Dataset Documentation

(Data Sheet + Data Dictionary)

Project: MetroGo

این سند با هدف مستندسازی کامل داده‌هایی تهیه شده است که در پروژه MetroGo مورد استفاده قرار می‌گیرند یا در آینده می‌توانند مبنای تحلیل‌های پیشرفته و مدل‌های AI/ML قرار گیرند. فرض بنیادین این سند آن است که داده صرفاً یک ورودی فنی برای سیستم نیست، بلکه یک دارایی راهبردی محسوب می‌شود که کیفیت، منشأ، مالکیت و محدودیت‌های آن می‌تواند به‌طور مستقیم بر اعتبار تصمیم‌های محصول، تجربه کاربر و حتی ریسک‌های حقوقی و تجاری پروژه اثر بگذارد. از همین رو، این مستند تلاش می‌کند با شفافیت کامل نشان دهد داده‌ها از کجا می‌آیند، تحت چه شرایطی قابل استفاده هستند، چه ضعف‌ها و سوگیری‌هایی دارند و چگونه باید تفسیر شوند.

منبع داده(Data Source)

داده‌های مورد استفاده در MetroGo از ترکیب چند منبع مختلف به دست می‌آیند که هر کدام ماهیت، سطح اطمینان و محدودیت‌های خاص خود را دارند. بخش مهمی از داده‌ها مستقیماً از تعامل کاربران با اپلیکیشن MetroGo تولید می‌شود. این داده‌ها شامل اطلاعات مربوط به جستجوی مسیر، انتخاب خط مترو، ایستگاه مبدأ و مقصد، زمان تقریبی سفر، زمان استفاده از اپلیکیشن، و در صورت وجود، رفتارهای مرتبط با خرید یا مشاهده بلیت هستند. این داده‌ها به‌صورت رویدادمحور (event-based) ثبت می‌شوند و بازتاب‌دهنده رفتار واقعی کاربران در شرایط عملیاتی هستند.

در کنار داده‌های رفتاری کاربران، MetroGoاز داده‌های ساختاری و مرجع نیز استفاده می‌کند. این داده‌ها شامل اطلاعات ثابت یا نیمه‌ثابت درباره خطوط مترو، ایستگاه‌ها، زمان‌بندی رسمی حرکت قطارها، ظرفیت خطوط، و در صورت دسترسی، داده‌های عمومی یا سازمانی منتشرشده توسط نهادهای مسئول حمل‌ونقل شهری است. این داده‌ها معمولاً از منابع رسمی یاAPI های معتبر دریافت می‌شوند و نقش «ground truth» یا مرجع را در بسیاری از تحلیل‌ها ایفا می‌کنند.

تفاوت این دو نوع منبع داده از نظر پایداری و ریسک کاملاً در این سند لحاظ شده است. داده‌های کاربرمحور پویا، وابسته به رفتار انسان و مستعد تغییرات ناگهانی هستند، در حالی که داده‌های مرجع رسمی معمولاً پایدارتر اما گاهی با تأخیر به‌روزرسانی می‌شوند. این تفاوت‌ها در طراحی هرگونه تحلیل یا مدل ML باید در نظر گرفته شوند.

مجوز استفاده و رضایت کاربر(Usage License & Consent)

از آنجا که بخش قابل توجهی از داده‌های MetroGo مستقیماً از کاربران نهایی به دست می‌آید، مسئله رضایت آگاهانه و محدودیت استفاده از داده‌ها اهمیت حیاتی دارد. جمع‌آوری داده‌ها صرفاً در چارچوب سیاست حریم خصوصی اپلیکیشن و پس از اعلام شفاف نوع داده‌های جمع‌آوری‌شده، هدف استفاده و مدت نگهداری انجام می‌شود. کاربران در زمان استفاده از سرویس از این موضوع آگاه هستند که داده‌های رفتاری آن‌ها ممکن است برای بهبود کیفیت خدمات، تحلیل الگوهای استفاده و توسعه قابلیت‌های جدید استفاده شود.

