chriskarta /
project-phase-3

Code Issues Pull requests Actions Projects Wiki Security Insights Setting

0 stars 0 forks 1 watching Branches Activity
Tags

Public repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ···

chriskarta Create .gitignore e9772e6 · 41 minutes ago

.ipynb_checkpoints	notebook finished	1 hour ago
.gitignore	Create .gitignore	41 minutes ago
Hotel-Reservations.csv	notebook finished	1 hour ago
README.md	Create README.md	1 hour ago
notebook.ipynb	notebook update	46 minutes ago
notebook.pdf	notebook finished	1 hour ago

README

Hotel Booking Cancellation Prediction

Overview

This project focuses on creating a predictive model to anticipate hotel booking cancellations using historical data. The model will be trained on a dataset sourced from Kaggle, which includes a vast collection of customer hotel reservations. The main goal is to design a machine learning system capable of reliably estimating the probability of a booking being canceled. Such predictions can assist hotels in optimizing inventory control, staff allocation, and revenue management.

The process will include data preprocessing, exploratory analysis, feature engineering, model development, and performance assessment. Key evaluation metrics like accuracy, precision, recall, and F1-score will be used to gauge the model’s effectiveness.

Additionally, the study will examine how different factors influence cancellations, uncovering patterns and opportunities for enhancement. Ultimately, this initiative offers significant benefits to the hospitality sector by enabling data-driven decisions. With an accurate cancellation forecast, hotels can allocate resources more efficiently, enhance guest experiences, and boost profitability.

Business Understanding

The hospitality industry faces significant revenue losses due to booking cancellations, which average 40% of reservations. To address this, a predictive model can help hotels forecast cancellations and take proactive measures—such as targeted discounts or optimized staffing—to minimize losses. By analyzing historical data, the model enables better inventory management and revenue strategies. This project provides business value by improving occupancy rates, enhancing customer satisfaction, and reducing financial risks. With accurate predictions, hotels can gain a competitive edge, ensuring efficient resource allocation and higher profitability. Ultimately, the model empowers hotels to make data-driven decisions, transforming cancellation challenges into opportunities for growth.

Data Understanding

To achieve this objective, I have utilized the Hotel Reservations Dataset obtained from Kaggle. The dataset contains information about hotel bookings made by customers, including various features such as the number of adults and children, the type of meal plan, the requirement for a car parking space, the type of room reserved, the lead time, the arrival date, the market segment type, and whether the booking was cancelled or not.

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.

Decision Tree Model

This achieved the results below:

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

Random Forest Model

This achieved the results below:

- 90.2% Accuracy

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

Recommendations

	Logistic	Decision Tree	Random Forest
Accuracy	80.4%	84.8%	90.2%
Precision	74.7%	86.3%	88.6%
Recall	63%	65%	81.4%
F1 score	68.4%	74.2%	84.8%
AUC	0.76	0.8	0.88

- Final Model Selection: I selected the Random Forest Classifier as the final model. This model offers a good balance between complexity and performance, achieving a precision score of 89%, which is a significant improvement over the Logistic Regression(74.7%) and Decision Tree models(86.3%).
- Key Factor - Lead Time: The primary factor influencing cancellations is lead time, which measures the number of days between booking and arrival. Longer lead times correlate with higher cancellation rates, as plans can change over time. Hotel management should consider monitoring lead times closely and potentially adjusting cancellation policies or offering incentives for early confirmations to reduce

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%