REGRESIÓN

Grupo 5

Manuel García Plaza
José Miguel Ramírez Muñoz

Índice

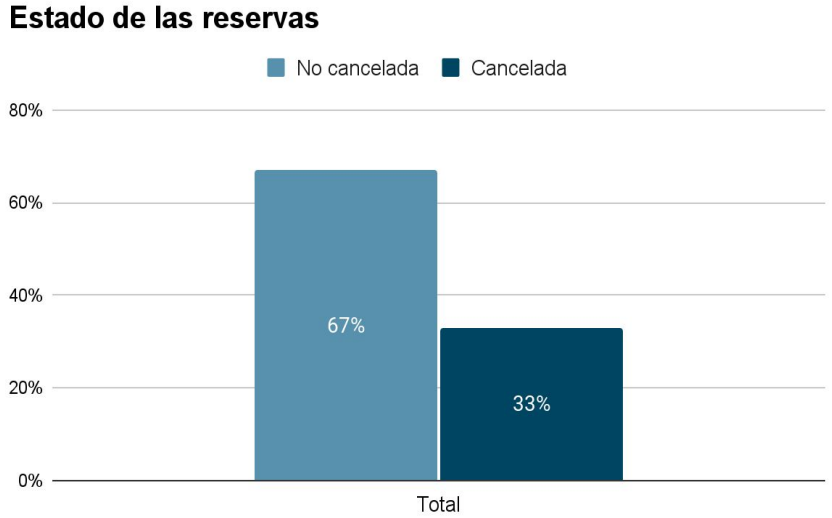
Objetivo para predecir: cancelaciones.

Es una variable binaria:

0: “reserva no cancelada”.

1: “reserva cancelada”.

Queremos **predecir** en función del resto de variables de cada reserva, si esta va a ser cancelada para poder **reducir pérdidas de dinero.**



Modelo con todas las variables.

- **R^2 =** 0,34
- **AIC =** 29796

Variables poco significativas:

- Número de reservas previas no canceladas.
- Menú tipo 2.

	Pr(>|z|)	
(Intercept)	< 2e-16	***
newdata$no_of_adults	0.000654	***
newdata$no_of_children	4.01e-05	***
newdata$repeated_guest	3.15e-06	***
newdata$no_of_previous_bookings_not_canceled	0.242000	
newdata$no_of_previous_cancellations	0.000329	***
newdata$avg_price_per_room	< 2e-16	***
newdata$no_of_week_nights	3.16e-05	***
newdata$no_of_weekend_nights	< 2e-16	***
newdata$type_of_meal_planMeal Plan 2	0.069819	.
newdata$type_of_meal_planNot Selected	4.64e-11	***
newdata$no_of_special_requests	< 2e-16	***
newdata$required_car_parking_space	< 2e-16	***

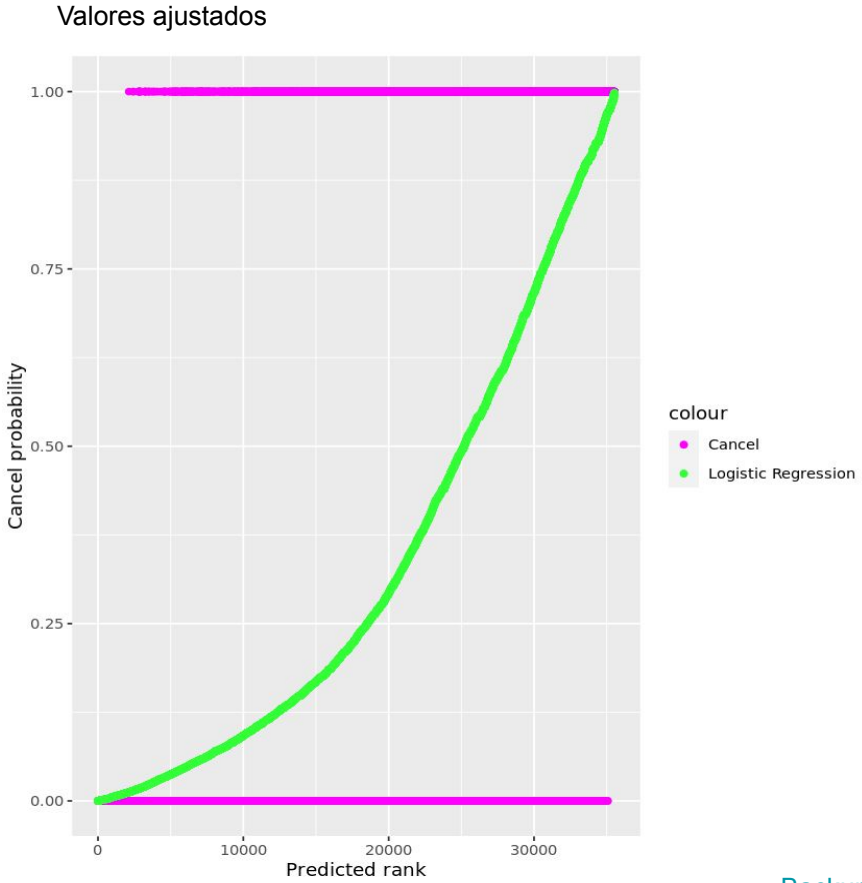
Modelo final.

Este modelo es el anterior eliminando la variable de reservas previas sin cancelar y cambiando la variable **“Meal Plan”** que pasa a ser binaria:

1 = “Selecciona”.
0 = “No Selecciona”.

- **R^2 =** 0,34 ; este modelo es un 34% mejor que el modelo trivial
- **AIC =** 29803

Mediante el método **step-wise no mejora**, luego este es el modelo elegido.



Backup

Modelos alternativos:

División por tipo de habitación:

Si hacemos 6 modelos (cada uno para un tipo de habitación) intentando buscar un mejor ajuste tenemos que el mejor modelo es el que coge datos del tipo de habitación 7 (117 individuos) con:

- **R^2 =** 0.71
- **AIC =** 86

En general el resto **no ajustan muy bien y** tienen **pocos datos**.

Summary

Solo con las variables de mayor correlación:

Si probamos con un modelo con estas 3 variables obtenemos:

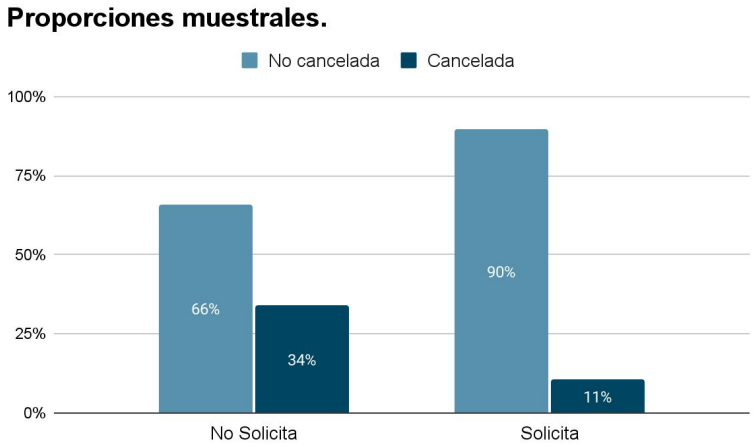
- **R^2 =** 0,24
- **AIC =** 34209

Tenemos **peores métricas** que en el modelo seleccionado.

