Rich and Cheap Bond Recommendation

Capstone Final presentation

May 9th, 2019
Columbia Vanguard Team

Columbia Team Vanguard

Jeff Khasin

Background Research / Domain Knowledge
EDA
JPM literature review

Addison Li

EDA
Bond Prediction
All Curve Fitting models
Backtesting

Steve McClain

Developed / evaluated recommendation models
Integrated user feedback into UI

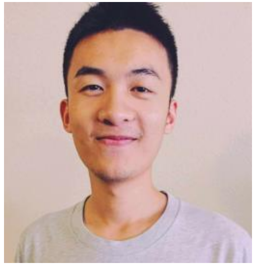
Naoto Minakawa

UI development
Implementation of interactive charts
Integration of recommendation models

Yilin Sun

EDA
Bond Prediction
Forward Shock Model
Backtesting
Model evaluation

Hangyu Zhou



Bond Prediction
Linear Mixed Effect Model
Model evaluation
Integration of Prediction Models into UI

Agenda

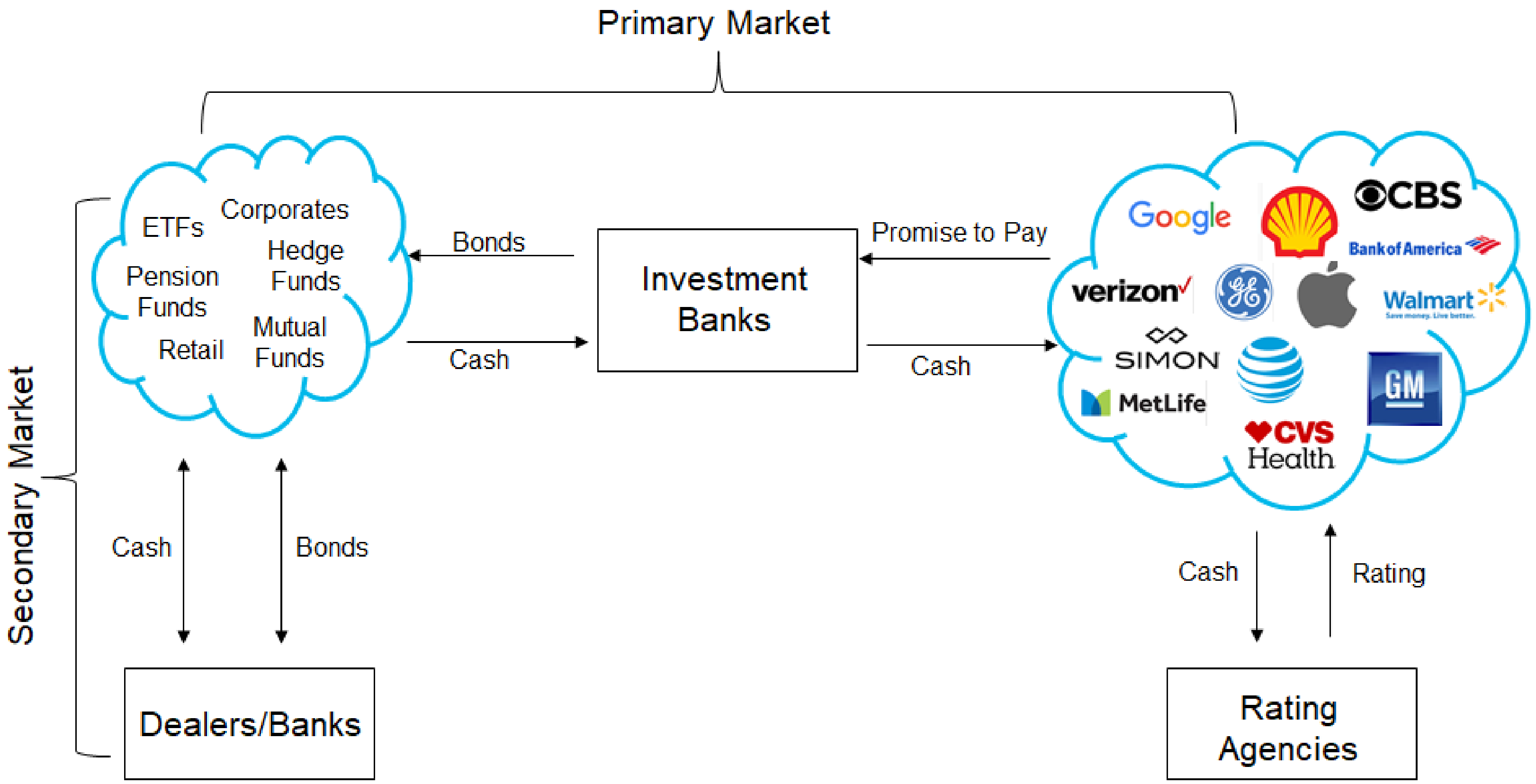
- Review problem statement
- Introduce background knowledge
- Share modeling methodology
- Discuss key business insights
- Examine future engineering / next steps

What are we solving for?

Portfolio Managers' dilemma:

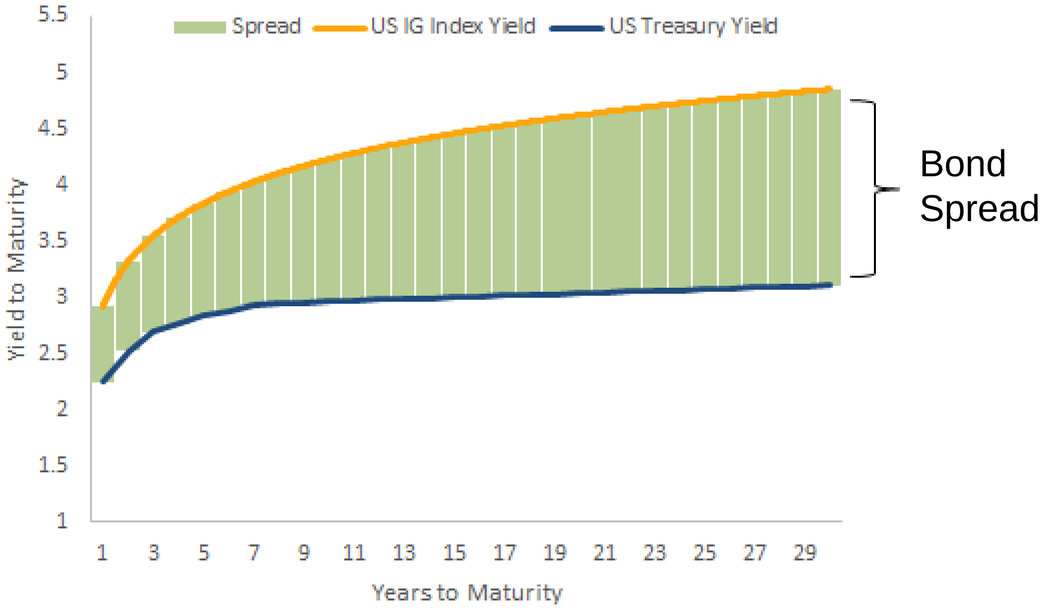
What is the next best bond to buy/short, when my first choice is unavailable?

