



View presentation slides for this project [here](#)

Overview

This project aims to develop a predictive model to forecast hotel booking cancellations using a Kaggle [dataset](#) of hotel reservations. The goal is to accurately predict cancellations, helping hotels optimize inventory, staff, and revenue strategies.

The project involves data cleaning, exploratory analysis, feature engineering, model selection, and evaluation using metrics like accuracy, precision, recall, and F1-score. By analyzing the factors contributing to cancellations, the project seeks to provide valuable insights for the hospitality industry, enabling better resource management and improved customer satisfaction.

Business Understanding

The hospitality industry faces significant challenges with booking cancellations, which can lead to lost revenue and lower occupancy rates. With cancellation rates rising to 40%, there's a clear need for a predictive model to forecast these cancellations accurately. Such a model would allow hotels to proactively address potential cancellations, optimize staff and inventory management, and offer targeted promotions to retain bookings. This project aims to provide hotels with a tool to enhance revenue strategies, increase customer satisfaction, and reduce the financial impact of cancellations, giving them a competitive edge.

Data Understanding

To build the predictive model, I used the Hotel Reservations [Dataset](#) from Kaggle, which includes data on customer bookings. Key features include the number of guests, meal plans, parking requirements, room types, lead time, arrival dates, and market segments. The dataset has 36,275 rows and 19 columns, with the target variable being `booking_status` (1 for canceled, 0 for not). I applied **OneHotEncoder** to categorical features and used **StandardScaler** for numerical features to prepare the data for modeling.

