A Comparative Study of Data Mining: A Review Ensemble Algorithms

TABLES OF CONTENT

TABLES OF TABLES

TABLES OF FIGURES

LIST OF ACRONYMS

No.	Acronym	Detail word
1	RF	Random Forest
2	GBC	Gradient Boosting Classifier
3	ETC	Extra Trees Classifier
4	LightGBM	Light Gradient-Boost Machine

1. Proposed model

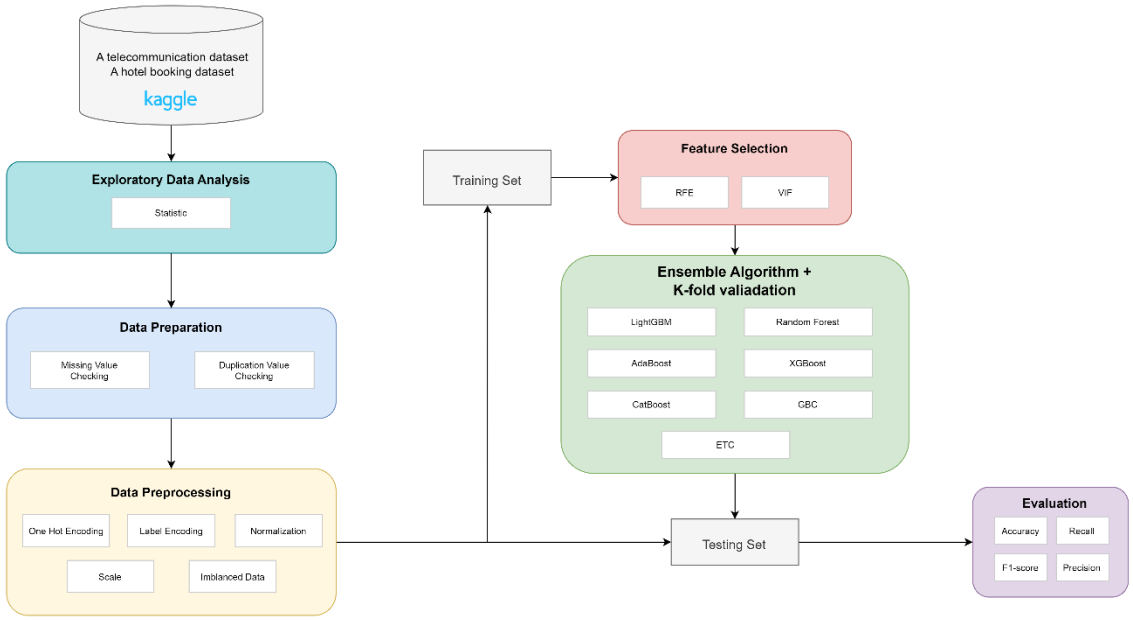

Figure 1 Proposed methodology framework

As shown in Figure 1, this section outlines the systematic approach undertaken in this study to analyze and predict key outcomes using machine learning techniques. The methodology encompasses a series of meticulously orchestrated stages, each contributing to the refinement and optimization of the analytical process. Beginning with dataset selection and acquisition, the journey progresses through exploratory analysis, data preparation, preprocessing, and advanced modeling techniques. The ultimate objective is to deliver accurate, reliable, and actionable insights through a judicious fusion of robust algorithms and comprehensive evaluations.

Stage 1: Dataset Acquisition and Selection

The initial phase of this study involved meticulous dataset acquisition from reputable sources such as Kaggle and gg data. From a pool of potential candidates, two distinct datasets were meticulously chosen: a comprehensive hotel booking dataset and a telecommunications dataset focusing on customer churn. The selection was driven by the relevance of the datasets to the study's objectives and the robustness of their contents.

Stage 2: Exploratory Data Analysis (EDA)

Before any substantive analysis, an exhaustive Exploratory Data Analysis (EDA) was conducted. This analytical step encompassed a systematic examination of data characteristics, distributions, patterns, and anomalies. This process provided critical insights into the fundamental attributes of the datasets, allowing for informed decision-making in subsequent phases.

Stage 3: Data Preparation and Quality Enhancement

The integrity of the datasets was a primary concern in this endeavor. Rigorous data preparation techniques were executed to ensure the highest quality standards. The identification and treatment of missing values were addressed with industry-accepted methods, and duplicates were systematically detected and eliminated to curtail the influence of redundant data.

