5 Neural Network Model

5.1 Experimental Setup

5.1.1 Model Architecture

This study constructed a feed-forward neural network (NN) which uses the Keras framework. The model architecture is as follows:

Input → Dense(64, ReLU) → Dropout(0.4) → Dense(32, ReLU) → Dropout(0.4) → Dense(1, Sigmoid)

- Activation functions: ReLU for hidden layers, and Sigmoid for the output layer;
- Loss function: Custom Focal Loss with parameters $\alpha = 0.8$ and $\gamma = 2$, designed to strengthen the model's learning on hard-to-classify samples.
- Optimizer: Adam (learning rate = 0.001).
- Regularization: Dropout rate at 0.4 for overfitting alleviation.
- Evaluation metrics: AUC, Precision, Recall, and MCC applied to evaluate the model performance. MCC served as the core metric, provides a fair evaluation index for severely class-imbalanced dataset.

During model training, an Early Stopping mechanism was applied with Matthews correlation coefficient(MCC) as the monitoring metric. Training will be terminated early if MCC does not improve for 10 consecutive epochs. This strategy effectively prevents the model from overfitting.

5.1.2 Experimental Design

Three Neural network models were conducted, and they form a progressive validation design, which corresponds to three layers of verification logic:

Table 2: Experimental Design and Research Questions

Experiment	**Research Objective**	**Key Question**
Baseline (No SMOTE)	Establish baseline to assess model performance on the original imbalanced data	Can the model identify minority-class samples without any adjustment on the distribution?
Tuned Model (SMOTE + Focal Loss)	Improve model's sensitivity to the minority-class through data rebalancing & Loss function	Does the combination of SMOTE and Focal Loss effectively mitigate class imbalance & enhance model generalization?
5-Fold Cross-Validation	Test model consistency, robustness, and stability across random splits	Are the improvements stable? Are the improvements not due to random variation?

After each training, model predictions on test dataset are evaluated over threshold intervals (0.01–0.99, step = 0.05). The thresholds yield for the highest MCC are selected as the classification cut-off.

The final outputs include the test data's MCC, AUC, and confusion matrix.

5.2 Experimental Results

5.2.1 Training and Validation Curve Analysis

Observations from the learning curves are chiefly as follows:

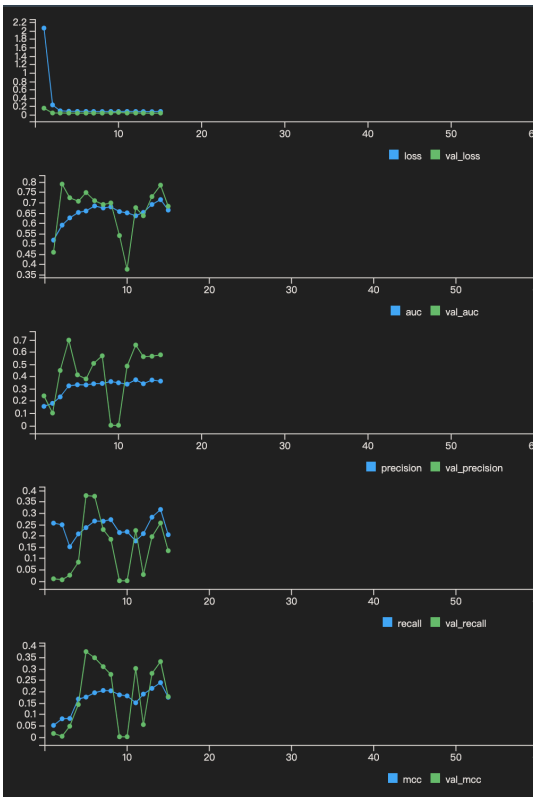

Figure 8: Learning Curve - Baseline Model

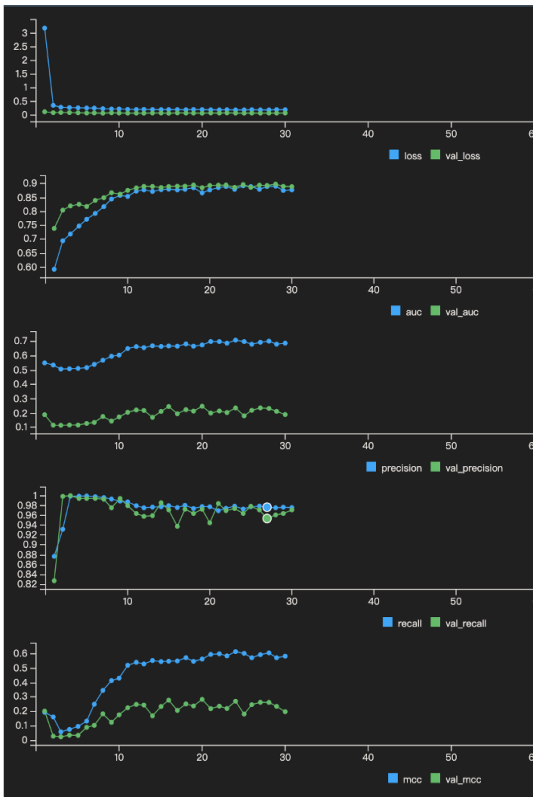

Figure 9: Learning Curve - SMOTE Model

- Baseline model: MCC, AUC, Precision, and Recall fluctuate significantly on both the train and validation dataset in later epochs, indicating poor generalization and possible overfitting of the baseline model.
- SMOTE model: Training and validation losses converged concurrently, which indicates that the SMOTE oversampling strategy and the application of Focal Loss function effectively mitigate overfitting and strengthen model stability concurrently.
- The learning rate (0.001) led to steady convergence after 30 epochs for both models.

