Analysis & Prediction: Mutual Funds vs. ETFs

By Joanne Chong

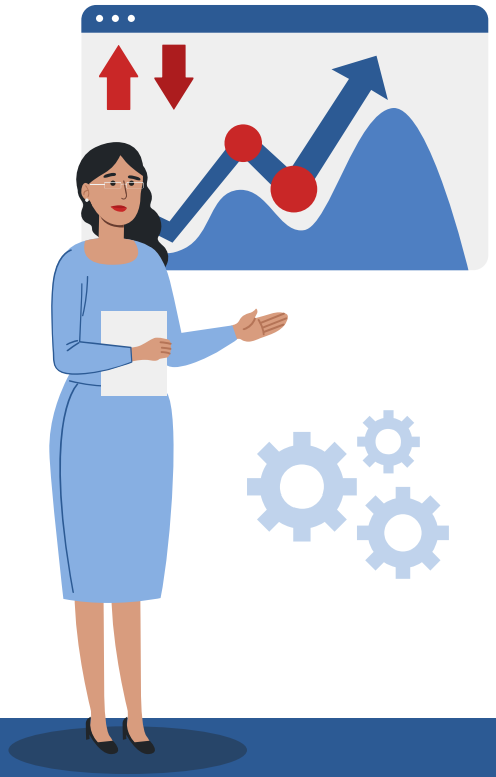


TABLE OF CONTENTS

01

BACKGROUND

Mutual funds vs. Exchange-traded funds (ETFs)

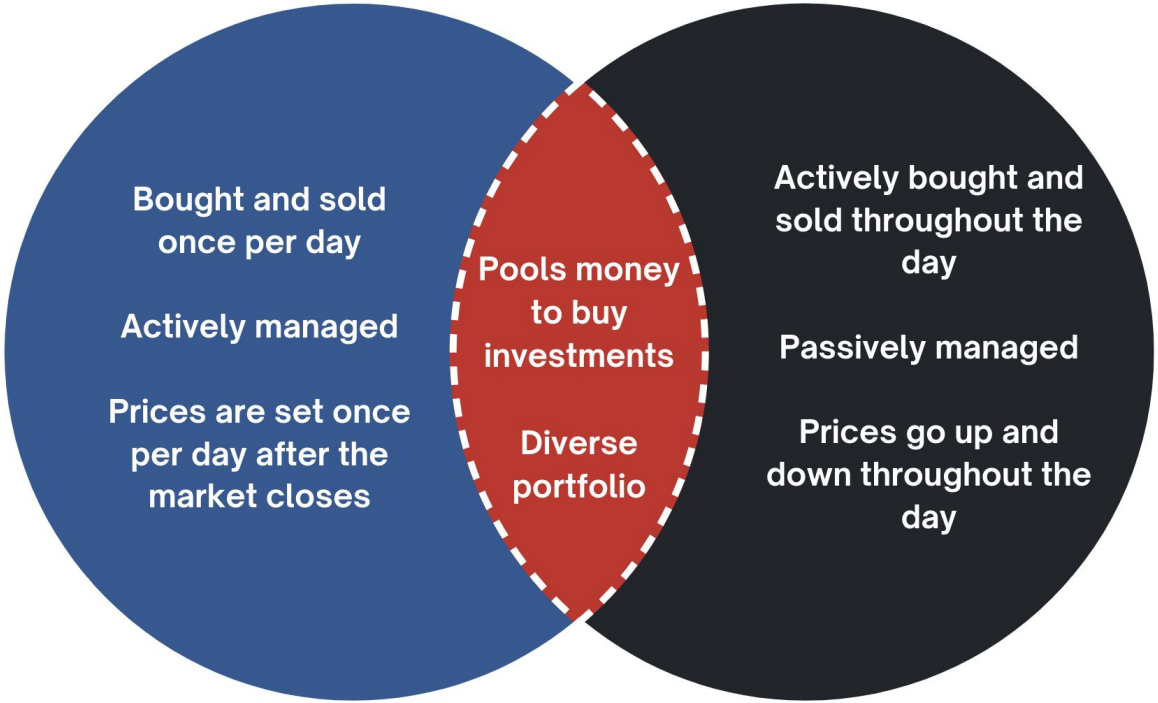


Diagram adapted from Ramsey Solutions

Declining sales of active mutual funds

- In 2022, investors pulled US$ 879 billion out of active mutual funds in the first 11 months
- Mutual funds' assets dropped 18% from the previous year, due to redemptions and market depreciation
- Active mutual funds reported consecutive monthly net outflows since October 2021
- Passive funds recorded US$51.6 billion in net inflows
- More than 400 exchange-traded funds (ETFs) launched in 2022

FINANCIAL TIMES

Market pressures add to US active mutual fund woes

The vehicles suffered outflows of $879bn in the 11 months to the end of November

Financial Times, 19 January 2023

Problem statement:

Which of these funds - mutual funds or ETFs - would make a better investment option?

How we'll address the problem:

1. **Build prediction models** that would accurately predict the performance of mutual funds vs. ETFs
2. Identify **influential factors** for the model prediction accuracy

The identified model and feature will then be used to **help fund managers and investors** make smarter investment decisions.

02

METHODOLOGY

About the datasets

Mutual Funds
23,782 data points of US-based mutual funds, with 298 columns columns that covered:

- Yearly and quarterly fund returns from 2000 to 2021
- Asset composition
- Share price and returns performance indicators
- Morningstar ratings and ESG ratings

ETFs
2,309 data points of US-based ETFs with 142 columns columns that covered:

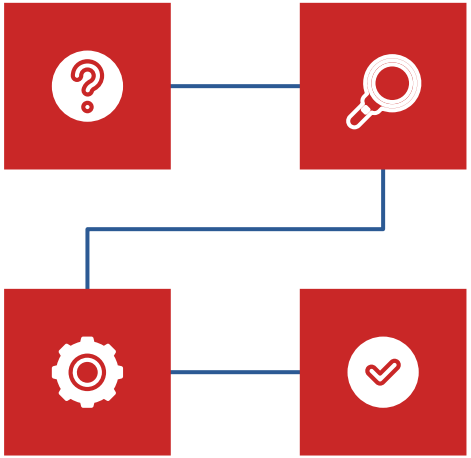
- Yearly returns from 2000 to 2020
- Asset composition
- Share price and returns performance indicators

