Business Report for Marketing Analytics

SID:

530699718

530554059

530324061

530684954

530697758

Table of content:

1. Introduction

A machine learning tool is always a reliable assistant for a human being. In this project, we will demonstrate how data analysis plays a relevant role in the business field, especially in campaign strategy planning.

Our task is to build classification models for our clients' market campaigns. Our clients for this task are one bank and one fashion store. We need to analyze the customers' data from our clients' former campaign records and help them build the machine learning model that can select the potential customers. Based on the model, we will also offer them constructive insight into how they can further improve their market campaign, like the most important features a customer interested in the campaign would have.

We have used the linear models in this project for the classification. The linear model is easy to interpret and has a high computational speed. The model also allows easy modification based on the original format. Furthermore, we used the tree-based models in our project, which also has an excellent interpretation. Moreover, it can automatically handle the nominal and ordinal variables without encoding; the model will not be affected by missing values or values that are too large or small. On the other hand, we also used the boosting methods to keep returning updated new tree models targeted to ease the residuals from the last models until we get the optimal models for classification. Finally, we use the methodology of model stacking, which combines a linear model with two non-linear models together to leverage the residuals.

After comparing these models, the random forest model outperforms the others in the bank dataset, making it qualified to be the best model. For the store's prediction, we selected the model stacking with the logistic regression l2 and the bagged tree model as our best model according to the AUC.

2. Problem Formulation and Objectives

From the decision theory perspective, the agent, in this case, would be a bank or store; they face a similar situation of selecting the target customer for a market campaign. The agency needs to decide whether to engage with the customer or not. Customers, on the other hand, can choose whether to engage or ignore the campaign. The agency risks completely wasting the operational cost if it targets inactive customers.

Certainly, our task is a prediction problem since the states of nature depend on customers' preferences based on their personal situations, like their incoming or age, and on the current market circumstance. The agency only introduces its service to the customers; it will unlikely shape the customers' final decision.

Here are the summaries of our primary objective for our project.

Model formation: building classification models that will accurately predict whether the customers will engage in the market campaign by giving the customers' basic information.

Decision-supporting: By analyzing past campaign data, our data analysis provides clients with thoughtful insight, such as the key features of activity customers, which our clients can refer to for their future marketing strategy.

Risk reductions: By responsible forecasting potential customers, avoid the risk of high expenses towards low-impact market effort, which is unfavorable for companies' resource allocation.

We aim to build the most efficient marketing campaign that accurately forecasts the customer's behavior. In this case, the machine learning tool will use thorough mathematical models to study the customers' data from the past marketing campaign and build the prediction model. Hence, the well-formed model can be used to classify the engaged and inactive customers to optimize the company's profit for future marketing.

3. Data understanding

3.1 Bank Dataset

3.1.1 Brief Description

Firstly, our research focuses on the phone campaign of the bank. In order to predict whether the campaign will be successful, before we start to build the machine learning model, it is important to do exploratory data analysis to select the proper factors.

The table below gives a brief description of the features of the bank dataset. In total, there are 20 columns in the bank datasets, including continuous, discrete, categorical, and binary types. From the table, we can tell there are variable clients of different ages in different situations as well as different months who have been phone-called several times in the phone campaign.

index	age	duration	campaign	pdays	previous
count	41188	41188	41188	41188	41188
mean	40.02	258.29	2.57	962.48	0.17
std	10.42	259.28	2.77	186.91	0.49
min	17	0	1	0	0
25%	32	102	1	999	0
50%	38	180	2	999	0
75%	47	319	3	999	0
max	98	4918	56	999	7
index	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188	41188	41188	41188	41188
mean	0.08	93.58	-40.5	3.62	5167.04
std	1.57	0.58	4.63	1.73	72.25
min	-3.4	92.2	-50.8	0.63	4963.6
25%	-1.8	93.08	-42.7	1.34	5099.1
50%	1.1	93.75	-41.8	4.86	5191

75%	1.4	93.99	-36.4	4.96	5228.1
max	1.4	94.77	-26.9	5.04	5228.1

3.1.2 ‘pdays’ and ‘duration’

After the brief description, our EDA research proceeds in more detail. If we check the ‘y’ column, that is, the result of this phone campaign. In total, 36548 of 41188 clients refused to subscribe to a term deposit. If we focus on the ”duration”, that is, the duration of the call, is highly correlated to the campaign so the result could be practically useless.

