3.8 XGBoost

3.8.1 Model Theory

XGBoost is an optimization algorithm based on Gradient Boosting Decision Trees. It improves prediction accuracy by using more precise calculation of gradient and regularization. Compared with traditional methods, XGBoost is better at handling nonlinear and high-dimensional sparse characteristics. so based on the bank marketing data samples' unbalanced features, XGBoost is highly suitable in terms of performance and stability, obviously, it is very compatible for our project**.**

3.8.2 Model Building

Specifically, in order to make sure the model can effectively process the raw data, and well as keep consistency of different models, we carried out systematic preprocessing on the data, mainly including categorical variable encoding, data spliting, and SMOTE oversampling.

3.8.3 Model Results

Firstly, we used all features to train a basic XGBoost Model. And the result is as table 3.10.1 follows.

Table 3.8.1：Performance results of the XGBoost basic Model (Model 1)

Model	MCC	F1-score	Precision	Recall	Accuracy
XGBoost (Full features)	0.5151	0.5961	0.5076	0.662	0.8537

The MCC of the basic model is 0.5151, while it is quite decent, so in the later modification we tried to understand the prediction rational, to improve fitness. Overall, Model 1 verified the initial effectiveness of XGBoost in bank marketing forecasting, providing a performance baseline for subsequent optimization.

3.8.4 Model Optimization

In this part, we use SHAP（SHapley Additive exPlanations）framework to help explain this model from global view down to individual customer..

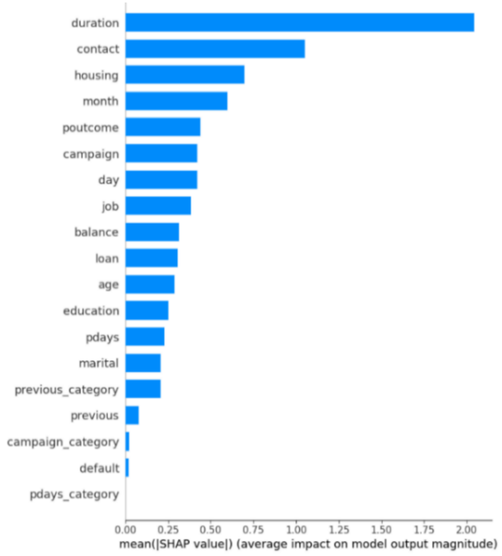

Figure 3.8.1 Global SHAP feature importance ranking chart

Figure 1 shows the distribution of the average SHAP values of the features. The x-axis represents the average contribution intensity of the features when predicting results. The features further to the right have a greater impact on the model's prediction. We can see that duration is the most influential positive feature and plays a decisive role in predicting whether customers will subscribe to marketing products. Secondly, would be housing (whether there is a mortgage), poutcome (the last marketing result), and month (the contact month). This result reveals the main decision variables, providing the guildline for selecting features.

Based on the global analysis, we depicted a SHAP bee Plot to find the idea of specific influence direction and intensity of different features. The result is quite similar to Figure 1.

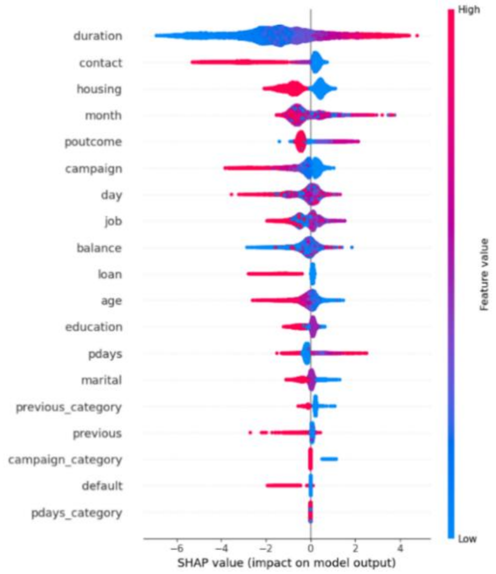

Figure 3.8.2 SHAP bee diagram

Then to further understanding the prediction process on a single decision sample, we plotted the SHAP Force Plot. In Figure 3, the red parts represent the features that drive "subscription" predictions, while the blue parts inverse. The overall predicted value is composed of the sum of the base predicted value and the contributions of each feature.