Note: *Datasets were taken from Kaggle*

Overview of the workflow

STEP 1 - CLEAN

Identify missing values & gaps in dataset



STEP 2- EXPLORE

Investigate patterns and discover trends

STEP 3 - MODEL

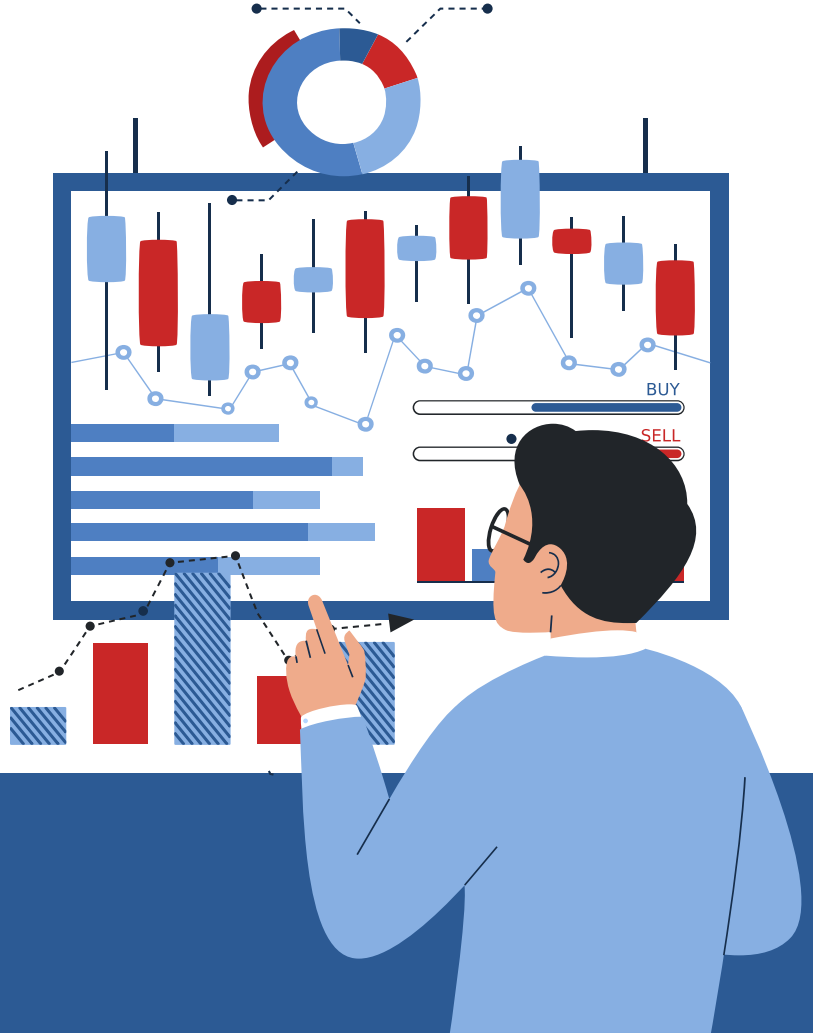
Run prediction trials and assess errors

STEP 4 - EVALUATE

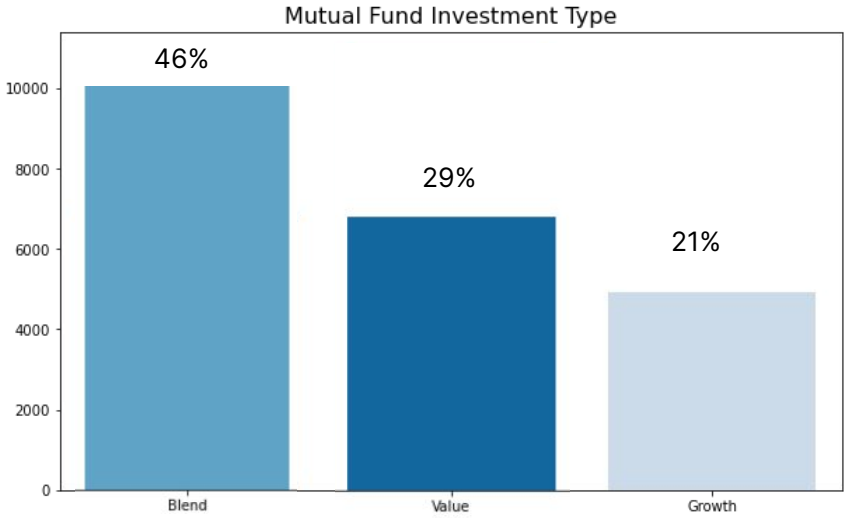
Select model with highest accuracy and features

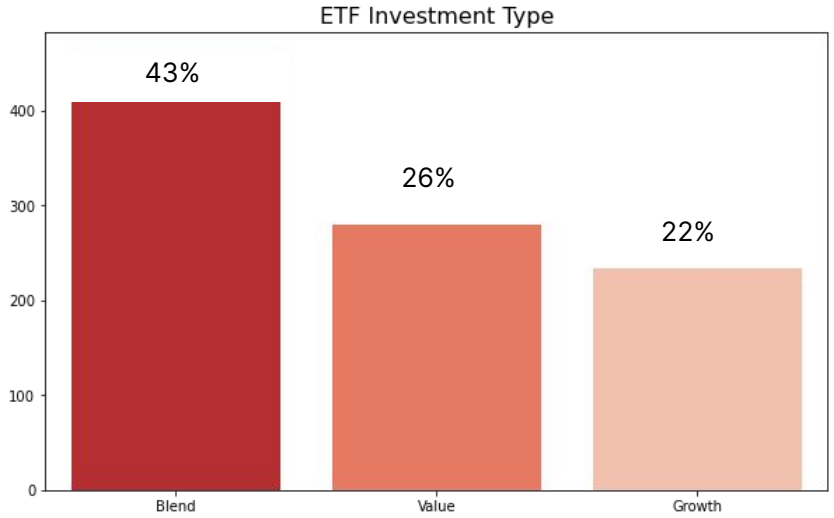
03

EXPLORATORY DATA ANALYSIS



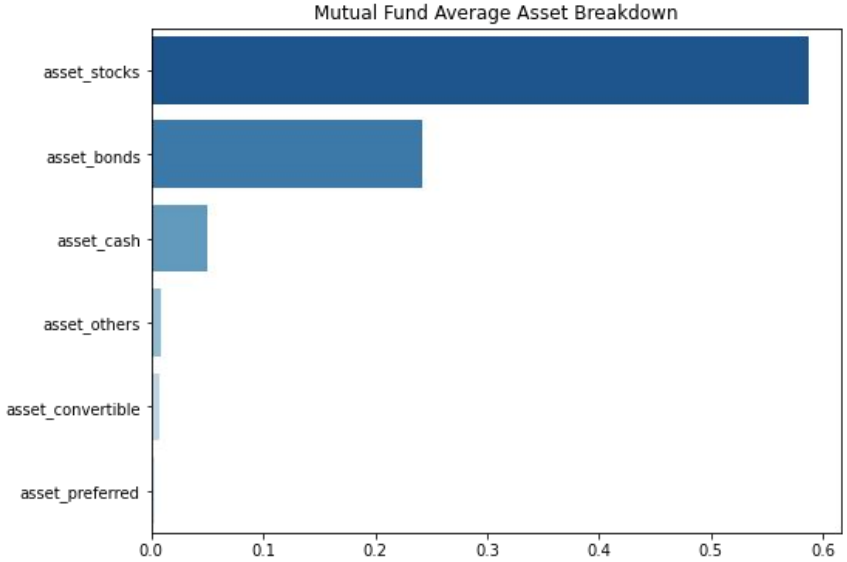
Blend portfolio is the most popular strategy



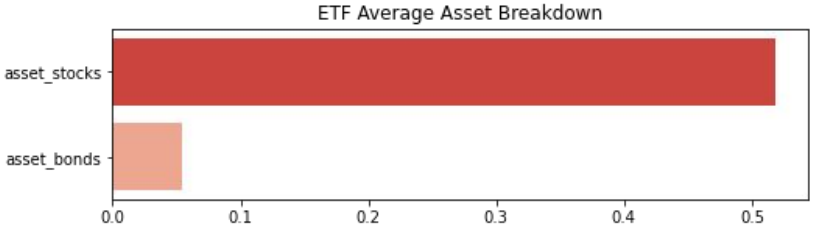


1. **Growth** are stocks or shares with expected high growth rates, e.g. tech and healthcare companies
2. **Value** are shares traded at lower prices, usually older and more established companies, e.g. finance or energy companies
3. **Blend** is a combination of the both

Stocks & bonds make up the most in assets



The top 3 avg. asset: **stocks (58.8%), bonds (24.2%) and cash (5%)**



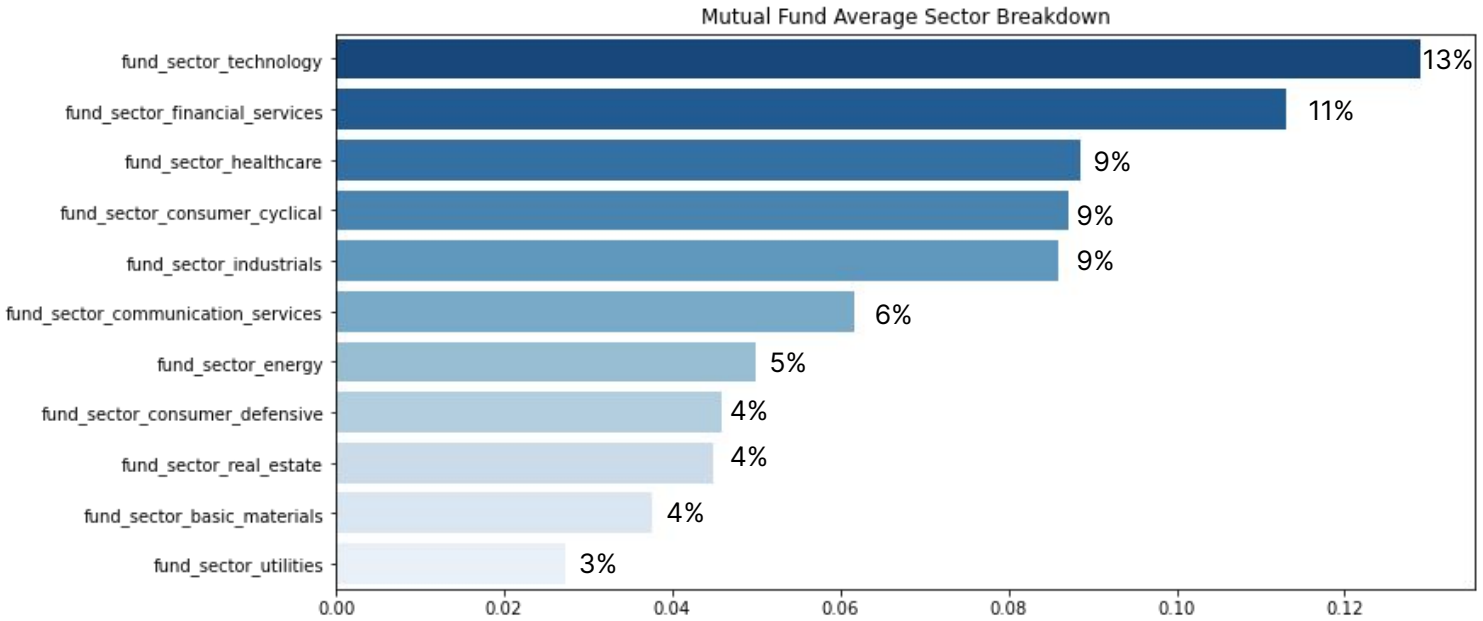
The top avg. asset: **stocks (51.8%) and bonds (5.5%)**

