QIML Insight：基于多源特征及机器学习的股票聚类模型

Original 全网Quant都在看 量化投资与机器学习 2022-05-24 15:58 Posted on 上海

收录于合集

#深度研读系列 21个 >

量化投资与机器学习微信公众号，是业内垂直于**量化投资、对冲基金、Fintech、人工智能、大数据**等领域的主流自媒体。公众号拥有来自**公募、私募、券商、期货、银行、保险、高校**等行业**30W+**关注者，荣获2021年度AMMA优秀品牌力、优秀洞察力大奖，连续2年被腾讯云+社区评选为“年度最佳作者”。

量化投资与机器学习公众号 **独家解读**

量化投资与机器学公众号 ***QIML Insight——深度研读系列*** 是公众号全力打造的一档**深度、前沿、高水准**栏目。

公众号**遴选**了各大期刊前沿论文，按照理解和提炼的方式为读者呈现每篇论文最精华的部分。QIML希望大家能够读到可以成长的量化文章，愿与你共同进步！

> **本期遴选论文**
>
> **来源：**The Journal of Financial Data Science Spring 2022
>
> **标题：**An Artificial Intelligence-Based Industry Peer Grouping System
>
> **作者：**George Bonne，Andrew W. Lo，Abilash Prabhakaran，Kien Wei Siah，Manish Singh，Xinxin Wang，Peter Zangari，Howard Zhang

核心观点

- 本文提出了一种基于数据驱动的行业分类方法，该方法以不同的粒度级别将类似的公司聚集在一起；
- 机器学习的技术可以从相关数据源中提取特征，并学习相关关系，从而识别出在样本外时期风险回报情况相似的公司。
- 历史收益相关性、GICS分类、10-K报告、规模、动量、资产负债率等基本因子对企业相似性的预测贡献最大。

行业分类体系在投资组合构建中有着非常广泛的应用，一个好的行业分类体系有以下两个特点：最小化组内股票的差距和最大化的组间股票区别。构建投资组合时，投资者往往通过分散行业配置来达到组合风险分散化的效果。但这种基于公司业务的分类体系，相对比较固定，在多变的市场环境及多样的市场观念下，很多时候属于同一行业的股票之间的并没有很高的相关性，反而不能行业的股票却有着较大相关性。这种情形下，投资组合在行业上的分散化效果就会大打折扣。

本文提出了一种数据驱动的，基于多维度的特征对股票进行行业聚类的方法。具体来说，是使用机器学习的模型对股票多维特征与未来相关性进行建模，从而建立一个可以预测未来股票间相关性的模型。相对传统行业分类体系，该方法能够构建更加动态有效的股票分类体系，及时反应市场最新的信息。

模型

我们先整体介绍下模型的框架，首先我们**预测的目标是两个股票未来收益率的相关系数**，所使用的特征是如表2所示的股票各维度的特征，其中包括以下几类：

- 股票过去252日的日度收益

- 股票过去12个月的月度收益

- 基本面因子：主要使用MSCI Barra US Total Market Model的描述因子（详见附录）

- 对10-K报告使用NLP算法提取的因子，包括TF-IDF和Doc2Vec两大类。其中TF-IDF又使用的SVD分解降到了100维；Doc2Vec的具体方法可以参考：
    - *https://markroxor.github.io/gensim/static/notebooks/doc2vec-wikipedia.html*