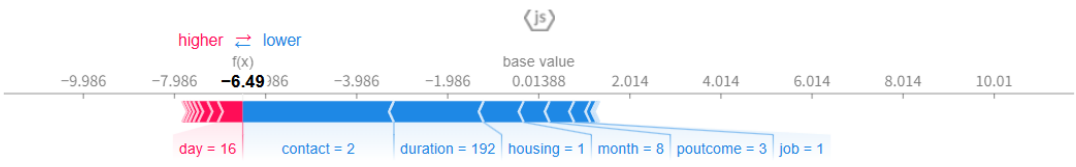

Figure 3.8.3 SHAP Force Plot

So, finally this customer will be predicted to not subscribe with f(x) approximately equals -6.49. As there are too many negative influencers, including short call duration, the mortgage and so on. The contact date, slightly increase the predicted value, but the overall impact is relatively small.

Based on the SHAP analysis results, these results helped us select better feature in Model 2, including duration, housing, poutcome, month, contact, previous, campaign, age and education. Some features were excluded due to limited impact or the possibility of introducing noise, such as marital, its contribution is quite small.

3.8.5 Model Comparison

In this project, we constructed and evaluated three models: XGBoost (Full features), XGBoost (Select Features), and the ensemble model (XGBoost + Random Forest). The model performance are as follows.

Table 3.8.2 XGBoost Results

Model	MCC	F1-score	Precision	Recall	Accuracy
XGBoost (Full features)	0.5151	0.5961	0.5076	0.662	0.8537
XGBoost （Select Features）	0.5096	0.5908	0.4776	0.7001	0.8432
XGB+RF	0.532	0.6045	0.5242	0.6728	0.8575

The three models reflect different modeling approaches. The baseline model utilizes all the data, but it is receptible to noise. Feature selection models focus on key variables. Although this enhances the ability to identify potential customers, it also leads to more misjudgments. In contrast, the ensemble model achieves a better balance between identifying potential customers and controlling marketing costs.

3.8.6 Model Conclusion

Based on the feature visualization and model comparison results, we suggested that bank firstly they could extend the call duration for high-value customers to enhance the conversion rate. Also, priority can be given to contacting customer groups that have no mortgage, use mobile phones.

