Predictive Analysis of Customer Booking Completion with British Airways: Harnessing Data Mining Techniques

Conference Paper · December 2023

DOI: 10.1109/SCOReD60679.2023.10563801

3 authors:

Umar Sunusi Umar
Asia Pacific University of Technology & Innovation
3 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Mafas Raheem
Asia Pacific University of Technology & Innovation
53 PUBLICATIONS **187** CITATIONS
SEE PROFILE

Muhammad Ehsan Rana
Asia Pacific University of Technology & Innovation
166 PUBLICATIONS **962** CITATIONS
SEE PROFILE

Predictive Analysis of Customer Booking Completion with British Airways: Harnessing Data Mining Techniques

Umar Sunusi Umar
School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia
tp068453@mail.apu.edu.my

Raheem Mafas
School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia
raheem@apu.edu.my

Muhammad Ehsan Rana
School of Computing
Asia Pacific University of Technology & Innovation
Kuala Lumpur, Malaysia
muhd_ehsanrana@apu.edu.my

2023 IEEE 21st Student Conference on Research and Development (SCOReD) | 979-8-3503-1882-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/SCOReD60679.2023.10563801

***Abstract*—In this study, advanced data mining techniques were applied to predict and categorize customer ticket sales for British Airways based on online behaviour. Leveraging a unique dataset comprising 50,000 observations and 14 characteristics provided by British Airways, sophisticated data preprocessing, including the management of outliers, character variables, and missing data, was undertaken. Two machine learning models Random Forest and Decision Trees were employed and evaluated for their performance. The Random Forest model exhibited remarkable accuracy, achieving 93%, underscoring its effectiveness in forecasting customer booking patterns. In contrast, the Decision Tree model yielded a moderate accuracy of 63.5%, with potential for improvement through hyperparameter tuning. These outcomes contribute to the ongoing discourse on refining customer booking predictions, showcasing the Random Forest model's prowess and identifying areas for enhancement in the Decision Tree model. The study concludes with actionable recommendations for British Airways, emphasizing timely customer engagement, flexible flight schedules, tailored packages, online platform improvements, enhanced in-flight experiences, group booking accommodations, optimized pricing strategies, and catering to dietary preferences.**

Keywords—Customer Ticket Purchases, British Airways, Machine Learning Models, Marketing Strategies, Airline Industry.

I. INTRODUCTION

Airlines work to enhance their marketing plans and boost client conversion rates in today's cutthroat aviation sector. A well-known airline, British Airways (BA), understands the value of using data-driven strategies to forecast client ticket sales [1]. Data mining techniques such as classification, clustering, statistical analysis, and prediction are effectively employed to uncover hidden information and internal connections [2]. This research examines British Airways' efforts to create a machine-learning model that can correctly forecast a customer's likelihood of completing a ticket purchase based on their online behaviour. For an airline to survive, remain competitive, be profitable, and experience sustainable growth in one of the most competitive business settings, it must improve its understanding of consumer demands and delivery of excellent service. Because service quality is viewed as the cornerstone of customer pleasure, airline firms make an effort to measure their service quality within the industry to establish and retain a competitive advantage by ensuring the happiness of their clients [3]. Customers may stay loyal to businesses if they receive high-quality service and aim for high customer satisfaction [4]. Retaining current customers also saves money because it is less expensive to do business with them already than to find new ones [5]. Through the constructive intent of satisfied consumers, businesses may attract new clients.

On April 1, 1974, British Airways was founded. Harmondsworth in London is home to the company's headquarters. At the beginning of 1987, the British government launched British Airways [6]. By its name, each airline kept flying. Because consumers are so important to BA operations, the firm actively seeks their input when developing service and product concepts. This helps the company better understand its target market. Imagine if the airline is unable to meet the needs of its clients and fails to communicate with them in a way that keeps them emotionally invested in the brand. Its reputation would suffer, and sales would increase more slowly in such a scenario. Customers place high importance on dependability, value for money, and user experience from the time they make a reservation until they reach their destination [7]. Consumer choices are being influenced by ethical and environmental issues more and more [8]. Over the long run, BA routes have continued to expand along with the intensifying competition in the carrier's industry. These days, BA Routes has been recognized as one of the leading airlines in green innovation variety and environmentally conscious important decisions. The primary carrier at the time to participate in the initiative of European countries reducing missions to do nursery inspections was BA Routes. Aside from the curiosity advancements, how people fly has seen enormous changes as a result of the BA routes [9]. The BA routes pioneered the custom of travellers generating their tickets.

In light of this, British Airways is proactively utilizing data analytics and predictive modelling to enhance its marketing initiatives [1]. BA can efficiently deploy resources, customize promotions, and improve the entire customer experience by properly forecasting consumer ticket purchases. The knowledge acquired from this study will help BA maintain its leadership position in the market while providing top-notch customer service and achieving sustainable development. The unique research issue of forecasting consumer ticket purchases for British Airways will be covered in this study. The study goals will be outlined, the methodology used, the data preparation process will be covered, the machine learning model's development will be discussed, its performance will be evaluated, key factors influencing customer purchase decisions will be identified through feature analysis, and actionable insights will be offered to improve conversion rates and BA marketing efforts. British Airways may use data-driven tactics to improve customer happiness, boost brand loyalty, and promote corporate growth in the fiercely competitive airline industry by using the results of this thorough investigation.

A. Problem Overview

British Airways' main goal is to overcome the difficulty of converting clients by comprehending the aspects that affect customers' decision-making. British Airways wants to find trends and insights that will help them with their marketing strategies and increase the possibility that customers will buy tickets by looking at customer behaviour on their website. An extensive dataset from British Airways that contains useful data on website visitor behaviour has been made available. This dataset contains a variety of characteristics of booking information such as the number of passengers, route type, flight hours, purchase lead, and other possibly pertinent elements. The dataset offers a rich supply of data for training and creating a prediction model while maintaining data privacy and adhering to any data protection requirements.

