# Report of costumer uptake prediction

## Methodology

### Data Balancing

The initial dataset has a significant class imbalance (0:39922 , 1:5289), with far more customers declining the offer than accepting it. To ensure our models were not biased towards the majority class, the **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to the training data. This technique creates a balanced set by generating synthetic samples for the underrepresented 'Yes' class, allowing the models to learn the patterns of both outcomes effectively.

### Models Evaluated

The following classification models were trained on the balanced data and evaluated:

- **LightGBM**: A fast and powerful gradient boosting model that builds a sequence of decision trees, with each new tree correcting the errors of the previous ones.
- **Random Forest**: An ensemble model that combines the predictions of many individual decision trees to make a more accurate and robust final decision.
- **Logistic Regression**: A statistical model that predicts a binary outcome by fitting a simple S-shaped curve to the data.
- **Neural Network (MLP Classifier)**: A model inspired by the human brain, composed of interconnected layers of nodes that can learn complex, non-linear patterns.
- **Linear Discriminant Analysis (LDA)**: A statistical classifier that finds a linear boundary between classes.
- **K-Nearest Neighbors (KNN)**: A simple model that classifies a new customer based on the majority vote of its closest neighbors in the training data.

- **Quadratic Discriminant Analysis (QDA)**: A more flexible version of LDA that uses a curved (quadratic) boundary to separate classes.

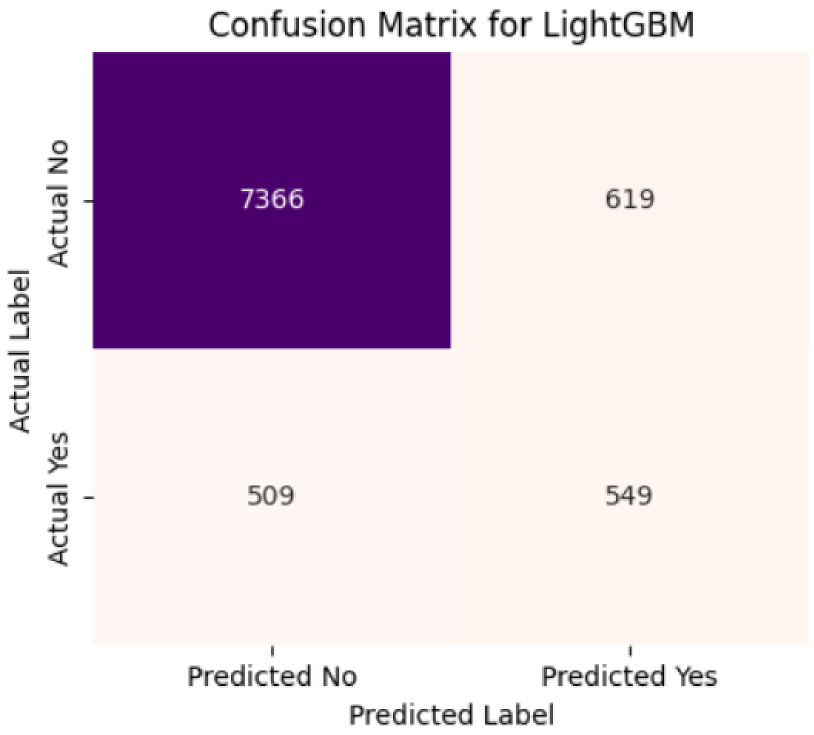
## 1. Conclusion

Based on a comprehensive evaluation, the **LightGBM model is the selected choice** for the marketing campaign. This decision is primarily driven by its superior **F1 Score of 0.4933**, which represents the most effective balance between our key business objectives.

The F1 Score is the most critical metric for this project because it harmonizes two competing goals:

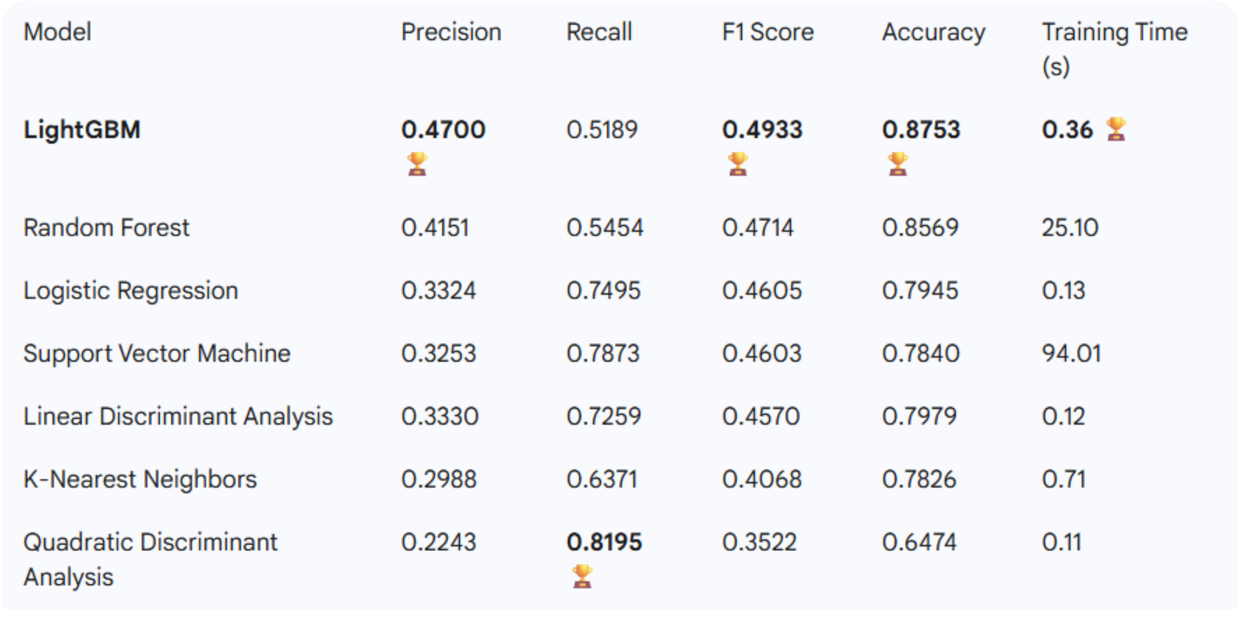
- **Precision (0.4700)**: This measures the efficiency of our campaign. A higher precision means fewer wasted calls to uninterested customers, which directly reduces costs and minimizes potential reputational damage.
- **Recall (0.5189)**: This measures the effectiveness of our campaign in capturing potential sales. A strong recall ensures we minimize opportunity loss by successfully identifying a large portion of customers who would have accepted the offer.

