Master's Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University

Author:	**Bc. Karolina Chalupova**	**Supervisor:**	**doc. PhDr. Jozef Baruník, Ph.D.**
E-mail:	64087028@fsv.cuni.cz	E-mail:	barunik@fsv.cuni.cz
Phone:	720 237 273	Phone:	776 259 273
Specialization:	NEF	**Consultant:**	**Mgr. Martin Hronec**
Defense Planned:	July 2020	E-mail:	martin.hronec@fsv.cuni.cz
		Phone:	606 681 623

Proposed Topic:

Can Machines Explain Equity Returns?

Motivation:

Explaining why different assets earn different average returns is one of the most studied problems in finance. While the academia is typically interested in identifying common risk factors in returns, the industry strives to *predict* the returns to identify which assets to buy and which to sell. The task is similar: build a model that correctly explains (predicts) asset returns.

A first such model is the 1970s' Capital Asset Pricing Model (CAPM), which attributes an asset's expected return to its exposure to the market risk factor. Fama and French (1993) find two additional sources of risk, resulting in a three-factor model, later augmented to five factors (Fama and French 2015). Since Fama and French's publications, hundreds of papers have been reporting discoveries of yet new factors: Harvey, Liu, and Zhu (2016) count 316 different published factors, considering top journals only, and Cochrane (2011, 1047) describes the state of the research as a "zoo of factors".

One of the issues with the zoo of factors is that the predictors were typically found controlling only for the usual three or five factors and not the rest of the zoo, which is one of the leading reasons why "most claimed research findings in financial economics are likely false" (Harvey, Liu, and Zhu 2016, 1). Search for a good model is further complicated by possible nonlinearities and factor interactions (Gu, Kelly, and Xiu 2018). Cochrane (2011, 1060) calls for solving this multidimensional challenge: "First, which characteristics really provide independent information about average returns? Which are subsumed by others? Second, does each new anomaly variable also correspond to a new factor formed on those same anomalies? (…) Third, how many of these new factors are really important?"

Meanwhile, practitioners in finance industry are celebrating unparalleled results when they predict stock returns with machine learning (ML) models. Gu, Kelly, and Xiu (2018, 1) find that ML offers "an improved description of expected return behavior" compared to the standard linear regression and an "unprecedented" R^2 in the United States. Tobek and Hronec (2018) confirm this finding internationally, with their ML models 4 times more profitable than linear regression and offering 2 times the Sharpe ratio.

As argued by Gu, Kelly, and Xiu (2018), ML is particularly suited for addressing the multidimensional challenge. First, the sheer amount of 316 candidate explanatory variables (even without considering unknown number and form of their interactions), traditional estimation techniques loose many degrees of freedom. Second, high correlation of characteristics induces multicollinearity, increasing the variance of the estimates. Third, allowing for nonlinarities seems crucial for predicting returns (Gu, Kelly, and Xiu, 2018). ML methods are often *build* for dimension reduction and nonlinearity.

For the academia to answer the multidimensional challenge and for finance ML practitioners to understand their models better, it is necessary to *interpret* ML models, i.e., be able to predict how they link inputs to returns

predictions. ML models are scarcely interpretable per se, but additional methods can be used to extract interpretable information from them.

Research Question:

1) Which characteristics provide information about average equity returns?
2) Which interactions of characteristics are important?
3) What is the effect of changing a characteristic's value on average returns?

Methodology:

First, I am going to obtain data from CRSP, Compustat and Datastream on publically traded equity. Second, I am going to calculate the characteristics proposed by prior research as return predictors. Third, I am going to train several ML models with these characteristics as features. In doing so, I am going to pay particular attention to high correlation of characteristics to answer the research question meaningfully. This includes considerations of stability of the feature selection. Finally, I am going to examine the models to answer the research questions.

For interpretable models, the research questions can be answered directly. These models include regularized linear regression, principal components regression or partial least squares. A disadvantage of linear models is their significantly lower predictive power compared to complex models (Gu, Kelly, and Xiu 2018).

For complex models, the following methods can be used to make them interpretable and tackle the respective research questions:

1) Feature Importance. Feature importance measures describe how much a model relies on a feature to produce a prediction. Model reliance (Fisher, Rudin, and Dominici 2018) measures feature importance by calculating error in prediction resulting from permuting a feature. Intuitively, if permuting a feature (randomly shuffling its values across examples), does not have an impact on prediction error, the feature is not important for the prediction.

