Decision System

No code AI/ML - MIT

05/15/2024

Presented by : Nehal Naik

Contents / Agenda

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

Executive Summary

High Cancellation Rate: The cancellation rate at INN Hotel Group is currently close to 33%, indicating a significant challenge. Immediate attention is required to proactively identify bookings with a higher likelihood of cancellation and implement appropriate measures to mitigate this issue.

Machine Learning Solution: A Random Forest-Prune based machine learning model has been developed and finalized for the purpose of predicting the likelihood of cancellation for existing bookings. This model will enable proactive management of cancellations by identifying potential cancellations well in advance.

Proactive Customer Engagement: Upon identifying bookings flagged as more likely to be canceled using ML based solution, INN Hotel Group can initiate targeted communication with clients, offering incentives such as complimentary breakfast or room upgrades. This proactive approach aims to reduce the high cancellation rate and enhance customer satisfaction.

Strategic Focus: Following a thorough analysis of available data, it is recommended that INN Hotel Group increase its focus on corporate clients and families. Currently, the proportion of business from these segments is considerably lower compared to individual clients. By targeting these segments effectively, INN Hotel Group can diversify its customer base and strengthen its business performance.

Executive Summary

Model development and selection results:

- Upon evaluating available data and problem statement, classification technique is used to flag exsiting booking's likellyhood of being cancelled as a part of the solution.
- Based on availbale historical data, 4 different classification models were tested and one was finalzed with better outcome in terms of better overall accuracy and other parameteres.
- **Random forest - Prune** based ML model was finalized based on the its accuracy to predict likelyhood of cancelation and for having minimal difference between results fo training set and testing seet.

Business Problem Overview and Solution Approach

Problem statement:

- The hotel industry grapples with a significant challenge: an increasing number of booking cancellations, disrupting revenue streams and inflating operational costs. While technological advancements have streamlined the booking process for clients, they've also facilitated easy, last-minute cancellations. INN Hotels Group, with its chain of hotels in Portugal, faces this dilemma and seeks a solution.

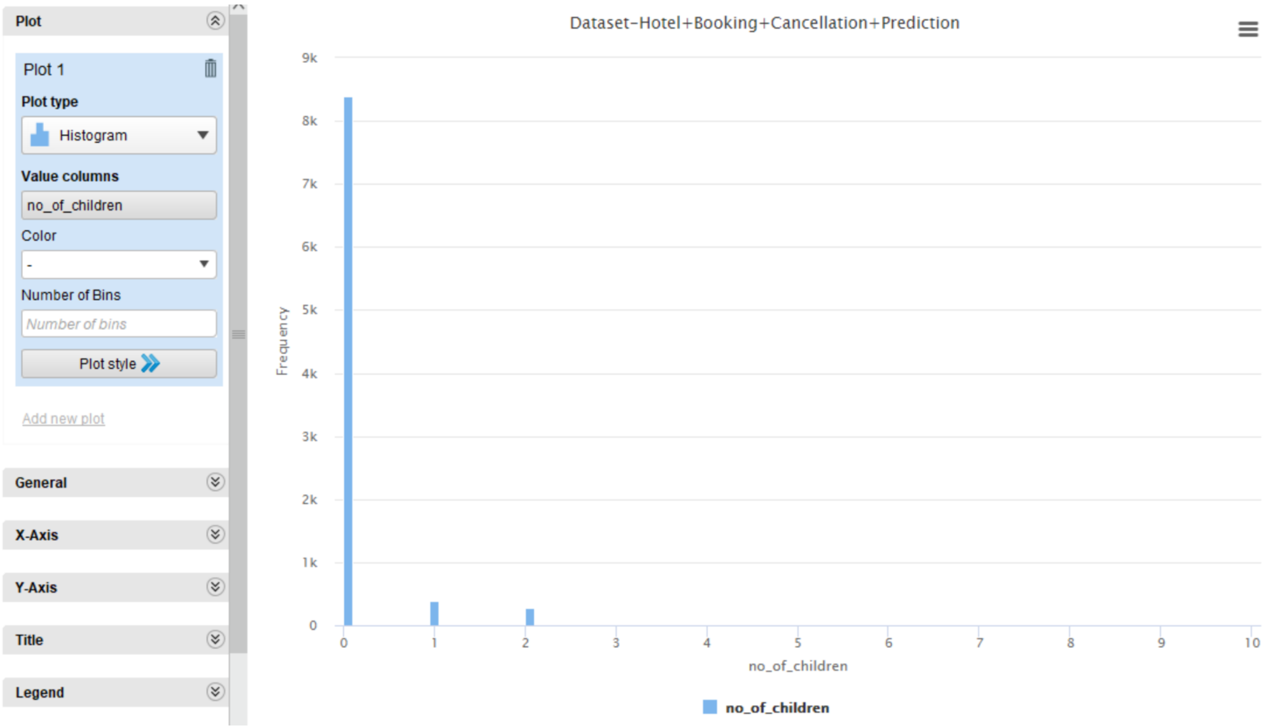
- Recognizing the urgency of the situation, there's a pressing need to find a solution that can predict which bookings are likely to be canceled. Such a solution empowers INN Hotels Group to react swiftly and devise effective strategies to mitigate losses stemming from cancellations.