Large portfolio dominates both mutual funds & ETFs

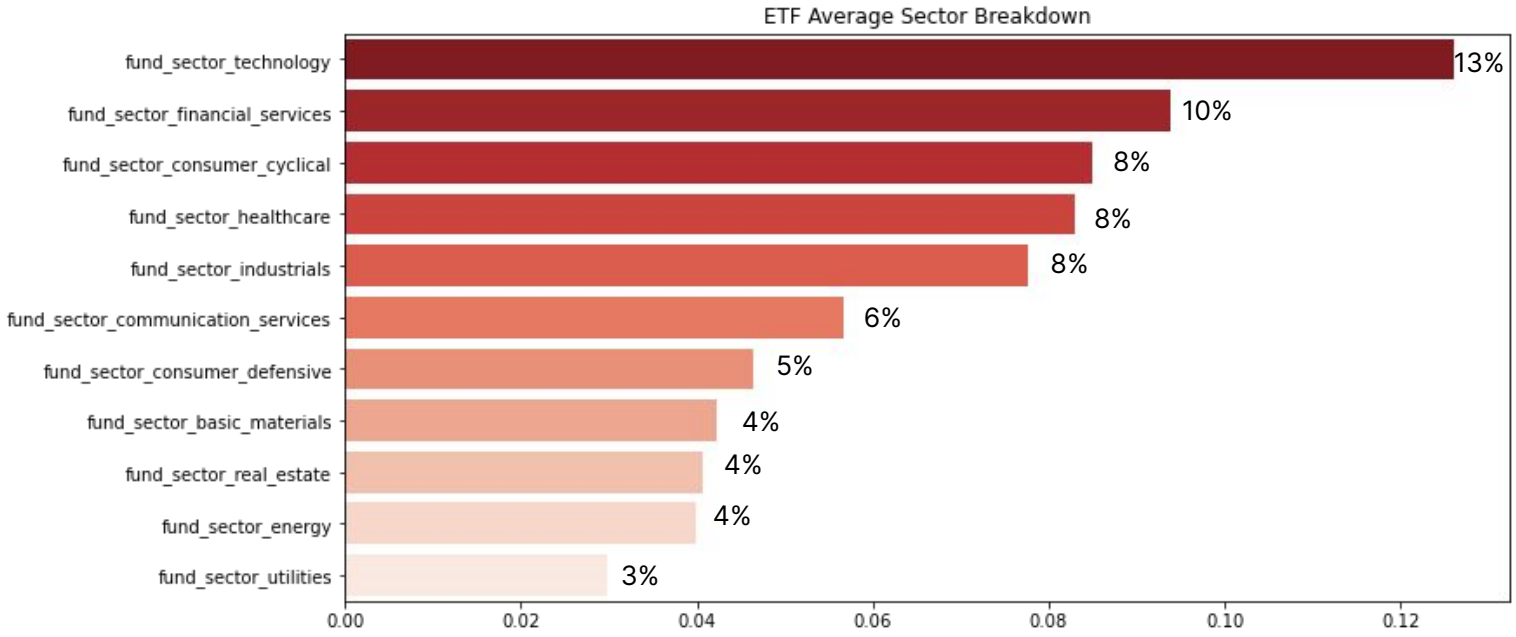
	Mutual Funds	ETFs
Large	• $6.2 mil avg. total net assets • 56% Blend, 22% Growth, 22% Value	• $3.5 mil avg. total net assets • 55% Blend, 23% Growth, 22% Value
Medium	• $3.8 mil avg. total net assets • 43% Blend, 23% Growth, 34% Value	• $0.8 mil avg. total net assets • 31% Blend, 27% Growth, 43% Value
Small	• $2.8 mil avg. total net assets • 26% Blend, 20% Growth, 53% Value	• $1.3 mil avg. total net assets • 41% Blend, 22% Growth, 36% Value

Note: Fund size is the total asset base, or total amount of money that is being managed and invested

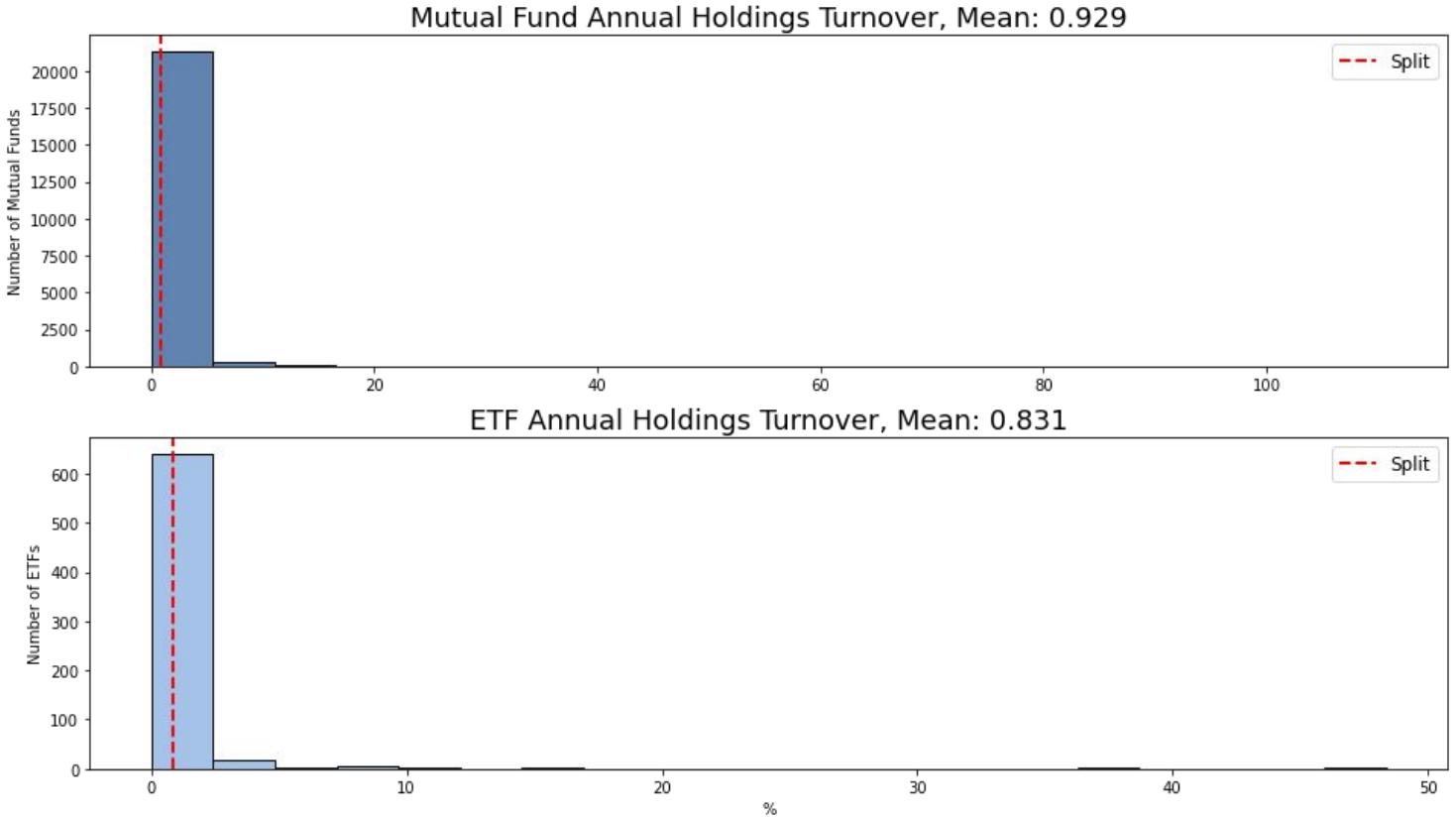
Tech, financials and healthcare tops mutual fund portfolio



