个人数据分析报告：
数字化转型的颠覆性创新效应研究

（2024-2025 学年第一学期）

学　　院　经济学院

课程名称　R 语言编程基础与金融数据分析

班　　级　22 金融 03 班

学　　号　24210819

姓　　名　周生瑞

任课教师　王皓

撰写日期　2024 年 1 月 10 日

附件地址 https://sr6688.github.io/

目录

引言.....4

理论分析与模型构建.....6

理论分析.....6

企业数字化转型与颠覆式创新.....6

企业数字化转型对颠覆式创新投入的影响.....6

基准模型设定.....7

数据分析模型.....7

弹性网模型.....7

随机森林模型.....8

神经网络模型.....10

变量说明.....17

实证分析.....20

描述性统计.....20

数据可视化.....20

相关性分析.....21

基准模型结果.....22

数据分析结果.....23

弹性网模型.....23

随机森林模型.....24

神经网络模型.....25

结论与建议.....27

图表索引

图表 1 岭回归估计量与 OLS 估计量在方差与偏差权衡下的选择.....8
图表 2 作为分段常值函数的决策树.....9
图表 3 特征变换.....13
图表 4 人工神经网络的神经元示意图.....14
图表 5 多个输出结果的感知机.....15
图表 6 多层感知机.....15
图表 7 双隐藏层神经网络.....16
图表 8 主要变量定义.....18
图表 9 样本结构.....19
图表 10 描述性统计.....21
图表 11 相关性分析.....22
图表 12 基准模型估计结果.....23
图表 13 弹性网模型估计结果.....24
图表 14 随机森林模型估计结果.....25
图表 15 神经网络训练结果.....26
图表 16 神经网络估计结果.....26

引言

随着数字化技术的迅猛发展，企业所处的商业环境发生了翻天覆地的变化。在数字经济时代，企业面临着前所未有的机遇与挑战。一方面，大数据、云计算、人工智能等数字技术的广泛应用，为企业提供了全新的生产方式、商业模式和创新路径，促使企业加快数字化转型的步伐，以提升自身的竞争力和适应市场变化的能力。另一方面，数字创新和颠覆式创新成为企业获取竞争优势的关键因素。数字创新通过信息、计算、沟通和连接技术的组合，带来新的产品、生产过程改进、组织模式变革以及商业模式的创建和改变。颠覆式创新则通过引入新的技术或商业模式，改变现有市场的竞争格局，对现有企业构成威胁。例如，Uber 通过基于智能手机的叫车服务，颠覆了传统出租车行业；Pandora 通过互联网广播服务，改变了音乐产业的商业模式。这些案例充分说明了在当前时代，企业必须积极拥抱数字化转型，开展数字创新和颠覆式创新，才能在激烈的市场竞争中立于不败之地。

学术界对企业数字化转型、数字创新和颠覆式创新的内涵和关系进行了广泛而深入的研究。在企业数字化转型方面，有学者认为其是企业利用数字技术推动业务模式创新、管理模式变革、商业模式重构与核心能力提升的过程（李健等，2022；），进而促进产业链、价值链、供应链和创新链的数字化融合，构建更加开放的数字生态。数字创新的特征包括收敛性、自生长性、数据同质化和可重新编程性（Yoo et al.，2012；Nambisan et al.，2017），其分类涵盖数字产品创新、数字过程创新、数字组织创新和数字商业模式创新（刘洋等，2020）。颠覆式创新则具有初始性能不足、价值网络变化、低端市场和新市场等特征（Christensen，1997）。在三者的关系上，数字技术作为颠覆性创新的驱动力，为颠覆性创新提供了新的工具和平台（Gawer，2013）；数字创新的自生长性促进颠覆，推动企业不断改进和优化产品，最终实现对现有市场的颠覆（Ansari and Krop，2012）；数字平台和生态系统为企业提供了新的商业模式和创新机会，支持颠覆性创新的产生和扩散（Ansari et al.，2016）。然而，现有文献在探讨企业数字化转型、数字创新和颠覆式创新时，多侧重于理论层面的分析（Yu and Hang，2009；Christensen and Bever，2014），虽然有部分学者针对理论原因进行了讨论（李玉花和简泽，2021；刘海兵等，2023），但对于三者在企业实践上的具体应用和相互作用机制

的研究相对不足，尤其是在不同行业、不同规模企业中的差异化表现和影响因素等方面（Ahuja et al.，2009），尚缺乏系统性和实证性的研究。

本文选用 2010-2022 沪深 A 股上市公司进行了实证分析。具体来说，本文采用了固定效应模型、弹性网模型、随机森林模型和神经网络模型等多种数据分析方法，以全面检验企业数字化转型、行业集中程度和企业融资约束对颠覆式创新的影响。其中，固定效应模型用于控制不随时间变化的个体异质性和年度特定影响，确保估计结果的稳健性。弹性网模型结合了拉索回归和岭回归，有效处理多重共线性问题，平衡偏差和方差。随机森林模型通过构建多棵决策树，提高预测准确性和稳定性，并通过变量重要性图和偏依赖图直观展示各变量对创新的影响。神经网络模型，特别是多层感知机，能够捕捉复杂的非线性关系，适用于处理复杂的特征交互。这些模型的综合运用，使得本文能够从多个角度全面评估企业数字化转型对颠覆式创新的影响，确保了研究结果的可靠性和深度。

本研究聚焦于企业数字化转型、数字创新和颠覆式创新的内涵、关系及其对企业发展的实际影响，具有重要的现实意义。首先，能够为企业提供数字化转型和创新的理论指导和实践参考，帮助企业更好地理解数字化转型的内涵和路径，明确数字创新和颠覆式创新的方向和重点，从而制定科学合理的战略规划，提升企业的创新能力和核心竞争力。其次，有助于丰富和完善相关领域的理论体系，推动学术界对企业数字化转型、数字创新和颠覆式创新的深入研究，为后续的理论创新和实践探索奠定基础。再次，对于政府相关部门制定产业政策和科技政策具有重要的参考价值，能够为政府在引导企业数字化转型、促进产业升级和推动经济高质量发展等方面提供决策依据，助力数字经济与实体经济的深度融合。最后，本研究还能够为高校和职业院校的人才培养提供方向指引，促使教育机构根据企业数字化转型和创新的需求，调整专业设置和课程体系，培养更多适应数字经济发展的高素质专业人才，为经济社会发展提供坚实的人才保障。

理论分析与模型构建

理论分析

企业数字化转型与颠覆式创新

数字技术的高技术性、广覆盖性和强渗透性是推动企业颠覆式创新的基础动力。企业实施数字化转型，不仅是数字技术的简单组合和运用，而是将数据作为新生产要素投入到企业生产经营中，包括生产资料的数字化渗透、生产关系的数字化重构和商业活动的数字化创新。第一，数字化转型能够帮助企业管理层形成数据驱动决策模式，高效精准地分析市场需求与企业前景，畅通企业创新决策信息渠道，增加企业创新意愿。第二，数字化转型能够降低信息成本，增强企业信息获取与处理能力，帮助企业高效率、低成本地使用信息，降低生产经营过程中的边际投入。同时，数字化转型能够降低信息不对称，帮助企业实现资源有效整合，压缩企业内外协同成本，提高数据边际要素生产率，提升企业运营效率。第三，数字化转型能够提高信息使用效率，加快信息扩散和知识溢出，提高信息流动的规模性、效率性和融合性，开放式创新网络的构成为企业创新活动开展提供良好环境。数字化转型不仅深刻改变着企业创新环境，还改变了企业传统研发创新模式，拓宽了企业创新边界，实现由传统内部研发向开放式、网络化创新转变。

