GREG: R OUTPUT AND COMMENTARY FOR FINAL PROJECT

RANDOM FORESTS

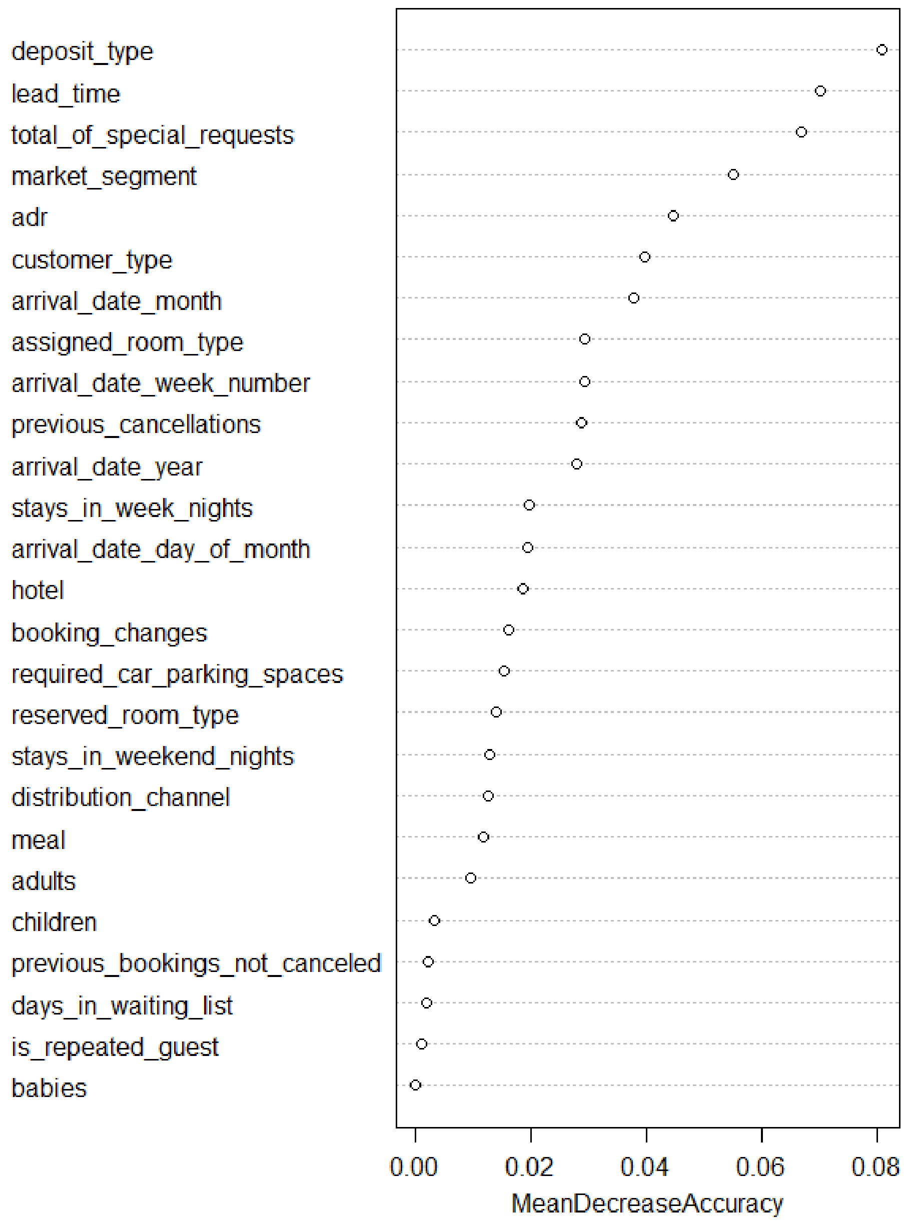
We first applied random forests to the original dataset. We dropped the country, agent, company, and reservation_status_date columns because randomForest in R does not handle categorical variables with more than 53 levels. We also dropped reservation_status because it was the same as the is_canceled response variable. Just 4 observations with missing values were dropped as well. We then refit random forests with subsets of the important variables to see the changes in accuracy.

After analyzing the combined data, we applied random forests to the two hotel types separately to discover their differences. We also produced some two-way frequency tables to examine important variables for the combined and separated datasets. This was another good way to find differences between the resort hotel and the city hotel.

Finally, we applied random forests using hotel type as the response variable (as in the Nest data homework) to identify where misclassifications were occurring.

Combined data random forests result:

```
> hotel1.rf$confusion
      0     1 class.error
0 70179  4987  0.06634649
1 10065 34155  0.22761194
> class.sum(hotel1$is_canceled,predict(hotel1.rf,type="prob")[,2])
 [,1]                               [,2]
 "Percent Correctly Classified = " "87.39"
 "Specificity = "                  "93.31"
 "Sensitivity = "                  "77.33"
 "Kappa ="                         NA
 "AUC= "                           "0.9365"
```



The overall accuracy for random forests is very good, but we have less than ideal sensitivity. We chose to refit random forests using the top 7 variables, and then using the top 4 variables.

7-variable model:

```
> hotel1.rf2$confusion
      0     1 class.error
0 69943  5223   0.0694862
1 16709 27511   0.3778607
> class.sum(hotel1$is_canceled,predict(hotel1.rf2,type="prob")[,2])
 [,1]                               [,2]
 "Percent Correctly Classified = " "81.64"
 "Specificity = "                  "93.04"
 "Sensitivity = "                  "62.26"
 "Kappa ="                         NA
 "AUC= "                           "0.8757"
```

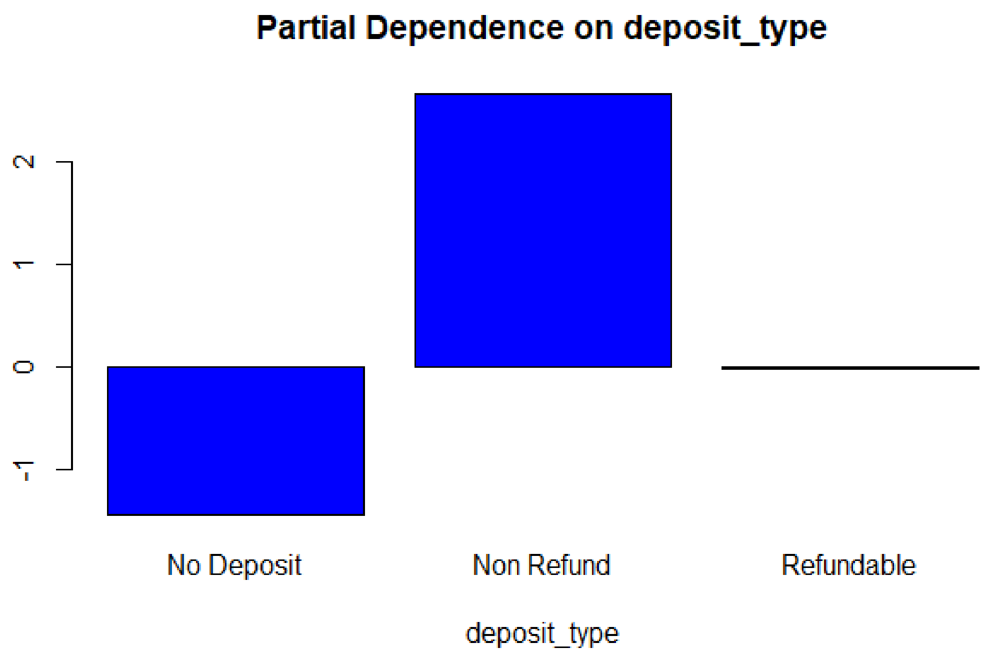
The overall accuracy decreased by about 6 percentage points, and the sensitivity decreased drastically.

