基于机器学习的公路工程概算造价预测模型

古 杰

（广东省高速公路有限公司项目部，广东 广州 510000）

摘 要 本研究提出了一种利用机器学习进行工程概算造价预测的方法，运用XGBoost算法建立概算预测体系。应用加法模型以及前向逐步求解选取重要变量，依靠梯度推进技术及损失函数实现非线性拟合。实证分析表明，该模型在RMSE、MAE及R^2等评价指标上均优于BP神经网络和K-Means聚类方法，具有较高的准确性和稳定性，证明了机器学习在预算管理中的实际应用价值。

关键词 机器学习；工程概算；造价预测；XGBoost算法；预算管理

中图分类号：TP242；U412 **文献标志码**：A DOI：10.3969/j.issn.2097-3365.2025.28.025

0 引言

在当前全球市场环境日益复杂多变的背景下，工程材料价格的不稳定性已经成为影响公路工程造价管理的一个重要因素[1]。材料费用占据了公路工程项目成本的主要部分，直接关系到工程的投资规模及经济效益[2]。宏观经济的不确定性加大以后，对材料价格变化做出及时而精确的判断就成为工程预算领域的一个迫切问题[3]。以往研究主要集中于搜集市场价格数据，预测未来走势。但针对市场价格高度波动的情况，传统造价预测手段遭受严峻挑战[4]。

本文建立以XGBoost算法为基础的机器学习方法，将历史材料价格变化与工程参数结合，建立成本预测模型。利用损失函数进行成本建模，弥补传统方法对复杂数据预测精准度不足的问题，增强机器学习的应用价值。

1 研究方法

1.1 XGBoost模型

XGBoost基于Boosting框架，通过集成多个决策树模型来提升预测精度，是一种结合加法模型与前向逐步优化策略的机器学习方法[5]。在训练过程不断迭代，每次生成一个新的树模型，用以修正前一轮模型未能准确拟合的部分，即对残差进行学习。整个过程可看作是不断叠加多个子模型，从而构建出一个强大的梯度提升树系统，形式如下：

$$f_M(x)=\sum_{m=1}^{M}T(x,\theta_m) \tag{1}$$

式（1）中，M表示所使用的决策树总数；$T(x,\theta_m)$表示第m颗树模型，θ_m是这棵树的参数集合。如果将初始模型$f(x)$设为0，那么第m棵树迭代时所得结果就是在此基础上的增量更新。为了确定每次迭代中的参数θ_m，需要通过最小化一个损失函数θ'_m来进行优化，形式如下：

$$\theta'_m=\arg\min_{\theta_m}\sum_{i=1}^{M}L\left[y_i,f_{m-1}(x_i)+T(x_i;\theta_m)\right] \tag{2}$$

式（2）中，x_i表示第i个输入特征，y_i是对应的真实值，L代表所使用的损失函数。通过前向逐步优化的方法构建出M棵决策树$T(x,\theta_m)$后，将这些树依次叠加，形成最终的梯度提升模型$f_M(x)$。

假设在梯度提升模型中，经过多轮迭代公式生成了第k颗树，此时模型的预测函数f_k已经建立，其对应的预测值$\hat{y}_i$可以用如下关系式表示：

$$\hat{y}_i=\varphi(x_i)=\sum_{k=1}^{K}f_k(x_i),f_k\in F \tag{3}$$

式（3）中，F表示有所有回归树组合的函数集合，其中每个$f(x)=w_{q(x)}q:R^m\to T,w\in R^T$，$q(x)$是一个将输入$x$映射到某颗决策树中某个叶子节点的函数；$R^m$表示输入空间被划分为的若干区域；$R^T$是对应$T$个叶节点的输入空间；$T$代表树中叶子的总数；$w$则是每个叶子节点对应的权重。在模型训练过程中，每轮生成一颗新树，用来拟合前一轮预测值与真实值之间的偏差，这个偏差即为模型当前的误差，通过定义一个损失函数$L(\phi)$来度量：

$$L(\phi)=l(\hat{y}_i,y_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{4}$$

式（4）中，$l(\hat{y}_i,y_i)$表示模型预测值与实际值之间的误差，而$\Omega(f_k)$是一个正则化项，用于控制模型复杂度。假设当模型迭代到第t棵树，并设初始模型$f_0(x_i)=\hat{y}_i^{(0)}=0$，那么第$t$次迭代后的预测结果$\hat{y}_i^{(t)}$就是基于前几轮模型的累加结果。此时，整个目标函数$Obj$累积到第$t$轮的表达形式如下：

$$Obj^{(t)}=\sum_{k=1}^{T}l\left[y_i,\hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right]+\Omega\left(f_k\right)+const \tag{5}$$