Secondly, the number of days that passed by after the client was last contacted from a previous campaign(‘pdays’) is worth noticing. If we divide the data into ‘No previous contact’ and ‘have previous contact’, as shown in the table, in which 0 means ‘No’ and 1 means ‘Yes’, we can easily calculate that there are only about 1.5% of clients subscribe to a term deposit without any previous contact, but 20.8% in the opposite situation.

y	0	1
pdays		
0	36000	3673
1	548	967

3.1.3 Crosstab Plot Analysis

If we focus on the crosstab plot, the situation is clearer. For ‘pdays’ that are shown above, it can be easily observed that in all clients that subscribe to a term deposit, the proportion of those who were previously contacted exceeded those who were not. The crosstab plot can also be helpful in describing the distribution of other variables, which the full table can be found in the appendix. For example, among different jobs, students and the elderly who are retired prefer the term deposit sales in the phone campaigns better, which has a noticeably higher proportion than other jobs. So, for the reality of the business, we can infer that these two parts of the clients may have lower demand for money nowadays but need to save money for the future and prefer to make some profit from the interest.

It may be surprising that the proportion of people who have a loan is similar to those who do not, we may have a common sense that people who need to pay back loans need more liquidity and may prefer demand deposit that they can withdraw money at any time, in this campaign, however, the situation could be different, where values in these columns stay stable when ‘y’ changes, which means they may not provide any extra information to our study. Still, we will leave these two variables to machine learning to check whether these factors will significantly influence the model.

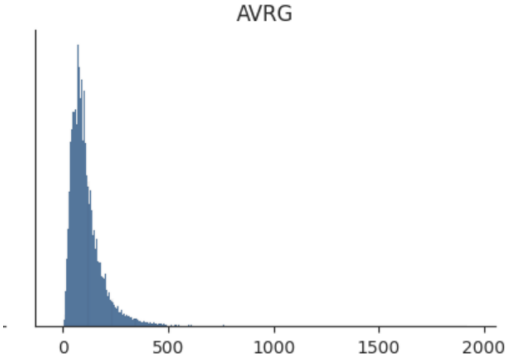
Finally, in this dataset, we need to check the mutual information between predictors and ‘y’, which describes how much information is communicated, on average, in one random variable about ‘y’. According to the table in the appendix, the mutual information of each variable is rather low in our research, this means that none of the variables can play a decisive role in our models, so the prediction could be rather difficult as we may need more complicated models. Back to the business reality, mutual information tells us that whether a client will choose to subscribe to a term deposit can be influenced by many factors, so it is not easy to succeed.

3.2 Fashion Store dataset

In this part, we will focus on the store dataset. The aim of our research is to find out whether the customer responded to the promotion, or to predict the relationship between ’RESP’ and other predictors. Most of the data types in this dataset are numerical, except for three categories of binary data.

	count	mean	std	min	25%	50%	75%	max
AMSPEND	21740	14.218	149.864	0	0	0	0	10642.72
PSSPEND	21740	147.822	395.139	0	0	0	127.938	11476.8
CCSPEND	21740	286.858	441.461	0.01	78	147.6	321.943	22511.49
AXSPEND	21740	24.257	113.856	0	0	0	0	4099.92
STORES	21740	2.340	1.603	1	1	2	3	19

The first thing that is worth noticing is that in different stores, there are significant differences in the amount spent. Obviously, in PS and CC stores, the spending amount is much higher than the other two stores. Also, in CC store, all the 21740 customers at least spend their money to purchase something. And in the last row of the table shown above we can tell that most people would like to store in less than 3 stores. This can help the discovery of customers' preferences in the market campaign.



Then the distribution of the average of each customer spent per visit will be discussed. The distribution of the average is rather positively skewed, which should

be noticed in our model building. Also, it reminds us that there are more than 50% of people spent less than 113 dollars, which is the mean of the data, so we may more focus on this part of people in our market campaigns.

Then we should focus on one of the binary data. It is very common to shop via the Internet nowadays, among customers, there are more people who use websites to purchase products. So we can infer that it will be more useful to make market campaigns through websites, such as place digital advertisements on different markets.

4. Feature engineering

4.1 Bank dataset

4.1.1 Detect useless features

As can be seen from the results of mutual information in EDA, there are some features that seem useless, such as ‘loan’, ‘housing’ and ‘day_of_week’ which have very low correlation with 0.000017, 0.000052 and 0.000344 respectively. In addition, as can be seen from the former EDA plots, values in these columns stay stable when ‘y’ changes, which means they do not provide any extra information for us to study. Also, EDA shows the feature ‘age’ and ‘campaign’ both have a very weak mutual information, which may also be considered as useless variables in later analysis.