B. Questions of the Study

- How can a machine learning model be built and trained effectively to capture the predictive power of the variables in predicting customer ticket purchases for British Airways?
- What evaluation technique can be employed to assess the performance of the trained machine learning model in predicting customer ticket purchases accurately and ensure its reliability?
- What are the key factors that significantly influence customers' purchase decisions on the British Airways website?
- What are the actionable insights derived from the analysis and findings that can be implemented to enhance British Airways' marketing strategies and increase customer conversion rates?

C. Aims and Objectives of the Study

This case study aims to create a reliable machine-learning model that can forecast client ticket purchases for British Airways based on their website behaviour.

- To build and train a machine learning model that captures the predictive power of the variable.
- To evaluate the model's performance using appropriate evaluation techniques to ensure reliability.
- To identify the key factors influencing customers' purchase decisions through feature analysis.
- To present actionable insights that can be used to optimize British Airways' marketing efforts and improve conversion rates.

D. Scope of the Study

To construct a prediction model specifically suited to British Airways' marketing needs, this case study will analyze customer behaviour on the British Airways website. The study scope excludes extraneous elements like market conditions and competitive evaluations that are not within the website's purview. British Airways wants to use the findings and recommendations from this study to enhance its marketing strategies and boost customer conversion rates. By addressing these objectives, this case study aims to provide British Airways with valuable insights and a data-driven approach to predict customer ticket purchases. These insights will enable British Airways to optimize its marketing strategies, allocate resources more effectively and enhance the overall customer experience.

II. Related Works

This section provides a review of pertinent prior studies that are directly connected to the issue of anticipating client ticket sales for British Airways based on online behaviour, as well as discussions of similar fields like hotel booking and the like. Studies that have used comparable datasets or applied comparable techniques to handle comparable research goals are the main emphasis. Later, the references included in this part will be compared to the findings of the current study, demonstrating the uniqueness and contribution of the latter. There is no report available for the same dataset that is currently being refined.

The study of [10] concentrates on developing a machine learning model to forecast consumer reservations with BA. The work investigates feature engineering and the performance evaluation of models. The paper offers perceptions of the use of several machine learning techniques. The accuracy of the Random Forest Classifier, Support Vector Machine, and Gradient Boosting used by the author is 84.41%, 85.00%, and 85.00%, respectively. The British Airways virtual internship, as explored by [11], delved into client concerns regarding flights, seating, service, and scheduling. The study proposed improvements, such as enhancing food services and refining refund processes. Utilizing XGBoost, the predictive model discerned pivotal variables influencing customers' completion or cancellation of booking procedures. This research offers valuable insights into optimizing customer experiences within the aviation sector. In the study of [12] utilizing the same dataset, the initial implementation of the Random Forest algorithm yielded poor predictive accuracy, with an AUC score of 0.53. Subsequently, the XGBoost method was employed, resulting in a more respectable AUC score of 0.78. Based on the project findings, [12] advised British Airways to investigate airports with higher conversion rates and concentrate their marketing efforts in those areas. In a study conducted by [13], the utilization of Random Forest for predictions on the same dataset involved the deliberate exclusion of two columns before proceeding with model building. As a result, the achieved accuracy was 62%.

In the comprehensive study of [14], the researcher embarked on a thorough exploration of a dataset aimed at predicting customer behaviours within the aviation sector. The analytical journey commenced with a meticulous data visualization process using Sweetviz, providing a nuanced depiction of the dataset's characteristics. Following this, rigorous data preprocessing measures were implemented to ensure the dataset's suitability for modelling. The modelling phase showcased the application of a diverse array of algorithms, including Random Forest Regressor, Logistic Regression, Linear Regression, Decision Tree Regressor, and SVR. Model tuning was executed with precision using StratifiedKFold, ultimately leading to the identification of Random Forest Regressor with StratifiedKFold as the superior performer among the tested models.

Beyond model evaluation, the study delved into feature importance, singling out "Route" as the pivotal factor influencing customer booking predictions. The model's effectiveness was quantitatively assessed through key metrics, revealing an impressive overall score of 0.8458, an R2 score

of 0.8458, a mean absolute error of 0.2883, and a mean squared error of 0.1591. These results not only underscore the model's robust performance but also lay a strong foundation for refining customer experiences in the aviation sector.

A comprehensive study of [15] investigated the application of machine learning classification algorithms to predict client loyalty in the hotel sector. The study aimed to provide actionable insights for hotel businesses to enhance their CRM programs. Employing the CRISP-DM technique, the authors implemented three classification algorithms random forest, decision tree, and logistic regression. The study rigorously compared their performances using a confusion matrix, reporting accuracy ratings of 57.83%, 71.44%, and 69.91%, respectively, for logistic regression, decision tree, and random forest algorithms. The decision tree method outperformed with a score of 71.44%, demonstrating superior performance in predicting consumer loyalty. The study successfully achieved its objectives and meticulously reported its findings. In a distinct context, [16] conducted a study to develop a model for forecasting airline loyalty based on antecedents identified in previous research. Utilizing the snowball sampling technique, the study involved 614 domestic airline passengers who responded to a questionnaire. The questionnaire, comprising 16 scale questions derived from prior studies, focused on factors influencing airline loyalty, including passenger pleasure, airline service quality, passenger perceived value, and airline image. The study employed covariance-based Structural Equation Modelling and Artificial Neural Network theory for predictive analysis, revealing an impressive 89% accuracy rate for the artificial neural network model in predicting airline loyalty.