Crash course: Bonds Market Structure

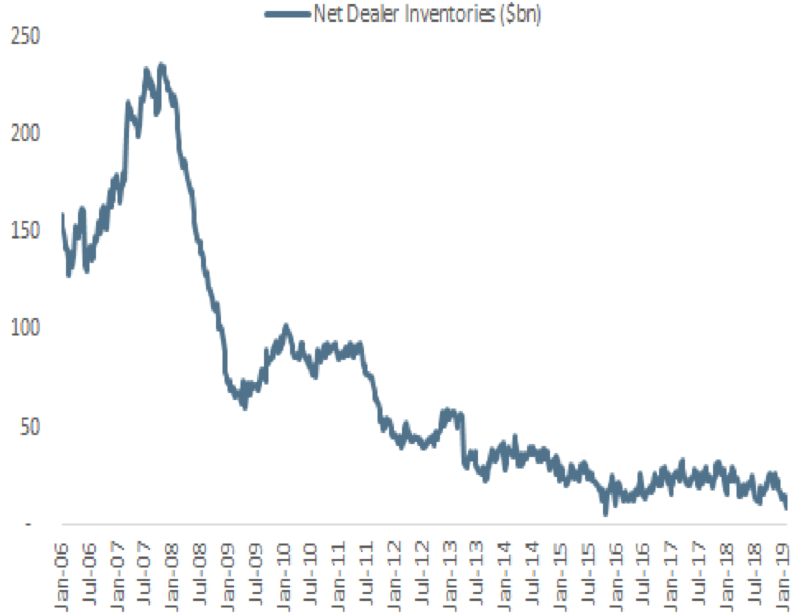


Data set: Important Bond Concepts and Features

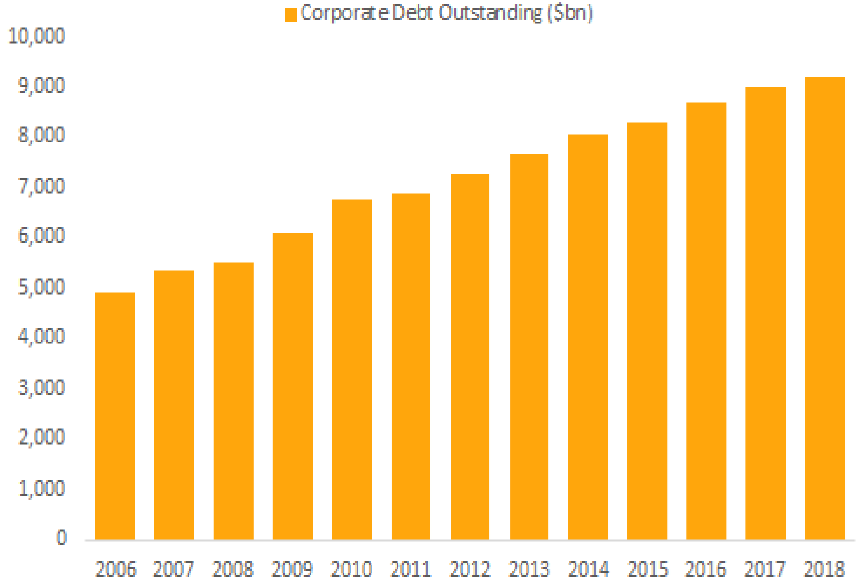
- Years to Maturity
- Coupon
- Yield and Spread
- Risk premium
- Duration
- Relative value



Crash course: Financial Crisis and Deteriorating Liquidity



Source: Federal Reserve Bank of New York



Source: SIFMA

What if ...

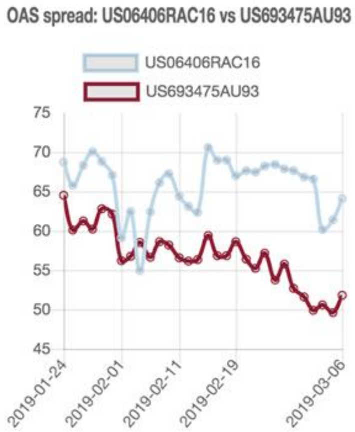
- When a bond is unavailable, we can provide a list of alternatives that match the portfolio manager's need?

- We can proactively identify which bonds are rich or cheap on the trading day?

Our solution enables traders without coding experience to get bonds recommendations easily

Searching for bonds that are similar to US06406RAC16:

