

基于时空行为特征工程与多模型统一评估的即时配送时长预测研究

闫淑凡

(齐鲁工业大学, 山东省济南市, 250353)

摘要: 即时配送履约效率的提升依赖对时长形成机制的精准刻画, 而传统人工规则与单一统计模型难以充分覆盖空间轨迹异质性、骑手个体差异与长尾分布的复杂影响。既有研究多侧重模型精度提升, 对时空特征的增量效应、多模型统一比较标准与误差场景分布的讨论相对不足。本研究基于 31415 条即时配送订单轨迹数据, 构建包含三类 Haversine 球面距离的时空行为特征体系, 在统一训练测试划分与 5 折交叉验证框架下对比 10 类回归模型的预测表现, 并借助 SHAP 方法解析核心影响因素的作用机制。研究发现: 时空行为特征可显著提升基础模型的预测精度, 其增量效应稳定存在于不同模型范式中; 树集成模型整体优于线性模型与神经网络基线, 随机森林取得最优 OOF 表现, RMSE 为 88.07、MAE 为 57.10、 R^2 为 0.6128; 模型误差随配送时长增加而扩大, 在高定位偏差、跨区域履约与高峰时段等场景下更为集中; 收货点到送达点距离、骑手个体差异、时间编码与全链路空间距离是影响时长预测的核心因素。研究表明, 将配送业务过程转化为可解释的时空特征, 结合集成学习与 SHAP 解释框架, 能够同时兼顾预测精度与业务可读性, 为即时配送精细化运营提供方法支撑。

关键词: 即时配送; 配送时长预测; 时空特征工程; Haversine 距离; 随机森林

中图分类号: F252

文献标识码: A

一、问题提出

(一) 即时配送时长预测的现实背景与管理需求

即时配送系统串联用户需求、商家供给、骑手运力与城市空间网络, 是外卖、商超到家、药品零售等本地生活服务的核心履约环节。近年来, 随着即时零售市场规模持续扩张, 平台对末端履约的精细化管理要求不断提升, 配送时长作为最直接的服务质量指标, 既是前端用户预计到达时间 (ETA) 展示的核心依据, 也是后端骑手调度、区域运力监控、异常订单识别与履约效率评估的关键输入^[2]。随着订单系统、GPS 定位系统与 AOI 地理信息系统的持续迭代, 平台积累了海量结构化履约数据, 利用机器学习从历史记录中挖掘时长形成机制, 已成为提升末端配送管理能力的主流路径。

传统履约管理多依赖人工经验规则, 通常根据直线距离、行政区域或固定时间窗设置时长阈值, 方法直观易执行, 但存在明显局限: 一是难以捕捉骑手初始位置、区域交通环境、AOI 内部通行条件等空间异质性的影响, 即使直线距离相近的订单, 实际完成时长也可能存

在显著差异；二是无法有效吸收高基数骑手、经纬度空间分区与异常长尾订单中的信息，对个体履约差异与极端场景的适配性不足；三是全局平均的估算逻辑难以适配日内时段波动、日期差异与区域运力变化带来的时长动态调整需求。随着配送订单密度提升与用户对时效敏感度的增强，传统规则的精度瓶颈日益凸显，亟需更精细的建模方法支撑履约决策。

机器学习模型为配送时长预测提供了更强的非线性拟合能力，随机森林、GBDT、LightGBM 等树集成模型能够通过多层分裂学习变量阈值与交互关系，适配表格化订单数据的复杂结构^[12]。然而现有实践与研究多侧重模型精度提升，对预测结果的业务解释不足，难以将黑箱预测转化为调度、运营人员可理解的决策依据。对于平台运营而言，可解释性与预测性能同样重要：模型不仅要给出时长数值，还要能够说明“哪些因素推高或拉低了时长”，从而支撑运力调配、异常复核与流程优化。因此，构建兼具预测精度与业务可解释性的配送时长建模流程，既具有理论方法价值，也对即时配送平台的精细化运营具有现实意义。

（二）时长形成机制与建模的核心挑战

本文研究任务为连续型配送时长回归，即基于订单属性、骑手信息、地理坐标与时间字段，预测骑手接单到订单送达的分钟级时长。从数据特征与业务机制看，该任务面临多重挑战，也构成了本研究需要回应的核心问题。

第一，目标变量呈显著右偏长尾分布。样本配送时长均值约 203.52 分钟，中位数约 175.00 分钟，少量极端订单时长超过 3000 分钟。长尾样本会显著放大 RMSE 等误差指标，要求模型不仅能拟合常规订单，也要具备识别异常履约场景的能力。单一全局模型往往在均值附近表现较好，但难以覆盖尾部极端样本，容易出现“整体指标尚可、局部误差失控”的问题。

第二，数据质量与变量结构复杂。数据中接单 GPS 经纬度各存在 955 条缺失记录，占总样本约 3.04%，直接影响空间距离特征的计算；同时包含高基数骑手 ID、AOI ID 等类别变量，以及时间字符串、经纬度坐标等多类型字段，增加了特征编码与建模复杂度。不同类型的变量的尺度差异、缺失模式与分布特征，对预处理与模型适配性提出了较高要求。

第三，时长形成具有非线性与多因素耦合特征。配送时长并非单一距离变量的函数，而是空间跨度、时间背景、区域场景、骑手个体与数据质量等多因素共同作用的结果。接单点到收货点距离反映骑手初始接近成本，收货点到送达点距离反映定位偏差与末端通行差异，

