# Quantitative Trading Models - Will Jane Street Hire Us?

Team A || <A0230632U Chua Peng Shaun, A0228402N Tan Chuan Xin, A0230521Y Teo Chen Ning>

### **Motivation**

We have selected the [**Jane Street Market Prediction**](https://www.kaggle.com/c/jane-street-market-prediction/)competition dataset from Kaggle as our project.

Quantitative trading is a highly lucrative industry - when companies can afford to pay interns $20k USD per month, you bet we’re interested in it! These days, mathematical and statistical models, machine learning and artificial intelligence are employed to make rapid trading decisions that seek to beat the stock market and provide investment returns in the billions of dollars.

With the influx of retail investors and market manipulation, the traditional way of analyzing the market no longer works as well - the market is way too volatile and fast moving for human analysts to maintain stable profits reliably. Instead, predictive machine learning models based off of data are more important nowadays to ensure consistent gains even in a volatile and unpredictable environment.

### **Statement of the Problem/Task**

Using machine learning techniques, can we create a quantitative trading model that maximises profits by correctly selecting buy/pass actions? In the scope of the Kaggle dataset, can our model achieve the perfect utility (profitability) score?

Building up to our goal, we aim to answer the following:

* Can we reduce the feature space of our dataset through data analysis and dimensionality reduction to discover a subset of features that represent the most salient aspects of the data?
* Will deep learning techniques outperform traditional machine learning in creating a quantitative trading model?
* How close can our quantitative trading model get to the maximum utility (profitability) score?

### **General Approach**

Step 1: Data set preparation

* Data analysis and exploration to learn about the dataset
* Dimensionality reduction (if applicable) to select salient features
* Data cleaning to remove unnecessary data
* Data preprocessing to conform and standardize data
* Perform our own train-test split (Kaggle submissions closing)

Step 2: Create the evaluation metric

* With the train set, calculate the maximum utility score achievable on the test set through the competition [evaluation metric](https://www.kaggle.com/c/jane-street-market-prediction/overview/evaluation)
* Set up pipeline to easily evaluate our future predictions

Step 3: Employing traditional machine learning

* Examples include linear regression, K Nearest Neighbours, decision trees, random forests, Support Vector Machines

Step 4: Employing deep learning

* Training a neural network/deep residual learning and comparing the utility scores against those of the traditional machine learning models to see which fares better

Step 5: Evaluation of all models, conclusion

* Compare our highest scoring model (traditional or deep learning) against the maximum utility score

### **Evaluation**

A test set will be generated through our data preprocessing step. The maximum utility score (profitability) achievable on the test set will be calculated, representing a ceiling on our model’s performance. The trained models will then be evaluated against the maximum utility score to determine the effectiveness of the model.

* A+ grade: Hired by Jane Street
* A grade: >60% of maximum utility (rather profitable model)
* C grade: >30% of maximum utility (mildly profitable model)

The percentages of maximum utility are simply an estimate, we do not know the ease by which we can achieve them. They may require adjustments based on the result of our initial experimentation with the simplest machine learning model.

### **Resources**

We will be relying on the **Jane Street Market Prediction** dataset which comprises a training set consisting of historical data and returns as well as metadata pertaining to anonymized features. There are numerous notebook submissions on the Kaggle competition site that we can utilise as a reference.

The dataset is approximately 5.8GB. GPU support is likely necessary, and online resources like Google Colab can be utilized to support our computation. Alternatively, NUS’ Tembusu computing cluster is a viable option for computing power.

Other resources that we will rely on would be free online resources on platforms like Medium to learn, and software like Tensorflow (Keras) to build and run our neural network model. Our strategy is to check out various forums (e.g. Reddit, StackOverflow, Medium) to find out how others have approached a similar problem of quantitative trading and what models they used.

### **Schedule / Role Assignment**

Members’ experience with Machine Learning:

* Chen Ning, Shaun - None
* Chuan Xin - Moderate exposure from past internship

Time for research and learning is needed, mostly consolidated within Recess Week to ensure all members are up to speed.

|  |  |  |
| --- | --- | --- |
| Week | Deliverables | Assignee |
| 5 | Project proposal  Data set analysis, exploration | Team  Team |
| 6 | Dimensionality reduction  Data cleaning, evaluation metric  Data preprocessing, train-test split | Chuan Xin  Chen Ning  Shaun |
| R | Curate list of topics to read up on  Research traditional ML, deep learning  Start traditional ML learnt to date | Chuan Xin  Team  Team |
| 7-8 | Traditional machine learning  Begin work on deep learning | Team  Chuan Xin |
| 9 | Interim Consultation | Team |
| 10-12 | Deep learning  Each team member to be assigned equal number of models to build the pipeline for and run code | Team |
| 13-14 | Code cleaning and refactoring  Result evaluation  Report preparation | Chen Ning  Shaun  Chuan Xin |
| 15 | Final Project Presentation | Team |
| 16 | Final Project Report | Team |

### 

**[End of proposal]**