Tech, financials and consumer cyclical tops ETF portfolio

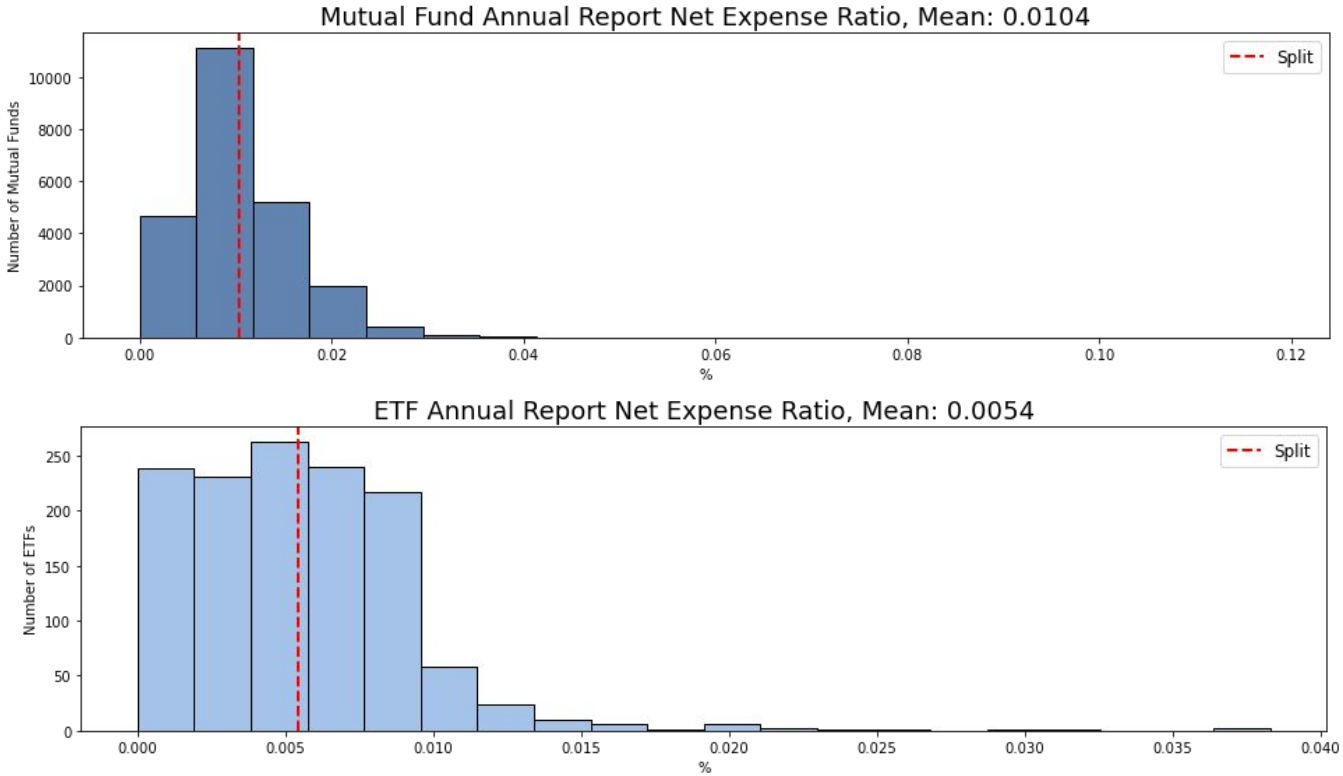


Mutual fund has higher annual holdings turnover



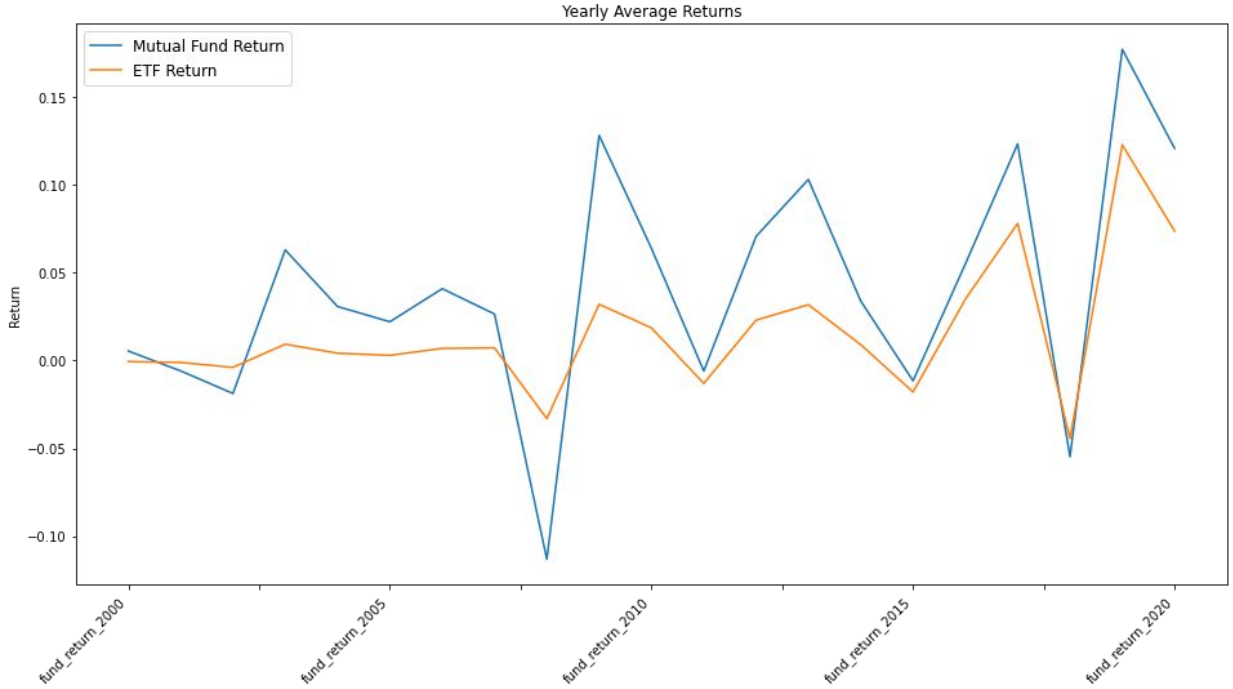
Note: The annual turnover refers to how many times an asset, security, or payment changed hands over a year-long period

Operating costs for ETF nearly half of mutual funds'



Note: Net expense ratio is the fee to manage the fund; it is calculated by taking the total fund costs divided by total fund assets