骑手 ID 隐含个体效率与路线熟悉度，时间变量捕捉日内与日期波动，各类因素存在复杂交互关系，难以通过全局线性关系刻画。

第四，业务落地对可解释性要求高。配送时长预测不仅要输出数值结果，还要能够支撑运营决策。黑箱模型的精度再高，若无法转化为业务语言，也难以落地应用。运营人员需要知道不同区域、不同骑手、不同时段时长差异来源，从而针对性优化流程，这要求建模过程必须配套可解释性分析，打通技术指标与业务语言的衔接。

（三）既有研究进展与研究缺口

1. 配送时长预测的研究脉络

配送时间建模源于车辆路径与城市物流研究，自经典卡车调度问题提出以来^[4]，车辆路径问题（VRP）始终是物流运作研究的核心议题。传统路径优化研究侧重给定需求与路网下的最优路径求解，关注算法的计算效率与解的最优性^[5]，带时间窗的路径拓展进一步验证了时间约束与空间位置共同决定履约效率的结论^[1]。而即时配送场景引入了高频订单、骑手实时位置与用户等待时间约束，时长预测的应用场景也从路径规划扩展至 ETA 展示、派单约束与异常预警^[3]，研究重心从“优化路径”转向“预测时长”。

既有研究普遍认可空间距离与时间背景是时长预测的基础因素，但仅依赖原始经纬度或距离阈值不足以解释复杂履约差异，骑手个体、区域结构、AOI 类型与定位偏差均会对完成时长产生显著影响。但现有研究多将这些因素作为控制变量纳入，较少从业务过程出发系统拆解配送全链路的时空特征，特征设计与履约流程的对应性有待加强。

2. 预测模型的方法演进

传统统计模型如线性回归、岭回归与贝叶斯线性回归结构清晰、训练高效，可作为基准模型^[7]，但难以表达空间非线性与高阶交互关系，在复杂履约场景下拟合能力有限。支持向量回归基于统计学习理论，通过核函数实现非线性拟合^[10]，但对特征尺度与参数设置敏感，更适配中小规模样本。决策树模型通过递归分裂生成局部分段规则，可解释性较强，但单树方差大、稳定性不足，容易出现过拟合问题^[11]。

集成学习通过组合多个基学习器提升泛化能力，是当前结构化数据建模的主流方法。其中随机森林基于 Bagging 思想，通过样本与特征的双重扰动训练多棵决策树并取平均，有效降低了单树方差^[12]；GBDT、XGBoost、LightGBM 与 CatBoost 基于 Boosting 思想，逐步拟

合残差以修正模型误差，在表格数据任务中表现突出^[13]。对于即时配送数据，树模型无需显式指定交互项即可通过分裂路径学习变量间的局部阈值与组合效应，天然适配空间、时间、类别变量的耦合结构。但现有研究多针对单一模型调优，缺少在统一数据、统一划分、统一指标下的系统性对比，不同模型的适配性差异缺乏清晰结论。

3. 特征工程与可解释性研究

特征工程是结构化建模的核心环节，直接决定模型的性能上限。即时配送场景的常见特征包括 Haversine 直线距离、GPS 缺失标记、时段编码、周末标记与对数变换特征等^[19]，但现有研究多侧重距离类特征构造，对“接单-收货-送达”全链路的空间点位关系拆解不足，较少将定位质量信息显式纳入特征体系，特征与业务过程的对应性有待加强。

可解释机器学习是提升模型落地性的关键支撑，现有方法包括特征重要性、偏依赖图、LIME、SHAP 等多种类型^[21]。其中 SHAP 基于合作博弈中的 Shapley 值思想^[20]，将模型预测拆解为各特征的贡献之和，既能实现全局特征重要性排序，也能开展单样本局部解释，具备坚实的理论基础与良好的一致性。针对树模型的 TreeSHAP 算法进一步提升了计算效率，可高效处理集成模型^[22]，成为表格模型解释的主流方法，相关专著也系统梳理了可解释机器学习的方法体系与应用场景^[24]。但现有配送时长研究中，可解释性分析多停留在特征重要性排名层面，较少结合业务流程解读特征的作用机制与非线性阈值，模型结果与运营决策的衔接不足。

总体而言，现有研究在模型方法上已形成较多积累，但仍存在三点不足：一是特征设计与配送业务过程的对应性不足，空间特征多停留在单点距离层面，全链路时空拆解的研究较少；二是多模型对比多在不同实验设置下开展，缺乏统一划分、统一特征、统一指标的系统性评估，结论可比性较弱；三是可解释性分析与业务决策的衔接不够紧密，误差的场景分布规律讨论不足，模型结果难以直接转化为运营建议。

基于上述研究缺口，本文提出三个核心研究问题：

RQ1：在基础特征之上，时空行为特征能够带来多大的增量预测效度，其作用在不同模型中是否稳定？

RQ2：不同类型的预测模型在即时配送时长任务中的表现有何差异，何种模型更适配高基数、长尾分布的订单数据？

RQ3: 统一预测模型的误差在不同场景下呈现何种分布规律, 核心特征对时长预测的作用机制是什么?