Another research of [17] delved into the airline sector to analyze customer churn risk and satisfaction using deep learning algorithms on survey data. The study considered perspectives from flight attendants and passengers to enhance predictive model accuracy. Employing various models such as KNN, Decision Tree, XGBoost, Random Forest, CNN, and CNN-LSTM, the research targeted customer churn risk and satisfaction, evaluating precision, recall, F1 score, and accuracy metrics. The results varied across models, with CNN-LSTM exhibiting the highest accuracy at 94% for customer churn risk and 90% for customer satisfaction. Furthermore, [18] tackled the challenging issue of predicting customer churn in their study. They proposed a six-step technique involving artificial intelligence and machine learning approaches, including boosting and ensemble methods, logistic regression, naive Bayes, support vector machine, random forest, and decision tree. Through meticulous evaluation using K-fold cross-validation, confusion matrix, and AUC curve, the study identified Adaboost and XGBoost classifiers as the most accurate, with respective values of 81.71% and 80.8%, outperforming other models and achieving the best AUC score of 84.

In client retention research, the study by [19] focused on first-time client retention in online reservation systems. They evaluated various prediction models, and the Generalized Additive Model (GAM) emerged as the most accurate, boasting an AUC value of 0.673. Despite its apparent simplicity, GAM outperformed computationally expensive ensemble learning techniques, including Bagging, Random Forest, Stochastic Gradient Boosting, and AdaBoost, which each had an AUC of 0.653. Notably, the sophisticated XGBoost technique achieved a competitive AUC of 0.672, nearly matching the performance of the GAM model. These findings underscore the effectiveness of diverse predictive models in addressing client retention challenges across sectors. In Another study of [20], five supervised machine learning algorithms Gaussian Naive Bayes, Support Vector Machine, K Nearest Neighbors, Decision Tree, and Random Forest Classifiers were scrutinized for predicting customer attrition using the Kaggle Churn_Modelling dataset. The Random Forest Classifier outperformed other algorithms in forecasting customer turnover, excelling in terms of accuracy, precision, and recall. This underscores the efficacy of machine learning approaches in effectively anticipating and addressing client turnover issues, particularly in the banking and corporate sectors. The Random Forest Classifier achieved an impressive accuracy of 87%, precision of 86%, and recall of 87% using default parameter settings on the Churn_Modelling dataset.

In a study by [21], the focus was on forecasting customer cancellations and their impact on service capacity planning. Backpropagation neural networks (BPNs) and general regression neural networks (GRNNs) were utilized to craft prediction models. The results revealed significant predictive capabilities of both BPN and GRNN models in anticipating client cancellations, providing valuable insights for managers to navigate the possibility of reservation cancellations. These models not only support dynamic service capacity scheduling to avoid surpluses and insufficient capacity but also demonstrate high specificity values, with GRNN at 87.14% and BPN at 80.00%, showcasing their accuracy in recognizing non-cancellations. The findings underscore how data mining approaches enhance forecast accuracy and facilitate effective service capacity management.

In their investigation into anticipating cancellations of reservations within the hospitality sector, specifically in resort hotels, [22] sought to overcome the challenges presented by booking cancellations. The study demonstrated the practicality of reliably forecasting booking cancellations with accuracy rates exceeding 90% by employing data science techniques such as data visualization, data mining, and machine learning. Framing booking cancellation prediction as a classification challenge, the study identified influential variables using data analytics and visualization tools alongside the mutual information filter. Leveraging machine learning, notably the decision forest algorithm, the research developed models to categorize reservations likely to be cancelled, achieving accuracy rates above 90% for all models. Notably, models for hotels 1, 2, and 3 exhibited accuracy rates of 98.6% and 97.4%, respectively, with AUC values exceeding 93.5%. These findings underscore the efficacy of machine learning techniques in crafting predictive models for booking cancellations, offering valuable insights for hotel management to refine cancellation rules, enhance demand estimation, and optimize overbooking strategies. This contributes to more precise pricing and inventory allocation decisions, ultimately elevating the overall performance of revenue management in the hotel sector.

The relevant studies covered in this part shed light on how machine learning methods are being used in the airline sector as well as other industries that include reservations or bookings to forecast customer behaviour. Similar datasets have been used by this research to create prediction models, which have also investigated different algorithms and assessment criteria. This analysis adds value by concentrating on British Airways in particular, utilizing a distinctive dataset

of online consumer behaviour, and assessing how well a machine learning model performs in accurately forecasting ticket purchases. Furthermore, to maximize British Airways' marketing efforts and raise conversion rates, it is important to pinpoint the critical variables that have a substantial impact on consumer purchasing choices. Table I provides a fast reference for readers to understand the benefits and drawbacks of various algorithms by summarizing the performance metrics of the prediction models employed in each research.

TABLE I. RELATED WORKS MODELS AND RESULT

Study	Methodology	Dataset Used	Performance Measures
[13]	Random Forest	British Airways	Accuracy 62%
[10]	Random Forest Classifier, Support Vector Machine, and Gradient Boosting	British Airways	Accuracy 84.41%, 85.00%, and 85.00%,
[11]	XGBoost	British Airways	Null
[12]	Random Forest, XGBoost	British Airways	Accuracy 0.53 and 0.78
[14]	RandomForestRegressor, LogisticRegression, LinearRegression, DecisionTreeRegressor, SVR	British Airways	Score: 0.8458, R2 Score: 0.8458, MAE: 0.2883, MSE: 0.1591
[15]	CRISP-DM, logistic regression, decision tree, random forest	Hotel sector	Accuracy 57.83%, 71.44%, and 69.91%,
[16]	Structural Equation Modelling, Artificial Neural Network	Airline passengers	Accuracy: 89% accuracy
[17]	kNN, Decision Tree, XGBoost, Random Forest, CNN, and CNN-LSTM.	Survey data from Korean airline users Customer churn risk and customer happiness	Accuracy: Range from 79% to 94% and 79% to 90%
[18]	Adaboost and XGBoost	Telecommunications industry	Accuracy: 81.71% and 80.8%
[19]	GAM, Bagging Random Forest, Stochastic gradient boosting, AdaBoost, and XGBoost	First-time customer retention	0.673, 0.614, 0.651, 0.653, 0.653, and 0.672
[20]	Gaussian Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest	Churn_Modelling dataset	Accuracy: 81%, 86%, 83%, 80% and 87%.
[21]	Back propagation neural networks, general regression neural networks	Service capacity planning	BPN: 80.00%, GRNN: 87.14%
[22]	Data science, data visualization, data mining, machine learning	Resort hotels (three hotels dataset)	Accuracy: >90%

III. METHODOLOGY

The Cross-Industry Standard Process for Data Mining (CRISP-DM) approach, a widely used framework for carrying out data mining initiatives, will be used in our trials. Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment are the six primary steps of the CRISP-DM process [23]. This process is visually represented in Fig. 1.

