PREDICTING HOTEL CANCELLATIONS

Letty Uy, Luke Kuller, Rocio Noriega,
Erica Kane & Trang Dinh

INTRODUCTION

For our final project, we decided to focus on and examine the “Hotel Booking Demand” dataset via Kaggle.com. Through this dataset we sought to explore this real-world problem, specifically related to hotel cancellations, which provoked our collective intrigue into providing a viable answer to the hospitality and tourism industry.

DESCRIPTION OF THE DATA-SET & OUR OBJECTIVE

- The "Hotel Booking Demand" dataset features 32 total columns.
- Our chosen dependent variable was the column, Is_Cancelled and we decided to utilize only columns, Distribution_Channel, Previous_Cancellations, Deposit_Type, Lead_Time, and Average_Daily_Rate or "ADR," for our independent variables.
- Our objective was to provide meaningful understanding and resolution to the issue of hotel cancellations.
- Through our findings, we will uncover trends and patterns that will enable us to provide relevant and impactful answers that allow the hospitality and tourism industry to better understand the various questions and aspects related to cancellations.

THE QUESTIONS

The reason why we chose the five independent variables we did, in relation to our decision variable of Is_Cancelled, is because we believed they would produce the strongest results for us to answer our objective of predicting hotel cancellations.
Regarding our independent variables, the questions in which our data will provide answers to are:

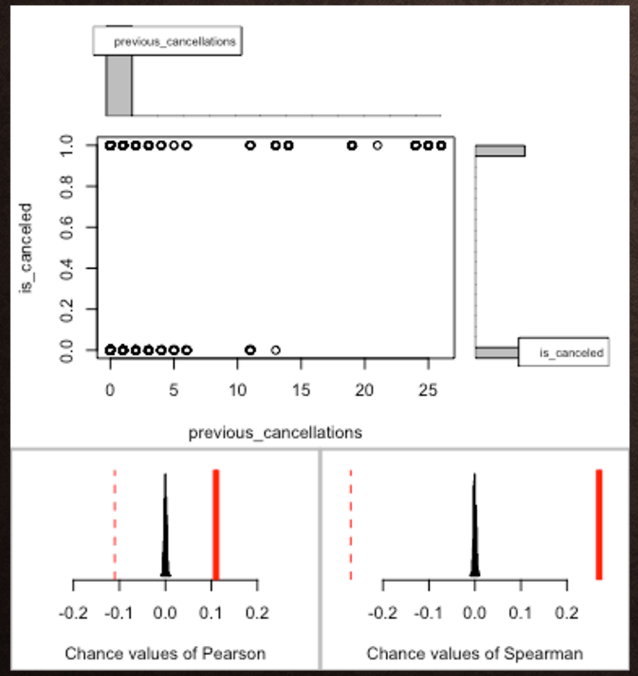
- Which travelers are more likely to cancel?
- Customers track records and if they're prone to canceling in the past.
- If the customer puts down a deposit, are they more likely to not cancel?
- Looking at the high or low lead times and whether we can see what the relationship is with cancellations.
- What ADR is more likely to result in cancellations?

ASSOCIATION ANALYSIS

Association between previous_cancellations (numerical) and is_canceled (converted boolean) using 119390 complete cases
Permutation procedure:

	Value Estimated	p-value
Pearson's r	0.1101328	0
Spearman's rank correlation	0.2702334	0

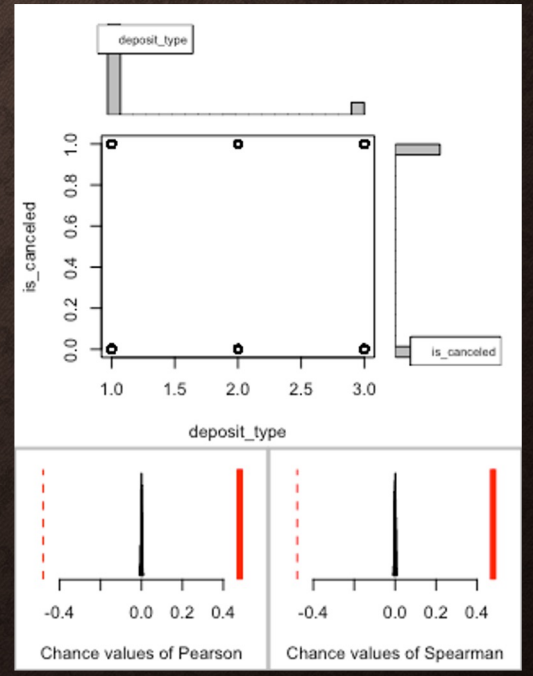
With 500 permutations, we are 95% confident that:
the p-value of Pearson's correlation (r) is between 0 and 0.007
the p-value of Spearman's rank correlation is between 0 and 0.007
Note: If 0.05 is in this range, increase the permutations= argument.



