دانشگاه صنعتی شریف

دانشکده مهندسی صنایع

تمرین دوم درس مدلسازی و تصمیم گیری داده محور

نگارش:

سینا نجفی، رضا گورانی

استاد :

سرکارخانم دکتر صدقی

آذر۱۴۰۴


بسم الله الرحمن الرحیم

عنوان	شماره

فصل اول: کلیات تحقیق

۱-۱) مقدمه

در سال‌های اخیر، با گسترش خدمات تحویل غذا و رقابت شدید میان پلتفرم‌های آنلاین، سرعت و دقت در فرآیند تحویل سفارش‌ها به یکی از مهم‌ترین شاخص‌های عملکردی کسب‌وکارها تبدیل شده است. تجربه‌ی کاربری مثبت و تحویل به‌موقع، علاوه بر افزایش رضایت مشتریان، نقش تعیین‌کننده‌ای در حفظ سهم بازار و ارتقای جایگاه رقابتی شرکت‌ها دارد. در همین راستا، تحلیل داده‌های عملیاتی و ساخت مدل‌های پیش‌بینی‌کننده می‌تواند ابزار ارزشمندی برای بهینه‌سازی فرآیند تحویل و تصمیم‌گیری‌های مدیریتی باشد.

در این تمرین، هدف اصلی مدل‌سازی زمان تحویل سفارش غذا و تعیین عوامل مؤثر بر آن با بهره‌گیری از روش‌های تصمیم‌گیری داده‌محور است. داده‌های ارائه‌شده شامل مجموعه‌ای از ویژگی‌های مرتبط با مسیر، شرایط ترافیک، وضعیت آب‌وهوا، ناحیه‌ی رستوران و مشتری، حالت حمل‌ونقل پیک، ساعت انجام تحویل و دیگر خصوصیات مؤثر بر زمان جابه‌جایی هستند. هدف این است که با تحلیل این داده‌ها و ساخت مدل‌های رگرسیون، نه‌تنها مقدار زمان تحویل را پیش‌بینی کنیم، بلکه میزان اهمیت و نقش هر یک از عوامل را نیز در این فرآیند ارزیابی نماییم.

این پروژه علاوه بر ارائه‌ی یک مدل قابل اتکا برای پیش‌بینی، در پی آن است که با استفاده از تحلیل آماری و روش‌های انتخاب ویژگی، مهم‌ترین متغیرها شناسایی شوند و تأثیرات متقابل احتمالی میان آن‌ها بررسی گردد. همچنین ارزیابی فروض رگرسیون، تحلیل باقیمانده‌ها و بررسی دقت مدل روی داده‌های آموزشی و آزمایشی، بخشی ضروری از مسیر دستیابی به مدلی معتبر و قابل استناد است.

۲-۱) پاکسازی و غربالگری اولیه داده ها

۲-۱-۱) کنترل نوع داده‌ها (Data Types)

در ابتدا، مطمئن شدیم که نوع هر ستون متناسب با ماهیت آن باشد:

ستون‌های عددی به صورت int/float ثبت شده‌اند.

ستون‌های طبقه‌ای (Categorical) مانند traffic_level, weather, delivery_mode به صورت object/categorical ثبت شده‌اند.

این موضوع اهمیت دارد زیرا نوع داده نادرست ممکن است در فرآیند One-Hot Encoding یا رگرسیون خطی ایجاد خطا کند.

۲-۱-۲) بررسی متغیرهای طبقه‌ای (Categorical Variables)

در اولین مرحله، متغیرهای طبقه‌ای از نظر وجود مقادیر نامعتبر، مقادیر خارج از دامنه، غلط‌های نوشتاری و ناسازگاری‌ها بررسی شدند. دسته‌های مورد انتظار برای هر ستون مطابق صورت‌مسئله مشخص بود و داده‌ها با آن تطبیق داده شد.

در این بررسی، یک نکته‌ی مهم مشاهده شد:

در ستون delivery_mode دو مقدار «Bike» و «Bicycle» وجود داشت که از نظر معنایی معادل یکدیگر هستند و وجود آن‌ها به صورت جداگانه می‌توانست باعث ایجاد دو دسته‌ی غیرواقعی و خطای مدل‌سازی شود. بنابراین تصمیم گرفته شد که این دو مقدار ادغام شوند و هر دو به صورت یکسان با عنوان "Bike" ثبت گردند.

غیر از این مورد، هیچ مقدار نامعتبر یا خارج از مقادیر مجاز در متغیرهای طبقه‌ای مشاهده نشد.

ـ زمان تحویل بر اساس سطح ترافیک (Delivery Time by Traffic Level)

نمودار باکس‌پلات نشان می‌دهد که زمان تحویل در سطح ترافیک High به‌طور قابل‌توجهی بزرگ‌تر از Low و Medium است. توزیع داده‌ها نیز نشان می‌دهد که با افزایش سطح ترافیک، علاوه بر افزایش میانه، دامنه‌ی تغییرات زمان تحویل نیز بیشتر می‌شود. این روند بیانگر نقش معنادار و مستقیم وضعیت ترافیک در فرآیند تحویل است و انتظار می‌رود در مدل رگرسیون نیز اثر قابل توجهی داشته باشد.

ـ تحلیل اثر نوع وسیله تحویل بر زمان تحویل سفارش

نمودار جعبه‌ای مربوط به زمان تحویل بر حسب نوع وسیله حمل‌ونقل نشان می‌دهد که بین سه دسته موتورسیکلت (Bike)، خودرو (Car) و اسکوتر (Scooter) تفاوت‌هایی در توزیع زمان تحویل وجود دارد، هرچند این تفاوت‌ها بسیار شدید و کاملاً جدا از هم نیستند.

بر اساس میانه‌های توزیع، مشاهده می‌شود که خودرو (Car) کمترین میانه زمان تحویل را دارد. این موضوع می‌تواند نشان‌دهنده سرعت بالاتر و پایداری بیشتر خودرو در مسیرهای طولانی‌تر یا شرایط خاص ترافیکی باشد.

در مقابل، اسکوتر (Scooter) دارای بیشترین میانه زمان تحویل است که می‌تواند به محدودیت‌های فنی، سرعت کمتر یا حساسیت بیشتر آن به شرایط مسیر و ترافیک مرتبط باشد. موتورسیکلت (Bike) نیز از نظر میانه زمان تحویل در موقعیتی بین این دو قرار دارد.

علاوه بر میانه‌ها، دامنه تغییرات زمان تحویل در هر سه دسته نسبتاً گسترده است و هم‌پوشانی قابل توجهی میان جعبه‌ها و بازه‌های داده مشاهده می‌شود. این هم‌پوشانی نشان می‌دهد که اگرچه نوع وسیله تحویل می‌تواند بر زمان تحویل اثرگذار باشد، اما این اثر به‌تنهایی تعیین‌کننده نیست و عوامل دیگری مانند طول مسیر واقعی، سطح ترافیک و شرایط محیطی نیز نقش مهمی ایفا می‌کنند.

در مجموع، نتایج این نمودار حاکی از آن است که نوع وسیله تحویل اثری متوسط بر زمان تحویل دارد. به بیان دیگر، این متغیر می‌تواند به‌عنوان یکی از متغیرهای توضیحی در مدل رگرسیون مورد استفاده قرار گیرد، اما انتظار نمی‌رود به‌تنهایی نقش غالب در پیش‌بینی زمان تحویل داشته باشد. به همین دلیل، بررسی اثر این متغیر در کنار سایر عوامل و همچنین در قالب اثرات تعاملی (برای مثال، تعامل نوع وسیله با طول مسیر یا سطح ترافیک) می‌تواند درک دقیق‌تری از نقش آن در فرآیند تحویل سفارش فراهم کند.

۲-۱-۳) بررسی متغیرهای عددی (Numeric Variables)

متغیرهای عددی شامل مواردی مانند distance_km, route_length_km, delivery_time_min و hour مورد ارزیابی قرار گرفتند تا مطمئن شویم در بازه‌های مجاز و مطابق با تعریف مسئله قرار دارند.

کنترل انجام‌شده نشان داد که تمام مقدارهای عددی در رنج معرفی‌شده در صورت‌مسئله قرار دارند و مقدار غیرطبیعی یا خارج از دامنه در این ستون‌ها مشاهده نشد.