4-variable model:

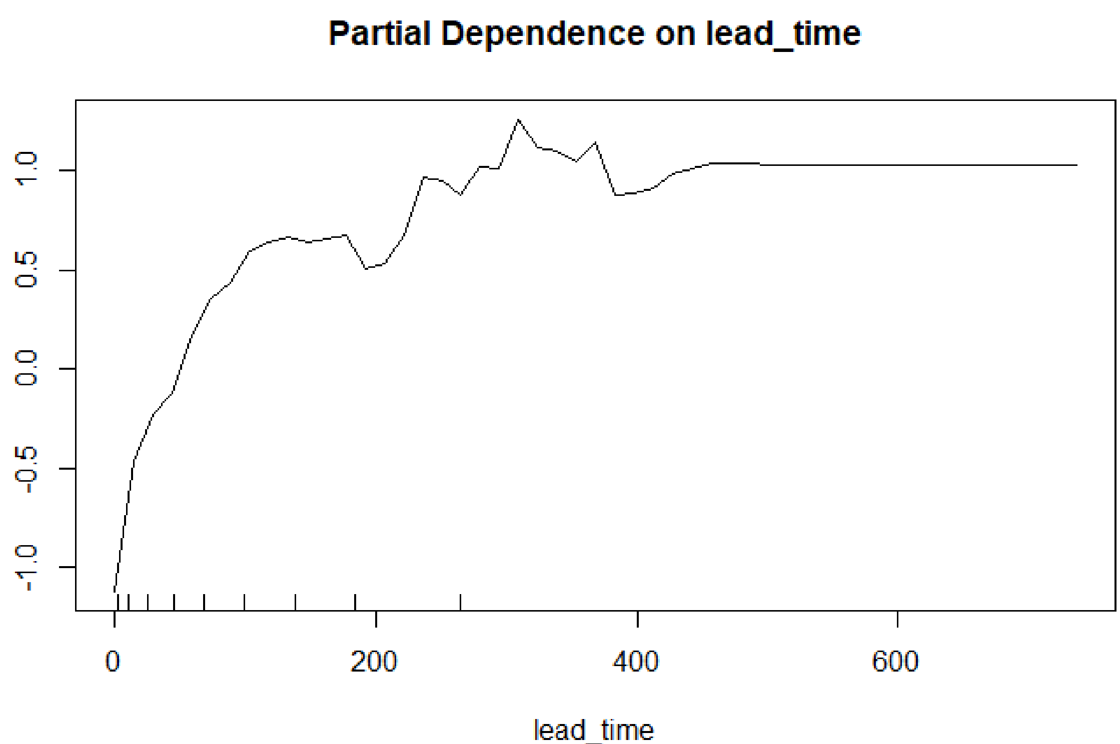
```
> hotel1.rf3$confusion
      0     1 class.error
0 68243  6923  0.09210281
1 17952 26268  0.40597015
> class.sum(hotel1$is_canceled,predict(hotel1.rf3,type="prob")[,2])
 [,1]                               [,2]
 "Percent Correctly Classified = " "79.16"
 "Specificity = "                  "90.78"
 "Sensitivity = "                  "59.43"
 "Kappa ="                         NA
 "AUC= "                           "0.7972"
```

Each metric decreased further by a few percentage points.

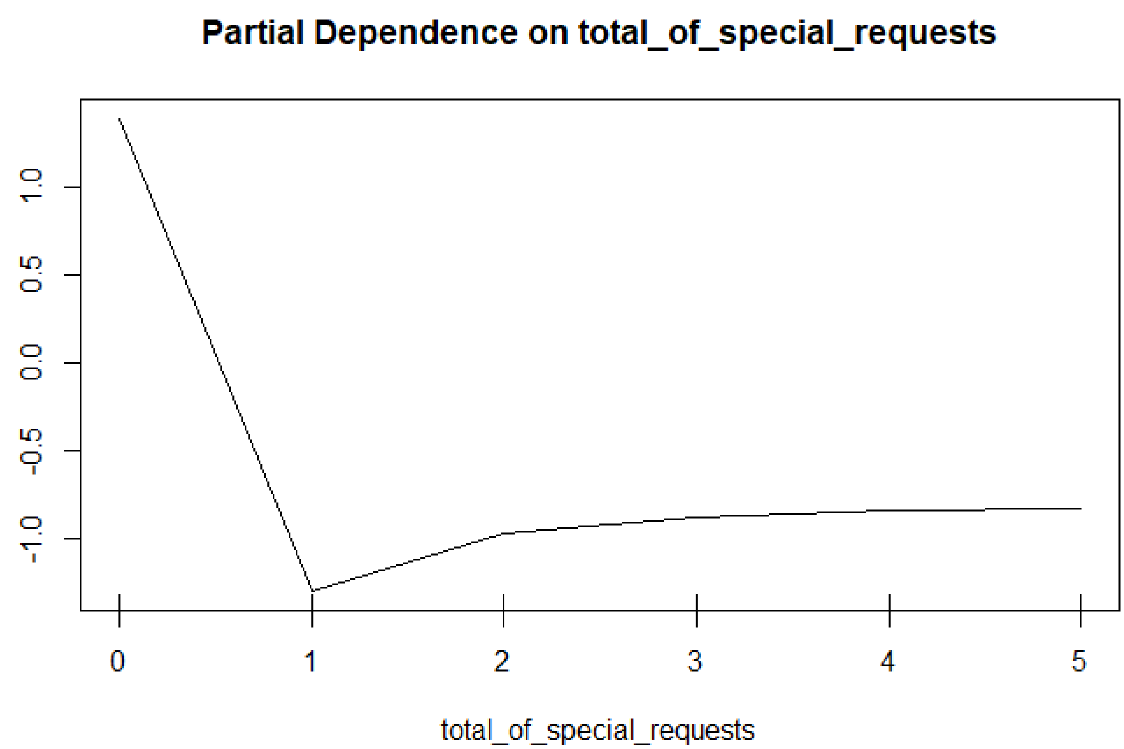
Partial Dependence Plots for Combined Data (Top 8 Variables, and Hotel):



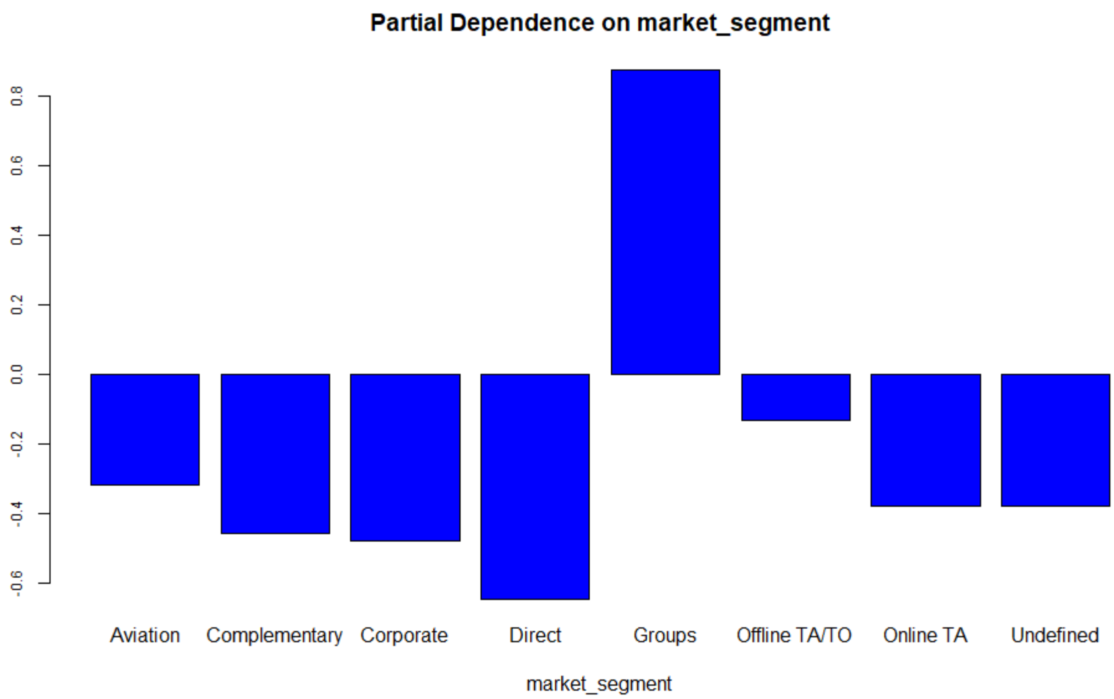
Bookings were much more likely to be canceled if deposit_type was Non Refund.



