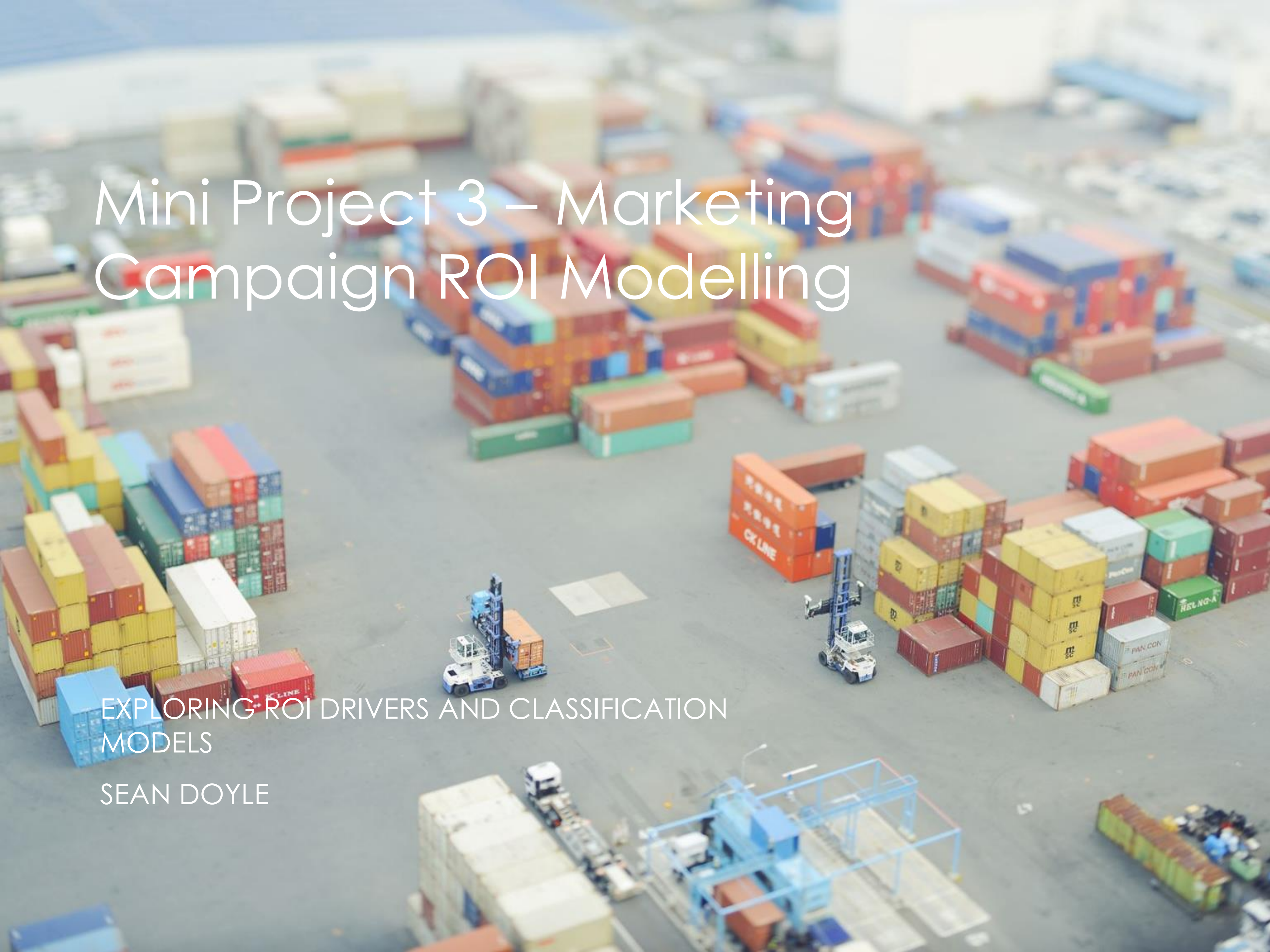


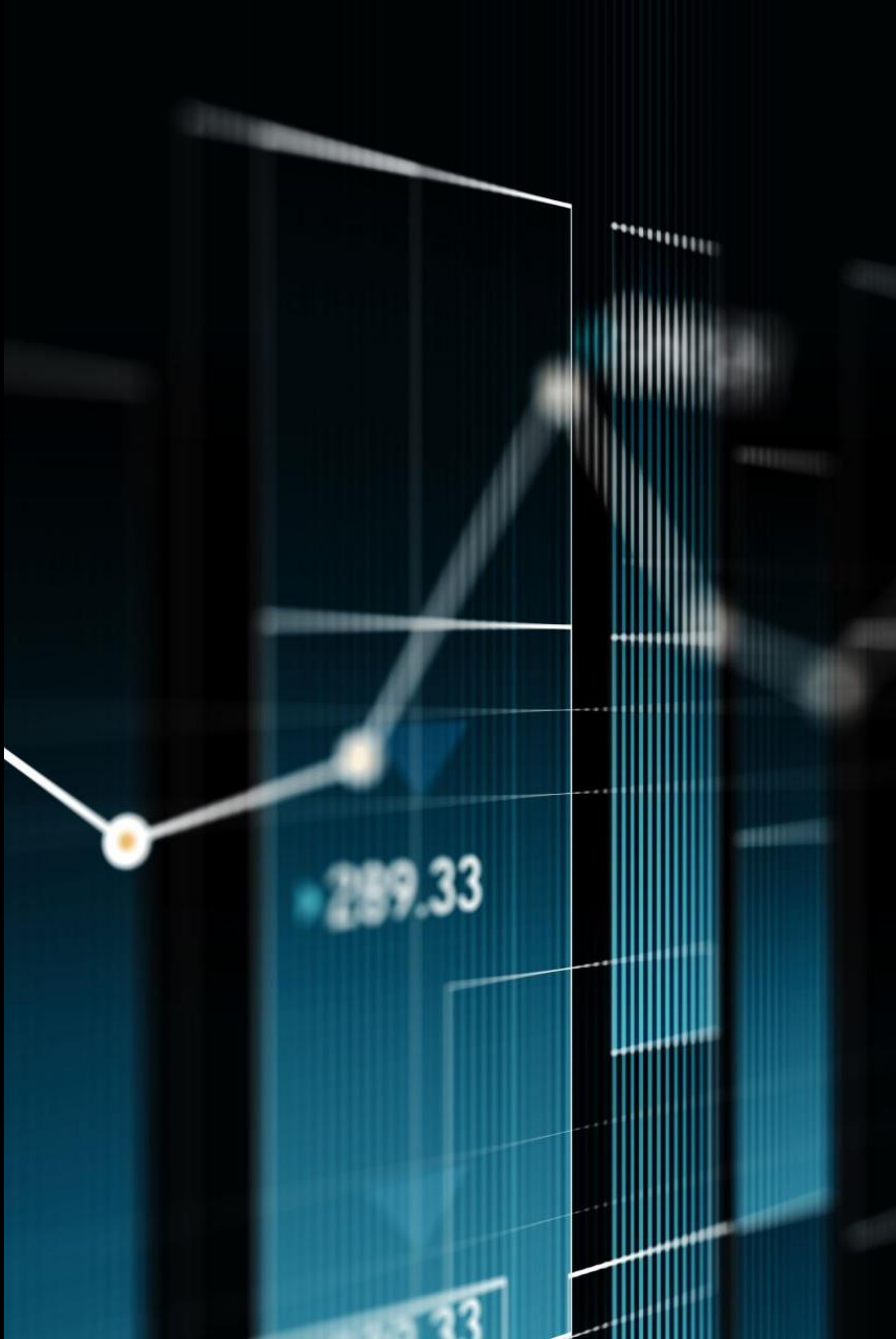
Mini Project 3 – Marketing Campaign ROI Modelling

