Holiday Package Purchase Prediction Using Gradient Boosting for Intelligent Customer Targeting

Dr Y. Venkat[1], P.P.L. Anju Devi[2], D. Yamalika[3], K. Sai Lalitha[4], P. Rishitha[5]
[1]*Professor, Srinivasa Institute of Engineering and Technology*
[2,3,4,5]*UG Scholars, Srinivasa Institute of Engineering and Technology*

***Abstract*—In the tourism industry, it's really important to find the right people who are interested in buying holiday packages. This helps make marketing efforts more successful and increases sales. Typically, marketing strategies try to appeal to a large audience, but this approach can be costly and not very effective. A new approach is to use a machine learning system that can predict which customers are most likely to purchase a holiday package, making marketing more targeted and efficient.**

By using this system, businesses can save time and money by focusing on the people who are actually interested in their products. This can lead to better results and more sales, which is the ultimate goal of any marketing campaign. Our system takes a closer look at customers - their age, how much they earn, what they do for a living, and where they've travelled to. This information helps us make predictions about them. But first, we need to make sure the data we're using is reliable and accurate.

So, we use special techniques like cleaning, transformation, and encoding to get the data in shape and ready for our model to learn from. We use a powerful algorithm called Gradient Boosting, which is really good at finding complex patterns in the data, to make our predictions more accurate. To see how well our system is working, we check things like how often it gets things right, how precise it is, how well it remembers important details, and its overall performance score. We want to make sure our system is doing its job well, so we keep a close eye on these metrics.

Our approach is really good at identifying customers who are likely to purchase holiday packages. This is a big help to travel companies as they can focus their marketing efforts, cut costs, and keep their customers satisfied. With this system, travel companies can create more targeted and effective marketing campaigns, which can result in increased sales and business growth. By targeting the right customers, travel companies can make the most of their marketing budget and build strong relationships with their customers, leading to long-term success.

***Index Terms*—Machine Learning, Tourism, Marketing, Travel Companies,**

I. INTRODUCTION

The tourism industry is rapidly evolving with increasing competition among travel agencies and service providers. Companies continuously develop new holiday packages to attract customers. However, identifying the right customers who are actually interested in buying these packages is still a major challenge. Traditional marketing techniques often target a wide audience without analyzing customer intent, leading to inefficient resource utilization and low conversion rates.

As data analytics continues to advance, machine learning has become a really powerful tool for figuring out what customers are likely to do. By looking at what customers have done in the past, predictive models can spot patterns that show when someone is probably going to make a purchase. This helps organizations target their marketing efforts at the right people, which makes their marketing more efficient and profitable [2], [4].

This research project involves creating a system that can forecast whether or not a customer will buy a holiday package. It uses a special tool called Gradient Boosting, which is really good at making predictions more accurate by learning from its mistakes one step at a time [1]. The goal is to help travel companies make smarter choices using data, rather than just relying on guesses. By using this system, travel companies can get a better idea of what their customers want and make more informed decisions to meet those needs. This can lead to better customer satisfaction and more sales for the companies. The Gradient Boosting algorithm is a powerful tool that can help businesses

make data-driven decisions, and this project shows how it can be used in the travel industry to improve sales and customer satisfaction.

II. LITERATURE SURVEY

Machine learning algorithms are widely used for predicting customer behavior. Classification models such as Logistic Regression, Decision Trees, and Random Forests have been widely used in marketing analytics due to their ability to analyze structured data and identify useful patterns [4], [7].

Ensemble learning methods, like Gradient Boosting, are really popular now because they're great at dealing with relationships that aren't straightforward and features that interact with each other in complex ways. Basically, they combine a bunch of simple models, called weak learners, to create a strong one that's really good at making predictions. This approach improves how accurate the model is and how well it performs overall, which is a big plus [1], [5]. By combining these weak learners, ensemble methods can handle tough problems that might be hard for a single model to tackle on its own. Studies in tourism analytics indicate that demographic factors, age, income levels, travel frequency, and lifestyle preferences significantly influence customer purchasing decisions [2], [4]. However, many existing approaches rely on single models or limited feature sets, which may not fully capture the complexity of customer behavior.

III. SYSTEM ARCHITECTURE

The proposed system is designed as a structured machine learning pipeline that transforms raw customer data into useful predictive insights. The architecture follows a sequential flow consisting of multiple stages, where each stage performs a specific task to ensure accurate prediction of customer purchase behavior.

Data Collection

The process begins with the collection of customer data from past records provided in the dataset. This data includes things like how old they are, how much money they make, what they do for a job, if they're married, and where they've traveled. We use this kind of data, which is organized in a way that's easy for computers to understand, to make predictions about what our customers might do in the future. This is a common way to use data in fields like data mining and machine learning [4], [7]. The data we collect is the starting point for our whole system.