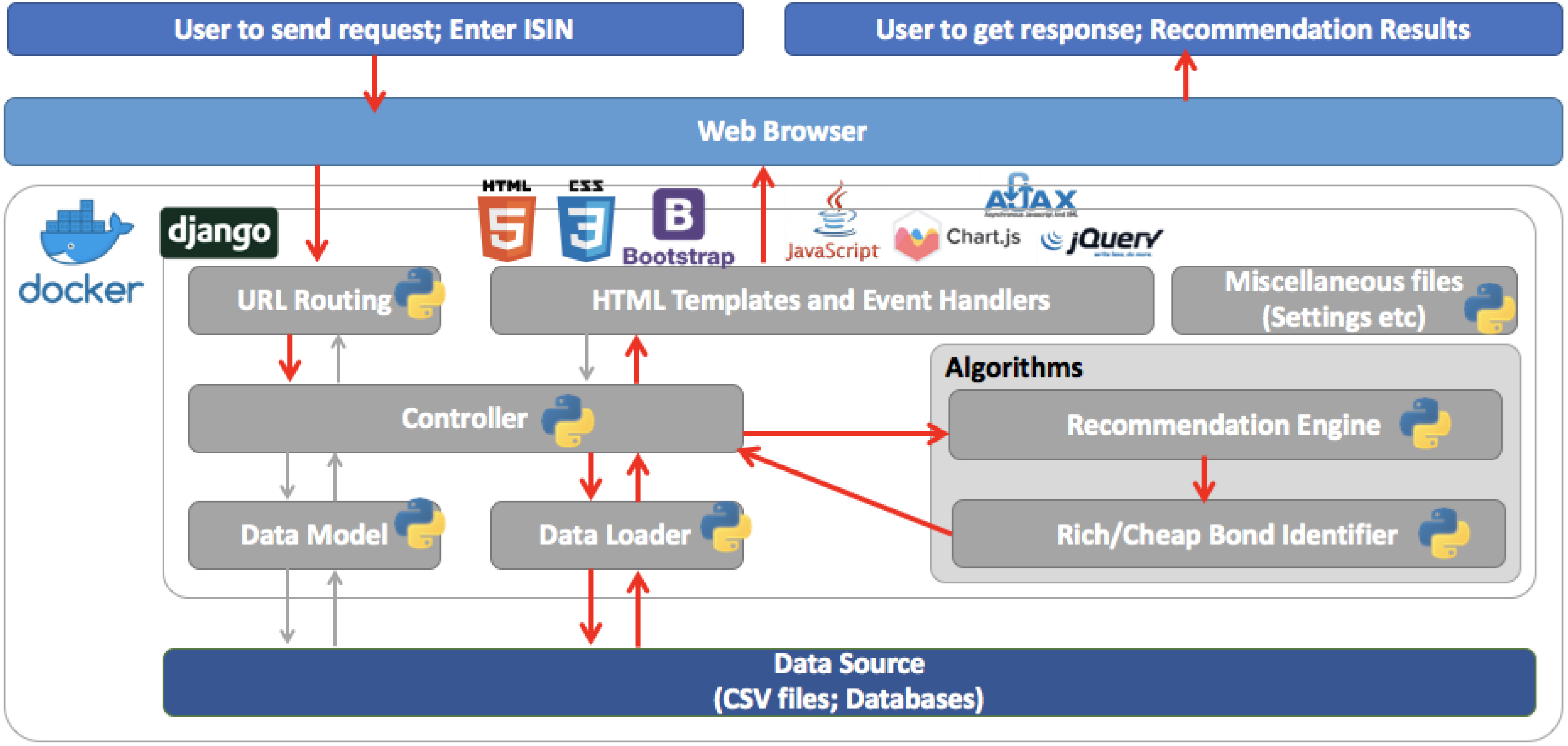
ISIN	Ticker	BCLASS3	Country	OAS	OAD	KRD 5Y	KRD 10Y	KRD 20Y	KRD 30Y	Yield to Mat	Cpn	Px Close	rich/cheap
US06406RAC16	BK	Banking	United States	59.99	3.01	1.04	0.0	0.0	0.0	3.04	2.661	98.84	



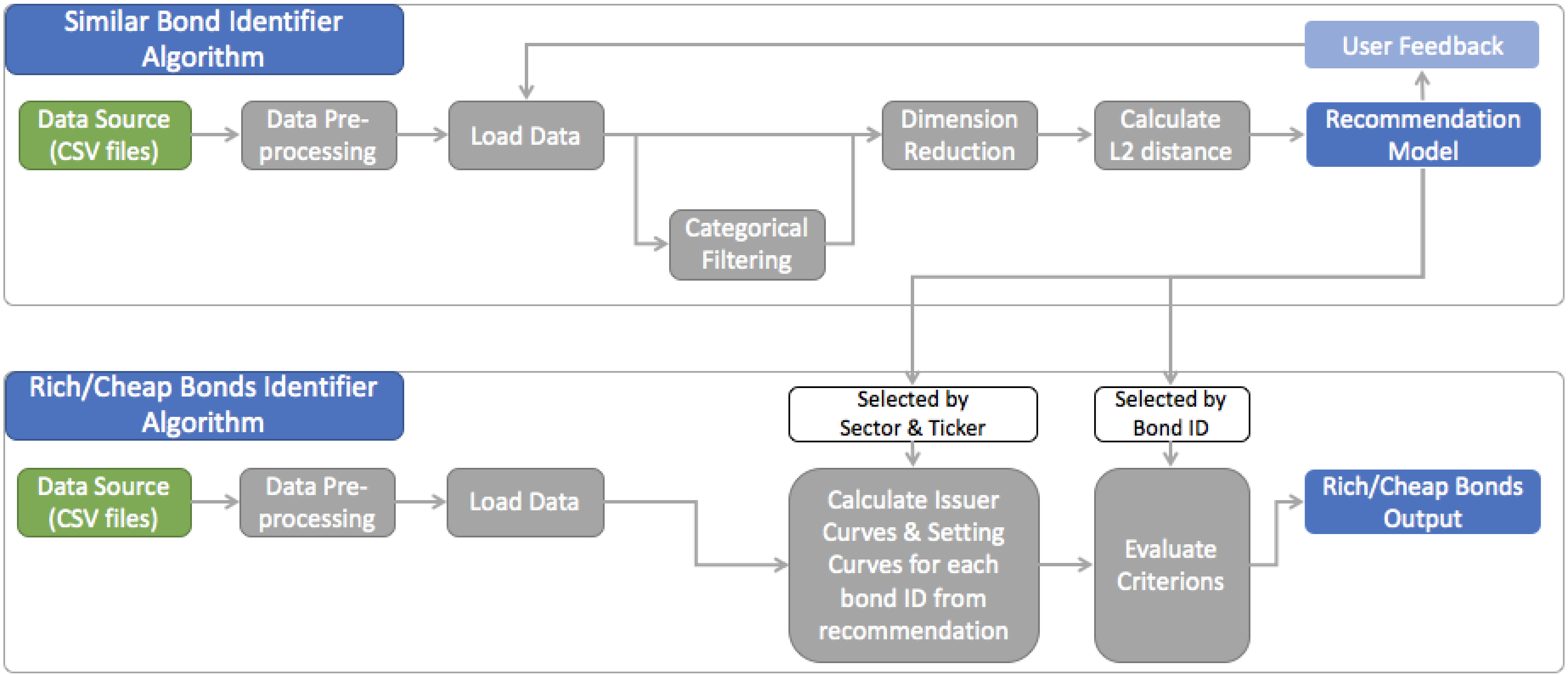
The bonds most similar to US06406RAC16 are:

ISIN	Ticker	BCLASS3	Country	OAS	OAD	KRD 5Y	KRD 10Y	KRD 20Y	KRD 30Y	Yield to Mat	Cpn	Px Close	rich/cheap	Feedback
US05531FAX15	BBT	Banking	United States	47.24	2.85	0.98	0.0	0.0	0.0	2.93	2.75	99.48	neither	
US693475AU93	PNC	Banking	United States	52.72	2.33	0.42	0.0	0.0	0.0	3.01	3.25	100.57	neither	Up
US06406RAC16	BK	Banking	United States	59.99	3.01	1.04	0.0	0.0	0.0	3.04	2.661	98.84	neither	Up

High-level System Architecture



High-level Algorithms Detail



Domain Knowledge + Unsupervised Learning

Categorical Filtering → **Dimensionality Reduction** → **Euclidean Distance Metric**

Categorical Filtering

Filter for bonds that match on key characteristics

Dimensionality Reduction

Mitigate “curse of dimensionality”

Reduce impact of highly-correlated features

Euclidean Distance Metric

Recommendations are the closest bonds in the vector space

Bootstrapping a feedback-based recommendation model

Categorical Group 1
(Insurance, United States, AA)

Bond	Distance to Bond A
A	0
B	1.5
C	2
…	…
D	12

Categorical Group 2
(Natural Gas, China, AAA)

Bond	Distance to Bond A
E	?
F	?

Bootstrapping a feedback-based recommendation model

Categorical Group 1
(Insurance, United States, AA)

Bond	Distance to Bond A
A	0
B	1.5
C	2
…	…
D	12

Feedback Training Set

Target Bond	Better Rec.	Worse Rec.
A	B	C

Categorical Group 2
(Natural Gas, China, AAA)

Bond	Distance to Bond A
E	?
F	?

Bootstrapping a feedback-based recommendation model

Categorical Group 1
(Insurance, United States, AA)

Bond	Distance to Bond A
A	0
B	1.5
C	2
…	…
D	12

Categorical Group 2
(Natural Gas, China, AAA)

Bond	Distance to Bond A
E	?
F	?

Feedback Training Set

Target Bond	Better Rec.	Worse Rec.
A	B	C
A	B	D
A	C	D
A	D	E
A	D	F

All bonds in Group 1 are closer to Bond A than *any* bond in Group 2

Bayesian Personalized Recommendation

- “Embed” each bond as a vector in latent space
- Measure bond similarity by computing dot products (higher is better)

Bond	Latent Vector
A	[0.20, 0.05, -0.25, ...]
B	[0.60, 0.15, -1.52, ...]
F	[-1.56, -2.71, 0.34, ...]

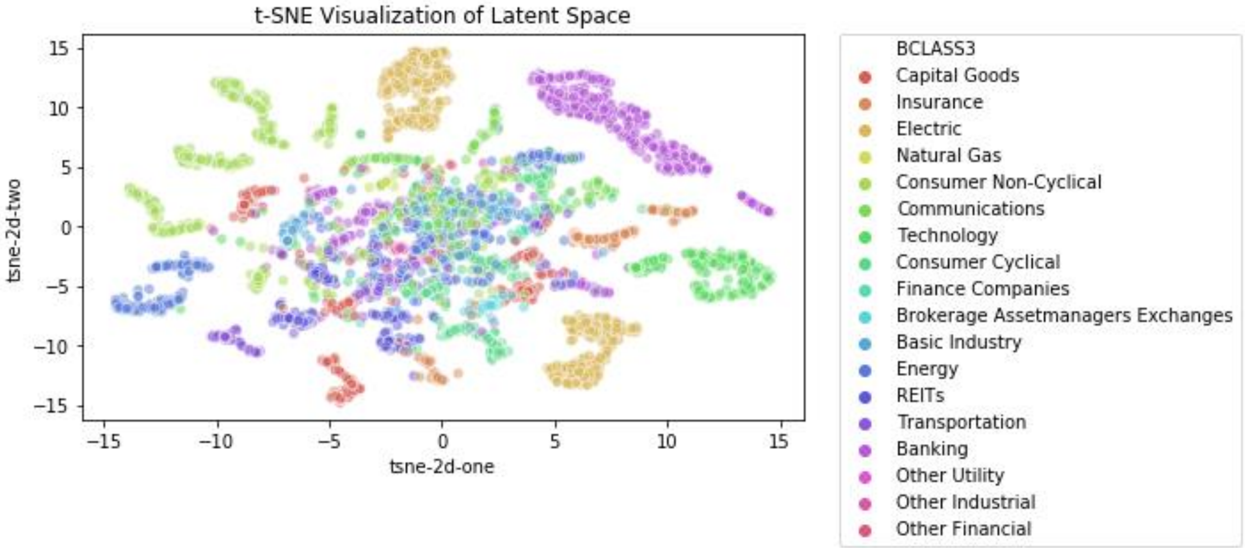
$$Simi(i, j) = \mathrm{i} \cdot \mathrm{j}$$

	Similarity Matrix						
	A	B	C	D	E	F	
A		1.81				-2.1	
B							
C							
D							
E							
F							

Learning problem: Find the embedding that maximizes the difference in the dot products between the target bond and the “better”/ “worse” recommendations given in the training set, subject to L2 regularization

$$\sum_{(u,i,j)\in B_s} ln\, \sigma(\hat{x}_{uij}) - \lambda_\Theta ||\Theta||^2$$

Model Validation



Qualitative Validation

The latent space recovered the concept of market sector

Latent Dimension	Area Under Curve (AUC)
8	0.9544
16	0.9645
32	0.9600

Quantitative Validation

The model can correctly order pairs of bonds from a hold-out set of rankings

Original filtering process:

All Bonds

Liquidity

G-spread

Issuer Curve Or / and Rating Curve

Our three filtering criteria:

