INN Hotels

Supervised Learning Classification

Mona Desai
Date: 12/10/2022

Contents / Agenda

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

Executive Summary

- **Conclusions, actionable insights:**
    - This facility seems to be primarily family based hotel with ~85% of total bookings
    - Cancellations rates are very high - ~33% (Though dont have data on Industry/Local baseline)
    - Market_segment_type, no_of_special_requests, no_of_family_members, total_days, repeated_guests, arrival_month, columns have big impact on cancellation.
    - Most of the data did not give any significant impact to show the reason for cancellation
    - All logistic regression models gave the generalized performance on the training and testing sets
    - The best regression Model seems to be providing us prediction with 0.70 F1 score
    - I chose Pre-pruned tree as the best model since it is giving the most comparable value for F1 on both training and testing sets

- **Recommendations:**
    - Hedge bookings with Lead time over 40d.
    - Increase Aviation & Corporate bookings to reduce cancellations
    - Collect data on reasons for cancellations
    - Incentivize repeat customers as they tend to cancel less
    - Lead_time, market_segment_type_online, number_of_special_requests, avg_price_per_room are the most important features to predict the cancellation.Keep a keen eye on these features

Business Problem Overview and Solution Approach

- **Problem**
  - High rate of booking cancellations
  - Finding out the best Logistic Regression Model which could predict cancellation with high accuracy and repeatability
- **Solution approach / methodology**
  - Exploratory Data Analysis
  - Model Building - Logistic Regression, Removing Multicollinearity, Optimal Threshold
    - Analyzing Odds and coefficients
    - Applying confusion matrix, AUC-ROC curve, Recall-Precision Curve
    - Best F1 value for Train and Test sets
  - Model Building - Decision Tree
    - Pre and Cost complexity Pruning
    - Applying Confusion Matrix
    - F1 score Vs. Alpha
    - Compare Accuracy, Precision, Recall, F1 values for Train and Test sets

EDA Results

- **Key results from EDA**
  - The Dataset contains different attributes of customer's booking details
  - The Data in the dataset
    - Collected in year 2017-2018
    - 36275 Rows
    - 19 Columns
  - 14 Numerical, 5 Categorical columns with no missing or duplicate values
  - Statistical Summary suggests a lot of variations in Data for each Numerical columns
  - lead_time ranges 0-443 days, shows a huge variation.
  - avg_price_per_room ranges 0 - 540 euros.
  - Booking_ID contains the unique value
- **Observation from Univariate Analysis**
  - lead_time is heavily right skewed.Average lead_time is 85 days
  - avg_price_per_room has lots of outliers and is right skewed
  - Some customers paids very high price for the booking. They could be traveling in the prime season or staying for a longer periods of time

EDA Results continued...

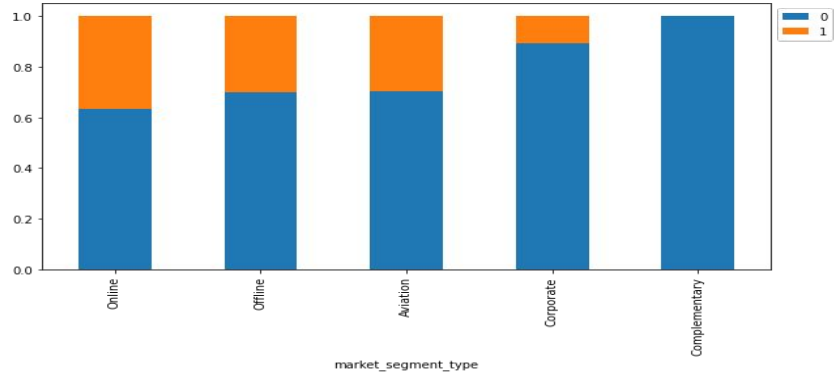
- **Observation from Univariate Analysis continued..**
  - 545 (~2%) reservations were made “0” avg_price_per_room. 354 “complementary and 191 “Online” market_segment_type
  - 72% had 2 adults and 93% bookings had no childrens
  - 97% bookings did not required car parking
  - 77% bookings had type_of_meal_plan 1
  - 64% : online, 29%: offline, ~6%: Corporate, 1% Complementary, 0.3%: Aviation bookings under market_segment_type
  - 11885/36275 (~33%) bookings were cancelled. That is really a lot.