EXPLORING ROI DRIVERS AND CLASSIFICATION MODELS

SEAN DOYLE

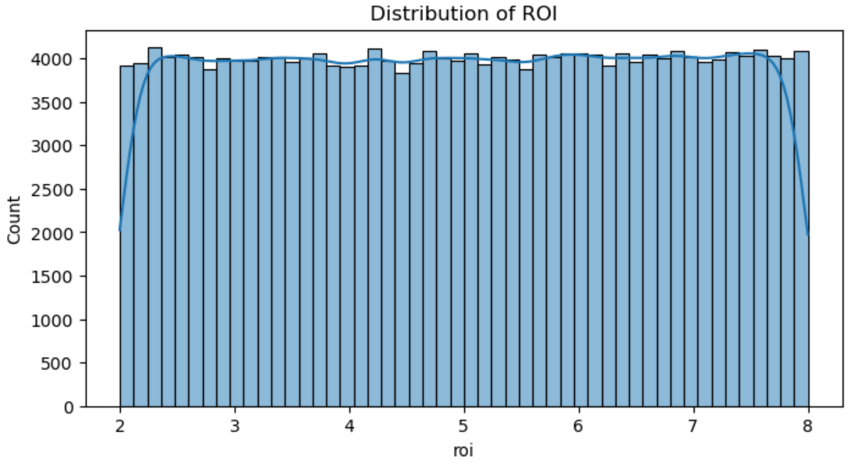
Business Question & Dataset

- Goal: Predict whether a marketing campaign will achieve a high ROI (ROI > 5).
- Synthetic dataset of 200,000 campaigns with 14 fields (company, channel, duration, location, language, metrics etc.).
- Target variable created as high_roi = 1 if ROI > 5, else 0 – roughly balanced classes.
- Focus for business: understand which levers (impressions, clicks, conversion rate, audience, channel) matter most.



ROI Distribution

- • ROI is almost uniformly spread between 2 and 8 – consistent with the synthetic design.
- • No clear group of ‘very low’ or ‘very high’ ROI campaigns – harder to find strong decision boundaries.

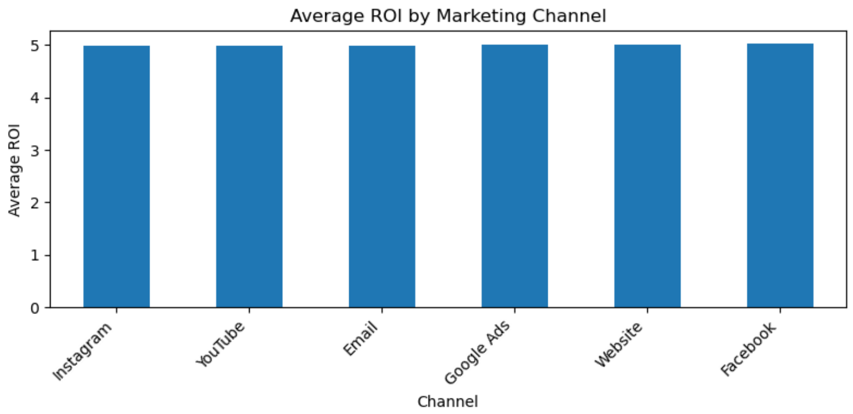


Average ROI by Channel

- • All channels have almost identical average ROI.

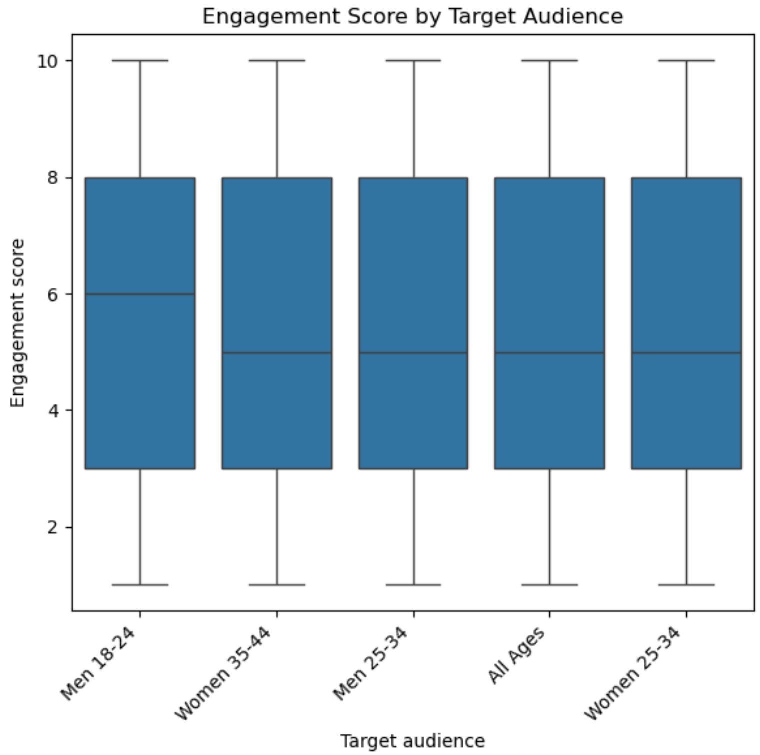
- • Because the data is synthetic, no single channel clearly outperforms the others.