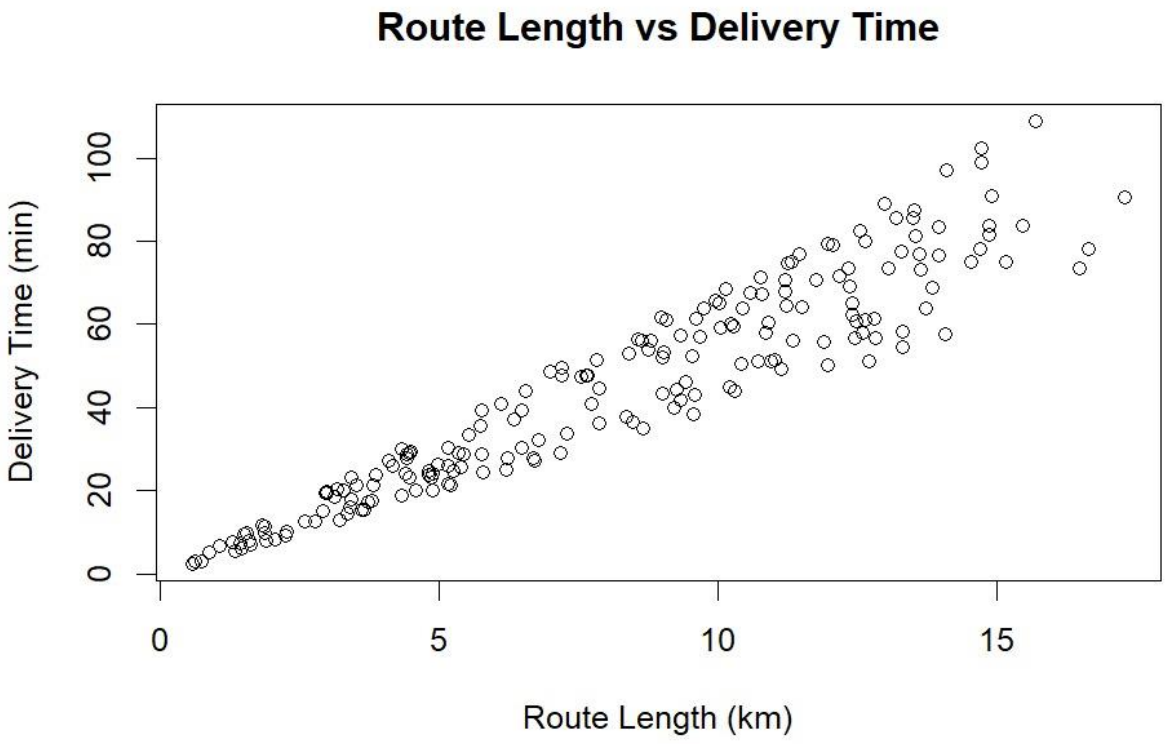
ـ توزیع زمان تحویل (Distribution of Delivery Time)

نمودار هیستوگرام نشان می‌دهد که زمان تحویل سفارش‌ها دارای یک الگوی نسبتاً پراکنده با یک تجمع مرکزی در محدوده‌ی حدود ۳۰ تا ۷۰ دقیقه است. توزیع به صورت کامل نرمال نیست و اندکی کشیدگی به سمت راست (Right-Skewed) مشاهده می‌شود، زیرا مقادیر بزرگ‌تر از ۸۰ دقیقه نیز هرچند با فراوانی کم، وجود دارند. این الگو بیانگر آن است که بیشتر تحویل‌ها در بازه‌ی زمانی میانگین انجام شده‌اند اما موارد دیرتر نیز قابل توجه‌اند. این موضوع اهمیت مدل‌سازی دقیق و بررسی باقی‌مانده‌ها در بخش رگرسیون را نشان می‌دهد.

_ رابطه فاصله مستقیم با زمان تحویل (Distance vs Delivery Time)

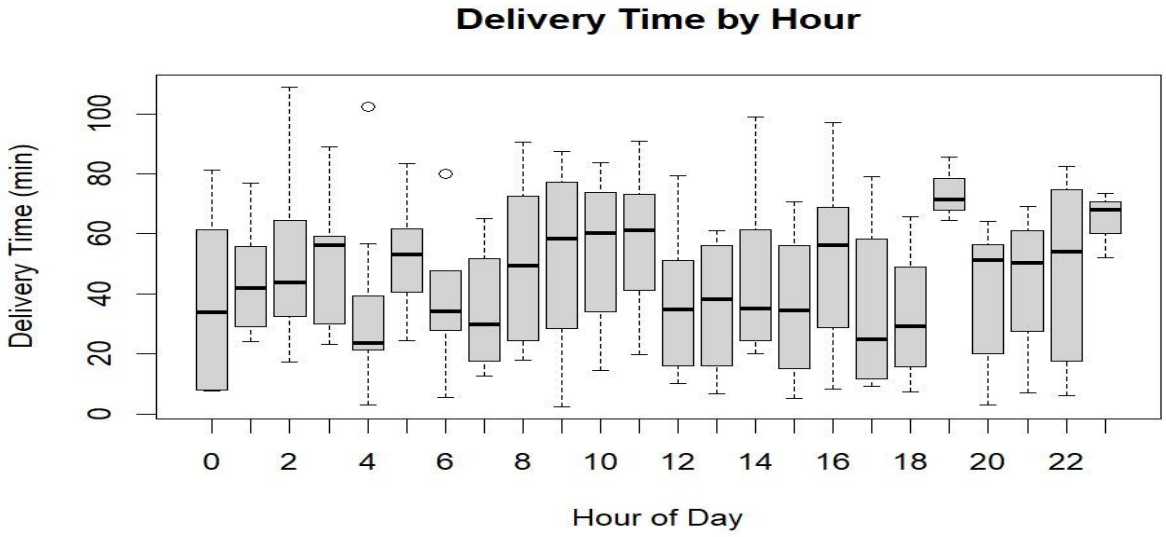
نمودار پراکنش مربوط به فاصله مستقیم از مبدا تا مقصد نشان می‌دهد که با افزایش فاصله، زمان تحویل نیز روند افزایشی دارد. با این حال پراکندگی نقاط به‌ویژه در فواصل بالاتر از ۸ کیلومتر بیشتر می‌شود که نشان می‌دهد تنها فاصله‌ی مستقیم نمی‌تواند به‌تنهایی توضیح‌دهنده‌ی بخش بزرگی از تغییرات زمان تحویل باشد. این رفتار نشان‌دهنده‌ی وجود عوامل تکمیلی مانند مسیر واقعی، شرایط ترافیکی و نوع وسیله‌ی نقلیه است.

ـ رابطه طول مسیر واقعی با زمان تحویل (Route Length vs Delivery Time)



این نمودار یکی از واضح‌ترین الگوها را نشان می‌دهد. بین طول مسیر واقعی و زمان تحویل یک رابطه‌ی خطی قوی و منظم وجود دارد و برخلاف فاصله مستقیم، پراکندگی نقاط بسیار کمتر است. این موضوع نشان می‌دهد که طول مسیر واقعی یکی از مهم‌ترین و قابل‌اتکاترین متغیرها برای پیش‌بینی زمان تحویل است. به دلیل هم‌بستگی بسیار بالا میان route_length و distance_km، احتمال وجود هم‌خطی (Multicollinearity) در مدل نیز مطرح می‌شود که در بخش رگرسیون باید بررسی شود.

_ تحلیل اثر ساعت تحویل بر زمان تحویل سفارش



برای بررسی نقش متغیر ساعت در زمان تحویل سفارش، این متغیر به دو شکل عددی و طبقه‌ای مورد تحلیل قرار گرفت تا الگوهای کلی و همچنین رفتار توزیعی داده‌ها در ساعات مختلف روز بهتر شناسایی شود.

در نمودار میانگین زمان تحویل بر حسب ساعت (نمایش عددی)، مشاهده می‌شود که تغییرات زمان تحویل در طول شبانه‌روز الگوی خطی و یکنواختی ندارد. میانگین زمان تحویل در برخی ساعات افزایش و در برخی ساعات کاهش می‌یابد، بدون آنکه روند صعودی یا نزولی مشخصی در کل بازه ۰ تا ۲۳ ساعت دیده شود. با این حال، در

برخی ساعات خاص به‌ویژه در ساعات عصر و شب میانگین زمان تحویل افزایش محسوسی دارد. این رفتار نشان می‌دهد که اثر ساعت بر زمان تحویل احتمالاً غیرخطی است و نمی‌توان آن را صرفاً با یک ضریب خطی ساده به‌درستی مدل‌سازی کرد.

برای بررسی دقیق‌تر، نمودار جعبه‌ای زمان تحویل بر حسب ساعت روز (نمایش طبقه‌ای) نیز ترسیم شد. این نمودار نشان می‌دهد که علاوه بر تفاوت در میانگین‌ها، پراکندگی زمان تحویل در ساعات مختلف به‌طور قابل توجهی متفاوت است. در برخی ساعات، به‌ویژه حوالی ساعات پرتردد روز، دامنه تغییرات زمان تحویل گسترده‌تر است که می‌تواند ناشی از نوسانات ترافیک، حجم سفارش‌ها یا محدودیت‌های عملیاتی باشد. در مقابل، در برخی ساعات دیگر، توزیع زمان تحویل فشرده‌تر و پایدارتر است.

نکته مهم این است که هم‌پوشانی قابل توجهی میان توزیع زمان تحویل در ساعات مختلف مشاهده می‌شود. این هم‌پوشانی نشان می‌دهد که اگرچه ساعت می‌تواند بر زمان تحویل اثر بگذارد، اما این اثر به‌تنهایی تعیین‌کننده نیست و تحت تأثیر عوامل دیگری مانند طول مسیر واقعی، سطح ترافیک و نوع وسیله حمل‌ونقل قرار دارد. به همین دلیل، انتظار نمی‌رود که متغیر ساعت به‌صورت یک متغیر عددی ساده در مدل رگرسیون اثر قوی و معناداری داشته باشد.

۲-۱-۴) بررسی داده‌های گمشده (Missing Values)

کل داده‌ها از نظر وجود مقادیر گمشده یا سلول‌های خالی ارزیابی شدند. نتیجه بررسی به این صورت است که: هیچ مقدار گمشده‌ای در هیچ یک از ستون‌ها وجود ندارد. این نکته یک مزیت محسوب می‌شود، زیرا نیاز به روش‌های جایگزینی داده وجود ندارد و ریسک انتقال خطا به مدل کاهش می‌یابد.

۲-۱-۵) بررسی سطرهای تکراری (Duplicate Rows)

وجود ردیف‌های تکراری می‌تواند باعث ایجاد بایاس در مدل و انحراف در تخمین ضرایب شود. بنابراین تمامی

ردیف‌ها با استفاده از شناسه‌ها و محتویاتشان بررسی شدند. نتیجه بررسی به این صورت بود که: هیچ سطر تکراری در مجموعه‌داده وجود ندارد. تمام ۱۹۴ مشاهده یکتا هستند و تکراری بودن داده‌ها در این مجموعه مشاهده نشد.