- **Insights and Observation from Bivariate Analysis**
  - Average price through through 5 segments lies between 75 and 110 Euros.
  - Aviation segments have no outliers.
  - Bookings from online, offline and corporate type has higher price than the average price.
  - Cancellations from those segments could cost a lot to the hotel

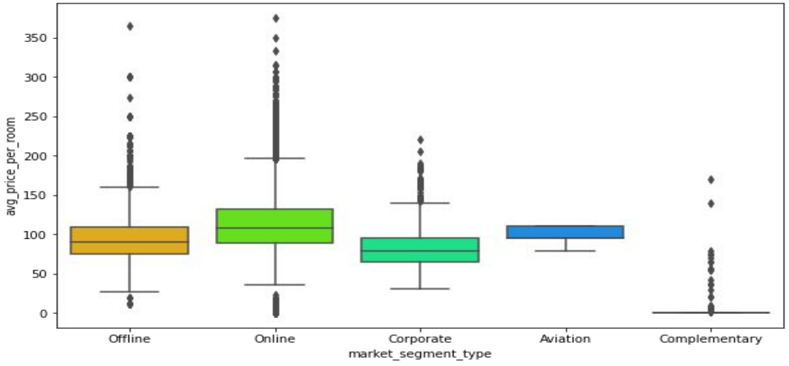
EDA Results continued...

- **Insights and Observation from Bivariate Analysis continued..**

“market_segment_type” vs. “booking_status”



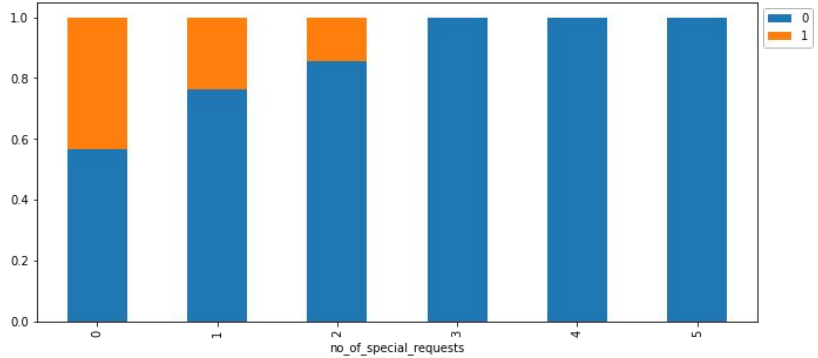
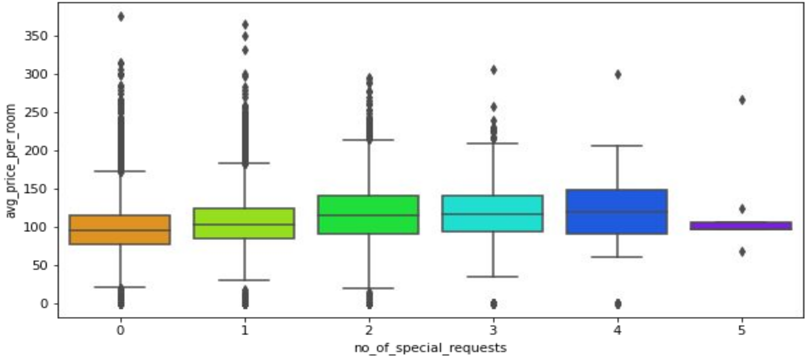
“market_segment_type” vs. “avg_price_per_room”



- Highest cancellations were from online bookings
- Online Booking
  - Average cost: 110 Euros
  - 75% of the online booking cost: ~200 Euros
  - Highest cost: ~375.
- Second highest cancellations were from offline bookings costs the second most loss.

EDA Results continued...

- **Insights and Observation from Bivariate Analysis continued..**

“Num_of_special_requests” vs. “booking_status”	“Num_of_special_requests”vs.“avg_price_per_room”
	

- The most cancellations has happened from the bookings where there were no special requests were made.
- 75% of the bookings costs for online range 75-375 euros
- No Cancellations from the bookings where 4-5 special requests were made

EDA Results continued...

- **Insights and Observation from Bivariate Analysis continued..**
  - ~115 euros avg_price_per_room is for those rooms whose booking is cancelled
  - ~90 euros avg_price_per_room is for those rooms whose booking is not cancellation
  - Average lead_time is ~125 days for the cancelation Vs ~40 days for non cancelation
  - Repeated Guests are hardly cancelling the booking Vs. New guests.
  - Reservations made for more than 8 days are not cancelled
  - Majority cancelled reservations are made for 3 and 4 days
  - Families having 4-5 members tend to cancel less.
  - ~40% bookings were only for one night.
  - Length of stay for 3-4 nights cancel the most.
  - August,September, October are the busiest months. Probably because Portugal has a beautiful weather during those months
  - Room costs more than double in the months of May-September. October and April has similar price range December-February months has the least costs and so are the cancellations.

