2024

Kartikesh Jadhao

7447587298

Report On Predictive Modeling

To provide a comprehensive Exploratory Data Analysis (EDA) focused on booking trends, customer demographics, and cancellation patterns,

1. Business Understanding

Problem Definition

"Provide actionable insights and recommendations for hotel management based on data analysis relevant to the hospitality industry. Consider aspects such as pricing strategies, customer segmentation, and marketing focus. Highlight methodologies and offer clear recommendations."

Scope

- The scope of this sample is to create a binary classification machine learning model which address the above redaction problem.
- We execute the project in jupyter lab.

Plan

We follow the steps of Agil Methodology. We will build two models. First on Logistic Regression and second Random Forest. The will Evaluate the model Performance. The Models will be of Classification Model.

Team Personnel

The project is executed by one **data scientist** and Git server for version control. Data scientist executes the various data science steps, creates and compares models, and validate the models.

Metrics

Performance of the machine learning models will be evaluated on the test set provided by the client. Accuracy is measured and reported using Accuracy score. IF Accuracy score of > 0.87will be considered acceptable and suitable for deployment.

2. Data Acquisition and Understanding

Raw Data

Data is provided to us in csv format.

There are a total of 119390 instances (prior to any filtering), mix of continuous and discrete (train=95512, test=23878) and 32 features.
TARGET: 'is_canceled'

FEATURES: 'hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
'arrival_date_month', 'arrival_date_week_number',
'arrival_date_day_of_month', 'stays_in_weekend_nights',
'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
'country', 'market_segment', 'distribution_channel',
'is_repeated_guest', 'previous_cancellations',
'previous_bookings_not_canceled', 'reserved_room_type',
'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
'company', 'days_in_waiting_list', 'customer_type', 'adr',
'required_car_parking_spaces', 'total_of_special_requests',
'reservation_status', 'reservation_status_date'

Data Exploration

Data exploration is performed using the Python 3 ,Pandas , Seaborn, Matplotlib

The location of the final data exploration report is here: X_final

Data Analysis

3. Modeling

Feature Engineering

Data cleanup: Removing columns and rows Prior to feature engineering, we removed three columns 'company', 'agent' ,'reservation_status_date'.

We did not remove any rows.

We fill null values in column 'children" as 0 assuming that they have now child with them.

One-hot encoding categorical features

Following categorical features were one-hot encoded using Scikit-learn's preprocessing.OneHotEncoder() function: hotel, arrival_date_month, meal, country, market_segment, distribution_channel, reserved_room_type, assigned_room_type, deposit_type, days_in_waiting_list, customer_type, reservation_status, reservation_status_date. we fit the transformation model on the dataset and then transformed the test set.

Saving processed data sets for modeling input

Training and test data sets were saved as X_train,Y_train for input into modeling (training data), and X_test, Y_test for model evaluation or deployment (test data).

Modeling training

We created two models : Logistic Regression and Random forest. In two different Jupyter Notebook.

Model evaluation

Accuracy of the models were measured using Accuracy Score on the test data set. Accuracy of both Logistic Regression and Random Forest models were > 0.87. We save both models ingit repository .

Version Control Repository

An Git repository is used to version control contents of this project.

Code Execution

In this example, we execute code in local compute environment . run in the Jypyter Notebook Server.

Executing a Python script in a local Python runtime is easy:

Python notebook files can be double-clicked from the project and run in the Jypyter Notebook Server.

Conclusion

Use Predictive Models to Anticipate Cancellations

Insight: Predictive models (e.g., logistic regression) can help identify factors leading to higher cancellation risks, such as long lead times, high ADR, and no-deposit bookings.

Recommendation:

- **Cancellation Risk**: Develop a **predictive cancellation risk score** based on historical data, lead time, ADR, deposit type, and customer behavior. This score can be used to offer cancellation-prone customers incentives to retain their booking, such as partial refunds, free upgrades, or flexible rescheduling.
-
- **Retention Strategies**: Once a high-risk booking is identified, send targeted offers (e.g., discount vouchers, loyalty points) to reduce the likelihood of cancellation.