二、研究设计

围绕时空特征的预测增益及统一模型的适用边界, 本研究依次完成数据整理、特征构造、多模型比较与解释分析。首先对原始订单数据进行清洗与预处理, 构造基础属性与时空气象两类特征; 其次在共同基线上比较时空特征带来的样本外预测增益, 再通过多维度误差切片考察统一模型的失准场景; 最后通过 SHAP 分析呈现最优模型的特征依赖与作用机制。

(一) 分析框架与研究问题

本研究从订单属性、骑手个体与区域场景三个层次组织变量。订单提供共同的空间形式与时间线索; 年龄、性别、编号与历史效率记录骑手的个人信息; 区域位置概括片区生活环境与相关交通经验。结果变量为订单从接单到送达的分钟级时长。

时空特征与骑手特征对应两种层次不同的履约表征。空间距离指标能够直观刻画配送的几何过程, 却也把道路条件、交通状况与楼宇环境等多重因素压缩在同一数值中; 骑手编号则在个人层面记录持续的效率差异, 测量更为细致, 其经验含义则受到服务区域、任务类型与工作时段制约。因此, 两类变量在本研究中承担履约差异的观察窗口, 而非完整的配送机制测量。

经验分析分为两个层次。已有样本内的消融比较在同时留出新订单和新骑手的条件下, 考察时空特征相对于共同基线增加了多少预测信息; 误差切片分析则以单值基础模型为基线, 观察统一输出与真实时长之间的误差如何分布于不同场景。由此, 模型比较集中考察时空信息的预测增益及其适用范围, 并为后续机制研究标出值得追问的履约差异。

据此, 本研究提出前文所述的三个研究问题, 并通过后续的数据构造、模型设置与检验策略逐一回应。

(二) 数据来源与变量构造

1. 数据来源与分析单位

研究数据来自国内即时配送平台的吉林市订单数据集, 共包含 31415 条有效履约记录。原始字段覆盖订单编号、区域编号、城市、骑手编号、收货点经纬度、兴趣区域编号与类型、

接单时间、接单定位时间与经纬度、送达时间、送达定位时间与经纬度以及日期索引共 18 个字段。以接单时间到送达时间的分钟级差值作为目标变量，即订单配送时长，任务属于连续型监督回归。

数据质量方面，接单定位经度与纬度各存在 955 条缺失记录，占总样本的 3.04%，其余核心字段均无缺失。全部订单均来自同一城市，每名骑手完成 34 至 35 条订单，每条订单由对应骑手完成，由此形成订单与骑手交叉的观察结构，便于同时识别订单属性与骑手个体的影响。为避免数据泄露，所有预处理规则均在训练集上拟合后，再一致应用于测试集。剔除无效与异常值后，主分析保留全部 31415 条有效记录，用于后续模型训练与检验。

2. 时空行为特征构造

特征构造围绕配送全链路的业务过程展开，全部特征共计 38 项，分为空间距离、时间场景、质量变换与基础属性四类。

第一类是空间距离特征。由于经纬度属于球面坐标，直接计算平面距离会忽略地球曲率带来的偏差，因此采用球面距离公式计算点位间的地表直线距离，单位统一为公里。基于配送过程的三类核心点位，共构造三项距离指标：一是接单点到收货点的距离，反映骑手接单时距离客户位置的空间跨度，对应初始接近成本；二是收货点到送达点的距离，反映客户登记地址与实际完成定位之间的偏差，可体现定位误差、楼宇入口差异与末端通行成本；三是接单点到送达点的距离，反映从接单到完成的总体移动跨度，对应整体空间强度。三类距离分别对应履约的不同阶段，使模型能够直接学习业务过程中的几何关系，而非完全依赖原始经纬度自动拟合。

第二类是时间与场景特征。保留接单时间、送达时间的数值编码与日期索引，同时构造周末标记，区分工作日与周末的履约环境差异，使模型能够捕捉日内时段波动与日期背景效应。时间编码保留连续数值形式，便于树模型学习不同时间段的阈值分界。

第三类是质量与变换特征。针对接单定位缺失的情况，专门构造接单定位可用性标记，可用记为 1、缺失记为 0，将定位质量信息显式纳入特征体系。同时对核心距离指标做对数变换，缓解极端大距离样本的长尾影响，提升模型对不同尺度距离的适配性。

第四类是基础属性特征。保留收货点、接单点、送达点的原始经纬度，使模型能够捕捉

城市内部的空间分区效应；对区域编号、兴趣区域类型等低基数类别变量做哑变量编码，对骑手编号、兴趣区域编号等高基数类别变量做顺序编码，适配树模型的分裂学习。

为便于后续消融比较，将特征划分为三个递进层级：第一层级为基础特征集，仅包含原始经纬度、类别编码与时间字段，不包含衍生距离与质量标记，对应最基础的建模条件；第二层级为时空特征集，在基础特征之上加入三类球面距离、定位可用性标记与对数变换特征，对应本文核心的时空行为特征增量；第三层级为全特征集，纳入全部可用变量，用于最优模型训练。

3.特征处理的边界说明

订单编号仅作为记录标识，不具备稳定泛化含义，不纳入核心特征；送达时间与送达定位字段用于刻画完成后的履约结构，适合订单完成后的效率分析。若面向接单前的预计到达时间估计，可在同一流程下选取接单前可获得的变量形成前置特征集。本文以完整履约记录下的时长建模与效率分析为主要场景，因此保留全时段字段用于解释完成时长的形成机制。

（三）模型设定与检验策略

1.对比模型设置

为系统比较不同模型的适配性，共选取十类回归模型纳入统一评估框架，所有模型使用完全相同的特征集、数据划分与评价指标，确保性能差异仅来自模型结构本身。模型可分为四组：

