Social Media Engagement Analysis & Predictive Modeling Report

1. Project Description

This project analyzes a Kaggle Social Media Engagement dataset to identify the key drivers of post engagement and to build predictive models for estimating total engagement per post.

The analysis is structured around practical business questions:

- Which platforms generate the most engagement?
- Which post formats perform best?
- Does sentiment impact engagement?
- How does posting time influence performance?
- Can engagement be predicted using available metadata?

The dataset consists of **100 posts and 9 columns**, including:

- `platform`
- `post_type`
- `post_time`
- `likes`
- `comments`
- `shares`
- `sentiment_score`

- `post_id`

The dataset is clean, with no missing values in the core engagement columns.

---

2. Data Preparation & Feature Engineering

2.1 Target Variable Creation

A unified engagement metric was created:

total_engagement = likes + comments + shares\textbf{total_engagement = likes + comments + shares}total_engagement = likes + comments + shares

This became the primary target variable for both analysis and modeling.

2.2 Time-Based Features

`post_time` was converted into a proper datetime format and used to derive:

- `post_hour`
- `post_dow` (day of week)
- `post_month`
- `is_weekend` (binary flag)
- `hour_bucket` (late night, morning, afternoon, evening)

These features allowed analysis of timing effects on engagement.

---

3. Exploratory Analysis & Key Findings

3.1 Platform Performance

Average engagement per post:

1. **Instagram** – highest average engagement
2. **Facebook** – second
3. **Twitter** – lowest

Interpretation:
Platform choice is a major driver of engagement outcomes in this dataset. Instagram appears to provide the strongest baseline performance.

3.2 Post Type Performance

Top-performing formats:

1. Polls
2. Videos
3. Carousels

Lowest-performing format:

- Text posts

Interpretation:
Interactive and rich-media formats generate stronger engagement than static text content. Content strategy should prioritize visual and interactive formats.

3.3 Sentiment Impact

Average engagement by sentiment:

- Negative – highest
- Positive – second
- Neutral – lowest

Interpretation:
Emotionally charged content (especially negative sentiment) correlates with higher engagement. This does not imply negative content is always preferable, but it suggests that emotional intensity influences reactions.

3.4 Timing Effects

Average engagement by time-of-day bucket:

Hour Bucket	Avg Engagement
Late Night	3558.85
Afternoon	2934.84
Morning	2796.68
Evening	2674.75

Interpretation:
Late-night posts performed best in this sample. Timing appears to be a meaningful—but not dominant—factor.

---

4. Predictive Modeling

Two models were built and evaluated using a standard train/test split:

- Ridge Regression (regularized linear model)
- Random Forest Regressor (tree-based ensemble)

4.1 Model Comparison

Model	MAE	RMSE	R^2
Ridge	1133.30	1455.36	0.2058
Random Forest	1220.81	1542.38	0.1080

4.2 Best Model: Ridge Regression

Ridge Regression achieved:

- **$R^2 \approx 0.206$**
- **MAE ≈ 1,133**
- **RMSE ≈ 1,455**

This means the model explains approximately **21% of the variance in engagement** on the test set.

Interpretation:

- There is measurable predictive signal in platform, post type, sentiment, and timing.
- However, nearly 80% of engagement variance remains unexplained.
- Engagement is influenced by unobserved factors such as:
    - Audience size
    - Content quality
    - Algorithm distribution
    - External events
    - Virality dynamics

In this small dataset, a simpler regularized linear model generalized better than a more flexible Random Forest.

---

5. Log-Transform Experiment

A log1p transformation of total engagement was tested to reduce skewness.

Results:

- RMSE: 1646.12
- MAE: 1392.45
- R^2 (original scale): -0.016
- R^2 (log scale): -0.042

Conclusion:
The log transformation reduced performance and produced negative R^2 values. In this dataset, log-scaling did not improve predictive alignment.

This is an important negative result and demonstrates appropriate model experimentation.

---

6. Diagnostic Output

Prediction diagnostics were exported for inspection:

`actual_vs_pred_and_residuals_log1p.csv`

Example rows:

Actual	Predicted	Residual
1883	3052.07	-1169.07
5482	3070.21	2411.79
4289	3096.69	1192.31

Residual analysis confirmed that predictions can deviate substantially for individual posts, especially high-engagement outliers.

---

7. Overall Conclusion

This project delivers two primary outcomes:

1️⃣ Actionable Content Insights

- Instagram performs best in this dataset.
- Polls and videos drive the strongest engagement.
- Negative sentiment correlates with higher average reactions.
- Late-night posting shows higher average engagement.

These insights can directly inform social media strategy.

2️⃣ Predictive Baseline Model

- Ridge Regression provides a stable and interpretable baseline.
- The model captures real patterns but demonstrates that engagement is only partially predictable from metadata.
- A large share of engagement variance is driven by factors not included in the dataset.

---

8. Limitations

- Small dataset (100 posts)
- No audience-size or follower data
- No reach/impressions data
- No content text analysis
- No algorithm exposure metrics