- 使用新闻共现矩阵提取的节点表征，具体就是对新闻共现的股票的邻接矩阵使用Node2Vec算法得到的每个股票对应的一个多维向量

- 原始的GICS行业分类

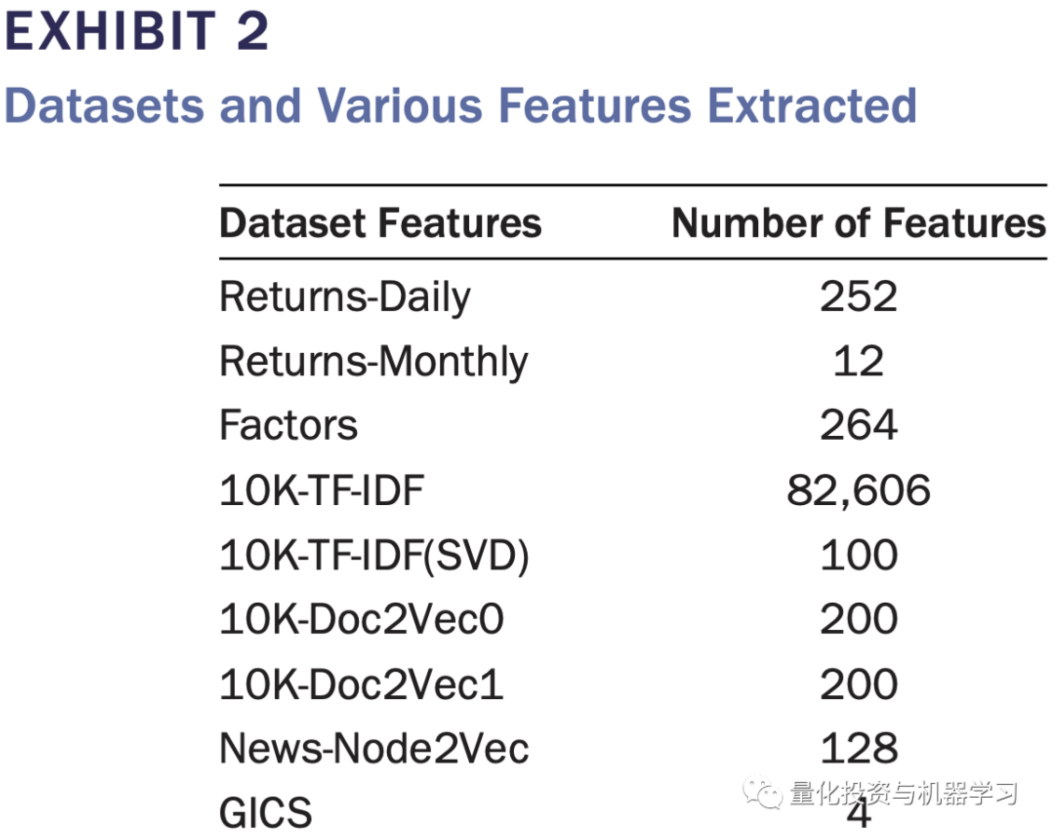


EXHIBIT 2

Datasets and Various Features Extracted

Dataset Features	Number of Features
Returns-Daily	252
Returns-Monthly	12
Factors	264
10K-TF-IDF	82,606
10K-TF-IDF(SVD)	100
10K-Doc2Vec0	200
10K-Doc2Vec1	200
News-Node2Vec	128
GICS	4

以上的特征并不直接作为模型输入的特征使用，因为预测的是任意两个股票未来收益率之间的相关性，所以准备训练数据时需要对任意两个股票的特征做如下处理：

- 计算两个股票每类特征的马氏距离（L1 Distance）

- 计算两个股票每类特征的cosine相似度

- 最终的特征为所有类别特征的马氏距离和cosine相似度构成的向量

> 比如使用了Returns-Daily、Returns-Monthly、Factors三类特征，那模型的输入就是以下6维向量：
>
> [两个股票Returns-Daily的马氏距离，
> 两个股票Returns-Daily的Cosine相似度，
> 两个股票Returns-Monthly的马氏距离，
> 两个股票Returns-Monthly的Cosine相似度，
> 两个股票Factors的马氏距离，
> 两个股票Factors的Cosine相似度]

比如我们使用2019年的某股票池中所有股票的特征数据，按照以上方法，可以加工出该股票池中任意两个股票对，如股票i和j的特征，然后再使用2020年的日度收益率计算任意两个股票对，如股票i和j的相关系数。这样我们就可以进行模型的训练，训练出来的模型就可以使用2020年的股票特征数据，预测它们之间任意两个股票在2021年收益率的相关系数。在模型的选择上，作者使用了**ridge regression、neural networks和XGBoos**t三种模型。

接着上面的例子，我们得到了某股票池2021年的预测的相关系数矩阵后，可以使用该预测的相关系数矩阵进行层次化聚类，从而生成动态的股票分类体系。**关于层次化聚类的层数及每个层次的聚类个数可以对齐传统的行业分类，比如GICS，这样也能方便我们对比该聚类方法与GICS行业分类体系。**

层次化聚类的具体实施可以使用scikit-learn中的实现：

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html#sklearn.cluster.AgglomerativeClustering

实证结果分析

作者对照GICS的前三级sector、industries及subindustries，使用了上述方法对股票进行了层次化聚类。也就是说层次化聚类时也分成了三个级别，每个级别中对应的聚类的数量与GICS对应，比如第一层聚类数量与GICS的sector的数量一致，也就是11个。

下表3和表4是股票聚类效果的对比，每一行表示不同的模型与特征集的组合在不同颗粒度下聚类的效果，如"Ridge:Factors"表示使用Ridge模型与Factors特征集的聚类效果，表中指标的意思表示该层次聚类下所有股票的平均相关性（使用样本外的收益率计算的相关系数）。

> 如最后一行XGBoost:ALL+GICS，Sector列的指标值是36.58，表示：使用XGBoost模型与所有特征数据进行聚类后，在Sector这个层聚类中，首先对每个聚类中的每个股票计算其与聚类中其他股票相关系数的均值，记为 $\bar{\rho}_i^k$；然后再计算该聚类中每个股票的 $\bar{\rho}_i^k$ 的均值得到 $\bar{\rho}^k$；最后计算所有11个sector的 $\bar{\rho}^k$ 的均值，即最后的指标值36.58

表3和表4的区别在于，表3中股票的数量取决于GICS行业分类体系中股票的数量，比如在GISC行业分类体系下，A股票在某个Sector中有N个同属这个Sector的股票；而在XGBoost:ALL+GICS的聚类算法下，A股票在某个Sector中有M个同属这个Sector的股票。那么在计算指标时使用的股票数量是M与N的最小值，也就是说与。而表4中，是使用全部M个股票。

因为表3中使用的股票数量为对应GICS分类中的股票数量，所以主要目的是与GICS进行对比，**我们可以看到XGBoost:ALL+GICS的聚类的相关性最高，效果最好。而参考表4，当使用各自聚类下所有股票数量时，在Sector和Industry层级，Ridge：ALL+GICS的聚类效果最好；XGBoost:ALL+GICS在Subindustries层级的聚类效果更好。**