Stage 4: Data Preprocessing for Model Readiness

The datasets underwent meticulous preprocessing to render them conducive to machine learning algorithms. Categorical variables underwent transformation via one-hot encoding and label encoding to capture their nuanced information. Furthermore, numerical features were normalized and scaled to mitigate the impact of differing magnitudes. To rectify class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied judiciously.

Stage 5: Optimal Feature Selection

To distill the datasets down to their most informative attributes, a comprehensive feature selection process was executed. The Recursive Feature Elimination (RFE) technique was employed to iteratively prune less relevant features, enhancing model efficiency. Simultaneously, the Variance Inflation Factor (VIF) analysis was conducted to identify and mitigate multicollinearity, ensuring model robustness.

Stage 6: Ensemble Algorithm Implementation with Cross-Validation

Ensemble algorithms, renowned for their ability to enhance predictive performance, were harnessed in this phase. The selected ensemble methods—namely, Random Forest, XGBoost, CatBoost, Gradient Boosting Classifier, Extra Trees Classifier, and Light Gradient-Boost Machine—were systematically implemented. The integration of K-fold cross-validation served to validate the models' performance across diverse data partitions, guarding against overfitting and affirming generalizability.

Stage 7: Rigorous Model Evaluation

A meticulous evaluation process was undertaken to ascertain the effectiveness of the ensemble algorithms. Performance assessment encompassed a comprehensive suite of metrics including Accuracy, Precision, Recall, and the F1-score. These metrics collectively provided a holistic evaluation of the models' capabilities in accurate classification and effective management of false positives and false negatives.

2. Experimental Result and Analysis

2.1. Dataset

The foundation of this study is built upon the careful curation of two distinct datasets, each offering unique insights into distinct domains of analysis.

The first dataset, termed the "Hotel Booking Dataset," constitutes a comprehensive repository of booking information pertaining to two distinct types of hotels: a city hotel and a resort hotel. Comprising 32 columns and a total of 119,390 transactions, this dataset encapsulates a wealth of booking-related attributes. Notably, the focal point of predictive interest is encapsulated within the "is_canceled" variable, a binary indicator of booking cancellation status. Through a meticulous examination of this dataset, we aim to unravel patterns and factors contributing to booking cancellations in the hospitality industry.

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_of_month	Day of month of arrival date	int64
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

These attributes encompass details spanning the booking process, guest demographics, and reservation specifics. Among these variables, noteworthy categories include booking characteristics such as cancellation status, lead time, and arrival date details. Additionally, the dataset captures key customer attributes like the number of adults, children, and babies associated with each booking. Factors like meal preferences, country of origin, market segment, and distribution channel contribute to comprehensive insights into guest behavior and hotel operations. Furthermore, the dataset provides an opportunity to analyze the impact of repeated guest status, previous cancellations, and bookings history on hotel occupancy and performance. The dataset's comprehensive nature enables an in-depth exploration of booking dynamics, customer preferences, and reservation trends, making it a valuable resource for understanding the intricate interplay between guests and hotels.

In parallel, the second dataset, christened the "Customer Churn Telecommunication Dataset," takes center stage with its insights into customer attrition trends within the telecommunications domain. This dataset's significance is underscored by its embodiment of 7,043 customer profiles, each encompassing 21 distinct features that offer a holistic representation of customer attributes. The focal point of predictive analysis within this dataset resides in the "Churn" variable, which serves as a pivotal binary indicator of customer churn. With an overarching objective to uncover factors influencing customer churn within the telecommunications sector, the study delves deep into the myriad attributes to discern patterns that prompt customer departure from service subscriptions.

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	object	Yes, No

		has dependents or not		
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

The data set includes information about:

- Services has been signed up for: phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

- Demographic information: gender, age range, and if they have partners and dependents

- Customer account information: tenure mean how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

- The last is our target, customers who left within the last month – the column is called Churn

2.2. Data Preparation

Both of the analyzed datasets exhibit distinct data characteristics in terms of duplication and null values. While no duplicate values are found within either dataset, some null values are present.

Table 3. Null value of Hotel booking dataset

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

In the Hotel Booking dataset, the 'country' column contains 488 null values, 'agent' has 16,340 null values, and 'company' records 112,593 null values. To mitigate the impact of these null entries, they are uniformly replaced with 0, facilitating smoother data processing.