این داده‌ها به هیچ عنوان خارج از دامنه اعلام‌شده مورد استفاده قرار نمی‌گیرند. به‌ویژه، در صورتی که داده‌ها در آینده برای آموزش یا ارزیابی مدل‌های AI/ML استفاده شوند، این استفاده همچنان در چارچوب بهبود سرویس و بدون شناسایی مستقیم افراد خواهد بود. داده‌ها پیش از هرگونه تحلیل پیشرفته، به‌صورت ناشناس‌سازی‌شده (anonymized یا pseudonymized) مورد استفاده قرار می‌گیرند تا ریسک نقض حریم خصوصی به حداقل برسد.

این سند همچنین تأکید می‌کند که داده‌های کاربران دارایی MetroGo نیستند، بلکه داده‌هایی هستند که با اعتماد کاربر و تحت شرایط مشخص در اختیار سیستم قرار گرفته‌اند. هرگونه ابهام در مالکیت یا استفاده از این داده‌ها می‌تواند این دارایی بالقوه را به یک بدهی حقوقی و اعتباری تبدیل کند؛ به همین دلیل شفافیت در این بخش آگاهانه بیش از حد حداقلی رعایت شده است.

کیفیت داده(Data Quality)

کیفیت داده یکی از محورهای اصلی این سند است، زیرا هرگونه تحلیل یا مدل ML به‌طور مستقیم تحت تأثیر کیفیت داده ورودی قرار دارد. داده‌های MetroGo در عمل با چالش‌های متعددی مواجه هستند. برای مثال، ممکن است برخی رویدادهای کاربر به‌طور کامل ثبت نشوند؛ کاربر ممکن است اپلیکیشن را ببندد، اتصال اینترنت قطع شود یا سیستم عامل اجازه ثبت کامل event را ندهد. این موارد منجر به داده‌های ناقص یا ناپیوسته می‌شوند.

علاوه بر این، خطاهای زمانی نیز ممکن است وجود داشته باشند؛ برای مثال اختلاف بین زمان ثبت‌شده در دستگاه کاربر و زمان سرور. این موضوع در تحلیل‌های حساس به زمان، مانند تخمین مدت سفر، اهمیت ویژه‌ای دارد. در این سند به‌صراحت ذکر می‌شود که داده‌های زمانی قبل از استفاده نیازمند نرمال‌سازی و اعتبارسنجی هستند.

کیفیت داده‌های مرجع نیز بدون نقص نیست. داده‌های رسمی حمل‌ونقل شهری ممکن است به‌روز نباشند یا تغییرات عملیاتی (مانند تأخیر یا تغییر مسیر) را به‌صورت بلادرنگ منعکس نکنند. بنابراین، حتی داده‌های «رسمی» نیز نیازمند تفسیر محتاطانه هستند.

) Missingnessداده‌های گمشده(

وجود داده‌های گمشده یک واقعیت اجتناب‌ناپذیر در سیستم‌های واقعی مانند MetroGo است. این سند به‌طور شفاف توضیح می‌دهد که داده‌های گمشده الزاماً تصادفی نیستند. برای مثال، کاربرانی که سفر خود را نیمه‌کاره رها می‌کنند یا از اپلیکیشن فقط برای مشاهده استفاده می‌کنند، ممکن است الگوی missingnessمتفاوتی نسبت به کاربران فعال داشته باشند.

این موضوع از نظر تحلیلی اهمیت زیادی دارد، زیرا حذف ساده این داده‌ها می‌تواند منجر به سوگیری شود. بنابراین، سیاست برخورد با داده‌های گمشده (حذف، جایگزینی، یا تحلیل جداگانه) باید متناسب با هدف تحلیل یا مدل ML تعریف شود و این تصمیمات در مستندات فنی بعدی ارجاع داده می‌شوند.