当模型迭代到第 t 棵树时，可以将公式中的常数部分省略，从而得到第 t 次迭代所需优化的统一目标函数表达式。

XGBoost 算法有良好的正则化效果，能有效地防止模型出现过拟合现象[6]。利用二阶导数值的加入使损失函数的计算更加精确，达到优化目标函数的目的。而且此算法可进行并行计算，极大地提高了运算速度。

1.2 建模流程

XGBoost 是一种增强型的梯度提升算法，将目标函数转换成一个二次规划问题，进而提高模型预测的准确性[7]。建模过程如下：

1. 数据预处理阶段。应用 SPSS 及 Python 软件对所选取的预测变量做多重共线性检验并做标准化，删除异常值。

2. 样本划分。预处理结束后，将数据集分割成训练集以及测试集两部分。

3. 模型调参。运用网格搜索配合交叉验证方法来优化 XGBoost 模型的参数组合，确定出训练集上效果最佳参数配置。

4. 模型评估与验证。选用决定系数（RE）、平均绝对误差（MAE）、均方误差（MSE）及均方根误差（RMSE）这些指标来评判预测的结果。并且从数据中随机选取一部分做样本，再做一次检验。

本文为了对模型的预测能力做出全面评价，选取了上述的各项误差评价指标来检验所建模型在实际工程估算中应用的可行性及泛化能力，具体公式为：

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{6}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_i\right| \tag{7}$$

$$R^2=1-\frac{\sum\left(\hat{y}_i-y_i\right)^2}{\sum\left(\overline{y}_i-y_i\right)^2} \tag{8}$$

$$MSE=\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2 \tag{9}$$

1.3 数据集选取

为了简化模型结构并提高预测效果，结合 XGBoost 损失方法选择频率超过 85% 的关键因子，作为模型的输入变量。最终选定了八个对造价影响显著的变量。其中，路基宽度、路面厚度、设计车速和填挖方体积属于定量变量，可直接用于建模；其余为定性变量，需进行数值化处理；而基础类型、结构形式和外立面材料为定性变量，需进行数值化处理后才能输入模型。表 1 为参数的赋值转换，包含变量的原始类别及对应的编码方式。

表 1 参数的赋值转换

定性参数	参数分类	赋值后变量
基础类型	土质路基、碎石路基、填石路基、软土地基	1、2、3、4
结构形式	沥青混凝土路面、水泥混凝土路面、复合式路面	1、2、3
外立面材料	波形钢护栏、混凝土护栏、缆索护栏	1、2、3

模型变量的统计特征如表 2 所示，本研究共整理了研究项目的 40 组样本数据，构建了响应的数据库，用作后续模型训练与验证的数据集。

表 2 模型变量的统计特征

变量	最小值	最大值	平均值	标准差	变量缩写
沥青单价（元 /㎡）	800	3 200	1 850	620	AP
路基填挖方量（㎡）	500	12 000	4 500	2 700	EV
高度	2.5	18	6.2	3.1	H
数量	1	8	3.4	1.6	Q
宽度	3.0	12.0	6.5	2.2	W
基础类型	1	4	2.3	0.9	FT
结构形式	1	3	1.8	0.7	ST
抗震烈度	6	9	7.2	0.8	EQ
外立面材料	1	3	2.1	0.6	WM

2 模型预测对比

2.1 项目情况简介

研究数据来源于纵横造价、广联达造价指标网及广东省交通运输工程造价信息平台公开的信息，选取了从 2015 年到 2024 年广东省内 200 多个公路工程新建及改造实例。为了让模型预测更准确，所有样本都经过 XGBoost 算法处理，只选取构造简单且改造标准一致的公路工程，将样本数据依照 7:3 的比例分成训练集和测试集。BP 神经网络以及 K-Means 聚类方法也是采取同样的方式来进行训练测试。建模和参数调整利用 Python 中的 XGBoost 库，并运用网格搜索法对模型进行优化。

2.2 模型对比

表 3 为不同预测模型指标评价，RMSE 以及 MAE 数值越小，说明预测值与真实值间的差距就越小；R2 值

越趋近于 1，就证明模型对造价变化的拟合度越高。可见，XGBoost 模型在上述三个指标均取得优异表现，体现了该模型的稳定性和精确性。

表 3　不同预测模型指标评价

模型	RMSE	MSE	MAE	R^2
BP 神经网络	2.31	2.15	1.43	0.72
K-Means 聚类	1.46	1.21	0.85	0.89
XGBoost	0.97	1.03	0.76	0.91

为确保模型间备可比性，本文在使用 BP 神经网络和 K-Means 聚类进行预测时，采用与 XGBoost 模型相同的样本序列划分方式，不再进行随机抽样分组。在设置训练初始参数时，最大训练轮数为 1 000，学习率为 0.1，收敛精度阈值为 0.5。图 1 为模型的预测结果与实际值差异，可见，XGBoost 对公路工程造价的预测相对误差基本控制在 ±15% 范围以内，而 BP 神经网络和 K-Means 聚类在 ±25% 和 ±30% 左右，XGBoost 预测精度更加满足实际需求。

