Joe Opitz & Quinn Springer

Professor Sitaula

STAT 491 – Project Proposal

4/4/2025

**Project Proposal**

**Project Identification**

Our goal for this project is to predict trends in the stock market. The stock market has long been known as being very unpredictable, but we believe that using the algorithms and techniques learned in class, we could come up with an efficient algorithm for it. This is a prevalent data science problem to try and tackle, and with its many applications, it could be beneficial for providing us with our own financial advice to decide what stocks to invest in.

**Data Collection and Preparation**

There is good, clean data easily accessible from Yahoo Finance through web scraping and API’s that we can use to analyze a specific stock. Most of the data does not contain null values, so we would only need to manipulate the data to predict future prices more easily. Before anything, though, we will separate our dataset into training and test sets. One common approach for this is to create lag-based features. A lag-based feature is where you grab the closing prices from the previous few days and store them as their own variable in the dataset. And using those features, you can then make predictions for the current day. There are other features that we could calculate, such as returns (used for outlier detection), log transformations, percentage-based returns, along with others. Although unsure what exactly we will use, we have a good baseline understanding.

**Exploratory Data Analysis**

We will be following many of the EDA processes discussed in class. First, we will examine the overall trend of the training data set closing values. We will then investigate individual features to identify a couple of things. One main thing to worry about is outliers. As the stock market can be very volatile, it will be essential to handle the outliers in a way that the models can perform as efficiently as possible. In addition, we may explore graphical and statistical methods to better understand the relationships between our predictors and the target variable. This could include visualizations like scatterplots, histograms, and boxplots to identify trends, distributions, and potential data quality issues. We might also experiment with deriving new features, such as moving averages or percentage changes, or applying binning techniques to uncover patterns that aren’t obvious in the raw data. Other tools like correlation matrices or contingency tables could help inform our feature selection and modeling choices. While we may not implement all of these approaches, they represent the types of techniques we plan to consider as time permits.

**Modeling & Analysis**

We will use a regression-based approach to predict a specific price. We will run tests using Multiple Regression models, decision trees, random forests, neural networks, and any others that we see as a good fit for this specific problem. Unsure if this will be covered in class, but would like to implement ensemble methods eventually. We will be using R2 and RMSE as our performance metrics to analyze model performance. With the new predictions that the models make, we will create graphs to better analyze the performance in relation to the actual values. We can utilize these graphs to display important aspects such as feature importance, permutation importance, and/or comparison graphs between the models R2 and RMSE results.

**Project Deliverables & Quality Assurance**

We will present our findings through both a structured written report and a 20-minute presentation, clearly summarizing our analysis, key results, and interpretations. To ensure clarity and reproducibility, we will maintain consistent formatting in our code and documentation throughout the project. All relevant code will be included in the final report, and we will structure our workflow in a way that makes it easy to follow and replicate. Additionally, we will reflect on the limitations of our analysis and discuss potential improvements or directions for future work.