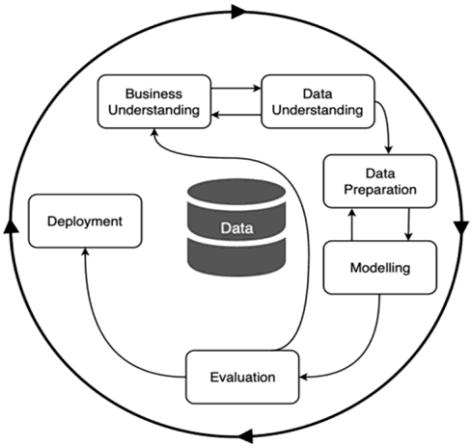

Fig. 1. CRISP-DM Method [24]

In the comprehensive exploration of predicting client ticket sales for British Airways based on internet behaviour, our study meticulously adheres to the CRISP-DM methodology. Beginning with a profound understanding of the business context, our objectives, participants, and success criteria were meticulously defined. The subsequent phases involved scrupulous data collection and examination, leveraging a dataset graciously provided by British Airways, undergoing meticulous cleaning, and preprocessing to address outliers and character variables. Feature engineering strategies were employed to extract pertinent characteristics. Informed by their proven efficacy in related domains and endorsed by British Airways, our choice of machine learning algorithms Random Forest and Decision Tree was judicious. The subsequent modelling phase, executed in RStudio using specialized libraries, will be followed by a rigorous evaluation employing diverse metrics like F1 score, recall, accuracy, and precision. The culmination involves meticulous documentation and finalization of the chosen model's hyperparameters, ensuring a robust foundation for predictive analytics tailored to British Airways' needs.

IV. EXPLORATORY DATA ANALYSIS AND DATA PRE-PROCESSING

A. Dataset Description

The dataset under scrutiny comprises 14 columns and 50,000 rows, encapsulating diverse facets of flight ticket information. Variables such as “num_passengers” delineate passenger counts, while “sales_channel” categorizes the reservation method, distinguishing between online and phone bookings. It captures the nature of travel through the “trip_type” column, indicating round-trip, one-way, or circular journeys. Crucial temporal aspects include “purchase_lead” denoting the days between booking and trip, “length_of_stay” revealing the duration at the destination, and “flight_hour” signifying the departure time. Day-specific details are catalogued in “flight_day,” alongside flight origin, destination in the “route” column, and booking origin in “booking_origin.” Consumer preferences are embedded in columns like “wants_preferred_seat” for specific seat choices, “wants_extra_baggage” indicating additional luggage requests, and “wants_in_flight_meals” reflecting preferences for in-flight meals. Moreover, a thorough examination of Fig.

2. reveals a dataset devoid of empty columns or missing values, substantiating its completeness and robustness.

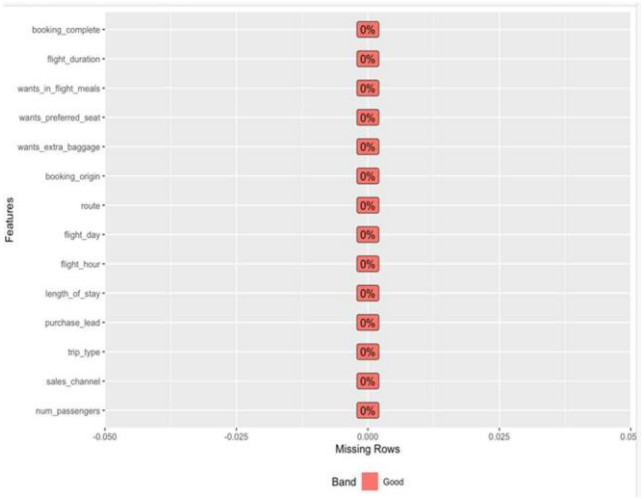

Fig. 2. Variables and Missing Values

B. Feature Selection

Feature subset selection represents a critical step in refining the dataset for enhanced efficiency in learning algorithms. This process involves the identification and elimination of superfluous or redundant features, ultimately diminishing the dataset dimensionality [25]. The primary objective is to bolster the efficacy and performance of learning algorithms by streamlining the input variables [26]. In adherence to this principle, columns 8 and 9 have been deliberately excluded from the dataset, a decision underlined by their perceived lack of relevance to the prediction task, as asserted by [13]. This strategic reduction in dimensionality is expected to contribute to the optimization of algorithmic operations, fostering a more streamlined and effective analytical process.

C. Encoding categorical Variables

In the pursuit of constructing robust prediction models, the transformation of categorical attributes into a numerical format is imperative, allowing for seamless integration into analytical processes. One widely adopted technique for this purpose is one-hot encoding, a method extensively recognized in diverse fields, as emphasized by [26]. This technique was applied meticulously to the dataset, where the "sales_channel" variable transformed into a binary representation, assigning a value of 1 for "Internet" as the sales channel and 0 otherwise. Similar encoding methodologies were employed for the "trip_type" variable, translating the categories into numeric values for enhanced interpretability. The resulting numerical representations, derived through strategic encoding, serve to optimize the utilization of categorical variables in subsequent modelling and analytical endeavours, thereby enhancing the overall efficacy of predictive models [27].