Mutual funds reports higher fund returns

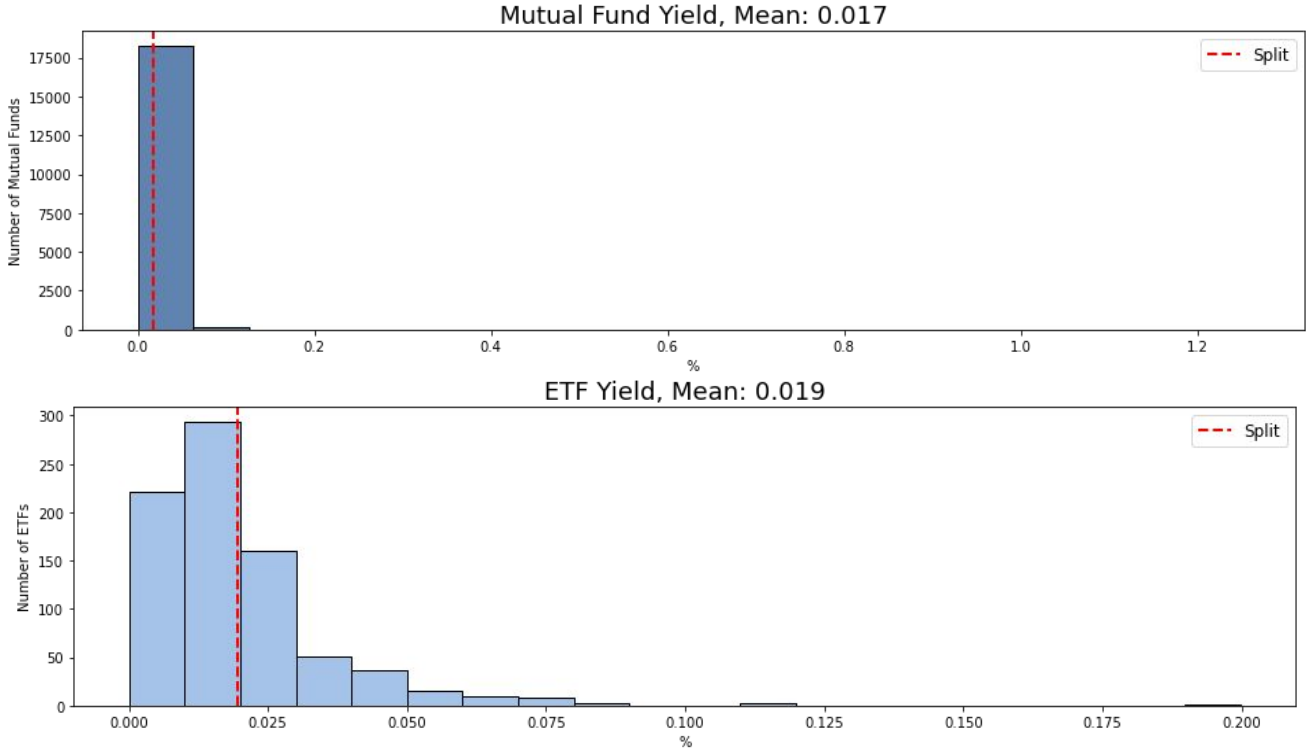


Mutual fund avg. YTD return: **9.2%**

ETF avg.
YTD return: **7.3%**

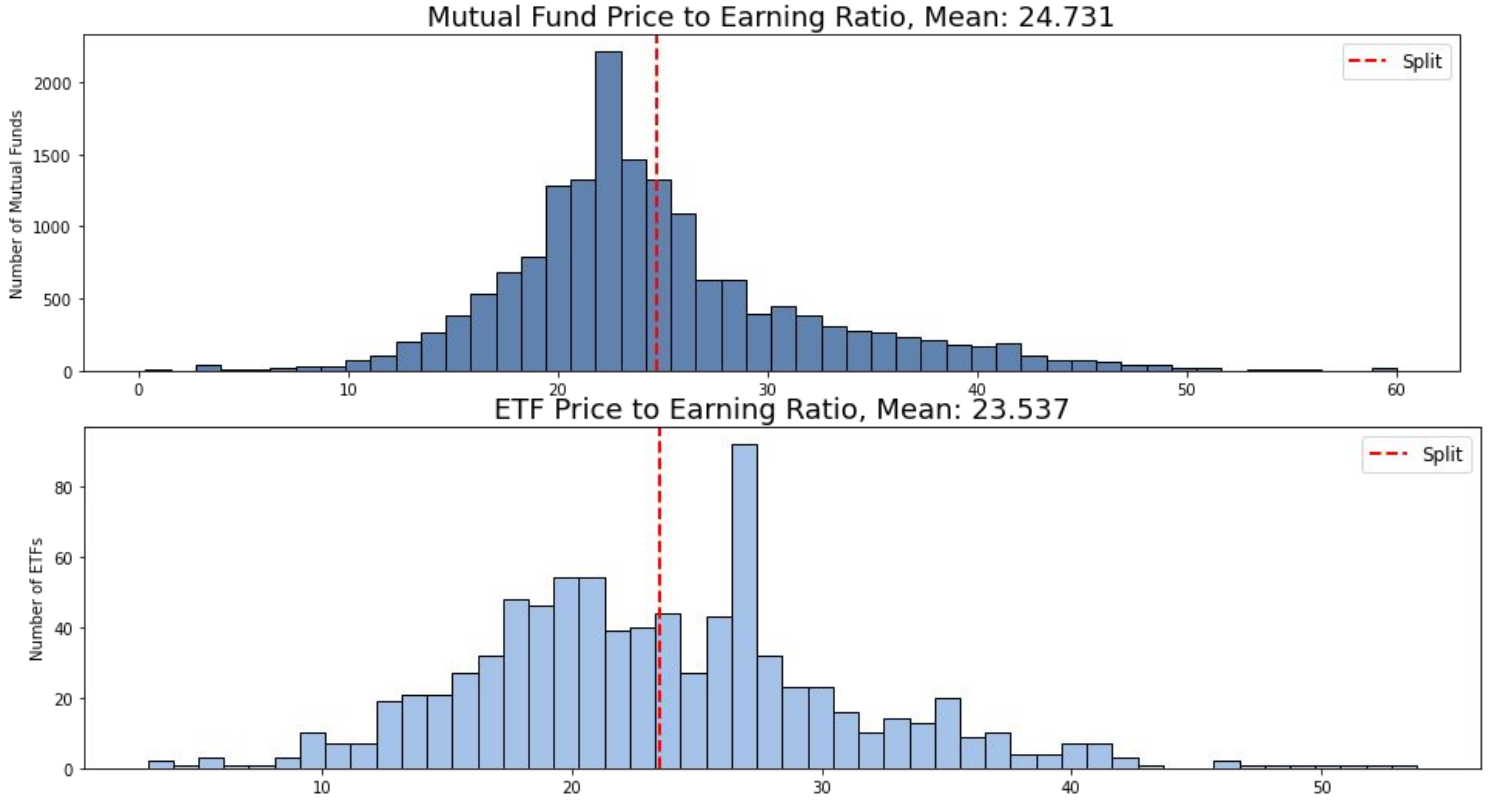
Note: Return is the change in investment over a time period

The yield from the funds only differ by 0.2%



Note: Yield is the income earned on an investment, such as interest and dividends

The price per share vary slightly more than 1 point



Note: Price-to-earnings (P/E) ratio is how much investors pay for a share compared to the earnings a company generates per share

Large fund generated alpha closest to zero

	Alpha ratio	
	Mutual Funds	ETFs
Large	-0.41	-0.40
Medium	-1.45	-2.46
Small	-2.49	-4.15

Note: Alpha refers to the investment's strategy to beat the market; positive alpha is desirable. Alpha of 0 means the fund is tracking well against benchmark.

Growth fund generated positive alpha

	Alpha ratio	
	Mutual Funds	ETFs
Blend	-1.32	-0.97
Growth	2.56	1.44
Value	-3.11	-4.30

Note: Alpha refers to the investment's strategy to beat the market; positive alpha is desirable. Alpha of 0 means the fund is tracking well against benchmark.

