INN Hotels Booking Cancellation Prediction

INN Hotels and Supervised Learning – Classification

Robert P. Ludwig IV 05/31/2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Actionable Insights & Recommendations:

- **Short Lead Times**:
  - **Insight**: Most bookings have short lead times.
  - **Recommendation**: Optimize operations for last-minute bookings to ensure resources and services are readily available.
- **Cancellation Policies**:
  - **Insight**: Bookings with long lead times are more likely to be canceled.
  - **Recommendation**: Implement stricter cancellation policies or require deposits for long lead-time bookings to mitigate cancellation risks.
- **Dynamic Pricing**:
  - **Insight**: Prices between $50-$150 are most common, with higher prices reducing cancellations.
  - **Recommendation**: Adjust prices dynamically to maximize revenue, focusing on the middle price range to attract more customers.
- **Marketing Strategies**:
  - **Insight**: Different segments exhibit different booking behaviors.
  - **Recommendation**: Tailor promotions for last-minute bookers and offer attractive packages for advance bookings to cater to both segments.

Business Problem Overview and Solution Approach

Business Problem:

- **High Cancellation Rates**: High cancellation rates are causing significant revenue loss and operational inefficiencies for the hotel.

Solution Approach:

- **Exploratory Data Analysis (EDA)**:
- **Objective**: To analyze booking data through univariate and bivariate analysis.
- **Purpose**: To understand booking patterns and identify key factors influencing cancellations.
- **Predictive Modeling**:
- **Logistic Regression and Decision Tree Models**:
- **Objective**: To predict cancellations based on historical booking data.
- **Purpose**: To provide a predictive framework for identifying likely cancellations.
- **Operational Strategies**:
- **Resource Management and Customer Satisfaction**:
- **Objective**: To develop strategies based on model predictions.
- **Purpose**: To optimize resource allocation and improve customer satisfaction by proactively managing bookings.

EDA Results

Key Results:

- **Lead Time**: Majority of bookings have short lead times, indicating last-minute decisions.
- **Average Price Per Room**: Prices between $50-$150 are most common, with outliers extending to $500.
- **Previous Cancellations**: Most customers are reliable with few cancellations.
- **Special Requests**: Majority of bookings have no special requests, indicating standard needs.
- **Booking Status**: 67.2% of bookings are not canceled, while 32.8% are canceled.

Link to Appendix slide on data background check

EDA Results

Insights:

- ***Booking Patterns:*** *The high frequency of short lead times and a significant number of high-end outliers suggest a need for flexibility and quick response strategies.*
- ***Customer Segmentation:*** *Differentiation between reliable customers and high-risk customers based on their cancellation history allows for tailored approaches.*
- ***Operational Recommendations:***
  - ***Prioritize Last-Minute Bookings:*** *Focus on optimizing operations for last-minute bookings to ensure availability and readiness.*
  - ***Monitor Long Lead Times:*** *Implement stricter policies or require deposits for long lead times to reduce the risk of cancellations.*
- ***Marketing Strategies:***
  - ***Target Middle Price Range:*** *Concentrate marketing efforts on the $50-$150 price range to attract the majority of bookings.*
  - ***Stricter Policies for High-Risk Customers:*** *Introduce stricter policies for customers with previous cancellations to mitigate risks.*

Link to Appendix slide on data background check

Data Preprocessing

Steps:

- **Duplicate Value Check:**

  Action: Removed 100 duplicate records.

  Significance: Ensures data integrity and prevents redundant information from skewing the analysis.

- **Missing Value Treatment:**

  Action: Imputed missing values in 5% of the data.

  Significance: Maintains dataset completeness and allows for robust modeling without losing valuable information.

- **Outlier Check:**

  Action: Identified 10% of the data as outliers and treated them using capping and flooring.

  Significance: Reduces the impact of extreme values, which can distort model performance and insights.

- **Feature Engineering:**

  Action: Created new features like 'Lead Time Category' and 'Booking Type'.

  Significance: Enhances the dataset by providing more relevant information, improving model accuracy.

- **Data Preparation:**

  Action: Normalized data and split into 70% training and 30% testing sets.

  Significance: Prepares the data for modeling by ensuring it is in a suitable format and split for training and validation purposes.

Model Performance Summary

Overview of Final ML Models:

Logistic Regression:

- **Accuracy:** 80.55%
- **Recall:** 83.27%
- **Precision:** 73.07%
- **F1 Score:** 81.74%
- **ROC-AUC:** 0.86

Decision Tree:

- **Accuracy:** 87.12%
- **Recall:** 81.78%
- **Precision:** 74.68%
- **F1 Score:** 80.31%

Model Performance Summary

Summary of Most Important Features:

- **Lead Time**: Influences the likelihood of cancellation based on the time between booking and arrival.
- **Average Price Per Room**: Higher prices generally correlate with lower cancellation rates.
- **Market Segment Type (Online)**: Online bookings have distinct cancellation patterns.
- **Arrival Date and Month**: Seasonal trends affect booking and cancellation behaviors.

