## I created Three models to predicted OR% , Models are Valid, Reliable, having Good Accuracy, Good Calibration and statistically valid

## ****Integrated Modeling Strategy – Email Open Rate Prediction****

**Models Used:**

1. **Binomial GLM** – interpretable baseline for understanding key feature effects.
2. **Mixed Linear Model (LMM)** – captures both fixed and campaign-level random effects.
3. **LightGBM** – advanced nonlinear model for highest predictive accuracy and ranking power.

### ****How to Use Them Together****

**1. Start with the GLM for interpretability.**

* Use it to **understand drivers** (e.g., tone, personalization, subject length).
* Provides clear coefficients and significance levels — useful for communication with marketing teams.

**2. Apply the Mixed Linear Model for campaign-aware benchmarking.**

* Adjusts for **group or campaign-specific effects**, giving fairer comparisons.
* Best for estimating “expected” OR when campaigns differ structurally (e.g., across brands or audiences).

**3. Use LightGBM for targeting and prediction.**

* **Highest accuracy and lift** (1.4–1.6× in top decile).
* Ideal for **ranking and prioritizing** future email variants or target segments.
* Use SHAP values for interpretable feature importance alongside GLM insights.

### ****Recommended Workflow****

1. **Exploration:** Use GLM and LMM to understand feature impact and control for random factors.
2. **Prediction & Optimization:** Use LightGBM for scoring new email variants and campaign targeting.
3. **Validation:** Cross-check LightGBM feature importance with GLM coefficients to ensure consistency and trust.
4. **Action:** Deploy LightGBM predictions to guide A/B testing and select high-performing subject lines, tones, or personalization strategies.

## Binomial GLM – Email Open Rate

* **Goal:** Predict email open rates (OR) using 13 features
* **Model Type:** Binomial GLM with logit link — appropriate for proportion data like open rate.
* **Data Setup:** Cleaned, winsorized, and one-hot encoded with ~13 categorical and numeric predictors.
* **Accuracy:**
  + Mean Actual OR = **0.1275**
  + Mean Predicted OR = **0.1109**
  + **MAE = 0.0285**, **RMSE = 0.0380** (≈3–4% error range).
* **Calibration:** Predictions align well with actuals across most deciles; mild underestimation at the top end.
* **Top-Decile Lift:**
  + Top 10% predicted OR = **0.1498** vs overall 0.1153 → **1.3× lift** in open rate.
* **Insights:** Model effectively distinguishes higher-performing campaigns; personalization and tone likely key drivers.
* **Use Case:** Estimate expected open rate for new email variants; guide A/B testing and targeting.

## LightGBM Model – Email Open Rate

* **Model Type:** Gradient Boosted Trees (LightGBM) – captures nonlinear feature effects and interactions automatically.
* **Accuracy:**
  + **MAE = 0.0193**, **RMSE = 0.0246**
  + Weighted MAE = **0.0128**, Weighted RMSE = **0.0169**
  + Explains about **43% of variance (R² = 0.43)** on hold-out data.
* **Calibration:** Predicted open rates closely align with actuals across deciles — well-calibrated even in high-performance bins.
* **Top-Decile Lift:**
  + **Unweighted Lift:** 1.42× (0.150 vs 0.106 overall)
  + **Weighted Lift:** 1.59× (0.151 vs 0.095 overall) — strong ability to rank top-performing campaigns.
* **Insights:**
  + Model effectively identifies high-engagement campaigns.
  + Handles nonlinearities and feature interactions better than GLM.
* **Business Impact:** Enables smarter targeting and prioritization of campaigns with the highest open rate potential.

## Mixed Linear Model – Email Open Rate Prediction

**• Goal:** Model and predict email open rates (OR) while accounting for both fixed and random effects across campaigns.  
**• Model Type:** Mixed Linear Model (LMM) — captures variability across campaign groups while modeling open rate as a continuous outcome.  
**• Data Setup:** Same cleaned and feature-engineered dataset as the GLM; includes 13 categorical and numeric predictors.

**• Accuracy:**

* **Train:** RMSE = 0.0262, MAE = 0.0194, MAPE = 16.46%, R² = 0.477 (Adj R² = 0.461)
* **Test:** RMSE = 0.0284, MAE = 0.0214, MAPE = 18.08%, R² = 0.409 (Adj R² = 0.330)  
  → Indicates solid generalization with moderate explanatory power and minimal bias (≈ 0).

**• Calibration:**  
Predicted and actual open rates align well across deciles, with only slight underestimation in the upper range.  
Example deciles: mean predicted OR rises from ~0.075 to 0.160, closely matching actuals (0.078–0.147).

**• Top-Decile Lift:**

* Top 10% predicted OR = 0.1458 vs overall 0.1280 → **1.14× lift** in open rate.

**• Insights:**

* The model captures key campaign-level variations and improves stability over purely fixed-effect models.
* Predictive strength (R² ≈ 0.41 on test) confirms that structural patterns—such as tone, personalization, and length—explain meaningful variance in open rates.

**• Use Case:**  
Estimate expected OR for new campaigns while adjusting for group-level differences (e.g., brand, audience, or category).