Quant Models and Back Test: An overview

This note provides an overview of quant models and back testing their performance with archival data from CRSP and Compustat.

I. Quant Model

The objective of a quant model is to predict stock returns based selected features and construct portfolios.

Sample of stocks: In class, use samples of all US listed common stocks. The samples can be restricted to subsets, such as small cap or large cap stocks, or consumer durables and staples.

Features: Choose features that are likely to capture misvaluation or sentiment and predict returns.

Examples:

- Book-to-market ratio
- Cash flow to price
- Price momentum (past returns)
- New stock Issues/ Repurchases

Predicted Returns

Predict returns as a function of stocks' features.

Approaches:

- Rank and aggregate features: e.g.. The JP Morgan's Q-Model
- Fitted values of Regressions
- Decision Tree
    - Random Forest
    - XGBoost
- LLM

Portfolios:

Long portfolios: Invest only in stocks with the largest predicted returns. Form portfolios that are:

- **Equal-weighted**: Invest equal dollar amounts in stocks with the largest predicted returns,
- **Market-cap weighted**: Invest dollar amounts proportional to market cap in stock with largest predicted returns, or
- **Risk managed:** optimized to minimize portfolio risk

Short portfolios: Short-sell stocks with smallest predicted returns.

Long/Short portfolios: Combination of long and short portfolios

II. Back Tests

Back tests examine the historical performance of quant models using archival data. Our primary data sources are:

1. CRSP: Market data at monthly and daily frequencies. For example,
    a. Returns
    b. Price
    c. Trading Volume
    d. Shares outstanding

2. Compustat: Data from financial statements (Balance sheet, Income statement and Cash flow statement), e.g.
    a. Sales
    b. Net income
    c. Current Assets
    d. Total Assets
    e. Current Liabilities
    f. Total Liabilities
    g. Book Value of Equity
    h. Cash flow from operations

Yahoo Finance is a data source for daily price and volume data, and selected recent financial statement data.

Back tests evaluate the historical performance of quant models using archival data. It is critically important to avoid any look-ahead bias when using archival data – Ensure that the information that the back tests use are available at the time the trade is executed.

CRSP and Compustat

- AAPL data on CRSP (see **'crsp_aapl.csv**')

PERMNO	date	SHRCD	EXCHCD	TICKER	PRC	RET	SHROUT	CFACPR	Market Cap
14593	1/31/2022	11	3	AAPL	174.78	-0.01571	16319441	1	2,852,311,898
14593	2/28/2022	11	3	AAPL	165.12	-0.05401	16319441	1	2,694,666,098
14593	3/31/2022	11	3	AAPL	174.61	0.057473	16207568	1	2,830,003,448

Each row of the CRSP data pertains to the month ending on 'date.' For example, the last row contains data for March, 2022. In this row 'RET' is the return for the month ending 3/31/2022. 'PRC' and 'SHROUT' data are AAPL's price and shares outstanding as of 3/31/2022.

Q. Market Cap is one of the features in your model. What is the feature value for AAPL that should be used to predict AAPL's return for March '22?

- AAPL data on Compustat (see 'cstat_aapl_ibm.csv')

LPERMNO	datadate	fyear	indfmt	consol	popsrc	datafmt	tic	curcd	ceq
14593	9/30/2019	2019	INDL	C	D	STD	AAPL	USD	90,488
14593	9/30/2020	2020	INDL	C	D	STD	AAPL	USD	65,339
14593	9/30/2021	2021	INDL	C	D	STD	AAPL	USD	63,090
14593	9/30/2022	2022	INDL	C	D	STD	AAPL	USD	50,672
14593	9/30/2023	2023	INDL	C	D	STD	AAPL	USD	62,146

Q. Suppose book value (ceq) is one of the features in your model. What is the ceq value for AAPL that should be used to predict AAPL's return for January '22?

Merge CRSP and Compustat data (CompustatCRSP_merge.ipynb)

- Use 'Permno' as stock id.

```
Cstat_book_eps.rename(columns = {'LPERMNO' : 'PERMNO'}, inplace = True) # Renaming "LPERMNO" for merging Cstat_book_eps with Returns data
```

- Assume that financial statement data are not available for at least four months after 'datadate'

```
Cstat_book_eps['date'] = Cstat_book_eps['date'].apply(lambda x: x + DateOffset(months=+5)) # Adding five months (using DataOffset library) assuming it takes at most 4 months for the data to reach the market
```

- Because CRSP data are monthly and Compustat data are annual, each Compustat 'datadate' would match with multiple CRSP 'date' records. Use `pd.merge_asof to merge`

```
merged_data = pd.merge_asof(Returns, Cstat_book_eps, by = 'PERMNO', left_on = 'date', right_on= 'date', tolerance=dt.timedelta(days = 365)) # Merging "Returns" & "Cstat_book_eps" dataframe on "PERMNO" & "date" with 1 year tolerance for date
```

- Output (aapl_merge.csv)

PERMNO	date	RET	marketcap as of	marketcap ('000)	datadate	ceq ($ m)	b2m
14593	9/30/2022	-0.12	8/31/2022	2,526,643,629	9/30/2021	63,090	0.0250
14593	10/31/2022	0.11	9/30/2022	2,203,381,335	9/30/2021	63,090	0.0286
14593	11/30/2022	-0.03	10/31/2022	2,439,350,814	9/30/2021	63,090	0.0259
14593	12/30/2022	-0.12	11/30/2022	2,354,878,708	9/30/2021	63,090	0.0268
14593	1/31/2023	0.11	12/30/2022	2,058,403,783	9/30/2021	63,090	0.0306
14593	2/28/2023	0.02	1/31/2023	2,282,948,430	9/30/2022	50,672	0.0222
14593	3/31/2023	0.12	2/28/2023	2,332,313,060	9/30/2022	50,672	0.0217