Chair of Financial Management and Capital Markets
TUM School of Management
Technical University of Munich

Building an Anomaly Library: Using LLMs to Track Cross-Sectional Return Predictors

Keywords: Empirical Asset Pricing, Return Predictors, Anomalies, SSRN / arXiv paper screening, Text mining and NLP, Generative AI

Project description

Empirical asset pricing has produced a fast-growing “anomaly zoo” of firm characteristics that predict cross-sectional stock returns. Many of these predictors first appear in working papers on platforms such as SSRN and arXiv, and the literature is now too large and dynamic to track manually. For researchers and practitioners, it is increasingly important to have a structured, continuously updated overview of which predictors exist, where they were tested, and how robust their results are.

In this project, the student will design and implement a Python-based tool that automatically builds and maintains an “anomaly database” from papers published on SSRN and arXiv. The tool will regularly scan new publications and use modern large language models to classify whether a given paper studies cross-sectional return predictors (e.g., firm-level characteristics predicting stock returns) or not. Relevant papers will be flagged and further processed.

For each identified paper, the tool will then extract key information using generative AI and convert it into structured database entries. Typical fields include, for example, the sample (time period, markets, assets), a concise description of the predictor, the underlying data sources (e.g., accounting, prices, analyst forecasts), the empirical methods (portfolio sorts, Fama–MacBeth regressions, machine learning models, etc.), etc. Together with the supervisors, the student will design the database schema and the extraction prompts so that the anomaly database can be extended in a consistent way.

Specifically, a central goal of this IDP is to create a pipeline that supports **continuous updating**: new SSRN/arXiv papers can be fetched at regular intervals, classified, summarized, and added to the database with minimal manual intervention. This requires not only integrating LLM-based components, but also careful attention to data quality, versioning, and validation (e.g., logging model outputs). Over the course of the project, the student will gain hands-on experience at the intersection of finance and natural language processing, working with large text corpora, APIs, and modern AI tools to build a research-grade data product.

The skills developed in this IDP—combining academic finance, LLM-based information extraction, and robust software engineering—are highly relevant for careers in quantitative asset management, hedge funds, research-driven FinTechs, and data science roles in banks or consulting, as well as for a potential future in empirical finance research.

Chair of Financial Management and Capital Markets
TUM School of Management
Technical University of Munich

What we are looking for

- Strong analytical and project management skills
- Determination and passion for your areas of expertise
- Good Python programming skills
- Interest to work at the intersection of finance and IT
- 1 or 2 persons

What we offer

- Knowledge in quantitative finance, corporate finance and machine learning
- Kick-off session including introduction to relevant finance and/or business topics
- Experience with IDPs
- Open dialogue and support
- Access to prime capital markets databases (Bloomberg, Datastream, Thomson Reuters, etc)
- Potential for publication and/or evaluation of future use cases
- Both single and group projects are possible

Confidentiality of Code and Data

This IDP is based on proprietary financial data and internal research infrastructure provided by the chair. The Python code, scripts and documentation developed in this IDP are **confidential** and become the property of the chair. They **must not be shared** with third parties and **must not be published** in any public or semi-public repository (e.g. GitHub, GitLab, Bitbucket, Kaggle, etc.).

By participating in this project, you agree to comply with these confidentiality requirements.

Interested?

Please send an e-mail with CV, academic transcript and your preference for this project to sebastian.mueller.hn@tum.de and to david.osterrieder@tum.de.

Questions?

In case of any (e.g. topic related) questions, please contact Prof. Dr. Sebastian Müller (sebastian.mueller.hn@tum.de) and/or David Osterrieder (david.osterrieder@tum.de).