Hotels Dataset

MGSC 310 - 02

Arshia Sarma | Betsy Heredia | Kayla Cho | Nawal Alasmari

Motivation

- Determine which factors would affect a hotel business and advance their profit
- Focusing on predicting why customers would cancel their reservation
  - The more guests that cancel = less profit earned
- Useful for hotels to find which factors of their operation to improve on to keep customers from cancelling

  - Profits will increase as more guests stay committed

Exploratory Analysis & Summary Stats

- Dataset consists of information from hotels from Portugal
- Total dataset → 119,390 observations and 32 columns
- Pruned variables agent, company, country, reservation_status, and reservation_status_date
- 37% cancelled their reservation (for is_canceled)
- After pruning, we also upsampled our model with more repeated copies of our minority class

```
Rows: 119,390
Columns: 32
$ hotel                         <fct> Resort Hotel, Resort Hotel, Resort Hotel, Resort Hotel, Resor…
$ is_canceled                   <fct> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ lead_time                     <int> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, 23, 35, 68, 18, 37, 68…
$ arrival_date_year             <int> 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2…
$ arrival_date_month            <fct> July, July, July, July, July, July, July, July, July, July, J…
$ arrival_date_week_number      <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 2…
$ arrival_date_day_of_month     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ stays_in_weekend_nights       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1…
$ stays_in_week_nights          <int> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 1, 1, 4…
$ adults                        <int> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1…
$ children                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0…
$ babies                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ meal                          <fct> BB, BB, BB, BB, BB, BB, BB, FB, BB, HB, BB, HB, BB, HB, BB, B…
$ country                       <fct> PRT, PRT, GBR, GBR, GBR, GBR, PRT, PRT, PRT, PRT, PRT, PRT, U…
$ market_segment                <fct> Direct, Direct, Direct, Corporate, Online TA, Online TA, Dire…
$ distribution_channel          <fct> Direct, Direct, Direct, Corporate, TA/TO, TA/TO, Direct, Dire…
$ is_repeated_guest             <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ previous_cancellations        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ previous_bookings_not_canceled <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ reserved_room_type            <fct> C, C, A, A, A, A, C, C, A, D, E, D, D, G, E, D, E, A, A, G, F…
$ assigned_room_type            <fct> C, C, C, A, A, A, C, C, A, D, E, D, E, G, E, E, E, E, G, G, F…
$ booking_changes               <int> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0…
$ deposit_type                  <fct> No Deposit, No Deposit, No Deposit, No Deposit, No Deposit, N…
$ agent                         <fct> NULL, NULL, NULL, 304, 240, 240, NULL, 303, 240, 15, 240, 240…
$ company                       <fct> NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, N…
$ days_in_waiting_list          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ customer_type                 <fct> Transient, Transient, Transient, Transient, Transient, Transi…
$ adr                           <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00, 107.00, 103.00, 82.00…
$ required_car_parking_spaces   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ total_of_special_requests     <int> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, 1, 0, 3, 0, 0, 0, 1, 1…
$ reservation_status            <fct> Check-Out, Check-Out, Check-Out, Check-Out, Check-Out, Check-…
$ reservation_status_date       <chr> "7/1/15", "7/1/15", "7/2/15", "7/2/15", "7/3/15", "7/3/15", "…
```

Exploratory Analysis & Summary Stats

```
> nlevels(hotels$agent)
[1] 334
>
> nlevels(hotels$company)
[1] 353
>
> nlevels(hotels$country)
[1] 178
```

