# Model Card – Ad Click Prediction with Bayesian-Tuned LightGBM

## Model Description

**Input:**  
The model receives a structured dataset containing customer demographics and contextual features. The input features include:

age (numerical): Age of the customer.

gender (categorical): Male or Female.

device\_type (categorical): The type of device used.

ad\_position (categorical): Placement position of the ad.

time\_of\_day (categorical): Time when the ad was displayed.

age\_bin (categorical): Age categorized into 5 discrete bins.

device\_hour (categorical): A composite feature combining device type and time of day.

**Output:**  
The model outputs a **probability score** (0 to 1) representing the likelihood of a user clicking on an advertisement. This is further thresholded to generate a binary prediction:

**1: Clicked the ad**

**0: Did not click the ad**

**Model Architecture:**  
The model is built using **LightGBM**, a gradient boosting framework based on decision trees. Key architectural and tuning elements include:

* Decision tree boosting with optimized split finding.
* Categorical feature handling via native LightGBM support.
* Hyperparameter tuning via **Bayesian Optimization** using optuna, which searched over learning rate, tree depth, number of leaves, sampling fractions, and scale\_pos\_weight.

## Performance

The model was evaluated on a held-out test set (20% split from original data). Metrics include:

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **AUC Score** | 0.6777 |
| **Accuracy** | 0.69 |
| **Precision** (Class 1) | 0.70 |
| **Recall** (Class 1) | 0.94 |
| **F1-score** (Class 1) | 0.80 |

**Evaluation Method:**

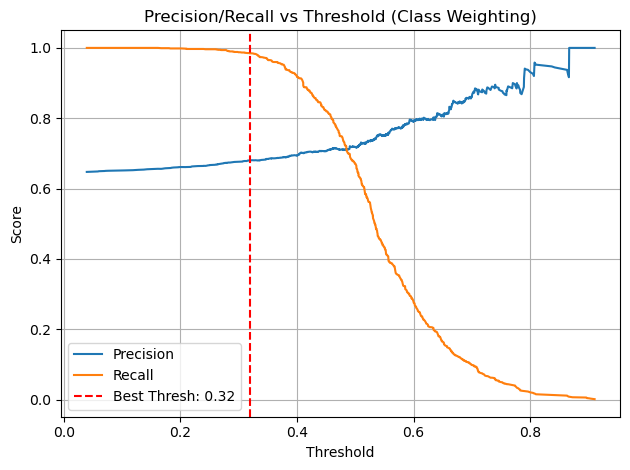
* Binary classification performance was assessed using standard metrics (AUC, precision, recall, F1).
* A decision threshold was optimized using F1 score on the validation set.
* Confusion matrix and precision-recall curve were analyzed to validate balance and utility.

## Limitations

* The model relies on a relatively small and potentially biased dataset (10,000 entries with demographic data).
* Some features, like browsing\_history, were excluded due to high missingness, which might have reduced contextual richness.
* Imputation strategies may introduce bias, especially for users with missing categorical data.
* Performance on unseen or very different distributions (e.g., mobile-first users, new ad positions) is untested.

## Trade-offs

* **Recall vs Precision**: The model prioritizes **high recall** for users who click ads, which is favorable in marketing contexts where missing a click opportunity is costly. However, this comes at the expense of **lower precision**, leading to more false positives.
* **Threshold Tuning**: The default threshold (0.5) was suboptimal. F1-optimized thresholding (around 0.45–0.50) was essential to balance performance.



* **Complexity vs Interpretability**: While LightGBM is more interpretable than neural networks, its complexity compared to logistic regression may be a barrier for non-technical stakeholders without additional SHAP analysis.