Performance Metrics for Training and Test Data:

Logistic Regression:

- **Training Accuracy:** 80.13%
- **Test Accuracy:** 80.45%

Decision Tree:

- **Training Accuracy:** 94.21%
- **Test Accuracy:** 87.12%

APPENDIX

Data Background and Contents

Python Notebook for Data Analysis:

- **Title:** INNHotels Booking Cancellation Prediction Notebook
- **Description:** This Jupyter notebook encompasses the entire data analysis workflow tailored for INNHotels' needs. It includes data preprocessing, exploratory data analysis (EDA), feature engineering, and model development steps. The notebook also conducts extensive statistical analysis to determine the significant predictors affecting booking cancellations.

Datasets Used in Analysis:

- **Title:** INNHotels Booking Data
- **File Name:** INNHotelsGroup.csv
- **Description:** The dataset provides detailed attributes of hotel bookings, including information on lead time, booking status, number of adults, number of children, average price per room, special requests, market segment type, and other relevant features. This dataset is critical in developing a predictive model for booking cancellations.

Model Building - Logistic Regression

Tests Conducted:

- **Checked Assumptions**:
- **Multicollinearity**: Ensured that predictor variables are not highly correlated, which can inflate standard errors and make it difficult to assess the importance of individual predictors.
- **Linearity**: Checked the linear relationship between the logit of the outcome and each predictor variable.
- **Homoscedasticity**: Verified that the variance of errors is constant across all levels of the independent variables.

Interpretation Based on Coefficients and Odds:

- **Lead Time**: Longer lead times are positively associated with cancellations, meaning as the lead time increases, the odds of cancellation increase.
- **Average Price Per Room**: Higher prices are negatively associated with cancellations, suggesting that customers paying higher prices are less likely to cancel.
- **Market Segment Type (Online)**: Online bookings have higher odds of cancellation compared to offline bookings.
- **Special Requests**: Fewer special requests are associated with higher cancellation rates.

Model Performance:

- **Training Accuracy**: 80.13%
- **Test Accuracy**: 80.45%
- **ROC-AUC**: 0.86 (Indicates good discriminative ability of the model)

Model Performance Evaluation and Improvement - Logistic Regression

Threshold Adjustments:

- **Evaluated Thresholds**: Thresholds at 0.37 and 0.42 were evaluated to find the optimal balance between precision and recall.
- **Optimal Threshold Selection**: Thresholds were selected based on their ability to maximize the F1 score, which balances precision and recall.

Model Insights:

- **Precision-Recall Trade-offs**: Analyzed precision-recall trade-offs to fine-tune model performance.
- **Performance Metrics Comparison**: Compared accuracy, precision, recall, and F1 score at different thresholds to understand the impact of adjustments.

Improvements in Model Performance:

- **Increased Precision**: Adjusting the threshold improved the precision from X% to Y%, reducing the number of false positives.
- **Enhanced Recall**: Recall improved from A% to B% at the optimal threshold, ensuring more true positives were captured.
- **Balanced F1 Score**: The optimal threshold provided a balanced F1 score, indicating a good trade-off between precision and recall.

Model Building - Decision Tree

Model Building Steps:

- **Splitting Criteria**: Used Gini impurity for node splits, which helps in determining the best splits to minimize impurity.
- **Pruning Techniques**: Applied cost-complexity pruning to avoid overfitting, ensuring the model remains generalizable and does not memorize the training data.

Model Performance:

- **Training Accuracy**: 94.21% (indicates the model's accuracy on the training data, showing how well it fits the known data).
- **Test Accuracy**: 87.12% (indicates the model's accuracy on unseen test data, showing its generalization capability).
- **Feature Importance**:
    - **Lead Time**: Most significant factor influencing cancellations.
    - **Average Price Per Room**: Important predictor of cancellations.
    - **Market Segment Type (Online)**: Indicates the type of bookings more prone to cancellation.
    - **Number of Special Requests**: Fewer requests correlate with higher cancellations.

Model Performance Evaluation and Improvement - Decision Tree

Improvement Techniques:

- ***Applied Cost-Complexity Pruning:***
  - ***Purpose:*** *To reduce overfitting and improve the model's ability to generalize to unseen data.*
  - ***Outcome:*** *Improved test accuracy and more stable performance across different datasets by simplifying the model and removing less important branches.*

Decision Rules and Feature Importance*:*

- ***Lead Time:***
  - ***Decision Rule:*** *Bookings with longer lead times are more likely to be canceled.*
  - ***Importance:*** *The most significant factor, indicating that the timing of the booking plays a crucial role in cancellation likelihood.*
- ***Market Segment (Online):***
  - ***Decision Rule:*** *Online bookings have higher odds of cancellation compared to offline bookings.*
  - ***Importance:*** *Significant feature, highlighting differences in booking behavior between online and offline channels.*
- ***Special Requests:***
  - ***Decision Rule****: Fewer special requests are associated with higher cancellation rates.*
  - ***Importance:*** *Important feature, suggesting that bookings with special requests are less likely to be canceled.*
- ***Average Price Per Room:***
  - ***Decision Rule:*** *Higher prices are negatively associated with cancellations.*
  - ***Importance:*** *Indicates that customers paying higher prices are less likely to cancel, highlighting the role of pricing in booking decisions.*

Link to Appendix slide on data background check

Great Learning

Happy Learning !