- • In real data we would expect some separation (e.g. email vs social media).



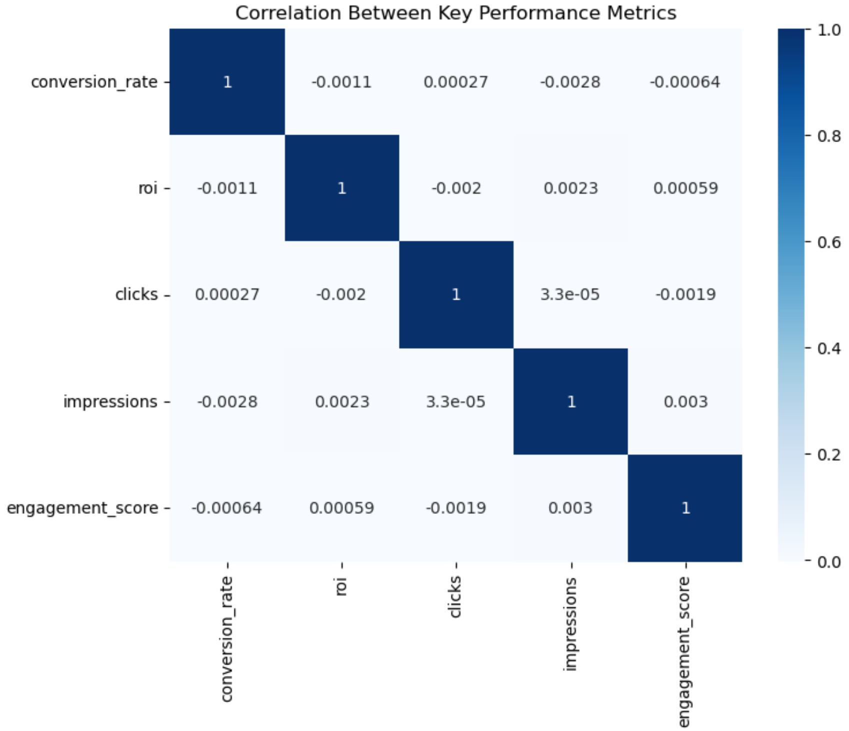
Engagement Score by Target Audience

- • Engagement scores are similarly distributed across target audiences.
- • No audience segment stands out as significantly more or less engaged.
- • Again this reflects the balanced, synthetic nature of the dataset.



Correlation Between Key Metrics

- • Conversion rate, ROI, clicks, impressions and engagement have near-zero pairwise correlations.
- • This means linear relationships between these metrics are very weak.
- • Low correlations partly explain why simple linear models struggle to separate high vs low ROI campaigns.