۲-۱-۶) تبدیل متغیرهای طبقه‌ای برای استفاده در مدل رگرسیون

در این مجموعه‌داده، تعدادی از متغیرها ماهیت طبقه‌ای دارند؛ مانند سطح ترافیک، نوع وسیله حمل‌ونقل، وضعیت هوا و ناحیه رستوران و مشتری. از آنجا که مدل رگرسیون تنها با متغیرهای عددی کار می‌کند، لازم بود این متغیرهای طبقه‌ای به شکلی مناسب برای ورود به مدل آماده شوند.

برای این منظور، در محیط نرم‌افزاری R متغیرهای طبقه‌ای به صورت «عامل» (با استفاده از دستور factor) تعریف شدند. با این کار، هنگام برازش مدل رگرسیونی، خود نرم‌افزار به‌طور خودکار برای هر متغیر طبقه‌ای چند متغیر دوحالته ایجاد می‌کند و یکی از طبقات را به عنوان طبقه مرجع در نظر می‌گیرد. در نتیجه، ضرایب به‌دست‌آمده برای هر طبقه، نشان‌دهنده اختلاف آن طبقه با طبقه مرجع در زمان تحویل هستند.

این رویکرد چند مزیت مهم دارد: از بروز خطا در تعریف دستی متغیرهای کمکی جلوگیری می‌کند، انتخاب طبقه مرجع را استاندارد و شفاف می‌سازد و کد را ساده‌تر و خواناتر نگه می‌دارد. به این ترتیب، متغیرهای طبقه‌ای به شکل مناسبی برای ورود به مدل رگرسیون آماده شدند و امکان تحلیل درست اثر آن‌ها بر زمان تحویل فراهم گردید.

۱-۲-۷) انتخاب نوع متغیر ساعت

متغیر «ساعت ثبت سفارش» یکی از عواملی است که می‌تواند الگوی تغییرات زمان تحویل را در طول شبانه‌روز توضیح دهد. با این حال، این متغیر را می‌توان به شیوه‌های مختلفی وارد مدل رگرسیونی کرد و انتخاب نادرست آن ممکن است باعث افزایش پیچیدگی مدل یا کاهش قابلیت تعمیم شود. از این رو، پیش از ورود به مدل‌سازی نهایی، لازم بود نحوه مناسب نمایش این متغیر به‌صورت تجربی بررسی شود.

در این مرحله، دو رویکرد متفاوت برای مدل‌سازی متغیر ساعت مورد آزمون قرار گرفت. در رویکرد نخست، ساعت به‌صورت یک متغیر عددی پیوسته در مدل وارد شد تا اثر خطی تغییر ساعت بر زمان تحویل بررسی گردد. در رویکرد دوم، ساعت به‌صورت یک متغیر دسته‌ای مدل شد تا امکان وجود الگوهای متفاوت در ساعات مختلف شبانه‌روز بدون فرض خطی بودن فراهم شود.

برای آن‌که مقایسه بین این دو رویکرد منصفانه و کنترل‌شده باشد، در هر دو مدل، متغیرهای «طول مسیر» و «فاصله» نیز به‌عنوان عوامل اصلی و پایه‌ای وارد شدند. بدین ترتیب، ارزیابی اثر متغیر ساعت در شرایط یکسان انجام گرفت.

نتایج حاصل از برازش مدل‌ها نشان داد که مدل مبتنی بر ساعت عددی، عملکرد بهتری روی داده‌های آزمون دارد؛ به‌طوری‌که ضریب تعیین آن برابر با ۰.۹۲۴ به دست آمد،

مدل: delivery_time_min = −۲.۸۷۷ + ۴.۸۷۶ * route_length_km + ۰.۹۴۶ * distance_km + ۰.۱۲۵ * hour_num

در حالی که این مقدار برای مدل مبتنی بر ساعت دسته‌ای برابر با ۰.۹۱۰ بود. با این حال، تصمیم‌گیری صرفاً بر اساس ضریب تعیین انجام نشد و سایر معیارهای مهم نیز مورد توجه قرار گرفت.

مدل: delivery_time_min = −۵.۷۸۲ + ۴.۷۹۲ * route_length_km + ۱.۰۸۳ * distance_km + ۲.۱۰۸ * hour_1 + ۳.۵۰۹ * hour_2 + ۷.۷۰۹ * hour_3 + ۵.۳۵۴ * hour_4 + ۳.۵۵۴ * hour_5 + ۶.۳۲۹ * hour_6 + ۰.۴۷۷ * hour_7 + ۰.۲۱۰ * hour_8 + ۲.۲۰۰ * hour_9 + ۳.۲۴۴ * hour_10 + ۳.۶۷۵ * hour_11 + ۳.۰۲۶ * hour_12 + ۰.۵۵۲ * hour_13 + ۲.۸۴۸ * hour_14 + ۴.۸۷۳ * hour_15 + ۷.۲۲۰ * hour_16 + ۴.۸۳۴ * hour_17 + ۵.۷۴۴ * hour_18 + ۱۴.۰۴۹ * hour_19 + ۳.۳۷۹ * hour_20 + ۱۱.۶۲۳ * hour_21 + ۷.۷۲۵ * hour_22 + ۰.۶۰۳ * hour_23

از نظر خطای پیش‌بینی، مدل عددی دارای خطای باقیمانده کمتر (Residual Standard Error پایین‌تر) نسبت به مدل دسته‌ای بود که نشان‌دهنده دقت بیشتر آن در پیش‌بینی زمان تحویل است. علاوه بر این، مدل دسته‌ای به دلیل ایجاد تعداد زیادی پارامتر (یک ضریب برای هر ساعت)، منجر به افزایش قابل توجه درجه آزادی مصرف‌شده شد که این موضوع خطر بیش‌برازش را افزایش می‌دهد، بدون آن‌که بهبود متناظری در عملکرد مدل ایجاد کند.

از منظر معنی‌داری آماری ضرایب نیز مشاهده شد که در مدل دسته‌ای، تنها تعداد بسیار محدودی از ضرایب ساعات معنی‌دار هستند و بخش عمده‌ای از آن‌ها از نظر آماری فاقد اهمیت‌اند. این مسئله نشان می‌دهد که افزودن متغیر ساعت به‌صورت دسته‌ای اطلاعات مؤثر چندانی به مدل اضافه نکرده است. در مقابل، مدل عددی با تعداد ضرایب کمتر، ساختار ساده‌تری داشته و از پایداری آماری بالاتری برخوردار است.

همچنین از نظر قابلیت تفسیر، مدل عددی مزیت قابل توجهی دارد؛ زیرا اثر ساعت به‌صورت یک روند کلی و قابل فهم بیان می‌شود، در حالی که مدل دسته‌ای نیازمند مقایسه هر ساعت با یک سطح مرجع بوده و تفسیر نتایج آن پیچیده‌تر و کمتر شهودی است.

با در نظر گرفتن هم‌زمان ضریب تعیین روی داده‌های آزمون، میزان خطای پیش‌بینی، سادگی ساختاری مدل، پایداری ضرایب و قابلیت تفسیر، در ادامه تحلیل‌ها تصمیم گرفته شد متغیر ساعت به‌صورت عددی در مدل‌ها استفاده شود و از نمایش دسته‌ای آن صرف‌نظر گردد. این انتخاب موجب ایجاد مدلی متعادل‌تر از نظر دقت، سادگی و قابلیت تعمیم شد.

۱-۳) تقسیم داده‌ها به مجموعه آموزش (Train) و آزمون (Test)

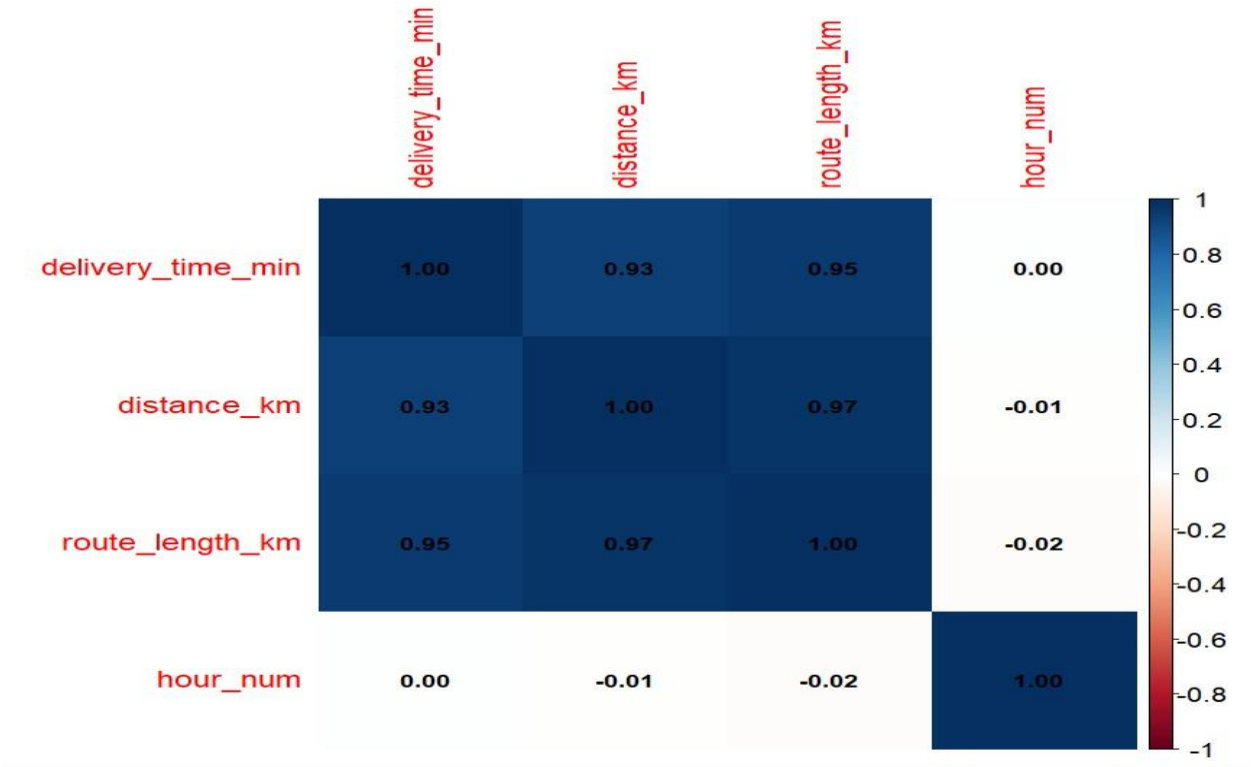
برای ارزیابی صحیح مدل‌های پیش‌بینی و جلوگیری از بیش‌برازش (Overfitting)، مجموعه‌داده به دو بخش آموزش و آزمون تقسیم شد. هدف از این تقسیم آن است که مدل تنها بر اساس داده‌های آموزشی برازش یابد و

