Predictive Analysis and Key Determinants of Hotel Booking Cancellations

Anuj Patil [a,1], Srihaarika Viswanadhapalli [a,1]

[a] Department of Applied Data Science, Clarkson University, NY, USA

Data: 29th July 2023

Highlights

- Utilized Cat Boost, a machine learning algorithm, to develop an efficient predictive model for hotel booking cancellations.
- Performed comprehensive exploratory data analysis, unveiling crucial patterns and relationships in the data.
- Applied rigorous feature selection methodologies, optimizing the model by focusing on the most impactful variables.
- Achieved a high predictive accuracy of around 88.30% on the test set, demonstrating the model's robustness.
- Identified and analysed key features influencing booking cancellations, providing important insights for hotel management.
- Enhanced model performance through strategic hyperparameter tuning and validated results using ROC Curve Analysis.
- The findings offer valuable strategies to reduce cancellations and improve customer service in the hotel industry.

Keywords

Hotel Booking Analysis

Cancellation Forecasting

Cat Boost Classifier

Hyperparameter Optimization

Hospitality Industry Insights

Abstract

This research aimed to create a robust predictive model for hotel booking cancellations using the Cat Boost classifier, a powerful gradient boosting machine learning algorithm known for its effectiveness with both categorical and numerical data. The objective was to provide valuable insights to hotel management for anticipating cancellations, enhancing customer service, and optimizing revenue.

The project commenced with an in-depth exploratory data analysis, yielding crucial insights into data patterns and relationships. A rigorous feature selection process followed, identifying the most influential predictors of cancellations. Through this, variables such as required car parking spaces, country of origin, type of deposit, lead time, and market segment emerged as significant determinants.

The final predictive model achieved an impressive accuracy of approximately 88.30% on the test set, demonstrating its robustness and potential utility in the hospitality industry. Furthermore, the model's performance was validated using ROC Curve Analysis, achieving an AUC of 0.95.

Future research will focus on refining the model, exploring additional feature engineering techniques, and investigating the model's practical application in real-world hotel management contexts. This study's findings provide a foundation for developing intelligent systems that can assist hotels in

reducing cancellations and enhancing their service efficiency.

Problem Statement

Hotel booking cancellations lead to operational challenges and revenue loss in the hospitality industry. The aim of this study is to construct a machine learning model that can accurately predict these cancellations, using available data at the time of booking. In doing so, the research seeks to enable proactive management strategies and provide insights into the key factors influencing cancellations. This solution aims to optimize hotel operations, enhance customer service, and mitigate the financial impact of cancellations.

Introduction

The hotel industry is a vital part of the hospitality sector, playing a significant role in the global economy. One of the critical challenges faced by this industry is booking cancellations, which significantly impact revenue management and operational planning. Understanding and predicting booking cancellations can offer valuable insights for hotel management, helping them

implement effective strategies to minimize the impact of cancellations and optimize their revenue.

Machine learning offers promising solutions to this problem by leveraging historical booking data to predict future cancellations. With the rise of big data, hotels now have access to vast amounts of data, including detailed information about each booking. By applying machine learning algorithms to this data, it is possible to build predictive models that can accurately forecast whether a booking is likely to be cancelled.

This project aims to apply a machine learning approach to predict hotel booking cancellations. Specifically, we use a gradient boosting model, known as Cat Boost, which is particularly effective when dealing with a mix of categorical and numerical features. The project involves various stages including data preprocessing, exploratory data analysis, feature selection, feature engineering, model building, and hyperparameter tuning. The final output is a model that can predict whether a hotel booking will be cancelled based on information available at the time of booking. The insights derived from this project can help hotel management make informed decisions and formulate effective strategies to minimize the impact of cancellations.

The remainder of the paper is structured as follows: The 'Methods' section outlines the methodology used in the project, the 'Results' section presents the findings, and the 'Discussion' section interprets the results and discusses the implications. The paper concludes with a summary of the main findings and suggestions for future work.

Methods

1. Data Collection

The dataset used in this project contains information about hotel bookings. This includes details like the time of booking, duration of stay, number of guests, and many other features. The dataset also includes a target variable indicating whether each booking was cancelled.