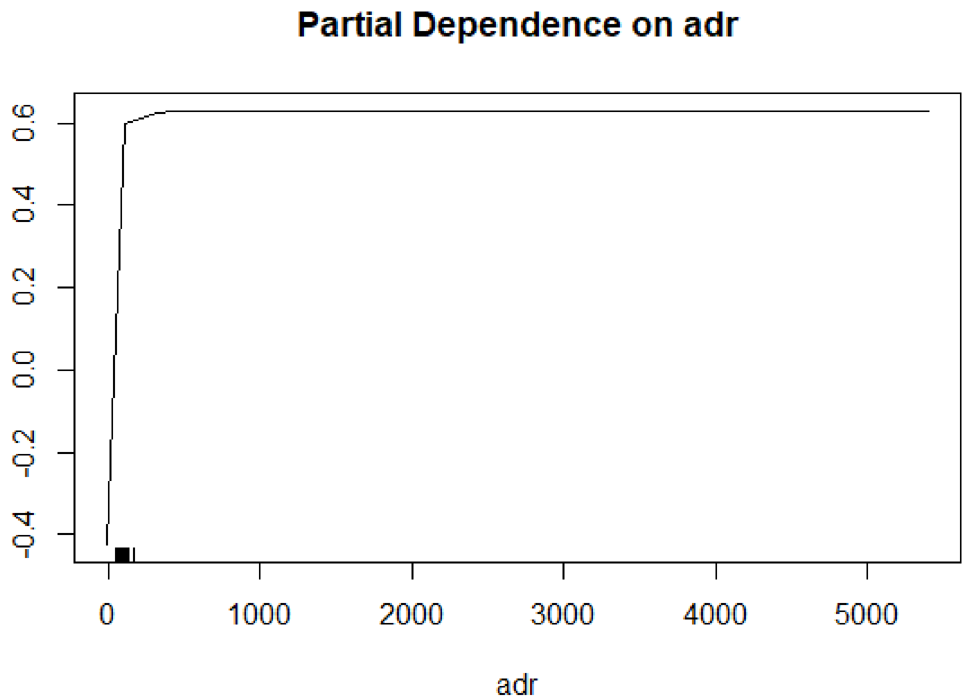
Cancelation was more likely for higher lead times between reservation and arrival. The likelihood changes most between about 0 and 100 days.



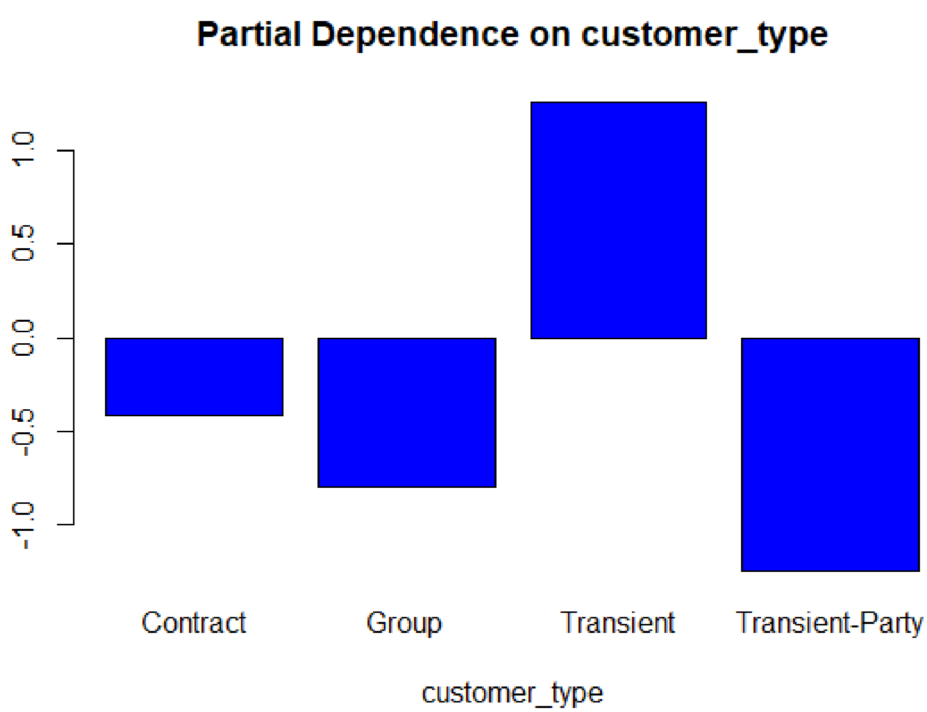
Cancelation dropped drastically where at least one special request was made.



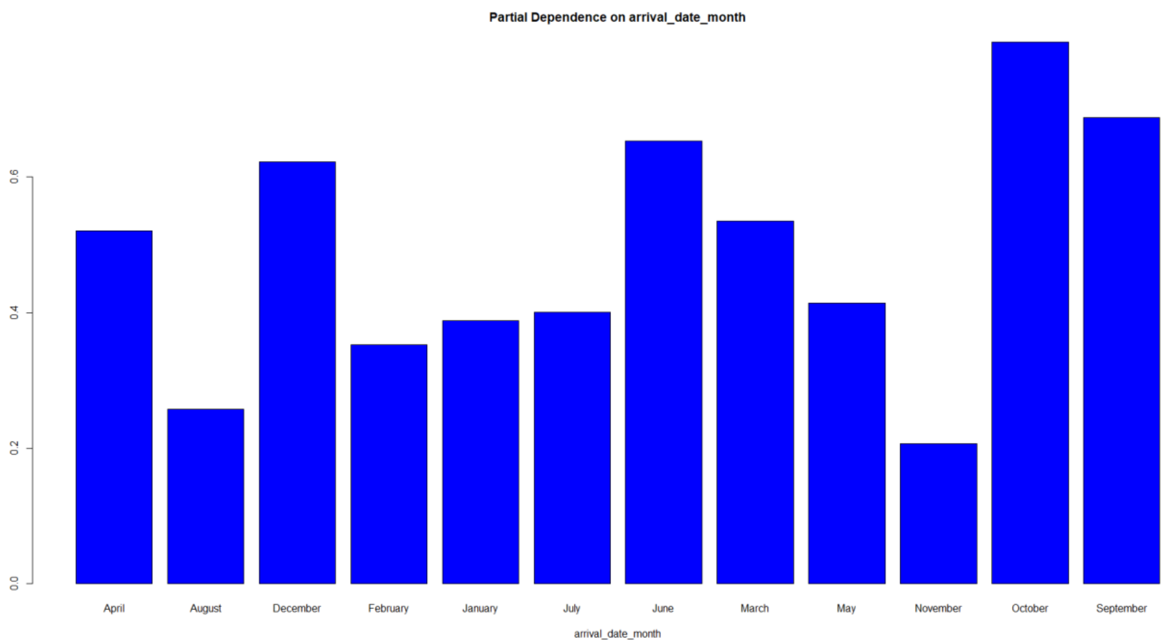
Groups were much more likely to cancel than any other market segment.