基于以上分析，本文提出以下假说：

H_1:企业数字化转型促进颠覆式创新。

企业数字化转型对颠覆式创新投入的影响

供应链集中度包括上游供应商集中度和下游客户集中度，供应链集中度增加会降低企业话语权，导致企业经营风险增加，不利于企业持续性创新投入。企业数字化转型能够利用数字专业技术加快企业信息获取和处理，使产业链中的企业信息更加透明，供需市场信息更加开放，企业间信息交流更加便捷。数字平台技术的兴起，促进企业与产业链中上下游企业开展更多新型合作方式，降低企业供应商集中度和客户集中度，提升企业话语权并降低经营风险，避免企业创新资源被侵蚀，促进企业持续性创新投入。

H_2:企业数字化转型通过降低供应链集中度为企业持续性创新投入提供内源资金保障。

创新活动需要大量资金作为研发经费，持续性创新活动更需要稳定的研发支出。企业数字化转型可以缓解创新投入压力，打破融资桎梏。基于信号传递理论，企业实施数字化转型向外界传递出大胆改变、积极部署的信号，易于拓展企业创新所需的外部资金渠道。数字经济改变传统金融服务模式，凭借数字化技术降低企业与金融机构之间的信息不对称，使金融机构和投资者能够更加精准地识别评估和降低投资企业风险，提高信贷资源分配效率。

H_3:企业数字化转型通过缓解融资约束减少企业颠覆式创新投入外源融资压力。

基准模型设定

本文首先采用以下回归模型来检验企业数字化转型对持续性创新的影响：

$$Innovation_{i,t} = a_0 + a_1 DT_{i,t} + a_2 Ind_{i,t} + a_3 Fin_{i,t} + \boldsymbol{\Lambda C_{it}} + \mu_i + \gamma_t + \varepsilon_{i,t}$$

其中，$DT_{i,t}$表示企业数字化转型指数，$Ind_{i,t}$表示行业集中程度，$Fin_{i,t}$表示企业融资约束，$\boldsymbol{C_{it}}$表示相关控制变量的集合；μ_i为个体固定效应；γ_t为年份固定效应；$\varepsilon_{i,t}$为随机误差项。这里我们主要关注系数a_1、a_2和a_3。

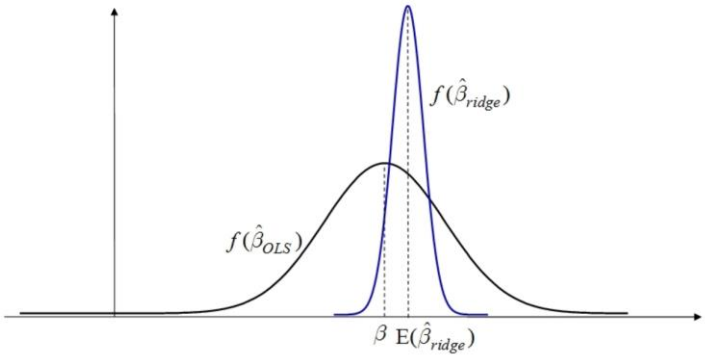
数据分析模型

弹性网模型

对于任意估计量 $\hat{\boldsymbol{\beta}}$ ，其均方误差可分解为方差与偏差平方之和：

$$MSE(\hat{\boldsymbol{\beta}}) \equiv E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] = Var(\hat{\boldsymbol{\beta}}) + [Bias(\hat{\boldsymbol{\beta}})][Bias(\hat{\boldsymbol{\beta}})]'$$

因此，使均方误差最小化，可视为在方差与偏差之间进行权衡（trade-off）。比如，一个无偏估计量（偏差为 0），如果方差很大，则可能不如一个虽然有偏差但是方差却很小的估计量。在（严格）多重共线性的情况下，虽然 OLS 估计量无偏，但其方差太大（无穷大），而岭回归虽有少量偏差，但可大幅减少方差，这使得岭回归估计量的均方误差（MSE）可能比 OLS 更小：



图表 1 岭回归估计量与 OLS 估计量在方差与偏差权衡下的选择

ZouandHastie（2005）将 Lasso 与岭回归相结合，提出弹性网（elastic net）估计量。在弹性网估计量损失函数中，同时包含L_1与L_2惩罚项：

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2$$

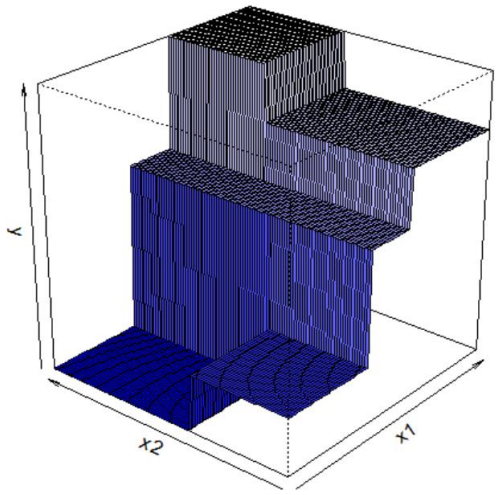
其中，$\lambda_1 \geqslant 0$与$\lambda_2 \geqslant 0$都是调节参数。由于λ_1与λ_2的取值范围均为无穷，不便于使用交叉验证选择其最优值。为此，定义$\lambda \equiv \lambda_1+\lambda_2, \alpha \equiv \frac{\lambda_1}{\lambda}$，可将损失函数等价地写为

$$\min_{\beta}(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda[\alpha\|\boldsymbol{\beta}\|_1+(1-\alpha)\|\boldsymbol{\beta}\|_2^2]$$

其中，$\lambda \geqslant 0$与$0 \leqslant \alpha \leqslant 1$为调节参数。这里，我们关注$\boldsymbol{\beta}$中的系数$a_1$、$a_2$和$a_3$。

随机森林模型

对于三维以上的特征空间，我们依然可用树状结构来表示，因为决策树每次仅使用一个变量进行分裂（splitting）决策树模型将特征空间分割为若干（超）矩形的终节点。在进行预测时，每个终节点只有一个共同的预测值。对于分类问题，此预测值为该终节点所有训练样本的最常见类别（most commonly occurring class）。对于回归问题，此预测值为该终节点所有训练样本的平均值。因此，在数学上，决策树为“分段常值函数”（piecewise constant function）。



图表 2 作为分段常值函数的决策树

对于回归问题，其响应变量y可为连续变量。因此，对于回归树，可使用“最小化残差平方和”作为节点的分裂准则。这意味着，在进行节点分裂时，希望分裂后残差平方和下降最多，即两个子节点的残差平方和的总和最小。为避免过拟合，对于回归树，也要使用惩罚项来进行修枝，即最小化如下目标函数：

$$\min_{T} \underbrace{\sum_{m=1}^{|T|} \sum_{\boldsymbol{x_i} \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2}_{cost} + \lambda \cdot \underbrace{|T|}_{complexity}$$