4.1.2 Delete features

There is a special column ‘duration’ that needs to be deleted. According to the dictionary for bank datasets, ‘duration’ indicates the last contact duration with the client, and it has a high effect on the output target. Also, the dictionary implies that if ‘duration’ = 0, then y=0, and after the end of the call y would be obviously known. This will cause data leakage if we use this feature in model training. Therefore, we deleted this feature. Additionally, we supposed that ‘pdays’ can lead to hidden data leakage because it represents number of days that passed by after the client was last contacted from a previous campaign.’pdays’=0 means client was not previously contacted. This can have similar effect with ‘duration’=0, so we finally decided to delete ’pdays’.

4.1.3 Deal with response variable

The response variable is textual data which cannot be used directly in the machine learning models. Since the response only has two outcomes: ‘yes’ and ‘no’, we simply use number 1 and 0 to denote ‘yes’ and ‘no’ respectively. In this way, we successfully transform this categorical data into numerical data which can be used in further analysis.

4.1.4 Deal with nominal variables

The dataset has several categorical features, such as ‘month’, ‘jobs’, ‘education’ and so on. As those categorical variables are all nominal that only indicate classification and do not involve sorting, we can directly use method ‘get_dummies’ to transform them into numerical types.

4.1.5 Feature scaling and transformation

Additionally, we need to do the data transformation and scaling. From the EDA plots, we find that the data in ‘age’, ‘campaign’, ‘emp.varr.rate’ and ‘cons.conf.idx’ is far from the ideal normal distribution. In this case, we select ‘yeo-johnson’ which is a flexible and effective method to deal with this problem. Since the data tend to be more normally distributed, it is more likely to have better performance when fitting models in the machine learning process.

Besides, we apply the ‘StandardScaler’ method to scale those continuous features like ‘emp.var.rate’, ‘nr.employed’, ‘cons.price.idx’ and so on to be in the same range. In this way, it can ensure that training and test data are uniformly standardized. Then, in the subsequent training or prediction of the machine learning process, the model can work on features with the same scale, avoiding the bias caused by some features because the value is too large or too small.

4.2 Fashion Store datasets

4.2.1 Delete features

Firstly, the columns 'ZIP_CODE' and 'CLUSTYPE' need to be deleted. While 'ZIP_CODE' can be a rough reflection of geographic location and regional characteristics, it can be too broad for predicting individual customer behavior. 'CLUSTYPE' is Micro vision lifestyle cluster type according to the data dictionary, but the data under this column are all integers. We do not know what the numbers represents for, so we delete this feature. Besides, we also consider the feature 'PERCRET', due to the return rate may reflect more customer satisfaction with the product or consumer behavior after purchase, it does not directly affect the customer's willingness to participate in future promotional activities. Customers with high return rates may be due to quality problems or purchase errors, and do not necessarily mean that they are not interested in the promotion. So, it will be deleted too.

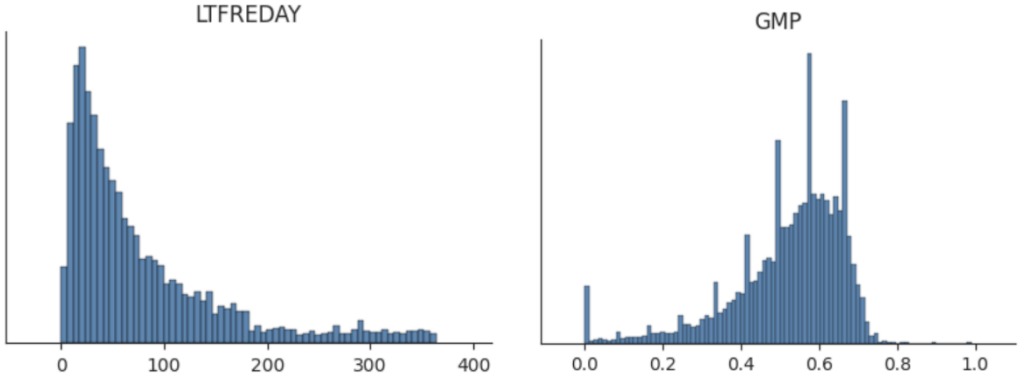
4.2.2 Missing value and data format

Applying the same method as we do in bank datasets to check the missing values and formats. There are no missing values in store datasets and no need for format adjustment. However, we notice there is a unique texture feature 'VALPHON', which cannot be directly used in the machine learning models. In this case, we transform its two outcomes: ‘Y’ and ‘N’ into 1 and 0 respectively.

