光华管理学院 Guanghua School of Management

因思想 而光华

现代预测观点与评价
Forecasting Perspectives and Evaluation

李丰

北京大学光华管理学院

https://feng.li/forecasting-with-ai

从点预测到概率预测

From point forecast to probabilistic forecast

预测值是随机变量

- 我们试图预测的东西 y_t 是**未知的**，我们可以把它想象成一个**随机变量 (random variable)**
- 例如，下个月的总销售额可能会有一系列的可能值，直到月底我们把实际销售额加起来，我们才知道这个值会是多少。所以在我们知道下个月的销售情况之前，这是一个随机的变量。

- **科学上：一切未知的情形都可以用随机变量来描述。**

- 我们用 y_t 表示时间 t 对应的观察值。假设将观察到的所有信息表示为 I，我们将 $y_t|I$ 表示为“**给定已知 I 情况下的随机变量 y_t**”。这个随机变量 $y_t|I$ **所有可能的取值**就构成了**预测分布**

- **讨论：在你的预测场景里面有哪些已知 I？**

点预测 (point forecast)

- 每当我们谈到“预测”时，通常指的是**点预测**，即**预测分布的平均值**，一般用 $\hat{y}_t$来表示。^读作 hat，是一顶魔术帽。
- **特点（魔术师“手的魔力”）**
  - **直观清晰：**只输出一个数字。
  - **易于比较与评估。**
  - 当不确定性影响较小时，一个数足够做决策。
- **局限（隐藏在魔术帽里）**
  - 仅给一个结果，却不告诉你“它可能错多少”
  - 没有预测分布信息，决策者无法评估“最坏情况”
  - 无法表达“有 90% 把握销量在某范围内”这样的度量。

概率预测 (probabilistic forecast)

- 现在魔术师变聪明了，他不只掏出一个数，而是展示帽子里的“所有可能”。

- **那么概率预测就是他从帽子里掏出一整个分布**
  - 告诉你“明天销量在 1000 ~ 1500 的概率是 80%，超过 1600 的概率是 10%”。

- 点预测 = “确定性思维”
  - 适合短期、低风险、稳定环境。

- 概率预测 = “不确定性思维”
  - 适合复杂系统、风险控制、资源配置决策。

需求与库存管理中的概率预测

- **商业场景：**零售商、电商平台或制造企业在做库存决策时，如果只根据点预测的销量安排生产或补货，很容易出现“缺货”或“积压”。

- **需求与库存管理**（Demand & Inventory Management）
  - 概率预测提供未来需求的分布（如 90% 置信区间），使得企业能设定安全库存水平（safety stock）。
  - **可计算“缺货风险概率”或“过量库存概率”**。
  - **常见指标**：Service Level, Expected Shortage, Expected Holding Cost。
  - **示例：**Amazon 的自动库存系统使用概率需求预测来动态设定补货阈值。

能源的概率预测

- **场景：**电力公司预测明日负荷或风电产出时，误差会直接影响系统稳定性和成本。
- 概率预测提供负荷/发电的概率分布，支持**稳健调度**与**备用容量配置**。
- 能源交易市场使用预测分布计算“**风险调整价格**”。

数据来源：国务院新闻办 “十四五”时期能源高质量发展成就新闻发布会
http://www.scio.gov.cn/live/2025/37113/index.html

现代预测观点：从点预测转向概率预测

- **技术与方法层面制约**：模型实现复杂、评价体系不是显而易见
- **认知与文化层面制约**：决策者偏好“确定答案”
  - 销售经理宁愿听“明天卖1200件”，而不是“有70%概率卖1000－1500件”。
- **管理决策体系制约：**很多企业或政府仍是“计划导向”（planning-based）而非“风险导向”。
  - “本着上述方针，一九八一年国民经济计划主要指标安排如下：工农业总产值，比一九八〇年预计增长百分之五点五。其中，工业总产值增长百分之六；农业总产值增长百分之四。”—摘自1981年《政府工作报告》
  - “今年发展主要预期目标是：国内生产总值增长5%左右。”—摘自2025年《政府工作报告》
  - 欧洲中央银行 Survey of Professional Forecasters:
    - 报告了点预测
    - 预测分布的中位数
    - 多个概率预测的聚合均值

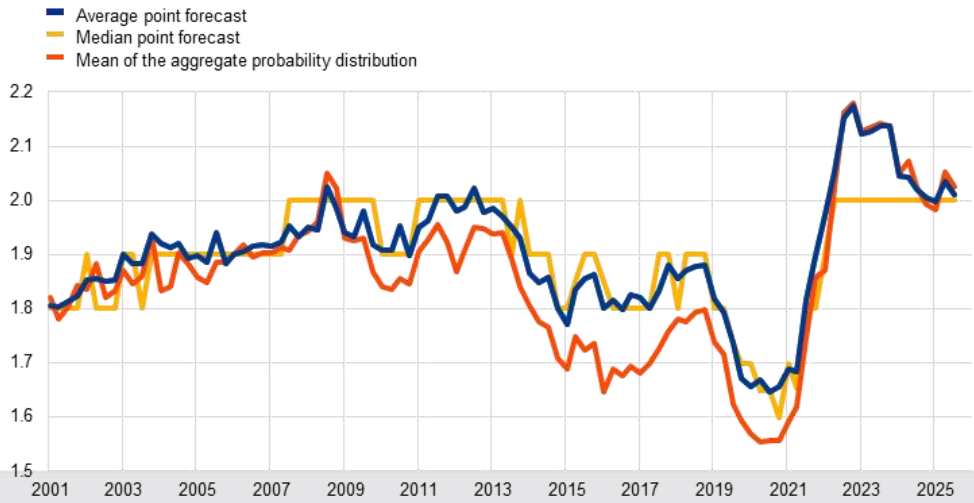


Figure source: https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/ecb.spf2025q3.en.html