4. Comparison and Conclusion

4.1 Evaluation Indicators

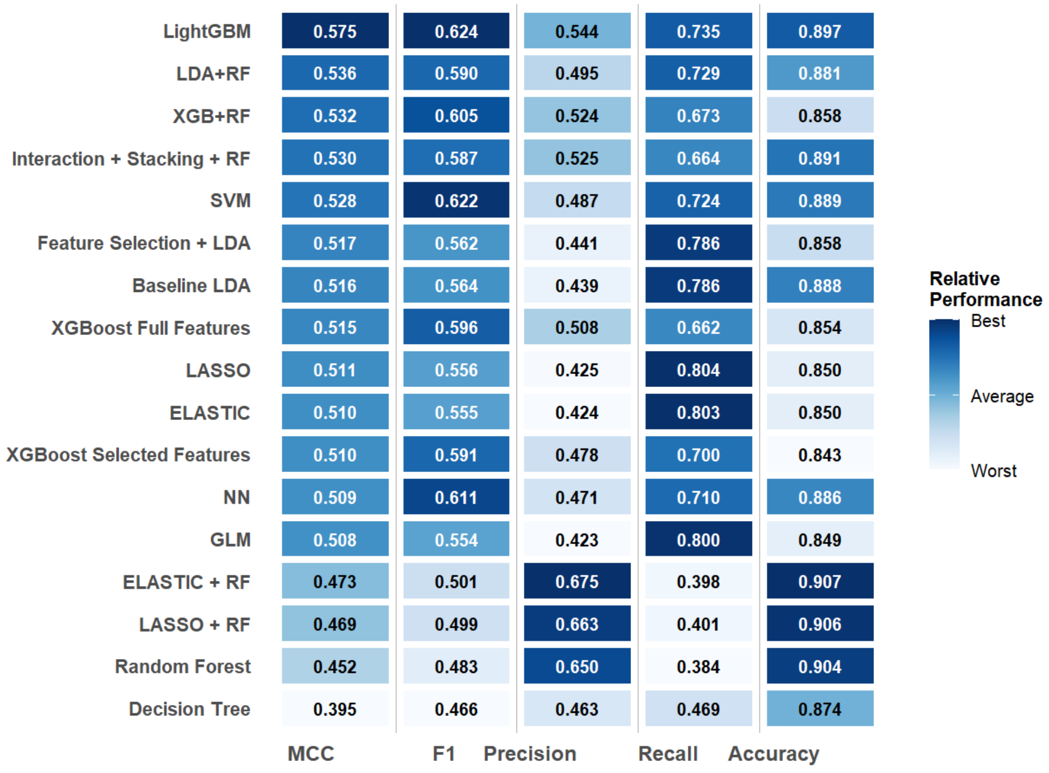

Figure 4.1.1 Model performance heatmap

To evaluate models' performance, we selected three metrics: Precision, Recall, MCC. Precision reflects the proportion of customers who are predicted to subscribe and actually subscribe, which means reducing invalid calls and saving marketing costs. Recall represents the proportion of actual subscribers that have been correctly identified, reflecting the model's ability to cover potential customers. And MCC is useful to comprehensively reflect the overall predictive ability and stability of the model.

4.2 Comparison of Model Performance

This project has established a model evolution path from simple to complex. All models are subject to 5-fold cross-validation to ensure fairness. The following table summarizes the performance of all model.

Table 4.1.1 Summary Table of Model Results

Model	MCC	F1	Precision	Recall	Accuracy
NN	0.5092	0.6107	0.4709	0.7101	0.8858
GLM	0.5080	0.5536	0.4235	0.8000	0.8492
SVM	0.5281	0.6216	0.4869	0.7240	0.8893
LASSO	0.5111	0.5560	0.4249	0.8045	0.8497
ELASTIC	0.5098	0.5553	0.4241	0.8030	0.8495
LDA+RF	0.5365	0.5903	0.4949	0.7290	0.8812
XGB+RF	0.5320	0.6045	0.5242	0.6728	0.8575
Decision Tree	0.3948	0.4661	0.4633	0.4688	0.8743
LightGBM	0.5752	0.6240	0.5439	0.7346	0.8967
LASSO + RF	0.4685	0.4991	0.6631	0.4006	0.9061
ELASTIC + RF	0.4734	0.5007	0.6753	0.3981	0.9073
LDA Full Features	0.5162	0.5645	0.4394	0.7864	0.8876
Random Forest	0.4523	0.4832	0.6503	0.3844	0.9038
XGBoost Full Features	0.5151	0.5961	0.5076	0.6620	0.8537
LDA Selected Features	0.5173	0.5623	0.4407	0.7864	0.8582
Interaction + Stacking + RF	0.5296	0.5869	0.5252	0.6641	0.8906
XGBoost Selected Features	0.5096	0.5908	0.4776	0.7001	0.8432

It is very obvious that whenever what kind of model we choose, the MCC values are not very high, concentrated around 0.5. This is because our bank marketing samples are very unbalance, only around 10% customer have been subscribed, which further show that our models have caught the signal in the real-world data.

4.3 Model Selection Strategy Based on Business Goals

4.3.1 Trade off

The core challenge here lies in the trade-off between precision and recall rate. Precision focuses on: "Among the customers we contact, how many are genuine subscribers?" This reflects cost efficiency. Recall focuses on: "Among all the genuine subscribers, how many have we successfully identified?" This reflects the coverage rate. And actually, these two important rates cannot be maximized simultaneously, and our XGBoost model (especially after feature selection) clearly demonstrates this dilemma.

Therefore, under this consideration, we categorize our model into three parts, which correspond to the real-world requirements.

