Project Work 2

Elias Eskelinen, Jarkko Komulainen, Matti Aalto, Vili Niemelä

November 11, 2025

1 Introduction

The aim of this project is to develop a model to predict the forward returns from the US stock market. To do this, we utilize a dataset (Hull et al. (2025)) containing US market data and daily returns data from buying the S&P 500 and selling it the following day.

2 Exploratory data analysis

The dataset consists of 9021 samples with 98 features. The features consists of a date id, which acts as an identifier for a single trading day, market dynamics and technical features, macro economic features, interest rate features, price/valuation features, volatility features, sentiment features, momentum features and dummy/binary features. There are also features for forward returns, i.e. the returns from buying the S&P 500 and selling it the following day, risk free rate i.e. the federal funds rate, and market forward excess returns i.e. forward returns relative to expectations.

The data contains plenty of missing data. Figure 1 shows that the first thousand datapoints have 85 missing features and only the last two thousand contain all the features. This is likely due to the data originating from a long period of time and data availability. Missing data must be taken into account in the model training phase and should be dealt with in data preprocessing.

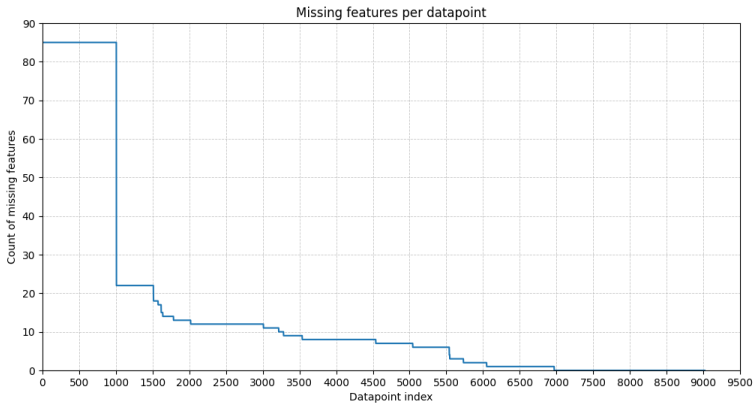

Figure 1: Missing features per datapoint.

Figure (2) shows the forward returns over time. From visual inspection, there does not seem to be any strong trends in the data, or even clear seasonality. However, the data shows that there are clear periods of higher volatility, where the variation in daily returns is high.

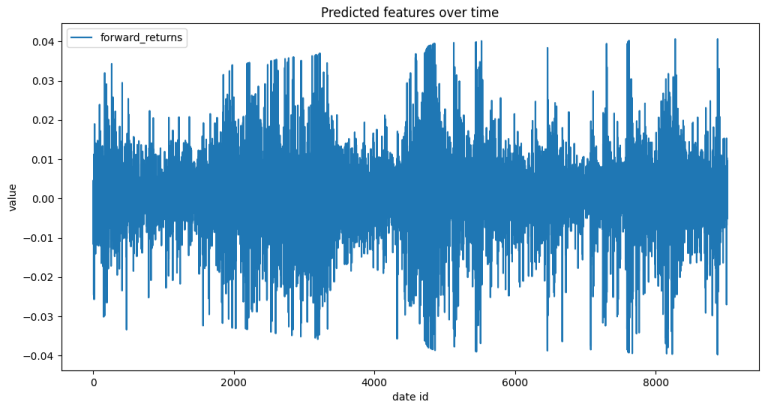

Figure 2: Forward returns time series.

The findings from figure (2) are supported by the distribution of forward returns observations, shown in the histogram (3). The values are distributed symmetrically around zero, with most of the density in the center. The high peak on the mean and symmetric spread around zero tells us, that most of the time the return is close to 0; high or low values are rarer, and getting positive returns is just as likely as getting negative returns.

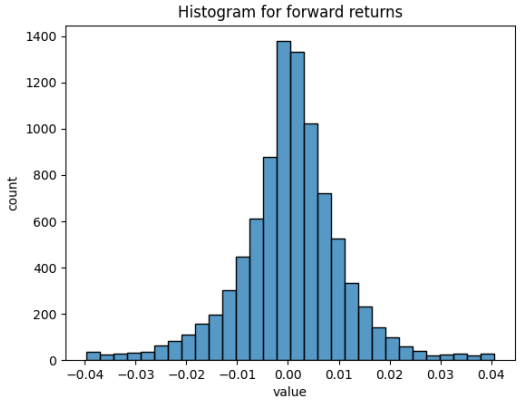

Figure 3: Forward returns histogram.

Figure (4) presents a correlation matrix of selected features in the data. Because of the high number of variables in the data, in order to keep the figure readable, the correlation matrix only shows the correlations for features which have strong ($|\text{corr}(x_i, x_j)| > 0.8,\ i \neq j$) positive or negative correlations with some other feature. The figure shows there are some strong correlations between the features, often between features of the same type. Notably, however, there are no strong correlations between the predictor features and forward returns.