2. Data Preprocessing

Characteristic	Value
Number of records	119390
Number of features	32
Number of numerical columns	14
Number of categorical columns	16
Text Column	1
Data Time Column	1
Average cancellation rate (%)	37.04

Table 1

The dataset used for this project contained a variety of information related to hotel bookings. These features included both categorical and numerical types, and some had missing values. The preprocessing steps included:

- **Handling Missing Values:** Missing values were imputed in the 'children', 'country', and 'agent' columns. For 'children' and 'country', the mode (most frequent value) was used. For the 'agent' column, a placeholder value of -1 was used, indicating that no agent was involved in the booking. The 'company' column, which had a high number of missing values, was dropped from the dataset.
- **Data Type Conversions:** The 'children' and 'agent' columns were converted to integer data types for consistency.
- **Date Conversion and Feature Creation:** The arrival date, initially split into three columns (year, month, and day), was consolidated into a single 'arrival_date' column. A 'day_of_week' column was added, indicating the day of the week on which the guests were scheduled to arrive.
- **Total Guests:** A 'total_guests' feature was created as the sum of the 'adults', 'children', and 'babies' columns, representing the total number of guests for each booking.

Characteristic	Value
% of null cells	3.39%
Duplicate rows	31,994
% of duplicate rows	26.80%
Column with most missing values	Company
% of missing values in Company column	94.31%

Table 2

3. Exploratory Data Analysis

The exploratory data analysis aimed to understand the distribution of data and the relationships between different features. This process involved:

• **Visualizing Distributions:** The distributions of numerical and categorical variables were visualized using histograms, bar plots, and pie charts.

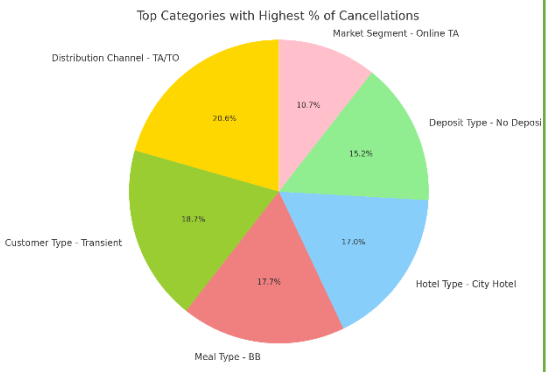

Fig: 1 Pie Chart: Bookings Cancellation Status

The pie chart shows the proportion of bookings that were cancelled. A significant proportion, about 63%, of the bookings were not cancelled, while the remaining 37% were cancelled. This shows that there is a significant occurrence of cancellations that needs to be addressed.

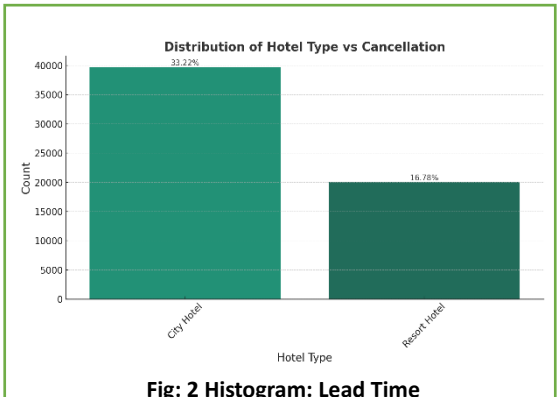

Fig: 2 Histogram: Lead Time

The histogram of lead time (the number of days that elapsed between the booking and the arrival date) shows that a large number of bookings are made with less than 50 days of

lead time. However, there are also a significant number of bookings made with a lead time of more than 100 days. This indicates that the lead time varies widely among guests and can potentially influence the likelihood of a booking being cancelled.