Total of 865 factors from agent, company and country

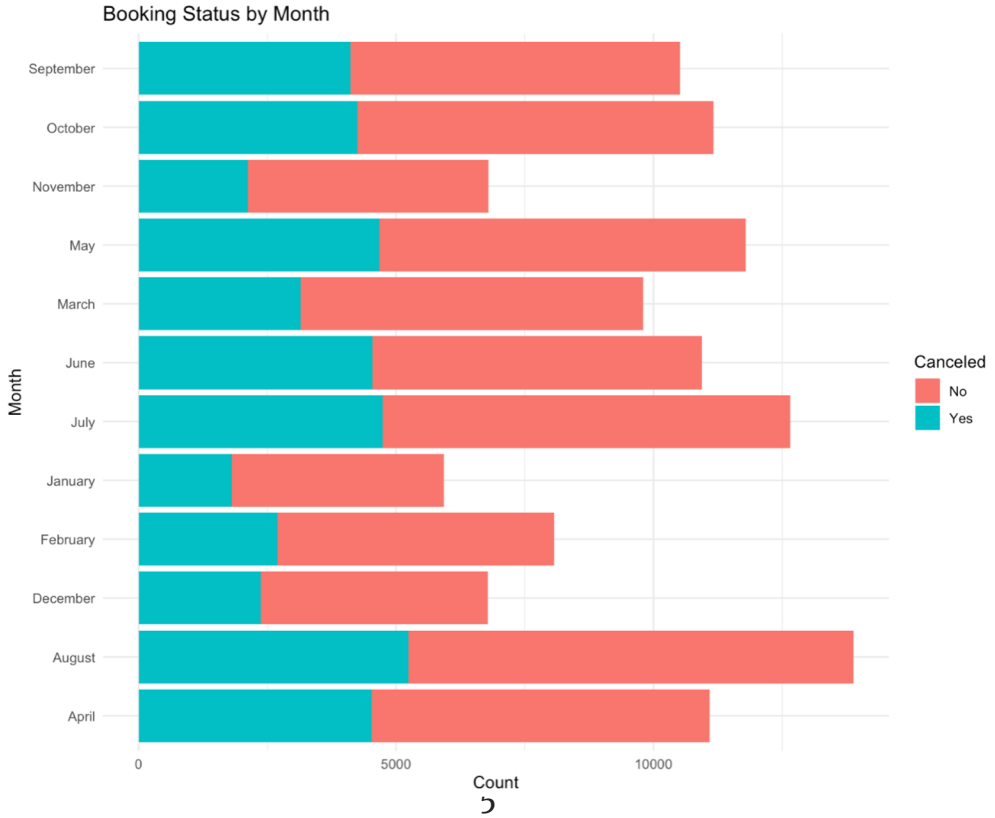
```
is_canceled
0:75166
1:44224
```

```
reservation_status
Canceled :43017
Check-Out:75166
No-Show  : 1207
```

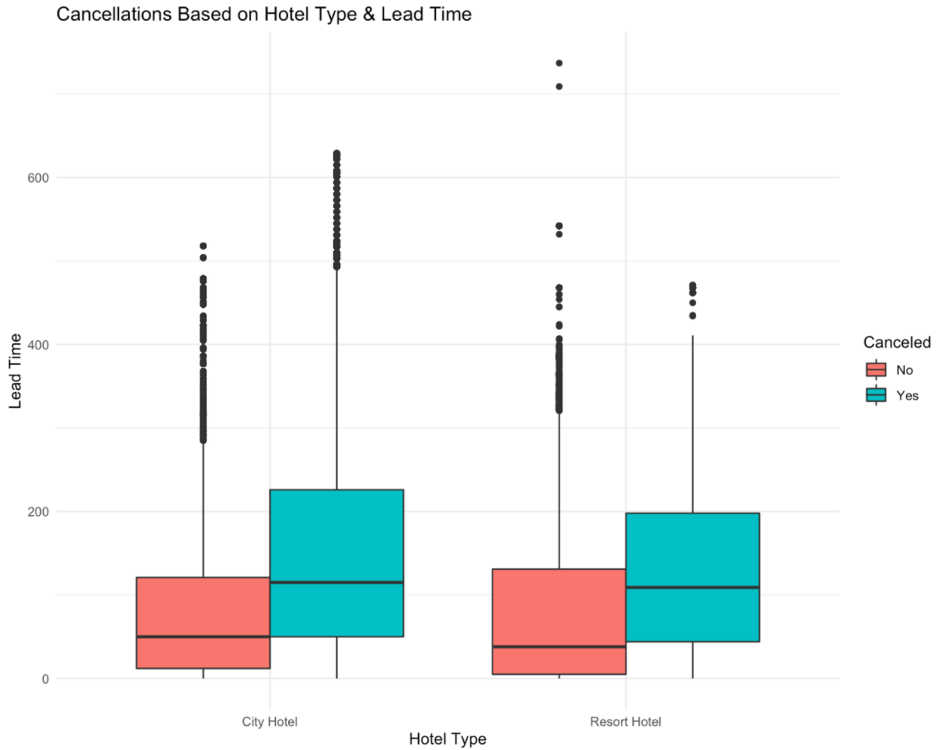
Multicollinearity

```
reservation_status_date
Min.   :2014-10-17
1st Qu.:2016-02-02
Median :2016-08-07
Mean   :2016-07-30
3rd Qu.:2017-02-08
Max.   :2017-09-14
```