Business Problem Overview and Solution Approach

Proposed solution

- Basic Exploratory Data Analysis (EDA) needs to be performed to find obvious connection and correlations among captured attributes to identify patterns and business insight.
- Looking at the problem statement and available data points, variation of Decision Tree methodology will provide much needed solution to predict likelihood of booking cancellation.
- From available data set, 70% of the records will be used as training set, while 30% will be used at testing set.
- Based on the performance of the methodology, final solution will be picked to be used.
- Will measure success of the final solution for next 6 months to decided if further improvement is needed.

EDA Results



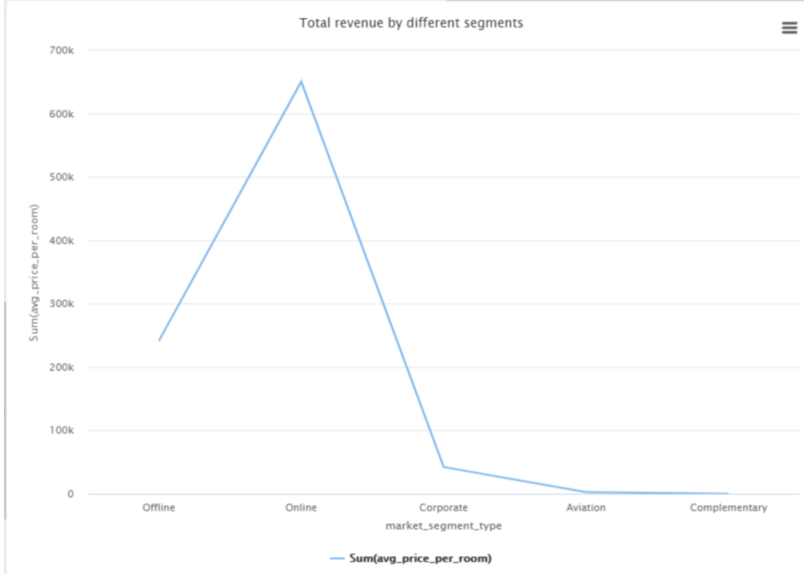
OBERVATIONS:

- More customers are coming without children.

SUGESSTIONS:

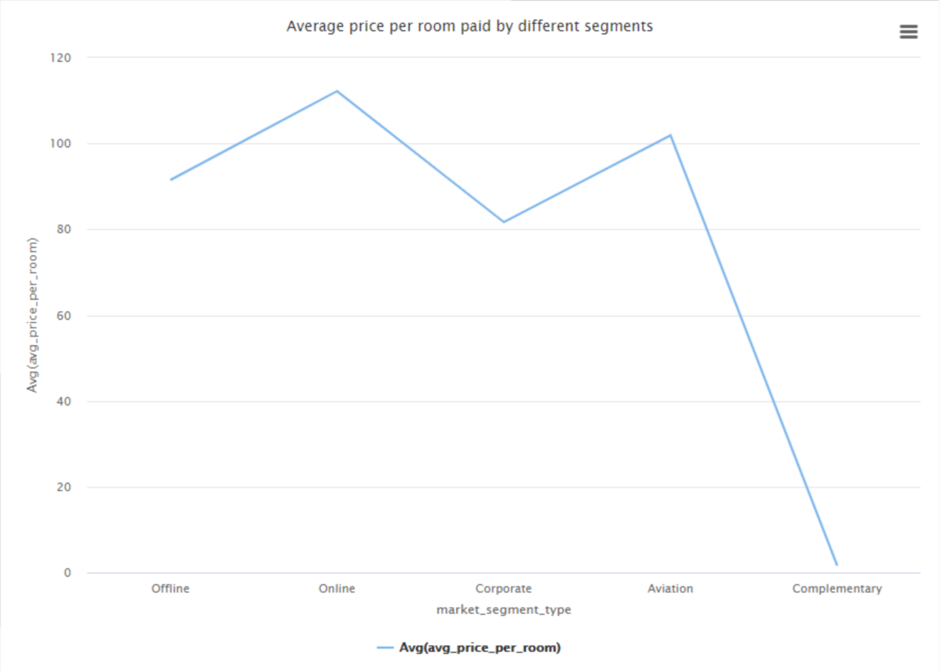
- Need to focus more on family segment as company might be missing out opportunity to attract families.