其中，R_m为第m个终节点，而$\hat{y}_{R_m}$为该终节点的预测值（此终节点的样本均值）。$\sum_{\boldsymbol{x_i} \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2$为第$m$个终节点的残差平方和，然后对所有终节点$m = 1,2,\cdots,T$进行加总，即为成本$R(T)$。

一方面，由于在每次节点分裂，仅使用一个变量，因此容易区分每个变量的贡献，考察该分裂变量使得残差平方和（或者基尼指数）下降多少。具体来说，对于每个变量，在随机森林的每棵决策树，可度量由该变量所导致的分裂准则函数的下降幅度。然后，针对此下降幅度，对每棵决策树进行平均，即为对该变量重要性的度量。将每个特征变量的重要性依次排列画图，即为变量重要性图（Variable Importance Plot）。这里我们关注数字化转型、行业集中、融资约束对于企业颠覆式创新的重要程度。

另一方面，我们依然感兴趣于每个变量对于y的边际效应（marginal effects）。比如，对于特征向量$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$，假设$y=f(\boldsymbol{x})$，但函数$f(\cdot)$无解析表达式（比如随机森林）。不失一般性，考虑第 1 个特征变量x_1对y的边际效应：

$$\frac{\partial y}{\partial x_1}=\frac{\partial f(x_1,x_2,\cdots,x_p)}{\partial x_1}$$

在上式中，边际效应$\frac{\partial y}{\partial x_1}$依赖于其他变量$(x_2,\cdots,x_p)$取值，一般来说不是常数（除非是线性模型）。为此，考虑在函数$y=f(x_1,x_2,\cdots,x_p)$中，将其他变量$(x_2,\cdots,x_p)$对于y的影响通过积分平均掉：

$$\phi(x_1)\equiv E_{x_2,\cdots,x_p}f(x_1,x_2,\cdots,x_p)$$

其中，期望算子$E_{x_2,\cdots,x_p}(\cdot)$对变量$(x_2,\cdots,x_p)$求期望，故在右边已将$(x_2,\cdots,x_p)$积分积掉，因此所得结果$\phi(x_1)$只是$x_1$的函数。当然，由于$f(\cdot)$无解析表达式，一般很难直接计算此期望。为此，使用统计学的常用估计方法，以样本均值替代总体均值$E_{x_2,\cdots,x_p}(\cdot)$可得：

$$\hat{\phi}(x_1)\equiv\frac{1}{n}\sum_{i=1}^{n}f(x_1,x_{i2},\cdots,x_{ip})$$

在式中，任意给定x_1均可计算$\hat{\phi}(x_1)$，并可画出$\left(x_1,\hat{\phi}(x_1)\right)$的图像，称为偏依赖图（Partial Dependence Plot）。这里我们关注数字化转型、行业集中、融资约束对于企业颠覆式创新的单个影响及其联合影响。

神经网络模型

求解最大间隔的超平面的约束极值问题为

$$\begin{aligned}\max_{\boldsymbol{\beta},\beta_0}\quad&\frac{2}{\|\boldsymbol{\beta}\|}\\ \text{s. t.}\quad&y_if(\boldsymbol{x}_i)\geqslant 1,i=1,\cdots,n\end{aligned}$$

最大化$\frac{2}{\|\boldsymbol{\beta}\|}$等价于最小化$\|\boldsymbol{\beta}\|$，而后者又等价于最小化$\frac{1}{2}\|\boldsymbol{\beta}\|^2=\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$。将“$f(\boldsymbol{x}_i)=\boldsymbol{\beta}_0+\boldsymbol{\beta}'\boldsymbol{x}_i$”代入约束条件，则最优化问题可写为

$$\begin{aligned}\min_{\boldsymbol{\beta},\beta_0}\quad & \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}\\ \text{s.t.}\quad & y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\geqslant 1, i=1,\cdots,n\end{aligned}$$

由于目标函数$\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}=\frac{1}{2}\left(\boldsymbol{\beta}_1^2+\cdots+\boldsymbol{\beta}_p^2\right)$为二次型，约束条件$y_i(\boldsymbol{\beta}_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\geqslant 1$为线性不等式约束，故为“凸二次规划”（convex quadratic programming）问题。为求解此问题，引入“原问题”（primal problem）的拉格朗日乘子函数 L_P：

$$\begin{aligned}\min_{\boldsymbol{\beta},\beta_0,\boldsymbol{\alpha}} L_P(\boldsymbol{\beta},\beta_0,\boldsymbol{\alpha}) \quad &= \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}-\sum_{i=1}^{n}\alpha_i[y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)-1]\\ &= \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}-\beta_0\sum_{i=1}^{n}\alpha_i y_i-\boldsymbol{\beta}'\sum_{i=1}^{n}\alpha_i y_i\boldsymbol{x}_i+\sum_{i=1}^{n}\alpha_i\end{aligned}$$

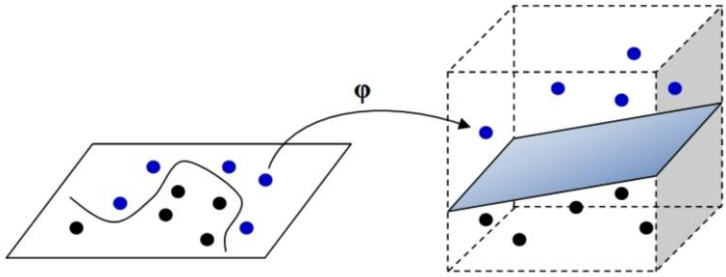
其中，$\boldsymbol{\alpha}\equiv(\alpha_1\cdots\alpha_n)'$ 为对应于约束条件 $y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\geqslant 1$ 的 n 个拉格朗日乘子。可以证明，原问题的最优解为拉格朗日乘子函数的 “鞍点”，故单独从拉格朗日乘子 $\boldsymbol{\alpha}$ 来看，则为最大化问题。

将一阶条件代回拉格朗日函数，可得到“对偶问题”（dual problem）L_D：

$$\begin{aligned}\max_{\boldsymbol{\alpha}} L_D &= \frac{1}{2}\underbrace{\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)'}_{=\hat{\boldsymbol{\beta}}'}\underbrace{\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)}_{=\hat{\boldsymbol{\beta}}} - \beta_0\underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{=0} \\ &\quad - \underbrace{\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)'}_{=\hat{\boldsymbol{\beta}}'}\underbrace{\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x}_i\right)}_{=\hat{\boldsymbol{\beta}}} + \sum_{i=1}^{n}\alpha_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i}\right)'\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i}\right) \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\left(\sum_{i=1}^{n}\alpha_i y_i \boldsymbol{x_i'}\right)\left(\sum_{j=1}^{n}\alpha_j y_j \boldsymbol{x_j}\right) \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i' x_j} \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle\end{aligned}$$

其中，$\boldsymbol{x_i' x_j}$为$\boldsymbol{x}_i$与$\boldsymbol{x}_j$的内积，可记为$\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle$。这意味着，特征向量$\{\boldsymbol{x}_i\}_{i=1}^n$仅通过相互之间内积的方式而影响最优解。这为第在支持向量机中使用“核技巧”（kernel trick）提供了方便。最大化问题是关于拉格朗日乘子$\boldsymbol{\alpha}$的二次（型）规划问题。求解此对偶问题，并将所得解$\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1 \cdots \hat{\alpha}_n)'$代回最优$\boldsymbol{\beta}$表达式可得，$\hat{\boldsymbol{\beta}} = \sum_{i=1}^n \hat{\alpha}_i y_i \boldsymbol{x}_i$。对于截距项$\beta_0$，可通过支持向量来求解。