) Biasسوگیری داده(

یکی از مهم‌ترین بخش‌های این سند، شناسایی و پذیرش سوگیری‌های ذاتی داده است. کاربران MetroGoنماینده کل جمعیت شهر نیستند. این کاربران معمولاً دسترسی به گوشی هوشمند، اینترنت و تمایل به استفاده از ابزارهای دیجیتال دارند. بنابراین، داده‌ها ممکن است رفتار گروه‌هایی از جامعه را کمتر یا بیشتر از واقعیت بازتاب دهند.

این سوگیری به‌ویژه در صورت استفاده از داده‌ها برای تصمیم‌گیری‌های کلان یا مدل‌های پیش‌بینی باید جدی گرفته شود. این سند تأکید می‌کند که داده‌های MetroGo بیشتر برای بهبود تجربه کاربران فعلی مناسب هستند و استفاده از آن‌ها برای تعمیم‌های اجتماعی گسترده نیازمند احتیاط و تکمیل با داده‌های دیگر است.

) Driftتغییر توزیع داده(

رفتار کاربران حمل‌ونقل شهری به‌شدت وابسته به زمان، شرایط اقتصادی، تغییرات شهری و حتی رویدادهای غیرمنتظره است. بنابراین، داده‌هایی که امروز جمع‌آوری می‌شوند ممکن است در چند ماه آینده دیگر نماینده رفتار واقعی نباشند. این پدیده که به‌عنوان data drift شناخته می‌شود، در MetroGoیک ریسک شناخته‌شده است.

این سند تأکید می‌کند که هرگونه مدل ML مبتنی بر داده‌های گذشته باید به‌صورت دوره‌ای بازبینی شود و استفاده بلندمدت بدون مانیتورینگ می‌تواند منجر به تصمیم‌های نادرست شود.

تعریف دقیق Features و(Labels (Data Dictionary

در بخشData Dictionary ، هر ویژگی داده‌ای به‌صورت دقیق تعریف می‌شود. برای هر feature مشخص می‌شود که دقیقاً چه چیزی را اندازه‌گیری می‌کند، چگونه تولید شده و چه محدودیت‌هایی دارد. این شفافیت از سوءتفسیر داده جلوگیری می‌کند؛ برای مثال، تفاوت بین «زمان تخمینی سفر» و «زمان واقعی ثبت‌شده» باید کاملاً روشن باشد.

در صورتی که داده‌ها برای آموزش مدل ML استفاده شوند، تعریف labels نیز به‌طور دقیق انجام می‌شود تا مشخص باشد مدل دقیقاً چه چیزی را یاد می‌گیرد و چه چیزی را نه.

تفسیر داده به‌عنوان دارایی در مقابل بدهی(Data as Asset vs Data as Liability)

در پروژهMetroGo ، داده به‌عنوان یک دارایی بالقوه تلقی می‌شود، اما این دارایی تنها در صورتی ارزشمند خواهد بود که ویژگی‌های بنیادی آن شفاف، قابل اتکا و مستند باشد. داده‌ای که منبع آن نامشخص است، کیفیت آن سنجیده نشده یا محدودیت‌های حقوقی و اخلاقی استفاده از آن مشخص نیست، به‌جای آنکه به مزیت رقابتی تبدیل شود، می‌تواند به یک بدهی پنهان اما پرهزینه بدل گردد. این بدهی ممکن است در قالب تصمیم‌های اشتباه محصول، نارضایتی کاربران، ریسک‌های قانونی یا حتی توقف توسعه برخی قابلیت‌ها در آینده خود را نشان دهد.

MetroGoبا این فرض طراحی شده است که داده نه فقط برای گزارش‌گیری، بلکه برای یادگیری سیستم از رفتار کاربران و بهبود مستمر تجربه سفر شهری استفاده شود. بنابراین، این سند به‌طور آگاهانه تلاش می‌کند مرز بین «داده سالم و قابل استفاده» و «داده مبهم و پرریسک» را مشخص کند. این رویکرد به تیم اجازه می‌دهد پیش از آنکه داده وارد چرخه تصمیم‌سازی یا مدل‌سازی شود، درباره کیفیت و پیامدهای استفاده از آن آگاهانه تصمیم بگیرد.