EDA Results



OBSERVATION:

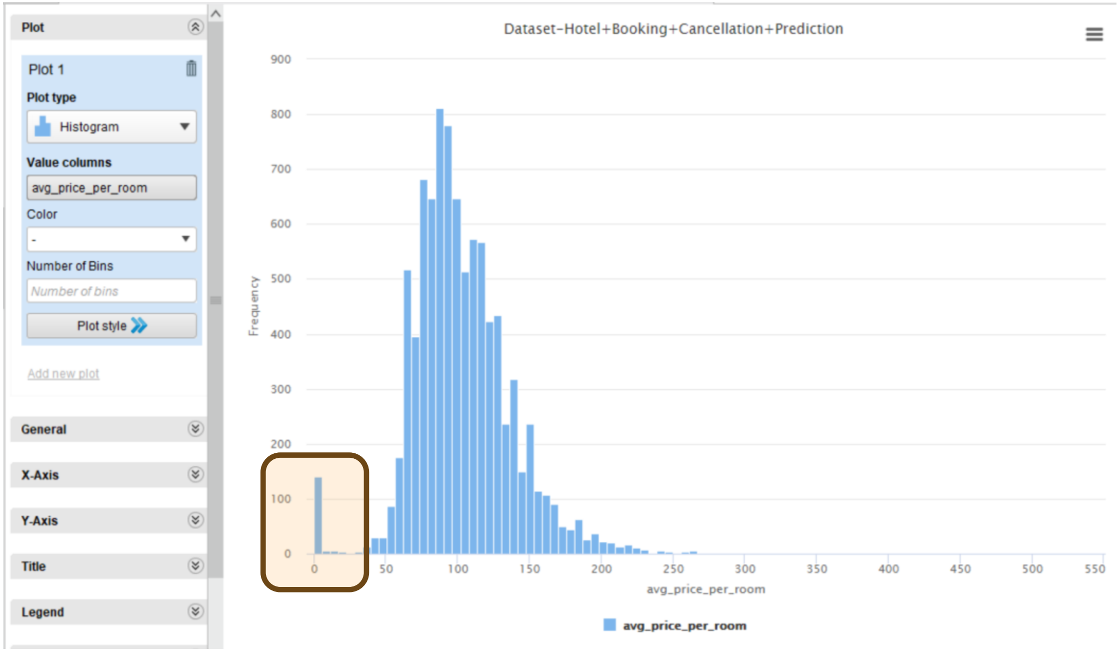
- Contribution from Corporate customer towards revenue is very small compared to Online and Offline, while average price paid is not much different.



SUGGESTIONS:

- There is a potential to increase corporate clients, which will improve revenue stream.

EDA Results



- Need to check data quality issue as many records with small amount of average price.
- Upon investigation, it was found that it came from complementary category

Model Performance Summary

Overview of the final ML model and its parameters

- Based on the problem definition, four classification models were be tested and one with the best accruacy, precesion and recall is picked.

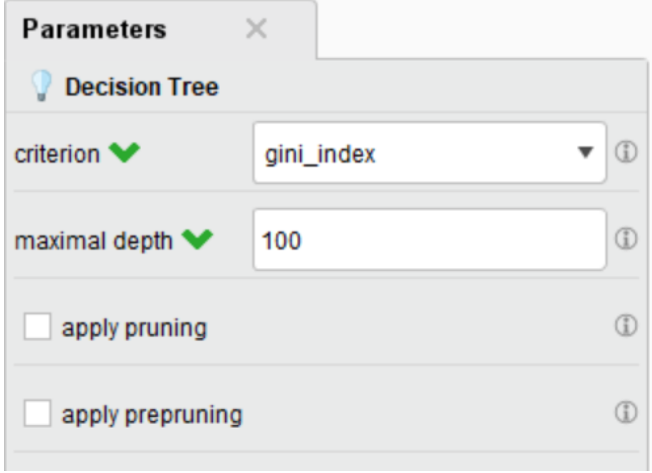
- Model used:
    - Decision Tree
    - Decision Tree – Prune
    - Random Forest
    - Random Forest - Prune