D. Outliers

The dataset under investigation in Fig. 3. reveals the presence of extreme values in the "purchase_lead" and "length_of_stay" columns, potentially exerting a substantial impact on the analytical outcomes. To address this, a recognized strategy from previous studies [26], [28], [29] involves the implementation of Winsorization. This technique systematically replaces values exceeding the 95th percentile with the value at that percentile, mitigating the influence of outliers. A meticulous comparison has been conducted between the outcomes below the 5th percentile and those below this threshold, ensuring a comprehensive examination of the data robustness and the effectiveness of the applied Winsorization method.

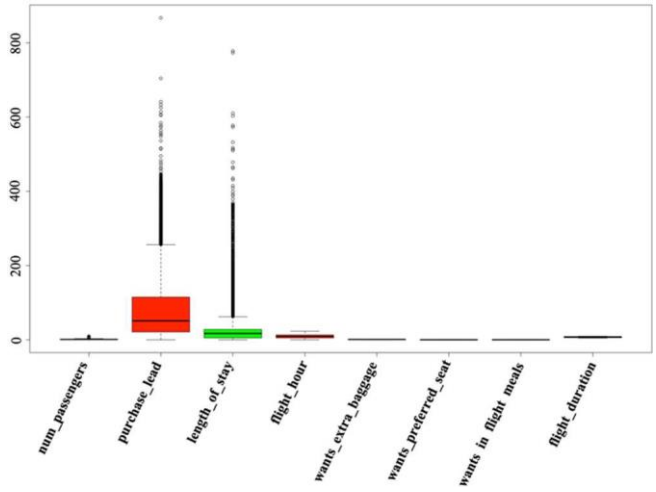

Fig. 3. Outliers Detection

E. Normalization

In the pursuit of enhancing the efficacy of machine learning algorithms, the process of data normalization is deemed indispensable. This pivotal step, exemplified in our study, specifically targets continuous features, aiming to mitigate issues arising from disparate value ranges. Recognizing the potential impediment, a wide range of values can pose to algorithmic performance, we employed Min-Max Scaling to normalize numerical attributes, including "num_passengers", "purchase_lead", "length_of_stay", "flight_hour", and "flight_duration". The utilization of Min-Max Scaling, as advocated by [30], [31], ensures a standardized range, typically between 0 and 1, fostering comparability and coherence among variables. This normalization technique acts as a catalyst for more efficient and meaningful data analysis within the realm of machine learning applications.

F. Correlation Matrix

Examining the relationships within the dataset reveals insightful associations between key variables [32]. Notably, Fig. 4. illustrates a nuanced pattern in the lead time between booking and departure concerning the number of passengers. A positive correlation of 0.22 between "num_passengers" and "purchase_lead" indicates a tendency for lead times to extend as the passenger count increases. Conversely, a negative correlation of -0.15 between "length_of_stay" and "num_passengers" implies that shorter stays align with a reduced number of passengers. Furthermore, notable connections emerge, with "wants_extra_baggage" exhibiting a reasonably strong correlation with both "wants_preferred_seat" and "wants_in_flight_meals" at 0.21 and 0.22, respectively. Lastly, a subtle positive association of 0.08 between "flight_duration" and "purchase_lead" suggests a potential link between extended lead periods and prolonged flight durations. These findings contribute to a nuanced understanding of intervariable dynamics within the dataset, guiding subsequent analytical and predictive endeavours.

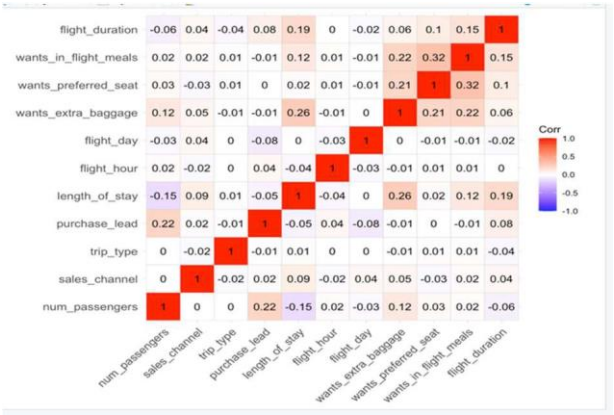

Fig. 4. Outliers Detection

G. Data Balancing

Fig. 5. illustrates a significant class imbalance in the dataset, particularly emphasizing a disproportion in the number of instances where customers completed the booking compared to those who did not. Acknowledging the impact of imbalanced class distributions on model training, we strategically employed oversampling techniques. This corrective measure successfully balanced the class labels, as evidenced by Fig. 5. The transformation achieved through oversampling ensures a more equitable representation of both classes, mitigating potential biases that could affect the predictive performance of the models [33]. This meticulous attention to data balance fortifies the reliability and generalizability of the subsequent predictive models, aligning with best practices in machine learning [34], [35],[36].

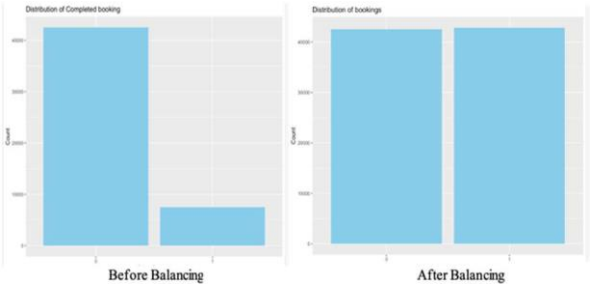

Fig. 5. Class Balancing with Oversampling Method

H. Data Splitting

Before proceeding with the implementation of predictive models, a crucial step involves the stratified division of the dataset into a training set and a testing set, maintaining a 70:30 ratio. This division ensures a robust evaluation of model performance. To foster result reproducibility, a predefined seed is established. Thirty per cent of the dataset is allocated to the testing set, selected randomly using a designated index, while the remaining rows constitute the training set as shown in Fig. 6. This meticulous partitioning strategy enables the training of machine learning models on one subset and the evaluation of their performance on an independent subset, fostering a rigorous assessment of predictive capabilities [37].