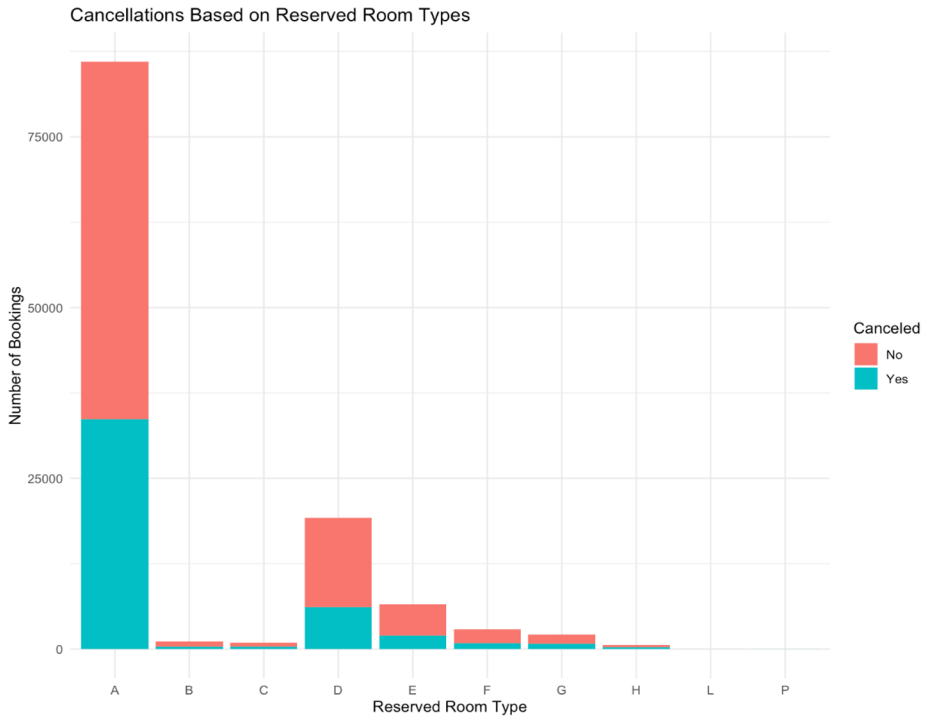
Data Visualizations: Booking Status v. Month



Data Visualizations: Hotel Type v. Lead Time



Data Visualizations: Cancellations v. Room Type



Logistic Model #1

- Predicting the categorical outcome of if a guest canceled their reservation
- Confusion matrix of training and test