Model Performance Summary

Overview of the final ML model and its parameters

Note: Rapidminer is used as No-code AI tool to build and test models

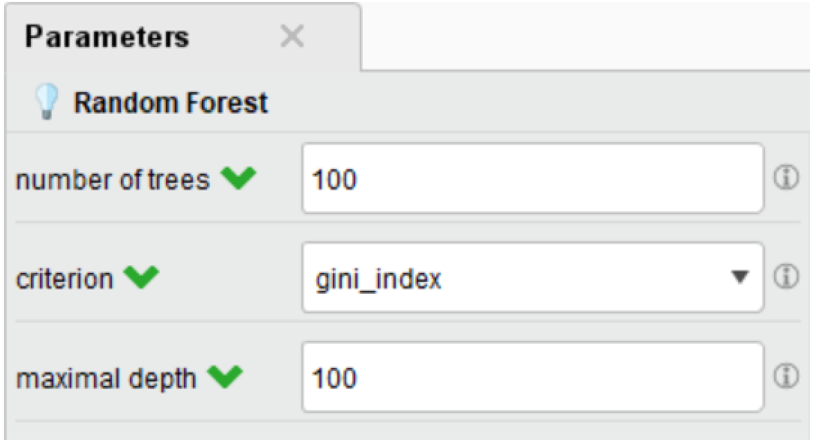
- **Decision Tree :** Following parameters are selected in cofiguration of this model to optimize runtime vs accuracy.



Model Performance Summary

Overview of the final ML model and its parameters

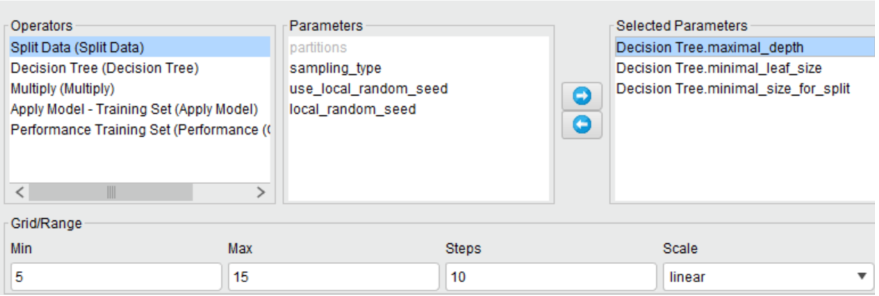
- **Random Forest :** Following parameters are selected in cofiguration of this model to optimize runtime vs accuracy.

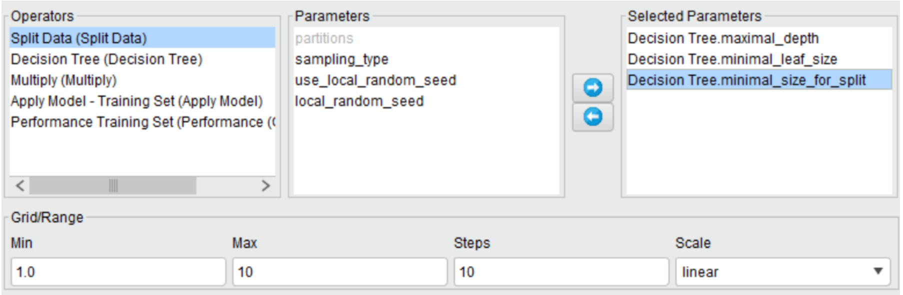


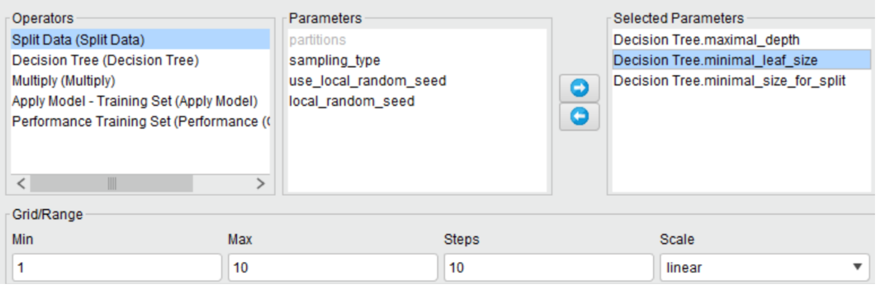
Model Performance Summary

Overview of the final ML model and its parameters

- **Decision Tree – Prune :** Following parameters are selected in cofiguration of this model to optimize runtime vs accuracy.