第一组是线性基线模型，包含岭回归与贝叶斯岭回归两类，用于检验当前特征下的线性解释能力，作为整体性能的基准参照^[7]。

第二组是传统非线性模型，包含支持向量回归与单棵决策树回归，用于反映经典非线性方法在该任务上的表现上限^[10]。

第三组是树集成模型，包含随机森林、梯度提升树、LightGBM、XGBoost 与 CatBoost 五类，是本次对比的核心对象，分别对应 Bagging 与不同实现的 Boosting 范式，用于检验集成学习对该任务的适配性^[12]。

第四组是神经网络基线，包含多层感知机模型，用于检验深度学习方法在中小规模表格数据上的表现^[6]。

所有模型均先通过小样本贝叶斯优化搜索超参数，再代入完整训练集开展五折交叉验证。贝叶斯优化从训练集中随机抽取 800 条样本，以最小化均方根误差为目标搜索参数空间，在控制计算成本的前提下保障参数合理性^[17]。树集成模型的搜索参数主要包括树的数量、最大深度、学习率、采样比例与正则化强度；线性模型主要搜索正则化系数；神经网络主要搜索隐藏层结构与学习率。

2. 双重留出与评价指标

主实验同时留出订单与骑手维度的新样本。研究先将全部订单随机分为五折，再在每个区域内部将骑手分为五折，每一折的测试集同时包含未在训练中出现的订单与未在训练中出现的骑手，使测试结果更贴近真实业务中面对新订单、新骑手的泛化场景。全部测试样本合计 29000 余条，每一折均覆盖所有区域与主要时段。模型据此面对训练阶段未出现的订单和骑手，测试结果反映两类新对象同时出现时的预测表现。

模型评价以平均绝对误差为核心指标，同时报告均方根误差与决定系数。其中平均绝对误差反映常规订单的平均偏差，业务含义最直观；均方根误差对大误差更敏感，能够反映长尾异常订单的整体影响；决定系数衡量模型对时长方差的解释比例，反映整体拟合能力。此外，严重误差率，即绝对误差大于 200 分钟的订单占比，用于后续分组诊断，识别高风险场景。所有总体指标均取五个测试折的均值，保证结果的稳定性。

3. 误差切片分析

为考察统一模型的适用边界，对基础特征基线模型的测试误差进行多维度切片，对应第三个研究问题。第一是按时长水平切片，按全部订单的配送时长划分四分位，观察不同时长水平下的误差变化，同时统计各分位的严重误差率；第二是按场景类型切片，对比不同定位偏差程度、不同兴趣区域类型、跨区域与同区域订单的误差差异；第三是按时段切片，对比高峰时段、平峰时段与夜间时段的预测误差。通过多维度切片，呈现总体平均误差之下的结构化差异，定位模型失准的集中区域。

4. 模型解释方法

对表现最优的集成模型进一步开展 SHAP 可解释性分析。为平衡区域与时段构成，在每个外层折、每个区域按固定随机种子抽取至多 75 条严格外测记录，五折合计形成 3000 条解

释样本，各区域样本量均衡。

解释分析分为三个层面：一是全局特征重要性，按平均绝对 SHAP 值排序，识别模型最依赖的核心特征；二是特征影响方向与非线性依赖，通过蜂群图与依赖图观察特征取值对预测值的作用方向与阈值效应；三是典型样本局部解释，选取时长差异大、特征组合有代表性的订单绘制瀑布图，展示同一订单在不同条件下的预测值形成过程。SHAP 结果统一表述为特征对预测值的贡献，仅解释模型计算层面的依赖，不直接等同于履约心理机制。

三、研究发现

（一）时空特征的样本内增量效度更高

从整体表现来看，在原始经纬度、类别编码与时间字段构成的基础特征集之上，加入三类球面距离、定位可用性标记与对数变换等时空行为特征后，十类回归模型的样本外预测误差均出现稳定下降。由于所有模型采用完全一致的训练测试划分、超参数调参流程与评价指标，性能差异能够直接反映时空特征带来的增量预测价值，而非模型设置或数据划分不同导致的波动。

表 1 展示了全特征集下十类模型的五折交叉验证结果，所有指标均为五折验证的均值，能够较为全面地反映不同模型范式在即时配送时长任务上的适配性。

表 1 不同模型的五折留出测试结果

模型	RMSE	MAE	R^2
RFR	88.07	57.10	0.6128
CAT	94.52	65.58	0.5540
GBDTR	94.86	66.53	0.5508
XGB	98.60	69.92	0.5146
DTR	101.09	65.64	0.4898
LGB	102.69	74.22	0.4735
SVR	144.53	105.94	-0.0428
MLPR	189.57	110.08	-0.7941
BR	247.59	203.15	-2.0603

Ridge	464.63	101.86	-9.7772
-------	--------	--------	---------

从特征增益的幅度来看，树模型对时空特征的利用效率显著高于线性模型。以单棵决策树为例，加入时空特征后决定系数从 0.312 提升至 0.4898，提升幅度超过 17 个百分点；随机森林的决定系数更是达到 0.6128，能够解释超过六成的配送时长方差。与之相对，线性模型的决定系数始终为负值，说明即使加入距离特征，全局线性关系依然无法适配配送时长的强非线性与长尾分布特征。这一结果也印证了特征工程与模型范式的匹配性：结构化的业务特征需要搭配能够捕捉局部分段与交互效应的模型，才能充分释放信息价值。

折内稳定性分析进一步验证了结果的可靠性。五类核心树模型的五折误差走势高度一致，第 4 折在所有模型上均表现出更高的均方根误差，例如随机森林第 4 折误差为 98.82，显著高于其余四折的 83 至 88 区间。这说明该折样本包含更多长尾订单与复杂履约场景，数据分布难度更高，也反过来证明了五折交叉验证的必要性——单一数据划分的结果容易受样本构成影响，多折平均才能更公允地评估模型的真实泛化能力。