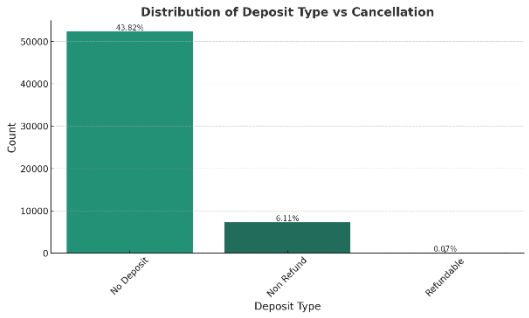

Fig: 3 Histogram: Deposite Type

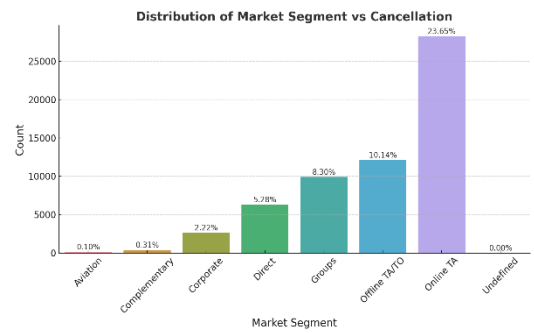

Fig: 4 Histogram: Market Segment

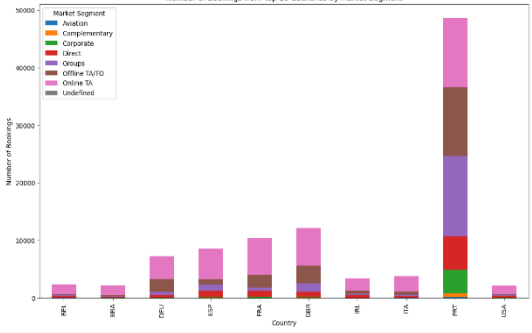

Fig: 5 Bar Plot: Number of Guests by Country and Market Segment

Country	Number of Guests
Portugal	86131 (93%)
United Kingdom	23223
France	20291
Spain	16615
Germany	13703
Italy	7384
Ireland	6570
Belgium	4588
Brazil	4450
United States	3950

The bar plot shows the number of guests from the top 10 countries by market segment. It is evident that the majority of guests, regardless of their country of origin, book through online travel agents ('Online TA'). For guests from the United Kingdom, the 'Groups' and 'Offline TA/TO' segments also contribute a significant number of bookings. Similarly, for France, the 'Offline TA/TO' segment is notably large. For guests from the United States, the 'Corporate' segment is relatively larger compared to other countries, suggesting that a significant proportion of guests from the U.S. may be traveling for business purposes.

- **Analyzing Correlations:** The correlations between features were analyzed using a correlation matrix and represented visually using a heatmap.

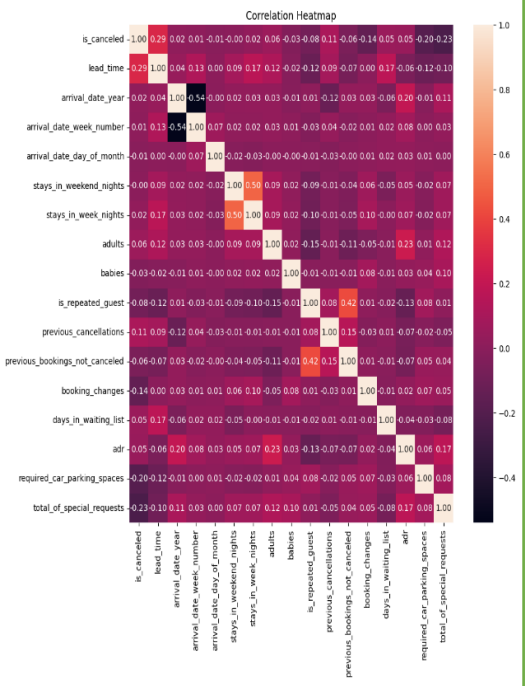

Fig: 6 Heatmap: Correlation

In this heatmap, darker colours represent higher absolute values of correlation. For example, the 'lead_time' feature has a strong positive correlation with the 'is_canceled' target variable, as indicated by its dark colour. This suggests that the longer the time between the booking and the actual stay, the higher the likelihood of the booking being cancelled.