Training

		Truth	
PREDICTION		**0**	**1**
	0	41,104	10,603
	1	19,073	49,486

Test

		Truth	
PREDICTION		**0**	**1**
	0	10,175	2,620
	1	4,813	12,457

Training

Accuracy: 75.32%
Sensitivity: 68.31%
Specificity: 82.35%

Test

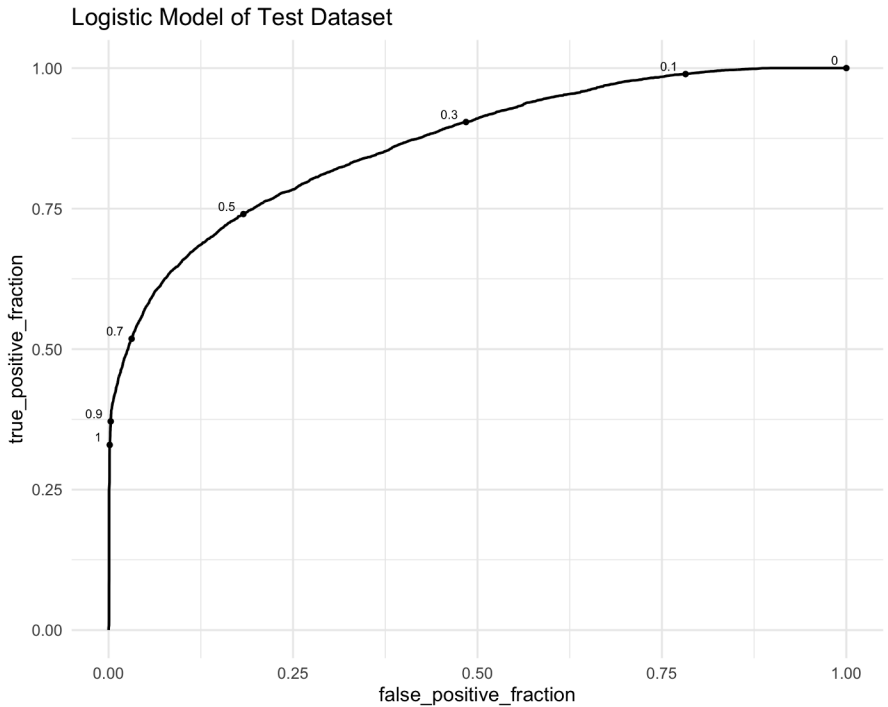
Accuracy: 75.27%
Sensitivity: 67.89%
Specificity: 82.62%