4.2.3 Feature scaling and transformation

Then, as the EDA plots indicate, many features like ‘AVRG’, ‘GMP’, ‘CLASSES’, ‘LTFREDAY’ and so on are right or left skewness. So, we use ‘StandardScaler’ in the scikit-learn library to standardize the numerical characteristics of the datasets, which is suitable for normalizing numerical features to a distribution with a mean of 0 and a

standard deviation of 1. In this way, the performance of the model can be improved.



5. Methodology

5.1 Bank Dataset

In this dataset, we have used linear models, tree-based models, and a model stack. In this part, we will focus on three models: the best linear model, the best tree-based model, and the stacked model.

5.1.1 Logistic Regression with L1 Regularization

First, Logistic regression can provide a clear explanation of the relationship between model parameters and results and is suitable for assessing which features have a significant impact on the predicted results. Maybe it Although the performance of the logistic regression model may not be optimal, it is suitable as a benchmark model. If a complex nonlinear model does not perform as well as logistic regression on this data set, then it has no practical value for a certain dataset.

To achieve better results, we need to configure some reasonable parameters for it. First, we apply L1 regularization, which is also called LASSO. It penalizes the absolute values of the model coefficients, which helps in feature selection by driving the weights of some features to zero and can finally avoid overfitting. In addition, we use cross-validation (k = 5) to select the optimal regularization parameter, ensuring that the model's regularization strength is best suited for the current dataset. This process can also reduce the risk of overfitting by evaluating the model performance across different training-test splits in one dataset. Finally, we also used 'neg_log_loss' as the scoring criterion. It is the negative log loss, which is a commonly used loss function in logistic regression.

However, this model also has some disadvantages. Although we have configured many parameters for it to obtain better performance, it is still a linear model, which means it is poor at dealing with complex datasets. Moreover, the response value in the bank dataset is imbalanced. ‘Yes’ only accounts for about 10%. In this case, Logistic Regression has a bias towards classes which have a number of instances. They tend to only predict the majority class data. The features of the minority class are treated as noise and are often ignored (Upasana, 2023). It tends to predict the results toward All Negative.

Another disadvantage is produced by L1 regularisation. In real world environments, we often have features that are highly correlated. For example, the

year our home was built and the number of rooms in the home may have a high correlation (Pykes, 2023). But in the process of training, if there are highly correlated features, L1 regularization usually only selects one of the features and ignores other highly correlated features, which is something that we might not want.

5.1.2 Random Forest

Random forest model is oriented to the tree model. Compared with the linear model, the tree model is more structured with different levels. The tree will split the variables into several subsets on every level, such as splitting the numerical data according to the data range and splitting the categorical variables based on the values. Compared with the linear model, the tree model is less affected by the outliers and can automatically handle the nominal and ordinal data.

The random forest model will return the result from a bundle of tree models. For every tree model in the random forest, the model will randomly select the feature for every split, which avoids choosing the dominant features each time. Overall, the random forest model will return a more accurate prediction than the tree model. For more details of our model, we initially set the model's criterion = 'entropy', max_features =8, n_estimators=1000, min_sample leaf = 5, random states =10.

After finishing the hyperparameter optimization based on minimizing the overall lose, we finalized the parameters by using the Gini index as our criterion, which means the pureness of every leaf is calculated by the formula $GiniIndex=1-\sum^{n}_{1}(p_i)^2$, and the model's every splitting will aim to have the highest pureness for the leaf. We will conduct min_samples leaf= 15 and max_feature as twelves, which means that every leaf must contain more than 15 samples, and for every split, the model will randomly select 12 features. Furthermore, the n_estimators =1000 means the model's outcomes will come from the demographic based on the result of 1000 trees.

Although the random forest model is more accurate than the tree model, it only demonstrates the importance of variance and is not interpreted in detail as the tree model. Random Forest models require more memory on the machine (Fratello, Tagiaferri, 2019).

5.1.3 Model Stacking

Multiple individual models have been utilized for analysis, each with its own strengths and weaknesses. Therefore, after analyzing with single models, a stacking approach has been implemented to potentially enhance prediction accuracy. In this stacking process, various models, referred to as base models, are employed and "stacked" together. The outputs from these base models are used to make the final prediction. These outputs serve as inputs to create the final estimate, allowing the stacking model to benefit from the advantages of all the base models used.