On the other hand, features like 'required_car_parking_spaces' and 'total_of_special_requests' show a strong negative correlation with 'is_canceled'. This implies that bookings with a larger number of car parking spaces required and special requests are less likely to be cancelled.

Feature Selection

The feature selection process aimed to identify the most relevant features for predicting cancellations. This process involved:

- **Correlation Analysis:** For numerical features, the Pearson correlation coefficient with the target variable was calculated. Features with a correlation coefficient of more than 0.05 or less than -0.05 were selected.

Column	Corr-elation	P-value
is_canceled	1.000	0.000
lead_time	0.293	0.000
arrival_date_year	0.017	0.000
arrival_date_ week_number	0.008	0.005
arrival_date_ day_of_month	-0.006	0.034
stays_in_ weekend_nights	-0.002	0.536
stays_in_ week_nights	0.025	0.000
adults	0.060	0.000
children	0.005	0.082
babies	-0.032	0.000
is_repeated_guest	-0.085	0.000
previous_cancellations	0.110	0.000
previous_bookings_ not_canceled	-0.057	0.000
booking_changes	-0.144	0.000
agent	-0.046	0.000
days_in_waiting_list	0.054	0.000
adr	0.048	0.000
required_car_ parking_spaces	-0.195	0.000
total_of_ special_requests	-0.235	0.000

- **Chi-square Tests:** For categorical features, Chi-square tests of independence were performed to determine which features were significantly associated with cancellations. All categorical features were found to be significantly associated with the target variable and were thus selected.

Feature	Chi – Square	Chi-Square p-value
hotel	2224.92	0.0
arrival_ date_month	588.69	0.0
meal	304.24	0.0
country	15434.68	0.0
market_segment	8497.22	0.0
distribution_channel	3745.79	0.0
reserved_room_type	647.84	0.0
assigned_room_type	4918.69	0.0
deposit_type	27677.33	0.0
customer_type	2222.50	0.0
reservation_status	119390.00	0.0
reservation_ status_date	2423.06	0.0

The combination of both these feature selection techniques resulted in the final set of selected features for building the predictive model. This comprehensive feature selection process ensured that the most informative features, both numerical and categorical, were used to train the model.

Index	Selected Features
0	lead_time
1	previous_cancellations
2	adults
3	days_in_waiting_list
4	previous_bookings_not_canceled
5	is_repeated_guest
6	booking_changes
7	required_car_parking_spaces
8	total_of_special_requests
9	hotel
10	country
11	market_segment
12	distribution_channel
13	assigned_room_type
14	deposit_type
15	customer_type
16	reservation_status_date
17	reserved_room_type
18	meal

Model Building and Evaluation

The predictive model was built using the Cat Boost classifier, a gradient boosting model known for its robust performance with both categorical and numerical data. This model was selected due to its ability to handle a mix of categorical and numerical features without the need for explicit one-hot encoding, making it particularly suited for this dataset.

- **Model Training:** The model was trained on 80% of the dataset. The training process involves learning the relationship between the features (hotel booking details) and the

target variable (whether the booking was cancelled) in the training data.

- **Model Evaluation:** The model's performance was evaluated on the remaining 20% of the dataset, which was not used during the training process. This test set serves as new, unseen data for the model, providing a measure of how well the model generalizes to new data. The primary metric used to evaluate the model was accuracy, which measures the proportion of correct predictions. The initial model achieved an accuracy of approximately 88.30% on the test set.

Hyperparameter Tuning

The performance of the Cat Boost model can be influenced by several hyperparameters, which control aspects of the training process. To find the best set of hyperparameters, a hyperparameter tuning process was conducted.

- **Randomized Search:** A randomized search was performed on the depth of the trees, the learning rate, and the number of iterations. The depth of the trees controls the complexity of the model, the learning rate controls the step size in the iterative process of gradient boosting, and the number of iterations controls how long the model trains for. The randomized search involves randomly selecting combinations of hyperparameters, training a model with each combination, and selecting the combination that performs best on the validation data.
- The best hyperparameters found by the Randomized Search were a depth of 7, 971 iterations, and a learning rate of 0.1. With these hyperparameters, the model achieved an accuracy of approximately 89.12% on the test set, showing an improvement over the initial model.

