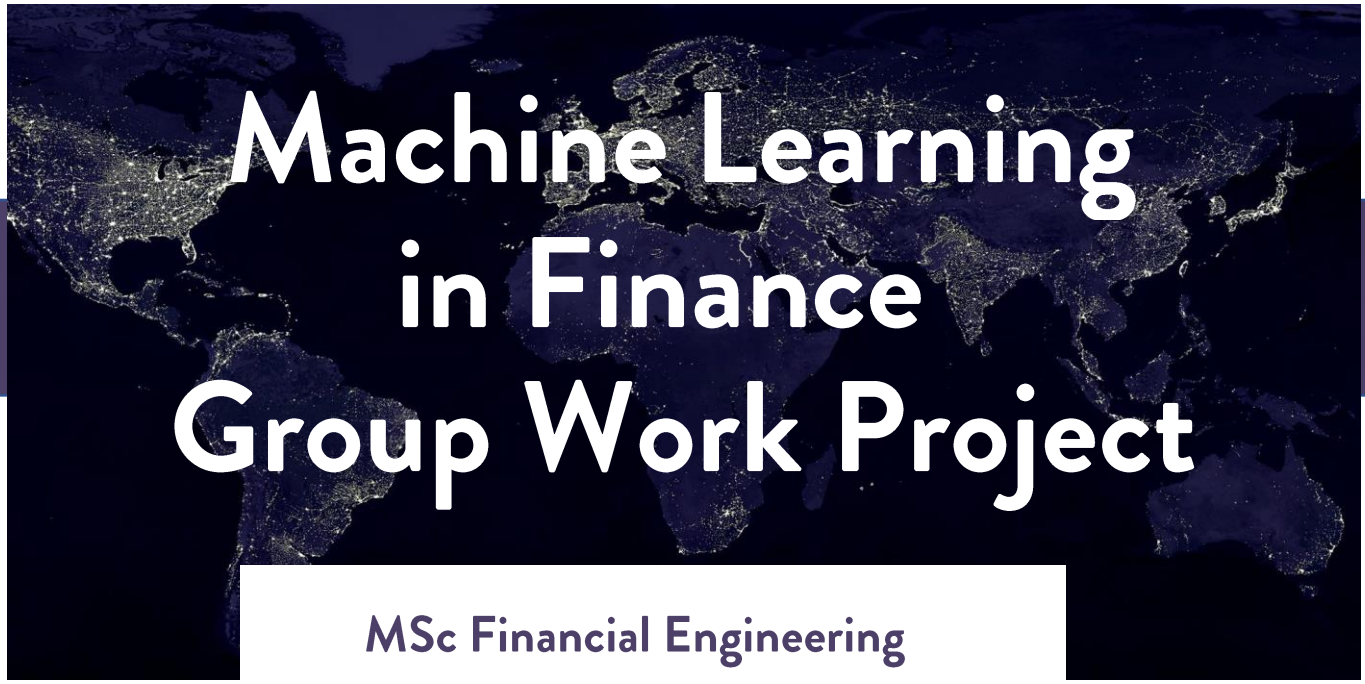
WORLDQUANT UNIVERSITY



Machine Learning in Finance Group Work Project

MSc Financial Engineering

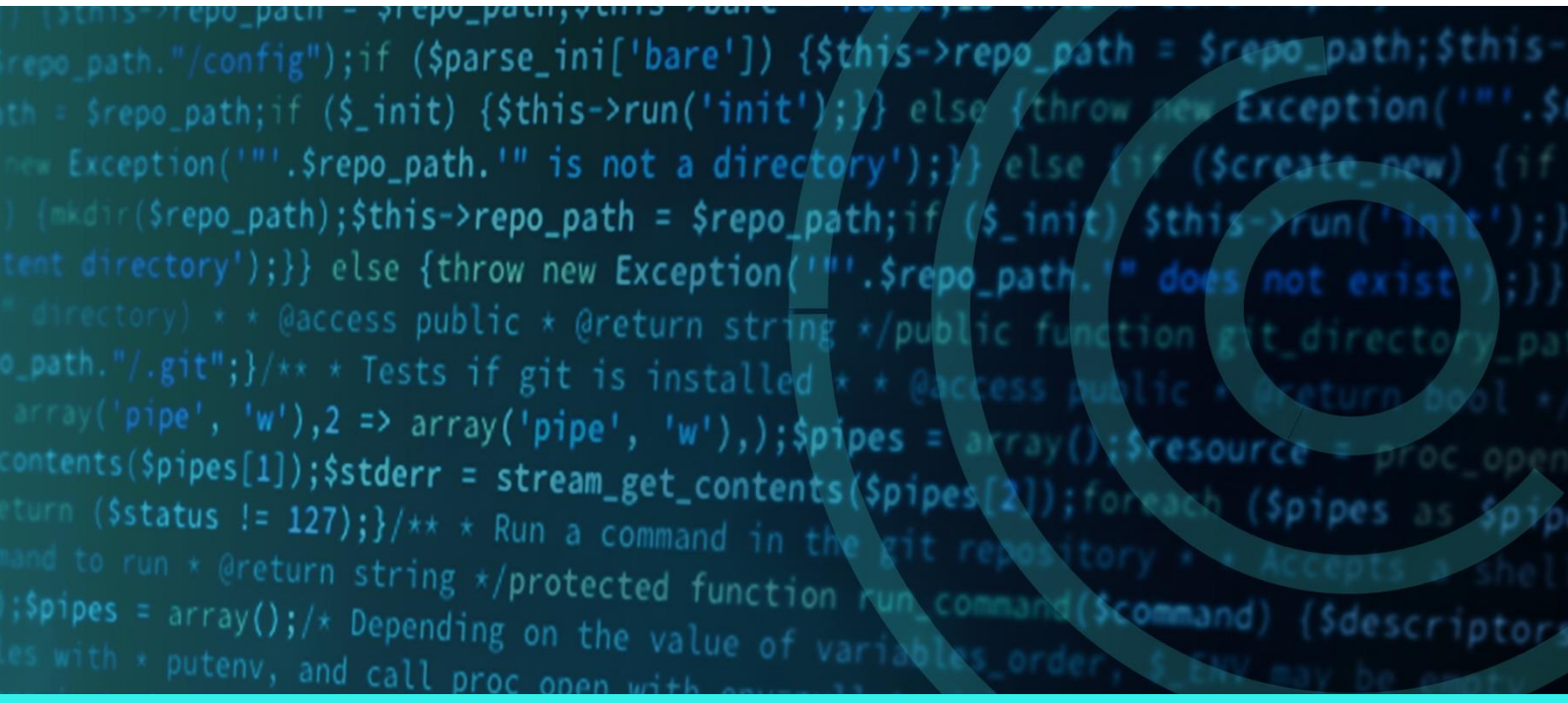


Table of Contents

1. Overview

In every course of the WQU Master's in Financial Engineering, students are required to complete a group work project. Groups are geographically banded and consist of 3-5 students who are able to communicate via a forum. All groups are given the same submission topics for their projects – topics designed to assess not only their understanding of the course content but also their skills of analysis and application.

You are required to make three group work submissions during the Machine Learning in Finance course. Within one week of the first and second submissions, your lecturer will provide detailed feedback, enabling you to improve the substance, clarity, cohesion and structure of your second and final submissions.

Your research should favor authoritative, scholarly sources, and you must reference all sources where relevant. Not only are you required to cite accurate and relevant facts, but you should also present your own clear logic when linking and contextualizing these facts.

All submission dates are published on the learning platform. If you have any questions, remember to post them on the "Ask your Lecturer" forum. The forum's most upvoted questions will be addressed in a live lecture hosted on the platform.

2. Submission Instructions

The ultimate goal of this project is for you and your group members to implement an end-to-end trading strategy using machine learning. The aim of each submission can be summarized as follows:

- **Submission 1** requires you to prepare your data by implementing financial data structures in the form of volume or dollar bars. This leads to data that has better statistical properties than traditional fixed time interval sampling.

- **Submission 2** deals with feature engineering and you will get a chance to implement your own ideas, in an attempt to provide useful features for a model.

- **Submission 3** focuses on using machine learning models in a trading strategy.

