# TABLE OF CONTENTS

1. **INTRODUCTION**  
   1.1 Background and Motivation  

2. **LITERATURE REVIEW**  

3. **PROJECT OBJECTIVES**  

4. **SCOPE**  

5. **METHODOLOGY**  
   5.1 Introduction  
   5.2 Data Preparation  
      5.2.1 Data Collection  
      5.2.2 Data Preprocessing  
      5.2.3 Exploratory Data Analysis  
      5.2.4 Feature Extraction and Engineering  
   5.3 Development of the Momentum Model  
   5.4 Machine Learning Model  
   5.5 Combining Models and Evaluation  

6. **PROJECT SCHEDULE AND MILESTONES**  

7. **REFERENCES**
1. INTRODUCTION

1.1. Background and Motivation

The financial markets have experienced significant growth over the last decade. Most importantly, the markets have opened up and are easily accessible by retail and institutional investors alike. As a result, financial forecasting is of particular interest to almost every market participant. In the present day, both institutional and retail market players have deployed large amounts of capital into developing their financial forecasting trading systems. This is done in hopes to beat the market and generate “alpha,” which refers to the excess returns earned over a benchmark return.

These forecasting systems constitute a small but major part of many quantitative trading strategies. A quantitative trading strategy, irrespective of its complexity, is responsible for generating trade signals in order to buy and sell securities. More often than not, it is extremely hard to generate such order signals with high accuracy. One main reason is that markets have become much more accessible to retail investors, and as a result majority of the investors bet on the same markets using similar strategies. In financial jargon, such trade is known as a crowded trade. These crowded trades make the market efficient by leaving little to no room for pricing inefficiencies.

As a result, quantitative traders in the search of superior returns develop forecasting models and strategies where they have an “edge” -- an approach, technique, or observation that creates a cash advantage over other market players.

Over the last two years, I have successfully developed and deployed systematic trading strategies across multiple derivative markets such as options and futures. My previous internship experiences have solidified my understanding of market flows, which I believe will provide me with an edge when I develop the forecasting model. This project is an invaluable opportunity for me to combine my knowledge in financial markets along with technical skills and build a robust financial forecasting system. I intend to develop a hybrid model which combines a machine learning approach along with a price momentum filter in order to significantly improve the accuracy of the forecast.
2. LITERATURE REVIEW

Financial forecasting has become an increasingly popular topic both in academia and the financial industry. Multiple attempts have been made to develop a highly accurate forecasting model. However, making such forecasts is extremely difficult given the noisy nature of the markets.

In practice, linear models or non-linear models are used to forecast time series data. However, these models are not self-sufficient because they assume a linear correlation structure among the different values of the time series [1]. As a result, these models are unable to capture any non-linear patterns.

New techniques have been introduced to perform forecasting on a non-linear time series, mainly artificial neural networks (ANNs). A popular forecasting technique involves combining the use of ANNs alongside statistical linear models such as autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroskedasticity (GARCH) volatility. However, most numerical time series data is neither purely non-linear nor purely linear [2].

Additionally, the majority of the traditional approaches involving machine learning models focus solely on optimizing, and in many cases over-optimizing the algorithms. As explained by Wang et al. [3], using a single model for forecasting purposes is not very effective. However, combining multiple individual models to develop a hybrid model significantly improves the prediction accuracy of forecasting systems. Similarly, Babu and Reddy [4] argue that the nature of the volatility of the time series data can first be understood by developing a filter, followed by the application of machine learning techniques to generate the final forecast of higher accuracy. The prediction accuracies have been compared using relevant metrics such as the mean squared error (MSE) and the mean absolute error (MAE), and show a significant improvement.

It is also worth noting that most of the hybrid models in the market rely entirely on quantitative data, without any regard for external factors such as market and investor sentiment. Non-quantitative data such as textual data often contains useful information that can lead of forecasts of higher accuracy. For instance, a change in tone of market experts or the government can at times signal or result in a shift in the investor sentiment [5].
3. PROJECT OBJECTIVES

The project aims to develop a forecasting model which identifies the market direction and generates systematic buy and sell signals on the index futures market. This means that the model’s output will be binary instead of a time series, therefore resulting in significantly higher accuracy. There are four main objectives of this project:

1. Quantify the trending nature of the market and identify the relationship between the price movement of index futures and the market volatility.
2. Establish a positive correlation relationship between market sentiment and price momentum, where one often leads to the other.
3. Model a price momentum filter with accuracy greater than 51% in the index futures market using systematic trend-following trading strategies and associated risk management techniques.
4. Show that a hybrid forecast model utilizing a momentum filter significantly improves the forecast accuracy in comparison to a standalone machine learning approach.