Association between deposit_type (numerical) and is_canceled (converted boolean) using 119390 complete cases
Permutation procedure:

	Value Estimated	p-value
Pearson's r	0.4686338	0
Spearman's rank correlation	0.4770607	0

With 500 permutations, we are 95% confident that:
the p-value of Pearson's correlation (r) is between 0 and 0.007
the p-value of Spearman's rank correlation is between 0 and 0.007
Note: If 0.05 is in this range, increase the permutations= argument.

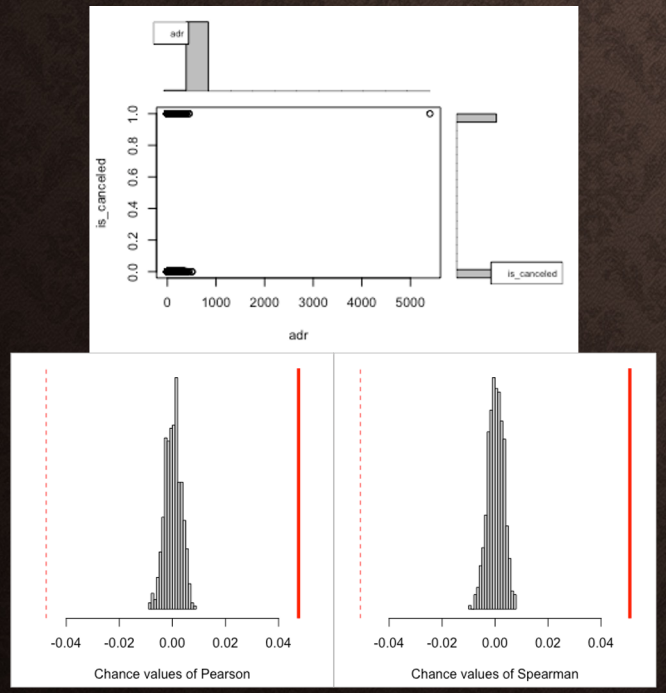


ASSOCIATION ANALYSIS

Association between adr (numerical) and is_canceled (converted boolean) using 119390 complete cases
Permutation procedure:

	Value Estimated	p-value
Pearson's r	0.04755660	0
Spearman's rank correlation	0.05087593	0

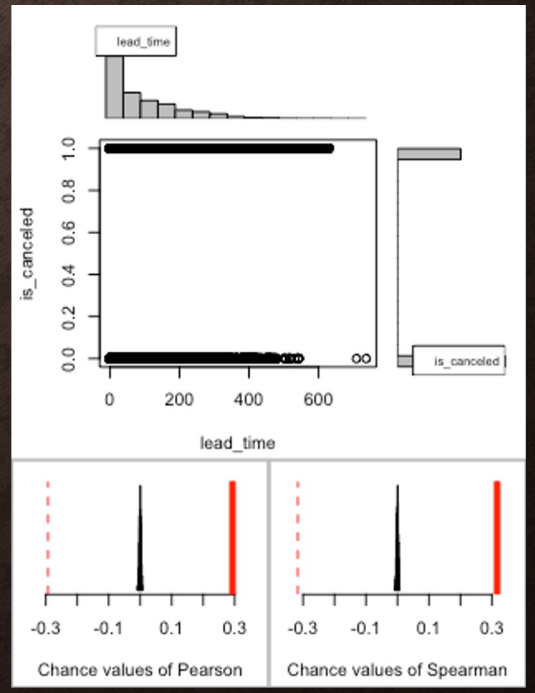
With 500 permutations, we are 95% confident that:
the p-value of Pearson's correlation (r) is between 0 and 0.007
the p-value of Spearman's rank correlation is between 0 and 0.007
Note: If 0.05 is in this range, increase the permutations= argument.