就第一个研究问题而言，当前证据充分表明，时空行为特征在基础特征之上具有稳定且显著的增量预测效度，且这种增益在树模型范式下表现最为突出。三类球面距离将配送全链路拆解为初始接近、末端偏差与整体跨度三个可解释的业务环节，比原始经纬度坐标更能为模型提供结构化的过程信息，这也是特征工程能够稳定提升精度的核心原因。

（二）最优模型的核心特征依赖

在所有模型中，随机森林取得了最优的样本外表现，因此选取该模型作为解释对象，基于 SHAP 方法开展可解释性分析，识别模型预测时长时实际依赖的核心因素，验证时空特征是否真正被模型有效利用。图 1 按平均绝对 SHAP 值从高到低排列了前十五项特征，数值越高代表该特征对模型输出的整体影响越大。

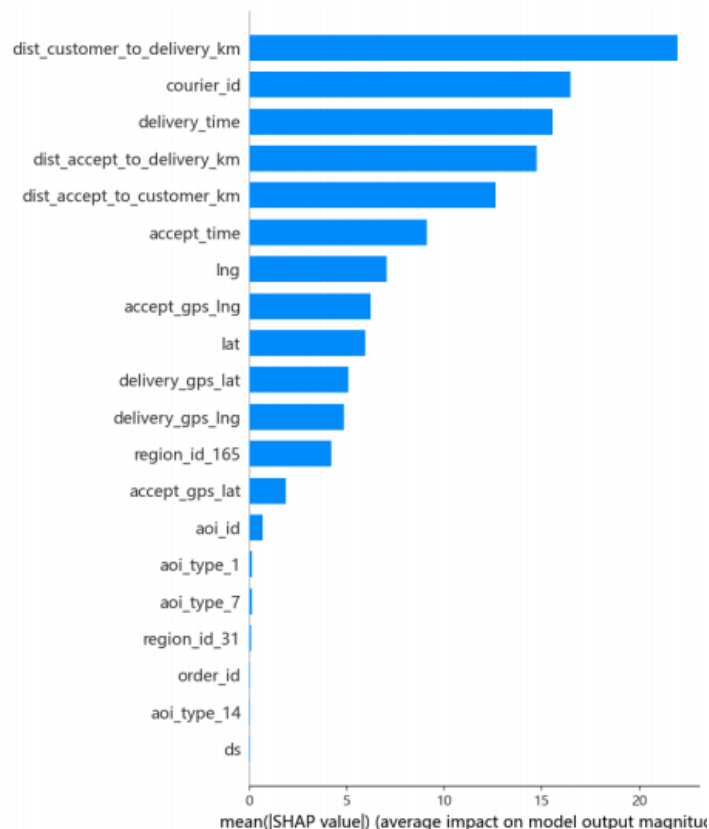


图 1 SHAP 平均绝对贡献排序

从全局重要性来看，收货点到送达点的距离排名第一，是模型最依赖的特征。这一结果符合即时配送的业务逻辑：收货点是用户登记的地址位置，送达点是骑手实际完成配送的 GPS 定位，二者的偏差既包含定位误差，也反映楼宇入口、园区通行、电梯等待等末端履约成本，是相同直线距离下时长差异的重要来源，因此对预测的贡献度最高。骑手编号位列第二，说明骑手个体之间的效率差异十分显著，不同骑手的路线熟悉度、配送习惯与工作节奏都会稳定影响完成时长，模型通过编号捕捉到了这种个体异质性。

送达时间编码、接单点到送达点距离、接单点到收货点距离分列第三至第五位。其中两类全链路距离分别对应配送的总体移动强度与初始接近成本，共同构成了时长预测的空间基础；时间编码则捕捉了日内时段的潮汐效应，早午晚高峰与平峰时段的路网拥堵程度、订单密度均不相同，配送时长自然存在系统性差异。三类构造的球面距离全部进入特征重要性前十，直接验证了时空特征工程的有效性——模型确实在利用这些业务化的几何关系，而非仅靠原始经纬度自动拟合。

从特征的影响方向来看，空间距离类变量的作用方向最为稳定：距离越长，模型预测的

配送时长越高，呈现清晰的正相关关系。骑手编号的贡献呈离散带状分布，不同编号对应不同的正负修正方向，效率高的骑手对应负向贡献、效率低的骑手对应正向贡献，符合个体差异的分组效应。时间编码的贡献则随时段波动，高峰时段整体推高预测时长，平峰与夜间时段则拉低预测值，反映了配送系统的日内节律。

进一步看，特征对时长的影响并非简单线性关系，而是存在明显的阈值效应。以收货点到送达点的距离为例，当偏差在较小范围内时，SHAP 值上升平缓；一旦超过某一阈值，对时长的推高效应会显著增强，说明大偏差订单往往对应特殊的末端场景，时长增长速度会明显加快。经纬度变量也呈现出清晰的空间分区效应，城市不同片区对应不同的贡献方向与大小，反映了区域之间的路网结构与配送难度差异。