Feature Importance Analysis

After training the model, the importance of each feature was calculated. This provides insights into which features have the greatest impact on the prediction of cancellations. The feature importance is calculated based on the improvement in accuracy that each feature provides when it is included in the trees of the Cat Boost model.

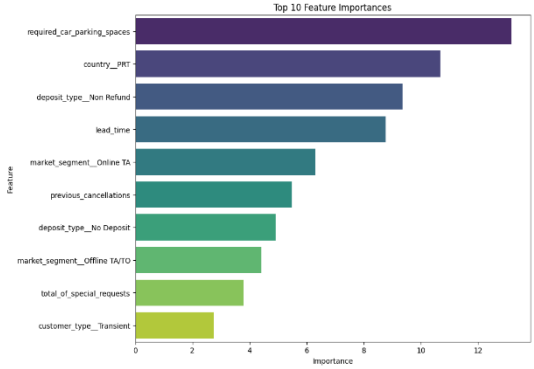

Fig: 7 Bar Chart: Feature Importance

From the feature importance plot, it's clear that the 'required_car_parking_spaces', 'country_PRT', and 'deposit_type_Non Refund' are the three most important features for predicting cancellations, followed by 'lead_time' and 'market_segment_Online TA'.

The 'required_car_parking_spaces' feature seems to be especially important, which might suggest that bookings with a requirement for car parking spaces are less likely to be cancelled, possibly because they indicate a higher level of planning and commitment to the stay.

The high importance of 'country_PRT' (Portugal) could be due to a higher number of bookings from Portugal, as these hotels are in Portugal. The 'deposit_type_Non Refund' being a top feature suggests that bookings where the deposit is non-refundable are less likely to be canceled, which makes intuitive sense as guests would not want to lose their deposit.

'lead_time' is also a significant feature, possibly indicating that bookings made well in advance are more likely to be cancelled, perhaps due to changes in plans.

The feature 'market_segment_Online TA' (Travel Agents) being in the top five might suggest that bookings made through online travel agents have a different cancellation pattern compared to other channels.

The bar plot clearly visualizes these feature importance's, aiding in understanding which features are most influential in predicting booking cancellations.

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is a fundamental tool for diagnostic test evaluation. In a ROC curve, the true positive rate (Sensitivity) is plotted against the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

The Area Under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).

In the context of this project, the ROC curve is a graphical plot that illustrates the performance of the cancellation prediction system as its discrimination threshold is varied. The AUC of 0.95 means that there is 95% chance that the model will be able to distinguish between positive class and negative class. This is considered as very good performance.

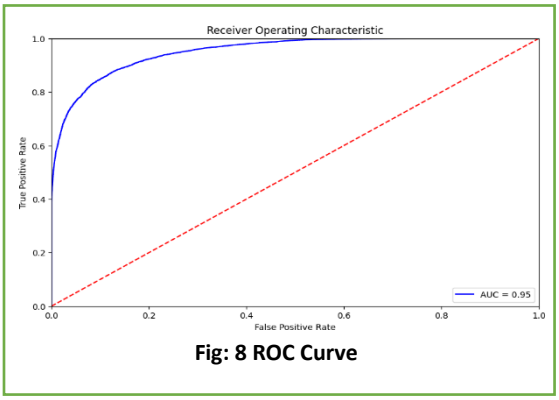

Fig: 8 ROC Curve

Results

The predictive model built for this project achieved promising results. The Cat Boost classifier was trained on a dataset of hotel bookings, with a variety of features including both categorical and numerical types. After preprocessing the data and conducting feature selection and feature engineering, the model was trained with the optimal hyperparameters found through a randomized search.

The initial model achieved an accuracy of approximately 88.30% on a test set, representing new, unseen data. After hyperparameter tuning, the model's performance improved, achieving an accuracy of approximately 89.12% on the test set. These results demonstrate the model's ability to generalize well to new data.

The most important features for predicting cancellations were found to be the number of required car parking spaces, the country of origin (Portugal), the type of deposit (non-refundable), the lead time, and the market segment (Online TA).