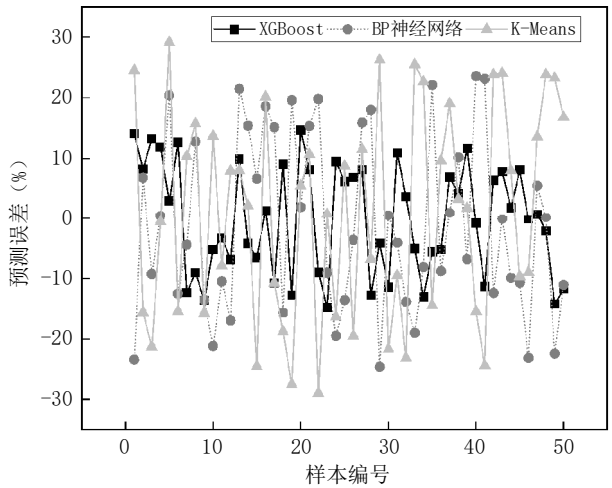

图 1　模型的预测结果与实际值差异

2.3　特征值排序

本文对主要变量做了特征重要性排序来辨识各种模型特征对造价预测的影响程度，图 2 为不同特征的重要性排名。沥青单价（AP）与路基填挖方量（EV）是所有变量里得分最高的两个，其重要性值分别是 0.297 与 0.214，这就证明二者是对公路工程概算造价起决定性作用的主要因素。抗震烈度（EQ）0.028 以及外立面材料（WM）0.018 尽管排位较后，但是也在某些特殊情况下对预算结果有辅助的作用，尤其是在施工周期复杂或是外立面设计等级较高的时候，潜在的成本变动是值得重视的。从模型变量重要性的分布来看，主要材料价格和施工规模类指标对工程造价的影响最为显著。这一排序对于工程项目的初步预算、设计方案的改进以及预算管理具有指导意义。

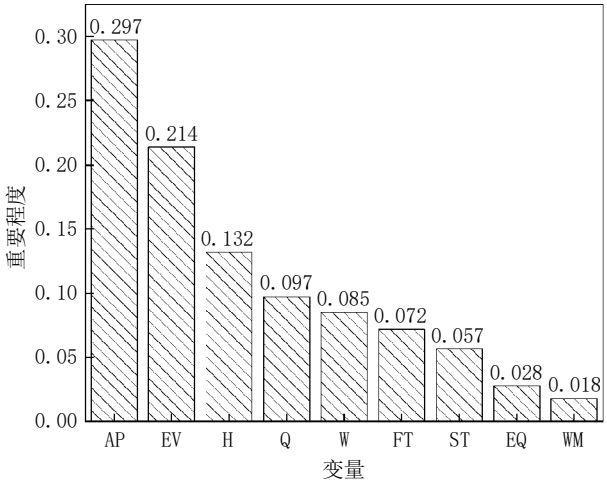

图 2　不同特征的重要性排名

3　结论

本文建立了一个以 XGBoost 为基础的机器学习预测体系，该模型于各项评价标准之中皆有优异表现，尤其是在 RMSE、MAE 及 R2 这几项指标上，XGBoost 算法较 BP 神经网络算法和 K-Means 聚类分析有着显著的优势。而且 XGBoost 算法预测的相对误差也较小，基本维持在 ±15% 之内。另外，利用特征值的重要性排序得出沥青单价（AP）与路基填挖方量（EV）是影响造价的主要变量。未来研究应当考虑加入动态政策、宏观经济环境等要素来改进机器学习模型，从而开发出更富时效性和预警功能的造价管理机制。

参考文献：

[1]　付碧芸 . 建筑材料价格波动背景下工程造价动态调整机制研究 [J]. 四川建材 ,2025,51(05):236-238.
[2]　李月明 . 提高公路工程材料价格信息合理性的相关建议 [J]. 低碳世界 ,2024,14(11):166-168.
[3]　沈艺宏 . 市政供水管道安装工程概算指标与造价预测 [J]. 中华建设 ,2024(08):34-36.
[4]　骆然 , 乔曙 , 王禹超 , 等 . 浅谈硬梁包水电工程概算执行情况分析与造价管控 [J]. 四川水力发电 ,2023,42(01):52-56,88.
[5]　郭婧娟 , 刘曜玮 . 基于 XGBoost 的高速公路工程概算预测方法研究 [J]. 公路交通科技 ,2023,40(03):58-68.
[6]　周波 , 刘云 , 李维嘉 , 等 . 基于融合 XGBoost 的变电工程造价数据预测算法 [J]. 沈阳工业大学学报 ,2025,47(03):317-323.
[7]　李松 , 郭秀伟 , 范克昌 .XGBoost 算法驱动下的建筑工程造价预测模型优化 [J]. 中国建筑金属结构 ,2025,24(11):172-174.