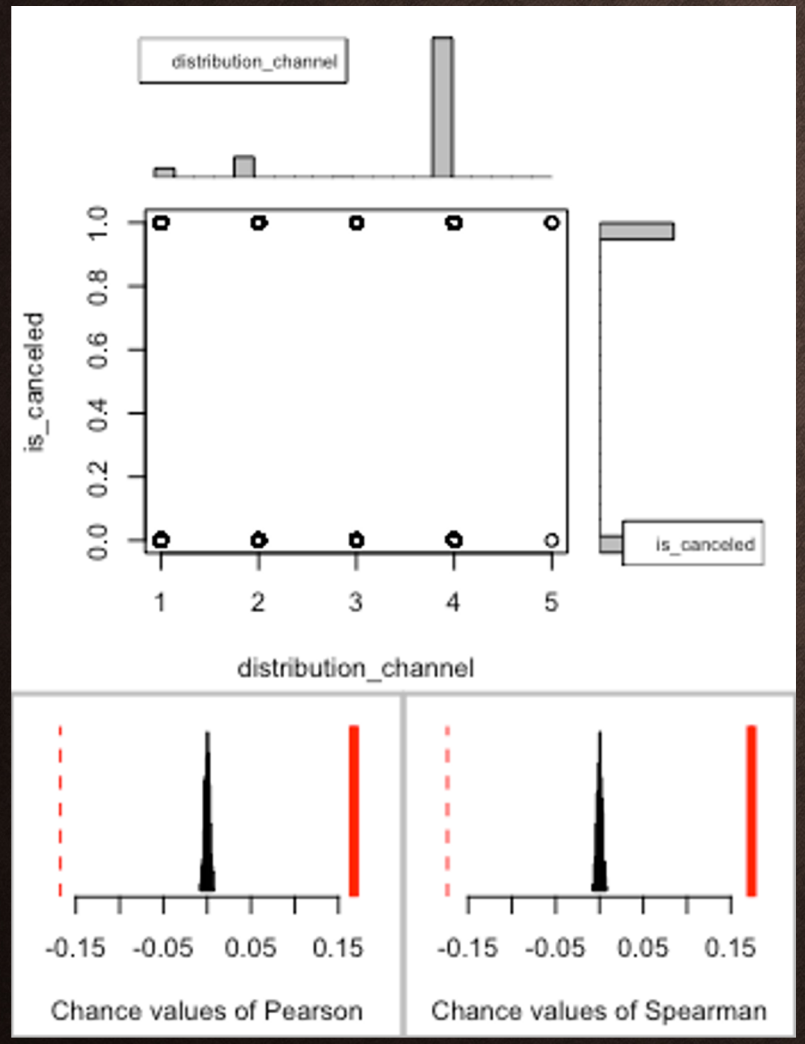
Association between lead_time (numerical) and is_canceled (converted boolean) using 119390 complete cases
Permutation procedure:

	Value Estimated	p-value
Pearson's r	0.2931234	0
Spearman's rank correlation	0.3166346	0

With 500 permutations, we are 95% confident that:
the p-value of Pearson's correlation (r) is between 0 and 0.007
the p-value of Spearman's rank correlation is between 0 and 0.007
Note: If 0.05 is in this range, increase the permutations= argument.



ASSOCIATION ANALYSIS



Association between distribution_channel (numerical) and is_canceled (converted boolean)
using 119390 complete cases
Permutation procedure:

	Value Estimated	p-value
Pearson's r	0.1676003	0
Spearman's rank correlation	0.1736620	0

With 500 permutations, we are 95% confident that:
the p-value of Pearson's correlation (r) is between 0 and 0.007
the p-value of Spearman's rank correlation is between 0 and 0.007
Note: If 0.05 is in this range, increase the permutations= argument.

ASSOCIATION ANALYSIS: LOG OUTPUTS

Previous_Cancellations

```
Call:
glm(formula = is_canceled ~ previous_cancellations, family = binomial,
    data = smallhotel)

Coefficients:
                      Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)         -0.642251     0.006261   -102.58     <2e-16 ***
previous_cancellations  2.060846     0.032777    62.88     < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157398  on 119389  degrees of freedom
Residual deviance: 151456  on 119388  degrees of freedom
AIC: 151460

Number of Fisher Scoring iterations: 6
```