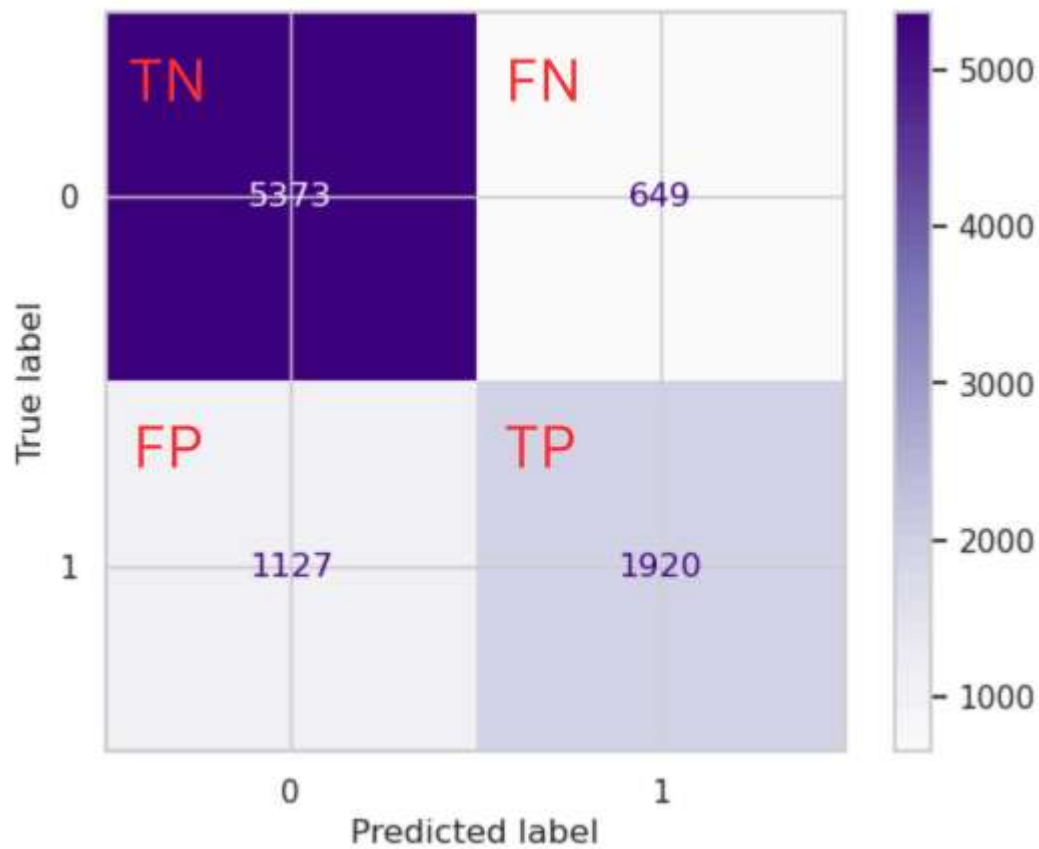
Modeling - Baseline Model

I began by creating a Logistic Regression model using scikit-learn's `LogisticRegression` class. The model was trained on the `x_train_transformed` and `y_train` data. This baseline model estimates the probability of booking cancellations (`booking_status`) based on input features, providing a foundation for comparison. The model's performance metrics are as follows:

- Accuracy of 80.4% - The percentage of correct predictions.
- Precision of 74.74% - The percentage of true positive predictions among all positive predictions.
- Recall of 63% - The percentage of true positive predictions among all actual positive predictions.
- F1 score of 68.4% - The harmonic average of precision and recall.
- ROC AUC score of 0.76 - Reflects the model's ability to distinguish between cancellations and non-cancellations.

These metrics provide a baseline to assess model performance and highlight areas for potential improvement.

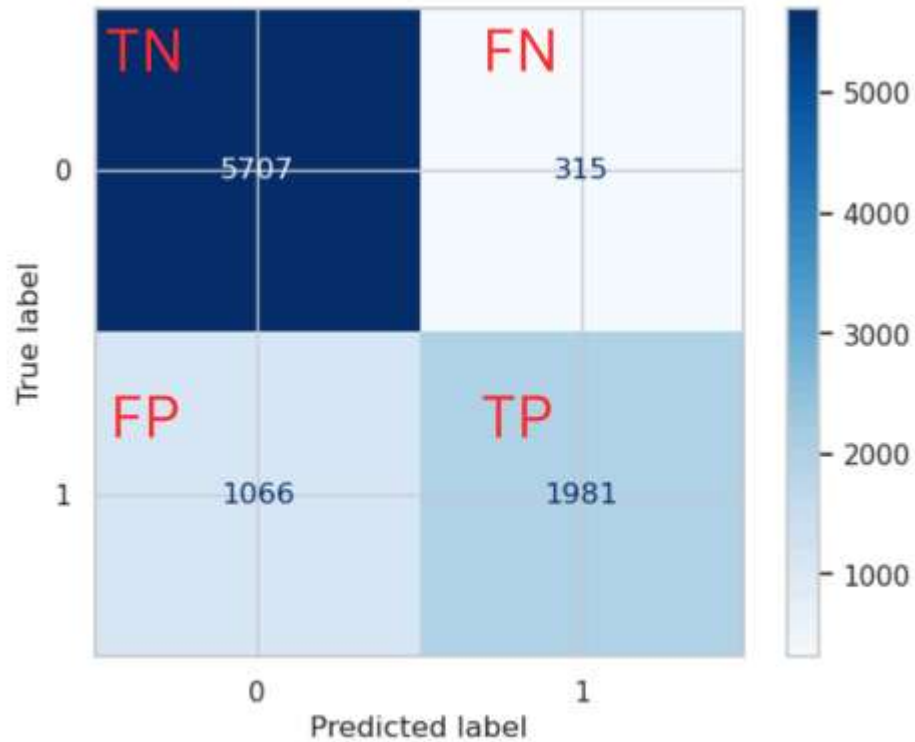
Baseline Model:



Decision Tree Model

Achieved:

- 85% Accuracy
- 86% Precision
- 65% Recall
- 74.2% F1 score
- 0.8 AUC



Random Forest Model

Achieved:

- 90.2% Accuracy
- 88.6% Precision
- 81.4% Recall
- 84.8% F1 score
- 0.88 AUC



Releases

No releases published

Packages

No packages published

Languages

- Jupyter Notebook 100.0%