04

MODELING & EVALUATION

Selected features for modeling

Predictors:

- Type of investment - Blend, Value, Growth
- Size of investment - Large, Medium, Small
- Fund sectors - tech, healthcare, financials, energy, etc.
- Asset types - stocks, bonds, cash
- Total net assets
- Annual holdings turnover
- Operating costs (net expense ratio)
- Fund yield
- Price-to-earnings ratio
- Bond maturity and duration
- Bond ratings
- Last dividend *(only mutual funds)*
- Morningstar overall ratings *(only mutual funds)*
- ESG score *(only mutual funds)*

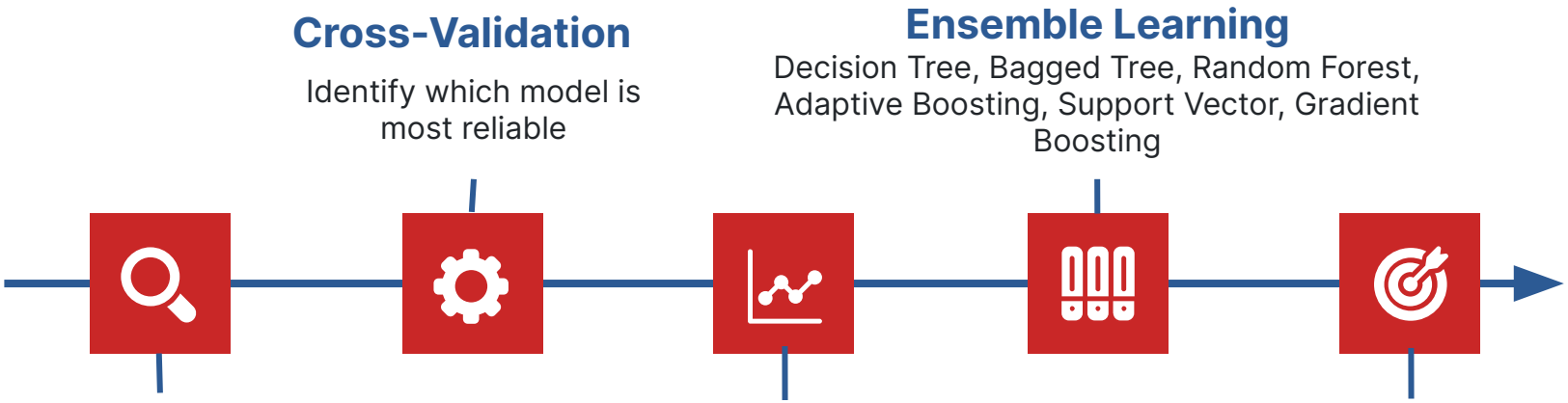
Target: Year-to-date return, Alpha ratio

Size of data for modeling:

Mutual funds:
22,586 rows & 40 columns

ETFs:
973 rows & 33 columns

Generating predictions workflow



Train/Test Split

Splitting the dataset for training

Cross-Validation

Identify which model is most reliable

Supervised Learning

Linear Regression, Lasso Regression, Ridge Regression

Ensemble Learning

Decision Tree, Bagged Tree, Random Forest, Adaptive Boosting, Support Vector, Gradient Boosting

Hyperparameter tuning

Finding set of optimal learning algorithm values to maximize accuracy

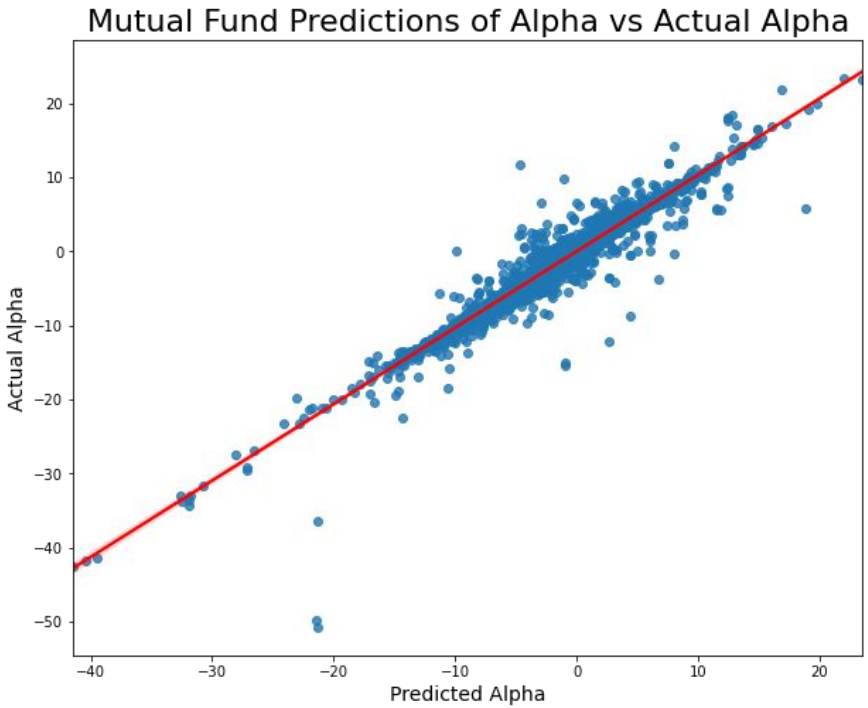
Alpha predictions achieved best scores

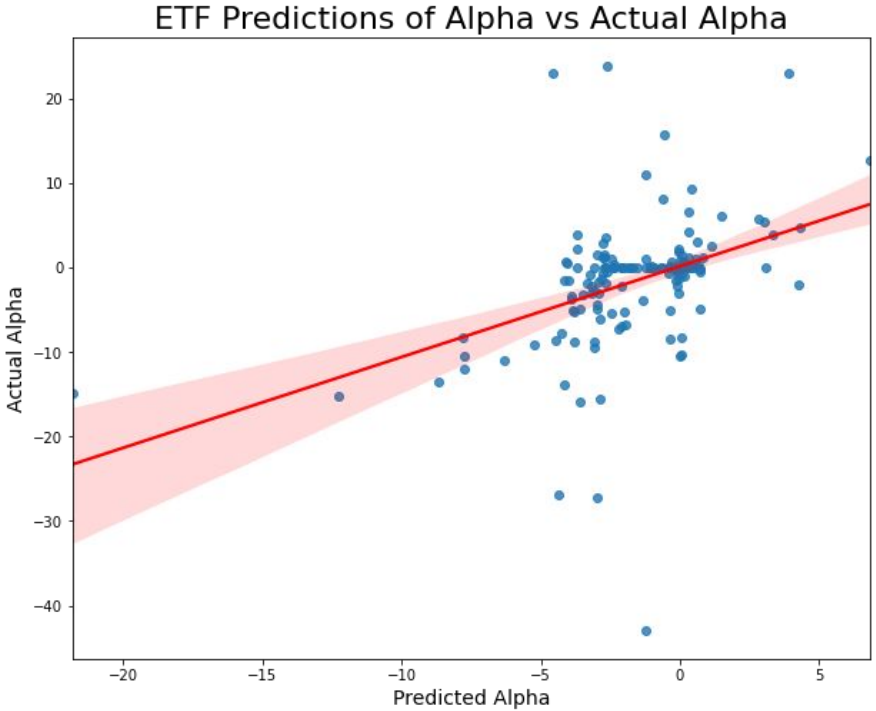
Using GridSearch to tune Random Forest model helped achieve these optimal parameters:

- N_estimators: 150, 200
- Max_depth: 4, None

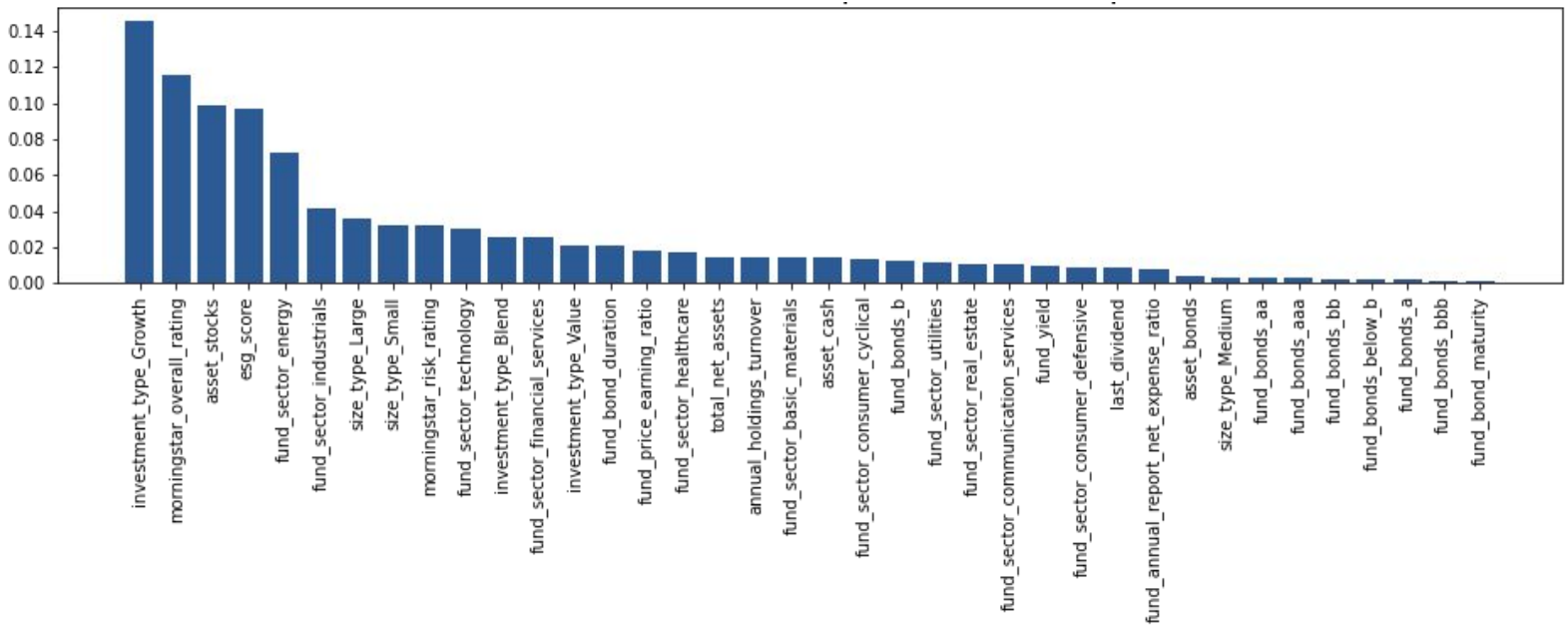
	Best score	R2 Train	R2 Test	RMSE Train	RMSE Test
Mutual fund prediction of YTD returns	0.9246	0.9919	0.9410	0.0072	0.0193
Mutual fund prediction of alpha	0.9290	0.9921	0.9294	0.4438	1.3637
ETF prediction of YTD returns	0.2315	0.5287	0.1998	0.0814	0.1341
ETF prediction of alpha	0.2685	0.5661	0.1871	3.6966	6.029