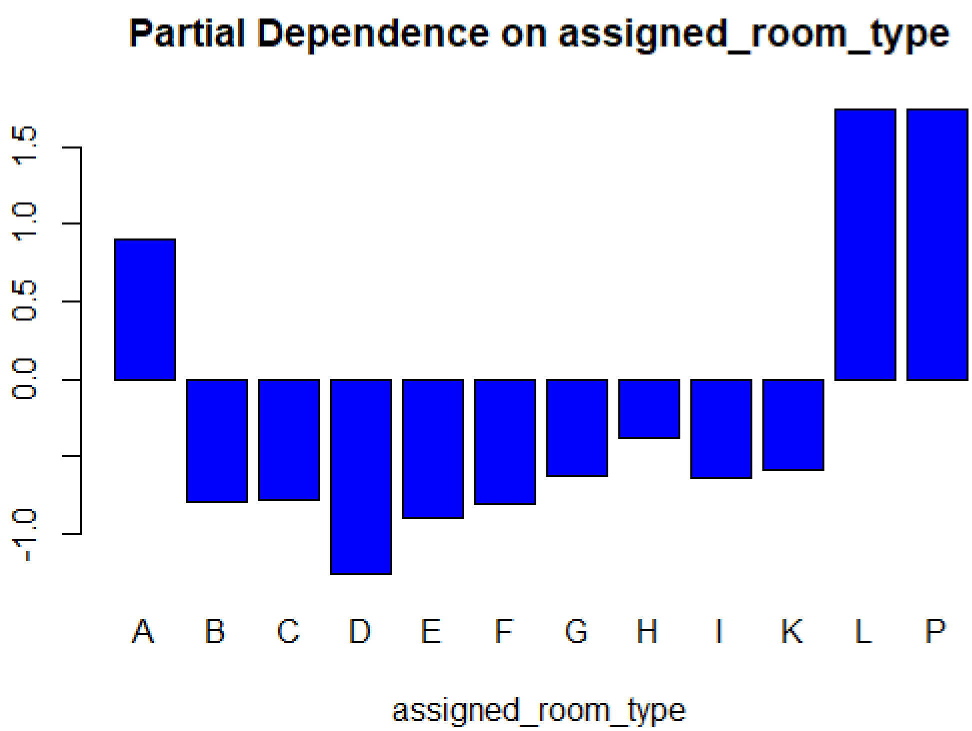
Cancelation was much more likely beyond an average daily rate of about 100 to 200.



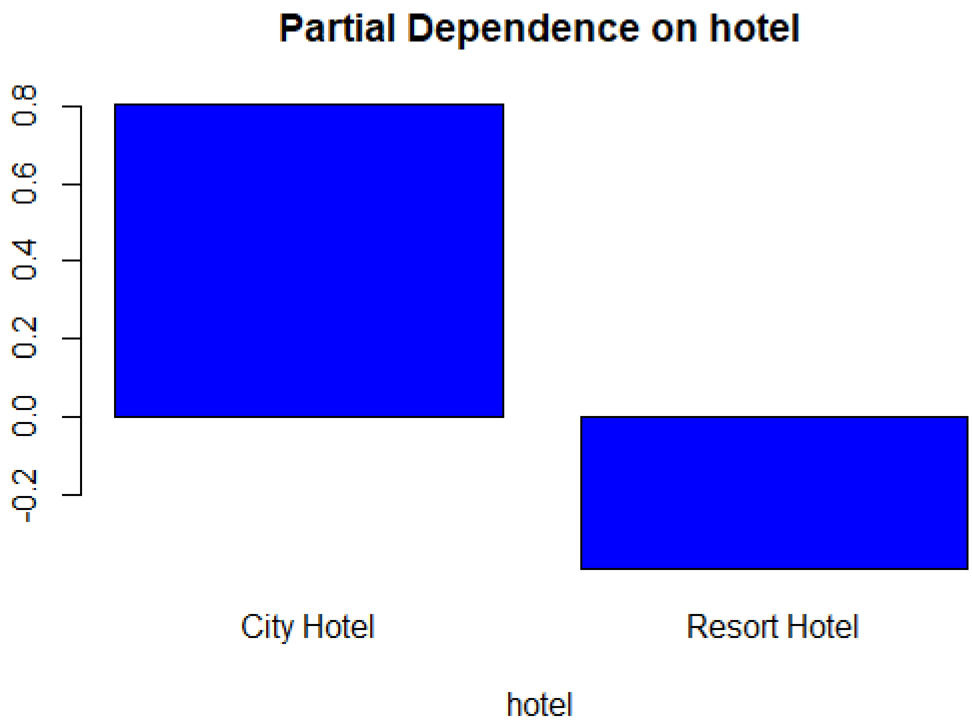
Transient customers were more likely to cancel – this was also shown in classification trees.



For the combined data, cancelations were more likely in October.



Room type A was more likely to be cancelled, and there were very few observations with room types L and P, which were all cancelled.

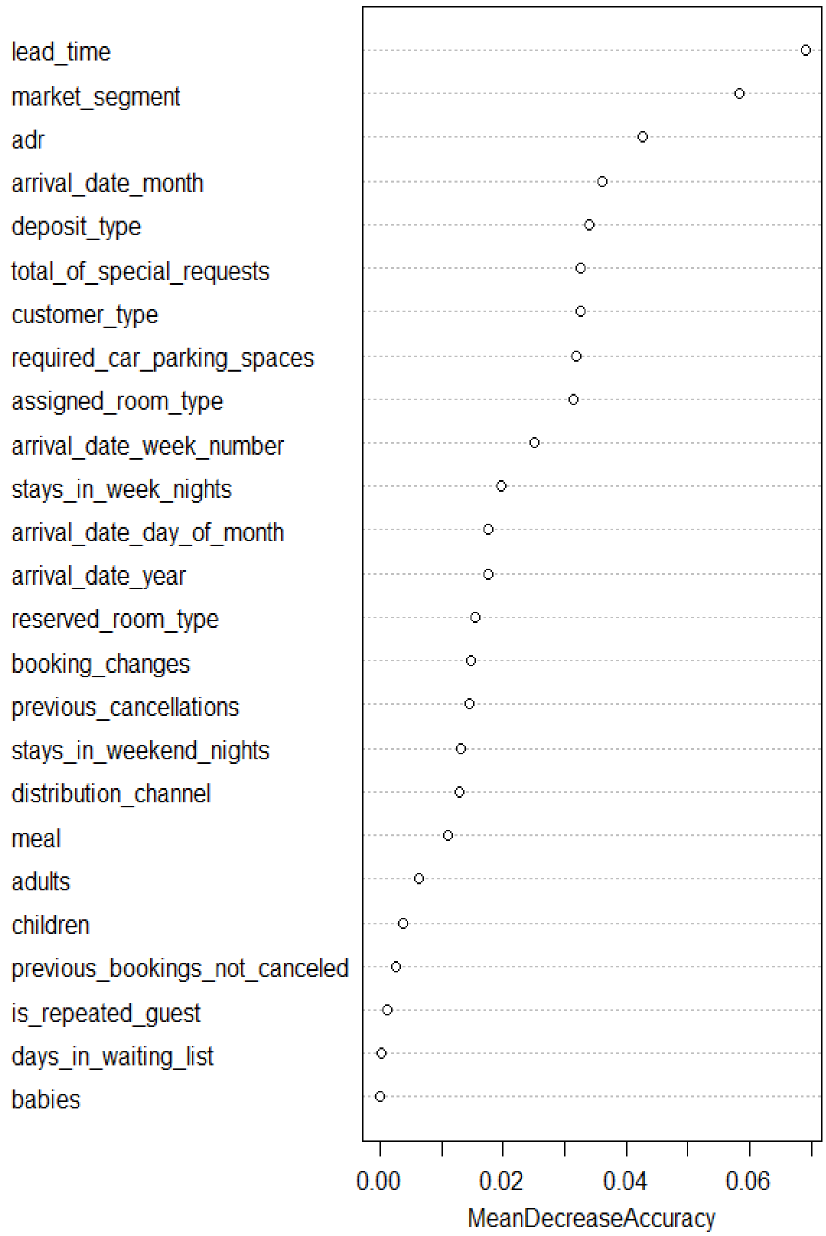


The city hotel had a higher cancelation rate than the resort hotel.