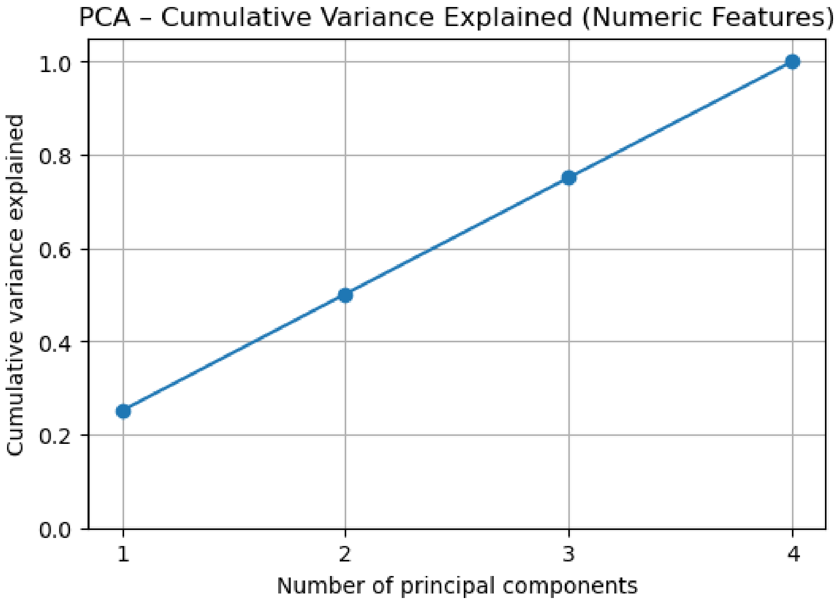
	conversion_rate	roi	clicks	impressions	engagement_score
conversion_rate	1	-0.0011	0.00027	-0.0028	-0.00064
roi	-0.0011	1	-0.002	0.0023	0.00059
clicks	0.00027	-0.002	1	3.3e-05	-0.0019
impressions	-0.0028	0.0023	3.3e-05	1	0.003
engagement_score	-0.00064	0.00059	-0.0019	0.003	1

PCA – Cumulative Variance Explained (Numeric Features)

- • The four numeric metrics can be compressed into 3–4 principal components with minimal information loss.

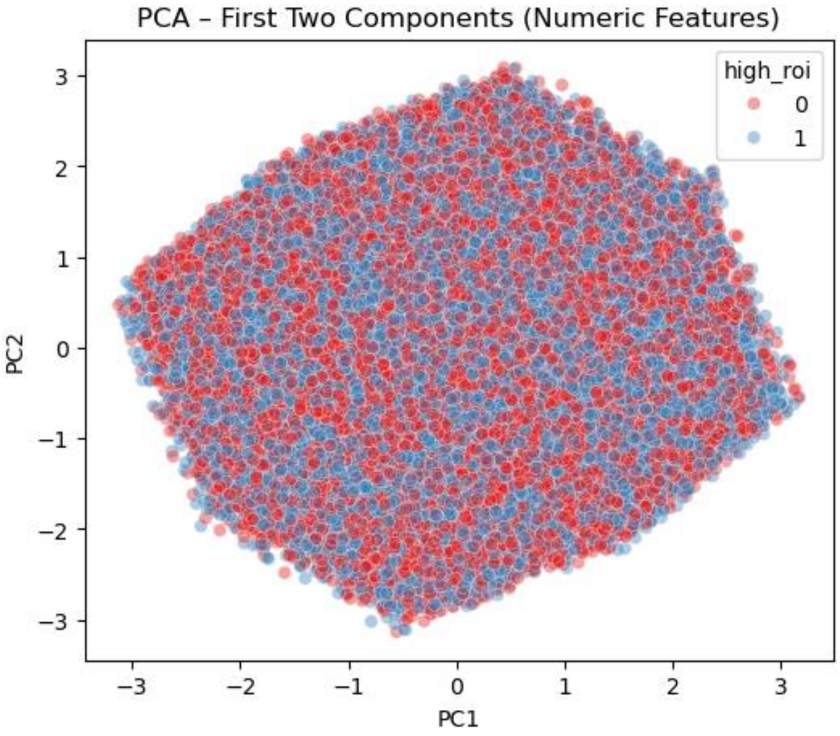
- • PC1–PC2 already capture around half of the variance; all 4 components capture almost 100%.

- • Useful mainly for visualisation rather than improving model performance on this dataset.



PCA – First Two Components (Numeric Features)

- • High-ROI (1) and low-ROI (0) campaigns are heavily mixed in PC1–PC2 space.
- • There is no clear cluster or boundary separating the two classes.
- • This visual pattern matches the near-random model performance we see later.