Figure XX The Process of Model Stacking

As the above picture indicates (Tpoint Tech, n.d.), for each base model, the training set is used to fit the model through cross-validation. Once fitted, the model can predict outcomes for both the validation set and the test set. At this point, all prediction results from the validation set are used as new inputs for the meta model, which serves as the Level 2 model in the stacking approach. The Level 2 model is then trained using these new inputs. By incorporating the earlier predictions from the test set, the final prediction is generated. Various methods can be employed to construct the meta model, such as linear regression or ridge regression.

In this case, we create a stacked classification model using ‘StackingClassifier’ from the ‘sklearn.ensemble’ module. Considering the timings and computational costs, we just select three models: Logistic Regression with L2 Regularization (can help prevent overfitting of the model and is suitable for high-dimensional data), XGBoost (can effectively handle various types of data structures, including sparse data, and has the ability to handle missing data) and CatBoost (very effective for processing data sets with a large number of class features without complex preprocessing). With the number of cross validations set to be 5, the error rate of the fitting model reaches the lowest compared to other models. Lastly, we use a simple model which is Logistic Regression as the Meta-model to synthesize the output of each base model to form the final prediction result.

Certainly, model stacking often enhances accuracy over a single model approach. However, selecting base models, determining the number of cross-validations, and choosing a higher-level model remain challenging due to the relatively high computational costs involved in stacking. Moreover, as new features are derived from other base models, there is a high likelihood that the stacking model may encounter issues with multicollinearity.

5.2 Fashion Store Dataset

Like the bank dataset, we will focus on three models: the best linear model, the best tree-based model, and the stacked model.

5.2.1 Logistic Regression with L2 Regularization

L2 regularization is quite similar to the L1 regularization illustrated above. In this model, we also used 'neg_log_loss' as the scoring criterion and used cross-validation to select the optimal regularization parameter. The difference is that L2 regularization (Ridge) tends to shrink the coefficients evenly instead of driving some of the

unimportant ones to zero by penalizing the sum of the squares of the model coefficients. Additionally, we set Cs=50 to ask the model to automatically test and find the best ‘C’ value from 50 different possibilities. 'C' value represents the inverse of the regularization strength. A smaller 'C' represents a larger penalty, which helps to achieve a better balance between model complexity and performance, preventing overfitting.

Since L2 regularization is a logistic regression model, it is also not good at handling imbalanced data sets, like mentioned above. Besides, compared to L1, the drawbacks of L2 regularization are not obvious. One disadvantage is that it shrinks all coefficients uniformly, which can reduce the interpretability of the model because this movement makes it difficult to identify which features are most important. It might underestimate the true effects of predictors that have strong influences and cause potential bias.

5.2.2 CatBoost

Gradient Boosted Decision Trees (GBDT) is an important tool to enable classification and regression tasks in big data, and CatBoost are an efficient method in GBDT. It is also very suitable for machine-learning scenarios that classify heterogeneous data (Hancock & Khoshgoftaar, 2020). Therefore, in this study, CatBoost fits well with the Fashion Store's purpose of understanding whether different customers respond to promotional activities.

In CatBoost, we set the value of iterations to 2000, which is the number of trees. The more iterations there are, the better the model is, but this may lead to overfitting. So, to solve this problem, we set learning_rate to a lower value (0.01). This allows the model to learn more slowly and accurately. In addition, we set the maximum depth of each tree to 6, which controls overfitting because deeper trees can learn more complete details of the data.

Hancock & Khoshgoftaar (2020) demonstrated that CatBoost is an extremely competitive model in the case of highly unbalanced data classification. In addition, CatBoost has a fairly high interpretability of the data (Jabeur et al., 2021). This can be found from the variable importance analysis chart, CatBoost has the highest importance scores for "LTFREDAY" and "DAYS", which also means that they are the most influential variables in the model prediction. It is an optimization algorithm for dealing with categorical variables. Automatic application of regularization ensures that model predictions are not affected by overfitting and fast learning can be achieved on the CPU and GPU (Jung Min Ahn et al., 2023). Based on these advantages, CatBoost provides accurate and effective predictions based on project data.

However, CatBoost also has disadvantages. CatBoost is affected by the size of the data and the complexity of the model, which may lead to an increase in its

training time (J. F. Hancock & Khoshgoftaar, 2020). Especially when using default hyperparameters or complex data sets, this can lead to longer training times, which can potentially lead to a loss of efficiency (Greeks for Geeks, 2021).