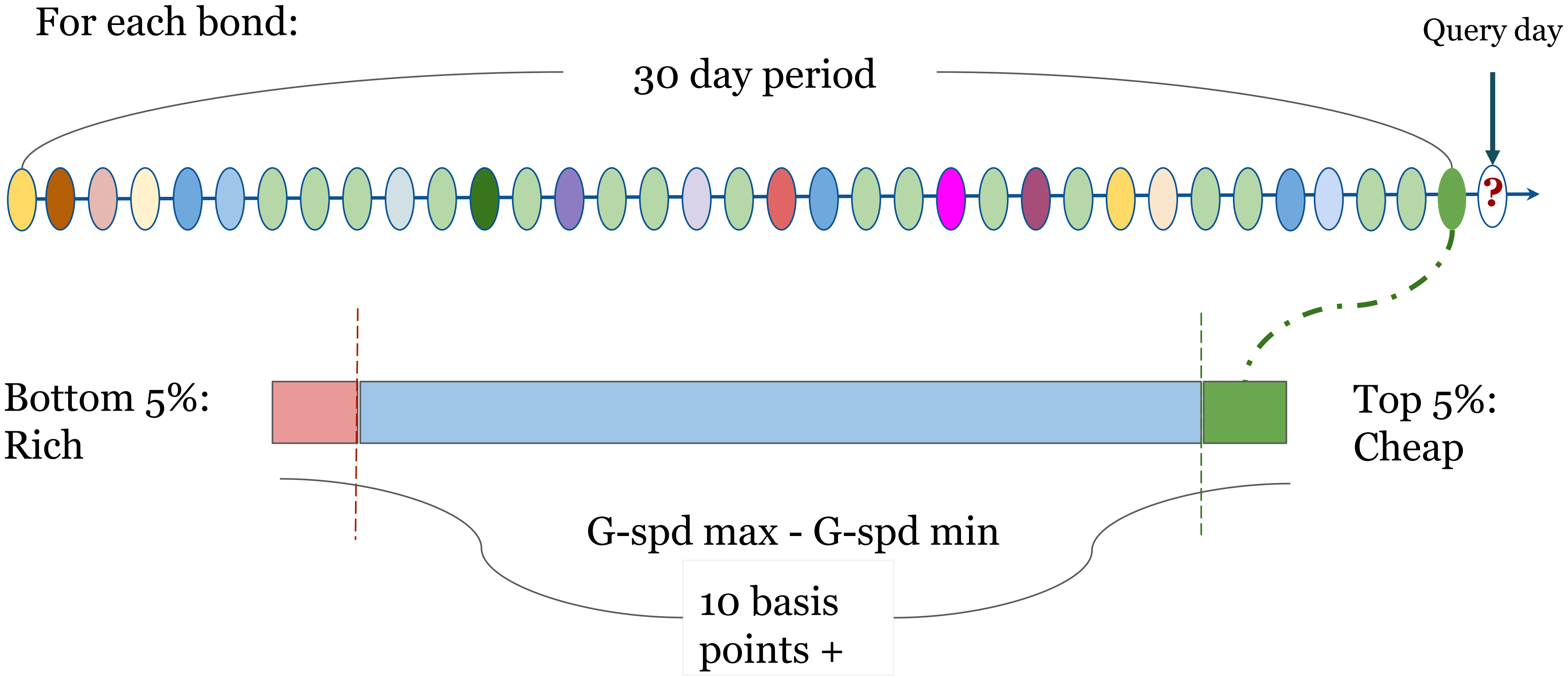
All Bonds

~~Liquidity~~

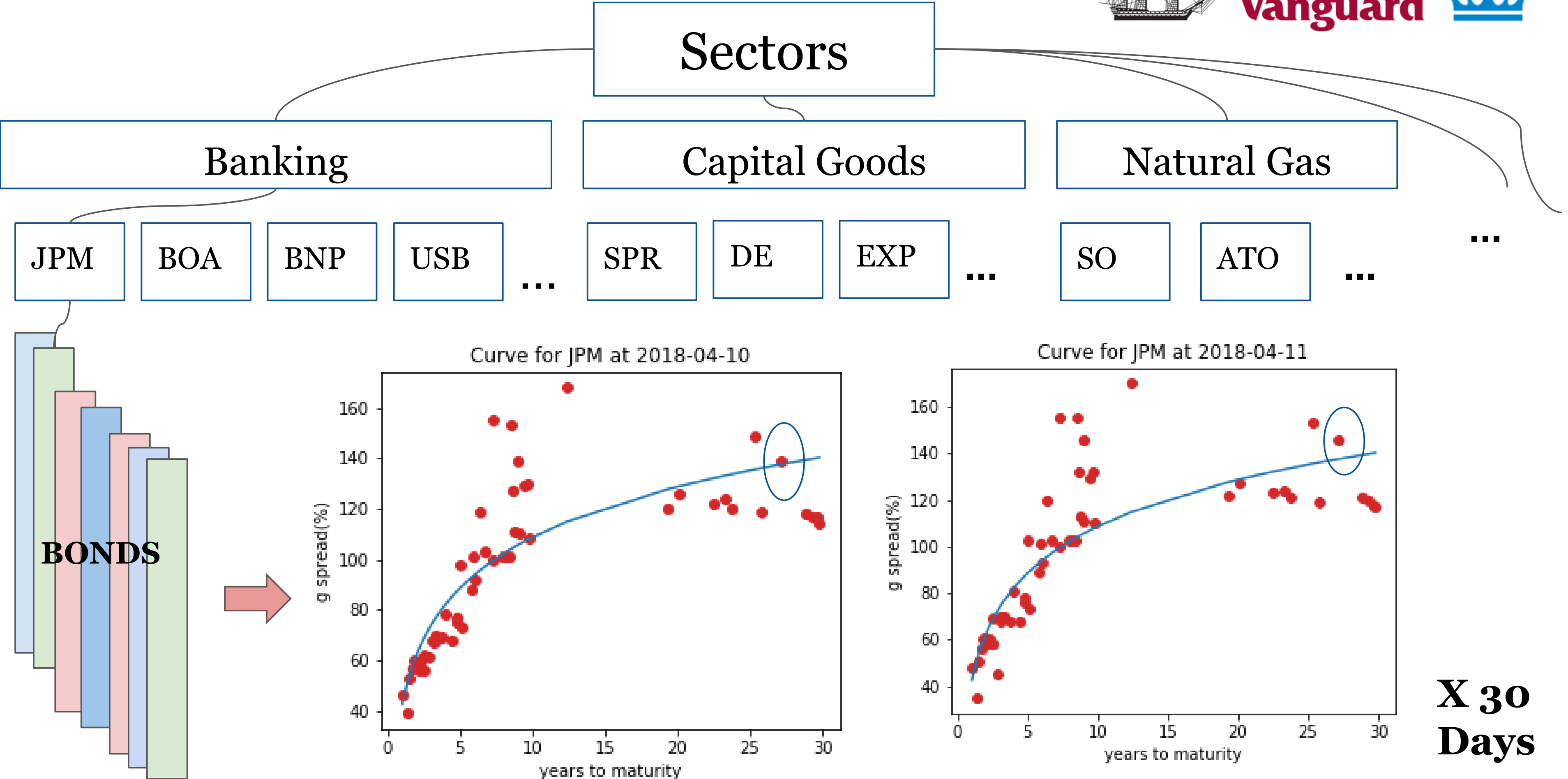
G-spread

Issuer Curve Or / and Rating Curve

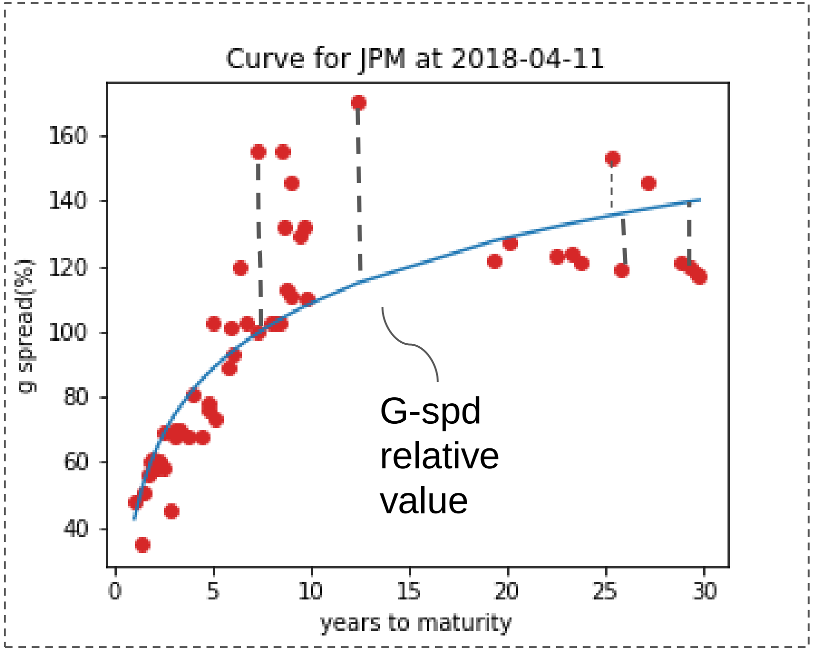
Bond prediction - G-spread



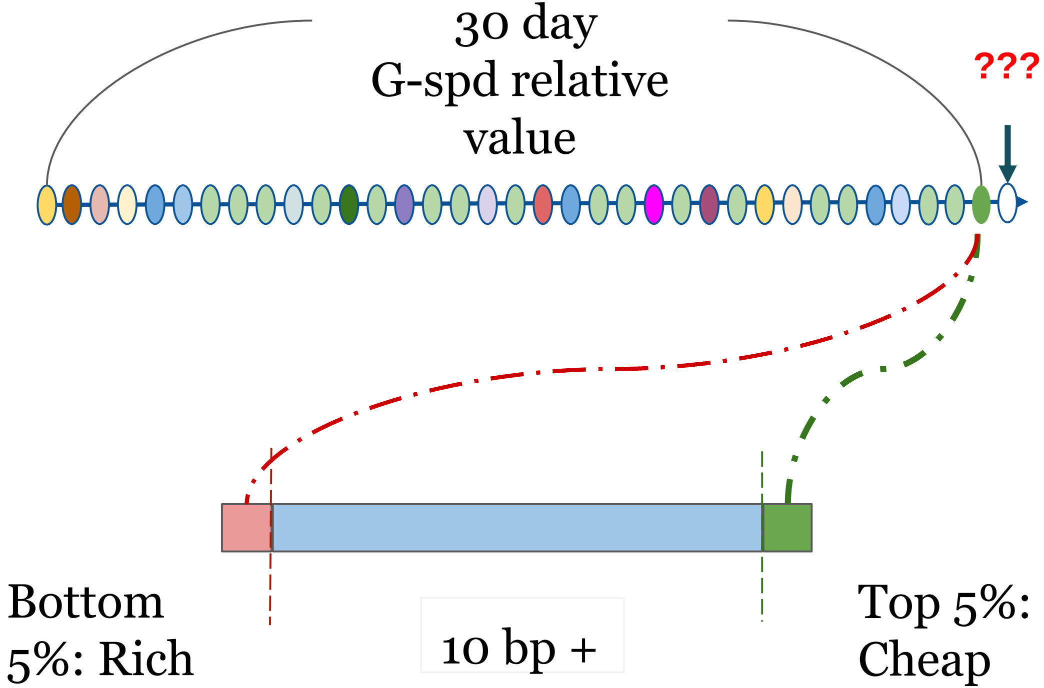
Bond Prediction - Issuer Curve Fitting



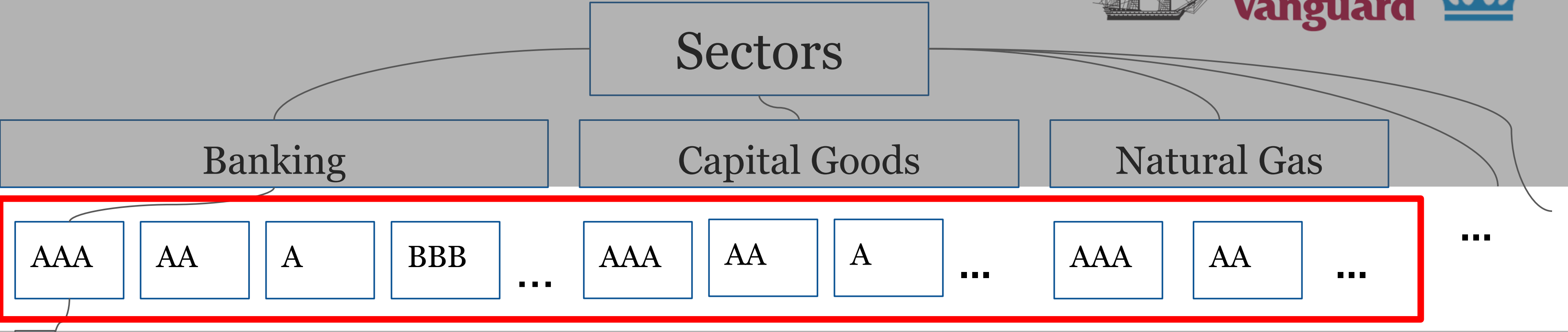
Bond prediction - after Curve fitting