2) Feature Interaction. Friedman and Propescu (2008) propose H-statistic to measure a) to what extent two features interact with each other, b) to what extent a feature interacts with all other features. The statistic ranges between 0 and 1 and measures the variance of model output explained by the interaction. An advantage of H-statistic is that it captures all functional forms of interactions.

3) Accumulated Local Effects (ALE). Apley (2016) develops ALE to measure how a feature on average influences the prediction. For a given feature, ALE first defines a grid of values. Second, for each grid segment, ALE looks at examples such that their feature values fall in the given grid segment. Third, it calculates average prediction pretending the feature values are slightly higher (lower) than in the grid segment. Fourth, it subtracts these two averages. Therefore, one can interpret ALE as "increasing value of this feature by this unit increases output by x units, holding other features fixed at the average of the group, where the group consists of examples with similar values of analyzed feature". The advantage of ALE over partial dependence plot is that it holds other features fixed at the average of the local examples group, as opposed to global averages: holding other features fixed at global average is unreasonable in case of highly correlated features.

Finally, a nonlinear yet interpretable model can be used, as in the new paper by Bryzgalova et al. (2019), which introduces a variant of random trees, so called asset-pricing trees, to tackle the question of which variables and interactions are important for predicting stock returns. Through a new method of pruning random trees, the authors are able to identify common risk factors in stock returns.

Expected Contribution:

My thesis would contribute to a deeper understanding of equity returns and of the ML models trying to predict them. All my research questions are designed to address the first part of Cochrane's (2011, 1060) multidimensional challenge: „First, which characteristics really provide independent information about average returns? Which are subsumed by others?". Answering this question is essential for understanding of equity returns behavior and building any model, explanatory or predictive. My thesis fits into this strand of literature in

the following ways.

Several studies have use ML to study my first research question. Gu, Kelly, and Xiu (2018) calculate feature importance of their 94 characteristics in estimating equity returns. Their different ML models agree on selection of the most important characteristics (price trends, return reversal, momentum, stock liquidity, stock volatility, and valuation ratios). Kelly, Pruit, and Su (2017) use instrumental principal component analysis to instrument risk factors from 36 characteristics, with result similar to Gu, Kelly, and Xiu (2018) that only a small fraction of characteristics is important, again including return reversal and valuation ratios. I expect to use these studies as a benchmark.

The academic contribution is twofold. While Gu, Kelly, and Xiu (2018) find that feature interactions seems crucial for returns prediction, I am not aware of a paper trying to identify, measure, and interpret the interactions. By studying my second research question, I am going to shed light on which interactions are important. In my third research question, I am going to shift from the task of *identifying* important features to *measuring their effec*t on predictions, which is a key step in interpreting return-predicting ML models.

In addition to the academic contribution, my thesis can have a practical impact. ML models are becoming the norm in the finance industry. To debug, improve, and sell an ML model, interpretability of the model is highly important. From a practitioner's perspective, my thesis can serve as a case study of some of the methods, issues, and results that occur when interpreting return-predicting ML models.

Outline:

1. Motivation: ML is suitable for answering Cochrane's (2011) multidimensional challenge and solving the current issues of asset pricing research.
2. Literature review: State of current asset pricing research, existing answers to Cochrane's (2011) multidimensional challenge.
3. Data: Calculation and inspection of features.
4. Methodology: ML models used, interpretation techniques.
5. Results: Answers to my three research questions.
6. Conclusion: Implications of my results, areas for further research.

Core Bibliography:

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. *arXiv preprint arXiv:1612.08468*.

Bryzgalova, S., Pelger, M., & Zhu, J. (2019). Forest through the trees: Building cross-sections of stock returns. *Available at SSRN 3493458*.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047-1108.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1-22.

Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. *arXiv preprint arXiv:1801.01489*.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3), 916-954.

Gu, S., Kelly, B., & Xiu, D. (2018). *Empirical asset pricing via machine learning* (No. w25398). National Bureau of Economic Research.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5-68.

Tobek, O., & Hronec, M. (2018). Does it Pay to Follow Anomalies Research? International Evidence.

Author

Supervisor