Research Proposal

Project Outline:

In order to carry out a research project within a relevant industrial domain, as well to best marry the student's own industrial experience in quantitative finance and statistics with Dr Heinis's excellent expertise in big data analytics, the student would like to choose a PhD research project on **financial time series forecasting using data mining and machine learning technologies**.

This project would entail three development/research stages:

1. Implementation of a software application/platform that perform financial forecasting using statistical learning methods, with a focus on data preprocessing techniques, current feature selection methods and current predictive models;

2. Adapt forecasting models to work with ticking data stream with a particular focus on tackling concept drifts embedded in financial time series data;

3. Design and implementation of an adaptive ensemble model that should possess better predictive power than contemporary methods and an ability to adapt to changing data features ;

Work carried out in Stage 1 would be relatively light and would mainly serve the purpose of creating a software environment in which the researcher can effectively carry out research work in Stage 2 and 3.

Stage One:

The project would firstly look to implement a fully functional software program in R or Python that would take in high quantity of high-dimensional, nonlinear financial time series data, apply various rigorous feature selection techniques to extract the most prevalent predictors from the data and use either regression-based or learning-based or ensemble models with the selected features to carry out prediction on the future price of one or more financial assets (stocks, FX, or even financial derivatives). Alternatively, the focus could be put towards predicting the direction where the prices would move (up or down), turning the problem into a classification problem.

This stage would consist of the following 3 steps:

1. *Data pre-processing and feature engineering.*

   Data pre-processing: Many questions will be explored here, such as what data samples? Formats? Cleansing? Clustering? Normalisation? Smoothing? Enforcing stationarity? And so on so forth.

   Feature engineering: By manually crafting additional features or transforming existing feature data to something more interpretable, we can provide a wider range of features for selection, potentially enhancing the success rate of regression/learning as was seen in [12]. The design of new features and their implications would be the main research topics here.

2. *Feature selection*

Tang and Lin stated that the predictive power of a financial forecasting system should come primarily from the features and not from the models themselves [1] and the abilities to reduce irrelevant and/or insignificant features are key to combating the overfitting problem. Work at this step would be imperative to the success of the forecasting framework and heavily affects the researcher's ability to make meaningful investigation into the predictive models that follow.

The main actions in this step are: Pick a list of well-known feature selection schemes [2] [3] and review their performances in the framework by using the caret package [4] or equivalent. Examine the out-of-sample performances using cross-validation techniques, and more importantly examine how they impact the subsequent predictive models.

3. *Predictive Models*

In this step numerous predictive models which are respectively based on regression schemes such as Ridge or LASSO regressions [5], variants of neural network [6], classification methods such as KNN [7] and Support Vector Machine [8], would be implemented using again the caret package in R [4] or equivalent from other languages such as Python.

As shown in the CRAN's Machine Learning Task View [9] there is currently a substantial list of data mining and machine learning methods worthy of consideration. It's impractical to perform an exhaustive search for the 'best' algorithm for a particular task, but it is certainly possible to arrive at some guidelines around what tends to work better for particular use cases and how their respective performances can be degraded or enhanced by the outcome of the feature selection step and why. Most importantly I would obtain important baseline results here to assist in the development of potentially more advanced models in the next stage's research. These would be the main objectives of this step.

Milestones for stage sign-off:

- Successful implementation of a fully functional software framework that can be configured to use multiple different schemes for feature selection as well as different statistical learning models for time series prediction.

When the student looks at the adaptive feature selection schemes in Stage 2 and ensemble prediction models in Stage 3, he would be using this software framework to evaluate the correctness and effectiveness of schemes and models that he is to develop. That's the main objective of this stage. Work required to achieve this goal would be light compared to that in Stage 2 and 3. This stage would sign off once the program running in the framework exhibits acceptable and measureable forecasting results.

Stage Two:

Majority of existing literature on financial time series forecasting point to using static data and indeed this is the case with many financial practitioners within the industry who carry out non-high frequency trading.

On the other hand, vast amount of streaming data is generated intra-day in the financial markets and if taken advantage of properly they can be used to largely accelerate the forecasting frequency to accommodate the need of intra-day price forecast. However, unlike in traditional data mining, data stream mining algorithms face many challenges and have to satisfy constraints such as bounded memory, single-pass, real-time response, and concept-drift detection [11]. The first goal of research at stage 2 would be about identifying the constraints in data stream that negatively impact the results of feature selection (30+ features would be involved) and the performance of predictive models in the student's forecasting framework. Efforts would be then put towards finding solutions to deal with these issues with an aim to adapt the program for streaming environments.