5.2.3 Model Stacking

The rationale of model stacking has been illustrated above in the bank part. Next, we will only introduce the choice of base models.

In the fashion store part, we use the logistic model with L2 regularization and Bagged Trees as base models. L2 regularization can prevent overfitting and enhance the model's generalization ability on unknown data, while Bagged Trees can capture the nonlinear characteristics of the data. About the meta-model, we still choose the Logistic Regression to synthesize the output of each base model to form the final prediction result.

6. Results

6.1 Bank

In the report, to show that different types of errors will lead to different business costs, we created a Loss Matrix and added different weights to the costs incurred by each type of error. When the model predicts that the customer will not subscribe, but the customer does, the loss weight, in this case, is 2; when the model predicts customer subscriptions, but the customer does not subscribe, the loss weight for this situation is 1. The main reason for this weighting is that the loss of a bank if it misses a potential customer is greater than the loss if the marketing fails to meet the expected target.

Actual/Predicted	Subscribe	Not Subscribe
Subscribe	0	2
Not Subscribe	1	0

In addition, according to the Loss Matrix, we calculate that the decision threshold tau is 0.33, this means if a customer makes a prediction about whether he will deposit money in the bank, if the prediction probability is greater than 33.3%, the client will be classified into the category of "will subscribe to a term deposit" and recorded as 1. If the predicted probability is less than 33.3%, the customer is placed in the "will not subscribe to a term deposit" category and recorded as 0.

In addition, in this report, we first evaluate the performance of 14 models. The model is evaluated based on its ability to predict whether clients will subscribe for term deposits. We used a total of 11 indicators, including accuracy, precision, recall rate, etc., to evaluate the models.

We will choose the optimal model with AUC as the main indicator and Loss, Precision, F1-score, and Cross-entropy as auxiliary indicators. The reason for choosing AUC as the main indicator is that AUC is a more accurate measurement

standard than Accuracy when there is an imbalance in the data set (Yang & Ying, 2022), which can be observed in the process of data understanding.

Validation results for the model are shown in the table below.

Model	Loss	Precision	F1-score	AUC	Cross-entropy
Logistic	0.176	0.514	0.451	0.788	0.276
LDA	0.178	0.468	0.468	0.786	0.324
QDA	0.179	0.461	0.469	0.782	0.401
KNN	0.172	0.517	0.469	0.777	0.517
Logistic_l1	0.178	0.512	0.443	0.790	0.276
Logistic_l2	0.178	0.510	0.443	0.789	0.276
Classification Tree	0.177	0.447	0.494	0.779	0.277
Bagged trees	0.179	0.462	0.469	0.783	0.321
Random Forest	0.169	0.506	0.491	0.808	0.268
GBM	0.171	0.528	0.472	0.793	0.285
XGBoost	0.171	0.515	0.476	0.795	0.283
LightGBM	0.172	0.504	0.476	0.806	0.267
CatBoost	0.169	0.527	0.482	0.803	0.273
Stack	0.171	0.582	0.450	0.802	0.274

By comparing the value of these indexes, we can observe that Random Forest has the best performance in all models because its AUC is the highest (0.808), which indicates that it is very strong in distinguishing different categories. Its cross-entropy is low (0.268), which indicates that the Random Forest performs well when separated between classes (Das & Chaudhuri, 2019). It also indicates that the probability distribution predicted by the model is close to the probability distribution of the real data, and the confidence of the predicted results is high. The accuracy (0.506) and F1-score (0.491) of the Random Forest are good in all models, which guarantees the effectiveness of the Random Forest model in practical applications (Chicco & Jurman, 2020). In addition, the observation results also show that the Loss value of Random Forest is the lowest (0.169). In the actual business or decision support system, the Random Forest model with low loss value is more likely to provide accurate and reliable prediction results, which increases confidence in decision-making.

In Model valuation, Logistic_l1, which has the best performance among the relatively simple linear models, was selected as the benchmark model, and it was entered into a test set together with Random Forest for total model valuation. The following table can be obtained by comparing the two results.

Model	Loss	Precision	F1-score	AUC	Cross-entropy
Logistic_l1	0.184	0.495	0.424	0.787	0.284

Model	Loss	Precision	F1-score	AUC	Cross-entropy
Random Forest	0.181	0.478	0.457	0.797	0.277

Through comparison, it can be found that Random Forest still performs well in most indicators. Although the value of AUC in Random Forest is 0.797, compared with the AUC in Model Validation, it is still at a higher value, and the main reason for the decline is that the data in the test set is never used by the model. Moreover, the data distribution of the test set is different from that of the validation set, which may also lead to the decline of AUC. Compared with Logistic_l1, Random Forest performs better on most of the indicators, which also proves that the model validation is correct and Random Forest can meet the project requirements.