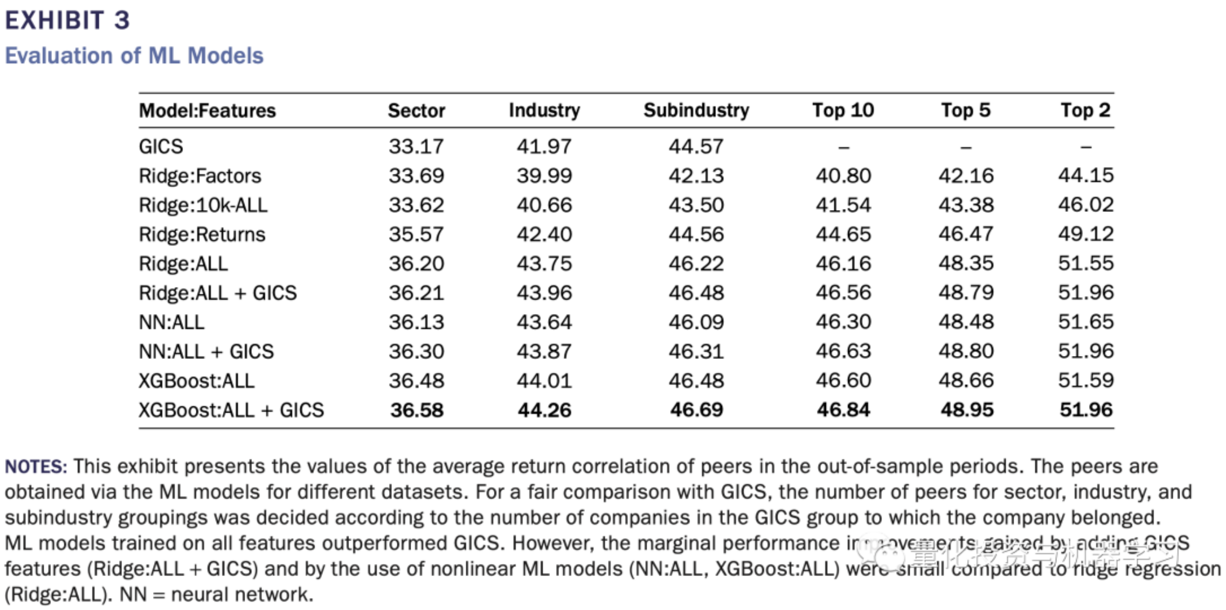


EXHIBIT 3

Evaluation of ML Models

Model:Features	Sector	Industry	Subindustry	Top 10	Top 5	Top 2
GICS	33.17	41.97	44.57	–	–	–
Ridge:Factors	33.69	39.99	42.13	40.80	42.16	44.15
Ridge:10k-ALL	33.62	40.66	43.50	41.54	43.38	46.02
Ridge:Returns	35.57	42.40	44.56	44.65	46.47	49.12
Ridge:ALL	36.20	43.75	46.22	46.16	48.35	51.55
Ridge:ALL + GICS	36.21	43.96	46.48	46.56	48.79	51.96
NN:ALL	36.13	43.64	46.09	46.30	48.48	51.65
NN:ALL + GICS	36.30	43.87	46.31	46.63	48.80	51.96
XGBoost:ALL	36.48	44.01	46.48	46.60	48.66	51.59
XGBoost:ALL + GICS	**36.58**	**44.26**	**46.69**	**46.84**	**48.95**	**51.96**

NOTES: This exhibit presents the values of the average return correlation of peers in the out-of-sample periods. The peers are obtained via the ML models for different datasets. For a fair comparison with GICS, the number of peers for sector, industry, and subindustry groupings was decided according to the number of companies in the GICS group to which the company belonged. ML models trained on all features outperformed GICS. However, the marginal performance improvements gained by adding GICS features (Ridge:ALL + GICS) and by the use of nonlinear ML models (NN:ALL, XGBoost:ALL) were small compared to ridge regression (Ridge:ALL). NN = neural network.