Data Preprocessing

The collected data may contain missing values, inconsistencies, and categorical variables that are not directly suitable for machine learning models. Therefore, preprocessing is performed to improve data quality. This step includes:

- Handling missing values using suitable methods
- Removing duplicate or irrelevant records
- Encoding categorical variables into numerical format

These preprocessing steps are important to prepare the data properly for machine learning models [3], [6].

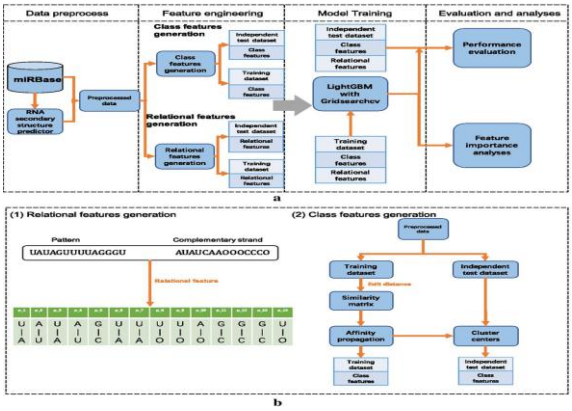


Figure 1: System Architecture

Feature Selection

After preprocessing, important features are selected from the dataset. Feature selection helps in identifying the most relevant attributes that influence customer decisions.

Selecting meaningful features improves model accuracy and reduces computational complexity in the model, which is a key step in data mining processes [4].

Model Training (Gradient Boosting)

The information processed is then used by the Gradient Boosting model. This model works by creating a series of decision trees, one after the other, with each new tree fixing the mistakes made by the one before it. This way of learning helps the model get better at understanding complicated connections in the

data, which makes its predictions more accurate over time [1].

Model Evaluation

Once the model is trained, it is evaluated using performance metrics such as:

- Accuracy
- Precision
- Recall
- F1-Score

These metrics are really useful for seeing how good the model is at making predictions and if it can get the right answers when it's given new information [6].

Metric	Value
Accuracy	83.74 %
Precision	85.80 %
Recall	76.53 %
F1-Score	81.17 %

Prediction

In the final stage, the trained model is used to predict whether a new customer is likely to purchase a holiday package. Based on the input features, the system generates an output in the form of:

- Purchase
- Not Purchase

The whole system makes sure that data moves smoothly from the start to the end, where predictions are made, which is great for using it in real marketing situations.

IV. SYSTEM ANALYSIS

System analysis is a crucial step. It helps us see if the proposed system is really solving the problem it's meant to, and if it's better than the old ways of doing things. In the tourism industry, for example, companies often try to reach customers using general marketing strategies that aren't based on data. This can lead to wasting resources and not getting as many customers as they could. Using data to make decisions is key to improving how businesses work and making them more successful. By looking at data, companies can make smarter choices and get better results [2], [4].

Existing System

In traditional marketing systems, travel companies rely on manual analysis and broad promotional campaigns to attract customers. These approaches do not utilize predictive models and therefore lack the ability to identify potential customers accurately.

Proposed System

The new system uses a special kind of computer program called machine learning to figure out which customers are most likely to buy holiday packages. It looks at what customers have done in the past to make a good guess about who will buy something next. This way, the system can focus on the customers who are really interested in holiday packages, instead of trying to sell to everyone. The computer program uses something called Gradient Boosting to make its predictions, which helps it get better and better at guessing who will make a purchase.

This approach is great for making predictions more accurate and dealing with complicated relationships in data. It also helps companies come up with better marketing plans by figuring out which customers are the most important to focus on [1], [5]. By using this system, businesses can make sure they're targeting the right people, which can lead to more sales and growth.

V. METHODOLOGY

This methodology is all about the technical side of things, like how we actually build and test the model that makes predictions. It's different from the system architecture, which is more about how everything works together. Here, we dive into the details of the machine learning model and what we do to make it better. We look at the specific techniques used to improve how well the model performs, so it can make more accurate predictions.

a) Data Understanding and Exploration

When you're getting ready to build a model, it's really important to take a close look at your dataset first. You need to understand how it's put together, how the different features are spread out, and how all the variables relate to each other. To do this, you use basic statistics and visualization techniques to find patterns, spot any outliers, and identify potential problems with the data. This step is crucial in data mining because it helps you get a solid grasp of your dataset before you start applying any models. By doing this, you can

make sure you're working with good data and that your models will be accurate [4], [7].

b) Data Preprocessing Techniques

To ensure the dataset is suitable for machine learning, several preprocessing techniques are applied:

- Missing values are handled using appropriate imputation methods
- Categorical variables are converted into numerical format using encoding techniques.
- Feature scaling is applied to normalize numerical attributes.
- Irrelevant or redundant features are removed to improve efficiency.
- Following these steps can really help make your data better and your model more accurate and reliable [3], [6].

