Adobe Behavior Simulation Challenge

Final Report

Team DDJ

Abstract

In the current digital marketing landscape, predicting and enhancing user engagement is crucial. Our project addresses this challenge through the task: Simulating customer behaviour by predicting the number of likes a tweet might receive based on its content. This task arises from brand's need for more intelligent ways to connect with their audience online. Understanding user engagement will enable marketers to refine their online strategies effectively.

1. Methodology for Predicting Likes

1.1 Dataset Overview

1. Training data: 300,000 examples
2. Test data: 10,000 examples
3. Dataset structure: 7 columns representing id, date, likes, content, username, media links, inferred company.

1.2 Data Preprocessing

1. Feature extraction from date:
2. Month
3. Is weekend
4. Day of week
5. Year
6. Image embedding generation:
7. Used EfficientNet-B0 for 1000-dimensional embeddings
8. Fine-tuned EfficientNet-B0 for 256-dimensional embeddings
9. Text embedding generation:
10. Utilized BERTweet for text embeddings

1.3 Approach Evolution

Initial Attempts:

1. Image-only model using EfficientNet-B0 embeddings
2. Text-only model using BERTweet embeddings
3. Experimented with various regression models:
4. LightGBM
5. XGBoost
6. Random Forest
7. CatBoost
8. Neural Networks

Cross-validation Strategy:

1. For unseen companies: Implemented groupKFold
2. For unseen dates: Removed year from the feature set

1.4 Final Approach

1. Combined BERTweet and EfficientNet-B0 embeddings
2. Utilized final embedding from EfficientNet-B0
3. Incorporated extracted date features
4. Trained two models on the combined features:
5. LightGBM
6. Neural Network
7. Created an ensemble of both models for final predictions

1.5 Challenges and Solutions

1. Challenge: Handling unseen companies in the test set
2. Solution: Concatenated content, username, and company string before passing it to BERTweet for embedding

2. Results and Evaluation

2.1 Performance Metrics

1. Achieved RMSE of 3.4k on the cross-validation set

3. Future Improvements

3.1 Feature Engineering:

1. Implement features for the number of hashtags and mentions in tweets
2. Incorporate video-specific features such as likes and duration

3.2 Model Optimization:

1. Conduct comprehensive hyperparameter tuning for all models in the ensemble

4. Conclusion

Our approach to the Adobe Behavior Simulation Challenge demonstrates the effectiveness of combining advanced embedding techniques with ensemble learning for predicting tweet engagement. By addressing the challenges of unseen companies and dates, we've created a robust model that can generalize well to new data. While our current results are promising, there's still room for improvement through additional feature engineering and model optimization.

This project highlights the complexity of predicting social media engagement and the importance of considering multiple factors such as text content, visual elements, and temporal features. As we continue to refine our approach, we aim to provide even more accurate predictions that can help brands optimize their social media strategies and connect more effectively with their target audiences.