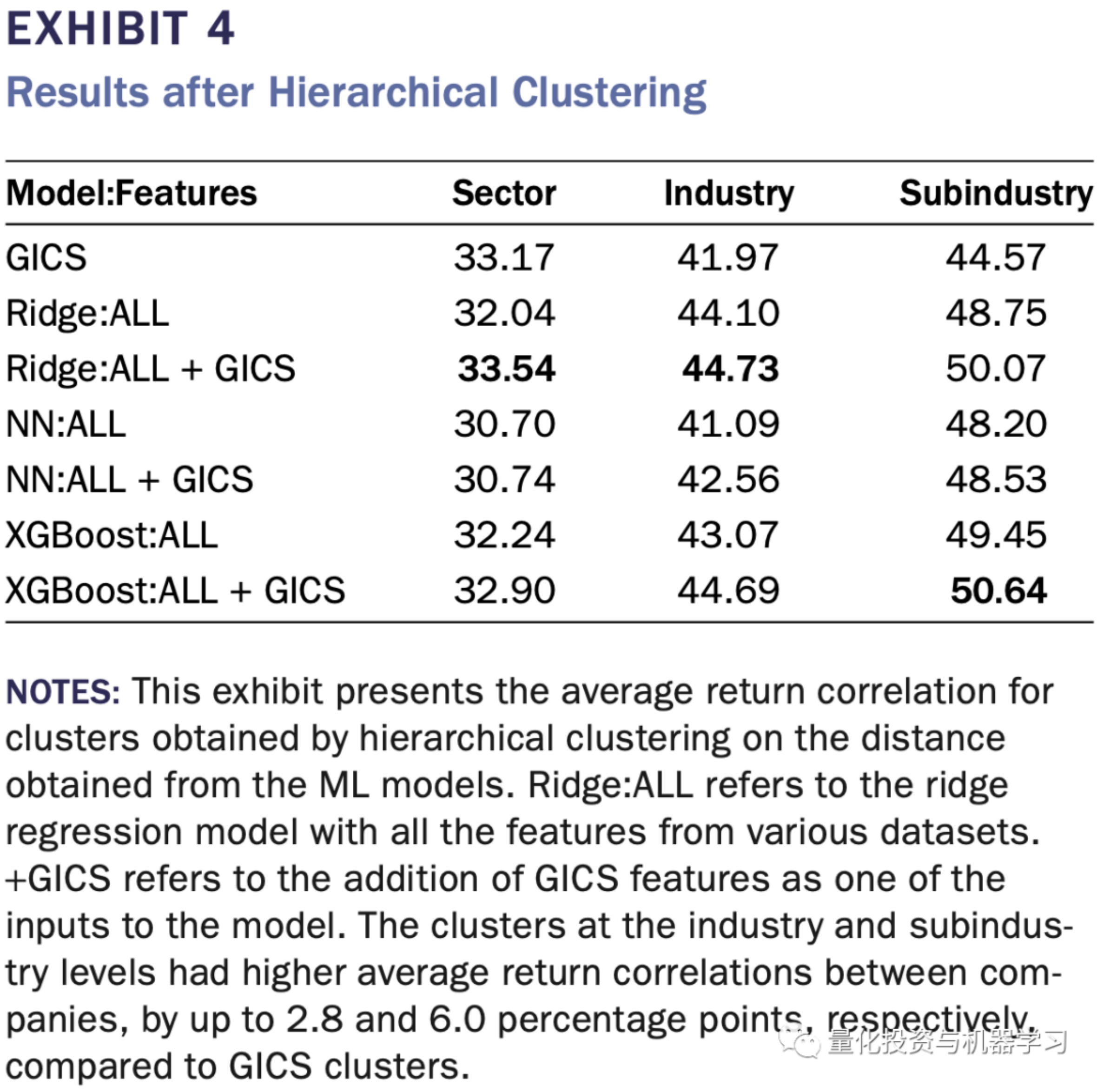


EXHIBIT 4

Results after Hierarchical Clustering

Model:Features	Sector	Industry	Subindustry
GICS	33.17	41.97	44.57
Ridge:ALL	32.04	44.10	48.75
Ridge:ALL + GICS	**33.54**	**44.73**	50.07
NN:ALL	30.70	41.09	48.20
NN:ALL + GICS	30.74	42.56	48.53
XGBoost:ALL	32.24	43.07	49.45
XGBoost:ALL + GICS	32.90	44.69	**50.64**

NOTES: This exhibit presents the average return correlation for clusters obtained by hierarchical clustering on the distance obtained from the ML models. Ridge:ALL refers to the ridge regression model with all the features from various datasets. +GICS refers to the addition of GICS features as one of the inputs to the model. The clusters at the industry and subindustry levels had higher average return correlations between companies, by up to 2.8 and 6.0 percentage points, respectively, compared to GICS clusters.

除了组内的相关性，本文对不同聚类的持续性进行了对比。如下表[Sector:90%, Ridge]对应的值35，表示，在使用Ridge模型时（使用所有特征），35%的Sector层次的聚类中的股票与下一期相比股票的变动小于90%。