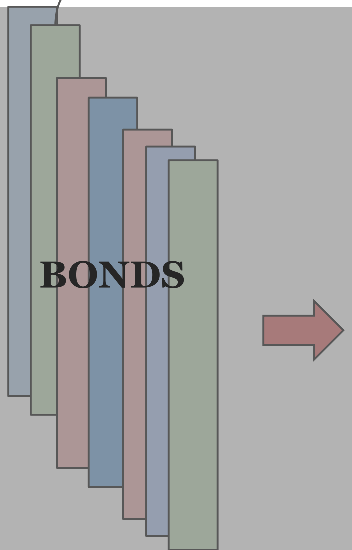


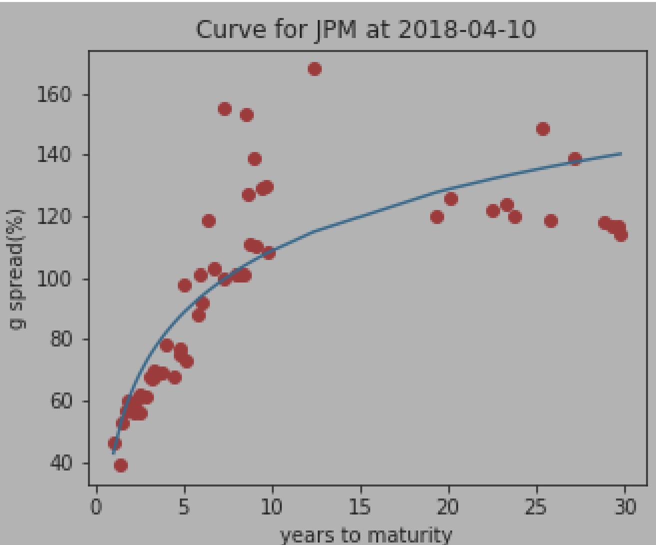
x 30 DAYS

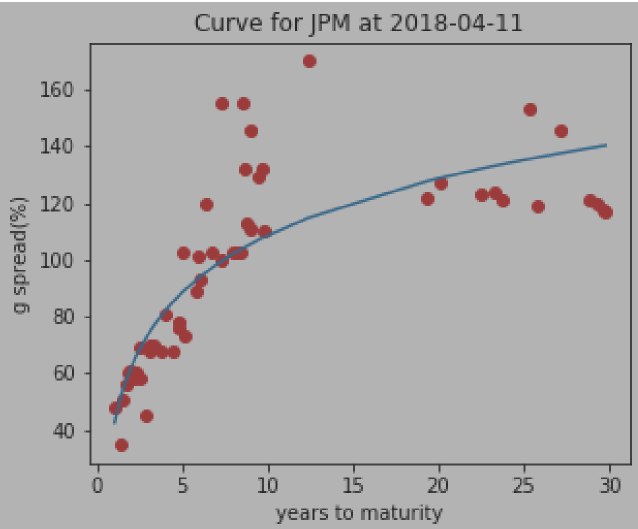


Bond Prediction - Rating Curve Fitting









X 30 DAYS

Bond Prediction - Curve Models

- Logarithmic Model:

$$Y = \beta_1 \log x + \beta_0 + \epsilon$$

- Forward Shock Model:

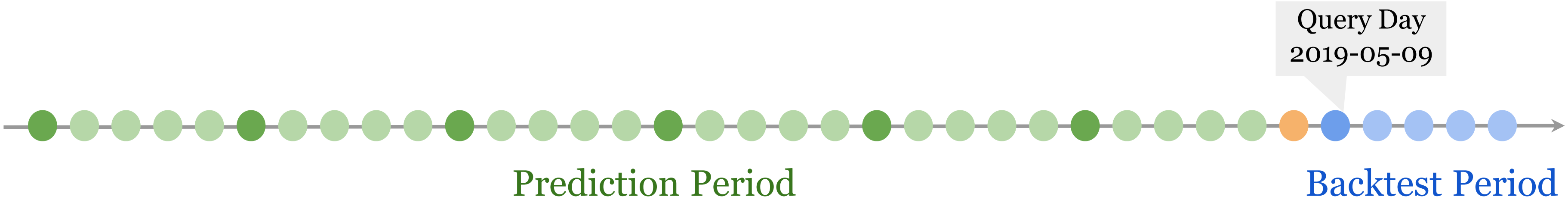
$$Y = \beta_2(1 - e^{-0.2x}) + \beta_1(1 - e^{-0.05x}) + \beta_0 + \epsilon$$

- Linear Mixed Effect Model:

$$\begin{aligned} Y =& \beta_2(1 - e^{-0.2x}) + \beta_1(1 - e^{-0.05x}) + \beta_0 + \\ & \gamma_2(1 - e^{-0.2x}) + \gamma_1(1 - e^{-0.05x}) + \gamma_0 + \epsilon \end{aligned}$$

Backtesting and Model Evaluation

Backtesting is a method for evaluating how well a model or strategy would have performed on historical data.