Separate Hotel Types Random Forests Results:

Resort hotel random forests result:

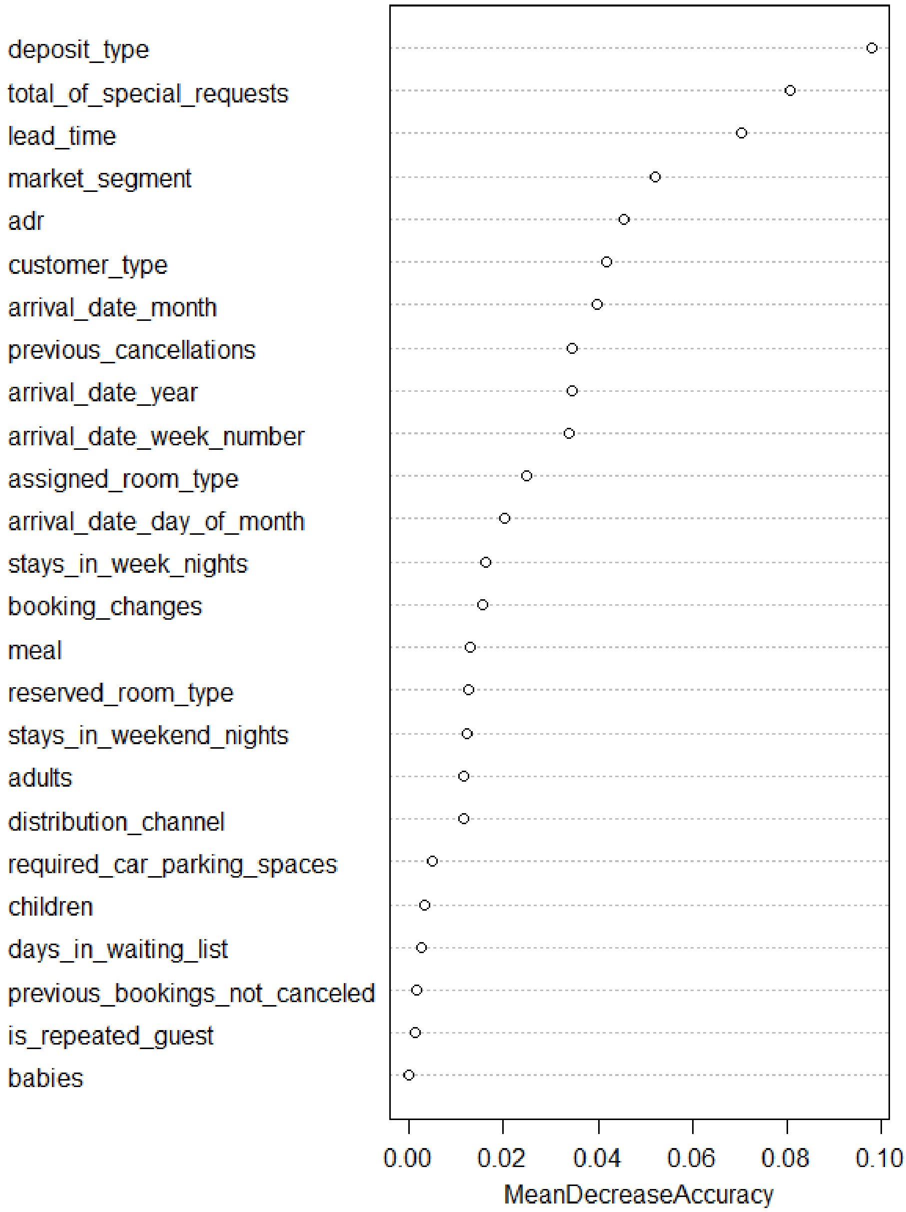
```
> ResortHotel.rf$confusion
      0    1 class.error
0 27265 1673  0.05781326
1  3256 7866  0.29275310
> class.sum(ResortHotel$is_canceled,predict(ResortHotel.rf,type="prob")[,2])
 [,1]                              [,2]
 "Percent Correctly Classified = " "87.71"
 "Specificity = "                  "94.17"
 "Sensitivity = "                  "70.89"
 "Kappa ="                         "0.6797"
 "AUC= "                           "0.9282"
```



Random forests on the resort hotel had the highest overall accuracy and specificity, but lower sensitivity. The important variables shifted around a lot, with lead_time now at the top and deposit_type fifth. This suggests the resort hotel is different from the city hotel.

City hotel random forests result:

```
> CityHotel.rf$confusion
      0     1 class.error
0 42825  3403  0.07361339
1  6558 26540  0.19813886
> class.sum(CityHotel$is_canceled,predict(CityHotel.rf,type="prob")[,2])
 [,1]                                [,2]
 "Percent Correctly Classified = " "87.45"
 "Specificity = "                  "92.59"
 "Sensitivity = "                  "80.26"
 "Kappa ="                         NA
 "AUC= "                           "0.9389"
```



We got slightly higher accuracy and sensitivity with the city hotel, and the most important variables stayed relatively the same as the combined data.

Frequency Tables for Important Variables

Cancelation by hotel type:

Cancelation by Hotel Type

The FREQ Procedure

Frequency / Percent / Row Pct

Table of hotel by is_canceled

hotel	is_canceled: 0	1	Total
City Hotel	46228 38.72 58.27	33102 27.73 41.73	79330 66.45
Resort Hotel	28938 24.24 72.24	11122 9.32 27.76	40060 33.55
Total	75166 62.96	44224 37.04	119390 100.00

The two hotels had very different overall cancelation rates (27.76% versus 41.73%), which is good evidence for treating them separately.

deposit_type:

Cancelation by Important Predictors:
deposit_type
Combined Data

The FREQ Procedure

Frequency / Percent / Row Pct

Table of deposit_type by is_canceled

deposit_type	is_canceled: 0	1	Total
No Deposit	74947 62.77 71.62	29694 24.87 28.38	104641 87.65
Non Refund	93 0.08 0.64	14494 12.14 99.36	14587 12.22
Refundable	126 0.11 77.78	36 0.03 22.22	162 0.14
Total	75166 62.96	44224 37.04	119390 100.00

Cancelation by Important Predictors:
deposit_type
Resort Hotels

The FREQ Procedure

Frequency / Percent / Row Pct

Table of deposit_type by is_canceled

deposit_type	is_canceled: 0	1	Total
No Deposit	28749 71.76 75.26	9450 23.59 24.74	38199 95.35
Non Refund	69 0.17 4.01	1650 4.12 95.99	1719 4.29
Refundable	120 0.30 84.51	22 0.05 15.49	142 0.35
Total	28938 72.24	11122 27.76	40060 100.00