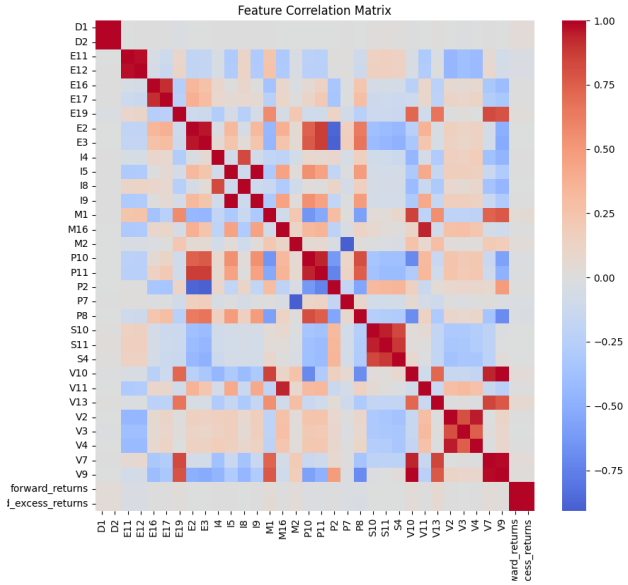

Figure 4: Correlation matrix for features with strong correlations.

2.1 Decomposition

Based on the Figure 2, the time series to predict does not seem to contain any clear long-term trend or seasonal pattern. In Figure 5 a zoomed trends of the time series can be seen. Based on the Figure, the time series does not seem to have either clear short-term trend or seasonal patterns. The time series seems to resemble a heteroscedastic Gaussian process.

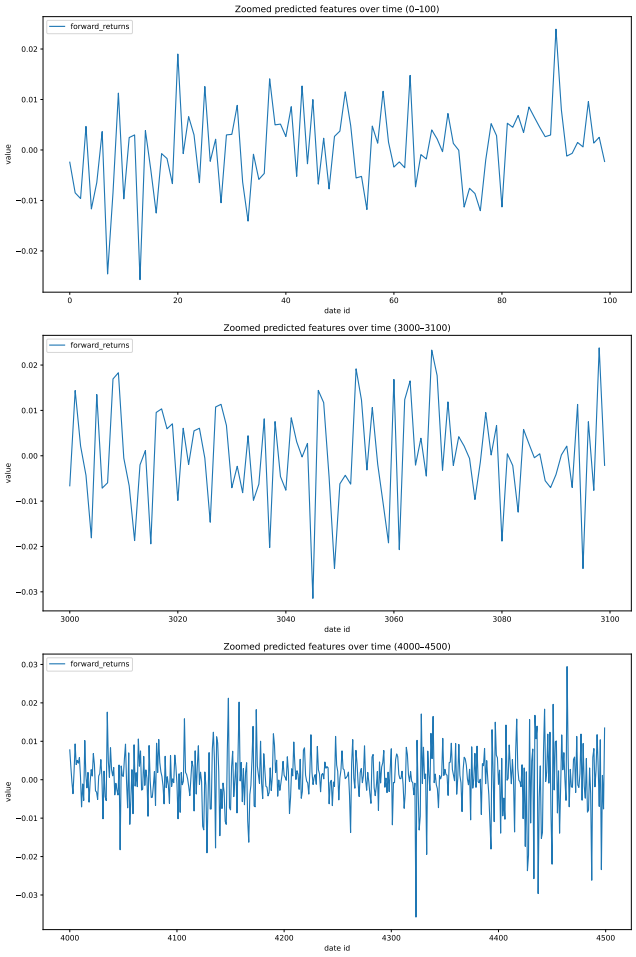

Figure 5: Zoomed forward returns.

We performed a regression analysis to the forward returns time series to confirm whether any linear trend can be identified. The estimated bias and trend coefficient terms are $[1.98082307 \times 10^{-8}, 3.82116136 \times 10^{-4}]$ which indicates that no long-term linear trend exists. As a conclusion, the forward return time series does not seem to have long-term trend or seasonal components to decompose.

2.2 Autocorrelation

The lag plot of forward returns looks like a round cloud and the ACF after lag 0 stays near zero, so yesterday's return doesn't help predict today's. When we fit a simple model that uses yesterday's return to predict today's, the residuals look like noise. Their ACF is flat, so the mean part seems fine. But when we look at the absolute residuals, the ACF shows clear positive correlation that fades slowly. This could be interpreted to mean the volatility clusters. So larger

moves and quieter periods tend to come in streaks. So the findings suggest that returns or their residuals themselves don't show autocorrelation, but the volatility does.