Fig. 6. Data Splinting Ratio [38]

V. Model Implementation and Validation

A. Model Implementation

1) Random Forest: a robust ensemble learning method, has been harnessed for the task of predicting and categorizing customer ticket sales for British Airways based on online behaviour. This model, characterized by the construction of 500 decision trees, operates by training each tree on a random subset of the dataset. During predictions, the results from these individual trees are amalgamated through a voting mechanism, endowing the model with resilience against overfitting and enhancing its adaptability to unseen data [39]. The comprehensive results of the Random Forest model are detailed in Table II, where performance metrics on training data are presented.

TABLE II. Random Forest Results

	Test Data	Train Data
Accuracy	83%	89%
Precsion	89%	94%
Recall	75%	84%
F1 Score	82%	89%

These Table II performance metrics illuminate the efficacy of the Random Forest model. The high accuracy values signify the model's proficiency in correctly predicting customer booking outcomes. Precision, measuring the accuracy of positive predictions, is notably high, indicating a low rate of false positives. Moreover, the recall metrics underscore the model's ability to identify actual positive instances, striking a balance between precision and recall. The F1 Score, a harmonized measure of accuracy and recall, further reinforces the model's robustness in predicting customer ticket purchases for British Airways [19].

2) Decision Tree Model: a fundamental component of machine learning, was employed to predict and categorize customer ticket sales for British Airways based on online behaviour. The model, designed to capture decision patterns in a tree-like structure, utilizes features such as flight details, passenger preferences, and purchase lead time to make predictions. The implementation of a single Decision Tree is chosen for its interpretability, enabling a clear understanding of the decision-making process [39]. The results of the Decision Tree Model are summarized in Table III.

TABLE III. Decision Tree Results

	Test Data	Train Data
Accuracy	62.2%	62.4%
Precision	61.0%	61.1%
Recall	67.6%	67.5%
F1 Score	64.2%	64.1%

The Table III metrics comprehensively assess the model's performance, indicating its ability to correctly identify positive instances (precision), capture all positive occurrences (recall), and provide a balanced measure (F1 score) that considers both precision and recall. Despite its simplicity, the Decision Tree demonstrates effectiveness in predicting

customer booking behaviour, contributing valuable insights for British Airways' strategic decision-making.

B. *Model Evaluation through Cross-Validation and Hyperparameter Tuning*

Cross-validation, a critical facet of model evaluation, stands as a pivotal measure in fortifying robustness by scrutinizing performance across diverse subsets of the dataset. Both the Random Forest and Decision Tree models underwent meticulous cross-validation to affirm their consistency and gauge their generalization prowess. To optimize model performance, a judicious exploration of hyperparameters was conducted through grid search and cross-validated tuning for both models [40]. This exhaustive evaluation process not only substantiates the model's efficacy but also hones its precision [41], culminating in heightened reliability for predicting customer ticket sales for British Airways.

1) Random Forest: The Random Forest model underwent meticulous refinement involving cross-validation and hyperparameter adjustment. This process is aimed at optimizing the "mtry" parameter, influencing the number of variables considered at each split, and is crucial for enhancing model performance [42]. Following cross-validation, the adjusted Random Forest model achieved an impressive accuracy of 93% as shown in Table IV showcasing its enhanced predictive capabilities. The confusion matrix in Fig. 8. provided a detailed breakdown of true positives, true negatives, false positives, and false negatives, forming the basis for various performance metrics. Fig. 7. outlines sample sizes, the number of predictors, and the performance evaluation across different "mtry" values. Notably, the model with the "mtry" value of 6 emerged as the optimal choice. This rigorous evaluation and fine-tuning process exemplify the commitment to delivering a highly accurate and reliable model for predicting customer ticket sales for British Airways.

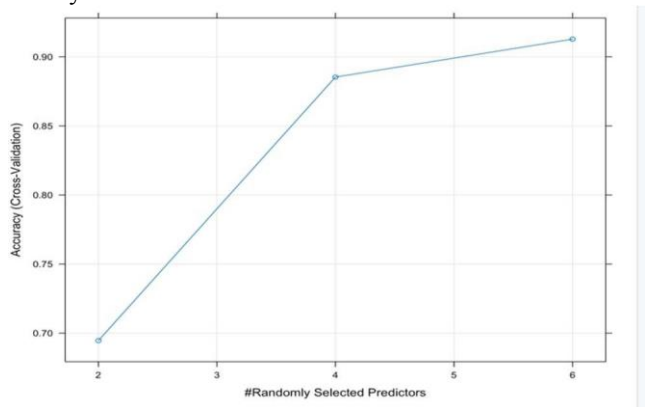

Fig. 7. Hyperparameter Tuning for Random Forest

Fig. 8. Confusion matrix

- **True Negatives (TN):** 11257 instances correctly predicted as customers who will not complete the booking.
- **False Positives (FP):** 218 instances wrongly predicted as customers who will complete the booking when they will not.
- **False Negatives (FN):** 1553 instances were wrongly predicted as customers who will not complete the booking when they will.
- **True Positives (TP):** 12578 instances correctly predicted as customers who will complete the booking.

TABLE IV. RANDOM FOREST ACCURACY AFTER CROSS-VALIDATION

RF Model	Test Data
Accuracy	93%

2) Decision Tree: In the process of optimizing the Decision Tree model, a crucial step involved employing cross-validation to fine-tune hyperparameters, specifically the Fig. 9. complexity parameter (cp). Through systematic experimentation with different values of cp, the optimal value of 0.001 was identified, indicating the minimal improvement required to split a node [43]. This hyperparameter adjustment significantly influenced the model's complexity. Subsequently, the Decision Tree model, with its refined parameters, was applied to the test dataset, and its performance was rigorously evaluated using a confusion matrix. The test accuracy, computed as 63.5%, is highlighted in Table V, demonstrating the model's ability to correctly predict outcomes.