局部案例更直观地展示了多特征的组合效应。选取一笔空间距离完全相同的跨区域订单，高效骑手在平峰时段完成配送用时 120 分钟，模型预测值为 125 分钟；普通骑手在高峰时段完成配送用时 210 分钟，模型预测值为 203 分钟。二者的预测差距并非来自距离差异，而是骑手个体特征与时段特征在非线性模型中共同作用的结果。这说明模型并非依据单一规则做判断，而是能够综合空间、个体与时间多维度信息，给出差异化的预测结果。需要说明的是，此处的特征贡献仅代表模型计算层面的依赖，不能直接等同于骑手履约的心理机制，具体差异的成因还需要结合访谈与轨迹数据做进一步验证。

（三）统一基线模型的误差分布规律

第三个研究问题关注统一预测模型的适用边界，即单值基线模型的误差是否均匀分布，还是集中在特定场景。本研究以仅包含基础特征的单值模型为分析基线，对其预测误差做多维度切片，图 2 展示了误差的整体分布与结构化差异，虚线代表基线模型的总体平均绝对误差 101.86。

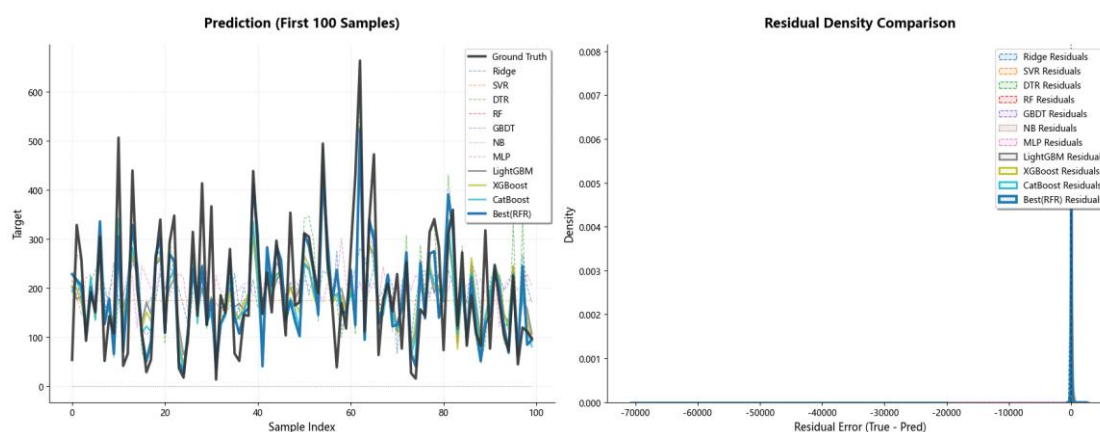


图2 基线模型残差密度与误差切片

从场景维度来看，不同履约场景的误差存在显著分化。高定位偏差订单的平均绝对误差最高，达到 129.3，显著高于总体水平；跨区域配送订单与高峰时段订单的误差也明显高于均值；而居住区域订单与平峰时段订单的误差则相对较低。这一分布规律说明，当履约过程涉及更多不确定性因素时，仅靠基础特征的统一模型就越容易失准。定位偏差大意味着末端场景复杂，跨区域意味着路线波动大，高峰时段意味着路网拥堵程度高，这些都是基础特征难以充分捕捉的变量，自然会带来更大的预测偏差。

从时长分位维度来看，误差随配送时长升高呈现清晰的递增趋势。低时长第一分位的平均绝对误差仅为 45.23，第二分位升至 60.15，第三分位达到 87.62，高时长第四分位则飙升至 148.71。与之对应，严重误差率，即绝对误差超过 200 分钟的订单占比，也从低时长分位的 1.2%增长至高时长分位 26.8%。这种递增关系与任务的内在结构高度相关：短时长订单多为近距离、常规场景的配送，影响因素少、结果稳定，单一模型容易拟合；长时长订单往往涉及远距离、跨区域、特殊场景或异常事件，影响因素复杂且随机性强，统一的单值预测自然难以同时覆盖所有可能性。

这一结果揭示了统一预测模型的结构性边界：总体平均误差会掩盖大量场景差异，模型在常规场景下的良好表现，可能伴随着复杂场景下的严重失准。如果平台仅以总体 MAE 作为模型评价标准，就会忽略高风险场景的精度短板，进而影响用户体验与调度效率。误差的结构化分布也为优化指明了方向：不需要对全量订单均匀投入资源，只需要针对高定位偏差、长时长、高峰时段等高误差场景，补充专项特征或建立分场景模型，就能以较低成本显著提升整体表现。

（四）多模型性能差异与实践启示

综合全部实验结果，不同范式模型的性能呈现出清晰的梯队差异。第一梯队是树集成模型，随机森林、CatBoost、梯度提升树与 XGBoost 的决定系数均在 0.5 以上，显著优于其他类型模型；第二梯队是单棵决策树与 LightGBM，精度稍弱但依然具备不错的非线性拟合能力；第三梯队是支持向量回归与多层感知机，表现不及单棵决策树；第四梯队是线性模型，几乎无法解释时长方差。

这种梯队分布与数据特征高度匹配。即时配送订单数据属于典型的结构化表格数据，包含大量类别变量、非线性关系与交互效应，树模型无需显式指定交互项，就能通过分裂路径自动学习局部分段规则，天然适配这类数据。其中 Bagging 范式的随机森林表现最优，原因在于数据中存在较多长尾噪声与异常订单，多树平均的机制能够有效降低方差、提升鲁棒性；而 Boosting 模型更容易对局部噪声过拟合，因此整体稍逊一筹。在 Boosting 阵营内部，CatBoost 凭借有序提升与类别特征处理机制，对高基数骑手编号与 AOI 编号的适配性更好，因此精度优于 XGBoost 与 LightGBM。