Cancelation by Important Predictors:
deposit_type
City Hotels

The FREQ Procedure

Frequency / Percent / Row Pct

Table of deposit_type by is_canceled

deposit_type	is_canceled: 0	1	Total
No Deposit	46198 58.24 69.53	20244 25.52 30.47	66442 83.75
Non Refund	24 0.03 0.19	12844 16.19 99.81	12868 16.22
Refundable	6 0.01 30.00	14 0.02 70.00	20 0.03
Total	46228 58.27	33102 41.73	79330 100.00

- **In all three cases, the Non Refund deposit type had extremely high cancelation rates. This was surprising because Non Refund indicates a deposit was made in the value of the total stay cost. One explanation may be that only 4.29% of resort hotel bookings were Non Refund, and only 16.22% of city hotel bookings were Non Refund, so it was a less typical option.**
- **This also highlights a difference between the resort hotel and the city hotel. As shown in the variable importance plots, deposit_type is less important for the resort hotel because over 95% of resort bookings were No Deposit, and slightly less Non Refund bookings were canceled.**

lead_time:

Cancelation by Important Predictors: lead_time Combined Data

The FREQ Procedure

Frequency
Percent
Row Pct

Table of LeadTime by is_canceled	is_canceled		
LeadTime	0	1	Total
a: under 15	23324 19.54 87.12	3448 2.89 12.88	26772 22.42
b: 15 and over	51842 43.42 55.97	40776 34.15 44.03	92618 77.58
Total	75166 62.96	44224 37.04	119390 100.00

Cancelation by Important Predictors: lead_time Resort Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of LeadTime by is_canceled	is_canceled		
LeadTime	0	1	Total
a: under 15	10404 25.97 91.26	996 2.49 8.74	11400 28.46
b: 15 and over	18534 46.27 64.67	10126 25.28 35.33	28660 71.54
Total	28938 72.24	11122 27.76	40060 100.00

Cancelation by Important Predictors: lead_time City Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of LeadTime by is_canceled	is_canceled		
LeadTime	0	1	Total
a: under 15	12920 16.29 84.05	2452 3.09 15.95	15372 19.38
b: 15 and over	33308 41.99 52.08	30650 38.64 47.92	63958 80.62
Total	46228 58.27	33102 41.73	79330 100.00

In all three cases, cancelation rates were much lower when reservations were made less than 15 days before arrival. The grouping cutoff of 15 days is based on the lead_time split from the Resort Hotel classification tree. The variable importance plot above also shows lead_time as the resort hotel's most important variable.

market_segment:

Cancelation by Important Predictors: market_segment Combined Data

The FREQ Procedure

Frequency
Percent
Row Pct

Table of market_segment by is_canceled	is_canceled		
market_segment	0	1	Total
Aviation	185 0.15 78.06	52 0.04 21.94	237 0.20
Complementary	646 0.54 86.94	97 0.08 13.06	743 0.62
Corporate	4303 3.60 81.27	992 0.83 18.73	5295 4.44
Direct	10672 8.94 84.66	1934 1.62 15.34	12606 10.56
Groups	7714 6.46 38.94	12097 10.13 61.06	19811 16.59
Offline TA/TO	15908 13.32 65.68	8311 6.96 34.32	24219 20.29
Online TA	35738 29.93 63.28	20739 17.37 36.72	56477 47.30
Undefined	0 0.00 0.00	2 0.00 100.00	2 0.00
Total	75166 62.96	44224 37.04	119390 100.00

Cancelation by Important Predictors: market_segment Resort Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of market_segment by is_canceled	is_canceled		
market_segment	0	1	Total
Complementary	168 0.42 83.58	33 0.08 16.42	201 0.50
Corporate	1958 4.89 84.80	351 0.88 15.20	2309 5.76
Direct	5635 14.07 86.52	878 2.19 13.48	6513 16.26
Groups	3362 8.39 57.61	2474 6.18 42.39	5836 14.57
Offline TA/TO	6334 15.81 84.77	1138 2.84 15.23	7472 18.65
Online TA	11481 28.66 64.76	6248 15.60 35.24	17729 44.26
Total	28938 72.24	11122 27.76	40060 100.00

Cancelation by Important Predictors: market_segment City Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of market_segment by is_canceled	is_canceled		
market_segment	0	1	Total
Aviation	185 0.23 78.06	52 0.07 21.94	237 0.30
Complementary	478 0.60 88.19	64 0.08 11.81	542 0.68
Corporate	2345 2.96 78.53	641 0.81 21.47	2986 3.76
Direct	5037 6.35 82.67	1056 1.33 17.33	6093 7.68
Groups	4352 5.49 31.14	9623 12.13 68.86	13975 17.62
Offline TA/TO	9574 12.07 57.17	7173 9.04 42.83	16747 21.11
Online TA	24257 30.58 62.60	14491 18.27 37.40	38748 48.84
Undefined	0 0.00 0.00	2 0.00 100.00	2 0.00
Total	46228 58.27	33102 41.73	79330 100.00

In all three cases, Groups had the highest cancelation rates. For the resort hotel, the 'Online TA' segment had a higher than average cancelation rate as well.

total_of_special_requests:

Cancelation by Important Predictors:
total_of_special_requests
Combined Data

The FREQ Procedure

Frequency
Percent
Row Pct

Table of total_of_special_requests by is_canceled	is_canceled		
total_of_special_requests	0	1	Total
0	36762 30.79 52.28	33556 28.11 47.72	70318 58.90
1	25908 21.70 77.98	7318 6.13 22.02	33226 27.83
2	10103 8.46 77.90	2866 2.40 22.10	12969 10.86
3	2051 1.72 82.14	446 0.37 17.86	2497 2.09
4	304 0.25 89.41	36 0.03 10.59	340 0.28
5	38 0.03 95.00	2 0.00 5.00	40 0.03
Total	75166 62.96	44224 37.04	119390 100.00

Cancelation by Important Predictors:
total_of_special_requests
Resort Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of total_of_special_requests by is_canceled	is_canceled		
total_of_special_requests	0	1	Total
0	15145 37.81 67.73	7216 18.01 32.27	22361 55.82
1	9209 22.99 78.00	2597 6.48 22.00	11806 29.47
2	3700 9.24 76.65	1127 2.81 23.35	4827 12.05
3	744 1.86 81.76	166 0.41 18.24	910 2.27
4	127 0.32 89.44	15 0.04 10.56	142 0.35
5	13 0.03 92.86	1 0.00 7.14	14 0.03
Total	28938 72.24	11122 27.76	40060 100.00

Cancelation by Important Predictors:
total_of_special_requests
City Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of total_of_special_requests by is_canceled	is_canceled		
total_of_special_requests	0	1	Total
0	21617 27.25 45.08	26340 33.20 54.92	47957 60.45
1	16699 21.05 77.96	4721 5.95 22.04	21420 27.00
2	6403 8.07 78.64	1739 2.19 21.36	8142 10.26
3	1307 1.65 82.36	280 0.35 17.64	1587 2.00
4	177 0.22 89.39	21 0.03 10.61	198 0.25
5	25 0.03 96.15	1 0.00 3.85	26 0.03
Total	46228 58.27	33102 41.73	79330 100.00

Cancelation decreased as special requests increased for the combined data as well as for the separate hotel types.

arrival_date_month:

Cancelation by Important Predictors: arrival_date_month Combined Data

The FREQ Procedure

Frequency
Percent
Row Pct

Table of arrival_date_month by is_canceled

arrival_date_month	is_canceled 0	is_canceled 1	Total
Apri	6565 5.50 59.20	4524 3.79 40.80	11089 9.29
Augu	8638 7.24 62.25	5239 4.39 37.75	13877 11.62
Dece	4409 3.69 65.03	2371 1.99 34.97	6780 5.68
Febr	5372 4.50 66.58	2696 2.26 33.42	8068 6.76
Janu	4122 3.45 69.52	1807 1.51 30.48	5929 4.97
July	7919 6.63 62.55	4742 3.97 37.45	12661 10.60
June	6404 5.36 58.54	4535 3.80 41.46	10939 9.16
Marc	6645 5.57 67.85	3149 2.64 32.15	9794 8.20
May	7114 5.96 60.33	4677 3.92 39.67	11791 9.88
Nove	4672 3.91 68.77	2122 1.78 31.23	6794 5.69
Octo	6914 5.79 61.95	4246 3.56 38.05	11160 9.35
Sept	6392 5.35 60.83	4116 3.45 39.17	10508 8.80
Total	75166 62.96	44224 37.04	119390 100.00

Cancelation by Important Predictors: arrival_date_month Resort Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of arrival_date_month by is_canceled

arrival_date_month	is_canceled 0	is_canceled 1	Total
Apri	2550 6.37 70.66	1059 2.64 29.34	3609 9.01
Augu	3257 8.13 66.55	1637 4.09 33.45	4894 12.22
Dece	2017 5.03 76.17	631 1.58 23.83	2648 6.61
Febr	2308 5.76 74.38	795 1.98 25.62	3103 7.75
Janu	1868 4.66 85.18	325 0.81 14.82	2193 5.47
July	3137 7.83 68.60	1436 3.58 31.40	4573 11.42
June	2038 5.09 66.93	1007 2.51 33.07	3045 7.60
Marc	2573 6.42 77.13	763 1.90 22.87	3336 8.33
May	2535 6.33 71.23	1024 2.56 28.77	3559 8.88
Nove	1976 4.93 81.08	461 1.15 18.92	2437 6.08
Octo	2577 6.43 72.49	978 2.44 27.51	3555 8.87
Sept	2102 5.25 67.63	1006 2.51 32.37	3108 7.76
Total	28938 72.24	11122 27.76	40060 100.00

Cancelation by Important Predictors: arrival_date_month City Hotels

The FREQ Procedure

Frequency
Percent
Row Pct

Table of arrival_date_month by is_canceled

arrival_date_month	is_canceled 0	is_canceled 1	Total
Apri	4015 5.06 53.68	3465 4.37 46.32	7480 9.43
Augu	5381 6.78 59.90	3602 4.54 40.10	8983 11.32
Dece	2392 3.02 57.89	1740 2.19 42.11	4132 5.21
Febr	3064 3.86 61.71	1901 2.40 38.29	4965 6.26
Janu	2254 2.84 60.33	1482 1.87 39.67	3736 4.71
July	4782 6.03 59.12	3306 4.17 40.88	8088 10.20
June	4366 5.50 55.31	3528 4.45 44.69	7894 9.95
Marc	4072 5.13 63.05	2386 3.01 36.95	6458 8.14
May	4579 5.77 55.62	3653 4.60 44.38	8232 10.38
Nove	2696 3.40 61.88	1661 2.09 38.12	4357 5.49
Octo	4337 5.47 57.03	3268 4.12 42.97	7605 9.59
Sept	4290 5.41 57.97	3110 3.92 42.03	7400 9.33
Total	46228 58.27	33102 41.73	79330 100.00

For the resort hotel, November, December, January, and March had significantly lower cancelation rates than the average of 27.76%. The combined data and the city hotels had less variation in monthly cancelation rates, so it is harder to predict the best times of year for them. This agrees with the variable importance plot for resort hotels, which shows arrival_date_month as more important.

Out-of-Bag Confusion matrix for RF with HOTEL as Response:

```
> hotelchanged.rf$confusion
              City Hotel Not Canceled Resort Hotel class.error
City Hotel         26095         6922           81  0.21158378
Not Canceled        3415        70702         1049  0.05938855
Resort Hotel         231         4114         6777  0.39066715
```

- **Most of the misclassifications were the two hotels being confused for Not Canceled. The hotels were not confused for each other very often. This is further reason to treat the resort hotel and the city hotel differently.**
- **This confusion matrix also shows the lower sensitivity and higher specificity we had with our other random forests fits.**

Random Forests Conclusions:

- **Random Forests gave us very good predictive accuracy for is_canceled compared to the other methods. It identified deposit_type, lead_time, total_of_special_requests, and market_segment among the most important variables for prediction cancelation.**
- **Our reduced models had decreased accuracy, but the 7-variable model was not bad apart from the decreased sensitivitiy.**
- **The resort hotel is different from the city hotel because:**
    1. **It had a much lower overall cancelation proportion.**
    2. **It had different proportions of deposit types.**
    3. **lead_time was a better cancelation predictor for the resort hotel.**
    4. **It had more variation in cancelation rates by month, which made arrival_date_month a better cancelation predictor.**
- **In each case examined, the Groups market segment had the highest cancelation rate.**
- **In each case examined, more special requests resulted in less cancelation.**

ADABOOST:

Combined Data:

```
> table(hotel1$is_canceled,round(hotel1.ada.xvalpr))

        0     1
  0 70701  4465
  1 17590 26630
> class.sum(hotel1$is_canceled,hotel1.ada.xvalpr)
 [,1]                                [,2]
 "Percent Correctly Classified = "  "81.53"
 "Specificity = "                    "94.06"
 "Sensitivity = "                    "60.22"
 "Kappa ="                           NA
 "AUC= "                             "0.8806"
```

Resort Hotel:

```
> table(ResortHotel$is_canceled,round(ResortHotel.ada.xvalpr))

        0     1
  0 27181  1757
  1  5161  5961
> class.sum(ResortHotel$is_canceled,ResortHotel.ada.xvalpr)
 [,1]                                [,2]
 "Percent Correctly Classified = "  "82.73"
 "Specificity = "                    "93.93"
 "Sensitivity = "                    "53.6"
 "Kappa ="                           "0.5247"
 "AUC= "                             "0.8805"
```

City Hotel:

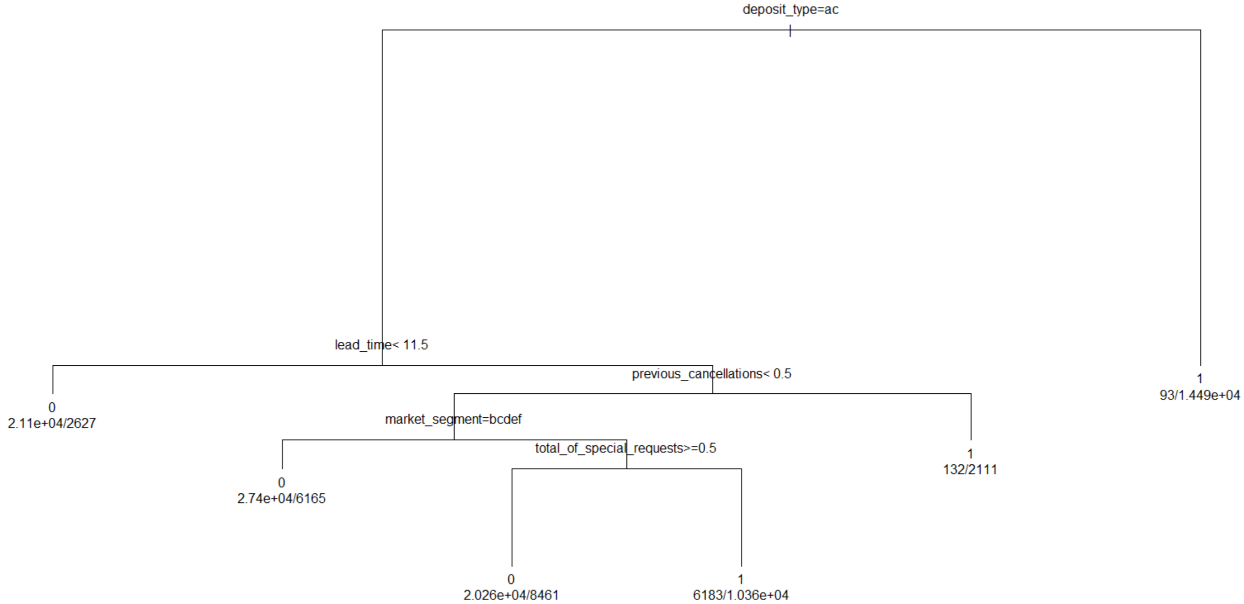
```
> table(CityHotel$is_canceled,round(CityHotel.ada.xvalpr))

        0     1
  0 42873  3355
  1 11378 21720
> class.sum(CityHotel$is_canceled,CityHotel.ada.xvalpr)
 [,1]                                [,2]
 "Percent Correctly Classified = "  "81.43"
 "Specificity = "                    "92.74"
 "Sensitivity = "                    "65.62"
 "Kappa ="                           NA
 "AUC= "                             "0.8866"
```

Adaboost Conclusion:

Adaboost gave similar results for each of our three cases, and it was slightly less accurate than random forests.