Table 4. Null value of churn telecommunication dataset

TotalCharges	11 null value

Conversely, in the Customer Churn Telecommunication dataset, only 11 null values are observed within the 'TotalCharges' column. Given the minor quantity of null values and their limited impact on the overall dataset, a decision has been made to remove these entries. This approach ensures the dataset's integrity and consistency, and while a small fraction of the data is discarded, it has a negligible effect on the dataset's overall analytical 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	…

In the initial stages of data preprocessing, our focus was directed towards effectively handling both object and numeric variables within the dataset. For the hotel booking dataset, encompassing variables like hotel, arrival_date_month, meal, and others, a two-pronged encoding approach was employed. Given the non-ordinal nature of these variables and their lack of inherent order, we opted for one-hot encoding for those with more than two unique values. For those with only two unique values, we utilized label encoding.

Turning our attention to the telecommunications dataset, the variables painted a distinct picture. The dataset encompassed attributes like customerID, gender, SeniorCitizen, and more. The process of encoding was tailored to this context. With both datasets in view, a comprehensive approach was taken to cater to their unique characteristics.

For numeric variables such as is_canceled, lead_time, MonthlyCharges, and more, a rigorous normalization process was executed. To enhance the stability and uniformity of these variables, normalization methods like square root and logarithmic transformation were applied. Subsequently, the data was further standardized to a compact range using the StandardScaler technique, enabling fairer comparison across variables.

In light of the potential imbalance present within both datasets, we took proactive measures to address this issue. Specifically, the Synthetic Minority Over-sampling Technique (SMOTE) was adopted. By generating synthetic instances of minority class samples, SMOTE effectively balanced the class distribution, enabling our models to be more robust and equitable in their predictions.

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

This comprehensive approach to data preprocessing ensured that both object and numeric variables from both the telecommunications and hotel booking datasets were optimally prepared for subsequent analysis. By considering the distinct attributes of each dataset, we enhanced the overall quality and reliability of our predictive models.

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

Following the intricate process of data preprocessing, particularly in the context of one-hot encoding, the number of variables expanded significantly. This augmentation in the number of variables can have a dual impact on machine learning tasks. On one hand, it may elongate the processing time, and on the other hand, it could potentially compromise the final performance of our predictive models. To circumvent these challenges and streamline our feature space, we adopted a two-fold approach, employing Recursive Feature Elimination (RFE) and Variance Inflation Factor (VIF).

RFE, a dimensionality reduction technique, systematically gauges the relevance of each variable to the model's predictive power. It iteratively eliminates the least significant features, gradually refining the feature space while maintaining or even enhancing model performance. By allowing us to focus on the most impactful variables, RFE effectively enhances both computational efficiency and model interpretability.

Complementing RFE, VIF assists in identifying multicollinearity within the dataset, which can distort the interpretability of the features and potentially lead to unstable model results. Variables that are highly correlated with each other can negatively influence the model's precision. By quantifying the interdependence between variables, VIF helps us in identifying and subsequently removing redundant features, thereby enhancing the robustness of our models.

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

Together, the synergistic use of RFE and VIF serves as a powerful strategy to curate a refined and optimal set of features from the enriched variable pool generated during the data preprocessing phase. This not only streamlines the computational demands but also nurtures the models' predictive prowess, ultimately resulting in more efficient and effective machine learning outcomes.

2.5. Ensemble algorithms application and evaluation

Having meticulously prepared and processed the input datasets, we now proceed to a critical phase where we harness the potential of ensemble algorithms. Specifically, we leverage a suite of powerful ensemble techniques, including Random Forest (RF), XGBoost (XGB), CatBoost (CatB), LightGBM, Gradient Boosting Classifier, and Extra Tree Classifier. These ensemble methods collectively draw upon a diverse set of algorithms, enhancing the predictive capabilities of our models.

In this phase, our primary focus is twofold: prediction and comparison. Through the utilization of ensemble algorithms, we aim to predict outcomes with greater accuracy and reliability. By integrating the insights of multiple algorithms, we mitigate individual shortcomings and achieve more robust results. Moreover, we undertake a thorough comparison of the ensemble algorithms to discern their respective strengths and limitations.

To quantitatively evaluate the performance of these ensemble models, we employ a comprehensive set of evaluation metrics. These metrics encompass accuracy, precision, F1-score, and recall. Accuracy provides an overall measure of correctness, precision gauges the proportion of true positive predictions, F1-score balances precision and recall, while recall captures the model's ability to correctly identify positive instances. This multi-dimensional assessment framework empowers us to make informed decisions about the suitability of each ensemble algorithm for our specific prediction task.

