Hotel Booking Cancellation Prediction

Decision Systems, Random Forests

3/10/24

Contents / Agenda

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

Executive Summary

- **Recommendation**:
- From the data, the Client should utilize the Pruned Random Forest to aggregate its data into understanding the factors that lead to a Cancellation
- The client should understand how other variables such as Special Requests and Meal Plans can improve the number of Arrivals – this would especially be helpful in the Summer when the Cancellations are highest, but the number of Arrivals is around the average

- **Rationale**:
- Lead Time carried the greatest weight and was the Root Node for each process
- The Pruned Random Forest took the longest time to execute (~3.5 minutes)
- The Pruned Random Forest had the lowest Δ between the Training and Testing dataset accuracy.
  - The Training dataset accuracy for the other 3 models were in the high 90%
  - The accuracy in the Testing set was much lower in the 80's
  - The accuracy for the Pruned Random Forest was 87.32% in the Training set and 85.04% in the Testing set

Business Problem Overview and Solution Approach

- Please define the problem
  - Historically, hotel cancellations have been an area of concern for overall revenue because hotels are unable to determine with great precision, the likelihood that a reservation will be cancelled. If the reservation is cancelled, the hotel will lose money depending on how close it is to the check-in date. To better understand if a cancellation will occur, we will need to better understand the other variables and their correlation to a cancellation.

- Please mention the solution approach / methodology
  - We will determine the likelihood that a customer will cancel their reservation by using Decision Trees and Random Forests to understand the factors that lead to a cancellation.
  - Our Root Node is Lead time, but we will measure other factors such as Arrival Month, Special Requests, Meal Plan, Price, and Number of Adults
  - We will first conduct the Decision Tree process before using the pruned Decision Tree
  - Second, we will conduct the Random Forest process before using the pruned Random Forest

EDA Results

Total Processing Time

- The chart on the right shows the amount of time it took to conduct each process. From the chart, we can see that the Pruned processes took significantly longer than the unpruned processes. Additionally, the Decisions Trees took much less time than the Random Forest processes.

Correlation Matrix

- The Correlation Matrix shows us which variables have Strong and Weak relationships. In the chart, we can see Repeated Cancellations and Previous Cancellations are highly correlated with Previous Bookings Not Cancelled: 0.509 and 0.499, respectively

Total Processing Time	
Process	Time to Execute (min:sec)
Decision Tree	0:01
Decision Tree – Pruned	0:18
Random Forest	0:05
Random Forest – Pruned	3:29

Attribut...	no_of_a...	no_of_c...	no_of_...	no_of_...	require...	lead_ti...	arrival_...	arrival_...	arrival_...	repeate...	no_of_p...	no_of_p...	avg_pri...	no_of_s...
no_of_a...	1	-0.022	0.118	0.099	-0.011	0.104	0.070	0.014	0.011	-0.193	-0.045	-0.118	0.290	0.191
no_of_c...	-0.022	1	0.019	0.015	0.047	-0.044	0.048	0.001	0.022	-0.036	-0.017	-0.020	0.330	0.116
no_of_w...	0.118	0.019	1	0.197	-0.043	0.052	0.062	-0.029	0.019	-0.060	-0.015	-0.015	-0.005	0.079
no_of_w...	0.099	0.015	0.197	1	-0.064	0.160	0.034	0.032	-0.003	-0.090	-0.015	-0.023	0.013	0.052
required...	-0.011	0.047	-0.043	-0.064	1	-0.072	0.029	-0.015	-0.004	0.115	0.028	0.063	0.053	0.084
lead_time	0.104	-0.044	0.052	0.160	-0.072	1	0.162	0.128	0.001	-0.143	-0.052	-0.077	-0.069	-0.099
arrival_y...	0.070	0.048	0.062	0.034	0.029	0.162	1	-0.355	0.015	-0.032	0.003	0.027	0.180	0.059
arrival_...	0.014	0.001	-0.029	0.032	-0.015	0.128	-0.355	1	-0.032	0.009	-0.044	0.004	0.054	0.103
arrival_d...	0.011	0.022	0.019	-0.003	-0.004	0.001	0.015	-0.032	1	-0.022	-0.007	0.002	0.016	0.016
repeated...	-0.193	-0.036	-0.060	-0.090	0.115	-0.143	-0.032	0.009	-0.022	1	0.397	0.509	-0.162	-0.022
no_of_pr...	-0.045	-0.017	-0.015	-0.015	0.028	-0.052	0.003	-0.044	-0.007	0.397	1	0.499	-0.059	0.002
no_of_pr...	-0.118	-0.020	-0.015	-0.023	0.063	-0.077	0.027	0.004	0.002	0.509	0.499	1	-0.097	0.019
avg_pric...	0.290	0.330	-0.005	0.013	0.053	-0.069	0.180	0.054	0.016	-0.162	-0.059	-0.097	1	0.183
no_of_s...	0.191	0.116	0.079	0.052	0.084	-0.099	0.059	0.103	0.016	-0.022	0.002	0.019	0.183	1