Scores and Takeaways

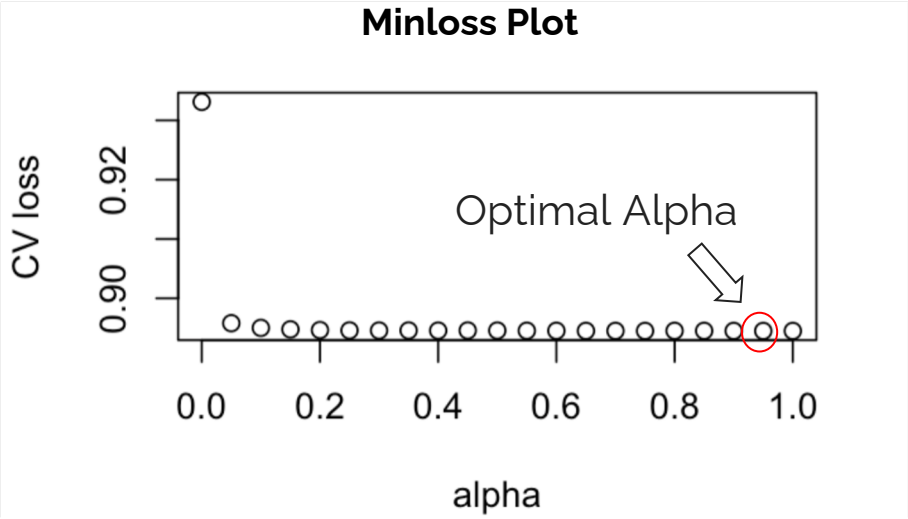
Conclusions

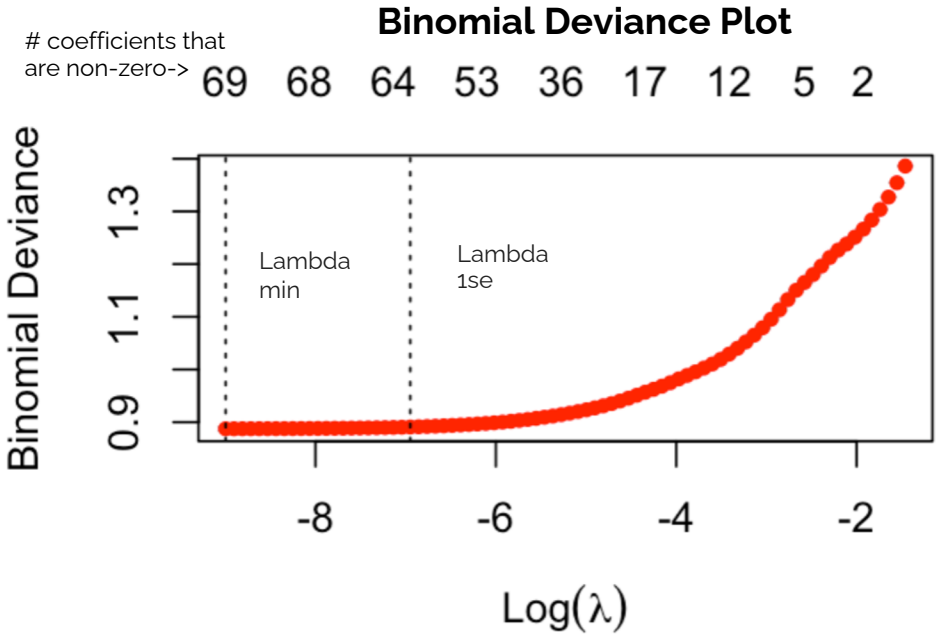
- Not the best model → ~0.26 R2 and average accuracy score
- Accuracy, sensitivity, and specificity of the training and test are similar, as well as their R2 and RMSE

	AUC	**R2**
Training	0.8644	0.2617
Test	0.8649	0.2609



Elastic Net Model





- Optimal Model Parameters:
- Alpha= 0.95

Lambda min: ~69* coefficients

Lambda 1se: ~64* coefficients

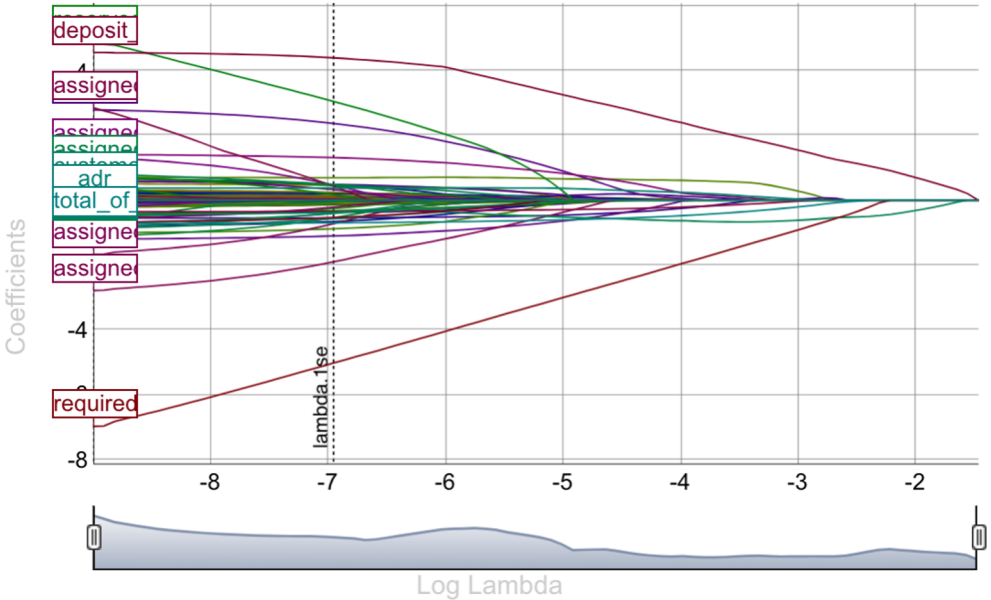
*includes factors

Elastic Net Coefficients

Noteable Coefficients

Required Car Parking Spaces	5.021
Deposit Type- Non Refund	4.394
Reserved Room Type- P	3.048
Previous Cancellations	2.368
Assigned Room Type - I	-1.9

Coefficients Plot



Elastic Net Coefficients Cont.