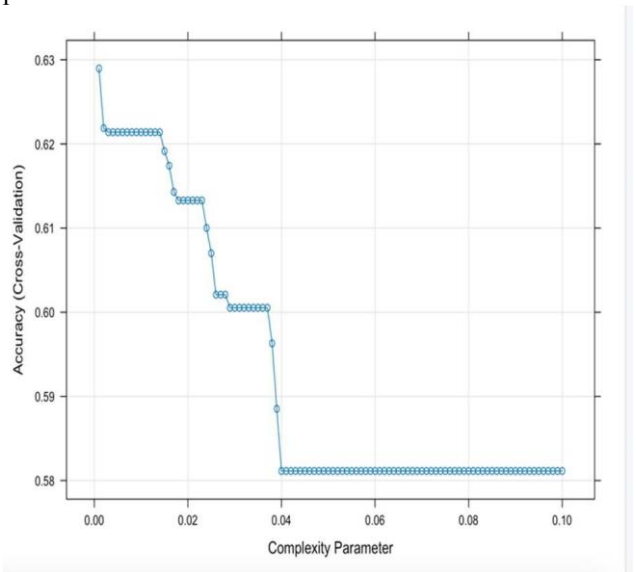

Fig. 9. Decision Tree Model Optimization and Evaluation

TABLE V. DECISION TREE ACCURACY AFTER CROSS-VALIDATION

Decision Tree	Test Data
Accuracy	63.5%

VI. Recommendations Based on Variable Importance

Identifying key factors influencing customer booking is crucial for model refinement. Flight details, customer preferences, and purchase lead time are pivotal variables. Fig. 10. visually depicts these important factors, guiding recommendations for optimizing predictive accuracy. Focusing on insights from these variables aims to enhance the model's effectiveness and inform strategic decisions for British Airways.

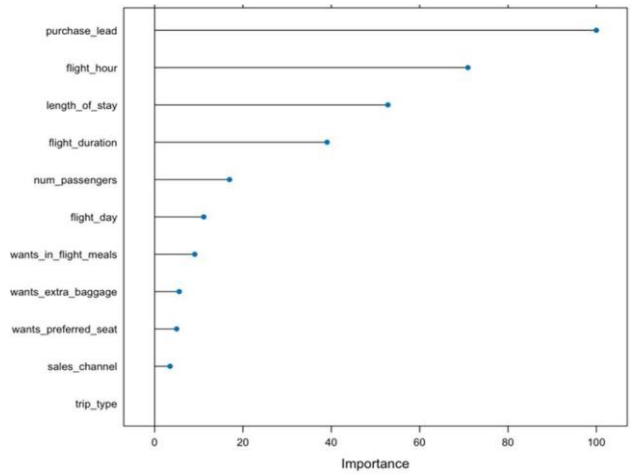

Fig. 10. Important Variables that Influence Customer Booking

Strategic management of the "Purchase_lead" parameter, representing the interval between initial client interaction and purchase, is imperative for British Airways. Implementation of targeted email marketing campaigns, featuring exclusive offers and time-limited incentives, can instill a sense of urgency and foster timely reservations. Disseminating pertinent information on flight updates further enhances customer engagement, shaping a proactive approach to bookings.

Recognizing the critical role of "Flight_hour" British Airways should meticulously tailor flight schedules to align with customer preferences throughout the day. Offering a diverse array of flight options based on peak periods and popular routes demonstrates adaptability to varying customer demands. Analyzing historical data and discerning customer preferences enables the airline to optimize schedules effectively, potentially augmenting reservations and overall customer satisfaction.

Addressing the "Length of stay" variable underscores the importance of catering to customers on extended vacations. British Airways can introduce bespoke packages inclusive of incentives such as affordable lodging and transportation, achieved through collaborations with hotels, tourist destinations, or local service providers. Strategic promotion of these packages through targeted marketing initiatives has the potential to amplify bookings, establishing British Airways as a comprehensive travel solution. Simultaneously, prioritizing the "flight duration" variable by enhancing the in-flight experience, encompassing comfortable seating, onboard entertainment, quality meals, and attentive service, effectively addresses concerns related to extended travel times. Emphasizing these aspects in marketing materials, customer testimonials, and across social media channels positions British Airways as a preferred choice for passengers seeking a gratifying travel experience.

VII. Conclusion

In conclusion, this study marks a significant stride in the realm of predicting customer ticket sales for British Airways. Leveraging sophisticated data mining techniques and employing a unique dataset comprising 50,000 observations and 12 features, our analysis culminated in the implementation and evaluation of two formidable machine learning models: Random Forest and Decision Tree. The outcomes stand as a testament to the efficacy of our approach, with the Random Forest model exhibiting an outstanding accuracy of 93%. This achievement positions our study at the forefront when compared to related works, surpassing existing models, whether utilizing similar datasets or different ones. Our commitment to rigorous evaluation, including cross-validation and hyperparameter tuning, further solidifies the robustness of our predictive models.

The implications of our findings extend beyond the scope of this study, providing British Airways with actionable insights to enhance customer engagement, optimize scheduling, and refine marketing strategies. This study not only contributes to the existing body of knowledge in data-driven decision-making for airlines but also establishes a benchmark for future endeavours in the dynamic intersection of data science and aviation.

References

[1] Jackson, S., & Tozer, J. (2020). A vision for data science at British Airways. *Impact*, *2020*(1), 15-19.

[2] Wei, W., & Rana, M. E. (2019). Software Project Schedule Management Using Machine Learning & Data Mining. International Journal of Scientific & Technology Research, 8(9), 1385-1389.

[3] Hameed, V. A., Rana, M. E., & Enn, L. H. (2023, July). Apriori Algorithm based Association Rule Mining to Enhance Small-Scale Retailer Sales. In *2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI)* (pp. 187-191). IEEE.

[4] Park, E., Jang, Y., Kim, J., Jeong, N. J., Bae, K., & Del Pobil, A. P. (2019). Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. *Journal of Retailing and Consumer Services*, *51*, 186-190.