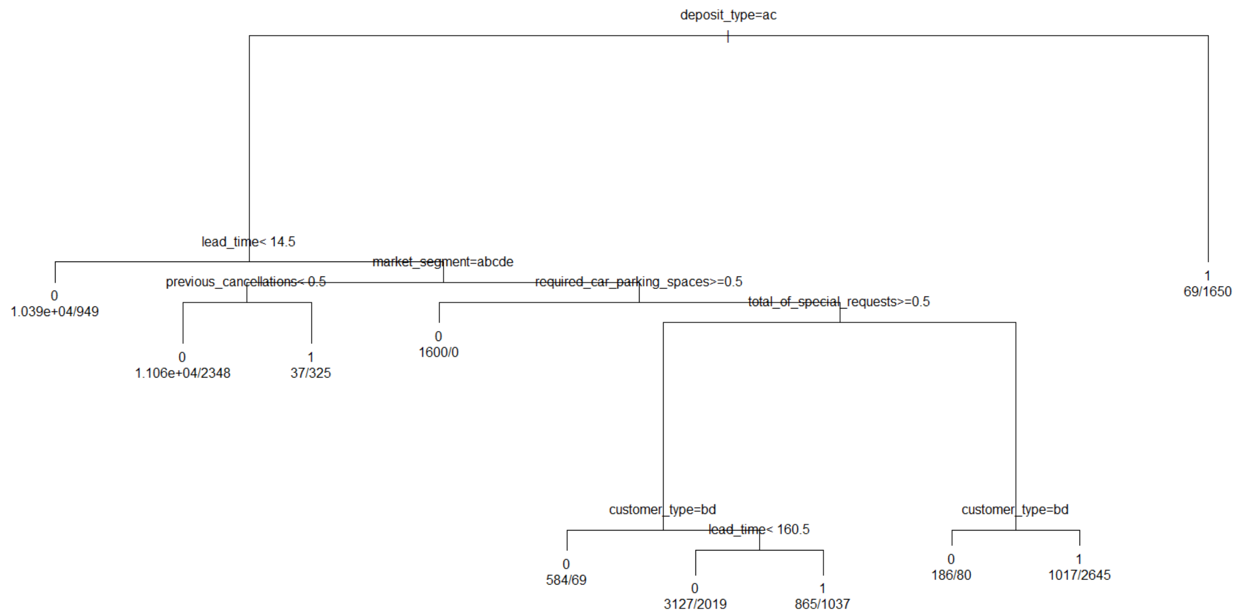
CLASSIFICATION TREES

Combined Data, 5 splits



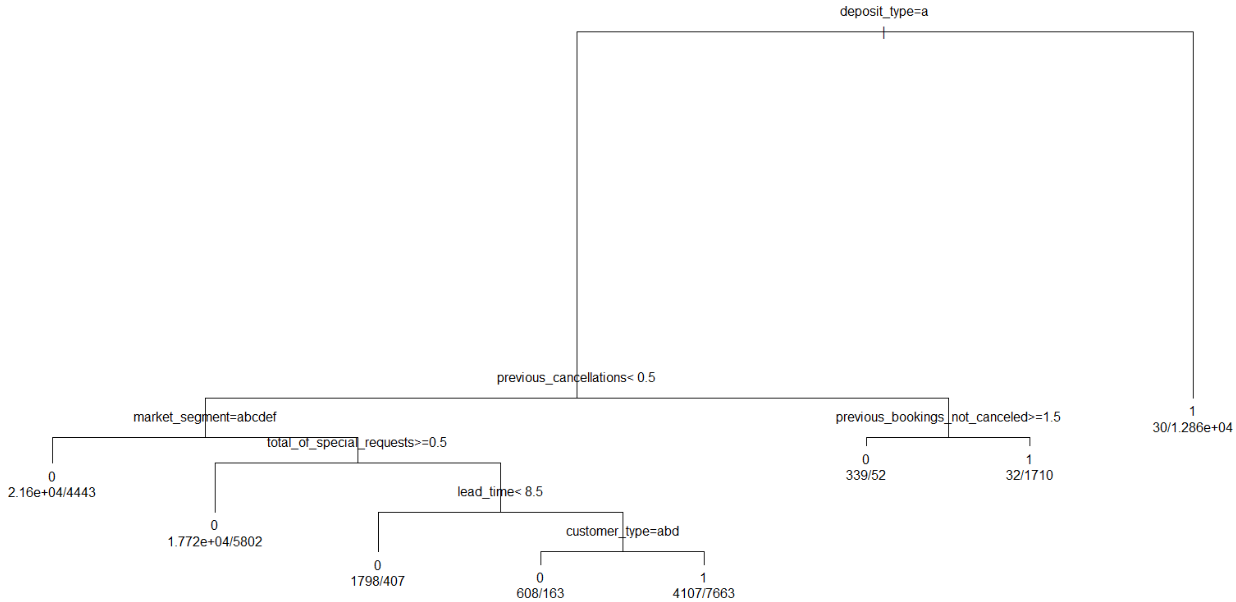

```
> table(hotel1$is_canceled,round(hotel1.rpart5.xval))

        0     1
  0 64738 10428
  1  9765 34455
> class.sum(hotel1$is_canceled,hotel1.rpart5.xval)
 [,1]                               [,2]
 "Percent Correctly Classified = "  "83.07"
 "Specificity = "                   "86.1"
 "Sensitivity = "                   "77.93"
 "Kappa ="                          NA
 "AUC= "                            "0.8244"
```

Resort Hotel Only, 7 splits




```
> table(ResortHotel$is_canceled,round(ResortHotel.rpart7.xval))

        0     1
  0 25362  3576
  1  3312  7810
> class.sum(ResortHotel$is_canceled,ResortHotel.rpart7.xval)
 [,1]                                 [,2]
 "Percent Correctly Classified = "   "82.8"
 "Specificity = "                    "87.61"
 "Sensitivity = "                    "70.27"
 "Kappa ="                           "0.5744"
 "AUC= "                             "0.7901"
```

City Hotel Only, 6 Splits




```
> table(CityHotel$is_canceled,round(CityHotel.rpart6.xval))

        0     1
  0 39421  6807
  1  6363 26735
> class.sum(CityHotel$is_canceled,CityHotel.rpart6.xval)
 [,1]                               [,2]
 "Percent Correctly Classified = "  "83.38"
 "Specificity = "                   "85.23"
 "Sensitivity = "                   "80.79"
 "Kappa ="                          NA
 "AUC= "                            "0.8326"
```

CROSSVALIDATED CONFUSION MATRIX FOR HOTEL AS RESPONSE