سپس عملکرد آن بر روی داده‌هایی که در مرحله آموزش مشاهده نکرده است، سنجیده شود. به این ترتیب می‌توان دقت مدل را به‌صورت بی‌طرفانه ارزیابی کرد و مقایسه نسخه‌های مختلف مدل را معتبرتر انجام داد.

در این پروژه، مجموعه‌داده شامل ۱۹۴ مشاهده بود. با الهام از رویه‌های رایج در مدل‌سازی، نسبت تقریبی ۸۰ درصد برای آموزش و ۲۰ درصد برای آزمون در نظر گرفته شد. بر این اساس:

۱۵۵ مشاهده به‌عنوان مجموعه آموزش (Train) و ۳۹ مشاهده به‌عنوان مجموعه آزمون (Test) در نظر گرفته شد. انتخاب این نسبت باعث می‌شود مدل از حجم مناسبی از داده برای یادگیری الگوها برخوردار باشد، در عین حال تعداد کافی مشاهده برای ارزیابی نهایی روی مجموعه آزمون باقی بماند. فرآیند تقسیم‌بندی به‌صورت تصادفی و با تنظیم یک seed ثابت انجام شد تا نتایج قابل تکرار باشند.

۱-۴) بررسی همبستگی های متغیر های عددی



پیش از ساخت مدل رگرسیون، لازم است روابط میان متغیرهای عددی بررسی شود تا مشخص گردد هر متغیر تا

چه اندازه با دیگری ارتباط دارد و آیا هم‌خطی شدید (وابستگی زیاد میان دو متغیر) در داده‌ها وجود دارد یا خیر. وجود چنین رابطه‌هایی می‌تواند بر تخمین ضرایب رگرسیون، معناداری آماری متغیرها و پایداری مدل تأثیر بگذارد.

به همین منظور، ابتدا ماتریس همبستگی میان متغیرهای عددی محاسبه و نمودار آن بررسی شد. این کار کمک می‌کند الگوهای کلی وابستگی در داده‌ها شناسایی شود و مشخص گردد کدام متغیرها اطلاعات مشابهی دارند. هدف این مرحله حذف متغیرها نیست، بلکه درک بهتر ساختار داده و آمادگی برای تفسیر نتایج مدل کامل و روش حذف پسرو است.

این بررسی یک گام ضروری پیش از مدل‌سازی است، زیرا نشان می‌دهد کدام متغیرها ممکن است اثر غالب داشته باشند و در صورت ورود هم‌زمان به مدل باعث تضعیف معناداری یکدیگر شوند. بر اساس این تحلیل اولیه، در مراحل بعدی از نتایج همبستگی برای تفسیر رفتار مدل و تصمیم‌گیری در روش‌های انتخاب متغیر استفاده شد.

۱. رابطه فاصله مستقیم با زمان تحویل

نتایج ماتریس همبستگی نشان داد که بین فاصله مستقیم بین رستوران و مشتری (distance_km) و زمان تحویل (delivery_time_min) یک همبستگی بسیار قوی وجود دارد (ضریب همبستگی در حدود ۰٫۹۳). این مقدار بیانگر آن است که هرچه مشتری از رستوران دورتر باشد، زمان تحویل نیز به‌طور میانگین بیشتر می‌شود و این افزایش تا حد زیادی رفتاری شبیه یک رابطه خطی دارد. به زبان ساده، فاصله مستقیم به‌تنهایی می‌تواند بخش قابل توجهی از تغییرات زمان تحویل را توضیح دهد و به عنوان یک متغیر پیش‌بین قوی مطرح شود؛ هرچند هنوز تنها عامل تعیین‌کننده نیست و سایر عوامل عملیاتی نیز در زمان نهایی تحویل نقش دارند.

۲. رابطه طول مسیر واقعی با زمان تحویل

در ادامه، همبستگی بین طول مسیر واقعی طی‌شده (route_length_km) و زمان تحویل بررسی شد. ضریب همبستگی به‌دست‌آمده در این حالت حتی از فاصله مستقیم نیز بزرگ‌تر و در حدود ۰٫۹۴۷ بود. این نتیجه نشان

می‌دهد که طول مسیر واقعی، بهتر و دقیق‌تر از فاصله مستقیم می‌تواند زمان تحویل را توضیح دهد. این موضوع از نظر منطقی نیز قابل انتظار است؛ زیرا ممکن است دو نقطه از نظر فاصله مستقیم نسبتاً به هم نزدیک باشند، اما به دلیل وجود خیابان‌های یک‌طرفه، چراغ‌های راهنمایی متعدد، دوربرگردان‌ها یا ترافیک سنگین، مسیر واقعی که پیک طی می‌کند بسیار طولانی‌تر و زمان‌برتر باشد. بنابراین، در تحلیل‌ها طول مسیر واقعی به‌عنوان شاخص اصلیِ مربوط به فاصله و یکی از مهم‌ترین متغیرها در پیش‌بینی زمان تحویل در نظر گرفته شد.

۳. هم‌خطی شدید بین فاصله مستقیم و طول مسیر واقعی

یکی از یافته‌های مهم دیگر، همبستگی بسیار شدید بین distance_km و route_length_km بود؛ به‌طوری‌که ضریب همبستگی میان این دو متغیر حدود ۰٫۹۷ به دست آمد. این مقدار نشان‌دهنده وجود هم‌خطی چندگانه بسیار قوی است؛ به این معنا که این دو متغیر تقریباً اطلاعات بسیار مشابهی را به مدل منتقل می‌کنند و تفاوت آن‌ها در عمل محدود است. حضور هم‌زمان چنین متغیرهایی در مدل رگرسیون می‌تواند باعث ناپایداری ضرایب، کاهش دقت برآورد و افزایش p-valueها شود و در نتیجه تفسیر مدل را دشوار کند. به همین دلیل، در مراحل بعد و به‌خصوص در فرآیند حذف پسرو، route_length_km به‌عنوان متغیر اصلی نگه داشته شد و distance_km از مدل نهایی کنار گذاشته شد؛ زیرا طول مسیر واقعی همبستگی قوی‌تری با متغیر پاسخ دارد و نماینده مناسب‌تری برای بُعد فاصله در مدل است.

۴. بررسی همبستگی متغیر ساعت با سایر متغیرها

بررسی همبستگی متغیر ساعت تحویل (hour) با زمان تحویل و سایر متغیرهای عددی نشان داد که ضرایب همبستگی در این مورد بسیار کوچک و نزدیک به صفر هستند. این نتیجه نشان می‌دهد که ساعت روز، حداقل به صورت یک رابطه خطی ساده، اثر مستقیم و قابل توجهی بر زمان تحویل ندارد. با این حال، این نکته به معنای بی‌اهمیت بودن کامل ساعت نیست؛ زیرا ممکن است ساعت در قالب الگوهای غیرخطی یا در قالب اثرات تعاملی با سایر متغیرها، مانند سطح ترافیک یا ناحیه مشتری، نقش پررنگ‌تری داشته باشد. به همین دلیل، متغیر ساعت

از تحلیل کنار گذاشته نشد و در ادامه‌ی کار، شیوه‌های مختلف نمایش آن (عددی، طبقه‌ای) و همچنین امکان وجود اثرات تعاملی آن با متغیرهای دیگر مورد بررسی قرار گرفت. این رویکرد کمک می‌کند نقش واقعی ساعت در کنار سایر عوامل به‌صورت کامل‌تر و دقیق‌تر ارزیابی شود.

۵.جمع‌بندی و نتایج تحلیل همبستگی

تحلیل ماتریس همبستگی میان متغیرهای عددی چند نتیجه مهم و سرنوشت‌ساز برای ادامه فرایند مدل‌سازی به همراه داشت. نخست آنکه رابطه میان فاصله مستقیم مشتری تا رستوران و زمان تحویل بسیار قوی بود. این موضوع نشان می‌دهد که افزایش فاصله به‌طور طبیعی موجب افزایش زمان تحویل می‌شود و فاصله می‌تواند یکی از متغیرهای توضیحی مهم باشد. با این حال، بررسی متغیر «طول مسیر واقعی طی‌شده» نشان داد که این متغیر حتی رابطه‌ای قوی‌تر با زمان تحویل دارد. از آنجا که مسیر واقعی شرایط خیابان‌ها، جهت حرکت، دوربرگردان‌ها و محدودیت‌های ترافیکی را بهتر از فاصله مستقیم منعکس می‌کند، طبیعی است که پیش‌بینی بهتری از مدت زمان تحویل ارائه دهد. به همین دلیل، در مدل‌های بعدی طول مسیر واقعی به‌عنوان متغیر اصلی مرتبط با زمان تحویل در نظر گرفته شد.