Data Preprocessing

- **Duplicate value check**
  - The dataset has duplicate values after dropping the column with the unique “Booking_ID”. However we do not need to take any actions for that
- **Missing value treatment**
  - There is no missing values
- **Outlier check (treatment if needed)**
  - There are quite a few outliers in the data.
  - However, we will not treat them as they are proper values
- **Feature engineering**
  - We drop the column with the unique “booking_id” because it will not have impact on the dependent variable

Data Preprocessing continued..

- **Data preparation for modeling**
    - “booking_status” is the dependent categorical variable.
    - We encoded canceled booking to “1” and not_canceled to “0” under the column “booking_status”
    - encode “categorical” data
    - Create dummies for the categorical data
    - To build the model on the train set, split the data into train and test sets in 70:30 part
    - Build the logistic regression model using the data from the train set

Model Performance Summary

- **Overview of the final ML model and its parameters**
  - Built three different models using “Default Threshold (0.5)”, “0.37 Threshold”, “0.42 Threshold”
  - Almost all the three models are performing well on both training and test data without the problem of overfitting
  - The model with the 0.37 threshold gives the best F1 scores, that's why it can be selected as the final model
  - Used Accuracy, Recall, Precision and F1 score to measure the accuracy of the model
- **Summary of most important features used by the ML model for prediction**
  - Logistic Regression and odds ratio
  - VIF and P value
  - Confusion matrix
  - AUC-ROC curve and Precision-Recall Curve
  - Optimal Threshold

Model Performance Summary continued...

- **Summary of key performance metrics for training and test data of all the models in tabular format for comparison**
  - The model with the 0.37 threshold gives the best F1 scores, that's why it can be selected as the final model
  - <u>Training set performance comparison</u>

Testing performance comparison:

	Logistic Regression-default Threshold (0.5)	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80345	0.79555	0.80345
Recall	0.70358	0.73964	0.70358
Precision	0.69353	0.66573	0.69353
F1	0.69852	0.70074	0.69852

  - <u>Test set performance comparison</u>

Training performance comparison:

	Logistic Regression-default Threshold	Logistic Regression-0.37 Threshold	Logistic Regression-0.42 Threshold
Accuracy	0.80545	0.79265	0.80132
Recall	0.63267	0.73622	0.69939
Precision	0.73907	0.66808	0.69797
F1	0.68174	0.70049	0.69868

Model Building - Logistic Regression

- **Please mention regarding the tests conducted to check the assumptions of Logistic Regression**
  - P-value analysis - Removed below columns as they had high values (> .5)
    - arrival_date, num_of_previous_booking_not_cancelled, meal_plan_3, room_type_3, market_segment_type_complimentory, market_segment_type_online.
  - VIF - after removing high P-value, VIF values of all remaining columns were ~1
- **Interpret the results based on coefficients and odds**
  - <u>Coefficients</u>
    - Increase in positive coefficient of Lead_time, no_of_children, no_of_adults will increase the chance of the booking being cancelled
    - Increase in negative coefficient no_of_special_requests, repeated_guetsts, required_car_parking_space will decrease the chance of cancellation

Model Building - Logistic Regression continued..

- **Interpret the results based on coefficients and odds**
  - Odds
    - Holding all other features constant a 1 unit change in no_of_special_request will decrease the odds of a booking beling cancelled ~0.22 times or a ~77%
    - Holding all other features constant a 1 unit change in required_car_parking_space will decrease the odds of a booking beling ~0.20 times or ~80%
- **Comment on the model performance**
  - Almost all the three models are performing well on both training and test data without the problem of overfitting
  - The model with the 0.37 threshold gives the best F1 scores, that's why it can be selected as the final model
- **Comment on the improvement in the model performance by changing the classification threshold**
  - The model with a threshold 0.37 is giving the best F1 score changing from the default threshold 0.5 and then tried one more threshold at 0.42

Model Building - Decision Tree

- **Please mention the model building steps of Decision Tree**
    - Encode the Categorical Features and determine the Decision column
    - Divide Dataset into Train and Test sets
    - Build the Baseline Model
    - Model evaluation setting prediction
        - FN-Predicting a customer will not cancel their booking but in reality, the customer will cancel their booking.
        - FP-Predicting a customer will cancel their booking but in reality, the customer will not cancel their booking.
    - Checking the Confusion Matrix and Model performance on Train and Test set
    - Checking the important features of the dataset which could impact the dependent variable

Model Building - Decision Tree continued...