Deposit_Type

```
Call:
glm(formula = is_canceled ~ deposit_type, family = binomial,
    data = smallhotel)

Coefficients:
                      Estimate   Std. Error   z value    Pr(>|z|)
(Intercept)         -0.925836     0.006857   -135.019     <2e-16 ***
deposit_type3        5.974727     0.104232     57.321     <2e-16 ***
deposit_type2       -0.326927     0.189107     -1.729     0.0838 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157398  on 119389  degrees of freedom
Residual deviance: 126130  on 119387  degrees of freedom
AIC: 126136

Number of Fisher Scoring iterations: 7
```

ASSOCIATION ANALYSIS: LOG OUTPUTS

Average_Daily_Rate

```
Call:
glm(formula = is_canceled ~ adr, family = binomial,
    data = smallhotel)

Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -0.7429287  0.0140628    -52.83     <2e-16 ***
adr          0.0020755  0.0001236     16.80     <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157398  on 119389  degrees of freedom
Residual deviance: 157115  on 119388  degrees of freedom
AIC: 157119

Number of Fisher Scoring iterations: 4
```

Lead_Time

```
Call:
glm(formula = is_canceled ~ lead_time, family = binomial,
    data = smallhotel)

Coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept) -1.166e+00  9.182e-03    -126.9     <2e-16 ***
lead_time    5.855e-03  6.137e-05      95.4     <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157398  on 119389  degrees of freedom
Residual deviance: 147158  on 119388  degrees of freedom
AIC: 147162

Number of Fisher Scoring iterations: 4
```

ASSOCIATION ANALYSIS: LOG OUTPUTS

Distribution_Channel

Call:
glm(formula = is_canceled ~ distribution_channel, family = binomial,
    data = smallhotel)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.26126	0.02951	-42.745	< 2e-16 ***
distribution_channel2	-0.29212	0.03667	-7.967	1.63e-15 ***
distribution_channel3	-0.17768	0.18522	-0.959	0.3374
distribution_channel4	0.89836	0.03021	29.734	< 2e-16 ***
distribution_channel5	2.64755	1.11842	2.367	0.0179 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

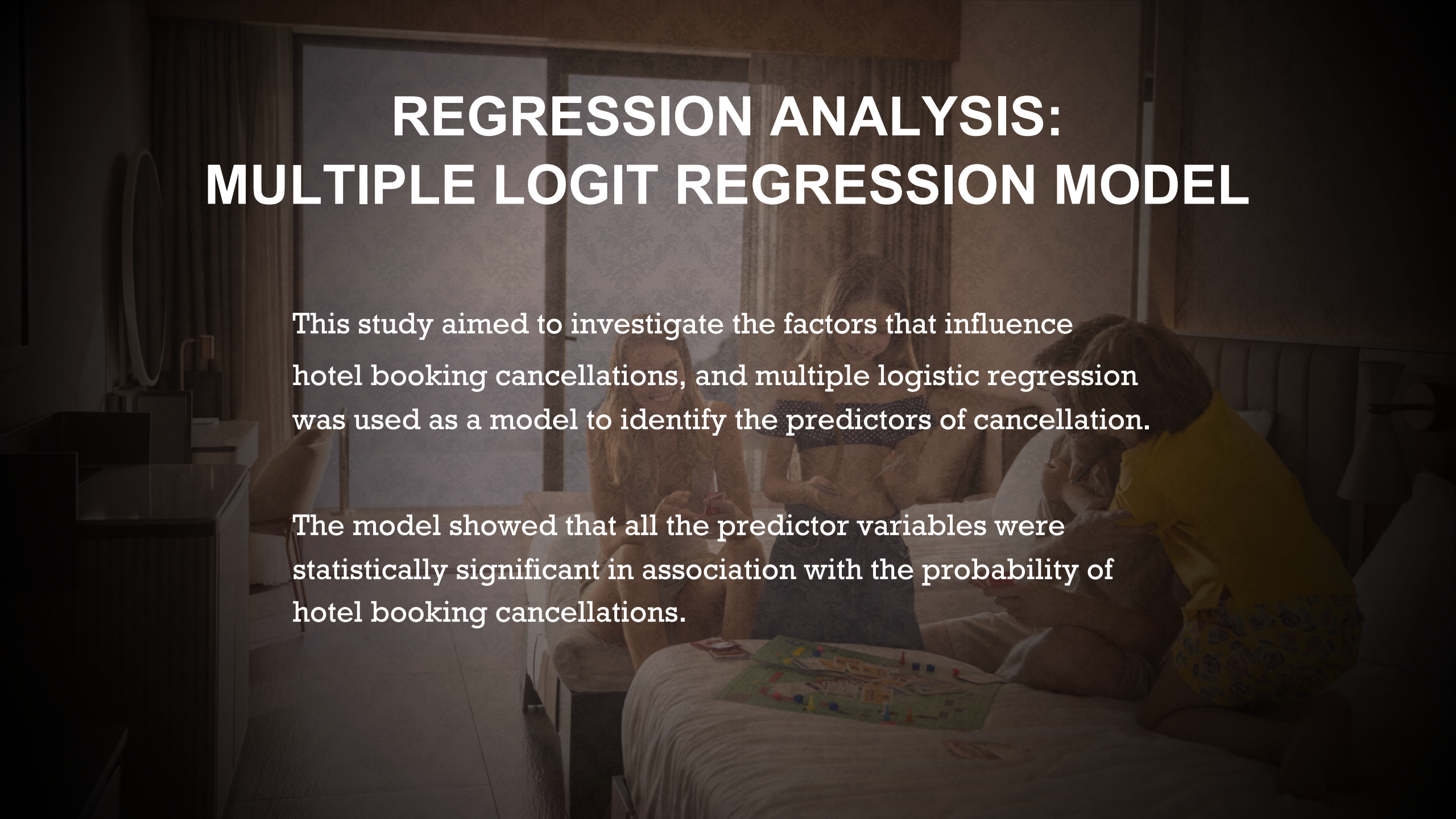
Null deviance: 157398 on 119389 degrees of freedom
Residual deviance: 153314 on 119385 degrees of freedom
AIC: 153324