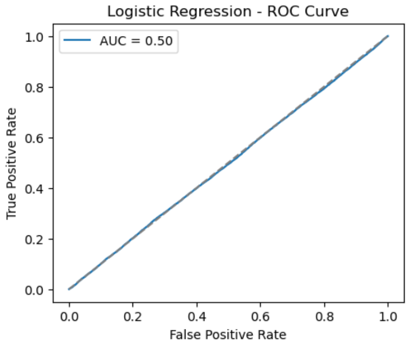


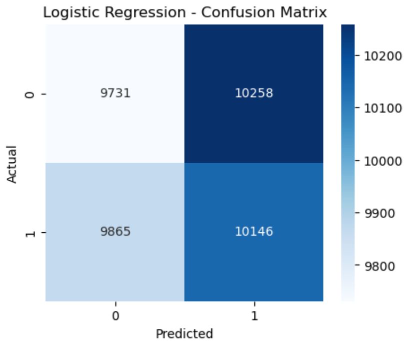
Logistic Regression – Results

- • Confusion matrix shows very similar counts in all four cells – the model is barely better than random.

- • ROC curve lies almost exactly on the diagonal (AUC ≈ 0.50), confirming no real predictive power.

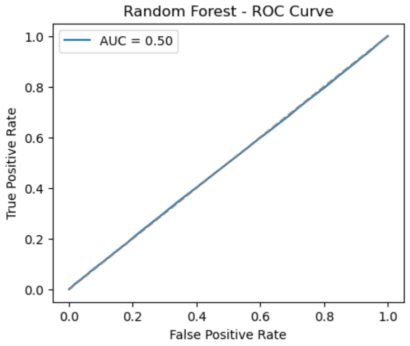
- • On this synthetic dataset, a linear decision boundary cannot meaningfully separate high vs low ROI.

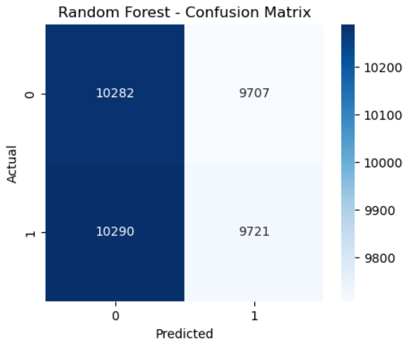




Random Forest – Results

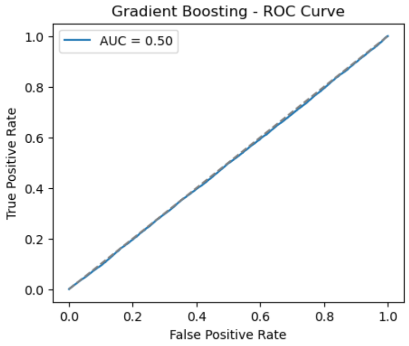
- • Random Forest confusion matrix is also close to balanced across cells, mirroring the label distribution.
- • ROC curve again tracks the diagonal (AUC ≈ 0.50) – the ensemble cannot exploit any strong patterns.
- • Trees mainly learn noise; performance does not justify deployment in a real-world setting.

