UMBC

Data 606 - Data Collection

Team Members:
Chandra Sekhar Katipalli
Sindura Reddy Challa
Sanjana Reddy Soma

Project 1- Real-Time Automated Data Pipeline for Advertising

Ipinyou - Numerical data (e.g., bid price, impressions, clicks) and categorical data (e.g., campaign IDs, ad slot IDs), with many categorical values hashed/unlabeled.
Criteo - Massive log for CTR prediction. Contains numerical & hashed categorical features.
Avazu - Mobile ad click-through data. Uses numerical data and hashed categorical features.

Data Sources
https://www.kaggle.com/c/avazu-ctr-prediction/data
https://ailab.criteo.com/criteo-1tb-click-logs-dataset/
https://contest.ipinyou.com/

- All data is anonymized (hashed/unlabeled).
- Aim shifts from Campaign Specific Predictions.
- Focuses on aggregated analysis and predictive modeling.

Due to anonymization (hashed/unlabeled categorical features), our analysis focuses on aggregated patterns instead of individual campaign-specific predictions.

Real-Time Automated Data Pipeline for Advertising

Integrating LLMs for Campaign Analysis

- Embedding Generation: Convert campaign names into numerical vectors (embeddings) to capture semantic meanings.

  Example:"Coca-Cola Holiday Cheer" → [0.12, 0.45, 0.67, ...]

  "Coke Xmas Special" → [0.11, 0.46, 0.66, ...]

- Similarity Assessment: Measure the cosine similarity between embeddings to determine the relatedness of campaigns.

  A cosine similarity close to 1 indicates high similarity.

- Clustering: Group campaigns with similar embeddings to reveal patterns and common themes.
- Cluster Labeling: Assign descriptive labels to each cluster to enhance interpretability.
- Predictive Modeling: Use insights from clusters to develop models that predict key performance indicators for new campaigns.

To facilitate this, we've collaborated with industry experts to generate a synthetic dataset that mirrors real-world advertising analytics.

Real-Time Automated Data Pipeline for Advertising

We are taking AI-generated data.- Generated a synthetic dataset reflecting real-world analytics

Schema Attributes

Commerce_CommerceorBrand
Day
Month_Number
Data_Stream
Advertiser_Name
Campaign_Name
Campaign_Key
Brand_Click_Sales
Video_Views
Add_to_Cart

Clean_Creative_Name_calc
Commerce_Service_Type
Commerce_Clean_Campaign_Name
Commerce_Funding_Source
Commerce_Partner
Commerce_Onsite_Offsite
Commerce_Channel
Commerce_Subbrand
Clean_Placement_Name_Calc
Commerce_Brand

Site_Commerce_Class
Retailer
Impressions
Clicks
Sales
Sale_Units
Revenue
Orders
Viewability_Percentage
Attributable_Sales

Real-Time Automated Data Pipeline for Advertising

Key Variables

impressions	The total number of times an advertisement was displayed to users.
clicks	The number of times users clicked on the advertisement.
sales	The total number of successful transactions generated from the ad.
sale_units	The number of individual product units sold through the advertisement
revenue	The total income generated from ad-related product sales (in currency)
attributable_sales	The revenue directly linked to the ad campaign's influence on purchases.
advertiser_name	The name of the company or entity running the ad campaign.
campaign_name	The specific marketing initiative or promotion being tracked.
clean_placement_name_calc	A standardized name for the ad's display location
media_buy_name	The method or channel used to purchase advertising space.

Project 2- Predictive Model for Highway Deterioration Forecasting

Primary Data Sources for Road Deterioration Prediction

Freight Analysis Framework (FAF4.5) Dataset	**Highway Performance Monitoring System (HPMS) Dataset**
● **Source:** U.S. Bureau of Transportation Statistics (FAF4.5) ● **Data Format:** CSV, Shapefiles ● **Total Variables:** 16 (**Selected 4 key variables**)	● **Source:** U.S. Federal Highway Administration (HPMS Data) ● **Data Format:** CSV, Shapefiles ● **Total Variables:** Varies by state submissions

LTPP (Long-Term Pavement Performance) Data Exclusion LTPP Data

- **Limited or inactive updates (2013–2018)** → Data is outdated and lacks recent trends.
- **Data incompatibility** → Inconsistent format and variables compared to other datasets.
- **Unable to combine datasets** → Due to structural differences, there will be challenges in merging with HPMS & FAF
- **Fewer road sections covered** → Limited geographic coverage reduces predictive model accuracy.

UMBC

FAF Dataset: Tracking Freight Movement

Column	Description	Why it matters?
dms_orig	Origin FAF region (where freight movement begins)	Starting point of the freight. Can be used to link datasets.
dms_dest	Destination FAF region (where freight movement ends)	Freight ending point. Can also be used to filter and link datasets
dms_mode	Mode of transport (Truck, Rail, Air, Water, etc.)	Helps determine if it is a mode of freight movement roadways, airways and seaways. We are concentrating on roadways
tons	Total weight of commodities shipped (in thousand tons)	Tons moved between the origin and destination. Deterioration may vary.

HPMS Dataset: Monitoring Pavement Conditions

Column	Description	Why it matters?
IRI	International Roughness Index	Helps determine the condition of the road.
AADT	Annual Average Daily Traffic	Average daily traffic movement on the section of road.
Pavement Type	Type of pavement (asphalt, concrete, etc.)	Materials used to build the section of road. Deterioration can vary based on materials.
Lane Miles	Total miles of lanes in a road segment	Shorter lanes can experience more deterioration.

THANK YOU