EDA Results (decision tree)

- In the chart below we can see the performance chart for the Pruned and Not Pruned Decision Trees.
- We can see that while the Accuracy of both Training data sets is 95%+, the Accuracy greatly falls for the Testing data sets below 85%.
- The precision fell significantly from 95%+ in the Training sets, to below 88% for the Testing sets.

	Training	Testing
Not Pruned	**accuracy: 99.65%** 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%	**accuracy: 83.20%** 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%
Pruned	**accuracy: 95.64%** 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%	**accuracy: 84.34%** 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%

EDA Results (random forest)

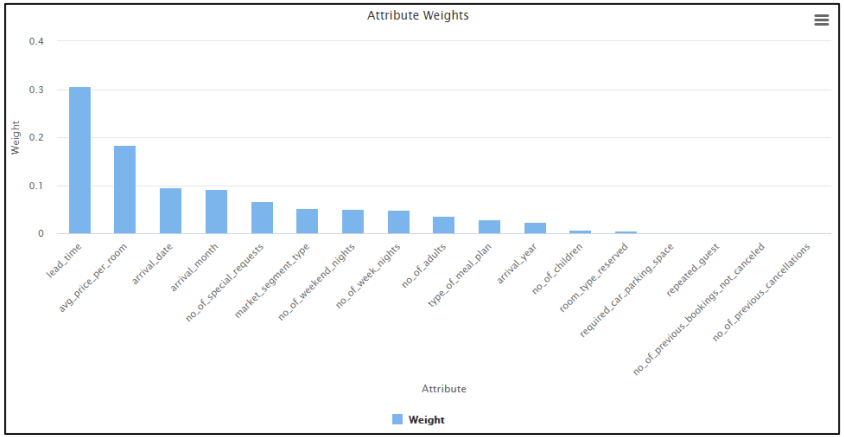
- In the Not Pruned Random Forest, the Accuracy improves in the Testing set, while retaining a high accuracy in the Training set. However, the precision falls from 98%+ to below 90%.
- The Pruned Random Forest has a lower accuracy and precision in the Training dataset, but the Δ is the smallest between both the Training and Testing data sets
- The Recall was very high for Not Cancelled, which indicates that it could find those values well.

	Training	Testing
Not Pruned	**accuracy: 98.47%** 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%	**accuracy: 87.94%** 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%
Pruned	**accuracy: 87.32%** 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%	**accuracy: 85.04%** 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%

Model Performance Summary

- Overview of the final ML model and its parameters
  - The Pruned Random Forest created 80 trees in 3:29 minutes
  - Lead Time had the greatest weight, which also caused it to be the Root Node in majority of scenarios
  - Other Attributes that had a weight of significance (>.005) had a tree created where it was the Root Node
- Summary of most important features used by the ML model for prediction
  - The Random Forest had a smaller Δ between Testing and Training Group Accuracy, which means we are minimizing the Overfitting of the model

RapidMiner File - Hotel Bookin [...] on (Random Forest - Pruned).rmp (5 results. Process results)
Completed: Mar 7, 2024 1:20:55 PM (execution time: 3:29)



***Note**: You can use more than one slide if needed*

APPENDIX

Data Background and Contents

- Please mention the data background and contents
- The data consistent in this assingment is the Customer booking details for INN Hotels Group
- The data is distinct on the Booking ID as its Primary Key
- It also highlights other details such as the number of Adults and Children, when the Booking occurred, and whether previous Cancellations had occurred

Model Building - Decision Tree / Random Forest

- Comment on the model performance of Decision Tree / Random Forest

- The best model for our objective is the Pruned Random Forest because it has the lowest variability in Accuracy, Recall, and Precision amogst the 4 models. Specifically, we want to limit the variance in Recall because we want to correctly yield Cancellation data.
- We were able to come to this conclusion because we can measure and change the weights of our Attributes in the model, and in the case of the Pruned Random Forest, aggregate the data

***Note**: You can use more than one slide if needed*

Slide Header

- Please add any other pointers (if needed)

Great Learning

Happy Learning !