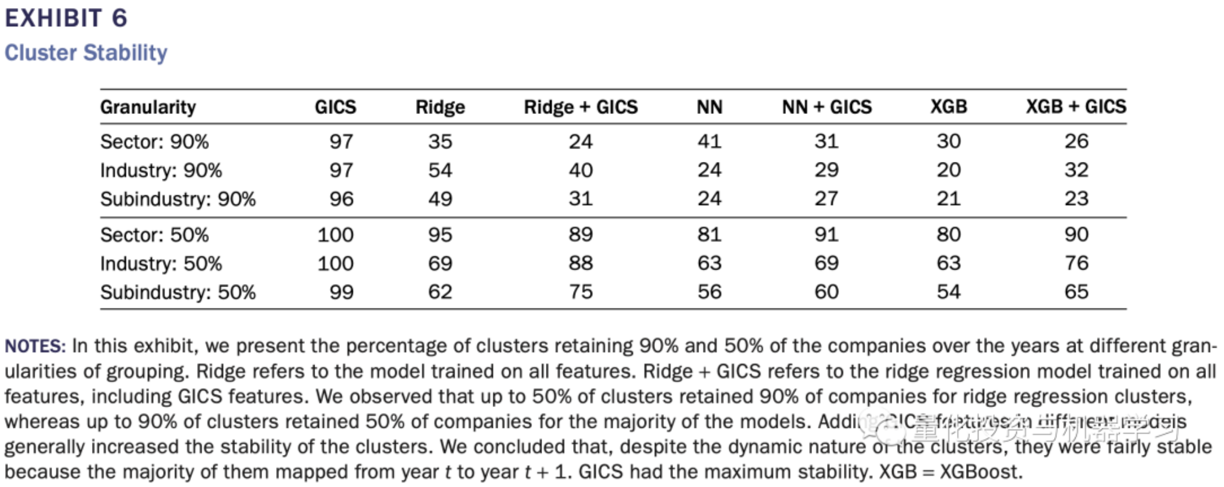


EXHIBIT 6

Cluster Stability

Granularity	GICS	Ridge	Ridge + GICS	NN	NN + GICS	XGB	XGB + GICS
Sector: 90%	97	35	24	41	31	30	26
Industry: 90%	97	54	40	24	29	20	32
Subindustry: 90%	96	49	31	24	27	21	23
Sector: 50%	100	95	89	81	91	80	90
Industry: 50%	100	69	88	63	69	63	76
Subindustry: 50%	99	62	75	56	60	54	65

NOTES: In this exhibit, we present the percentage of clusters retaining 90% and 50% of the companies over the years at different granularities of grouping. Ridge refers to the model trained on all features. Ridge + GICS refers to the ridge regression model trained on all features, including GICS features. We observed that up to 50% of clusters retained 90% of companies for ridge regression clusters, whereas up to 90% of clusters retained 50% of companies for the majority of the models. Adding GICS features in different models generally increased the stability of the clusters. We concluded that, despite the dynamic nature of the clusters, they were fairly stable because the majority of them mapped from year *t* to year *t* + 1. GICS had the maximum stability. XGB = XGBoost.

我们期望，属于同一聚类的公司将对不同的系统因子作出相似的反应。因此，每个公司聚类可以被认为是一个因子，可以解释系统冲击共同的因素。我们分析了从ML模型中获得的聚类作为因子，并评估了它们的同质性和样本外多样化效益：

$$R_{j,t} = A(t) + \sum_{i=1}^{N} L_{i,t} D_{j,i,t} + \epsilon_{j,t}$$

其中 $R_{j,t}$ 为股票收益， $D_{j,i,t}$ 为股票聚类暴露因子，当t时刻股票j属于聚类 D_i 时，该值为1，不属于为0（类似因子模型中的行业暴露因子）。

下表7中展示了不同聚类模型下，**聚类暴露因子收益的截面方差均值，方差越大说明不同聚类的收益区别越大，分散效果就越好。**可以看出不同模型的分散效果均好于GICS本身。

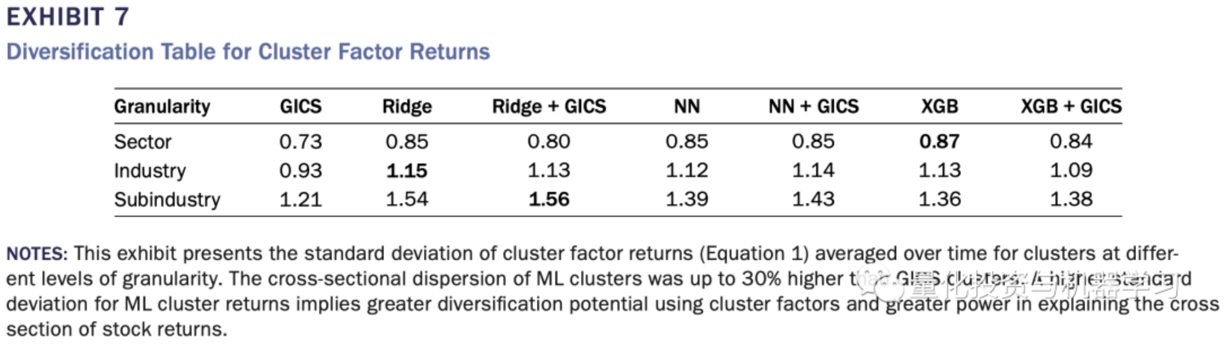


EXHIBIT 7

Diversification Table for Cluster Factor Returns

Granularity	GICS	Ridge	Ridge + GICS	NN	NN + GICS	XGB	XGB + GICS
Sector	0.73	0.85	0.80	0.85	0.85	**0.87**	0.84
Industry	0.93	**1.15**	1.13	1.12	1.14	1.13	1.09
Subindustry	1.21	1.54	**1.56**	1.39	1.43	1.36	1.38

NOTES: This exhibit presents the standard deviation of cluster factor returns (Equation 1) averaged over time for clusters at different levels of granularity. The cross-sectional dispersion of ML clusters was up to 30% higher than GICS clusters. A higher standard deviation for ML cluster returns implies greater diversification potential using cluster factors and greater power in explaining the cross section of stock returns.