نکته مهم دیگر، وجود همبستگی بسیار شدید میان فاصله مستقیم و طول مسیر واقعی است. این میزان وابستگی نشان می‌دهد که هر دو متغیر تقریباً یک نوع اطلاعات را منتقل می‌کنند و ورود هم‌زمان آن‌ها به مدل می‌تواند باعث ایجاد مشکل هم‌خطی شود. چنین وضعیتی موجب ناپایداری ضرایب، کاهش معناداری آماری و تضعیف کیفیت مدل خواهد شد. بنابراین، بر اساس این یافته، در مدل اصلی تنها طول مسیر واقعی نگه داشته شد و فاصله مستقیم از مدل پایه حذف گردید، مگر در شرایطی که به‌صورت اثر تعاملی با سایر متغیرها نیاز به بررسی داشته باشد.

همچنین مشاهده شد که متغیر ساعت، برخلاف انتظار، همبستگی قابل توجهی با زمان تحویل یا دیگر متغیرهای عددی ندارد. این موضوع نشان می‌دهد که ساعت روز اثر خطی مستقیمی بر زمان تحویل ندارد، اما احتمال وجود

نقش غیرخطی یا تعاملی آن قابل چشم‌پوشی نیست. به همین دلیل، این متغیر به‌طور کامل حذف نشد و در مراحل بعدی مدل‌سازی در قالب نمایش‌های مختلف (عددی، طبقه‌ای یا چرخه‌ای) و همچنین به‌صورت اثر تعاملی مورد ارزیابی قرار خواهد گرفت.

در مجموع، نتایج این بخش مسیر مدل‌سازی را روشن ساخت. طول مسیر واقعی به‌عنوان مهم‌ترین متغیر عددی در مدل باقی می‌ماند، فاصله مستقیم تنها در صورت نیاز در تعامل‌ها بررسی می‌شود، و متغیر ساعت با رویکردی دقیق‌تر و در قالب مدل‌های تعاملی یا نمایش‌های غیرخطی تحلیل خواهد شد. این یافته‌ها اساس تصمیم‌گیری در مراحل بعدی، شامل ساخت مدل کامل، حذف پسرو و مدل‌های تعاملی، را تشکیل می‌دهند و موجب می‌شوند مدل نهایی رفتار واقعی داده‌ها را بهتر منعکس کند.

۱-۵) مدل رگرسیون با تمام متغیر ها

در این بخش ابتدا یک مدل کامل برای پیش‌بینی زمان تحویل غذا بر اساس تمام متغیرهای موجود ساخته شد. هدف از این مرحله آن بود که مشخص شود هر یک از متغیرهای موجود تا چه اندازه در توضیح تغییرات زمان تحویل نقش دارند و کدام عوامل تأثیر قوی‌تری بر پاسخ دارند. به همین منظور، همه عوامل شامل مسافت، طول مسیر، سطح ترافیک، نوع وسیله، وضعیت آب‌وهوا، منطقه رستوران، منطقه مشتری و ساعت روز به‌طور هم‌زمان وارد مدل شدند.

نتیجه مدل کامل نشان داد که تقریباً تمام متغیرها، به‌جز طول مسیر، اثر آماری معنی‌داری بر زمان تحویل ندارند. تنها متغیر route_length_km با فاصله قابل توجهی مهم‌ترین عامل مؤثر بر زمان تحویل بود. مقدار ضریب تعیین در داده‌های آموزش حدود ۰.۸۹۷ به دست آمد، اما مهم‌تر از آن، مقدار R^2 روی داده‌های آزمون که به‌صورت دستی محاسبه شد برابر با ۰.۹۱۷بود که نشان‌دهنده توان پیش‌بینی نسبتاً مناسب مدل است.

پس از بررسی مدل کامل، مطابق دستور از روش حذف پسرو (Backward Elimination) برای ساده‌سازی مدل استفاده شد. در این روش، در هر مرحله کم‌اثرترین متغیر حذف می‌شود و معیار AIC بررسی می‌شود تا مدلی بهینه از نظر تعادل بین دقت و سادگی به دست آید. روند حذف به ترتیب شامل حذف آب‌وهوا، منطقه مشتری، منطقه رستوران، ترافیک، نوع وسیله و نهایتاً ساعت روز بود. در پایان تمام متغیرهای کم‌اثر کنار گذاشته شدند و تنها یک متغیر باقی ماند: طول مسیر.

مدل پس از بکوارد:

$$\text{delivery\_time\_min} \sim -1.05 + 5.58 * \text{route\_length\_km}$$

مدل نهایی حاصل از بکوارد، با وجود سادگی بسیار زیاد، عملکردی تقریباً برابر با مدل کامل داشت. مقدار R^2 روی داده‌های آزمون برای مدل ساده‌شده برابر با ۰.۹۱۸ به دست آمد که حتی اندکی بهتر از مدل کامل بود. این نتیجه اهمیت طول مسیر را به‌عنوان اصلی‌ترین عامل تعیین‌کننده زمان تحویل نشان می‌دهد و بیانگر آن است که اضافه کردن متغیرهای متعدد نه‌تنها کمکی به بهبود دقت نمی‌کند، بلکه تنها باعث پیچیدگی بیش‌ازحد مدل می‌شود.

در نهایت می‌توان نتیجه گرفت که مدل ساده‌شده مبتنی بر روش حذف پسرو، از نظر سادگی، تفسیرپذیری و همچنین پایداری در پیش‌بینی روی داده‌های جدید، انتخاب مناسب‌تری نسبت به مدل کامل است. این مدل بدون از دست دادن دقت، ساختاری مختصر و قابل فهم ارائه می‌دهد و به همین دلیل به‌عنوان مدل نهایی انتخاب شد.

۱-۶) تحلیل و انتخاب متغیرهای تعاملی (Interaction Effects)

پس از ساخت مدل کامل و مدل ساده‌شده با روش حذف پسرو، مرحله بعدی تحلیل به بررسی تعامل میان متغیرها اختصاص یافت. هدف از این بخش آن بود که مشخص شود آیا ترکیب هم‌زمان دو متغیر می‌تواند الگوی دقیق‌تری

از زمان تحویل ارائه دهد یا خیر؛ زیرا در مسائل واقعی، اثر یک متغیر اغلب ثابت نیست و ممکن است بسته به شرایط دیگر تغییر کند.

برای این منظور، در مجموع ۱۴ تعامل میان متغیرهای کلیدی داده‌ها ساخته شد. هر تعامل در قالب یک مدل مستقل ارزیابی گردید و از سه جنبه مورد بررسی قرار گرفت:

۱. معنی‌داری ضرایب (آزمون t)

۲. بهبود خطا و R^2 روی داده‌های آزمون

۳. قابل‌توجیه بودن از نظر منطقی و عملی

۱-۶-۱) تعامل‌هایی که اثر قابل‌توجهی نداشتند

بخش زیادی از تعامل‌ها، علیرغم اینکه از نظر تئوری ممکن بود اثرگذار باشند، در عمل تأثیر چندانی بر مدل نداشتند. چند نمونه از این موارد عبارت‌اند از:

۱) تعامل طول مسیر و نوع وسیله (route_length × delivery_mode)

ضرایب تعاملی تقریباً صفر و کاملاً غیرمعنی‌دار بودند.

آزمون R^2 روی داده‌های آزمون نیز هیچ بهبود چشمگیری نشان نداد.

این نتیجه از نظر عملی هم منطقی بود: سرعت متوسط وسایل مختلف تفاوت زیادی ندارد و طول مسیر برای همه تقریباً مشابه عمل می‌کند.

۲) تعامل فاصله و آب‌وهوا (distance × weather)

هیچ‌یک از ضرایب تعاملی معنی‌دار نبودند و مقدار R^2 مدل تقریباً با مدل پایه برابر شد.

این تعامل می‌توانست معنی‌دار باشد، اما به نظر می‌رسد تأثیر آب‌وهوا در این مجموعه داده چندان متغیر یا شدید

نبوده است.

۳) تعامل ساعت و ترافیک (hour × traffic_level)

با وجود اینکه انتظار می‌رود تأثیر ساعت در شرایط ترافیکی متفاوت باشد، داده‌ها چنین الگویی را تأیید نکردند.

ضرایب معنی‌دار نبودند و R^2 مدل حتی از مدل ساده‌تر هم بدتر شد.

این نشان می‌دهد که اثر ساعت به‌صورت مستقل بهتر کار می‌کند و ترکیب آن با ترافیک منجر به بهبود مدل نمی‌شود.

۴) تعامل طول مسیر و منطقه رستوران (route_length × restaurant_zone)

نتایج این بخش به این صورت بود که: ضرایب غیرمعنی‌دار و اثر آماری بسیار ضعیف بود. منطقه رستوران در این داده‌ها پراکندگی زیادی ندارد و همین موضوع باعث کم‌اثر بودن تعامل شده است.

۱۲ تعامل‌ها بررسی شده در این بخش معمولاً یکی از دو مشکل را داشتند:

ضرایب غیرمعنی‌دار بود یعنی داده شواهد کافی برای وجود اثر تعاملی ارائه نمی‌کرد.

عدم بهبود R^2 روی داده‌های آزمون یعنی حتی اگر ضرایب کمی معنی‌دار بودند، عملکرد مدل بهتر نمی‌شد.

