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

## I built three simple apps (in notebook itself) to make prediction easy for each model

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

**Models Used:**

* **Binomial GLM:** Interpretable baseline for understanding key feature effects.
* **Mixed Linear Model (LMM):** Captures both fixed and campaign-level random effects.
* **LightGBM (Final Model):** Advanced nonlinear model delivering the highest predictive accuracy and ranking power.

**How to Use Them Together**

1. **Start with the GLM for Interpretability**  
   Use it to understand key drivers such as tone, personalization, and subject length.  
   It provides clear coefficients and significance levels, making it ideal for communicating insights to marketing teams.
2. **Apply the Mixed Linear Model for Campaign-Aware Benchmarking**  
   This model adjusts for group or campaign-specific effects, ensuring fair and balanced comparisons.  
   It’s best suited for estimating “expected” open rates when campaigns differ structurally (for example, across brands, categories, or audiences).
3. **Use the Final LightGBM Model for Targeting and Prediction**  
   The final LightGBM model achieves the highest accuracy and lift (R² ≈ 0.70, top-decile lift up to 2.1×).  
   It is ideal for ranking and prioritizing future email variants or target segments.  
   SHAP values can be used to interpret feature importance alongside GLM insights, combining predictive power with interpretability.

**Recommended Workflow**

* **Exploration:** Use GLM and LMM to understand feature effects and account for random or structural variations.
* **Prediction & Optimization:** Apply the LightGBM model to score new email variants, predict open rates, and identify high-performing campaigns.
* **Validation:** Cross-check LightGBM feature importance with GLM coefficients to ensure consistency and transparency.
* **Action:** Deploy LightGBM predictions to guide A/B testing, optimize subject line tone and personalization, and focus on campaigns with the greatest open rate potential.

## 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 = 0.0190**, **RMSE = 0.0228**
  + Weighted MAE = **0.0175**, Weighted RMSE = **0.0199**
  + Explains about **70% of variance (R² = 0.696)** 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:** 1.65× (0.170 vs 0.103 overall)
  + **Weighted Lift:** 2.09× (0.170 vs 0.081 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( grouped by subject line )