4.3.2 Customer Acquisition

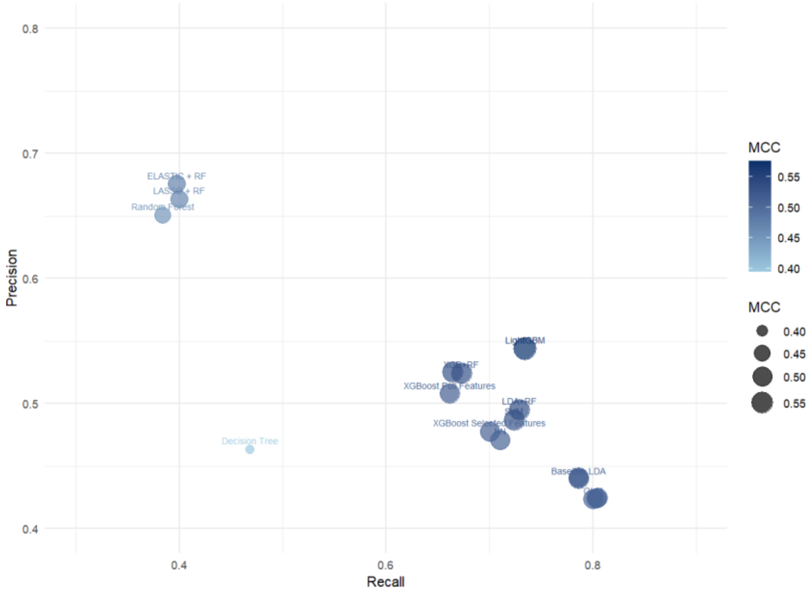

Figure 4.1.2 Precision-Recall scatter point

If the bank wants to prioritize customer acquisition, we recommend using a high-recall model such as LASSO. In this project, this model can cover about 80% of subscription customers which shows its ability of effectively capturing potential opportunities. However, the trade-off is lower precision, which will result in a large number of ineffective calls. Therefore, this strategy is suitable when the bank aims to expand its market.

4.3.3 Cost Control

If the bank wants to control its marketing costs, we recommend using a high-precision model such as ELASTIC + RF. This type of model focuses on each call, aiming to use minimal resources to acquire more actual subscriptions. It precisely identifies the customers with the highest potential to subscribe. However, this approach will result in a lower recall rate, as many potential customers will be missed. Therefore, this strategy is suitable when the bank is in a budget-constrained period.

4.3.4 Comprehensive Optimal

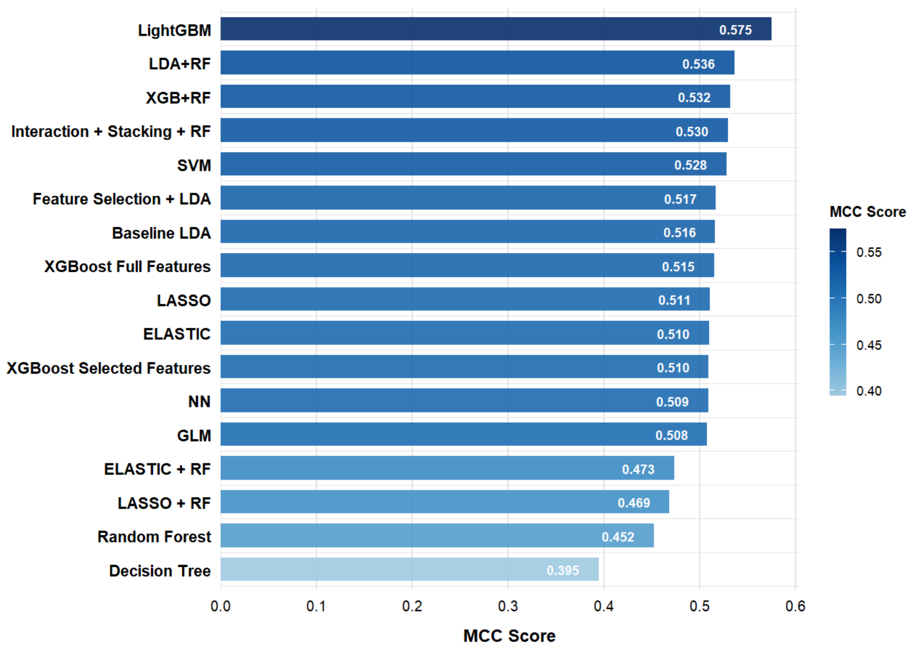

Figure 4.1.3 Model performance Ranking by MCC

The last model is for comprehensive optimal orientation, a balanced model (such as LightGBM in our project). This type of model performs best in MCC, can simultaneously balance considerable customer coverage and acceptable cost efficiency, providing a reliable guarantee for achieving stable performance.

Overall, the bank can choose the model that best suits its current development stage or business period.

4.3.5 Real-life cases

Take Trust Bank as an example. This bank offers new users a $60 no-threshold registration coupon without any requirements from Sep to Oct this year suddenly. In the early stage (it opened in 2022), its marketing strategy might focus on cost control, but after several years of money accumulation (from 2022 to 2025), now it pays more attention to the growth of the customer base. This phased strategy adjustment is similar to the modeling idea in this project: in the early stage, it leans towards cost-sensitive strategies, while in the customer expansion stage, it emphasizes a high recall rate to cover as many potential customers as possible.

4.4 Future Work

We also have some suggestions for further improve prediction accuracy. Firstly, feature engineering. Introduce time series features (such as customer historical marketing interactions) and external economic indicators (such as interest rates) to capture more dynamic behavioral patterns. In addition, we can do some model innovation in the future, like combing highly interpretable GLM and highly predictive LightGBM, to simultaneously enhance the business interpretability and predictive performance of the model.