1. Background

Factor investing has enjoyed its success as one of the forefront strategies in equity investment. The paradigm started out as mainly using fundamental analysis to analyze companies, as seen in the three-factor model (Fama and French 1993), consisting of the market, size, and value factors. The model is an expansion of Eugene Fama's and Kenneth French's Capital Asset Pricing Model, explaining a stock's character by only measuring the risk and return relative to the market. Presently, factor investing has expanded its analysis with the inclusion of technical and quantitative aspects, employing factors such as momentum (Asness 1994; Carhart 1997; Jegadeesh and Titman 1993) and volatility. Therefore, factor investing utilizes the three pillars of investing: fundamental, technical, and quantitative analyses. From this standpoint, the goal of the model is to implement a portfolio that enhances diversification, drives above market-relative returns, and decreases risk. Due to different risk premium characteristics, some factors might favor certain market conditions, causing those factors to fail. But because of diversification benefits, each factor's risk will compensate each other, making the overall model effectively timeless. There are many attempts of bending the original factor model from time to time, adding and tweaking factors with the goal to accomplish a greater risk premium. For example, time series momentum (Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen 2011), is an alternative strategy to factor investing's cross sectional momentum, capturing an individual asset's momentum anomaly instead of analyzing a group of assets' momentum.

2. Problem Statement

Throughout the majority of asset management history, factor investing was executed by separating each factor in silo portfolios; one might have a portfolio for value and a separate portfolio for quality. Recent studies have illustrated that with the birth of quantamentals investing, investors have attempted to combine fundamental and technical factors under a quantitative framework. The term quantamentals is a portmanteau combining "quant"itative and fund"amental" investing; it defines a fusion of computer and mathematical models with fundamental methods' that analyzes company characteristics. The framework is still young and research on the topic is scarce, which may imply that 1. The underlying practical approach to quantamentals is idiosyncratic and confidential to each investing entity and 2. There are a plethora of approaches (no ideal way) to execute a quantamentals model. This proceeding induces an opportunity to utilize data science to explore the risk premium of equity factor investing under the quantamentals framework and to introduce a quantamentals strategy using machine learning.

3. Data

We will be basing our research on the Russell 3000 universe of stocks, since it contains the top 3000 stocks ranked by market capitalization. This is in conjunction with our size factor, and ensures sufficient trading volume. Our data's time window will start in 2002-01 and end in 2020-09.

3.1 Data Sources

All data sources are paid subscriptions.

Norgate Data - Norgate is a bias-free data provider that provides economic indicators, historical constituents, and end-of-day price data. Their data is survivorship bias-free, which is extremely crucial since we are basing our analysis on a universe like Russell 3000. Price data is also back-adjusted if needed, with the option of 'total return' adjustment, where price is accounted for dividends and splits. We will be keeping track of the active/delisted constituents of Russell 3000 and will be using daily price data from Norgate to backtest and create price derivatives (i.e. momentum factor) if needed. Norgate data is stored locally, and it provides software to keep an asset's price data as a CSV. However, we will be using an api to access the data.

Sharadar - Sharadar provides numerous datasets including historical fundamentals and price data. They are also survivorship bias-free. We will be utilizing Sharadar's fundamentals dataset to analyze companies' factors like quality and value. Particularly, we will be using the 'As Reported' data, since this is forward looking-bias free and contains metrics as reported on every quarter and is not a restatement of a company's report. Data is stored locally as CSVs.

3.2 Concerns

Since we will be using two different data providers, one obstacle that arises is the difficulty in merging and cleaning data. Symbol names are different in each provider, therefore a dictionary reference in joining technical and fundamental features has to be created. Also, a company may fail to report their financials resulting in missing data. Because of this, we will be using trailing 12 months as reported data. Other concerns will be addressed in the report.

4. Solution

Referring back to factor investing, the method attempts to rank assets based on different characteristics. Traditionally, investors would group assets according to a single characteristic, by selecting assets to invest by ranking by factor magnitudes. We remedy for the lack of factors interaction effect by using a K-means Clustering model. This way, we let the algorithm determine the valid clusters of different factors and we will evaluate the risk premium by backtesting on each cluster. By using a (unsupervised) clustering model, we are essentially creating a quantamentals model that groups assets and combines fundamental and technical factors without having constraint rules that might overfit in unseen data. We are also using clustering to determine unknown interactions; i.e. assets with high quality and high momentum might provide better profits than assets with high value and high quality.

5. Benchmark model

We evaluate the profitability of our model by comparing the model's backtest performance with investment benchmark ETFS such as the SPY (SP500) and VTHR (Russell 3000), and individual factor model performance.

6. Evaluation Metrics

Model performance will be compared with other benchmarks by observing the Sharpe Ratio, Annual Returns, Cumulative Returns, Annual Volatility and Maximum Drawdown.

7. Project Design

Factors that are included are value, quality, volatility and momentum. Implementation of features and metrics used are defined in more detail in our project report.

7.1 Strategy Overview

Our investments will be rebalanced monthly. All stocks' position size will be equally weighted.

7.1.1 Data Preprocessing: Importing data/metrics and creating features that represent our factors. This will result in a wide data format consisting of stock's factor features. Our data also keeps tracks of active and delisted companies, and historical constituents of Russell 3000.

7.1.2 On our rebalance day, we select only assets that are included in the Russell 300 universe.

7.1.2 We then group those assets into clusters using K-means with the goal to discover unknown interaction effects between certain features. A particular concern is selecting the number of K clusters, and we propose using either the elbow method or silhouette score. Some exploratory data analysis will be employed to see how the algorithm clusters assets.

7.1.3 We'll be adding an extra step by using a universe trend filter. For example, if the SP500 is on an uptrend we will be in the market, and conversely in a down trend, we will turn our investment switch off.

7.1.4 Backtest the model to evaluate which cluster (which characteristics) is the most profitable and compare with benchmark models described earlier.