[5] Singh, R., & Khan, I. A. (2012). An approach to increase customer retention and loyalty in B2C world. *International journal of scientific and research publications*, *2*(6), 1-5.

[6] (N.d.). Britishairways.com. Retrieved November 9, 2023, from https://www.britishairways.com/content/information/about-ba/history-and-heritage/explore-our-past

[7] Hussain, R. (2016). The mediating role of customer satisfaction: evidence from the airline industry. *Asia Pacific Journal of Marketing and Logistics*, *28*(2).

[8] Shen, Z. (2022, December). Analysis of British Airway's Hedging Strategies. In *2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022)*(pp. 307-311). Atlantis Press.

[9] Kumar, S., & Anuj Thapliyal, S. (2022). How can British Airways better utilize management information systems to address its organizational weaknesses?. *Central European Management Journal*, *30*(3), 387-391.

[10] Kevin Kibe. (2023, January, 25). *Classification Model to Predict Customer Booking with an Airline.* https://keviinkibe.medium.com/predicting-customer-booking-with-british-airways-using-machine-learning-d5045fa07836

[11] RaffelRavionaldo. (2022, November 22). *British-Airways-Virtual-Internship* github https://github.com/RaffelRavionaldo/British-Airways-Virtual-Internship

[12] nugi.info. (2023, May 7). Airline Customer Booking Prediction (British Airways Virtual Experience).https://nugi.info/?p=1

[13] ImperfectSensei. (2022, November 14). *Data Science Intern at British Airways!* [Video]. https://www.youtube.com/watch?v=p7b8IRHMZts

[14] *British Airways Customer Booking prediction*. (2023, June 15). Kaggle.com; Kaggle.

https://www.kaggle.com/code/seunayegboyin/british-airways-customer-booking-prediction

[15] Hamdan, I. Z. P., & Othman, M. (2022). Predicting Customer Loyalty Using Machine Learning for Hotel Industry. *Journal of Soft Computing and Data Mining*, *3*(2), 31-42.

[16] Singh, B. (2021). Predicting airline passengers' loyalty using artificial neural network theory. *Journal of air transport management*, *94*, 102080.

[17] Park, S. H., Kim, M. Y., Kim, Y. J., & Park, Y. H. (2022). A deep learning approach to analyze airline customer propensities: the case of South Korea. *Applied Sciences*, *12*(4), 1916.

[18] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, 1-24.

[19] Chou, Y. C., & Chuang, H. H. C. (2018). A predictive investigation of first-time customer retention in online reservation services. *Service Business*, *12*, 685-699.

[20] GANGADHARAN, C. K., ALEX, R., & SABU, M. (2021). CHURN PREDICTION-A COMPARATIVE ANALYSIS WITH SUPERVISED MACHINE LEARNING ALGORITHMS.

[21] Huang, H. C., Chang, A. Y., & Ho, C. C. (2013). Using artificial neural networks to establish a customer-cancellation prediction model. *Przeglad Elektrotechniczny*, *89*(1b), 178-180.

[22] Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25-39.

[23] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.

[24] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., ... & Flach, P. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048-3061.

[25] Ponnusamy, R. R. A., Rana, M. E., Manickavasagam, S. A., & Hameed, V. A. (2023, July). PSO-SVM Based Algorithm for Customer Churn Prediction in the Banking Industry. In 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI) (pp. 220-225). IEEE.

[26] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1-45.

[27] Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Frontiers in Energy Research, 9,* 652801.

[28] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, *175*(4), 7-9.

[29] Lista, L. (2017). Statistical methods for data analysis in particle physics (Vol. 941). New York: Springer.

[30] Hamadani, A., Ganai, N. A., Raja, T., Alam, S., Andrabi, S. M., Hussain, I., & Ahmad, H. A. (2021). Outlier removal in sheep farm datasets using winsorization. *Bhartiya Krishi Anusandhan Patrika*, *36*(4), 334-337.

[31] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111-117.

[32] Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020, December). Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry. In 2020 23rd International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[33] Wagavkar, S. (2023, March 17). *Introduction to the correlation matrix*. Built In. https://builtin.com/data-science/correlation-matrix

[34] Ismail, L., & Materwala, H. (2022, August). From Conception to Deployment: Intelligent Stroke Prediction Framework using Machine Learning and Performance Evaluation. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)* (pp. 1-7). IEEE.

[35] Branco, P., Torgo, L., & Ribeiro, R. P. (2019). Pre-processing approaches for imbalanced distributions in regression. *Neurocomputing*, *343*, 76–99. https://doi.org/10.1016/j.neucom.2018.11.100

[36] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0192-5

[37] Reitermanova, Z. (2010, June). Data splitting. In *WDS* (Vol. 10, pp. 31-36). Prague: Matfyzpress.

[38] Sydorenko, I. (2021, March 17). *Machine learning and training data: What you need to know*. Labelyourdata.com; Label Your Data. https://labelyourdata.com/articles/machine-learning-and-training-data

[39] Afrianto, M. A., & Wasesa, M. (2020). Booking prediction models for peer-to-peer accommodation listings using logistics regression, decision tree, K-nearest neighbor, and random Forest classifiers. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 123-32.

[40] Al-Abdaly, N. M., Al-Taai, S. R., Imran, H., & Ibrahim, M. (2021). Development of prediction model of steel fiber-reinforced concrete compressive strength using random forest algorithm combined with hyperparameter tuning and k-fold cross-validation. *Eastern-European Journal of Enterprise Technologies*, *5*(7), 113.

[41] Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. *arXiv preprint arXiv:1803.11266*.

[42] Soper, D. S. (2021). Greed is good: Rapid hyperparameter optimization and model selection using greedy k-fold cross validation. *Electronics*, *10*(16), 1973.

[43] Mantovani, R. G., Horváth, T., Cerri, R., Junior, S. B., Vanschoren, J., & de Carvalho, A. C. P. D. L. F. (2018). An empirical study on hyperparameter tuning of decision trees. *arXiv preprint arXiv:1812.02207*.