The sections below provide all of the detail you'll need to complete each submission. Good luck!

2.1 Submission 1: Data Preparation

Raw tick data is hard to come by, but it is essential that you practice some of the techniques taught in Module 7. All students are required to download a set of raw tick data and create financial data structures. All students are required to download this set of raw tick data, containing 20 days of S&P500 E-mini features, and use it to create financial data structures. It is certainly possible for you to use this dataset for Submission 1; then, for Submissions 2 and 3, end-of-day data from Yahoo and Google Finance. However, it would be ideal to use the same tick data from start to finish. Students should consider cryptocurrencies because their tick data is more readily available and usually has longer timeframes.

The following questions were sourced from Chapter 2 of the textbook *Advances in Financial Machine Learning* (2018) by Dr Marcos Lopéz de Prado.

On a series of raw tick data:

1. Create tick, volume, and dollar bars. (Bar must have open, high, low, and close values.) Students can do this from first principles or clone the following repo for an implementation.

2. Count the number of bars produced by tick, volume, and dollar bars on a weekly basis. Plot the time series of the bar count. What bar type produces the most stable weekly count? And why?

3. Compute the serial correlation of each bar type and report back on which method has the lowest serial correlation.

4. Apply the Jarque-Bera normality test on returns from the three bar types. What method achieves the lowest test statistic?

Write a 500-word report on your findings and research. Make sure to use the Harvard reference style.

2.2 Submission 2: Feature Selection and Engineering

In this section students need to decide which features are helpful in predicting the target variable – for example, serial correlation, momentum, technical analysis indicators (such as RSI), and signals from trend-following strategies (such as the moving average crossover).