6.2 Fashion Store

For the fashion store, we also established a loss matrix to assign different weights to the losses caused by each kind of error. When the model predicted that the client would not respond to the promotion, but did, the weight of the loss was 2. Otherwise, the weight is 1. The decision threshold can be calculated as 33.3%. If the probability of prediction is greater than 33.3%, the client will be classified as 1 in the "will respond" category; If the probability of the prediction is less than 33.3%, the client is placed in the "will not respond" category, denoted as 0.

Actual/Predicted	RESP	Not RESP
RESP	0	2
Not RESP	1	0

In Model Validation, 11 indicators are also used to evaluate the performance of 14 models, and the following data is obtained:

Model	Loss	Precision	F1-score	AUC	Cross-entropy
Logistic	0.234	0.487	0.508	0.843	0.324
LDA	0.253	0.490	0.441	0.815	0.350
QDA	0.261	0.420	0.471	0.784	1.107
KNN	0.275	0.448	0.367	0.766	0.459
Logistic l1	0.235	0.487	0.507	0.843	0.324
Logistic l2	0.233	0.488	0.510	0.843	0.324
Classification Tree	0.246	0.558	0.434	0.818	0.335

Bagged trees	0.237	0.473	0.508	0.846	0.322
Random Forest	0.233	0.503	0.505	0.846	0.320
GBM	0.243	0.510	0.468	0.846	0.338
XGBoost	0.250	0.475	0.460	0.837	0.350
LightGBM	0.231	0.503	0.509	0.846	0.322
CatBoost	0.232	0.510	0.504	0.849	0.319
Stack	0.233	0.517	0.496	0.850	0.325

Through comparative analysis of the data, it can be seen that the AUC value of the Stack model is the highest, 0.850, which also indicates that the Stack has the strongest ability to distinguish between positive and negative categories and indicates that the model has a high sensitivity for prediction. In addition, although the Stack model is not the highest in Precision and F1-score, its high AUC performance is as expected, and it is often able to integrate the advantages of the individual models effectively to improve performance.

In the Model valuation, we chose Logistic_l2 as the benchmark model. Comparing Logistic_l2 with Stack in the test set, we can get the following table:

Model	Loss	Precision	F1-score	AUC	Cross-entropy
Logistic_l2	0.243	0.540	0.525	0.844	0.337
Stack	0.239	0.576	0.523	0.847	0.341

Compared to Logistic_l2, we can find that the AUC value is still higher, and it shows better performance on most key indicators. Although the cross-entropy is slightly higher than Logistic_l2, it also shows that the Stack model is more balanced in performance and can better meet the requirements of the actual situation for various performance indicators. Therefore, the comparison results in model validation are correct, and the Stack model should be selected.

7.Business Recommendation

Based on the Exploratory Data Analysis and model results of each dataset, we extract several insights and summary four kinds of customers who are most responsive to marketing campaigns.

Highly Responsive Customer: There is some significant evidence that demonstrates the people who actively responded to previous market campaigns have much more possibility to respond to future market campaigns in both datasets.

For the bank dataset, the most obvious variable to show the trend is “poutcome”, which means the outcome of the previous marketing campaign. From the crosstab plots, we can observe that over 60% of clients who positively responded to previous market campaigns subscribe to a term deposit this time. In the store dataset, we can also tell the same trend from regression plots of the variables “RESPONDED” and “RESPONSERATE” which are the number and rate of promotions responded to in the past year. The customers who responded to the previous promotions have a higher possibility to purchase in the campaign in this dataset. We suggest that the marketing team should engage with them more frequently to maximize sales opportunities.

Long-time/Loyal Customer: In the EDA of the store dataset, the variable 'DAYS'—(number of days the customer's data has been recorded) displays a positive correlation with customer responsiveness. Additionally, it is identified as a significant variable in the feature importance plots from various boosting models. On the other hand, 'LTFREDAY' (lifetime average days between visits) exhibits a strong negative correlation with customer responsiveness, and it consistently ranks as the most important feature in the feature importance plots of most models utilized in our analysis. The two variables measure how long the individuals have been a customer of a company and how loyal they are. To give practical suggestions, we need to integrate two features from bank dataset, 'previous' and 'campaign', which measure the number of contacts made with a customer before and during the current campaign respectively. They all show highly positive correlation with customer responsiveness in the regression plots. Based on these, we can reach a conclusion that companies should focus on those long-time/loyal customers and keep high contacted with them.