Number of Fisher Scoring iterations: 4

REGRESSION ANALYSIS: ULTIPLE LOGIT REGRESSION MODEL

- No deposit & a refundable deposit had no statistically significant prediction ability.
- Having a non-refundable deposit increased the likelihood of cancelling.
- A longer lead time increased the likelihood of cancelling.
- As previous cancellations increased, so did the likelihood of cancelling.
- As adr increased, so did the likelihood of cancelling.
- Distribution channels "TA/TO" and undefined increased the likelihood of cancelling.

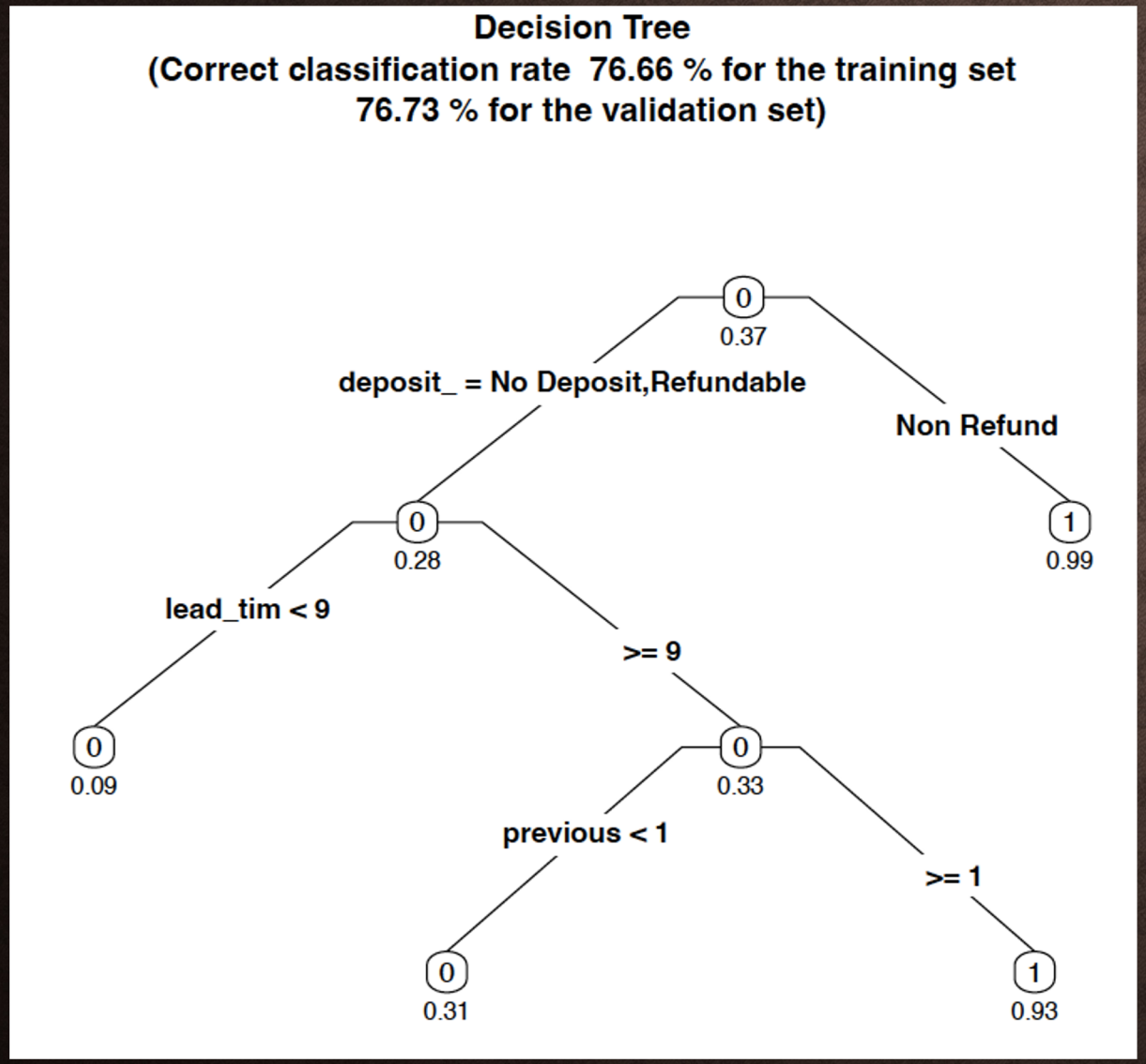
```
Call:
glm(formula = is_canceled ~ lead_time + distribution_channel +
    previous_cancellations + deposit_type + adr, family = binomial,
    data = smallhotel)

Coefficients:
                          Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)           -2.502e+00    3.994e-02   -62.641    <2e-16 ***
lead_time              2.828e-03    7.424e-05    38.085    < 2e-16 ***
distribution_channel2 -8.596e-02    4.529e-02    -1.898    0.05770 .
distribution_channel3  3.088e-01    1.879e-01     1.644    0.10017
distribution_channel4  7.988e-01    3.924e-02    20.353    < 2e-16 ***
distribution_channel5  3.597e+00    1.127e+00     3.192    0.00141 **
previous_cancellations 1.331e+00    3.779e-02    35.216    < 2e-16 ***
deposit_type3          5.618e+00    1.046e-01    53.720    < 2e-16 ***
deposit_type2          8.498e-02    1.946e-01     0.437    0.66231
adr                    5.596e-03    1.436e-04    38.965    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157398  on 119389  degrees of freedom
Residual deviance: 119114  on 119380  degrees of freedom
AIC: 119134

Number of Fisher Scoring iterations: 7
```



REGRESSION ANALYSIS: MULTIPLE LOGIT REGRESSION MODEL