1. Select at least four explanatory variables and perform the necessary transformations so that they are useful in the model phase. You are encouraged to use more than four variables. Investigate feature engineering techniques such as PCA and encoding target variables using one-hot encoding.

2. Write a short paragraph about each technique investigated and show an implementation of it in a Jupyter Notebook. Make sure to include references that indicate where the ideas were sourced.

3. At this stage groups should take the opportunity to familiarize themselves with the cross-validation techniques for forecasting financial time series – for example, traditional k-fold cross-validation versus walk forward analysis, and Purged K-Fold CV. Write a short paragraph explaining each technique researched. Research at least three (they don't have to be the 3 mentioned here).

Helpful resources

The following techniques are covered in Dr Lopéz de Prado's book (an implementation of the first and second techniques can be found on Github, and a relevant blog post can be found here):

1. The triple-barrier method (Labeling)
2. Meta-labeling
3. Fractionally Differentiated Features

The following papers provide insights into using technical analysis for features:

1. Kim, K.J. (2003). 'Financial Time Series Forecasting Using Support Vector Machines'. *Neurocomputing*, 55(1-2), pp.307-319.

2. Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). 'Predicting Stock Market Index Using Fusion of Machine Learning Techniques'. *Expert Systems with Applications*, 42(4), pp.2162-2172.

3. Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). 'Predicting Stock And Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques'. *Expert Systems with Applications*, 42(1), pp.259-268.

4. Kara, Y., Boyacioglu, M.A. and Baykan, Ö.K. (2011). 'Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks And Support Vector Machines: The Sample of The Istanbul Stock Exchange'. *Expert systems with Applications*, 38(5), pp.5311-5319.

PCA as a technique was covered in Module 2.

There are also many blogs that provide some insights:

1. Quantopian
2. QuantStart
3. QuantInsti
4. Robot Wealth

2.3 Submission 3: Modeling and Strategy Development

Note: The final submission is likely to take significantly more time than the first two, so remember to prepare well in advance.

Using what you have learnt from Submissions 1 and 2, implement a trading strategy using machine learning. We recommend that students focus on classification – for example: trying to forecast if a stock will move up and down, above some threshold such as the 90-day standard deviation.

Modeling

1. Decide on an algorithm or group of algorithms (for example, ensemble techniques).
2. Fit the model.
3. Show that it works out of sample, and use appropriate cross-validation techniques.
4. Provide the following performance metrics:
    (a) ROC curves,
    (b) Confusion Matrix,
    (c) Precision, Recall, F1-Score, Accuracy, and AUC.
5. Analysis of metrics and report.

Fund factsheet

Create a fund factsheet for your new investment strategy. Have a look at examples of popular funds found online and create a fact sheet with all the bells and whistles. It must at a minimum include (Pyfolio can be used):

1. Maximum Drawdown
2. Annualized Returns
3. Sharpe Ratio
4. Plot the Equity Curve

2.4 Deliverables

Groups must:

1. Create a solutions document that tracks the challenges, solutions, and academic papers. See example provided.
2. Create a solution design/report as a holistic view on your research and findings. See example provided.
3. Create three Jupyter Notebooks, one for each part, saved as both .ipynb and .html.
4. Create a ReadMe document. See example provided. For further assistance, here is:
   (a) A template to make a good ReadMe.md
   (b) A beginner's guide to writing a Kickass ReadMe

2.5 Additional Notes

1. It would be ideal to use the same data for Submissions 1, 2, and 3. Submission 1 is the only section that must be completed with raw tick data. Submissions 2 and 3 may be done using end-of-day data. If you find it challenging to source raw tick data, an easier option is, cryptocurrencies. Helpful links include:
   (a) http://api.bitcoincharts.com/v1/csv/
   (b) http://kibot.com/buy.aspx
   (c) 20 days of raw tick data, E-mini S&P 500 futures. Sourced from Tick Data LLC: https://s3-us-west-2.amazonaws.com/tick-data-s3/downloads/ES_Sample.zip

2. Students have all the tools they need for all three submissions. The use of platforms like Quantopian is permitted but not a requirement. Here are some of the advantages of the advantages of using Quantopian:
   (a) Free financial time series data.
   (b) Vibrant and helpful community.
   (c) Free backtester and research environment, plus additional libraries such as Pyfolio and AlphaLens.

   If you do choose to use Quantopian, be sure to save your notebooks locally as well as submit them.

3. Students are to comply with the PEP8 style guide. Groups will be penalized for:
   (a) Not complying with the style guide.
   (b) A lack of comments.
   (c) Bad variable names.

   In short, other practitioners must be able to easily read and understand your code.

4 Ensure that all of your code runs from top to bottom. Markers will penalize code that doesn't execute.

5 Report all of your findings and research in your solutions document.

6 We at WQU are aware that achieving an out-of-sample accuracy greater than 55% is not a simple task. Please therefore make sure that you display as much of your research process as possible, as well as how you developed your ideas.