**Datasheet: RetailRocket E-Commerce Dataset**

Adapted from Gebru et al., "Datasheets for Datasets" (2021), for capstone project transparency.

**Motivation**

- Purpose: The RetailRocket dataset supports my capstone project by providing e-commerce user interaction data to train DistilBERT for generating personalized ad creatives and predicting click-through rates (CTR). It enables analysis of user behavior (e.g., views) to simulate ad engagement.

- Creators: RetailRocket, a personalization platform, collected the data to advance recommendation system research.

- Funding: Not disclosed; dataset is publicly available via Kaggle.

- Gap Filled: Offers a lightweight (~94.2 MB), anonymized e-commerce dataset suitable for academic projects, unlike proprietary or larger datasets (e.g., Amazon Product Data).

- Comments: Chosen for its compatibility with Azure ML Studio’s Free Tier and manageable preprocessing time (~4 minutes for 50,000 records).

**Composition**

- Instances: User interactions (views, add-to-cart, purchases) with fields: `visitorid`, `itemid`, `timestamp`, `event`, `transactionid`.

- Count: ~2.75 million interactions; subset of 50,000 view events used for Week 2 to reduce compute load.

- Sample: First 50,000 view events (non-random, sequential by timestamp), sufficient for initial DistilBERT testing.

- Sample Representativeness: Focuses on view events, excluding add-to-cart (~1.2%) and purchases (~0.5%), potentially underrepresenting conversion behaviors.

- Data Type: Raw interaction logs (numerical IDs, timestamps, categorical events).

- Labels: Event type (`view`, `addtocart`, `transaction`) used as proxy for CTR prediction.

- Missing Data: ~10% of `transactionid` fields are `NaN` (expected for non-purchase events); handled by filtering to view events.

- Relationships: Implicit user-item interactions (e.g., multiple views by a `visitorid` for an `itemid`).

- Splits: Planned 80% train, 10% validation, 10% test, based on chronological timestamps.

- Errors/Noise: Potential duplicate events (e.g., repeated views); mitigated by deduplication in preprocessing.

- External Resources: Sourced from Kaggle (https://www.kaggle.com/retailrocket/ecommerce-dataset); no additional metadata used.

- Confidentiality: Fully anonymized; no personally identifiable information.

- Sensitive Content: None identified (e.g., no demographic or sensitive user data).

- Subpopulations: Not explicitly defined; assumes general e-commerce users.

- Individual Identification: Not possible due to anonymized IDs.

- Comments: Subset simplifies processing for Week 2 milestones; full dataset may be used in later weeks for robustness.

**Preprocessing**

- Cleaning: Removed rows with missing `visitorid` or `itemid` (~0.1% of data).

- Transformations: Filtered to `view` events; sampled 50,000 records; added synthetic ad text (`Discover item [itemid] now!`).

- Tools: Pandas in Azure ML Studio (Python 3.10 - PyTorch and TensorFlow environment).

- Comments: Preprocessing takes ~4 minutes on `pupbhavani\_training` (8 cores, 56 GB RAM), suitable for Free Tier constraints.