除了比较聚类暴露因子的收益，本文还比较了聚类内及聚类间，相关基本面因子的离散度。如下表8和9所示。**整体可以看出，GICS行业分类体系，在多个基本面因子的组内离散度要好于机器学习模型，但机器学习模型相比GICS在组间的离散度更大。**

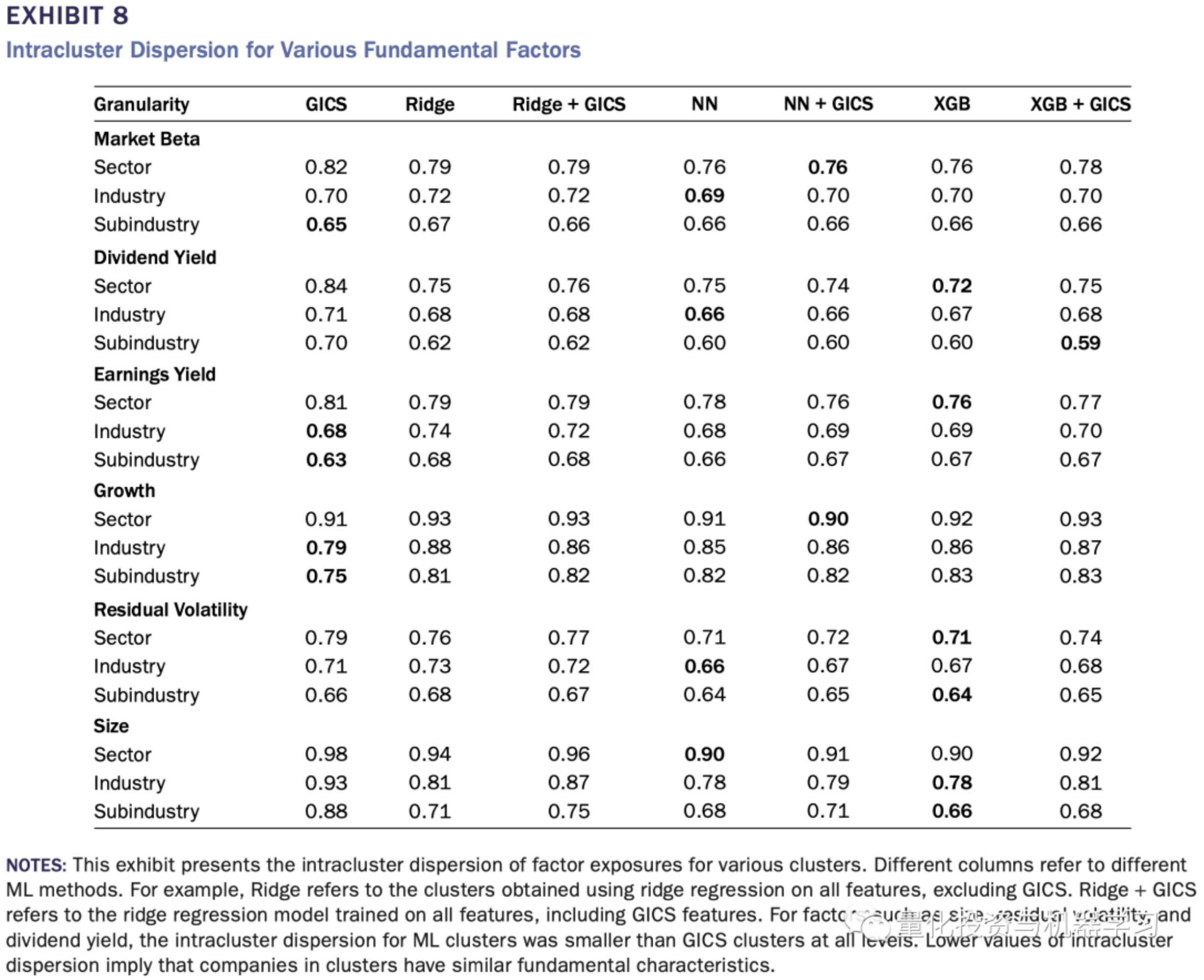


EXHIBIT 8

Intracluster Dispersion for Various Fundamental Factors

Granularity	GICS	Ridge	Ridge + GICS	NN	NN + GICS	XGB	XGB + GICS
Market Beta							
Sector	0.82	0.79	0.79	0.76	**0.76**	0.76	0.78
Industry	0.70	0.72	0.72	**0.69**	0.70	0.70	0.70
Subindustry	**0.65**	0.67	0.66	0.66	0.66	0.66	0.66
Dividend Yield							
Sector	0.84	0.75	0.76	0.75	0.74	**0.72**	0.75
Industry	0.71	0.68	0.68	**0.66**	0.66	0.67	0.68
Subindustry	0.70	0.62	0.62	0.60	0.60	0.60	**0.59**
Earnings Yield							
Sector	0.81	0.79	0.79	0.78	0.76	**0.76**	0.77
Industry	**0.68**	0.74	0.72	0.68	0.69	0.69	0.70
Subindustry	**0.63**	0.68	0.68	0.66	0.67	0.67	0.67
Growth							
Sector	0.91	0.93	0.93	0.91	**0.90**	0.92	0.93
Industry	**0.79**	0.88	0.86	0.85	0.86	0.86	0.87
Subindustry	**0.75**	0.81	0.82	0.82	0.82	0.83	0.83
Residual Volatility							
Sector	0.79	0.76	0.77	0.71	0.72	**0.71**	0.74
Industry	0.71	0.73	0.72	**0.66**	0.67	0.67	0.68
Subindustry	0.66	0.68	0.67	0.64	0.65	**0.64**	0.65
Size							
Sector	0.98	0.94	0.96	**0.90**	0.91	0.90	0.92
Industry	0.93	0.81	0.87	0.78	0.79	**0.78**	0.81
Subindustry	0.88	0.71	0.75	0.68	0.71	**0.66**	0.68