۱-۶-۲) تعامل‌هایی که به مدل نهایی راه یافتند

در میان تمام تعامل‌های آزموده‌شده، تنها دو تعامل از هر نظر برتر شناخته شدند:

۱) تعامل فاصله و سطح ترافیک (distance × traffic_level)

این تعامل نشان داد که اثر افزایش فاصله در شرایط ترافیکی مختلف یکسان نیست.

به‌عنوان مثال، افزایش ۲ کیلومتر فاصله در شرایط ترافیک سبک ممکن است تنها چند دقیقه اضافه کند، اما همان

مقدار افزایش در ترافیک سنگین می‌تواند تأثیر بسیار بیشتری داشته باشد.

این تعامل در مدل R^2 را به‌طور محسوسی افزایش داد و ضرایب آن منطقی و قابل‌تفسیر بودند.

۲) تعامل طول مسیر و منطقه مشتری ($route\_length \times customer\_zone$)

این تعامل نیز عملکرد مدل را بهبود داد. نتایج نشان داد که طول مسیر در مناطق مختلف اثر متفاوتی بر زمان تحویل دارد.

این موضوع از نظر عملی قابل‌توجیه است: چون بعضی مناطق ممکن است مسیرهای کندتر، شلوغ‌تر یا پرپیچ‌وخمی داشته باشند، بنابراین همان افزایش طول مسیر می‌تواند در منطقه‌ای خاص اثر بیشتری بر زمان تحویل داشته باشد.

فرآیند انتخاب به‌صورت چندمرحله‌ای بود:

۱. ساخت مدل مستقل برای هر تعامل

۲. بررسی معنی‌داری آماری ضرایب تعاملی

۳. محاسبه R^2 روی داده‌های آزمون و مقایسه با مدل پایه

۴. بررسی منطق عملی و قابل‌توضیح بودن اثر تعاملی

۵. انتخاب تعامل‌هایی که هم از نظر آماری و هم از نظر عملکردی برتر بودند

مدل ترکیبی حاصل از دو تعامل منتخب (مدل MIX) با مقدار $R^2 \approx 0.935$ بهترین عملکرد را ثبت کرد؛ به طوری که بالاتر از تمام مدل‌های دیگر و حتی مدل کامل اولیه است.

۱-۷) مدل رگرسیون ارائه شده

در این بخش به ارزیابی و تحلیل مدل نهایی پیش‌بینی زمان تحویل پرداخته شد تا مشخص شود این مدل تا چه اندازه توانسته رفتار واقعی داده‌ها را بازتاب دهد و در مقایسه با مدل‌های ساده‌تر و کامل‌تر چگونه عمل می‌کند. هدف این تحلیل، ارائه تصویری دقیق از کیفیت پیش‌بینی، میزان خطا و قابلیت اتکای مدل در استفاده واقعی است.

پس از ساخت مدل کامل شامل تمامی متغیرها و سپس ساده‌سازی آن با روش حذف پسرو، مجموعه‌ای از تعامل‌ها نیز بررسی شد. از میان ۱۴ تعامل ارزیابی‌شده، دو تعامل «ترافیک × فاصله» و «منطقه مشتری × طول مسیر» معنادارتر و کاراتر تشخیص داده شدند و به مدل نهایی افزوده شدند. این مدل ترکیبی با نام مدل میکس اجرا شد و عملکرد آن روی داده‌های آزمون سنجیده شد.

مدل نهایی: delivery_time_min ~ 2.18 + 3.85*route_length_km + 1.28*distance_km + 2.51*traffic_levelLow + 0.69*traffic_levelMedium - 6.66*customer_zoneEast - 5.01*customer_zoneNorth - 6.61*customer_zoneSouth - 5.31*customer_zoneWest - 0.16*distance_km:traffic_levelLow + 0.08*distance_km:traffic_levelMedium + 1.16*route_length_km:customer_zoneEast + 0.92*route_length_km:customer_zoneNorth + 0.82*route_length_km:customer_zoneSouth + 0.92*route_length_km:customer_zoneWest

نتایج نشان داد که مدل میکس توانست مقدار تقریباً ۹۳٫۵ درصد از تغییرات زمان تحویل را روی داده‌های آزمون توضیح دهد. این مقدار بالاتر از مدل کامل (حدود ۹۱٫۷٪) و تقریباً برابر مدل بسیار ساده‌ی تک‌متغیره (۹۱٫۸٪) است، اما برتری مدل نهایی در این است که علاوه بر دقت بیشتر، ساختاری معنادار و قابل‌تفسیر دارد و قادر است اثرات واقعی ترافیک، فاصله و منطقه مشتری را بهتر بازتاب دهد.

برای آنکه ارزیابی تنها بر اساس (R^2) نباشد، میزان خطا نیز بررسی شد. خطای استاندارد باقیمانده در مدل

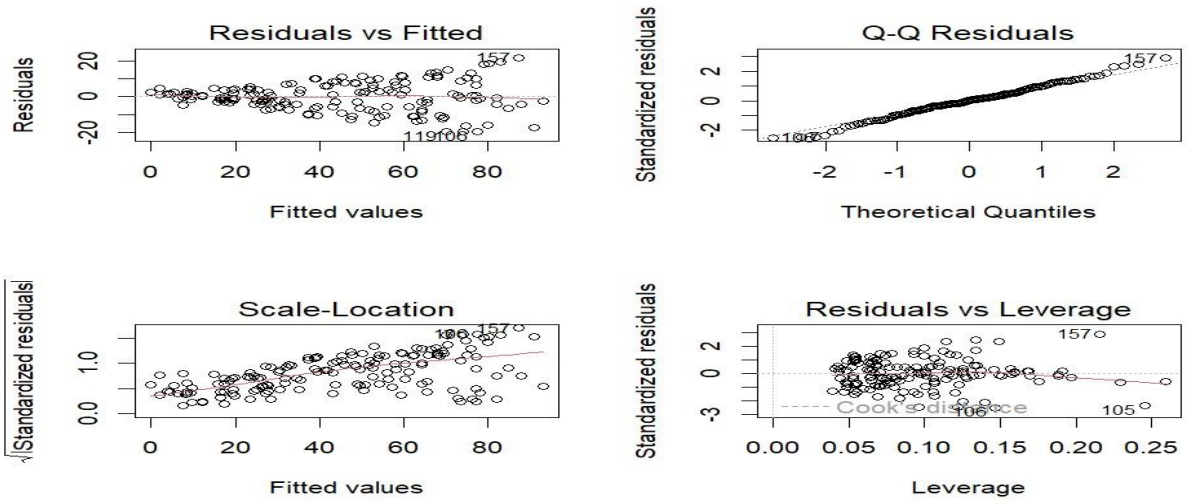
نهایی حدود ۸٫۴۷ دقیقه به دست آمد. این مقدار تقریباً معادل انحراف خطا در پیش‌بینی است؛ یعنی مدل به‌طور متوسط حدود ۸ تا ۹ دقیقه با زمان واقعی تحویل اختلاف دارد. با توجه به اینکه میانگین زمان تحویل در داده‌ها در حدود ۴۵ تا ۵۰ دقیقه است، این سطح خطا نسبتاً قابل‌قبول بوده و نشان‌دهنده توان مناسب مدل در کاربردهای عملی است. همچنین این میزان خطا نسبت به مدل کامل تقریباً ثابت مانده ولی در عوض پیش‌بینی مدل روی داده‌های آزمون بهبود یافته است.

مقایسه خطا میان مدل‌های مختلف نشان داد که مدل ساده‌ی تک‌متغیره از نظر خطای باقیمانده کمی بهتر از مدل کامل بود، اما چون هیچ اثر زمینه‌ای مانند ترافیک یا منطقه مشتری را منعکس نمی‌کرد، ارزش تبیینی کمتری داشت. از سوی دیگر، مدل کامل اگرچه شامل متغیرهای زیادی بود، اما بسیاری از آن‌ها اثر معناداری نداشتند و همین موضوع باعث کاهش تفسیرپذیری می‌شد. مدل نهایی با حفظ تعادل میان سادگی و دقت، بهترین عملکرد را در بین گزینه‌ها ارائه کرد.

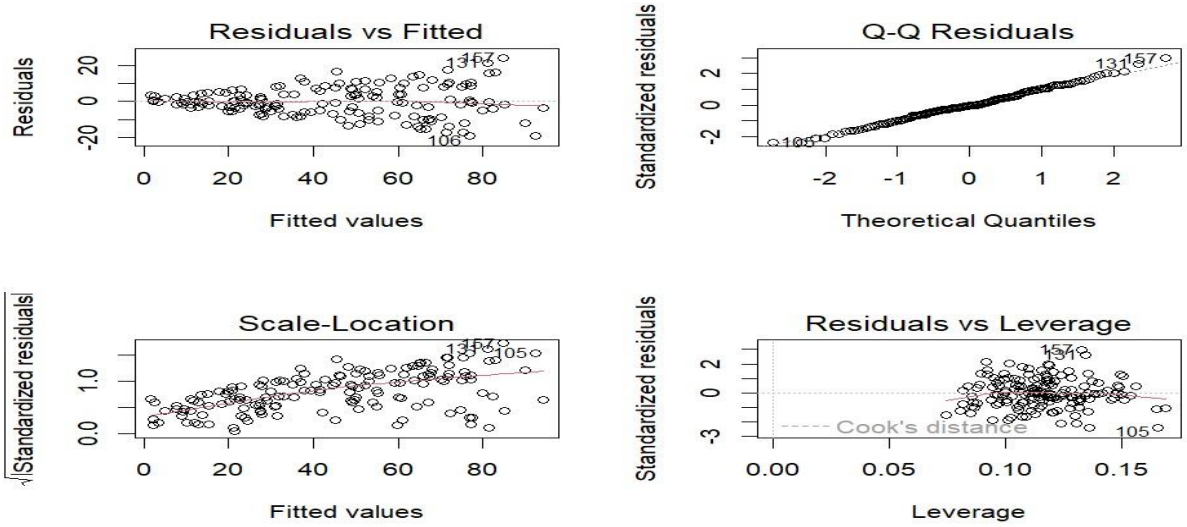
در مجموع می‌توان گفت مدل نهایی انتخاب‌شده، هم از نظر قدرت پیش‌بینی و هم از نظر قابلیت توضیح‌دهی، مناسب‌ترین گزینه برای تحلیل زمان تحویل در این پژوهش است. این مدل خطایی قابل‌پذیرش دارد، نسبت به داده‌های جدید عملکرد بسیار خوبی ارائه می‌دهد و به کمک تعامل‌های انتخاب‌شده، جنبه‌های پیچیده‌تری از رفتار سیستم را نیز مدل‌سازی می‌کند؛ از این رو می‌تواند پایه‌ای معتبر برای تصمیم‌گیری‌های مدیریتی و عملیاتی در زمینه بهینه‌سازی مسیر و زمان‌بندی ارسال باشد.