预测的科学评价

Forecast Evaluation

点预测的评价准则

- 点预测给出的是一个确定值 $\hat{y}_t$，我们希望它尽量接近**真实值** y_t。**预测误差**定义为：

$$e_t = y_t - \hat{y}_t.$$

- 评价指标的核心目的：衡量**预测误差的大小**、**方向**与**稳定性**。
- 这里我们先假设**真实值** y_t是能已知的。

平均绝对误差（MAE, Mean Absolute Error）

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|$$

- **含义：** 平均每次预测偏离了多少。
- **类比：** 像一个“平均距离计”：每次你射箭，计算箭头离靶心的平均距离。
- **特点：**
  - 对所有误差一视同仁；
  - 不夸大极端值的影响；
  - 适合对称损失场景。
- 库存与需求预测
  - 在电商、零售场景中，预测销量偏高或偏低的成本相似。
  - MAE能反映“平均偏差量”，例如平均每天少/多预测了多少件商品。

平均偏差与平均百分比误差

平均偏差（ME, Mean Error）与平均百分比误差（MPE）分别定义为

$$\mathrm{ME} = \frac{1}{T}\sum e_t, \qquad \mathrm{MPE} = \frac{100\%}{T}\sum \frac{e_t}{y_t}$$

- **含义：** 衡量预测的系统性偏差（bias）。
- **类比：** 像“风向标”：正表示高估，负表示低估。
- **作用：** 检查预测整体偏向。
- 适用场景：
  - 供应链与产能规划：判断预测是否长期高估（导致库存积压）或低估（导致缺货）。
  - 宏观经济模型评估：检查模型是否存在系统性乐观或悲观偏差。
  - 能源调度系统：判断预测是否持续低估发电负荷，帮助修正偏差方向。

均方误差与均方根误差（MSE / RMSE）

$$\mathrm{MSE} = \frac{1}{T}\sum e_t^2, \qquad \mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

- **含义：** 惩罚大误差更重，因为平方会放大错误的比重
- **类比：** 像“射箭罚分系统”：离靶心越远，罚分指数级上升。
- **特点：**
  - 强调极端误差
  - 常用于科学、工程类预测
  - 对异常值敏感
- **适用场景：**
  - 能源负荷与电力需求预测：大偏差会造成电网调度风险或备用成本暴涨，因此需要惩罚大误差。
  - 气象与温度预测：极端预测错误（例如温度差10℃）对实际影响巨大。

平均绝对百分比误差 (MAPE, Mean Absolute Percentage Error)

$$\text{MAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

- **含义：** 衡量预测偏离相对真实值的比例。
- **类比：** 像“体重误差的百分比报告”：预测70kg，实际72kg，误差约2.8%。
- **特点：** 易解释，但不能处理 $y_t = 0$，对小基数样本敏感。
- **适用场景：**
  - 财务与销售预测：以百分比形式衡量预测偏差，方便管理层理解：“预测误差控制在5%以内”。
  - 宏观经济指标预测（如GDP、通胀率）：数值量纲不同，用相对误差衡量更公平。
  - 不同规模的预测对象可在同一标准下比较精度。

对称平均绝对百分比误差（sMAPE, Symmetric MAPE）

$$\text{sMAPE} = \frac{100\%}{T}\sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2}$$

- **含义：** 使“高估”和“低估”影响对称。
- **类比：** “双方折中比例误差”。类似高考标准化分数，让双方误差影响相等。

- **适用场景：**
  - 电商销售预测：预测值和真实值都可能小，sMAPE在销量低时比MAPE稳定。
  - 短期时间序列预测：对高低波动敏感度对称，常用于竞赛和基准评估（如M4、M5竞赛）。
  - 中低量级行业数据（如机场客流、医院门诊量）
  - 适合非零但波动明显的序列。

比例误差类指标（Scale-Free Metrics）

MASE（Mean Absolute Scaled Error）：

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\mathrm{MAE}_{\mathrm{naive}}}$$

- **类比：** 像“预测能力得分比”：你的模型比“瞎猜”好多少。若 $\mathrm{MASE} < 1$，表示优于基准预测。MASE就像“你的预测成绩相对于全班平均水平的倍数”。
- **适用场景：**
  - 多产品、多地区的预测基准评估
  - 不同序列单位不同（如吨、件、元），MASE能统一比较预测性能。