چرخه تولید داده درMetroGo

داده در MetroGo به‌صورت پیوسته و در طول چرخه تعامل کاربر با محصول تولید می‌شود. این چرخه از لحظه‌ای آغاز می‌شود که کاربر اپلیکیشن را باز می‌کند و شامل مراحل مختلفی مانند جستجوی مسیر، مشاهده پیشنهادها، انتخاب مسیر، و در صورت وجود، استفاده از قابلیت‌های تکمیلی مانند اطلاع از زمان حرکت یا اختلالات احتمالی است. هر یک از این مراحل می‌تواند داده‌ای تولید کند که درک دقیق آن برای تحلیل رفتار کاربر ضروری است.

با این حال، مهم است تأکید شود که داده‌های تولیدشده الزاماً بازتاب‌دهنده نیت کامل کاربر نیستند. برای مثال، ممکن است کاربر چندین مسیر را صرفاً برای مقایسه مشاهده کند، بدون آنکه قصد استفاده واقعی از همه آن‌ها را داشته باشد. بنابراین، تفسیر داده‌ها بدون در نظر گرفتن زمینه رفتاری می‌تواند منجر به برداشت‌های نادرست شود. این سند تلاش می‌کند این تفاوت بین «سیگنال داده‌ای» و «واقعیت رفتاری» را شفاف کند.

محدودیت‌های ذاتی داده‌های کاربرمحور

داده‌هایی که مستقیماً از کاربران جمع‌آوری می‌شوند، به‌شدت وابسته به شرایط بیرونی هستند. عواملی مانند کیفیت اتصال اینترنت، نوع دستگاه، نسخه سیستم‌عامل، و حتی سطح سواد دیجیتال کاربران می‌توانند بر کامل بودن و دقت داده‌ها اثر بگذارند. برای مثال، کاربران با دستگاه‌های قدیمی‌تر ممکن است برخیevent ها را به‌طور ناقص ارسال کنند یا کاربران در شرایط شلوغ شهری ممکن است تعامل کوتاه‌تری با اپلیکیشن داشته باشند.

این محدودیت‌ها به این معناست که نبود داده در برخی بخش‌ها لزوماً به معنای نبود رفتار نیست. در MetroGo، این موضوع به‌عنوان یک اصل تحلیلی پذیرفته شده و در طراحی هرگونه استفاده تحلیلی یا ML از داده‌ها لحاظ می‌شود. این سند تأکید می‌کند که تحلیل داده بدون درک این محدودیت‌ها می‌تواند تصویری ساده‌انگارانه و حتی گمراه‌کننده از رفتار کاربران ارائه دهد.

داده‌های حساس و سطح ریسک آن‌ها

اگرچه MetroGo به‌طور مستقیم داده‌های بسیار حساس مانند اطلاعات هویتی رسمی یا مالی کاربران را جمع‌آوری نمی‌کند، اما داده‌های رفتاری مرتبط با الگوی جابه‌جایی شهری می‌توانند به‌طور غیرمستقیم حساس تلقی شوند. الگوهای سفر، ساعات تردد و مسیرهای پرتکرار می‌توانند اطلاعات معناداری درباره سبک زندگی کاربران آشکار کنند. از همین رو، این داده‌ها با سطح بالایی از احتیاط مدیریت می‌شوند.

این سند به‌صراحت بیان می‌کند که حتی در مراحل اولیه توسعه، داده‌های رفتاری نباید صرفاً به‌عنوان داده‌های بی‌خطر در نظر گرفته شوند. رویکرد MetroGo بر این اساس است که پیشگیری از ریسک‌های حریم خصوصی باید از ابتدا در طراحی داده لحاظ شود، نه اینکه پس از بروز مشکل به آن پرداخته شود.