NOTES: This exhibit presents the intracluster dispersion of factor exposures for various clusters. Different columns refer to different ML methods. For example, Ridge refers to the clusters obtained using ridge regression on all features, excluding GICS. Ridge + GICS refers to the ridge regression model trained on all features, including GICS features. For factors such as size, residual volatility, and dividend yield, the intracluster dispersion for ML clusters was smaller than GICS clusters at all levels. Lower values of intracluster dispersion imply that companies in clusters have similar fundamental characteristics.

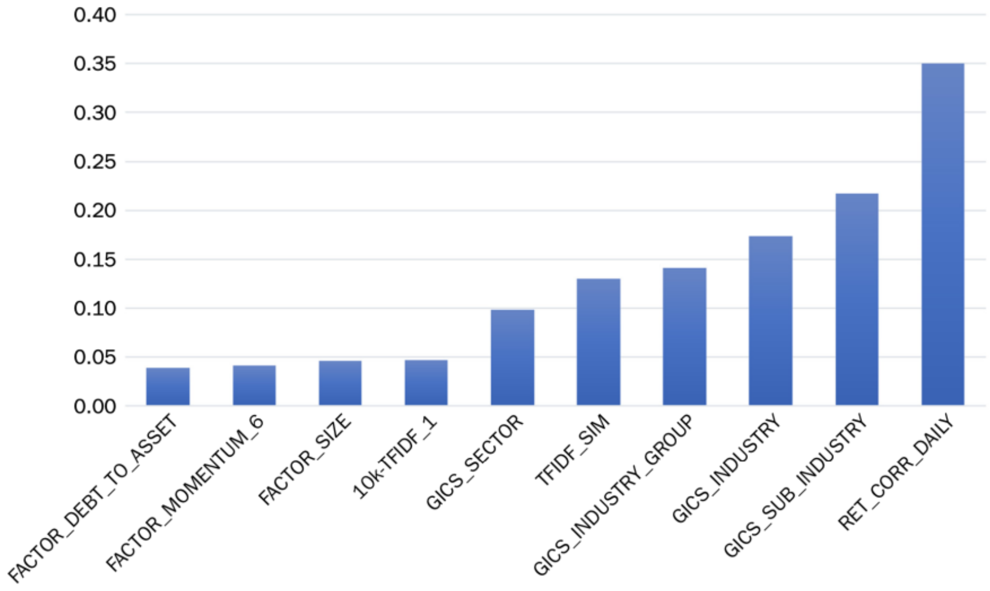
EXHIBIT 9
Diversification

Granularity	GICS	Ridge	Ridge + GICS	NN	NN + GICS	XGB	XGB + GICS
Market Beta							
Sector	0.69	0.77	0.74	**0.81**	0.77	0.80	0.76
Industry	0.68	0.69	0.70	**0.73**	0.74	0.73	0.72
Subindustry	0.72	0.74	0.75	0.75	**0.76**	0.75	0.75
Dividend Yield							
Sector	0.60	0.73	0.68	**0.76**	0.75	0.76	0.73
Industry	0.66	0.66	0.65	**0.73**	0.72	0.73	0.70
Subindustry	**0.73**	0.69	0.68	0.71	0.71	0.72	0.71
Earnings Yield							
Sector	0.44	0.66	0.62	0.69	0.66	**0.71**	0.67
Industry	0.50	0.67	0.67	0.68	**0.70**	0.69	0.67
Subindustry	0.57	0.78	0.78	0.78	0.78	**0.79**	0.78
Growth							
Sector	0.34	**0.41**	0.38	0.40	0.38	0.40	0.39
Industry	0.45	0.48	0.46	**0.50**	0.48	0.47	0.46
Subindustry	0.53	**0.63**	0.63	0.61	0.60	0.57	0.57
Residual Volatility							
Sector	0.49	0.66	0.63	0.72	0.72	**0.73**	0.69
Industry	0.53	0.68	0.68	0.70	**0.71**	0.70	0.69
Subindustry	0.61	0.78	**0.79**	0.78	0.79	0.78	0.79
Size							
Sector	0.22	0.32	0.23	**0.37**	0.34	0.36	0.30
Industry	0.44	0.58	0.49	0.62	0.58	**0.63**	0.57
Subindustry	0.62	0.70	0.68	0.73	0.69	**0.73**	0.72

NOTES: This exhibit presents the standard deviation of a cluster's fundamental value averaged over time for clusters at different levels of granularity. Different columns refer to different ML methods. For example, Ridge refers to the ridge regression model trained on all features, excluding GICS. Ridge + GICS refers to ridge regression trained on all features, including GICS. For factors such as residual volatility, size, and earnings yield, the cross-sectional standard deviation was around 40% higher than GICS clusters. Higher standard deviation of fundamental factor exposures for ML clusters implies higher cross-sectional dispersion in cluster factors.

通过以上分析，我们发现整体上Ridge: ALL+GICS的模型表现更优。下图12和13分别给出了Ridge: ALL+GICS模型中特征的重要性，及不同类别特征的重要性。可以看出，日度收益率在预测未来股票相关性系数上是最重要的，然后还有股票本身所属的GICS行业分类。而在特征大类的重要性上，收益率数据依然是最重要的，排在后面的就是基本面因子。

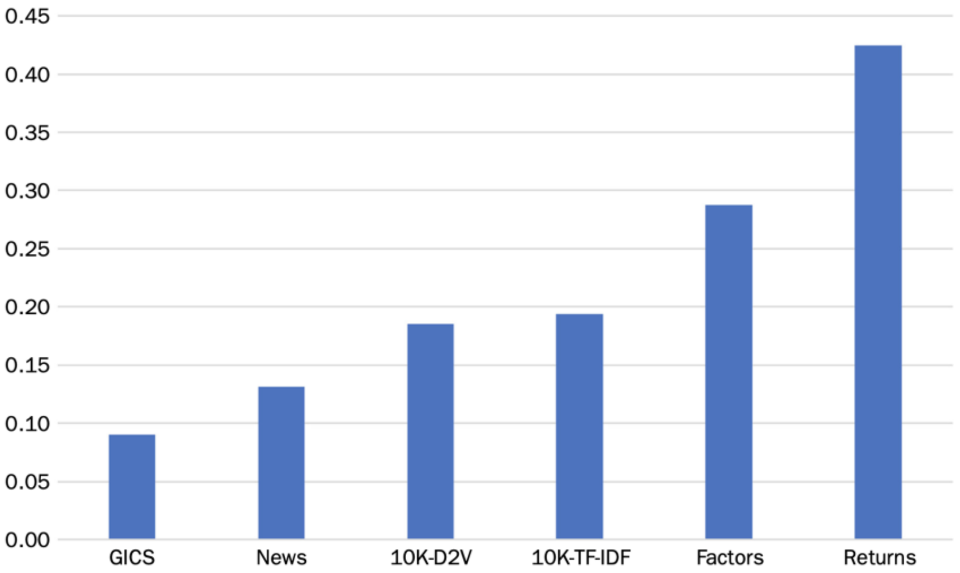
EXHIBIT 12
Feature Importance



NOTES: The exhibit shows the 10 most significant features for the ridge regression model aggregated over testing years. The return correlation as calculated from historical data was most important, followed by certain GICS features. Some fundamental factors such as size, momentum, and the debt-to-asset ratio were also found in the 10 most important features.

EXHIBIT 13

Feature Importance Aggregated by Dataset



NOTES: This exhibit shows the importance of features obtained by the absolute value of coefficients from different datasets and aggregated over years for the ridge regression model. The returns dataset was most important, followed by factor exposure and 10-K information.

总结

本文给我们最大的启示，是使用机器学习模型及多源数据对现有的行业分类体系进行改进，最终达到组合投资中更优的分散性。由于市场的对于不同公司的认知是不断变化的，传统的行业分类的更新可能跟不上市场的变化，从而导致同属一个行业的公司，表现并不一定会更相关。

其次，也给我们提供了一个新的思路，股票间的相关性是可预测的。我们之前一直将机器学习模型用于收益的预测，本文关于相关性的预测，给机器学习提供了一个新的可以尝试的应用场景。

收录于合集 #深度研读系列 21

< 上一篇 · 组合优化哪家强，HRP当自强！

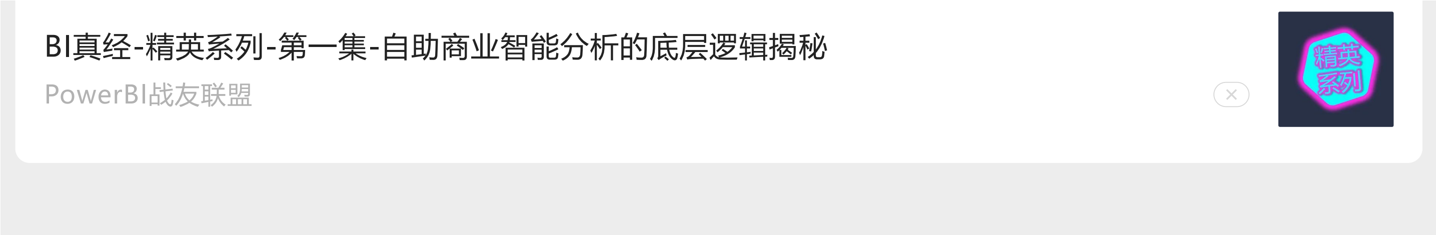
People who liked this content also liked

ETF拯救世界丨今天跌的还可以

提升之路

被Linux之父骂醒？英伟达破天荒开源GPU内核驱动，网友：活久见

机器之心



BI真经-精英系列-第一集-自助商业智能分析的底层逻辑揭秘

PowerBI战友联盟