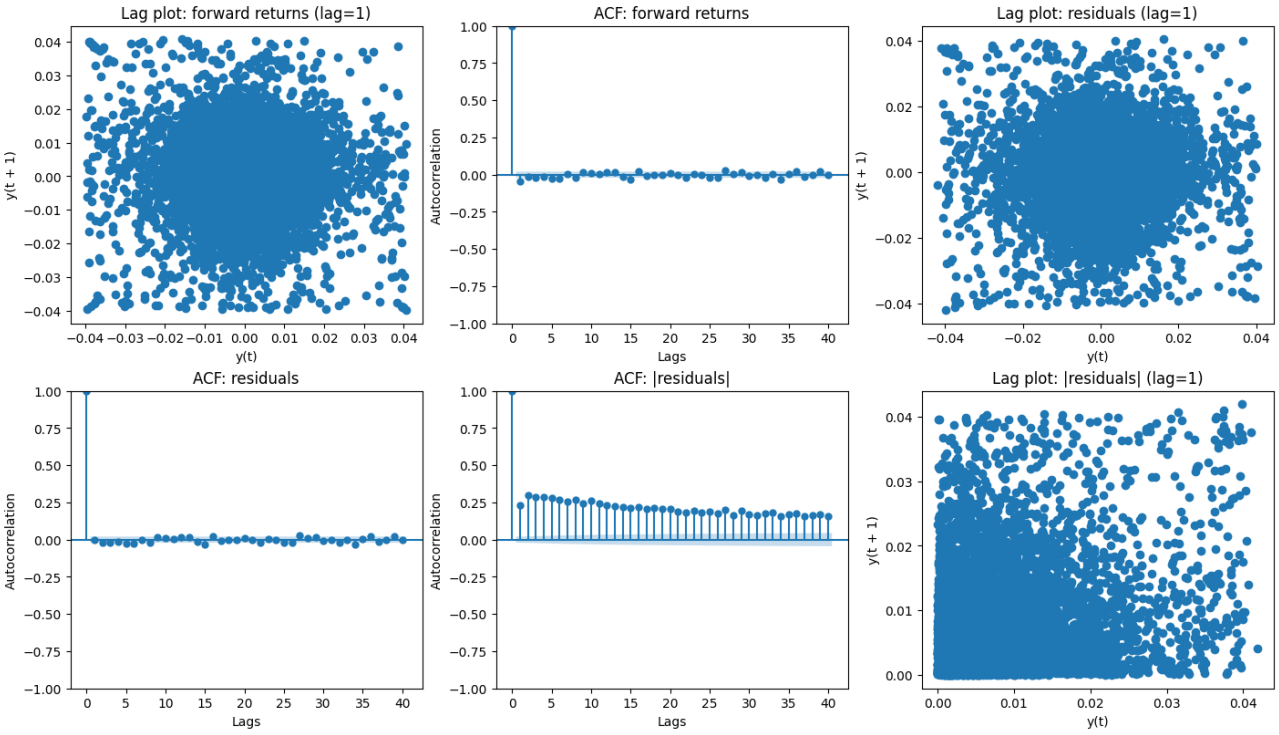

Figure 6: Autocorrelation plots.

2.3 Data partitioning for model formation

Partitioning a time-series dataset for model formation has its difficulties as the data has to keep its temporal structure. Our time-series is quite long and the variation seems to stay in a certain region. To perform robust training and validation for our model, we will use cross-validation in the form of a regular time-series split.

In a time-series split, we start with a smaller training data and validate on the next fixed amount of datapoints. Then we add the last validation set to our training set and validate on the next points. In this way, we use the most training data and we mimic the real life situation where we use all historical data to predict the future. The number of validation datapoints will be decided later, but we will start with 500. And the size of the first training data will be around a 1000 datapoints. These can change depending on the computational cost of the model and feature analysis. With this method, there is a possibility that the model learns patterns in the validation data and overperforms. If this seems to happen, we will test a slightly different form of splitting called the blocked time-series split, where the training data is always a fixed size and moves along with the validation data.

Our time-series has 9020 datapoints but many missing values in the beginning as we can see in figure 1. Because of this, we might have to ignore the first thousand datapoints, at least in the beginning. After we have performed an analysis on the features, we can better decide if we can use the first thousand datapoints and if we need to ignore more datapoints or predict some missing features. A little over two thousand of the last datapoints contain all the features and those datapoint will be used in the feture analysis.

References

Hull, B., Bakosova, P., Lanteigne, L., Shah, A., Sinclair, E. C., Fast, P., Raj, W., Janecek, H., Dane, S., and Howard, A. (2025). Hull tactical - market prediction. https://kaggle.com/competitions/hull-tactical-market-prediction. Kaggle.