1. INTRODUCTION

In today's dynamic financial landscape, the ability to discern and capitalize on robust investment opportunities is essential for achieving success. The decisions surrounding the purchase and sale of stocks and derivatives hinge upon precise evaluations of these financial instruments. Historically, these determinations were made through manual analysis by seasoned financial professionals. However, the advent of technology has ushered in a new era of opportunities for retail investors. Within this evolution, the field of data science has been increasingly drawn to the exciting realm of quantitative trading, where investment decisions are executed programmatically, guided by predictions from rigorously trained models.

While a myriad of quantitative trading efforts currently exist for analyzing financial markets and crafting investment strategies, the fundamental challenge resides in the quest for an accurate and reliable model. This model, which stands as the bedrock of success in quantitative trading, must exhibit a high degree of accuracy. The acquisition of both historical and real-time financial data is vital for these strategies, yet this resource often remains frustratingly inaccessible, particularly for individual investors.

2. PROBLEM DESCRIPTION

At the heart of this project lies the captivating world of quantitative trading, where investment strategies are constructed and executed based on predictive models. This sophisticated methodology relies heavily on the availability of both historical and real-time data – a resource that frequently eludes individual investors. The fundamental problem to be addressed in this project is the creation of a solution that can harness predictive power, providing accurate and dependable forecasts for stock market prices. Such insights are invaluable in guiding non-professional investors toward well-informed financial decisions.

For these aspiring investors, the stakes are high. They cannot afford the luxury of missteps or underperforming investments. As such, the paramount metric for optimization in this context is accuracy. The overarching objective of this project extends beyond mere quantitative trading. It seeks to ignite a newfound enthusiasm for this field among non-professional investors, thereby democratizing their access to the realm of programmatic investment strategies and portfolio analysis. In doing so, it fosters an environment where precision and informed decision-making are not exclusive to financial professionals but are attainable by all with a passion for smart investing.

3. EXISTING SOLUTION

While there have been attempts to tackle the challenge of quantitative trading, one notable prior solution utilized the LightGBM [1] algorithm in conjunction with 5 K-Fold Cross validation [2]. This approach incorporated the measurement of feature importance and yielded commendable results, achieving a precision score of 90 percent and an impressive F1-score of 94. This approach marked a significant step in the right direction in addressing the problem.

4. PROPOSED METHOD

Our proposed approach aims to build upon the foundation laid by the existing solution, taking into account the Data Science and Machine Learning pipeline we have been introduced to in this cohort. To address the problem of accurately predicting stock market prices for the benefit of non-professional investors, we plan to introduce a new method that leverages the Random Forest algorithm [3]. Our methodology will encompass the following key stages:

1. Data Sourcing: We will gather comprehensive and up-to-date financial data of the stock market from a Kaggle repo [link][4] – The whole Team

2. Data Cleaning and Preparation: Robust data cleaning and preprocessing will be carried out to ensure the quality and accuracy of the dataset. This stage is critical to the success of the model, as it enhances the model's capacity to make precise predictions- Gamaliel, Ayorinde

3. Machine Learning Model: We will employ the Random Forest algorithm to construct a predictive model. Random Forest is known for its capacity to handle complex data, mitigate overfitting, and offer superior performance. - Chito, Gamaliel

4.Feature Reduction with PCA: To further enhance the model's efficiency and address potential dimensionality challenges, we will employ Principal Component Analysis (PCA) [5] to reduce the number of features while preserving the most valuable information.

5. Hyperparameter Tuning with Grid Search CV [6]: To optimize the model's performance, we will utilize Grid Search Cross Validation to systematically explore various hyperparameters. This approach will enable us to fine-tune the model for superior accuracy and robustness.

6. Model Deployment: In the event that it is applicable, we will explore the possibilities of deploying our model to make real-time predictions or integrate it into an accessible platform for non-professional investors. Model deployment ensures that the benefits of our solution can be readily accessed and utilized. We will leverage the use of Streamlit for deployment. – Ayorinde, Chito

5. PROPOSED TIMELINE

1. Data Sourcing- 1 day
2. Data Cleaning and Prep- 3 days
3. Modelling- 4 days
4. Model Deployment- 3 days

TEAM MEMBERS

1. Fola Animashaun- Mentor
2. Gamaliel Okudo – Team Representative
3. Ayorinde Alase
4. Chito

REFERENCES

1. LightGBM Documentation -Welcome to LightGBM's documentation! — LightGBM 4.1.0.99 documentation
2. K-Fold Cross Validation - sklearn.model_selection.KFold — scikit-learn 1.3.1 documentation
3. Random Forest Algorithm - Random Forest Algorithms - Comprehensive Guide With Examples (analyticsvidhya.com)
4. Stock Price Data , Kaggle - Stock market prediction (kaggle.com)
5. Principal Component Analysis -Principal Component Analysis – How PCA algorithms works, the concept, math and implementation | ML+ (machinelearningplus.com)
6. Grid Search CV-Principal Component Analysis – How PCA algorithms works, the concept, math and implementation | ML+ (machinelearningplus.com)