High-Value Customer: In the store dataset, it can be found from the regression plots that there is a significant positive correlation between "MON" (total net sales on a customer) and customer responsiveness. In other words, customers with higher consumption amounts are more likely to conduct corresponding marketing activities. In terms of the number of products purchased, consumers 'CLASSES' and 'STYLES'(the number of types and items bought by the customer in one time) also show a certain positive correlation, which also indicates that the more products customers purchase, the greater their response to them. In terms of customer consumption amounts, no matter how long the consumption amount was in the past, there is a significant positive correlation with responsiveness, which indicates that customers with higher consumption amount will respond more actively to marketing activities. In the feature importance analysis, it can also be found that these variables prove to be the more critical factors in predicting customer response. By focusing on these metrics, Fashion stores can more effectively identify high-value customers and

develop more targeted marketing strategies for them, to improve customer marketing response and customer retention. Additionally, companies should ensure that the quality and frequency of communication are carefully managed to avoid customer fatigue or disruption.

Older Customers: We found that although preliminary exploratory data analysis did not show a clear association between age and subscription to term deposit, advanced nonlinear models, such as Random Forest, Gradient Boosting and LightGBM reveal an interesting phenomenon: older customers are more likely to subscribe to a term deposit. The model's feature importance score reflects the strength of each variable's influence on the predicted outcome, and age was identified as an important feature in this analysis. Senior customers typically possess greater accumulated wealth and have better control over their income. In this case, it may be helpful for companies to design specific marketing campaigns for older customer groups because they are more sensitive to such campaigns, which may include special offers on financial products or services for retirees.

Finally, there are features that measure economic conditions. They can also provide some insights, not about certain types of customers, but rather about specific periods that are suitable for carrying out marketing campaigns. There are two features ‘emp.var.rate’ and ‘nr.employed’ in the bank dataset that indicates employment variation rate and number employed respectively. They all show highly negative correlation with whether customers will subscribe to a term deposit in regression plots and high feature importance in most of the tree-based models. Low values of the two features represent a high unemployment rate and a poor economic condition. In such scenarios, individuals prefer stable investments, such as term deposits, gold, and insurance products. Companies offering these products can intensify the efforts of their marketing campaigns during these periods to attract more customers.

Overall, our analysis highlights the potential of machine learning in interpreting complex market dynamics, especially in details that traditional statistical methods fail to reveal. Through the proper application of these advanced machine learning methods, enterprises will not only be able to better understand customer behavior, but also be able to more effectively adjust their marketing campaigns to respond to the changing business environment.

References

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-019-6413-7

Das, R., & Chaudhuri, S. (2019, September 15). *On the Separability of Classes with the Cross-Entropy Loss Function*. ArXiv.org. https://doi.org/10.48550/arXiv.1909.06930

Fratello, M., & Tagliaferri, R. (2019). Desicion Trees and Random Forests. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1.

Greeks for Geeks. (2021, January 20). *CatBoost in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/catboost-ml/#limitations-of-catboost

Hancock, J. F., & Khoshgoftaar, T. M. (2020). Performance of CatBoost and XGBoost in Medicare Fraud Detection. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 572–579. https://doi.org/10.1109/icmla51294.2020.00095

Hancock, J., & Khoshgoftaar, T. M. (2020). Medicare Fraud Detection using CatBoost. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. https://doi.org/10.1109/iri49571.2020.00022

Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Jung Min Ahn, Kim, J., & Kim, K. (2023). Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. *Toxins*, *15*(10), 608–608. https://doi.org/10.3390/toxins15100608

Pykes, K. (2023, August 4). Fighting overfitting with L1 or L2 regularization: Which one is better? https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization

Tpoint Tech. (n.d.). The basic architecture of stacking [Online Image]. In javatpoint. Retrieved May 20, 2024, from https://www.javatpoint.com/stacking-in-machine-learning

Upasana. (2023, September 11). Imbalanced data : How to handle imbalanced classification problems. https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

Yang, T., & Ying, Y. (2022). AUC Maximization in the Era of Big Data and AI: A Survey. *ACM Computing Surveys*, *55*(8). https://doi.org/10.1145/3554729