۱-۸)بررسی مفروضات مدل رگرسیون

به‌منظور ارزیابی اعتبار نتایج مدل‌های رگرسیونی و اطمینان از صحت استنباط‌های آماری، فرضیات اصلی رگرسیون خطی برای دو مدل مدل کامل (full_model) و مدل منتخب نهایی (m_mix) مورد بررسی قرار گرفت. برای این منظور، چهار نمودار تشخیصی استاندارد شامل «باقی‌مانده‌ها در برابر مقادیر برازش‌شده»، «نمودار Q-Q باقی‌مانده‌ها»، «نمودار مقیاس–مکان» و «باقی‌مانده‌ها در برابر اهرم» ترسیم و تحلیل شد.



۴ نمودار مربوط به مدل میکس



۴ تمودار مربوط به فول مدل

نخست، نمودار باقی‌مانده‌ها در برابر مقادیر برازش‌شده برای هر دو مدل بررسی شد. در هر دو حالت، نقاط به‌صورت نسبتاً تصادفی در اطراف خط صفر پراکنده شده‌اند و خط روند ترسیم‌شده تقریباً افقی و نزدیک به صفر است. این الگو نشان می‌دهد که فرض خطی بودن رابطه بین متغیرهای توضیحی و زمان تحویل به‌طور قابل قبولی برقرار است. همچنین الگوی مشخصی مانند قیف یا افزایش منظم پراکندگی با افزایش مقادیر برازش‌شده مشاهده

نمی‌شود، که این موضوع بیانگر برقراری فرض همسانی واریانس باقی‌مانده‌ها در هر دو مدل است.

در گام بعد، نمودار Q-Q باقی‌مانده‌ها برای بررسی فرض نرمال بودن خطاها مورد توجه قرار گرفت. در هر دو مدل، نقاط عمدتاً بر روی خط قطری قرار گرفته‌اند و تنها در دو انتهای توزیع انحراف‌های جزئی مشاهده می‌شود. این میزان انحراف برای داده‌های واقعی کاملاً طبیعی است و نشان می‌دهد که توزیع باقی‌مانده‌ها به‌طور کلی به توزیع نرمال نزدیک است. بنابراین، فرض نرمال بودن باقی‌مانده‌ها برای هر دو مدل قابل قبول ارزیابی می‌شود.

نمودار مقیاس–مکان نیز الگوی خاصی از افزایش یا کاهش سیستماتیک پراکندگی باقی‌مانده‌ها را نشان نمی‌دهد و این موضوع مجدداً فرض همسانی واریانس را تأیید می‌کند. علاوه بر این، نمودار باقی‌مانده‌ها در برابر اهرم نشان داد که اگرچه چند مشاهده دارای اهرم نسبتاً بالاتری هستند، اما هیچ نقطه‌ای به‌طور جدی فراتر از خطوط مرجع فاصله کوک قرار نگرفته است. بنابراین، مشاهده‌ای با نفوذ بسیار بالا که بتواند نتایج مدل را به‌طور اساسی تحت تأثیر قرار دهد، شناسایی نشد.

در مورد فرض استقلال باقی‌مانده‌ها نیز، با توجه به اینکه هر مشاهده مربوط به یک سفارش مستقل است و داده‌ها ماهیت سری زمانی یا تکراری ندارند، می‌توان فرض استقلال خطاها را منطقی و قابل قبول دانست.

در مجموع، بررسی نمودارهای تشخیصی نشان می‌دهد که فرضیات اصلی رگرسیون خطی شامل خطی بودن، همسانی واریانس، نرمال بودن باقی‌مانده‌ها و نبود نقاط با نفوذ شدید، برای هر دو مدل full_model و m_mix به‌طور قابل قبولی برقرار هستند. از این رو، نتایج به‌دست‌آمده از این مدل‌ها از نظر آماری معتبر بوده و می‌توان به تفسیر ضرایب و مقایسه عملکرد آن‌ها اطمینان داشت. همچنین با توجه به برقراری این فرضیات، نیازی به اعمال تبدیل‌های اضافی یا استفاده از روش‌های جایگزین در این مرحله احساس نمی‌شود.

به‌طور کلی، تحلیل نمودارهای تشخیصی نشان داد که مدل نهایی از نظر مفروضات رگرسیون معتبر است و می‌توان به ضرایب برآوردشده، آزمون‌های آماری و پیش‌بینی‌های مدل اعتماد کرد. با توجه به دقت بالای مدل ($R^2 \approx$ ۰.۹۳۵) و پایداری خطاها، مدل انتخاب‌شده هم از نظر عملکرد پیش‌بینی و هم از لحاظ آماری دارای کیفیت

مطلوب است.

در ادامه به اقدامات لازم در زمان نقض فرض ها میپردازیم:

ـ اگر فرض خطی بودن نقض می‌شد (الگوی منحنی در نمودار Residuals vs Fitted):

می‌توانستیم ازتبدیلات غیرخطی روی پیش‌بینی‌کننده‌ها استفاده کنیم (مثلاً اضافه کردن (route_length_km^2) به مدل برای گرفتن اثرات درجه دوم).

می‌توانستیم از مدل‌های غیرخطی مانند رگرسیون چندجمله‌ای یا الگوریتم‌های یادگیری ماشین (مثل Random Forest) استفاده کنیم.

ـ اگر فرض همسانی واریانس نقض می‌شد (الگوی قیفی در نمودار Residuals vs Fitted):

یک راه حل رایج، تبدیل لگاریتمی متغیر هدف است. یعنی به جای delivery_time_min، از log(delivery_time_min) در مدل استفاده کنیم. این کار معمولاً واریانس را تثبیت می‌کند.

استفاده از رگرسیون وزنی (Weighted Least Squares) که به نقاط با واریانس کمتر، وزن بیشتری می‌دهد.

ـ اگر فرض نرمال بودن باقی‌مانده‌ها نقض می‌شد (نقاط در نمودار Q-Q از خط دور بودند):

رگرسیون خطی نسبت به نقض این فرض، به خصوص با حجم نمونه مناسب، نسبتاً مقاوم (Robust) است.

اما اگر نقض شدید بود، تبدیلاتی مانند تبدیل Box-Cox روی متغیر هدف می‌توانست به نرمال‌سازی خطاها کمک کند.

۱-۹) بررسی متغیر های مورد نیاز دیگر

نتایج مدل نهایی نشان داد که متغیرهای مرتبط با مسیر، به‌ویژه طول مسیر و برخی تعامل‌های مکانی، نقش اصلی را در تبیین زمان تحویل ایفا می‌کنند. با این حال، باقی‌ماندن بخشی از خطای پیش‌بینی و واریانس توضیح‌داده‌نشده

نشان می‌دهد که فرآیند تحویل تنها به عوامل حمل‌ونقل محدود نمی‌شود و متغیرهای مهمی خارج از داده‌های فعلی در این فرآیند دخیل هستند. در صورت امکان دریافت داده‌های تکمیلی از صاحب کسب‌وکار، می‌توان انتظار داشت که هم دقت پیش‌بینی و هم قدرت تبیین مدل به‌طور قابل توجهی افزایش یابد.

مهم‌ترین داده پیشنهادی، اطلاعات مربوط به فرآیند آماده‌سازی سفارش در رستوران است. در داده‌های موجود، متغیر زمان تحویل ترکیبی از «زمان آماده‌سازی غذا» و «زمان سفر پیک» است. در نتیجه، مدل فعلی قادر به تفکیک این دو منبع تأخیر نیست و مشخص نمی‌شود که افزایش زمان تحویل ناشی از کندی عملکرد رستوران بوده یا مشکلات حمل‌ونقل. در اختیار داشتن زمان آماده‌سازی غذا این امکان را فراهم می‌کند که متغیر هدف به «زمان واقعی سفر» بازتعریف شود. این کار باعث حذف نویز ناشی از عملکرد رستوران شده و منجر به ساخت مدلی بسیار دقیق‌تر برای پیش‌بینی زمان حمل‌ونقل می‌شود. علاوه بر این، می‌توان یک مدل مستقل برای پیش‌بینی زمان آماده‌سازی غذا طراحی کرد که از نظر مدیریتی ارزش بالایی دارد.

دسته دوم داده‌های پیشنهادی، اطلاعات زمانی و مکانی دقیق‌تر است. متغیر ساعت در مدل فعلی بخشی از الگوی زمانی را توضیح می‌دهد، اما تفاوت رفتار سیستم در روزهای مختلف هفته لحاظ نشده است. افزودن متغیر «روز هفته» می‌تواند الگوهای متفاوت ترافیک و حجم سفارش در آخر هفته‌ها و روزهای کاری را به‌طور مستقیم وارد مدل کند و بخشی از واریانسی را که توسط ساعت به تنهایی توضیح داده نمی‌شود، پوشش دهد.