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

As shown in the table , among the ensemble algorithms evaluated for Churn prediction, CatBoost emerged as the standout performer, showcasing the highest precision, accuracy, recall, and F1-score of 0.81, 0.80, 0.80, and 0.80, respectively. Its ability to accurately identify positive instances and minimize false negatives makes it a reliable choice for Churn prediction scenarios.

On the other hand, while all models displayed commendable performance, AdaBoost exhibited relatively lower accuracy, precision, and recall, with scores of 0.78, 0.79, and 0.78, respectively. Despite this, it still provides viable predictions, especially when a balanced approach between precision and recall is sought.

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**

In the context of hotel booking prediction as shown in the table , the ensemble algorithms were evaluated, yielding notable variations in their performances. Among these, Extra Tree Classifier (ETC) stood out as the frontrunner, showcasing the highest accuracy, precision, recall, and F1-score of 0.88 across the board. Its ability to provide consistently high predictions in all aspects underscores its suitability for accurate hotel booking predictions.

On the other hand, AdaBoost and Random Forest (RF) exhibited relatively lower scores across the metrics, with accuracy, precision, and recall hovering around 0.81. Despite this, both models still offer respectable predictions, albeit with room for improvement in terms of precision and recall.

In our thorough analysis of both Churn prediction and hotel booking prediction, a variety of ensemble algorithms were rigorously evaluated to ascertain their efficacy in delivering accurate and reliable predictions. For the Churn prediction task, a comprehensive comparison revealed that XGBoost, CatBoost, and Gradient Boosting Classifier (GBC) emerged as the top performers, boasting consistent accuracy, precision, recall, and F1-score values of approximately 0.80. These models demonstrated their ability to make well-rounded predictions, highlighting their suitability for Churn prediction applications.

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

The Churn prediction task has revealed a spectrum of model performances, ranging from the least effective to the most promising solutions. AdaBoost, while achieving a precision of 0.75, displayed notable shortcomings in terms of recall and F1-score, yielding scores of 0.84 and 0.79, respectively. This suggests that while AdaBoost may excel in capturing specific cases accurately, it may not be as comprehensive in identifying overall churn instances. Conversely, XGBoost, CatBoost, and LightGBM demonstrated superior capabilities, with recall values exceeding 0.82. These models, with precision and F1-scores hovering around 0.80 and 0.81, respectively, strike a favorable balance between identifying churn instances and achieving an overall strong performance.

This evaluation of Churn prediction models underscores the varying degrees of effectiveness among different algorithms. XGBoost, CatBoost, and LightGBM emerge as the most reliable options, delivering a harmonious blend of precision, recall, and F1-score. Conversely, AdaBoost lags behind in terms of recall, potentially hindering its ability to capture comprehensive churn instances. This analysis offers valuable insights into model selection for Churn prediction tasks, equipping decision-makers with the knowledge needed to make well-informed choices for their specific applications.

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**

Among the evaluated models for predicting hotel booking cancellations, we observe distinct variations in performance. AdaBoost emerged as the model with the lowest precision, recall, and F1-score values, indicating that while it can identify certain instances of booking cancellations accurately, it may not be as effective in capturing a broader range of cancellations. On the other end of the spectrum, XGBoost, CatBoost, and LightGBM demonstrated superior predictive capabilities. These models consistently achieved high precision values, indicative of their accuracy in classifying actual booking cancellations.

This assessment of hotel booking cancellation prediction models emphasizes the importance of precision and recall trade-offs. AdaBoost, while still demonstrating acceptable performance, lags behind in terms of recall, suggesting that it may miss some booking cancellations. Conversely, XGBoost, CatBoost, and LightGBM showcase a

robust balance between precision and recall, making them reliable choices for identifying booking cancellations. This analysis equips decision-makers with valuable insights into selecting the most effective model for their specific application needs.

Comparing the performance of the two prediction scenarios, where we focused exclusively on churn prediction and hotel booking cancellation prediction, reveals interesting insights. In the churn prediction task, the models exhibited varying levels of precision, recall, and F1-score. XGBoost demonstrated the highest precision and recall, underlining its capacity to effectively identify instances of churn. Conversely, AdaBoost displayed lower precision and recall values, indicating that it may miss certain cases. Moving to the hotel booking cancellation prediction, the models showcased their capabilities in predicting booking cancellations. Notably, the ensemble models, especially Random Forest, XGBoost, CatBoost, and Extra Tree Classifier, performed consistently well, offering high precision, recall, and F1-score values. While the specific priorities of each prediction scenario can influence model selection, these results provide a comprehensive understanding of how each model excels in different contexts.