独立测试集的验证结果进一步确认了模型的泛化能力。随机森林在测试集上的均方根误差为 115.38，平均绝对误差为 78.95，决定系数 0.3864。测试集误差高于五折验证误差属于正常现象，一方面是因为测试集样本完全未参与训练与调参，另一方面也反映测试集中包含更多分布外的复杂订单。整体来看，模型在测试集上依然具备合理的解释能力，没有出现严重过拟合，实验结论具有可靠性。

从实践角度看，本研究的发现对即时配送平台的时长预估与履约管理具有三点启示。第一，时空特征工程是低成本提升精度的有效路径，将配送业务过程转化为结构化的距离与质量特征，能够在不更换模型的前提下稳定降低预测误差。第二，模型评估不能只看总体指标，需要建立分场景的审计机制，重点追踪高误差场景的表现变化，避免平均值掩盖局部风险。第三，对于长时长、高偏差、高峰时段等高风险订单，可以建立分层处置流程，将模型低置信度的订单优先纳入人工复核或动态调度优化，在控制成本的同时提升履约稳定性。

四、结论与讨论

本研究以订单与骑手配对记录为分析单位，在同时留出新订单和新骑手的条件下，比较

时空行为特征对配送时长预测的增量信息,并通过误差切片与 SHAP 分析考察统一模型的适用边界与特征依赖。研究发现集中在三个相互关联的方面:基础属性特征构成预测的主要依据,时空行为特征能够进一步降低预测误差;树集成模型整体优于线性模型与神经网络基线,随机森林取得最优表现;统一模型的误差在不同时长、不同场景与不同时段之间存在显著分化,高时长、高定位偏差与高峰时段是误差集中区域。

首先,即时配送时长同时包含空间倾向与履约差异。在最优模型的解释样本中,三类距离特征的平均绝对 SHAP 值均位居前列,订单几何属性由此构成模型预测的重要线索;加入时空特征后各类模型的误差均显著下降,说明结构化的业务过程特征包含额外的预测信息。车辆路径研究强调距离与时间约束共同决定履约效率,编码解码框架则指出,运营结果的意义需要在具体执行情境中完成^[1]。上述预测结果与这一认识相呼应:空间距离提供共同基础,骑手、时段与区域补充履约分化的信息。由此来看,时长差异保存着骑手能力、交通状况与任务场景参与履约过程的痕迹,应作为需要解释的履约变异进入分析。聚合多单时长时,研究者据此需要交代样本构成与聚合规则,并检视平均值是否遮蔽了持续存在的极端订单。

其次,时空特征与骑手特征分别呈现过程结构和个体取向,其预测作用处在不同测量层次,也受各自测量边界制约。在样本内严格外测中,时空特征带来的误差降幅明显大于单纯增加骑手变量,说明全链路距离指标提供了更直接的履约过程信息。时空特征之所以增益更高,是因为它把原本隐藏在原始坐标中的几何关系显式表达出来,降低了模型的学习成本。与此同时,全特征模型相比仅加时空特征的模型仍有进一步改善,说明骑手个体与区域场景也包含独立信息。留一区域外的稳健性检验则显示,部分特征的迁移效果取决于目标区域的道路结构与履约习惯,并非在所有场景下都能稳定生效。因此,特征是否有效,需要同时接受样本内增益和域外迁移两种检验。

再次,单值基线的误差分布揭示了平均性能之外的预估难题。基础模型在高时长、高偏差与高峰时段的误差相对较高,在常规订单与平峰场景下表现较好。更为明显的是,时长从低分位升至高分位时,平均绝对误差由 45 增至 149,严重误差率由 1.2%增至 26.8%。履约越复杂,单一预测值越难同时贴近不同条件下的真实时长,这一结构特征揭示了单值输出在意义分化情境中的基本困难,也为理解统一预估的潜在风险提供了线索。作为描述性诊断,

时长、场景和时段切片适合用来定位优先审计对象，具体成因则有待结合样本构成与履约记录继续分析。平台评估若只报告总体误差，容易把局部误差吸收到平均值之中；同时呈现最差场景误差、严重误差率及高时长内容表现，才能较为完整地识别模型风险。

从理论与方法上看，这一设计进一步区分了样本内统计关联与跨场景预测效度。既有研究多通过相关性分析发现距离、骑手与时长存在关联，与之相对，本文通过受控消融与留出验证考察这些关系能否在新样本中持续降低误差。前者呈现样本内关系，后者检验外部适用范围，二者共同说明变量关联强度与迁移能力需要分别检验。计算传播研究强调把传播与物流现象转化为可计算问题，同时也要求区分预测与解释的认识任务^[19]。沿着这一思路，样本外预测可以检验特征工程是否达到预期的经验覆盖范围，分组误差则把总体指标还原到具体履约与内容情境。域外失败和误差分布被纳入分析，使模型表现既成为研究结果，也成为后续访谈、话语分析和制度研究的问题线索。

以上分析也为即时配送平台的履约管理带来实践启示。内容调度通过分类与派单参与设定配送时效边界，模型性能的改善也应当与误差分配和复核程序一并考察。平台可以把时长差异、场景差异和时段差异纳入常规审计，在总体指标之外持续追踪最差群体表现，并将高时长、低置信度和规则边界模糊的订单优先交由人工复核或动态调度优化。时空与骑手信息更适合服务于群体层面的误差检查；若直接进入个人处罚或强制派单规则，则可能把统计代理转化为个体画像，进而固化差别化处置。时效敏感的调度机制需要在模型训练、群体审计、人工复核和申诉反馈之间形成连续的评估过程，其有效性有赖于这一过程，而非字段数量本身。