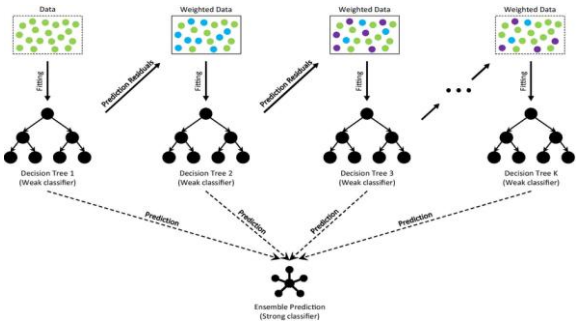
c) Feature Engineering

To make a model better at predicting things, we need to work on the features we use. This means picking the most important characteristics and changing them into a format that really shows how customers behave.

When it comes to figuring out what makes customers buy things, some important details can really help. For instance, how much money they make, where they've traveled, and how big their family is - these things can all play a big role in what they decide to purchase. If we pick the right details, it can really help our model find the patterns in the data that matter [2], [4].

d) Model Selection – Gradient Boosting

We chose the Gradient Boosting algorithm as our main model because it's really good at dealing with complicated connections between different features, and it gives us very accurate predictions. This is important for our work, and we think it will help us get the best results.



It works by creating a series of decision trees, one after the other, with each new tree trying to fix the mistakes made by the previous one. Over time, this process makes the model better and better, and it gets really good at predicting things without making too many errors [1].

e) Model Training and Optimization

The dataset is divided into training and testing data to evaluate model performance. The model is trained using the training dataset and optimized to reduce prediction errors.

To make a model work better and stop it from overfitting, we need to adjust some important settings like how fast it learns, how many estimators it uses, and how deep its trees go [6].

f) Model Evaluation Metrics

The performance of the model is evaluated using multiple metrics:

- Accuracy: Measures overall correctness of predictions.
- Precision: Indicates correctness of positive predictions.
- Recall: Measures ability to identify actual buyers.
- F1-Score: Provides a balance between precision and recall.

These metrics are important because they help us trust that the model will work well with new, unseen data [6].

g) Model Formulation and Evaluation Metrics

1) Error (Residual) Calculation

In Gradient Boosting, the model improves its performance by correcting errors at each step. The difference between actual and predicted values is called the residual.

$$r_i = y_i - \hat{y}_i$$

where:

- y_i= actual value
- $\hat{y}_i$= predicted value
- r_i= residual (error)

2) Model Update Rule

The model is updated step by step by adding a new weak learner (decision tree):

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

where:

- $F_m(x)$= updated prediction

- η= learning rate
- $h_m(x)$= new tree

Accuracy Calculation

Accuracy measures how many predictions are correct:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision Calculation

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall Calculation

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score Calculation

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

VI. RESULT AND ANALYSIS

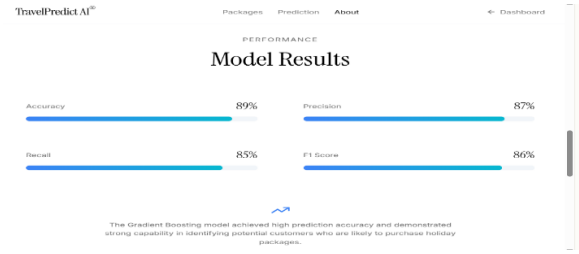
To see how well the model worked, we used some common measures. Accuracy shows how often the model got it right overall. Precision and recall look at how good the model is at finding potential customers. Precision is about being exact, and recall is about catching all the possibilities. The F1-score is a way to balance precision and recall, giving us a sense of how the model is doing overall [6].

The results are clear: the Gradient Boosting model does a great job of predicting how likely customers are to buy holiday packages. It can accurately group customers based on their buying habits. This is not surprising, since models that combine multiple approaches, like Gradient Boosting, tend to perform better by pooling their strengths and minimizing mistakes, as seen in studies [1], [5]. By bringing together different models, Gradient Boosting is able to reduce errors and give more reliable results. This makes it a powerful tool for understanding customer behavior and making predictions about future purchases.