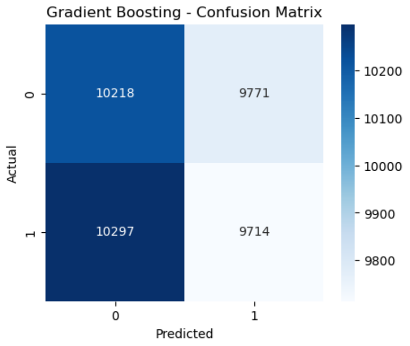




Gradient Boosting – Results

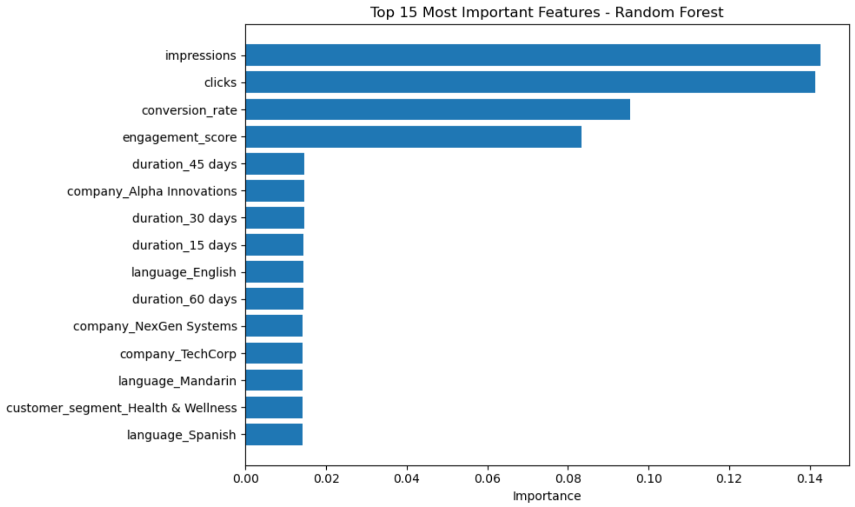
- • Gradient Boosting shows the same story: near-symmetric confusion matrix, no strong bias to either class.
- • ROC curve has AUC ≈ 0.50, indicating chance-level discrimination.
- • More complex boosting does not help because the underlying data lacks a signal to learn.





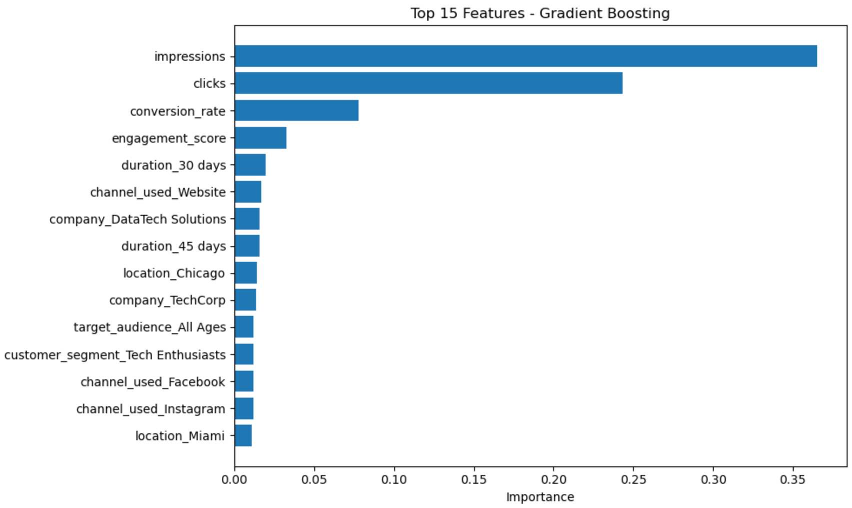
Top 15 Features – Random Forest

- • Impressions, clicks, conversion_rate and engagement_score are ranked as most important.
- • Durations and company/language dummies contribute smaller amounts of importance.
- • However, because overall model performance is random, these ‘importances’ should be interpreted cautiously.



Top 15 Features – Gradient Boosting

- • Gradient Boosting also highlights impressions, clicks, conversion_rate and engagement_score as key drivers.
- • Channel and location dummies appear with modest importance, but their effects are weak.
- • Both models broadly agree on which levers matter most in the synthetic design of the dataset.



Summary

EDA and PCA show that the synthetic dataset was designed to be very balanced, with weak relationships between metrics and ROI.

All three models (Logistic Regression, Random Forest, Gradient Boosting) achieve AUC ≈ 0.50 – effectively random guessing.

Conclusion

Feature importance still surfaces plausible levers (impressions, clicks, conversion_rate, engagement_score), but results are not reliable without real data.

Next steps

Incorporate time-based behaviour, seasonality, and campaign history.

Engineer more meaningful features such as cost per click, cost per impression, campaigns per brand, and customer lifetime value.

Try regularisation, hyperparameter tuning, and more complex ensemble methods.

- Any question????