This study aimed to investigate the factors that influence hotel booking cancellations, and multiple logistic regression was used as a model to identify the predictors of cancellation.

The model showed that all the predictor variables were statistically significant in association with the probability of hotel booking cancellations.



ADDITIONAL TECHNIQUES: DECISION TREE

The results are consistent with the multiple logit regression, showcasing the most significant and impactful variables towards predicting cancellations.

OUR FINDINGS

- The travelers that are most likely to cancel are those that went through travel agents as opposed to booking directly.
- The higher the number of previous cancellations a person has, the higher the likelihood of cancelling.
- If the customer puts down a non-refundable deposit, they are more likely to cancel.
- The greater the lead time between booking and date of stay, the higher the likelihood of cancelling.
- As ADR (room rate) increased, so did the likelihood of cancelling.

RECOMMENDATIONS

To decrease cancellations while balancing the need to maximize profit we recommend:

- Removing non-refundable deposits from rooms that have a low likelihood of canceling
  - rooms booked directly by the guest
  - rooms booked with a short lead time
  - rooms booked with a low ADR

- Limiting how far in advance rooms can be booked and adding a non-refundable deposit on rooms booked well in advance of the stay.

- Adding non-refundable deposits on rooms booked by travel agents and by guests who have previously cancelled stays.

- Utilizing dynamic pricing based on supply & demand to reduce cancellations due to finding a lower ADR room without lowering pricing across the board.


BOOKIN
HOTEL