Last but not least, I hope to improve my understanding of systematic trading and financial forecasting upon the completion of this project. Not only will this project provide me with an opportunity to work on something I’m passionate about, but it will also allow me to gain practical experience in new machine learning techniques which I plan to work on throughout this project.

4. SCOPE

In order to keep the project on track and manageable, it will only deal with market indices based in the United States of America (U.S.A.). The two major equity market index futures -- E-mini S&P 500 and E-mini Nasdaq 100 will form the foundation of this project, for which the forecasts will be made.

Additionally, an element of market volatility will be added to the model. This will be done using price data of futures of a popular volatility index known as the VIX. The VIX measures the market’s expectation of volatility based on S&P 500 index options. The algorithm implementation will first involve the development of a price momentum model, secondly, a machine learning model using ensemble methods, and finally a sentiment analyzer.

Instead of utilizing a single model, the project will combine multiple models in tandem to produce a forecast of the highest accuracy.
5. METHODOLOGY

5.1. Introduction

The first step towards building an accurate forecasting model involves the collection of high-quality data and preparing it for further use. Therefore, data preparation will be one of the most important aspects of this project and will involve multiple steps, including but not limited to -- data collection, data preprocessing, exploratory data analysis, and feature engineering. The next step will involve the implementation of a momentum model which will identify price trends in the market. Subsequently, more sophisticated machine learning models will be implemented on the price data alongside sentiment analysis on alternative data. These models combined together will provide us with a forecast of high accuracy. Throughout the project, the Python programming language will be used for all the tasks.

5.2. Data Preparation

For the momentum model, machine learning model, and sentiment analysis, we need to collect high-quality data. Given that we are working with financial data, it becomes even more important to source data from reliable sources as inconsistencies can lead to inaccurate forecasts. The data preparation process for this project will consist of four steps.

5.2.1. Data Collection

Data collection is an important aspect of every forecasting process. As mentioned previously, the project’s sample space has been narrowed down to two major U.S. equity market indices - E-mini S&P 500, E-mini Nasdaq 100 (futures contracts), and a volatility measure using the VIX futures. The U.S. markets have been chosen because they are the most developed financial markets with high trading volumes. This makes the market prices efficient and access to data is easier.

The model will rely on two types of data -- numerical and alternative. Numerical data will comprise a historical time series of daily prices and volumes for the market indices and VIX futures. On the other hand, alternative data will mostly comprise relevant text scraped from social media websites such as Twitter.
5.2.2. Data Preprocessing

Data preprocessing involves transforming raw data into an understandable format for the model. This will involve making sure that the data is accurate, identifying any missing values, and converting it to an appropriate format (e.g., the price should be a floating-point number instead of a string). Here, I will use an external library such as Pandas in Python, to format the data and store it in a tabular format using DataFrame. For alternative data, this will involve formatting and storing in a way that can later be utilized by the sentiment analysis model.

5.2.3. Exploratory Data Analysis

The numerical time series data will now be subjected to preliminary analysis. In particular, I intend to verify the trending nature of the markets, which will require the calculation of a metric known as the Hurst Exponent. Additionally, a test of stationarity will be performed using the Augmented Dickey-Fuller method, to ensure that the statistical properties of the time series remain constant. If the time series is not stationary, it will be transformed into one. Lastly, correlations between the time series will be measured and statistics such as the average and median price movements will be captured.

