Can Machines Explain Stock Returns?

Thesis Abstract

Karolína Chalupová

January 5, 2021

Recent research shows that neural networks predict stock returns better than any other model. The networks' mathematically complicated nature is both their advantage, enabling to uncover complex patterns, and their curse, making them less readily interpretable, which obscures their strengths and weaknesses and complicates their usage. This thesis is one of the first attempts at overcoming this curse in the domain of stock returns prediction. Using some of the recently developed *machine learning interpretability* methods, it explains the networks' superior return forecasts. This gives new answers to the long-standing question of which variables explain differences in stock returns and clarifies the unparalleled ability of networks to identify future winners and losers among the stocks in the market. Building on 50 years of asset pricing research, this thesis is likely the first to uncover whether neural networks support the economic mechanisms proposed by the literature. To a finance practitioner, the thesis offers the transparency of decomposing any prediction into its drivers, while maintaining a state-of-the-art profitability in terms of Sharpe ratio. Additionally, a novel metric is proposed that is particularly suited to interpret return-predicting networks in financial practice. This thesis offers a usable and economically explainable account of how machines make stock return predictions.