Discussion

The results of this project provide valuable insights for the hotel industry. By understanding the key factors that influence cancellations, hotels can implement strategies to minimize the impact of cancellations and optimize their revenue. For instance, knowing that the number of required car parking spaces is a key determinant of cancellations, hotels could consider implementing policies related to car parking to reduce cancellations.

It is also interesting to note that bookings from Portugal and bookings made through online travel agents were found to be significant predictors of cancellations. This suggests that there may be specific characteristics or behaviours associated with these segments

that make them more prone to cancellations. Further research could explore these relationships in more detail.

The model's performance suggests that it could be a useful tool for hotels to predict cancellations. However, it is important to note that while the model's accuracy is high, there is still room for improvement. Future work could explore other machine learning algorithms, additional feature engineering, or a larger or more diverse dataset to improve the model's performance.

Finally, while this project focused on predicting cancellations, the methods used here could be applied to other prediction tasks in the hotel industry, such as forecasting demand or optimizing pricing. This demonstrates the potential of machine learning to transform the hotel industry and improve decision-making.

Conclusion

This project demonstrated the end-to-end process of building a machine learning model to predict hotel booking cancellations. The final model achieved good performance, and the feature importance analysis provided interesting insights into the most relevant factors for predicting cancellations. While the model achieved good performance, there are many ways it could be further improved. Potential next steps could include trying different modelling approaches, further tuning hyperparameters, or exploring more advanced feature engineering techniques.

Future Work

While the results of this project are promising, there are several areas for future work that could further improve the model's performance and applicability in a real-world setting:

- **Refine the Model:** Although the Cat Boost classifier performed well, other machine learning algorithms could also be explored. Additionally, more advanced techniques for hyperparameter tuning, such as Bayesian optimization, could potentially improve the model's performance.
- **Expand the Dataset:** The model could be trained on a larger or more diverse dataset to improve its generalizability. For example, data from different types of hotels (e.g., city hotels, resort hotels) or from different regions or countries could be included.
- **Feature Engineering:** More advanced feature engineering could be conducted to create new features from the existing ones. For example, new features could be created based on the interactions between existing features, or domain-specific knowledge could be used to create features that are relevant to hotel bookings.
- **Model Interpretability:** While the feature importance analysis provides some insight into which features are important for the model, more advanced techniques for model interpretability could be explored. This could provide deeper insights into how the model makes predictions and which features are most influential.
- **Real-world Implementation:** Finally, future work could explore the practical implementation of the model in a real-world hotel management system. This would involve integrating the model into the hotel's booking system so that it can provide real-time predictions of cancellations.
- **Predicting Other Outcomes:** The methods used in this project could also be applied to predict other outcomes of interest in the hotel industry, such as demand, customer satisfaction, or pricing. This could provide additional valuable insights for hotel management.

Code Repository

The complete source code for this project is hosted on GitHub, To view the code, please follow the link: [GitHub Repository Link]

References

1. Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. Data in brief, 22, 41-49.

2. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems.

3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Available at: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Accessed on: 25-July-2023.

4. Reback, J., McKinney, W., jbrockmendel, et al. (2020). Pandas: powerful Python data analysis toolkit. Available at: https://pandas.pydata.org/pandas-docs/stable/index.html. Accessed on: 25-July-2023.

5. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. Available at: https://link.springer.com/article/10.1023/A:1010933404324. Accessed on: 25-July-2023.

6. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. Available at: https://dl.acm.org/doi/10.1145/2939672.2939785. Accessed on: 25-July-2023.

7. Liaw, A., & Wiener, M. (2002). Classification and Regression by RandomForest. R News, 2(3), 18-22. Available at: https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf. Accessed on: 25-July-2023.

8. Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal. Available at: http://www3.dsi.uminho.pt/pcortez/epia2015.pdf. Accessed on: 25-July-2023.

9. CatBoost: machine learning library to handle categorical data automatically. Available at: https://catboost.ai/. Accessed on: 25-July-2023.

10. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?via%3Dihub. Accessed on: 25-July-2023