These results highlight that, under unbalanced data circumstances, the binary classification model may struggle to capture minority-class signals. Both SMOTE oversampling and Focal Loss emphasize heavily and continuously on the minor examples. As a result, validation performance stability is maintained in later stages in baseline models.

5.2.2 Five-Fold Cross-Validation Results

The results of five-fold cross-validation are shown below:

Table 3: Model Performance across 5-Fold Cross-Validation

Fold	**MCC**	**AUC**
1	0.479	0.902
2	0.512	0.909
3	0.517	0.903
4	0.505	0.905
5	0.497	0.901
Mean	**0.502**	**0.904**

The fluctuations remained in a vert small range, while MCC within 0.04 and AUC within 0.01. This indicates that the model's structure and optimization strategy are both stable and generalized. The consistent results across different data partitions also statistically confirm the robustness and reproducibility of the model.

5.2.3 Overall Model Performance

Table 4: Comparison of Model Performance with and without SMOTE

Model	**SMOTE Used**	**Test MCC**	**Test AUC**
Baseline	No	0.319	0.748
SMOTE	Yes	0.473	0.887
5-Fold CV (Average)	Yes	0.502	0.904

Comparing 3 models, the Baseline model with original datasets shows a strong bias toward the majority class. This model exibits a high variance on the validation dataset and very limited generalization capability.

After applying SMOTE oversampling, model MCC improved approximately by 47% and AUC also increased by 0.14. This suggests that the data balancing substantially enhances the model discriminative capacity and improve the overall performance stability.

5-Fold CV is further applied and results showcase model robustness, with minimal variance among metrics, proving strong reproducibility for different datasets.

5.3 Error Type(False Negative/Positive) Analysis

To gain both modeling and business insight, error samples in the test set were analyzed.

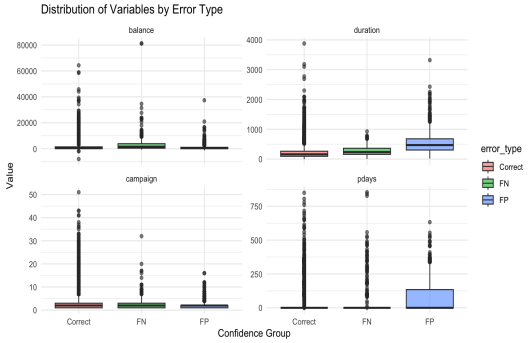

Figure 10: Boxplot for Error Types

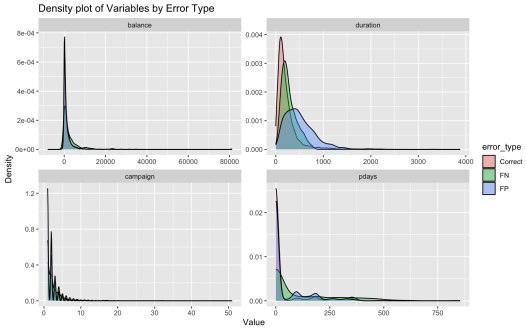

Figure 11: Density Plot for Error Types

Two main error types were identified:

5.3.1 False Negatives

Boxplot analysis shows that FN samples are densely located among customers with a lower balance and shorter duration. Density plot further confirms the finding. FN curves are concentrated near zero. This demonstrates model insensitivity towards the low-balance or low-duration customers.

It is indicated that the model is more likely to interpret low balance and short duration as implicit signals of weak purchasing power or low interest, leading to a misclassification as "non-subscribers."

However, this group may also include potential high-value clients, such as young or newly onboarded customers with low current balance but high future potential. Therefore, from a business perspective, these "underestimated active customers" require differentiated marketing strategies to improve overall conversion rates. For instance,

- reinforcing digital-channel engagement for short-duration customers, and
- offering flexible financial products or preferential rates for low-balanced clients.

5.3.2 False Positive

The duration median of FP samples is significantly higher than that of true subscribers. For the density plot, FP curves move rightward toward higher values, reinforcing the interpretation of model bias toward long-duration and high-balance customers and suggesting that the model over-relies on duration as a predictor of customers' subscription intent.

In model training, long duration was learned as a strong positive signal, but in real-world practice, this may not always stay true. Long calls may indicate customers comparing or evaluating the financial products which they are not necessarily intending to buy. Or in the other

hand, the conversations might extend by the sales' scripts rather than customer real interests. Thus, these long-duration-called but non-subscribing clients became false positives.

Apart from duration, pdays (days since last contact) also varied widely among the FP group. Outliers possibly caused the model to overestimate the willingness of long-uncontacted customers to subscribe. While some of these clients do respond positively upon re-engagement, majority in real cases represent low-response groups.

This indicates the model's feature dependency bias. The model over-weights frequent features like "duration" and underweights critical variables such as previous and campaign. Both duration and pdays are process-based indicators reflecting outreach interaction behavior rather than intrinsic customer intent.

The model's misinterpretation of these features as intent proxies inflates FP rates. To address this, feature-layer enhancement and semantic enrichment are needed to transform behavioral indicators into interest-behavior proxies, in a bid to reduce these Type I errors.

From a business standpoint, future improvement should include customer engagement modeling at the feature level, or aligning predictive modeling with marketing practice and achieve joint optimization of analytics and business strategy.