小结：点预测的好坏就像打靶的表现

- MAE 看平均离靶距离
- RMSE 惩罚“飞得最远的箭”
- ME 看整体偏左还是偏右
- MAPE 告诉你偏离比例
- MASE 看你比“闭眼射箭”的人好多少

没有真实值 y_t的预测评价

Forecast evaluation in the future horizon

预测评价中“无真实值”的本质难题

- 在真实世界的预测应用中，我们面临一个根本问题：**未来尚未发生，因此我们无法知道真实值 y_t**。
- 所有的时间序列预测都是在“未知的未来”上进行的外推。
- **时间序列预测评价与机器学习评价存在本质不同**
  - 机器学习的**训练集**与**测试集**都来自同一个静态数据生成过程
  - **机器学习测试集中的“真实标签”是已知的**，因此可以通过精确的误差度量评估模型性能。
  - 机器学习的测试集是“看得见的未来”，而**时间序列预测的目标是“尚未发生的未来”。**

- 预测问题中，真正关心的是下一期、下一个季度、下一年的值，这些值在建模与验证时都尚不存在。

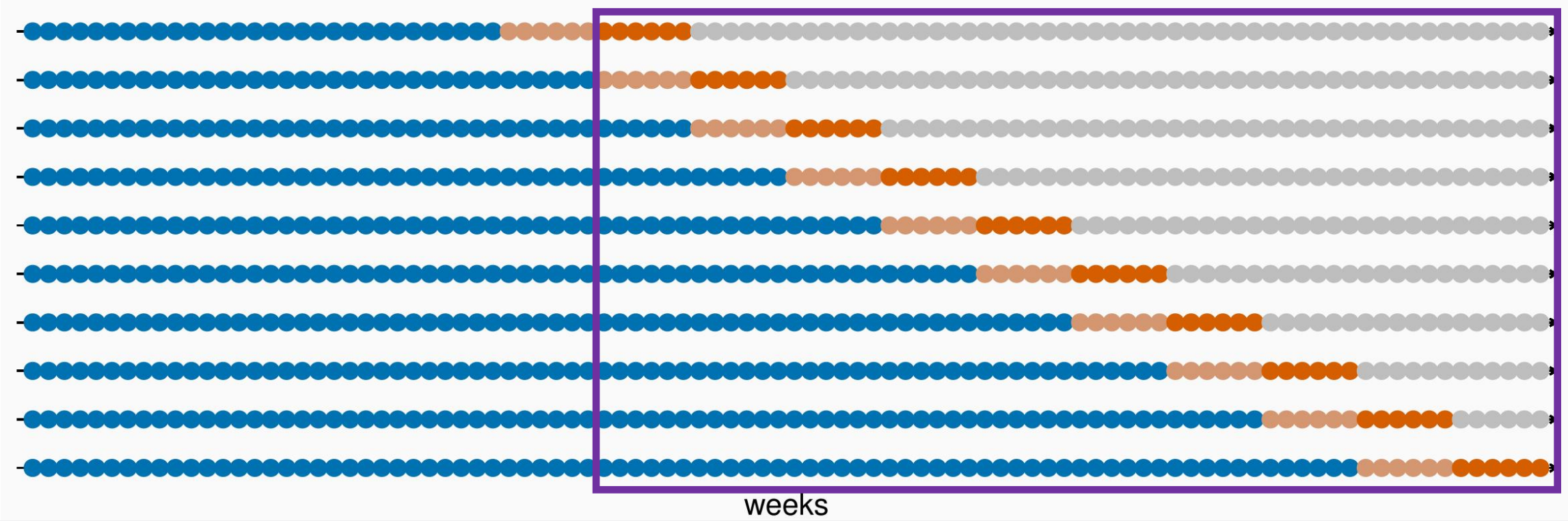
- 我们怎么办？评价模型依赖“**过去的未来**”——即已经发生的历史样本的未来。
- 如果一个预测方法在过去的未来中表现不错，那大概在真的未来中也可以。

时间序列交互验证(time series cross-validation)

- 时间序列交互验证
  - 我们不能随意打乱样本顺序进行随机分割（这会破坏时间依赖性）
  - 因此采用时间滚动式（rolling or expanding window）交叉验证
  - 在每一折（fold）中，用早期的观测值作为训练集（the past of the past）
  - 用紧随其后的时间段作为测试集（the past of the future）
  - 模型在每一折上滚动更新并重新评估。
- **把过去的过去当作训练集，过去的未来当作测试集的**

十折交叉验证设计

- **十折交叉验证**(Ten-fold cross validation)：整个时间序列被分为十个连续时间片，每一折都向前滚动一次。
- **预测区间**(Forecast horizon)：1–60 天，即模型要预测未来最长60天的结果(**紫色区间**)。
- **训练窗口**(Training window)：**蓝色部分**不断扩展，每一折的训练集比前一折多包含额外的 6 天数据。
- **测试窗口**(Testing window)：**淡黄色区域**为测试区间，可以计算各类预测误差
- **评估窗口**(Evaluation window)：**橙色区域**为评估区间，对应战略或规划层面的预测期。



案例实践

- 请结合提供的代码测试模型的预测准确度