تعریف عملیاتی Drift در داده‌هایMetroGo

تغییر توزیع داده‌ها در MetroGo نه‌تنها محتمل، بلکه تقریباً اجتناب‌ناپذیر است. تغییرات فصلی، تعطیلات، تغییر ساعات کاری، توسعه خطوط جدید مترو یا حتی رویدادهای اجتماعی می‌توانند الگوی استفاده کاربران را به‌طور اساسی تغییر دهند. این تغییرات ممکن است به‌صورت تدریجی یا ناگهانی رخ دهند.

این سند تأکید می‌کند که داده‌های جمع‌آوری‌شده در یک بازه زمانی خاص نباید بدون بازبینی در بازه‌های زمانی بعدی استفاده شوند. هرگونه مدل یا تحلیل مبتنی بر داده باید به این سؤال پاسخ دهد که داده مورد استفاده متعلق به چه دوره‌ای است و آیا هنوز نماینده شرایط فعلی سیستم هست یا خیر. نادیده گرفتن drift می‌تواند باعث شود سیستم بر اساس الگوهایی تصمیم بگیرد که دیگر وجود خارجی ندارند.

Biasساختاری ناشی از کانال دسترسی

یکی از منابع مهم سوگیری در داده‌هایMetroGo ، کانال دسترسی به محصول است. کاربرانی که MetroGoرا نصب و استفاده می‌کنند، الزاماً نماینده تمام استفاده‌کنندگان مترو نیستند. این کاربران معمولاً آگاه‌تر، دیجیتال‌محورتر و در برخی موارد جوان‌تر از میانگین جامعه هستند. این واقعیت به این معناست که داده‌ها ممکن است برخی نیازها یا مشکلات گروه‌های دیگر را کمتر منعکس کنند.

این سند تأکید می‌کند که این سوگیری باید در تفسیر داده‌ها و تصمیم‌های محصول لحاظ شود. برای مثال، اگر داده‌ها نشان دهند که یک مسیر خاص کمتر استفاده می‌شود، این نتیجه نباید بدون بررسی زمینه به‌عنوان کاهش نیاز واقعی تفسیر شود؛ ممکن است کاربران آن مسیر کمتر از ابزارهای دیجیتال استفاده کنند.

تعریف دقیقFeature ها در زمینه کاربردی

درMetroGo ، هر feature داده‌ای تنها یک ستون عددی یا متنی نیست، بلکه بازنمایی یک مفهوم عملیاتی در دنیای واقعی است. برای مثال، «زمان تخمینی سفر» ترکیبی از داده‌های مرجع، الگوریتم‌های محاسباتی و فرضیات سیستم درباره شرایط عادی تردد است. این مقدار لزوماً زمان واقعی تجربه‌شده توسط کاربر نیست و این تفاوت باید در هرگونه تحلیل یا مدل‌سازی در نظر گرفته شود.

این سند به‌طور مفصل توضیح می‌دهد که هر feature چگونه تولید می‌شود، چه فرض‌هایی پشت آن قرار دارد و در چه شرایطی ممکن است دقت آن کاهش یابد. این سطح از شفافیت به تیم کمک می‌کند از استفاده نادرست یا بیش‌ازحد از داده‌ها جلوگیری کند.

تعریف Labels و ریسک‌های مرتبط با آن‌ها

در صورتی که داده‌ها برای آموزش مدل‌های ML استفاده شوند، تعریف labels اهمیت ویژه‌ای پیدا می‌کند. درMetroGo ، labelsممکن است بر اساس رفتار مشاهده‌شده کاربر تعریف شوند، نه نیت واقعی او. برای مثال، انتخاب یک مسیر لزوماً به معنای رضایت از آن مسیر نیست. این تفاوت ظریف اما مهم، در این سند به‌عنوان یک ریسک بالقوه برای مدل‌سازی شناسایی شده است.

این سند تأکید می‌کند که labels باید با آگاهی کامل از محدودیت‌های داده تعریف شوند و نتایج مدل‌ها نباید بدون تفسیر انسانی و زمینه‌ای مورد استفاده قرار گیرند.