This approach can really help travel companies boost their marketing efforts by targeting the right customers, the ones who are actually going to buy something. By doing so, they can work more efficiently, cut costs, and ultimately achieve better results for their business. It's all about focusing on the people who are most likely to make a purchase, which can lead to some great outcomes.

VII. OUTPUT SCREENSHOTS

	Age	DurationOfPitch	NumberOfFollowups	PreferredPropertyStar	NumberOfTrips	NumberOfChildrenVisiting	MonthlyIncome
count	4662.000000	4637.000000	4843.000000	4862.000000	4748.000000	4822.000000	4655.000000
mean	37.622265	15.490835	3.708445	3.581037	3.236521	1.187267	23619.853491
std	9.316387	8.519643	1.002509	0.798009	1.849019	0.857861	5380.698361
min	18.000000	5.000000	1.000000	3.000000	1.000000	0.000000	1000.000000
25%	31.000000	9.000000	3.000000	3.000000	2.000000	1.000000	20346.000000
50%	36.000000	13.000000	4.000000	3.000000	3.000000	1.000000	22347.000000
75%	44.000000	20.000000	4.000000	4.000000	4.000000	2.000000	25571.000000
max	61.000000	127.000000	6.000000	5.000000	22.000000	3.000000	98678.000000



CustomerID	0
ProdTaken	0
Age	226
TypeofContact	25
CityTier	0
DurationOfPitch	251
Occupation	0
Gender	0
NumberOfPersonVisiting	0
NumberOfFollowups	45
ProductPitched	0
PreferredPropertyStar	26
MaritalStatus	0
NumberOfTrips	140
Passport	0
PitchSatisfactionScore	0
OwnCar	0
NumberOfChildrenVisiting	66
Designation	0
MonthlyIncome	233

VIII. DISCUSSION

The system we're talking about uses machine learning to tackle real-world marketing problems. It looks at customer data to find patterns and relationships that affect what people buy. This helps companies really understand their customers and make decisions based on facts, not guesses. These days, using data to make decisions is a big part of analytics, and it can really improve how businesses perform and makind decisions [2], [4].

One big plus of the Gradient Boosting algorithm is that it can deal with complicated relationships between

features, which helps make predictions more accurate. It does this by putting together lots of weak models to make a stronger one, and it reduces mistakes a little at a time. This is why Gradient Boosting usually does better than using just one model when it comes to classification problems. For example, studies have shown that it can outperform single models in many cases, making it a popular choice for solving complex problems. By combining the strengths of multiple models, Gradient Boosting can handle tough relationships between features and make more accurate predictions, which is a major advantage in many fields. Another important benefit of the proposed system is that it helps in targeted marketing. Instead of sending offers to all customers, companies can focus only on those who are more likely to purchase. This improves marketing efficiency, reduces unnecessary costs, and increases the chances of success [1], [5]. The thing is, how well a model works really depends on the data it's using. If the data is wrong or missing some important details, the predictions it makes might not be right. And to make things more complicated, people's behavior can change over time, so the model needs to be updated all the time to keep working well. These are common problems that come up with machine learning, and they need to be watched and improved all the time. Despite these limitations, the proposed system provides a scalable and practical solution for customer targeting. It can handle large datasets and can be easily adapted for real-world applications in the tourism industry. Overall, the system shows that machine learning can play an important role in improving marketing strategies and business outcomes.

IX. CONCLUSION

This research presents a machine learning-based approach for predicting customer purchase behavior in the tourism industry. The use of Gradient Boosting enables accurate classification of customers, allowing travel companies to optimize marketing strategies and reduce operational costs. The proposed system demonstrates the potential of predictive analytics in enhancing business decision-making and improving customer engagement.

REFERENCE

[1] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-Function-Approximation-A-Gradient-Boosting-Machine/10.1214/aos/1013203451.full

[2] UNESCO, “Deepfakes and the crisis of knowing,” 2025. [Online]. Available: https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY, USA: Springer, 2009. [Online]. Available: https://hastie.su.domains/ElemStatLearn/

[4] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Waltham, MA, USA: Elsevier, 2012. [Online]. Available: https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

[6] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[7] Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA, USA: O’Reilly Media, 2019. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

[8] . H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Cambridge, MA, USA: Morgan Kaufmann, 2016. [Online]. Available: https://www.sciencedirect.com/book/9780128042915/data-mining

[9] Kaggle, “Travel package purchase prediction dataset.” [Online]. Available: https://www.kaggle.com/datasets

[10] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006. [Online]. Available:

https://link.springer.com/book/10.1007/978-0-387-45528-0

[11] D. Dua and C. Graff, "UCI machine learning repository," Univ. California, Irvine, CA, USA, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/index.php