همچنین، داده‌های ترافیکی موجود در سطحی کلی (کم، متوسط و زیاد) ثبت شده‌اند، در حالی که ترافیک ماهیتی پویا و لحظه‌ای دارد. دسترسی به داده‌های ترافیک لحظه‌ای، مانند زمان تخمینی سفر بدون ترافیک در مقایسه با زمان واقعی، می‌تواند جایگزین مناسبی برای متغیر کیفی ترافیک باشد. این نوع داده عددی و پیوسته، اثر ترافیک را با دقت بسیار بیشتری وارد مدل کرده و به‌طور مستقیم باعث کاهش خطای پیش‌بینی می‌شود.

دسته سوم، داده‌های مربوط به جزئیات سفارش و پیک است. حجم یا تعداد اقلام سفارش می‌تواند بر زمان تحویل اثرگذار باشد، زیرا سفارش‌های بزرگ‌تر معمولاً نیازمند بسته‌بندی دقیق‌تر یا حمل با احتیاط بیشتر هستند. افزودن

این متغیر عددی می‌تواند بخشی از خطاهای باقی‌مانده مدل را توضیح دهد. علاوه بر این، ویژگی‌های پیک مانند سابقه کاری یا امتیاز عملکرد می‌توانند تفاوت‌های فردی در سرعت تحویل را تبیین کنند. در غیاب این اطلاعات، این تفاوت‌ها به‌صورت نویز در مدل ظاهر می‌شوند.

در مجموع، داده‌های پیشنهادی فوق این امکان را فراهم می‌کنند که مدل از یک چارچوب صرفاً توصیفی فراتر رفته و به مدلی تفکیکی و دقیق‌تر تبدیل شود؛ مدلی که بتواند سهم هر بخش از فرآیند تحویل، شامل آماده‌سازی، حمل‌ونقل و عوامل انسانی، را به‌صورت جداگانه تحلیل کند. چنین مدلی علاوه بر بهبود معیارهای آماری مانند ضریب تعیین و کاهش خطا، بینش عملی ارزشمندی برای بهینه‌سازی عملکرد کسب‌وکار فراهم خواهد کرد.

۱۰-۱) بررسی عوامل تاثیرگذار بر مدت زمان تحویل

بر اساس کل فرآیند انجام‌شده، از تحلیل‌های اولیه داده‌ها (EDA) تا ساخت مدل‌های مختلف و انتخاب مدل نهایی، می‌توان عوامل مؤثر بر مدت زمان تحویل را به ترتیب اهمیت زیر جمع‌بندی کرد:

مهم‌ترین و قوی‌ترین عامل مؤثر بر مدت زمان تحویل، طول واقعی مسیر (route_length_km) است. این متغیر در تمام مدل‌های ساخته‌شده، چه در مدل کامل، چه در مدل ساده‌شده با حذف پسرو و چه در مدل نهایی تعاملی، همواره دارای اثر بسیار قوی و معنادار آماری بوده است. ضرایب بزرگ این متغیر و p-value بسیار کوچک آن نشان می‌دهد که افزایش طول مسیر، به‌طور مستقیم و قابل‌توجهی باعث افزایش زمان تحویل می‌شود. همچنین حذف سایر متغیرها در فرآیند Backward Elimination بدون افت محسوس در دقت پیش‌بینی، اهمیت بنیادی این متغیر را تأیید می‌کند.

عامل مهم بعدی، منطقه مشتری (customer_zone) است، اما نه به‌صورت یک اثر مستقل، بلکه در تعامل با طول مسیر. نتایج مدل‌های تعاملی نشان داد که تأثیر هر کیلومتر افزایش مسیر در همه مناطق یکسان نیست. به‌طور خاص، در منطقه East، تعامل route_length_km با customer_zone به‌صورت معنادار ظاهر شد، به این معنا که در این منطقه، افزایش طول مسیر باعث افزایش شدیدتری در زمان تحویل می‌شود. این یافته

نشان می‌دهد که ویژگی‌های زیرساختی یا ترافیکی مناطق مختلف، نقش مهمی در عملکرد سیستم تحویل دارند.

در کنار این عوامل، سطح ترافیک (traffic_level) نیز در تحلیل‌های اکتشافی اولیه اثر قابل توجهی نشان داد، اما در مدل‌های نهایی به دلیل هم‌پوشانی اطلاعاتی با متغیرهایی مانند طول مسیر و زمان روز، اثر مستقل آن کاهش یافت و در فرآیند انتخاب مدل حذف شد. این به معنای بی‌اهمیت بودن ترافیک نیست، بلکه نشان می‌دهد که اطلاعات آن تا حد زیادی توسط متغیرهای قوی‌تر مدل جذب شده است.

در مقابل، عواملی مانند وضعیت آب‌وهوا، نوع وسیله تحویل و منطقه رستوران در این مجموعه داده تأثیر معنادار و پایداری بر زمان تحویل نداشتند. این نتیجه لزوماً به معنای بی‌اثر بودن این عوامل در دنیای واقعی نیست، بلکه نشان می‌دهد که در داده‌های موجود، اثر آن‌ها نسبت به عوامل قوی‌تری مانند مسیر و زمان، قابل تفکیک نبوده است.

ـ نحوه تصمیم گیری و راهکارهای عملی برای کاهش مدت زمان تحویل بر اساس نتایج آماری

بررسی p-value به‌تنهایی برای تصمیم‌گیری کاملاً ناکافی است. p-value فقط به ما می‌گوید که آیا اثر مشاهده‌شده یک متغیر از نظر آماری تصادفی است یا خیر، اما اطلاعاتی درباره شدت اثر، اهمیت عملی و کاربرد مدیریتی آن ارائه نمی‌دهد. در فرآیند تصمیم‌گیری، علاوه بر p-value، باید به موارد زیر توجه کرد:

اول، اندازه ضریب (Estimate). یک متغیر ممکن است p-value کوچکی داشته باشد اما ضریب آن بسیار کوچک باشد و در عمل تأثیر ناچیزی بر زمان تحویل بگذارد. در مقابل، route_length_km هم p-value بسیار کوچک دارد و هم ضریب بزرگ، که آن را به مهم‌ترین عامل عملی تبدیل می‌کند.

دوم، عملکرد مدل روی داده‌های آزمون (Test). نتایج شما نشان داد که برخی مدل‌های ساده‌تر، با وجود داشتن متغیرهای کمتر، عملکردی برابر یا حتی بهتر از مدل‌های پیچیده‌تر روی داده‌های تست دارند. این موضوع اهمیت

جلوگیری از بیش‌برازش و تمرکز بر قدرت پیش‌بینی واقعی مدل را نشان می‌دهد.

سوم، قابلیت کنترل از دید کسب‌وکار. برخی عوامل مانند منطقه مشتری یا زمان سفارش قابل کنترل مستقیم نیستند، اما عواملی مانند مسیر، تخصیص پیک و سیاست‌های عملیاتی قابل مدیریت‌اند. بنابراین تصمیم‌گیری باید بر عواملی متمرکز شود که هم اثر قوی دارند و هم امکان مداخله مدیریتی در آن‌ها وجود دارد.

با توجه به اینکه طول مسیر مهم‌ترین عامل قابل کنترل است، مؤثرترین راهکارها باید حول مدیریت مسیر و شرایط وابسته به آن طراحی شوند.

نخست، بهینه‌سازی پویا و هوشمند مسیرها. سیستم‌های مسیریابی باید نه‌تنها کوتاه‌ترین مسیر، بلکه سریع‌ترین مسیر را با در نظر گرفتن منطقه مشتری و بازه زمانی روز انتخاب کنند. نتایج مدل تعاملی نشان می‌دهد که یک مسیر یکسان در مناطق مختلف یا ساعات مختلف می‌تواند اثر متفاوتی بر زمان تحویل داشته باشد.

دوم، تخصیص منطقه‌ای منابع. از آنجا که اثر طول مسیر در برخی مناطق (مانند East) شدیدتر است، می‌توان پیک‌های باتجربه‌تر، شیفت‌های بیشتر یا وسایل نقلیه سریع‌تر را به این مناطق اختصاص داد، به‌ویژه در ساعات اوج.

سوم، مدیریت تقاضا بر اساس زمان. به‌جای تمرکز بر عواملی مانند آب‌وهوا که غیرقابل کنترل هستند، می‌توان با سیاست‌های تشویقی، مشتریان را به سفارش در ساعات خلوت‌تر ترغیب کرد. این کار باعث توزیع یکنواخت‌تر تقاضا و کاهش فشار عملیاتی در ساعات اوج می‌شود. برای مثال میتوان گفت:

اگرچه ممکن است در داده‌ها مشاهده شود که در روزهای آفتابی زمان تحویل کمتر است، اما این رابطه صرفاً یک همبستگی است، نه یک رابطه علّی مستقیم. هوای آفتابی به‌احتمال زیاد با شرایطی مانند ترافیک روان‌تر یا رانندگی آسان‌تر همراه است، نه اینکه خودِ آفتابی بودن مستقیماً باعث کاهش زمان تحویل شود. بنابراین تصمیم‌هایی مانند ارائه تخفیف صرفاً بر اساس وضعیت هوا، از نظر تحلیلی و مدیریتی استراتژی مناسبی نیستند. تمرکز باید بر عواملی

باشد که یا قابل کنترل‌اند یا می‌توان رفتار مشتری را نسبت به آن‌ها مدیریت کرد، مانند زمان سفارش و مسیر تحویل.