Conditional GAN-based Sales Prediction Model Using Reviews and Sales Data – Proposal

Li Siqi 122090271

Li Chudikang 122040057

Yichun Wang 122090550

1. Project Overview

Traditional recommendation systems are typically trained on historical data such as user behavior or sales records. However, they often struggle to perform well in rapidly changing market environments. Our project introduces a novel approach to recommendation by leveraging a Conditional Generative Adversarial Network (cGAN) to expand the training dataset.

We propose a two-stage algorithm named Arthurmend. Firstly we use cGAN algorism to generate future data before training the recommendation model. Then we combine the future date with the original dataset to form an augmented training set.

This project can give the recommendation system with a degree of temporal awareness and predictive capability, making it better suited for the current state of the market.

2. Data

Required Data:

- Sales data (Date + daily sales)
- Review data (Date + review text + rating)
- Temporal features (weekdays, weekend, holidays ...)

Source of datasets: Yelp Dataset, Kaggle e-commerce datasets ... etc.

Data preprocessing: Normalization, sentiment extraction from reviews (positive/negative), computing review counts and average ratings, and one-hot encoding of temporal features.

3. Method

To predict future sales with high accuracy, we propose a Conditional Generative Adversarial Network (cGAN) architecture that integrates multi-modal historical features as conditions to generate context-aware 7-day sales forecasts. The condition y comprises three key components: historical sales trends, review sentiment embeddings, and temporal features. Historical sales trends are encoded using a Long Short-Term Memory (LSTM) network, capturing patterns over a 30-day window to model temporal dependencies in sales data. Review sentiment embeddings are extracted using a pre-trained BERT model, which processes customer review texts to produce dense representations of positive or negative sentiment, supplemented by statistical features such as average ratings and review counts. Temporal features, including day-of-week, holidays, and promotional events, are encoded via one-hot vectors or learned embeddings to capture periodic and event-driven sales fluctuations.

In the cGAN framework, the Generator takes a random noise vector and the condition as inputs to produce a synthetic 7-day sales sequence. The Discriminator evaluates whether a given sequence (real or generated) is realistic under the same condition, ensuring the generated sales align with the contextual features. The training process alternates between optimizing the Discriminator to distinguish real sales sequences from generated ones and updating the Generator to produce sequences that better fool the Discriminator. To ensure training stability, we will also adopt the Wasserstein loss with gradient penalty (WGAN-GP) as our creative ideas.

Architecture Overview:

```
            Condition y (Sales + Reviews + Time)
                        |
                        ▼
Noise z ──→ G(z|y) ──→ Fake future sales ──┐
                                           ▼
```

Real future sales ─→ D(x|y)

Key Modules:

A. Feature Extraction: LSTM encodes past 7-day sales trends; BERT extracts sentiment embeddings from reviews; statistical features include average rating and review counts; time features are encoded via embeddings or one-hot.
B. Conditional Generator: Takes noise z and condition y (features) to generate future 7-day sales sequences.
C. Conditional Discriminator: Receives both real/fake sequences and the same condition y to judge whether the sequence is realistic under that context.

Training alternates between training the conditional discriminator and generator, following the standard cGAN objective functions.

4. Evaluation

To evaluate our approach, we will compare the performance of the recommendation system under two training settings: one using the original dataset, and the other using the cGAN-augmented dataset.

The results are then analyzed through quantitative metrics (such as accuracy, precision, and recall) as well as qualitative evaluation, which examines the interpretability and practical effectiveness of the recommendations.

This evaluation helps us assess how much improvement the cGAN-based data expansion contributes to the overall recommendation performance.

5. Expected Results

The proposed Conditional Generative Adversarial Network (cGAN) model is expected to yield significant improvements in sales forecasting accuracy and interpretability, offering both quantitative and qualitative benefits.

Quantitatively, the model aims to achieve a 10% reduction in Mean Absolute Error (MAE) compared to a normal recommendation system, leveraging the integration of multi-modal features (sales, review sentiment, and temporal factors) and data augmentation with generated sales sequences. This improvement is anticipated to be particularly pronounced during periods of strong sentiment fluctuations, such as surges in negative reviews, where traditional models often underperform. Qualitatively, the model will provide actionable insights into the relationship between customer sentiment and sales trends. Analysis is expected to reveal that spikes in negative sentiment, particularly associated with keywords like "defective" or "poor quality," precede sales declines by 1-2 weeks, enabling businesses to anticipate demand shifts.

6. Potential Challenges and Solutions

Challenge	Solution
Unstable GAN training	Use WGAN-GP or gradient penalty for stability
Limited data	Leverage public datasets and data augmentation
Long review texts	Truncate to 128 tokens and use pretrained BERT embeddings
Suboptimal performance	Fallback to conditional LSTM + attention mechanism