5.2.4. Feature Extraction and Engineering

Once the data has been collected, preprocessed, and analyzed, the feature engineering process will play a key role in determining the success of the machine learning model. Feature engineering will be deployed not only to identify and optimize key features but also to derive new features from existing ones. This will allow us to experiment with a wide range of features but narrow down the ones which produce the best results without the risk of overfitting. For best results, I intend to derive features

5.3. Development of the Momentum Model

Once the data has been prepared, I will begin the development of a momentum model. This model will utilize a combination of relevant market indicators, but mainly the volume-weighted moving averages. The model will utilize the processed numerical price and volume data in a time series format to perform two main tasks:

1. Validate and quantify the strength of price momentum.
2. Identify the preliminary direction of the price trend by assigning it binary values (for example, 1 for an uptrend and 0 for a downtrend).

The momentum model’s output is critical to the analysis because it will help in cross-validating and combining the results once the machine learning models are applied in the later stages of the project. This model will be implemented from scratch using Python and external libraries such as NumPy, which will enable us to perform mathematical functions on the data.

5.4. Machine Learning Model

The next step in the project will involve the development of a machine learning model which will combine two functionalities -- price forecasting, and sentiment analysis. For the purpose of price forecasting, an iterative process will be deployed to determine the most suitable numerical parameters for the machine learning model. Alongside, sentiment scores will be generated based on non-numerical/textual data, which mainly comprises text feeds from social media such as Twitter.

The model’s primary purpose will be to solve a classification problem based on numerical data and sentiment. I aim to deploy simple regression models using scikit-learn to identify the direction of the price trend, and then progressively use PyTorch if the need arises. In progressive stages, other ML models will also be implemented and accuracy across the methods will be compared. Such experiments will allow us to finalize the model which produces the best results without overfitting.

To fetch the data, I plan to use Tweepy, which is an easy-to-use Python library for accessing the Twitter API. In tandem, the TextBlob library will be used to perform basic sentiment analysis of textual data.

5.5. Combining Models and Evaluation

Lastly, the results from both the momentum model and the machine learning model will be combined in order to produce a final forecast for the direction of the price. The momentum model will act as a filter for the forecast of the machine learning model, and only values where the forecast is similar will be returned. Therefore, combining these models will significantly reduce the errors and improve the forecast accuracy compared to a standalone model.

The forecast of the machine learning model will be evaluated using common metrics such as the Root Mean Squared Error (RMSE). However, it is not final yet and I plan to use an iterative approach along the way to identify the most suitable metrics for this project.
6. PROJECT SCHEDULE AND MILESTONES

I plan to follow an agile development process -- iterative and effective, thus ensuring that all deliverables are finished on time. Below is a brief timeline for the project.

<table>
<thead>
<tr>
<th>Date</th>
<th>Goals and Internal Deadlines</th>
</tr>
</thead>
</table>
| September 14, 2021 | ● Finalize topic and methodology  
● Conduct initial research and literature review |
| September 30, 2021 | ● **Deliverable 1**: Set up a project website  
● **Deliverable 1**: Finish Detailed Project Plan |
| October 14, 2021  | ● Developing a web scraper for Twitter (textual data)  
● Collecting and consolidating historical price data for index futures (S&P 500, Nasdaq 100, VIX) |
| October 28, 2021  | ● Data Preprocessing                                                                       |
| November 14, 2021 | ● Feature Extraction  
● Exploratory Data Analysis - advanced statistical analysis and trend capturing |
| November 28, 2021 | ● Building the price momentum model  
● Initial testing of the momentum model using price data, followed by iterative improvements |
| December 14, 2021 | ● Development of Machine Learning models -- both for price data and sentiment analysis       |
| December 28, 2021 | ● Combining and deploying both models in tandem  
● **Deliverable 2**: Finalise preparation for the first presentation |
| January 14, 2022  | ● **Deliverable 2**: Finalise Detailed Intermediate Report                                 |
| January 28, 2022  | ● **Testing Phase #1** - Rigorous performance analysis of the final model                   |
| February 14, 2022 | ● **Testing Phase #2** - Final software and code review, followed by further improvements   |
| February 28, 2022 | ● **Testing Phase #3** - Deployment of the model on the live market for final testing        |
March 14, 2022
- Preparation of the Project Poster
- Begin preparation of the final presentation

March 28, 2022
- Final review of the report, presentation, and the entire project

April 14, 2022
- Deliverable 3: Finish the Final Report
- Deliverable 3: Prepare the Final Presentation

7. REFERENCES