Higher accuracy in mutual fund alpha prediction



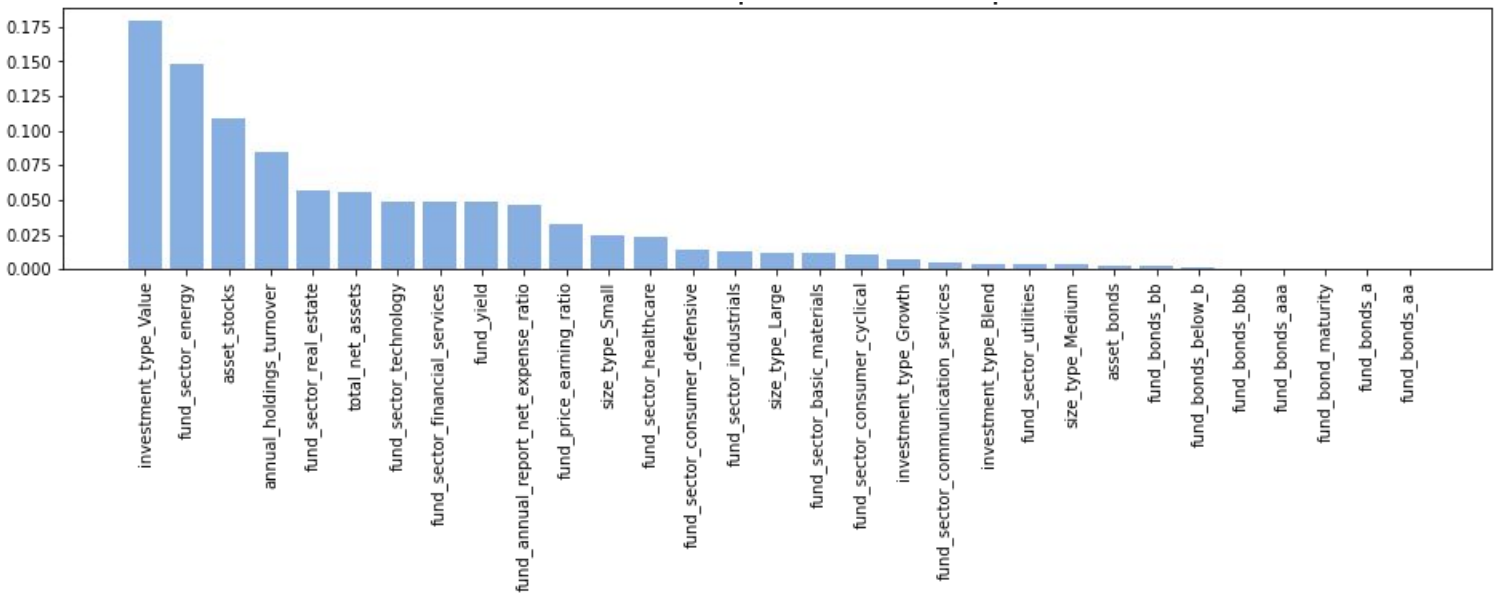


Top features in mutual fund alpha prediction



Top 10 features: Investment type (Growth), Morningstar overall and risk ratings, stocks asset breakdown, ESG score, fund sectors (energy, industrials, technology), Investment size (Large, Small)

Top features in ETF alpha prediction



Top 10 features: Investment type (Value), fund sectors (energy, real estate, technology, financial services) stocks asset breakdown, annual holdings turnover, total net assets, fund yield, operating costs

05

CONCLUSION

In summary

Solely based on prediction accuracy models*, the **mutual fund is a better investment option**.

The top 10 features are:

1. Investment type (Growth)
2. Morningstar overall rating
3. Stocks asset breakdown
4. ESG score
5. Fund sector (Energy)
6. Fund sector (Industrials)
7. Size fund (Large)
8. Size fund (Small)
9. Morningstar risk rating
10. Fund sector (Technology)

*This does not mean that ETF is not a worthy investment option

Overcoming limitations

Lack of quality ETF dataset

Looked at external research to understand sentiments on ETF

Practical use of prediction model

Kept prediction for short-term and remove features that are older

More info needed on the changing portfolio

Kept annual holdings turnover rate as a feature

Recommendations to maximize investment

For fund managers:

1. Re-look at current active mutual funds and optimize features
2. Consider converting active funds to index funds and measure performance
3. Become a holistic financial advisor and adapt counsel based on investor's needs

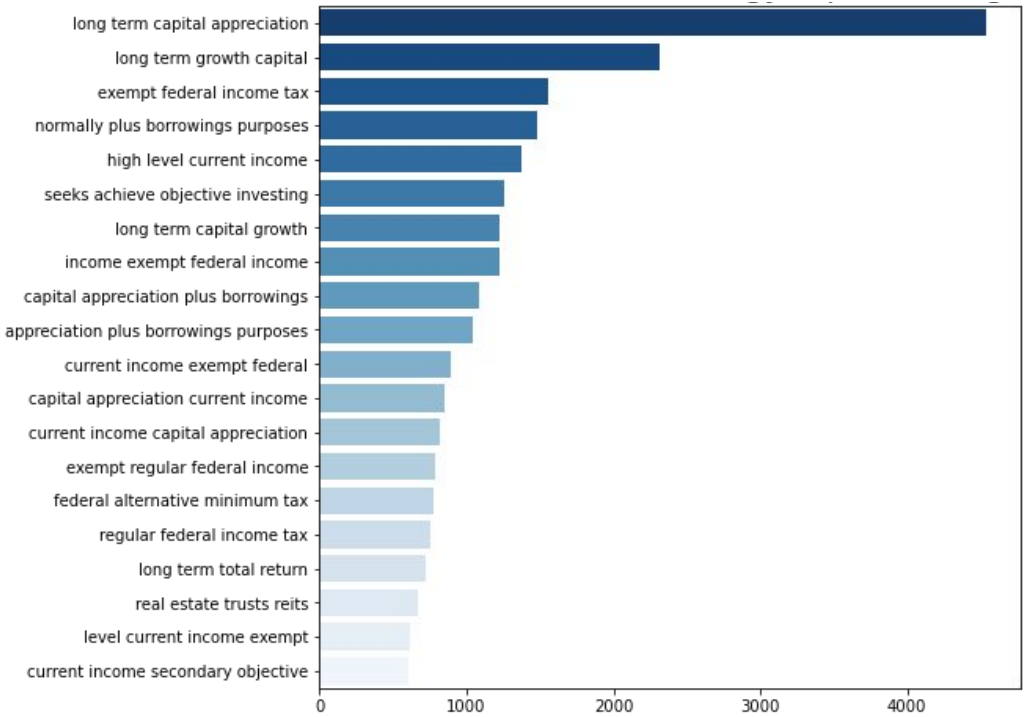
For investors:

1. Re-look at portfolio and check if features are optimal for returns
2. Prioritize investment goals, e.g. engage fund manager if want to outperform market
3. Keep up with new products and changes to make informed decisions

THANK YOU!