As part of this exercise, the student will review existing platforms that support analysis on streaming data, such as Apache Kafka, etc. Technical aspects of these platforms as well as how they can help with the predictive program and associated streaming data analysis will be studied.

The data source used would be Bloomberg which is the financial industry's standard data source. In order to create data stream the plan is to feed Bloomberg's static historical data to Apache Kafka which will replay them as data stream into our program. Bloomberg's historical intra-day stock prices have a minimum interval of 1 second. If the need arises to test how well our application can cope in high speed environment the plan is to replay the data faster, up to millisecond.

Since not all financial predictors have granular enough data to support intra-day trading, the group of eligible financial predictors to be used for data stream research would be slightly different from those used in stage 1.

When searching for solutions to adaptive feature selection problems, the student will pay extra attention to approximation methodologies.

Milestones for stage sign-off:

- Successful design and implementation of a novel approach to adaptive feature selection on data stream which tackles concept drift problem;

- Successfully adapt the predictive models designed in earlier stages to streaming data environment;

Stage Three:

Joerg Wichard looked at hybrid ensembles of models for time series forecasting where his ensemble combined forecasts of several different models in a weighted mean and showed such a practice increased the robustness of the model and provided an increased vote to the final forecasts [10]. The objective of Stage 3 of research would be therefore looking at developing smart and improved ways of aggregating the predictions of multiple models set up in stage one with a view to achieve a prediction accuracy that exceeds any individual model.

As part of the research at this stage we would be also looking at adaptive combinations of models and features. Which and how many features are involved in the predictive modelling have profound impacts on the forecast results. As characteristics of the underlying time series data change, the optimal combinations change accordingly too. The design of high performing predictive models can not be separated from the ability of adaptively selecting the most optimal features.

Milestones for stage sign-off:

- Successful design and implementation of a novel ensemble predictive model that is both high performing and able to adapt to changing data features;

References:

[1] Tang, L.C. and Lin, Q.M., 2016, *Stock Selection Based On a Hybrid Quantitative Method*. Open Journel of Statistics, 6, 346-362

[2] Deron Liang, Chih-Fong Tsai, Hsin-Ting Wu, *The Effect of Feature Selection on Financial Distress Prediction*, Knowledge-Based Systems, 73 (2015) 289-297

[3] Serena Ng, 2012, *Variable Selection in Predictive Regressions, Handbook of Economic Forecasting*, Volume 2, Part B, 2013, Pages 752-789

[4] Max Kuhn, *Building Predictive Models in R Using the caret Package*, Journal of Statistical Software, 2008, Volume 28, Issue 5

[5] Devavrat Shah, Kang Zhang, 2014, *Bayesian Regression and Bitcoin*, MIT

[6] Leandro S. Maciel, Rosangela Ballini, 2008, *Design a Neural Network for Time Series Financial Forecasting: Accuracy And Robustness Analysis*, https://www.cse.unr.edu/~harryt/CS773C/Project/895-1697-1-PB.pdf

[7] Amir Navot, Lavi Shpigelman, Naftali Tishby, Eilon Vaadia, 2006, *Nearest Neighbor Based Feature Selection for Regression and its Application to Neural Activity*, neural information processing systems

[8] Saahil Madge, 2015, *Predicting Stock Price Direction using Support Vector Machines*, Independent Work Report Spring 2015

[9] https://cran.r-project.org/web/views/MachineLearning.html

[10] Joerg Wichard, *An Adaptive Forecasting Strategy with Hybrid Ensemble Models*, Proceedings of the WCCI 2016, Vancouver, Canada 2016

[11] Hai-Long Nguyen, Yew-Kwong Woon, Wee-Keong Ng, *A Survey on Data Stream Clustering and Classification*, Springer-Verlag London 2014

[12] Cramer, S., Kampouridis, M., Freitas, A.A., *Feature Engineering for Improving Financial Derivatives-based Rainfall Prediction*, IEEE World Congress on Computational Intelligence, Vancouver, Canada (2016)