1. Required car parking spaces
2. Deposit Type
3. Reserved Room Type
4. Previous Cancellations
5. Assigned Room Type
6. Distribution Channel
7. Market Segment
8. Total of Special Requests
9. Meal
10. Customer Type
11. Booking Changes
12. Previous Bookings
13. Is Repeated Guest
14. Adults
15. Arrival Date Month

24 Variables

16. Guests
17. Hotel
18. Babies
19. Children
20. Stays in Week Nights
21. Stays in Weekend Nights
22. Arrival Date Year
23. ADR
24. Lead time

80 Coefficients -> 61 Coefficients

Elastic Net Coefficients Cont.

Non-Important Variables

- **Days in waiting list**
- **Arrival date week number**
- **Arrival date day of month**

Combined Coefficients

- **Reserved_room_type_Other**
  - Reserved_room_type (C,D,H, L)

- Market_Segement_Other

- Market_Segment (Aviation, Groups, Undefined)

- Distribution_Channel_Other

- Distribution_Channel (Corporate, TA/TO, Undefined)

```
hotels_train_clean <- hotels_train %>% mutate(market_segment =
fct_other(market_segment, drop= c("Aviation","Direct", "Groups",
"Undefined")),
                                              reserved_room_type
=fct_other(reserved_room_type, drop= c("C", "D", "H", "L")),
distribution_channel=fct_other(distribution_channel,
drop=c("Corporate", "TA/TO", "Undefined")))


hotels_test_clean <- hotels_test %>% mutate(market_segment =
fct_other(market_segment, drop= c("Aviation","Direct", "Groups",
"Undefined")),
                                              reserved_room_type
=fct_other(reserved_room_type, drop= c("C", "D", "H",
"L")),distribution_channel=fct_other(distribution_channel,
drop=c("Corporate", "TA/TO", "Undefined")))
```