更一般地，对于决策边界非线性数据，考虑对特征向量$\boldsymbol{x}_i$进行变换，比如将$\boldsymbol{x}_i$变换为$\boldsymbol{\varphi}(\boldsymbol{x}_i)$；其中，$\boldsymbol{\varphi}(\boldsymbol{x}_i)$为多维函数（维度可以高于$\boldsymbol{x}_i$），甚至可以无限维函数。这意味着，可将训练样本$\{\boldsymbol{x}_i, y_i\}_{i=1}^n$变换为$\{\boldsymbol{\varphi}(\boldsymbol{x}_i), y_i\}_{i=1}^n$，目的是希望在$\boldsymbol{\varphi}(\boldsymbol{x}_i)$的特征空间（feature space）中，可以得到线性可分的情形，参见图 12。但难点在于，对于高维数据，一般不知道变换$\boldsymbol{\varphi}(\cdot)$的具体形式。



图表 3 特征变换

根据上述推导，支持向量机的估计结果仅依赖于$\langle \boldsymbol{\varphi}(\boldsymbol{x}_i), \boldsymbol{\varphi}(\boldsymbol{x}_j) \rangle$，即$\boldsymbol{\varphi}(\boldsymbol{x}_i)$与$\boldsymbol{\varphi}(\boldsymbol{x}_j)$的内积，而不必知道$\boldsymbol{\varphi}(\cdot)$。

支持向量回归（support vector regression，简记 SVR）的基本思想是，将支持向量机的合页损失函数移植到回归问题。记回归函数（超平面）为$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}$，并以此函数预测连续型响应变量$y$。SVR的目标函数为：

$$\min_{\boldsymbol{\beta},\beta_0} \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta} + C\sum_{i=1}^{n} \ell_\varepsilon[y_i - f(\boldsymbol{x}_i)]$$

其中，$C > 0$为正则化参数（regularization parameter），$z_i \equiv y_i - f(\boldsymbol{x}_i)$为残差（residual）；而$\ell_\varepsilon(\cdot)$为$\varepsilon$-不敏感损失函数（$\varepsilon$-insensitive loss function）：

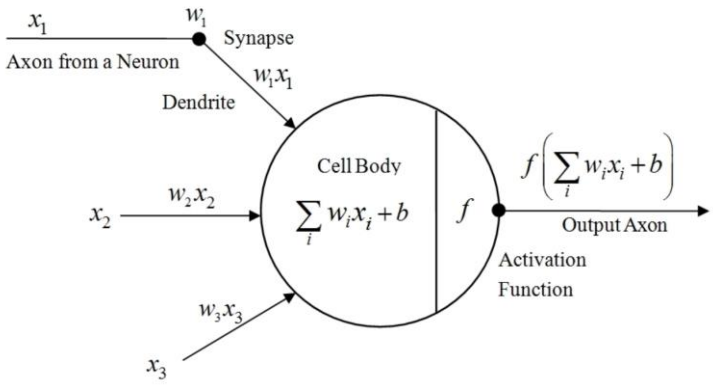
$$\ell_\varepsilon(z_i) = \begin{cases} 0 & \text{当 } |z_i| \leqslant \varepsilon \\ |z_i| - \varepsilon & \text{当 } |z_i| > \varepsilon \end{cases}$$

其中，$\varepsilon > 0$ 也是调节参数。

对于二分类问题，考虑使用分离超平面“$b + \boldsymbol{w}'\boldsymbol{x} = 0$”进行分类，而响应变量$y \in \{1, -1\}$。如果$b + w'\boldsymbol{x} > 0$则预测$y = 1$。反之，如果$b + w'\boldsymbol{x} < 0$则预测$y = -1$（$b + \boldsymbol{w}'\boldsymbol{x} = 0$可随意预测）。显然，正确分类要求$y_i(b + \boldsymbol{w}'\boldsymbol{x}_i) > 0$。反之如果$y_i(b + \boldsymbol{w}'\boldsymbol{x}_i) < 0$，则为错误分类。从某个初始值$(\boldsymbol{w}_0, b_0)$出发，感知机希望通过调整参数$(\boldsymbol{w}, b)$，来使得模型的错误分类最少。具体来说，感知机目标函数为最小化所有分类错误观测值的“错误程度”之和（即“负裕度”（negative margin）$-y_i f(x_i)$）：

$$\min_{\boldsymbol{w},b} L(\boldsymbol{w},b) = -\sum_{i\in M} y_i(b + \boldsymbol{w}'\boldsymbol{x}_i)$$

其中， M 为所有错误分类（misclassified）个体下标的集合。

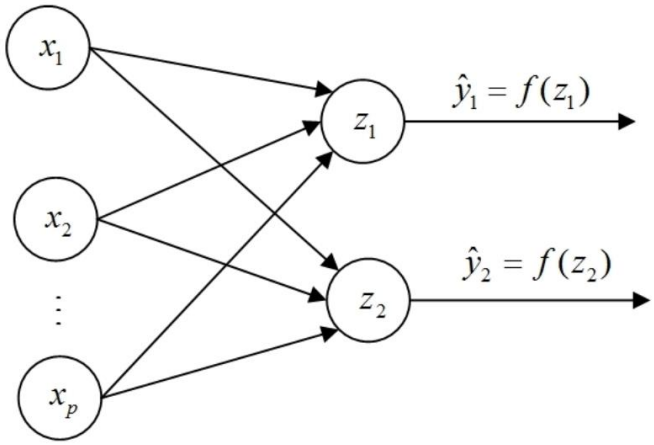


图表 4 人工神经网络的神经元示意图

感知机算法的直观解释为，当一个样本点被错误分类，即出现在分离超平面“$b + \boldsymbol{w}'\boldsymbol{x} = 0$”的错误一侧时，则调整参数$(\boldsymbol{w}, b)$，使得分离超平面向该误分类点一侧移动，以减少此误分类点与超平面的距离，直至正确分类为止。可以证明，对于线性可分的数据，则感知机一定会收敛，参见李航（2019）。这表明，只要给予足够的数据，感知机具备学得参数$(\boldsymbol{w}, b)$的能力，仿佛拥有“感知”世界的能力（比如，自动将事物分类），故名“感知机”。然而，从不同的初始值出发，一般会得到不同的分离超平面，无法得到唯一解。另一方面，如果数据为线性不可分，则感知机的算法不会收敛。感知机更严重的缺陷是，其决策边界依然为线性函数。

只要引入多层神经元，经过两个及以上的非线性激活函数迭代之后，即可得到非线性决策边界。在此，非线性的激活函数是关键；因为如果使用线性的激活函数，则无论叠加或嵌套多少次（相当于微积分的复合函数），所得结果一定还是线性函数。

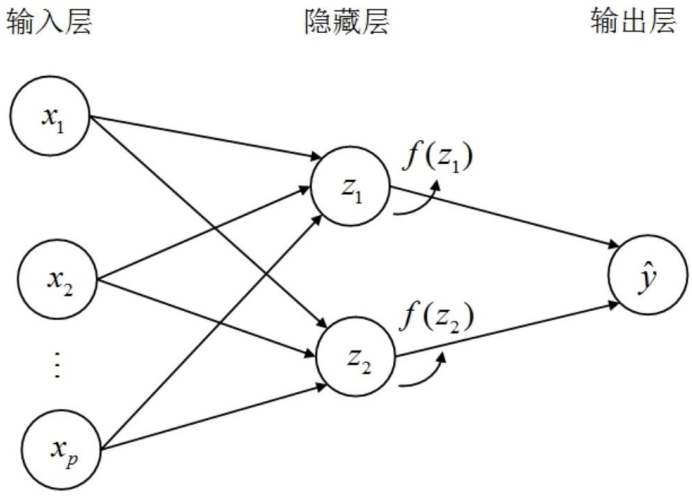
首先, 考虑具有多个输出结果（multi-output）的感知机，如图所示。



图表 5 多个输出结果的感知机

图中，共有两个输出（响应）变量，$\hat{y}_1$与$\hat{y}_2$。其中，$z_1 \equiv b_1 + \sum_{i=1}^{p} w_{i1}x_i$与$z_2 \equiv b_2 + \sum_{i=1}^{p} w_{i2}x_i$，均为在施加激活函数之前的加总值；而$f(\cdot)$为激活函数。

其次，图中的多个输出结果可重新作为输入变量，经过加权求和后，再次施以激活函数。



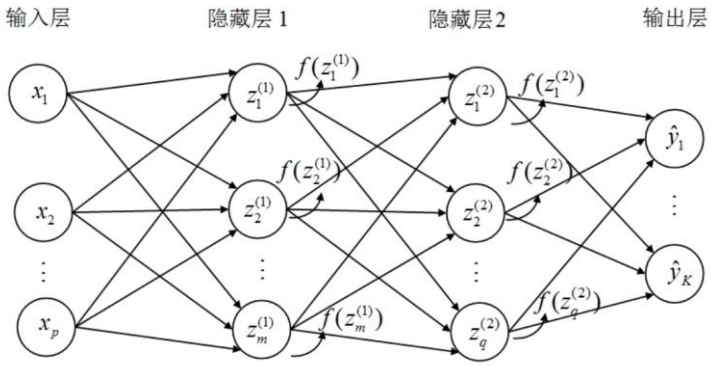
图表 6 多层感知机

最终输出结果为：

$$\hat{y} = f\left[b^{(2)} + w_1^{(2)} f(z_1) + w_2^{(2)} f(z_2)\right]$$

即对$f(z_1)$与$f(z_2)$再次加权求和，然后再施加激活函数$f(\cdot)$。显然，该函数所对应的决策边界为非线性的。图中，最左边为输入层 （input layer），中间为隐藏层（hidden layer），而最右边为输出层（output layer）。之所以将中间层称为“隐藏层”，因为该层的计算在算法内部进行，从外面并不可见。

当然，隐藏层可以有更多的神经元，而不仅是图中的两个隐藏神经元；而输出层也可有多个输出结果。在概念上，只要包含一个隐藏层，即为多层感知机（Multilayer Perceptron，简记 MLP），也就是“神经网络模型”。更一般地，神经网络模型可以有多个隐藏层。



图表 7 双隐藏层神经网络

如图所示，即为基本的神经网络结构。这种标准神经网络，称为前馈神经网络（feed forward neural network），因为输入从左向右不断前馈；也称为全连接神经网络（fully connected neural network），因为相邻层的所有神经元都相互连接。当然，针对特殊数据类型可能还需要一些特别的网络结构，比如，卷积神经网络（适用于图像识别）、循环神经网络（适用于自然语言等时间序列）等。另外，如果神经网络的隐藏层很多，则称为深度神经网络（deep neural networks），简称深度学习（deep learning）。

Cybenko （1988）与 Hornik et al.（1989）使用“泛函分析”（functional analysis），证明了神经网络的“通用近似定理”（universal approximation theorem）。其主要结论为，包含单一隐藏层的前馈神经网络模型，只要其神经元数目足够多，则可以以任意精度逼近任何一个在有界闭集上定义的连续函数。实际上，包含单隐藏层的前馈神经网络所代表的函数可写为

$$G(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i f(\boldsymbol{w}_i' \boldsymbol{x} + b_i)$$

其中,$(\boldsymbol{w}_i, b_i)$为第i个神经元的权重与偏置参数,$f(\cdot)$为激活函数,α_i为连接隐藏层与输出层的参数，而m为神经元的数目。通用近似定理表明，形如$G(\boldsymbol{x})$的函数在定义于有界闭集上的连续函数的集合中是“稠密的”（dense），这意味着对于任意有界闭集上的连续函数，都可以找到形如$G(\boldsymbol{x})$的函数（也即单隐层的前馈神经网络），使二者的距离任意接近。

综上所述，本文利用神经网络模型来刻画数字化转型、行业集中和融资约束在企业颠覆式创新实践上的具体相互作用机制。

变量说明

张陈宇等（2020）认为与现有技术相比，发明专利创新必须具有显著的实质性特征和显著的改善;实用新型和外观设计专利不会因新颖性和非显而易见性的不足而受到审查。因此，本文使用发明专利申请数来衡量颠覆式创新（$Innovation_D$），参考现有文献的做法（张陈宇等，2020），我们将企业申请的专利个数加 1 进行对数化处理。

针对主要解释变量，关于企业数字化转型程度（$Trans$），本文借鉴已有研究成果（吴非，2021）通过 Python 爬虫功能归集整理了样本内沪深 A 股上市企业的年度报告，并以此作为数据池供后续的特征词筛选。由于这类数据具有典型的“右偏性”特征，本文将其进行对数化处理，从而得到刻画企业数字化转型的整体指标。关于市场集中程度，本文选用赫芬达尔指数（HHI）来衡量企业面对的市场结构。关于融资约束，本文选用KZ指数（KZ）来衡量企业的融资困境。

本文从公司基本治理属性和公司结构特征选取企业特征变量作为企业层面控制变量，选取企业规模（$Size$）、企业年龄（Age）、资产负债率（Lev）、总主营业务增长率（$Grow$）、资产收益率（ROA）、现金资产占比（$Cashflow$）、资本密集度（$Capital$）、董事会规模（$Board$）、股权集中度（$Tops$）、机构投资者占比（Ins）、独立董事占比（Idr）等企业层面的控制变量。本文还考虑到地区经济因素对于企业数字技术创新的影响，选取区域人口规模（Pop）、经济发展水平

（GDP）、教育资源（Edu）、人力资本（$Human$）、产业结构（$Stru$）、财政收入（Tax）、政府支出（$Finance$）等区域层面的控制变量。

主要变量的度量方法如表所示。

图表 8 主要变量定义

变量类型	变量符号	变量名称	计算方法
被解释变量	颠覆式创新	$Innovation_D$	ln（发明专利申请数量+1）
解释变量	数字化转型程度	$Trans$	借鉴（吴非，2021）筛选企业年报特征词并且对数化处理
	市场集中程度	HHI	赫芬达尔指数
	融资约束	KZ	KZ指数
控制变量	企业规模	$Size$	ln（年末总资产）
	企业年龄	Age	ln（企业成立年限）
	资产负债率	Lev	负债总额占资产总额的比重
	主营业务增长率	$Grow$	主营业务收入增长率
	总资产收益率	ROA	总资产收益率
	现金资产占比	$Cashflow$	净现金流量/总资产
	资本密集度	$Capital$	总资产/营业收入
	董事会规模	$Board$	ln（董事会人数-）
	股权集中度	$Tops$	前五大股东持股比例
	机构投资者占比	Ins	金融机构持股比例
	独立董事占比	Idr	独立董事人数/董事会人数
	区域人口规模	Pop	ln（总人口）
	经济发展水平	GDP	地区生产总值/总人口
	教育资源	Edu	每百万常住人口拥有的高等学校数量
	人力资本	$Human$	高等学校在校生数/总人口
	产业结构	$Stru$	第三产业的生产总值/第二的生产总值

	财政收入	Tax	地区税收收入/地区国有及规模以上非国有工业企业数量总和
	政府支出	$Finance$	一般公共预算支出/地区国有及规模以上非国有工业企业数量总和

有关上市公司年报来自深圳证券交易所以及上海证券交易所的公开信息。有关企业基本信息、专利数据、财务指标以及公司治理的相关数据来自 CSMAR 数据库。有关地区基本信息、经济发展指标取自中经网统计数据库。有关数字经济发展相关法规数据来自北大法宝数据库。此外，本文对初始样本进行了如下处理：（1）剔除 ST、PT 类公司样本；（2）剔除金融行业公司；（3）剔除 2018 年及以后上市的公司样本；（4）剔除资产负债率小于 0 或大于 1 的企业；（5）剔除关键变量严重缺失的样本。经过整理，本文得到 2010—2020“公司—年度”26667 个样本观测值。

图表 9 样本结构

VarName	Obs	Mean	SD	Min	Median	Max
Innovation_D	26667	1.80	1.582	0	1.609438	8.842893
Trans	26667	1.24	1.367	0	.6931472	6.251904
HHI	26667	0.21	0.187	.038272	.140248	1
KZ	26667	1.44	2.254	-10.72232	1.572158	14.82334
Size	26667	22.07	1.353	13.7633	21.88793	28.63649
Age	26284	2.02	0.925	0	2.197225	3.401197
Lev	26667	0.45	1.204	.00708	.425527	178.3455
Grow	26667	5.55	824.433	-2.733488	.10389	134607.1
ROA	26667	0.03	0.436	-48.31592	.035798	22.00512
Cashflow	26667	0.16	0.129	-.059826	.123742	.999993
Capital	26667	3.09	16.893	.0876	1.932579	1764.061
Board	26667	2.13	0.201	.6931472	2.197225	2.890372
Tops	26667	53.08	15.520	.8109	53.0751	99.23

Ins	26667	44.57	24.553	.0001	46.47	101.1401
Idr	26667	37.52	5.632	0	35.71	100
Pop	26667	8.54	0.660	5.706186	8.661986	9.443355
GDP	26667	2.78	2.266	.4307114	2.027267	14.33843
Edu	26667	0.00	0.000	.0001138	.0001926	.0004434
Human	26667	203.93	43.826	79.91829	195.5115	412.6295
Stru	26667	1.45	0.977	.4995993	1.153471	5.296818
Tax	26667	3012.04	3616.627	382.9714	1590.291	18655.3
Fin	26667	6538.50	10833.555	498.3971	3300.515	160217.6

实证分析

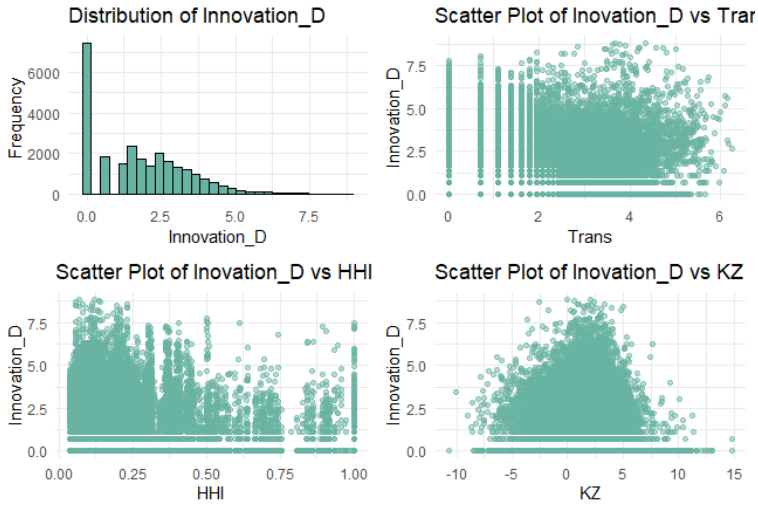
描述性统计

数据可视化

Innovation_D 的分布呈现右偏，大部分数据集中在 0 到 3 之间，显示出大多数企业的创新得分相对较低。然而，也存在一些极端值，尤其是在 5 以上的区域，尽管数量不多，但这些数据点可能代表了创新得分特别高的企业。在与其他变量的关系上，Innovation_D 与 Trans 之间的关系在散点图中显示出一定的分散性，这表明转型程度与创新得分之间可能没有直接的线性关系。与 HHI 的关系则显示出，在市场垄断程度较高（HHI 接近 1）时，企业的创新得分相对集中，这可能意味着在竞争较少的市场中，企业的创新活动更为一致。而与 KZ 指数的关系则表明，创新指数较高的企业在创新得分上也表现出一定的集中性，这可能反映了创新能力与创新活动之间的正相关性。

在散点图中，Innovation_D 与 Trans 的关系显示出，在 Trans 值较低时，Innovation_D 的值分布较为分散，这可能意味着在转型程度较低的企业中，创新得分的差异较大。而在 Trans 值较高时，Innovation_D 的值趋于集中，这可能表明转型程度较高的企业在创新得分上表现出一定的一致性。与 HHI 的关系则显示出，在市场竞争较为激烈的环境下（HHI 值较低），企业的创新得分差异较大，而在垄断或寡头市场下（HHI 值接近 1），创新得分相对一致，这可能反映了市

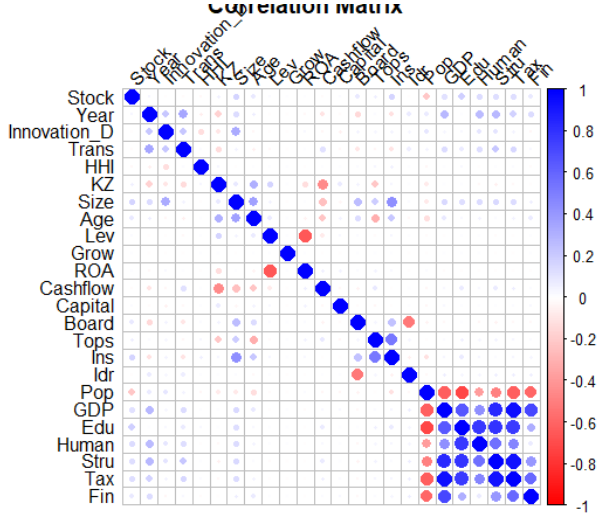
场结构对企业创新活动的影响。与 KZ 指数的关系则进一步证实了创新指数与创新得分之间的正相关性，即创新指数较高的企业在创新得分上也表现出较高的一致性。这些散点图提供了对 Innovation_D 与其他变量关系的直观理解，尽管没有明显的线性关系，但可以观察到在特定条件下创新得分的分布特征。



图表 10 描述性统计

相关性分析

在相关性分析图中，颜色和圆圈的大小代表了相关性的强度和方向，其中蓝色表示正相关，红色表示负相关，颜色越深表示相关性越强。从图中可以看出，“Stock”与“Year”之间有很强的正相关性，这可能意味着随着年份的增加，股票的表现也在增加。“Innovation_D”与“Trans”、“HHI”和“KZ”之间也显示出一定的正相关性，这可能表明创新度、转型程度和市场集中度之间存在某种联系，初步验证了研究假说。总的来说，这张图提供了一个快速查看多个变量之间关系的视角，但是具体的相关性强度和意义还需要结合具体的业务背景和数据进行深入分析。



图表 11 相关性分析

基准模型结果

基准回归结果显示，固定效应模型在解释颠覆式创新（Innovation_D）方面具有一定的统计显著性，F 统计量为 61.1243，对应的 p 值远小于 0.05，表明模型中至少有一个解释变量对创新有显著影响。然而，模型的 R 平方值为 0.050966，说明模型仅解释了创新变异的 5.1%，而调整后的 R 平方为-0.095742，这可能暗示模型可能存在过度拟合或多重共线性问题。在各个解释变量中，数字化转型（Trans）、市场集中度（HHI）、公司规模（size）、杠杆率（Lev）、资产收益率（ROA）、董事会规模（Board）、高管人数（Tops）、机构持股比例（Ins）和人力资本（Human）对创新有显著影响，其中大部分变量的 p 值小于 0.05，显示出正向影响，而现金流（Cashflow）、地方财政收入（Tax）和地方财政支出（Fin）则未表现出显著性。值得注意的是，虽然财务约束 KZ 指数（KZ）和人口（Pop）对创新的影响为负，但 KZ 的影响在统计上是显著的。整体而言，虽然模型中的一些变量对创新有显著影响，但模型的解释力有限，可能需要进一步的变量选择和模型调整来提高其预测创新的准确性。

```
Twoways effects Within Model

Call:
plm(formula = Innovation_D ~ Trans + HHI + KZ + Size + Age +
    Lev + Grow + ROA + Cashflow + Capital + Board + Tops + Ins +
    Idr + Pop + GDP + Human + Stru + Tax + Fin, data = panel,
    effect = "twoways", model = "within")

Unbalanced Panel: n = 3490, T = 1-11, N = 26284

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-4.64769 -0.39145  0.00000  0.39947  4.58753

Coefficients:
            Estimate  Std. Error t-value  Pr(>|t|)
Trans     5.2052e-02  7.6069e-03  6.8427 7.970e-12 ***
HHI       1.0600e-01  6.0583e-02  1.7497 0.0801770 .
KZ       -1.4708e-02  4.1563e-03 -3.5388 0.0004028 ***
Size      3.5983e-01  1.2237e-02 29.4054 < 2.2e-16 ***
Age       2.7429e-02  1.9606e-02  1.3990 0.1618217
Lev       1.1435e-02  5.5116e-03  2.0747 0.0380291 *
Grow      7.9690e-07  5.8489e-06  0.1362 0.8916274
ROA       6.3302e-02  1.5062e-02  4.2026 2.648e-05 ***
Cashflow -4.9728e-02  6.7303e-02 -0.7389 0.4599941
Capital   2.9485e-04  3.0132e-04  0.9785 0.3278290
Board     1.3913e-01  5.6652e-02  2.4558 0.0140652 *
Tops     -2.8288e-03  9.9480e-04 -2.8436 0.0044644 **
Ins       2.4962e-03  7.0262e-04  3.5527 0.0003820 ***
Idr       1.0207e-03  1.6184e-03  0.6307 0.5282514
Pop      -2.2993e-01  2.2269e-01 -1.0325 0.3018278
GDP      -4.3780e-03  1.7285e-02 -0.2533 0.8000535
Human    -1.6028e-03  5.9564e-04 -2.6909 0.0071320 **
Stru     -9.8671e-03  4.6407e-02 -0.2126 0.8316235
Tax       4.8075e-06  1.1780e-05  0.4081 0.6832042
Fin      -8.7885e-06  3.2595e-06 -2.6963 0.0070163 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    13447
Residual Sum of Squares: 12762
R-Squared:      0.050966
Adj. R-Squared: -0.095742
F-statistic: 61.1243 on 20 and 22764 DF, p-value: < 2.22e-16
```

图表 12 基准模型估计结果

数据分析结果

弹性网模型

弹性网回归模型的结果揭示了各个预测变量对颠覆式创新创新（Innovation_D）的影响程度。独立模型中的正系数，如 Trans（数字化转型）、Size（规模）、ROA（资产回报率）、Idr（独立董事占比）、Human（人力资本）、Stru（产业结构）、Tax（财政收入）表明这些因素与创新呈现正相关，即随着这些因素的增加，企业的创新能力也倾向于增强。相反，HHI（行业集中度）、KZ（财务约束）、Age（企业年龄）、Capital（资本密集度）、Tops（股权集中度）显示出负系数，意味着这些因素与创新负相关，可能暗示着随着行业集中度的提高、企业年龄的增长、资本密集程度的增加、股权集中的提高，企业的创新能力可能

会下降。值得注意的是，Lev（杠杆率）、Grow（成长性）、Cashflow（现金资产占比）、Board（董事会规模）、Ins（机构持股）、GDP（地区经济发展）、Edu（教育水平）这些变量的系数未被估计，可能是因为它们与其他变量高度相关或在模型中被排除，这里一定程度上表明弹性网回归具有克服多重共线性缺陷的功能，而且回归系数支持假说。

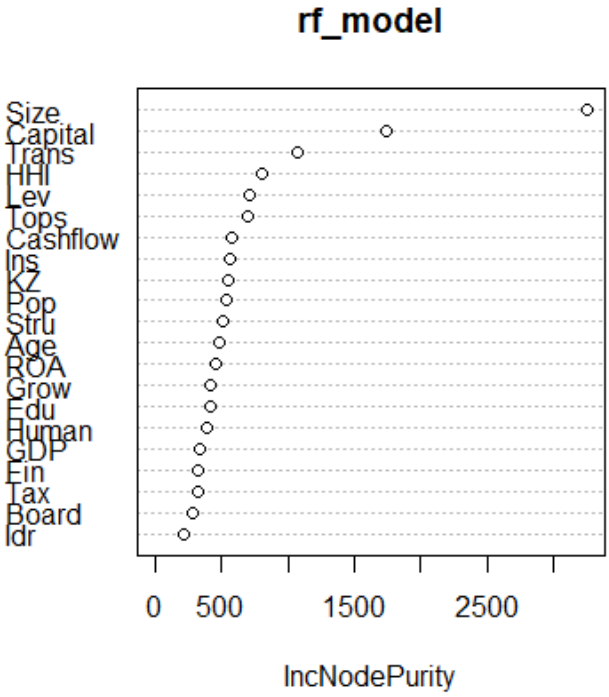
```
23 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -1.016436e+01
Stock       -1.121419e-07
Trans        1.785272e-01
HHI         -8.537891e-01
KZ          -3.655404e-02
Size         4.471667e-01
Age         -2.451683e-01
Lev          .
Grow         .
ROA          1.356237e-02
Cashflow     .
Capital     -1.398915e-03
Board        .
Tops        -8.038207e-03
Ins          .
Idr          1.539548e-03
Pop          2.953846e-01
GDP          .
Edu          .
Human        1.852247e-03
Stru         6.745561e-02
Tax          .
Fin          .
```

图表 13 弹性网模型估计结果

随机森林模型

随机森林模型的估计结果显示，Size（企业规模）在所有变量中对模型预测创新（Innovation_D）的贡献最大，其增加节点纯度（IncNodePurity）的值最高，表明规模较大的企业可能在创新方面表现更为突出。紧随其后的是 Capital（资本密集度）、Trans（数字化转型）、Lev（杠杆率）、Tops（高管人数）和 Cashflow（现金资产占比），这些变量也对模型的预测能力有显著影响，而 Tax（财政收入）、Board（董事会规模）和 Idr（机构持股比例）等变量的重要性相对较低，对模型的贡献较小。这表明在预测企业创新能力时，企业的规模和财务状况是关键

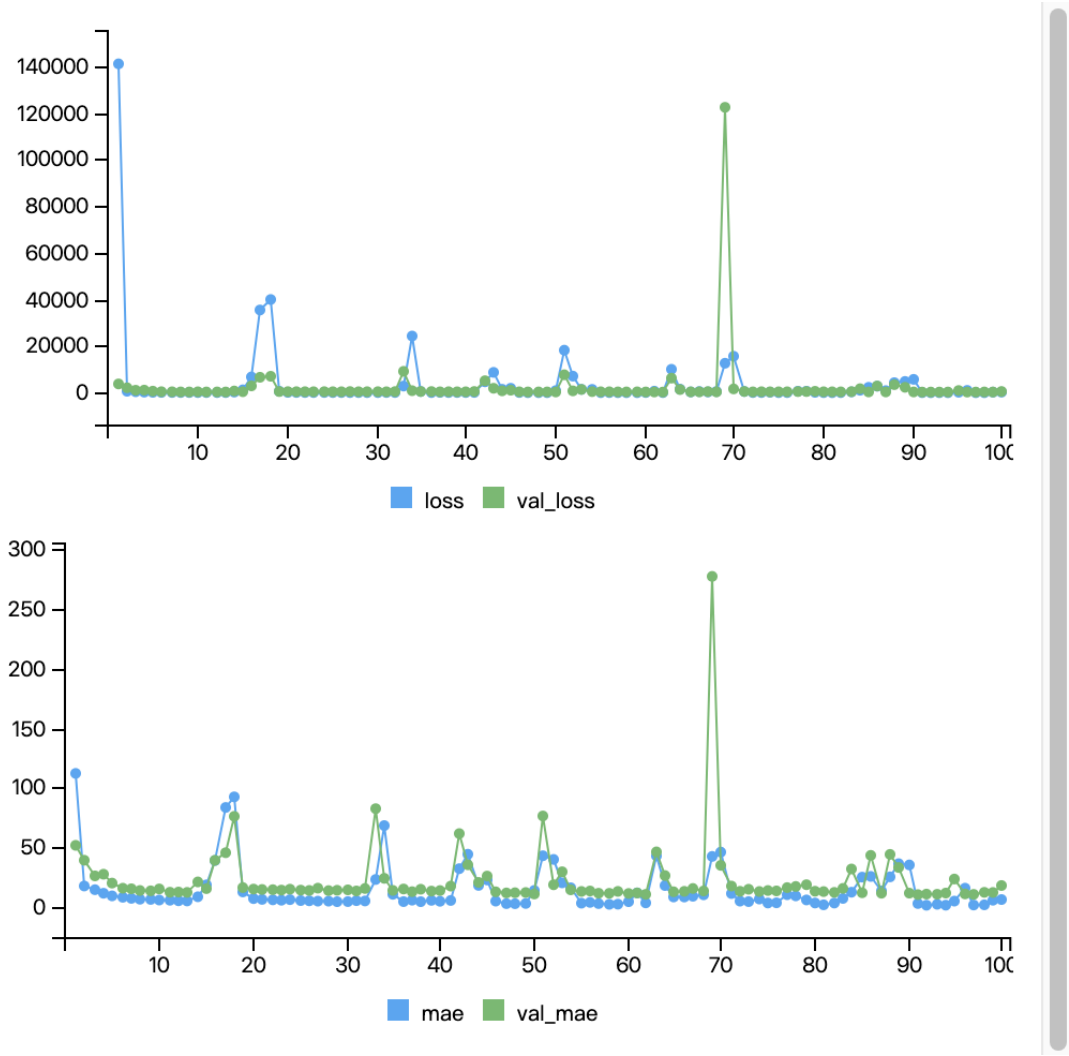
因素，而税率、董事会规模和机构持股比例的影响则不那么显著。整体而言，随机森林模型通过评估变量在节点分裂中对纯度提升的贡献，为我们提供了对企业创新影响因素的深入理解，有助于识别和优化影响创新的关键驱动力。



图表 14 随机森林模型估计结果

神经网络模型

本文神经网络模型由两层组成，第一层为包含 64 个单元的隐藏层，使用 ReLU 激活函数，第二层为 1 个单元的输出层，使用线性激活函数，模型总共有 1537 个可训练参数。从训练和验证的损失（loss 和 val_loss）曲线可以观察到，模型的训练过程存在一定程度的波动，尤其在某些时期（如第 70 轮附近），验证损失突然显著增高，表明模型在这部分数据上可能存在过拟合或数据不平衡现象。此外，验证集的平均绝对误差（val_mae）与训练集的平均绝对误差（mae）也表现出类似的波动趋势，进一步反映了模型的稳定性问题。未来将针对训练数据和模型结构进一步优化，例如加入正则化项或调整学习率，或在数据上进行预处理来改善训练效果。



图表 15 神经网络训练结果

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 64)                1472
 dense_7 (Dense)             (None, 1)                 65
=================================================================
Total params: 1537 (6.00 KB)
Trainable params: 1537 (6.00 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

图表 16 神经网络估计结果

结论与建议

本文通过对上市公司企业数据的市政分析，重点关注了企业数字化转型、市场集中度以及企业融资约束等多维度的数据特征。分析发现，创新能力的表现存在较大差异，可能与资源获取、政策支持和技术应用的差距有关。数据分析表明，企业在资源利用和产出间的关系存在改进空间，同时，不同领域的数据协同效应不足，限制了整体经济效率的提升。

基于数据分析结果，建议政策制定者重点优化资源配置机制，推动资源的精准匹配，尤其在低效率领域提高政策支持的针对性。此外，应加大对创新活动的扶持力度，通过科技投入、金融支持和人才激励等手段，提升个体的创新能力和成果转化率。建议推动数字化转型，利用数据技术实现资源利用的动态监控和智能调配，提高经济活动的整体效率。同时，加强跨领域协作机制，促进不同领域间资源和数据的高效协同，全面提升经济和社会发展的综合效益。