While other models like Quadratic Discriminant Analysis had a higher recall, their extremely low precision would have led to an inefficient and costly campaign. The LightGBM model represents the best compromise, providing a strong ability to find interested customers while maintaining the highest efficiency of all models tested. The primary trade-off is a slightly longer training time than simpler models, but this is a negligible factor as the model only needs to be trained once before the campaign.



## 2. Discussion

The decision to select LightGBM was made after a careful analysis of the trade-offs presented by the other models.



| Model | Precision | Recall | F1 Score | Accuracy | Training Time (s) |
|---|---|---|---|---|---|
| **LightGBM** | **0.4700** 🏆 | 0.5189 | **0.4933** 🏆 | **0.8753** 🏆 | **0.36** 🏆 |
| Random Forest | 0.4151 | 0.5454 | 0.4714 | 0.8569 | 25.10 |
| Logistic Regression | 0.3324 | 0.7495 | 0.4605 | 0.7945 | 0.13 |
| Support Vector Machine | 0.3253 | 0.7873 | 0.4603 | 0.7840 | 94.01 |
| Linear Discriminant Analysis | 0.3330 | 0.7259 | 0.4570 | 0.7979 | 0.12 |
| K-Nearest Neighbors | 0.2988 | 0.6371 | 0.4068 | 0.7826 | 0.71 |
| Quadratic Discriminant Analysis | 0.2243 | **0.8195** 🏆 | 0.3522 | 0.6474 | 0.11 |

- **Random Forest**: This was a very strong contender and the second-best model with an F1 Score of 0.4714. However, the LightGBM model slightly outperformed it in both precision and F1 score, making it the more optimal choice.
- **Logistic Regression & Linear Discriminant Analysis (LDA)**: These models were fast and produced decent F1 scores (~0.46). From a business perspective, their high recall (~74%) was attractive as it meant very low opportunity loss. However, their low precision (~33%) indicated that roughly 2 out of every 3 calls made would be to an uninterested customer. This level of inefficiency was deemed too costly and potentially damaging to the brand's reputation.
- **K-Nearest Neighbors (KNN)**: This model was outperformed by the top models in every key metric and was therefore not a competitive choice for this business problem.
- **Quadratic Discriminant Analysis (QDA)**: This model was not selected despite having the **highest recall (0.8195)**. A business interpretation of this result is that while the model is excellent at finding almost every potential customer, it does so by being extremely aggressive. Its very low precision (0.2243) means that nearly 80% of the calls made would be wasted. Such a high volume of unwanted calls would be financially inefficient and would almost certainly lead to significant reputational damage, making it a poor choice for a sustainable business strategy.
- **Neural Network (MLP Classifier)**: This model achieved a final F1 Score of 0.4509, which was lower than our top-performing models like LightGBM and Random Forest. Its performance did not justify its significantly longer training time (over 2 minutes), making it a less practical and efficient choice for this specific problem.

## 3. Next Steps

To further improve upon the selected LightGBM model, the following recommendations are proposed:

1. **Advanced Feature Engineering**:
   - **Create Interaction Features**: We can create more sophisticated features that capture the relationships between variables. For example, an age_balance_ratio or balance_per_day feature might provide more predictive power than the individual features alone.
   - **Binning Numerical Features**: Grouping numerical features like age into bins (e.g., '18-30', '31-45', etc.) can sometimes help the model capture non-linear relationships more effectively.
2. **Hyperparameter Tuning**:
   - We should perform an exhaustive hyperparameter search (using RandomizedSearchCV or GridSearchCV) specifically for the LightGBM model. While we tuned the Random Forest, a dedicated tuning process for LightGBM on the new feature engineered dataser could unlock further performance gains and push the F1 score higher.
3. **Explore Additional Data**:
   - If possible, acquiring more data, especially for customers who accepted the offer (the minority class), would be highly beneficial. A larger and more balanced dataset would allow the model to learn the characteristics of interested customers with greater confidence, leading to a more robust and accurate model.