Elastic Net Model Results

Train		Truth	
PREDICTION		0	1
	0	40860	10505
	1	19317	49584

Test		Truth	
PREDICTION		0	1
	0	10135	2578
	1	4853	12499

Conclusions

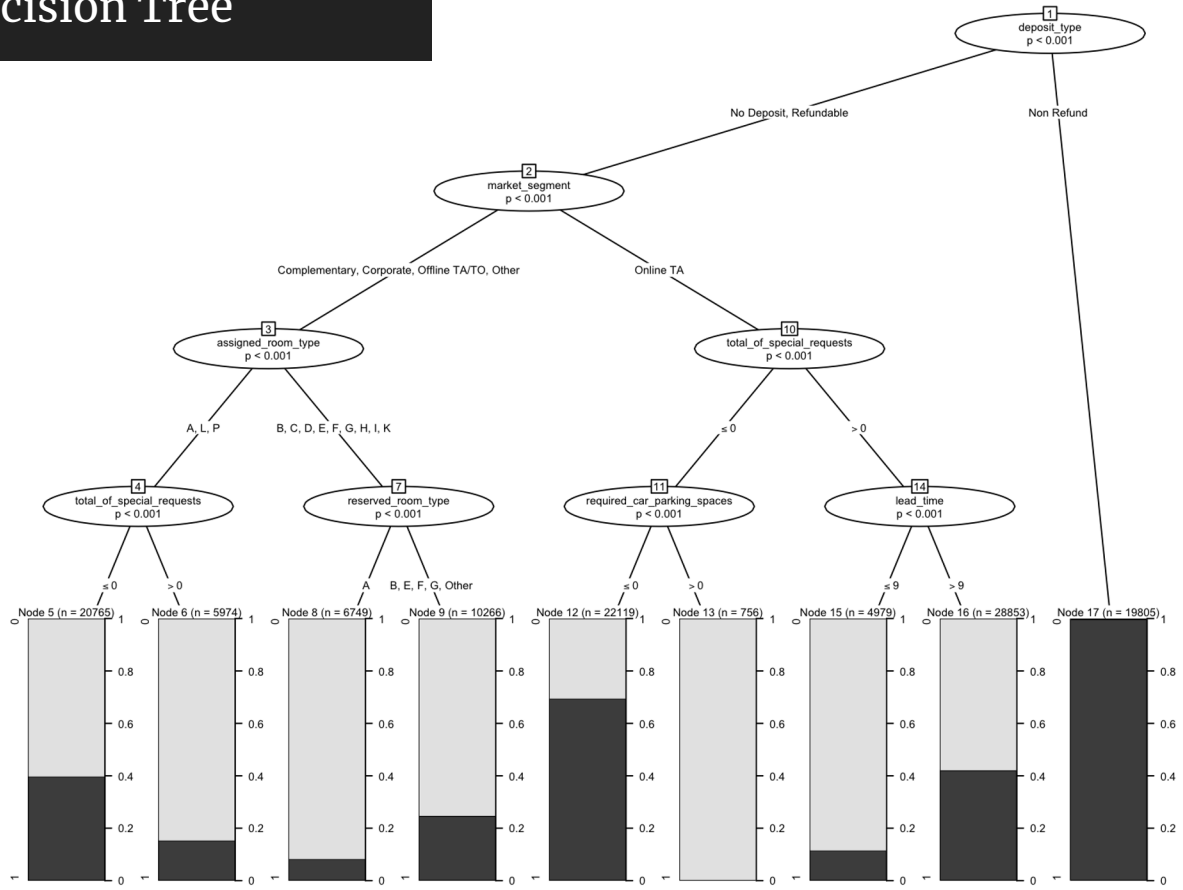
- Similar performance as Logistic Regression (Test: 75.27%)
- Slightly better test performance
- Higher Specificity -> Better at predicting if a booking is not canceled

Accuracy
Train: 75.2033%
Test: 75.2836%

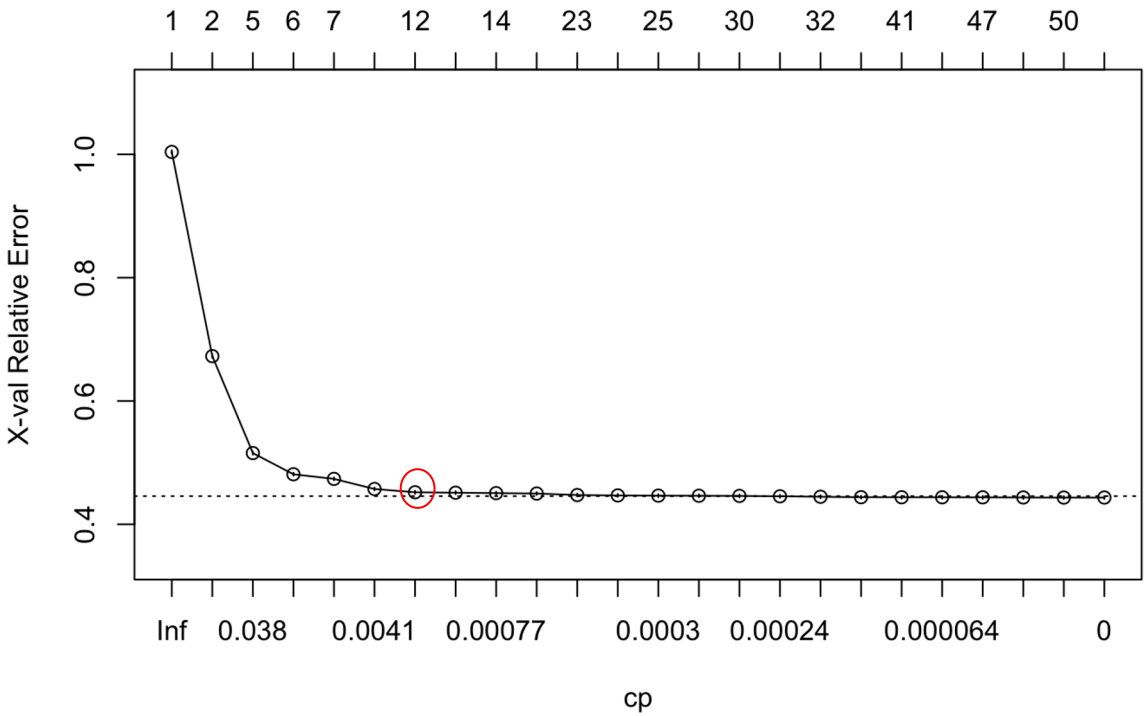
Sensitivity
Train: 67.8997%
Test:67.6208%

