  To evaluate the success of the practical outcome there will be need for:  Accuracy: Measure the model's accuracy in predicting stock market movements, such as returns or price changes.  Precision and Recall: Measure the model's ability to correctly identify stock market movements and avoid false predictions.  F1 Score: A weighted average of precision and recall to provide a single metric for model performance.  Confusion Matrix: Analyze the number of true and false positives and negatives to gain insights into the model's performance.  Profit and Loss (P&L): Calculate the financial gain or loss from using the model's predictions to make investment decisions.  Return on Investment (ROI): Measure the financial return from using the model's predictions, compared to a benchmark or alternative investment strategy.  Sharpe Ratio: Measure the risk-adjusted return from using the model's predictions, compared to a benchmark or alternative investment strategy.  Backtesting: Evaluate the model's performance over a historical period of time to assess its ability to predict stock market movements.  Model Stability: Evaluate the model's stability over time, its ability to generalize to new data and its resilience to changing market conditions. |

|  |
| --- |
| Research Ethics |
| I confirm that Ethical approval will be sought from the University of Sunderland Research Ethics Committee. |

|  |
| --- |
| References |
| [Include a list of citations]  Data Collection:  a. Kaggle Datasets: https://www.kaggle.com/datasets  b. Yahoo Finance API: https://finance.yahoo.com/quote/AAPL/history?p=AAPL  c. Alpha Vantage API: https://www.alphavantage.co/  Time Series Analysis:  a. "An Introduction to Time Series Analysis and Forecasting" by Douglas C. Montgomery and Cheryl L. Jennings (book)  Sentiment Analysis:  a. "Sentiment Analysis and Opinion Mining" by Bing Liu (book)  b. "A Survey of Opinion Mining and Sentiment Analysis" by Xiaojun Wan and Ying Liu (paper)  c. Vader Sentiment Analysis library for Python: https://github.com/cjhutto/vaderSentiment  Feature Engineering:  a. "Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists" by Amanda Casari (book)  b. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman (book)  Model Selection and Training:  a. "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems" by Aurélien Géron (book)  b. "Deep Learning" by Ian Goodfellow, Yoshua Bengio and Aaron Courville (book)  c. "An Introduction to Statistical Learning: with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (book)  Model Validation:  a. "Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach" by Kenneth P. Burnham and David R. Anderson (book)  b. "The Bias-Variance Tradeoff" by Scott Fortmann-Roe (blog post)  c. "Overfitting and Underfitting in Machine Learning" by Jason Brownlee (blog post)  Model Deployment:  a. "Productionizing Machine Learning Models" by Tom Fawcett (blog post)  b. "Streaming Algorithms for Data Processing" by Michael J. Franklin and Ali Ghodsi (book)  c. "Deploying Machine Learning Models: A Complete Guide" by Shubham Jain (book)  Model Monitoring:  a. "Model Monitoring for Machine Learning" by Rahul Agarwal and Jake Dancho (blog post)  b. "Continuous Model Monitoring and Feedback in Machine Learning" by Emily Fox and Carlos Guestrin (paper)  c. "Real-Time Analytics with Storm and Cassandra" by Saurabh Singh (book) |