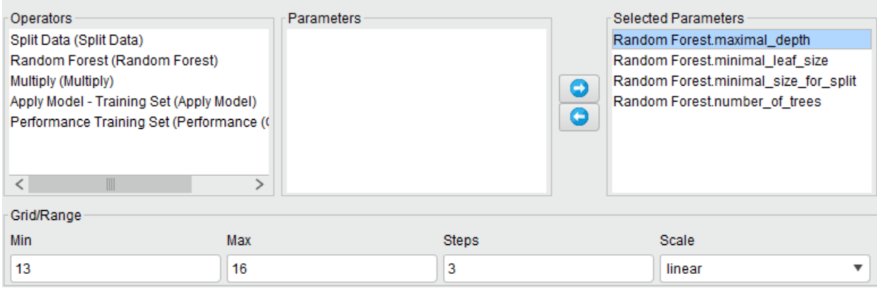


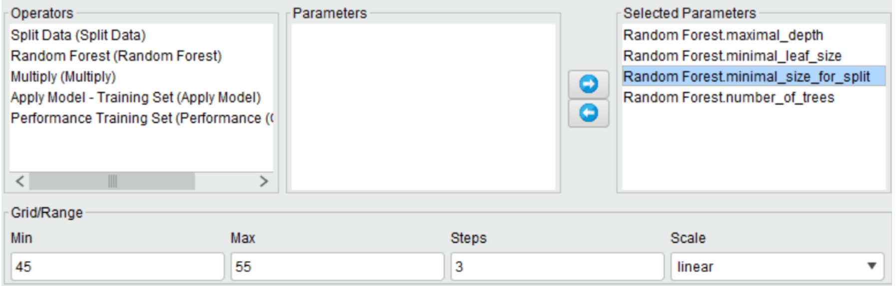


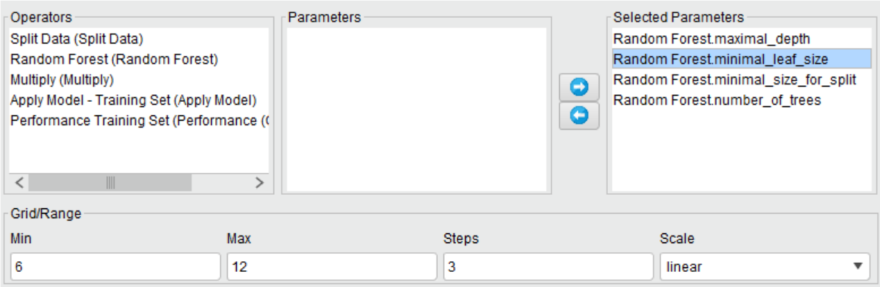
Model Performance Summary

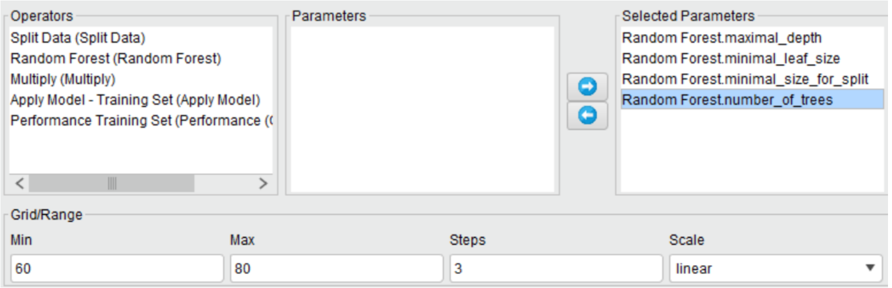
Overview of the final ML model and its parameters

- **Decision Tree – Prune :** Following parameters are selected in cofiguration of this model to optimize runtime vs accuracy.









Model Performance Summary

Summary of most important features used by the ML model for prediction

Attribute	Weights
lead_time	0.21651
no_of_weekend_nights	0.14017
no_of_week_nights	0.12891
avg_price_per_room	0.11180
arrival_date	0.09728
arrival_month	0.08634
type_of_meal_plan	0.06042
room_type_reserved	0.05738
no_of_adults	0.04563
no_of_children	0.02352
arrival_year	0.01165
market_segment_type	0.01002
required_car_parking_space	0.00696
no_of_special_requests	0.00323
no_of_previous_bookings_not_canceled	0.00014
repeated_guest	0.00003

- Attribute weights are calcuated to determine which attributes will add more value when used to build the tree.
- Table shows importance of these attributes in order of prioirty.
- Classification models will used these weights when bilding tree based on other parameters.