- **Please mention the model building steps of Decision Tree continued..**
    - Pre-Pruning
    - Checking the Confusion Matrix and Model performance on Train and Test set
    - Visualizing the Decision Tree
    - Checking the rules and important features
    - Cost Complexity Pruning
    - F1 Score Vs. Alpha for train and test sets
- **Comment on the model performance**
    - Among all three baseline, pre-pruning and post-pruning trees, the pre-pruning model has the best comparable F1 score for both Test and Train sets
    - Accuracy, Recall and precision value matches the best too for Pre-pruning model.

Model Performance Evaluation and Improvement - Decision Tree

- **Comment on the model performance continued..**
  - Training performance Comparison

Training performance comparison:

	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

  - Testing performance Comparison

Testing performance comparison:

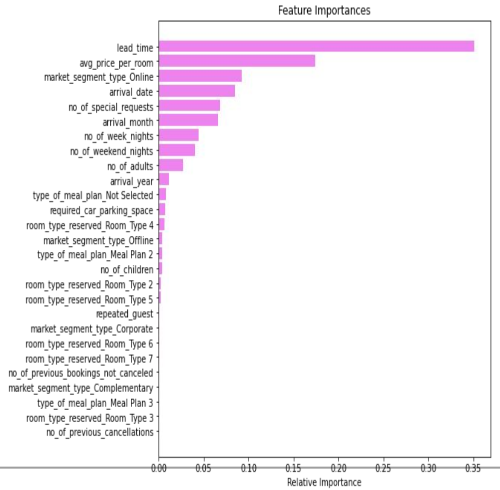
	Decision Tree sklearn	Decision Tree (Pre-Pruning)	Decision Tree (Post-Pruning)
Accuracy	0.87118	0.83497	0.86879
Recall	0.81175	0.78336	0.85576
Precision	0.79461	0.72758	0.76614
F1	0.80309	0.75444	0.80848

- Decision tree models with pre-pruning and post-pruning both are giving equally high recall scores on both training and test sets.
- However, we will choose the pre pruned tree as the best model since it is giving the most comparable value for f1

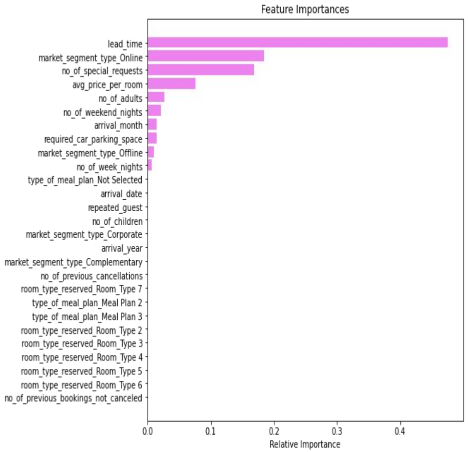
Model Performance Evaluation and Improvement - Decision Tree Continued...

Features Importance

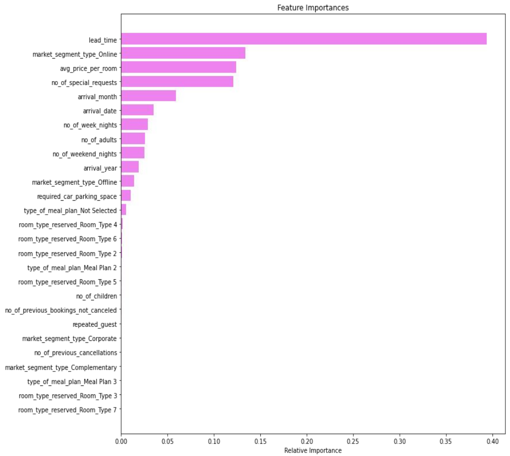
Baseline



Pre-Pruning



Post-Pruning



Lead_time, market_segment_type_online, number_of_special_requests, avg_price_per_room are the most important features to predict the cancellation

Model Performance Evaluation and Improvement - Decision Tree Continued...

- **Decision rules**
  - From the Decision Tree (Pre-pruned)- Lead_time>151.50,avg_price_per_room>100.4,arrival_month< =11.50,number_of_special_requests< =2.50, then the reservation is most likely going to be cancelled
  - The Hotel should keep a keen eye for these values in order to know if the booking is going to be cancelled