3. Conclusion

In the ever-evolving landscape of machine learning, ensemble algorithms have emerged as powerful tools, transforming the predictive analytics landscape. This study delved deep into the realm of ensemble algorithms, showcasing their prowess in two distinct prediction scenarios: churn prediction and hotel booking cancellations. Through meticulous evaluation, we gained valuable insights into their performance and adaptability across various contexts.

Our approach encompassed a comprehensive methodology, commencing with the meticulous preprocessing of raw data, followed by variable encoding, scaling, and the application of ensemble techniques. This structured approach enabled us to effectively compare and evaluate models across a spectrum of scenarios. Notably, each ensemble algorithm exhibited unique strengths, outperforming others in specific prediction contexts.

The results highlighted the nuanced performance of ensemble algorithms in diverse prediction settings. From the high precision and recall of XGBoost and CatBoost in churn prediction to the consistent excellence of Random Forest, XGBoost, CatBoost, and Extra Tree Classifier in hotel booking cancellation prediction, the ensemble models demonstrated their adaptability and reliability.

While this study underscored the success of ensemble algorithms, certain limitations should be acknowledged. The incorporation of deep learning models holds promise for tackling intricate datasets and further refining predictions. Additionally, customizing model parameters through parameter tuning is recommended to fine-tune efficiency and accuracy, enhancing the overall performance of these models.

In conclusion, this study not only affirms the popularity and productivity of ensemble algorithms but also provides an encompassing methodological framework that navigates data handling, transformation, and ensemble application. Through thorough evaluation and comprehensive insights, this analysis contributes to the expanding understanding of ensemble algorithms' potential in modern machine learning.

References

[1] Brownlee, J. (2021, April 26). A gentle introduction to ensemble learning algorithms. MachineLearningMastery.com. https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

[2] Singh, A. (2020, April 20). 4 boosting algorithms you should know - GBM, XGBoost, LightGBM & CatBoost. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/02/4-boosting-algorithms-machine-learning/#:~:text=4%20Boosting%20Algorithms%20You%20Should,GBM%2C%20XGBoost%2C%20LightGBM%20%26%20CatBoost

[3] Verma, J. (2023, February 9). How to normalize data using scikit-learn in Python. DigitalOcean. https://www.digitalocean.com/community/tutorials/normalize-data-in-python

[4] Singh, A. (2020, April 20). 4 boosting algorithms you should know - GBM, XGBoost, LightGBM & CatBoost. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/02/4-boosting-algorithms-machine-learning/#:~:text=4%20Boosting%20Algorithms%20You%20Should,GBM%2C%20XGBoost%2C%20LightGBM%20%26%20CatBoost

[5] BlastChar. (2018, February 23). Telco customer churn. Kaggle. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[6] F1 score in Machine Learning: Intro & Calculation. V7. (n.d.). https://www.v7labs.com/blog/f1-score-guide#:~:text=F1%20score%20is%20a%20machine%20learning%20evaluation%20metric%20that%20measures,prediction%20across%20the%20entire%20dataset.

[7] How extra trees classification and regression algorithm works. How Extra trees classification and regression algorithm works-ArcGIS Pro | Documentation. (n.d.). https://pro.arcgis.com/en/pro-app/3.0/tool-reference/geoai/how-extra-tree-classification-and-regression-works.htm

[8] R, S. E. (2023, July 5). Understand random forest algorithms with examples (updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[9] Brownlee, J. (2021a, March 16). Smote for imbalanced classification with python. MachineLearningMastery.com. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[10] Brownlee, J. (2020, August 27). Recursive feature elimination (RFE) for feature selection in Python. MachineLearningMastery.com. https://machinelearningmastery.com/rfe-feature-selection-in-python/

[11] Team, T. I. (2023, May 11). Variance inflation factor (VIF). Investopedia. https://www.investopedia.com/terms/v/variance-inflation-factor.asp

[12] What is exploratory data analysis?. IBM. (n.d.). https://www.ibm.com/topics/exploratory-data-analysis

[13] Saini, A. (2023, August 4). Master the adaboost algorithm: Guide to implementing & understanding adaboost. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/