Key Points

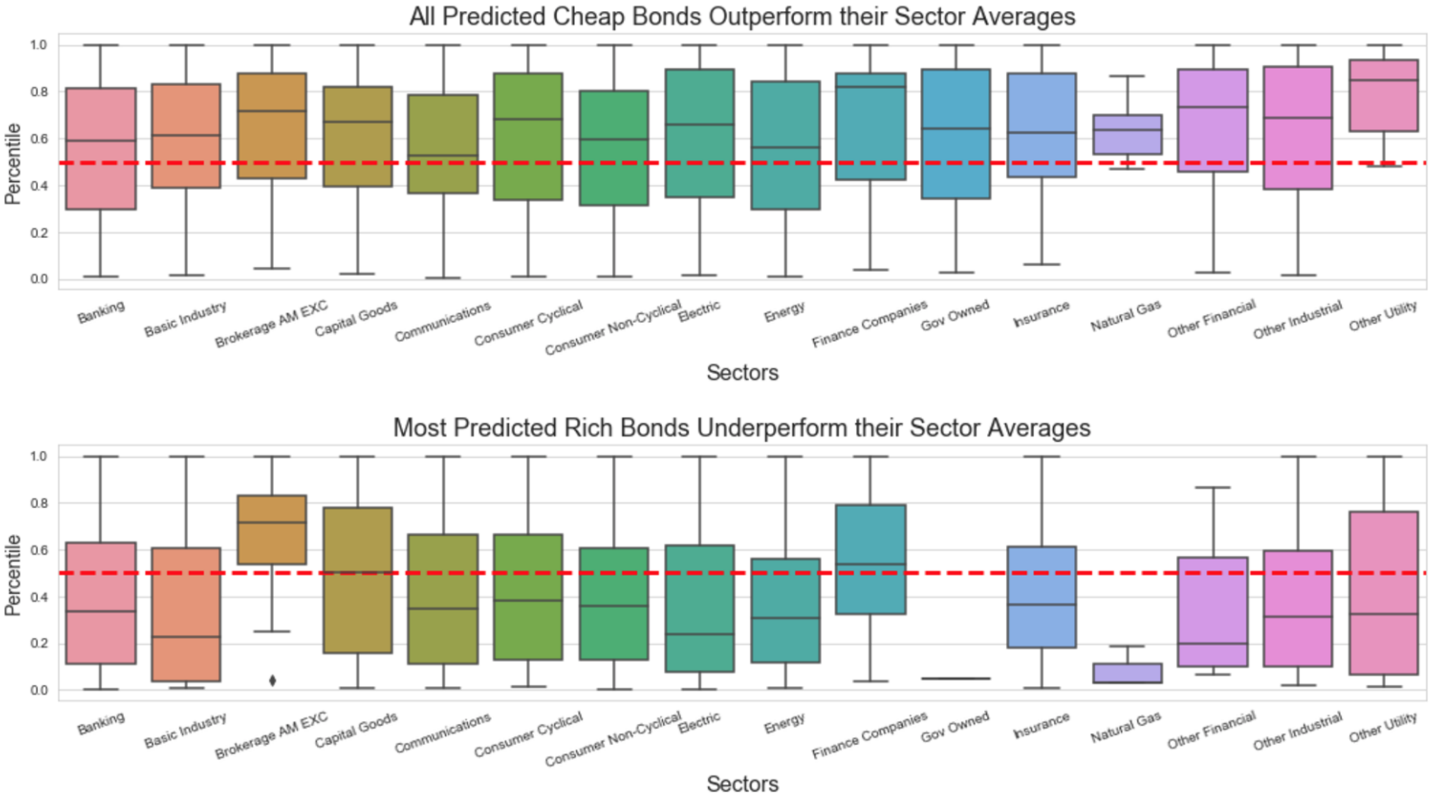
Metric duration-adjusted cumulative excess return (not price return or simple cumulative excess return)

Assumption a one-week bond holding period (one-week equals to five trading days)

Data daily excess return and one-week average duration

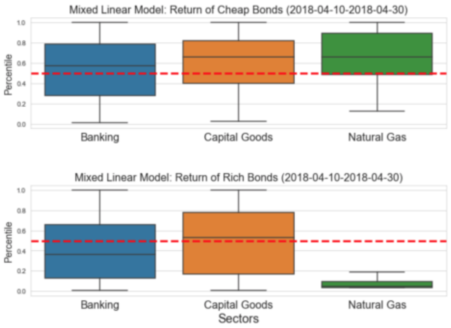
$$Return = \frac{r_1 + r_2 + r_3 + r_4 + r_5}{avg(duration)}, \ where\ r_i\ is\ daily\ excess\ return$$

Better Performance on Cheap Bonds; Natural Gas sector performed the best

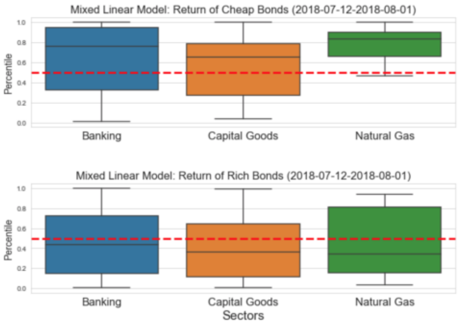


Prediction influenced by seasonality

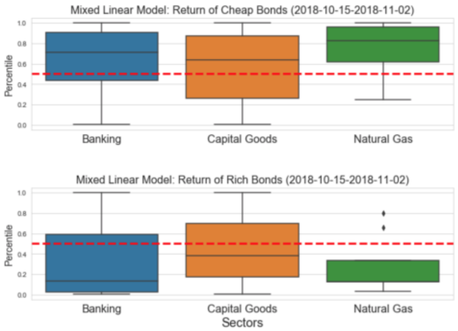
Columbia University
Vanguard®



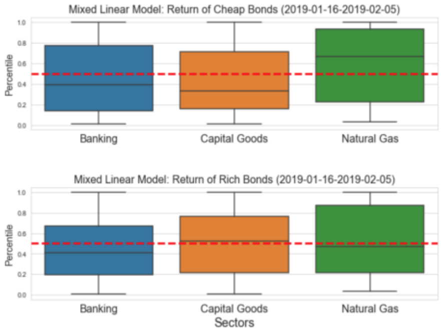
Period 1
2018-04-10 to 2018-04-30



Period 2
2018-07-12 to 2018-08-01



Period 3
2018-10-15 to 2018-11-02



Period 4
2019-01-16 to 2019-02-05

It appears to be a seasonal influence that affect our prediction accuracy.

Forward Shock Model gives a better prediction.

Logarithmic Model

	Outperform	Underperform
Predicted Cheap	62.01%	37.99%
Predicted Rich	41.39%	58.61%
	Actual cheap bonds	**Actual rich bonds**

Forward Shock Model

	Outperform	Underperform
Predicted Cheap	66.30%	33.70%
Predicted Rich	41.12%	58.88%

Linear Mixed Effect Model

	Outperform	Underperform
Predicted Cheap	62.05%	37.95%
Predicted Rich	41.53%	58.47%

Logarithmic Model	Forward Shock Model	Linear Mixed Effect Model
Average Number of Cheap Bonds per Day		
30	18	33
Average Number of Rich Bonds per Day		
82	81	91

Future steps

Recommendation:

- Investigate other matrix factorization approaches that scale better to large datasets (e.g. Hierarchical Poisson Factorization)

Prediction:

- Incorporate liquidity data
- Back testing: back testing full year, or multi- year data

UI:

- Allow a new bond that is not in our system yet to acquire a similarity score
- Allow user feedback to interact with the recommendation engine dynamically

Thank you!

Questions ? Fire away!