Questions?

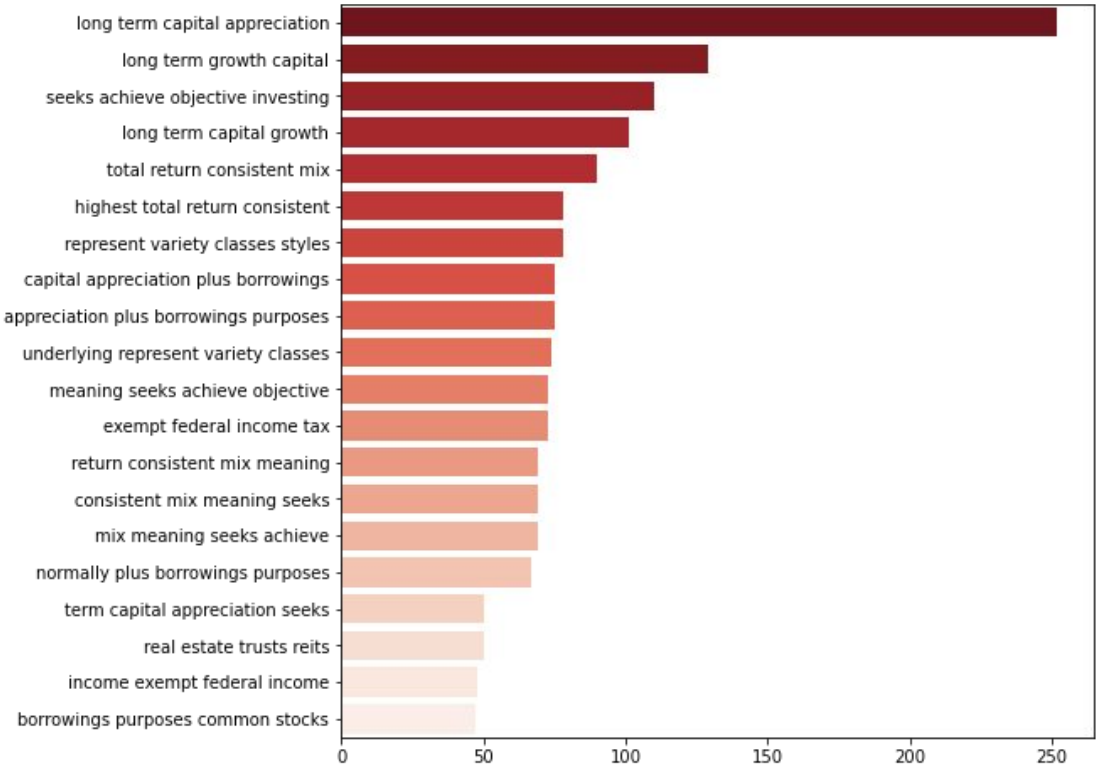
Mutual funds are focused on growth & income



Top four-grams for mutual fund investment strategy:

1. ‘Long term capital appreciation’
2. ‘Long term capital growth’
3. ‘High level current income’

ETFs are focused on growth & consistency



Top four-grams for ETF investment strategy:

1. ‘Long term capital appreciation’
2. ‘Long term capital growth’
3. ‘Total return consistent mix’

#1: Mutual fund prediction of YTD fund returns

	Model	R2 Train Set	R2 Test Set	R2 Cross-Val	RMSE Train Set	RMSE Test Set	RMSE Cross-Val
0	Linear	0.551085	0.537027	0.549006	0.053978	0.053983	0.054111
1	Lasso	0.551057	0.537056	0.548978	0.053980	0.053982	0.054113
2	Ridge	0.551079	0.537153	0.548994	0.053979	0.053976	0.054111
3	K-Nearest Neighbor	0.912874	0.844068	0.828484	0.023780	0.031329	0.033376
4	Decision Tree	0.999996	0.927732	0.878123	0.000165	0.021328	0.028963
5	Bagged Tree	0.986431	0.933892	0.914410	0.009385	0.020399	0.023835
6	Random Forest	0.991478	0.940528	0.923787	0.007437	0.019348	0.022255
7	Adaptive Boosting	0.454416	0.444243	0.406901	0.059507	0.059146	0.062197
8	Support Vector	0.636011	0.596786	0.598631	0.048605	0.050379	0.051028
9	Gradient Boosting	0.781322	0.752056	0.758162	0.037674	0.039506	0.039614

Random Forest model generated the **highest R-squared** value and l**owest RMSE**

#2: Mutual fund prediction of alpha

	Model	R2 Train Set	R2 Test Set	R2 Cross-Val	RMSE Train Set	RMSE Test Set	RMSE Cross-Val
0	Linear	0.509265	0.494115	0.505501	3.492973	3.649483	3.505888
1	Lasso	0.509245	0.494104	0.505245	3.493043	3.649525	3.506821
2	Ridge	0.509247	0.494066	0.505517	3.493035	3.649662	3.505854
3	K-Nearest Neighbor	0.924871	0.879200	0.849927	1.366706	1.783361	1.931691
4	Decision Tree	0.999999	0.884993	0.861671	0.003749	1.740070	1.870518
5	Bagged Tree	0.986900	0.929497	0.915072	0.570703	1.362410	1.437591
6	Random Forest	0.991795	0.930564	0.928253	0.451672	1.352065	1.337694
7	Adaptive Boosting	0.516632	0.533077	0.488069	3.466654	3.506129	3.544304
8	Support Vector	0.778146	0.767722	0.749632	2.348577	2.472919	2.496498
9	Gradient Boosting	0.795217	0.783886	0.773124	2.256412	2.385323	2.378591

Random Forest model generated the **highest R-squared** value and l**owest RMSE**

#3: ETF prediction of YTD fund returns

	Model	R2 Train Set	R2 Test Set	R2 Cross-Val	RMSE Train Set	RMSE Test Set	RMSE Cross-Val
0	Linear	0.268677	0.190142	0.151607	0.101456	0.134901	0.108935
1	Lasso	0.248466	0.161359	0.184742	0.102848	0.137277	0.107765
2	Ridge	0.258832	0.177874	0.199403	0.102137	0.135919	0.107059
3	K-Nearest Neighbor	0.456785	0.160804	0.194200	0.087440	0.137323	0.106185
4	Decision Tree	1.000000	-0.111882	-0.104982	0.000000	0.158067	0.128271
5	Bagged Tree	0.869881	0.174890	0.209913	0.042795	0.136165	0.100282
6	Random Forest	0.909068	0.238645	0.329086	0.035775	0.130799	0.097676
7	Adaptive Boosting	0.548978	0.124286	0.201751	0.079675	0.140279	0.104696
8	Support Vector	0.581657	0.194567	0.137529	0.076734	0.134532	0.108990
9	Gradient Boosting	0.811119	0.242856	0.271927	0.051561	0.130437	0.098872

Random Forest model generated the **highest R-squared** value and l**owest RMSE**

#4: ETF prediction of YTD alpha

	Model	R2 Train Set	R2 Test Set	R2 Cross-Val	RMSE Train Set	RMSE Test Set	RMSE Cross-Val
0	Linear	0.295833	0.172844	0.177045	4.709325	6.081976	5.065165
1	Lasso	0.279899	0.178622	0.218697	4.762306	6.060696	4.985858
2	Ridge	0.284403	0.180979	0.224202	4.747389	6.051994	4.967163
3	K-Nearest Neighbor	0.464900	0.216672	0.230635	4.105236	5.918652	5.054052
4	Decision Tree	1.000000	0.285206	-0.212497	0.000000	5.653811	5.827413
5	Bagged Tree	0.864997	0.224446	0.266490	2.062014	5.889209	5.088017
6	Random Forest	0.909565	0.259708	0.320265	1.687680	5.753771	4.727968
7	Adaptive Boosting	0.431418	0.097794	0.030571	4.231724	6.351902	5.288922
8	Support Vector	0.239499	0.128783	0.188329	4.894074	6.241861	5.107508
9	Gradient Boosting	0.800791	0.277257	0.247990	2.504812	5.685163	4.874230

Random Forest model generated the **highest R-squared** value and l**owest RMSE**