Specificity
Train: 82.5176%
Test: 82.9011%

Decision Tree



RPart


1 2 5 6 7 12 14 23 25 30 32 41 47 50
X-val Relative Error
0.4 0.6 0.8 1.0
Inf 0.038 0.0041 0.00077 0.0003 0.00024 0.000064 0
cp

Decision Tree

Max Depth = 4

	RMSE	RSQ
Train	0.5141	0.2443
Test	0.5147	0.2438

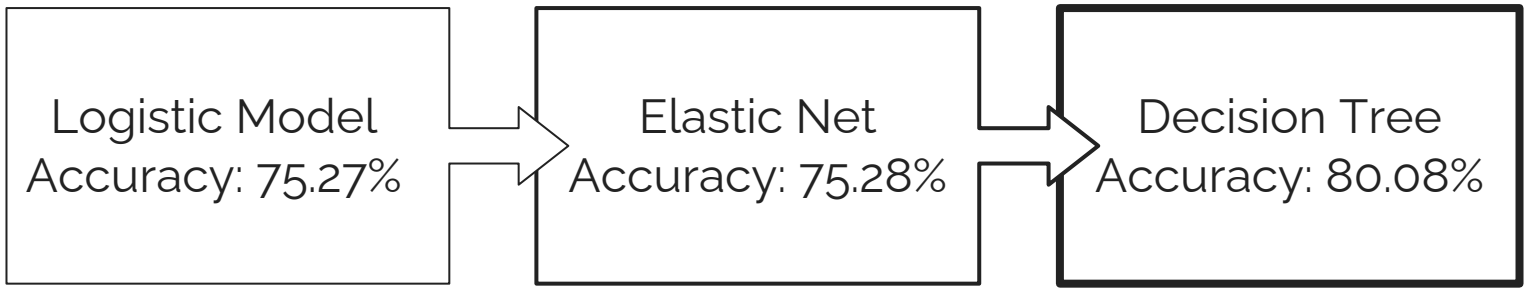
		Truth	
		0	**1**
PREDICTION	**0**	13,280	6,257
	1	1,708	8,820

Max Depth = 12

	RMSE	RSQ
Train	0.4427	0.3698
Test	0.4463	0.3620

		Truth	
		0	**1**
PREDICTION	**0**	11,932	2,932
	1	3,056	12,145

Comparison of Models



Conclusion and Final Takeaways

- Understand your dataset and what variables you want to predict
- Remove outliers, use the entire training sample for the elastic net model and model the outcome variable with a random forest.
- Logistic model with all variables performed better than with top 10 variables.
- Company should focus on reserve room type and people who cancelled their bookings in the past.
- Future directions:
    - Predict market segment
    - Predict distribution channel

Thanks!

ANY QUESTIONS?

Team Hotels

Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash