A Comparative Study of Data Mining: A Review Ensemble Algorithms

1. Proposed model

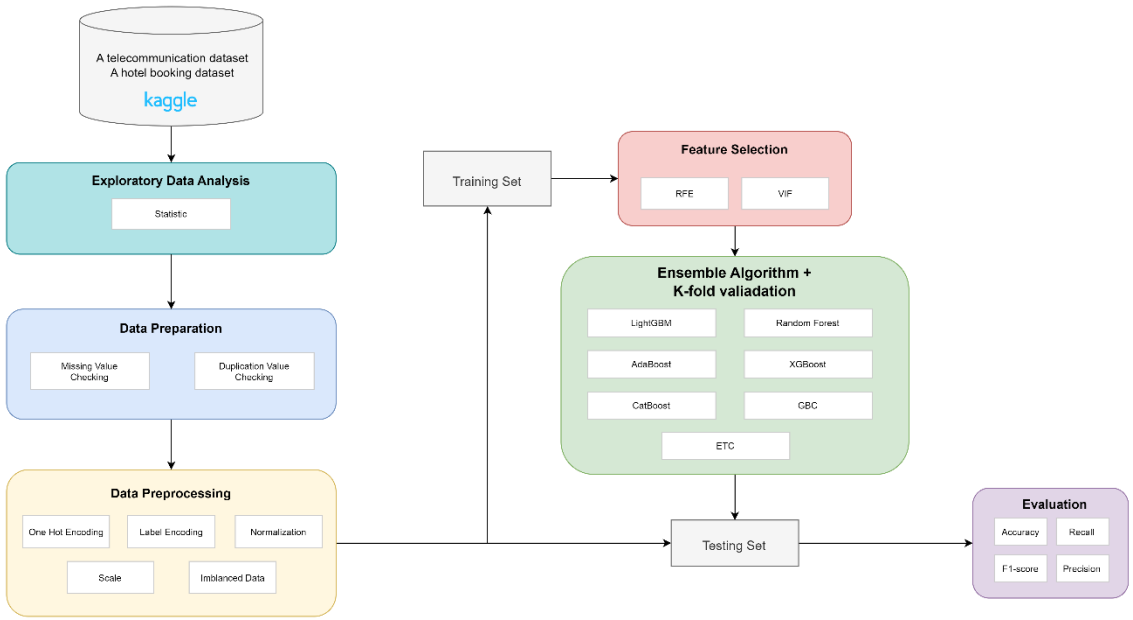

Figure 1 Proposed methodology framework

2. Experimental Result and Analysis

2.1. Dataset

Table 1. The detail information of the hotel booking dataset

No	Column	Description	Data Type
0	hotel	Type of hotel	object
1	is_canceled	Booking Cancellation Status	int64
2	lead_time	Lead time before booking (days)	int64
3	arrival_date_year	Year of arrival date	int64
4	arrival_date_month	Month of arrival date	object
5	arrival_date_week_number	Week number of arrival date	int64
6	arrival_date_day_	Day of month of arrival date	int64

	of_month		
7	stays_in_weekend_nights	Number of weekend nights stayed	int64
8	stays_in_week_nights	Number of week nights stayed	int64
9	adults	Number of adults	int64
10	children	Number of children	float64
11	babies	Number of babies	int64
12	meal	Type of meal	object
13	country	Country of origin	object
14	market_segment	Market segment	object
15	distribution_channel	Booking distribution channel	object
16	is_repeated_guest	Whether guest is a repeated guest	int64
17	previous_cancellations	Number of previous cancellations	int64
18	previous_bookings_not_canceled	Number of previous bookings not canceled	int64
19	reserved_room_type	Reserved room type	object
20	assigned_room_type	Assigned room type	object
21	booking_changes	Number of changes to booking	int64
22	deposit_type	Type of deposit	object
23	agent	ID of booking agent	float64
24	company	ID of company	float64
25	days_in_waiting_list	Number of days in waiting list	int64
26	customer_type	Type of customer	object
27	adr	Average daily rate	float64

28	required_car_parking_spaces	Number of required parking spaces	int64
29	total_of_special_requests	Total number of special requests	int64
30	reservation_status	Reservation status	object
31	reservation_status_date	Date of reservation status	object

Table 2. The detail information of the telecom dataset

No	Column name	Description	Data type	Unique value
1	customerID	Customer ID	object	7043 unique values
2	gender	Whether the customer is a male or a female	object	Male, Female
3	SeniorCitizen	Whether the customer is a senior citizen or not	int64	1,0
4	Partner	Whether the customer has a partner or not	object	Yes, No
5	Dependents	Whether the customer has dependents or not	object	Yes, No
6	tenure	Number of months the customer has stayed with the company	int64	
7	PhoneService	Whether the customer has a phone service or not	object	Yes, No
8	MultipleLines	Whether the customer has multiple lines or not	object	Yes, No, No phone service

9	InternetService	Customer's internet service provider	object	DSL, Fiber optic, No
10	OnlineSecurity	Whether the customer has online security or not	object	Yes, No, No internet service
11	OnlineBackup	Whether the customer has online backup or not	object	Yes, No, No internet service
12	DeviceProtection	Whether the customer has device protection or not	object	Yes, No, No internet service
13	TechSupport	Whether the customer has tech support or not	object	Yes, No, No internet service
14	StreamingTV	Whether the customer has streaming TV or not	object	Yes, No, No internet service
15	StreamingMovies	Whether the customer has streaming movies or not	object	Yes, No, No internet service
16	Contract	The contract term of the customer	object	Month-to-month, One year, Two year
17	PaperlessBilling	Whether the customer has paperless billing or not	object	Yes, No
18	PaymentMethod	The customer's payment method	object	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)

19	MonthlyCharges	The amount charged to the customer monthly	int64	
20	TotalCharges	The total amount charged to the customer	int64	
21	Churn	Whether the customer churned or not	object	Yes or No

2.2. Data Preparation

Table 3. Null value of Hotel booking dataset

country	488 null value
agent	16340 null value
company	112593 null value

Table 4. Null value of churn telecommunication dataset

TotalCharges	11 null value

2.3. Data Preprocessing

Table 5. A part of original hotel booking dataset

	hotel	**is_canceled**	**lead_time**	**arrival_date_year**	**arrival_date_month**	**arrival_date_week_number**	**…**
0	Resort Hotel	0	342	2015	July	27	…
1	Resort Hotel	0	737	2015	July	27	…

2	Resort Hotel	0	7	2015	July	27	…

Table 6. A part of original customer churn telecommunication dataset

	customerID	**gender**	**SeniorCitizen**	**Partner**	**Dependents**	**tenure**	**…**
7038	6840-RESVB	Male	0	Yes	Yes	24	…
7039	2234-XADUH	Female	0	Yes	Yes	72	…
7041	8361-LTMKD	Male	1	Yes	No	4	…

Table 7. Result of using SMOTE for Telecommunication dataset

	After SMOTE
length of the data	10326
Churn	5163
Not Churn	5163

Table 8. Result of using SMOTE for hotel booking dataset

	After SMOTE
length of the data	150022
No	75011

Yes	75011

Table 9. A part of final hotel booking dataset

	hotel	year	month	day	meal_BB	meal_FB	…
0	1	2015	7	1	1	0	…
1	1	2015	7	1	1	0	…
2	1	2015	7	2	1	0	…

Table 10. A part of final customer churn telecommunication dataset

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	…
0	0	0	1	0	1	0	…
1	1	0	0	0	34	1	…
2	1	0	0	0	2	1	…

2.4. Feature selection

Table 11. The finals variables of the two datasets after using RFE and VIF

	Before Feature Selection	After Feature Selection
Hotel booking dataset	53 variables	23 variables
Telecommunication dataset	39 variables	18 variables

2.5. Ensemble algorithms application and evaluation

2.5.1. Predict all cases

Table 12. Report result of Churn prediction for all cases

Algorithms	Accuracy	Precision	Recall	F1-Score
RF	0.78	0.78	0.78	0.78
XGBoost	0.80	0.80	0.80	0.80
CatBoost	**0.80**	**0.81**	**0.80**	**0.80**
LightGBM	0.79	0.80	0.79	0.79
AdaBoost	0.78	0.79	0.78	0.78
GBC	0.80	0.80	0.80	0.80
ETC	0.80	0.80	0.80	0.80

Table 13. Report result of hotel booking prediction for all cases

Algorithms	Accuracy	Precision	Recall	F1-Score
RF	0.81	0.82	0.81	0.80
XGBoost	0.87	0.87	0.87	0.87
CatBoost	0.86	0.86	0.86	0.86

LightGBM	0.83	0.84	0.83	0.83
AdaBoost	0.81	0.82	0.81	0.81
GBC	0.83	0.83	0.83	0.82
ETC	**0.88**	**0.88**	**0.88**	**0.88**

2.5.2. Predict target variable

Table 14. Report result of Churn prediction for target value

Algorithms	Precision	Recall	F1-Score
RF	0.77	0.79	0.78
XGBoost	0.79	0.82	**0.81**
CatBoost	**0.80**	0.82	**0.81**
LightGBM	0.77	**0.84**	0.80
AdaBoost	0.75	**0.84**	0.79
GBC	0.78	0.83	0.80
ETC	0.79	0.82	0.80

Table 15. Report result of hotel booking prediction for target value

Algorithms	Precision	Recall	F1-Score
RF	0.90	0.68	0.78
XGBoost	0.90	0.82	0.86
CatBoost	0.90	0.81	0.85
LightGBM	0.89	0.76	0.82
AdaBoost	0.85	0.75	0.80
GBC	0.88	0.75	0.81
ETC	**0.90**	**0.86**	**0.88**