Model Performance Summary (Decision Tree)

- Following tables clearly shows that Precision and Class Recall drops significantly (~23%) when model is run on Test data set.
- Overall accuracy on test data is reported at 83.20% which is not very comforting.
- This points a common phenomena of model overfitting the data during training.

Decission Tree - Training Set

	true Canceled	true Not_Canceled	class precision
pred. Canceled	2069	11	99.47%
pred. Not_Canceled	11	4258	99.74%
class recall	99.47%	99.74%	99.65%

Decission Tree - Testing Set

	true Canceled	true Not_Canceled	class precision
pred. Canceled	684	250	73.23%
pred. Not_Canceled	207	1579	88.41%
class recall	76.77%	86.33%	83.20%

Model Performance Summary (Decision Tree – Pruned)

- Following tables clearly shows that Precision and Class Recall drops (~16%) when model is run on Test data set.
- Overall accuracy on test data is reported at 84.34% which is more than non-pruned version of decision tree provided, but not very comforting.
- As we are focusing on cancellation problem, having 76.09% recall and precision on test data does not provide confidence in the model.

Decission Tree Pruned- Training Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	1888	85	95.69%
pred. Not_Canceled	192	4184	95.61%
class recall	90.77%	98.01%	95.64%

Decission Tree Pruned - Testing Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	678	213	76.09%
pred. Not_Canceled	213	1616	88.35%
class recall	76.09%	88.35%	84.34%

Model Performance Summary (Random Forest)

- Following tables shows that Precision and Class Recall remained very strong when running model on test data.
- Overall accuracy on test data is reported at 87.94% which is the best among the four models ran so far. But difference between Training and Test is almost 10%, which indicates overfitting and adds some doubts on the model performance on future unseen data.

Random Forest - Training Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	2006	23	98.87%
pred. Not_Canceled	74	4246	98.29%
class recall	96.44%	99.46%	98.47%

Random Forest - Testing Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	687	124	84.71%
pred. Not_Canceled	204	1705	89.31%
class recall	77.10%	93.22%	87.94%

Model Performance Summary (Random Forest – Pruned)

- Following tables shows that Recall for Positive case is very low (66%)while running model on test data.
- Overall accuracy on test data is reported at 85% which is very good, and it is also not much different than training set. This brings more confidence in the model and should be used as a final model

Random Forest - Pruned - Training Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	1453	178	89.09%
pred. Not_Canceled	627	4091	86.71%
class recall	69.86%	95.83%	87%

Random Forest - Pruned - Testing Set

	true Canceled	true Not_Canceled	**class precision**
pred. Canceled	590	106	84.77%
pred. Not_Canceled	301	1723	85.13%
class recall	66.22%	94.20%	85%

APPENDIX

Data Background and Contents

- Following is the shcema in which booking data is captured.

Column	Description
Booking_ID	the unique identifier of each booking
no_of_adults	Number of adults
no_of_children	Number of Children
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
type_of_meal_plan	Type of meal plan booked by the customer
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1- Yes)
room_type_reserved	Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels Group
lead_time	Number of days between the date of booking and the arrival date
arrival_year	Year of arrival date
arrival_month	Month of arrival date
arrival_date	Date of the month
market_segment_type	Market segment designation.
repeated_guest	Is the customer a repeated guest? (0 - No, 1- Yes)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
no_of_special_requests	Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
booking_status	Flag indicating if the booking was canceled or not.

Model Building - Decision Tree / Random Forest

Based on the test conducted, Random forest model behaved little bit better than Decisoin Tree model

Overall Accuracy

	Decision Tree	Decision Tree - Prune	Random Forest	Random Forest - Prune
Training	99.65%	95.64%	98.47%	87.32%
Test	83.20%	84.34%	87.94%	85.04%

Model Performance Evaluation and Improvement - Decision Tree / Random Forest

Pruning technique not only helped in reducing complexity of the tree and its runtime but provided little better correlation between training and test results.

Overall Accuracy

	Decision Tree	Decision Tree - Prune	Random Forest	Random Forest - Prune
Training	99.65%	95.64%	98.47%	87.32%
Test	83.20%	84.34%	87.94%	85.04%

Happy Learning !