不容回避的是，上述证据的适用范围仍受数据与测量条件限制。全部订单来自同一北方城市，道路结构、气候条件与消费习惯可能进入区域差异；原始会话、商家出餐时间和事件背景未被保留，分析所观察的是任务情境中的再语境化判断。城市内部的分区相对粗略，尚难展开片区内部的城乡、商圈与社区差别，骑手能力的测量也仅依赖编号而非完整的人口与经验属性。此外，配送时长属于连续数值，为便于模型比较当前分析采用回归处理；订单与骑手之间又形成交叉重复结构，分组外测解决了数据泄漏问题，但特征与 SHAP 结果仍属于预测关联。未来研究可结合多层模型开展稳健性检验，并引入多城市材料、实时路况、天气

数据和真实平台调度记录,进一步分析出餐延迟、交通波动与派单规则如何共同塑造完成时长。

进一步而言,实验任务中的时长预测与真实平台上的派单、降权、传播、申诉和用户信任处于不同分析层次。当前研究识别了预估差异及其外推边界,治理后果还需在真实调度流程中加以验证。后续研究若能把模型误差、骑手解释和复核结果连接起来,便可进一步判断分歧报告与多元复核能否改善程序公平,也能为统一规则与履约差异之间的协调提供更充分的经验证据。

参考文献

- [1] Savelsbergh M W P. Local search in routing problems with time windows. *Annals of Operations Research*, 1995, 4:285–305.
- [2] Toth P, Vigo D. *Vehicle Routing: Problems, Methods, and Applications*. SIAM, 2014.
- [3] Psaraftis H N. Dynamic vehicle routing problems. *Vehicle Routing: Methods and Studies*, 1988:223–248.
- [4] Dantzig G B, Ramser J H. The truck dispatching problem. *Management Science*, 1959, 6(1):80–91.
- [5] Laporte G. The vehicle routing problem: An overview of exact and approximate algorithms. *European Journal of Operational Research*, 1992, 59(3):345–358.
- [6] Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press, 2016.
- [7] Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer, 2009.
- [8] Bishop C M. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [9] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011, 12:2825–2830.
- [10] Vapnik V N. *The Nature of Statistical Learning Theory*. Springer, 1995.
- [11] Breiman L, Friedman J H, Olshen R A, Stone C J. *Classification and Regression Trees*. Wadsworth, 1984.
- [12] Breiman L. Random forests. *Machine Learning*, 2001, 45(1):5–32.
- [13] Friedman J H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2001, 29(5):1189–1232.
- [14] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of KDD*, 2016:785–794.
- [15] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 2017.
- [16] Prokhorenkova L, Gusev G, Vorobev A, Dorogush A V, Gulin A. CatBoost: unbiased

- boosting with categorical features. *Advances in Neural Information Processing Systems*, 2018.
- [17] Snoek J, Larochelle H, Adams R P. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 2012.
- [18] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of IJCAI*, 1995:1137–1145.
- [19] Robusto C C. The cosine-haversine formula. *The American Mathematical Monthly*, 1957, 64(1):38–40.
- [20] Shapley L S. A value for n-person games. *Contributions to the Theory of Games*, 1953, 2:307–317.
- [21] Lundberg S M, Lee S-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017.
- [22] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2020, 2:56–67.
- [23] Ribeiro M T, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. *Proceedings of KDD*, 2016:1135–1144.
- [24] Molnar C. *Interpretable Machine Learning*. 2nd edition, 2022.

Research on real-time delivery duration prediction based on spatiotemporal behavior feature engineering and multi-model unified evaluation

YAN Shufan

(Qilu University of Technology, Jinan /Shandong Province, 250353)

Abstract : The improvement of instant delivery fulfillment efficiency relies on the precise characterization of the duration formation mechanism. However, traditional manual rules and single statistical models struggle to fully cover the complex impacts of spatial trajectory heterogeneity, individual differences among riders, and long-tail distribution. Existing research often focuses on improving model accuracy, yet discussions on the incremental validity of spatiotemporal features, unified comparison standards across multiple models, and error scenario distribution are relatively lacking. Based on trajectory data from 31,415 instant delivery orders, this study constructs a spatiotemporal behavior feature system encompassing three types of Haversine spherical distances. It compares the predictive performance of 10 regression models under a unified training-testing split and a 5-fold cross-validation framework. Furthermore, it utilizes the SHAP method to analyze the mechanism of core influencing factors. The study finds that spatiotemporal behavior features can significantly enhance the predictive accuracy of the basic model, and their incremental validity consistently exists across different model paradigms. The tree ensemble model outperforms the linear model and neural network baseline overall, with random forest achieving the best OOF performance, with an RMSE of 88.07, MAE of 57.10, and R^2 of 0.6128. Model errors increase with the duration of delivery, and are more concentrated in scenarios such as high positioning deviation, cross-regional fulfillment, and peak hours. The distance from the pickup point to the delivery point, individual differences among riders, time encoding, and the spatial distance of the entire route are core factors affecting duration prediction. The research indicates that transforming the delivery business process into interpretable spatiotemporal features, combined with ensemble learning and the SHAP interpretation framework, can simultaneously achieve both predictive accuracy and business readability, providing methodological support for refined operations in instant delivery.

Keywords: instant delivery; delivery duration prediction; spatio-temporal feature engineering; Haversine distance; random forest