Summary

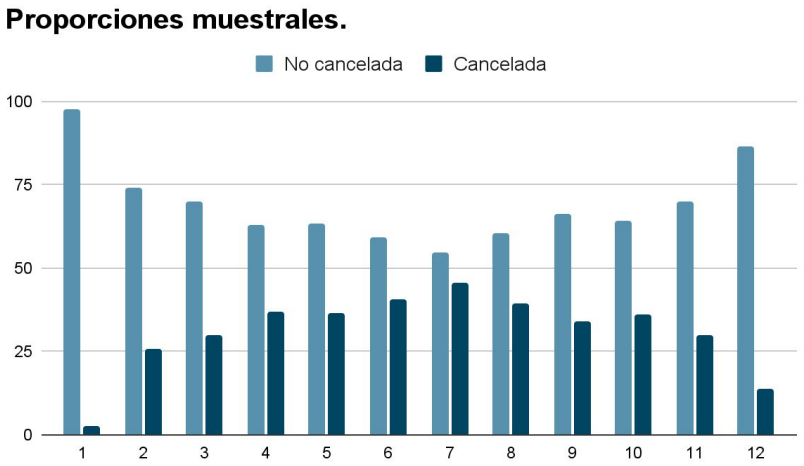
Tests ANOVA significativos en los factores:

Parking



p-valor ANOVA **< 0,05**.

Mes de llegada

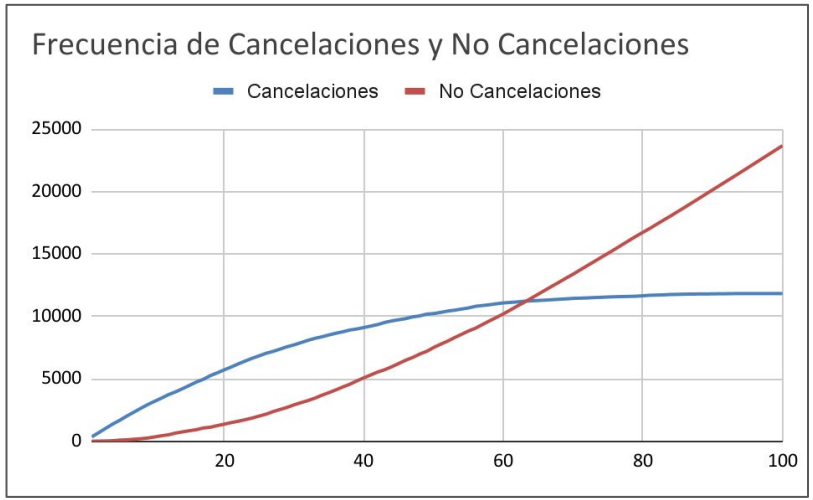


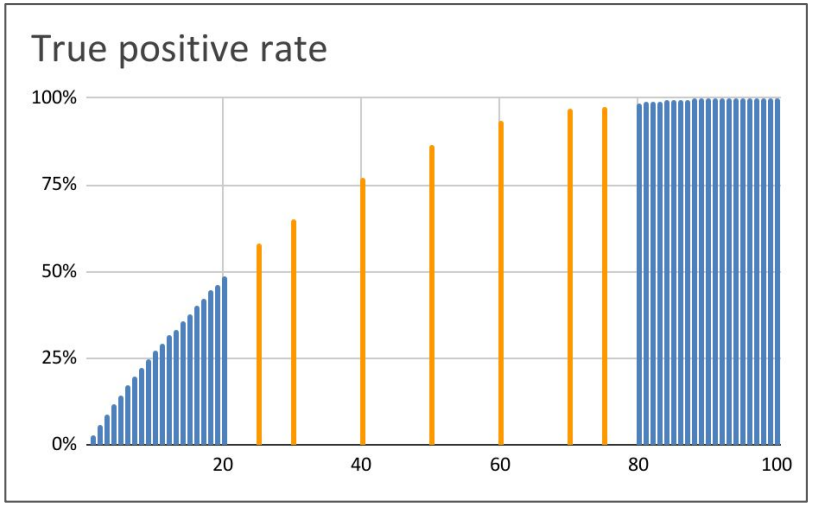
Todos los **p-valores** de ANOVA son **menores** que **0,05** excepto dos.

Efectivamente, existen diferencias estadísticamente significativas en ambas variables.

Backup

Métricas binarias (por percentiles top de scores)

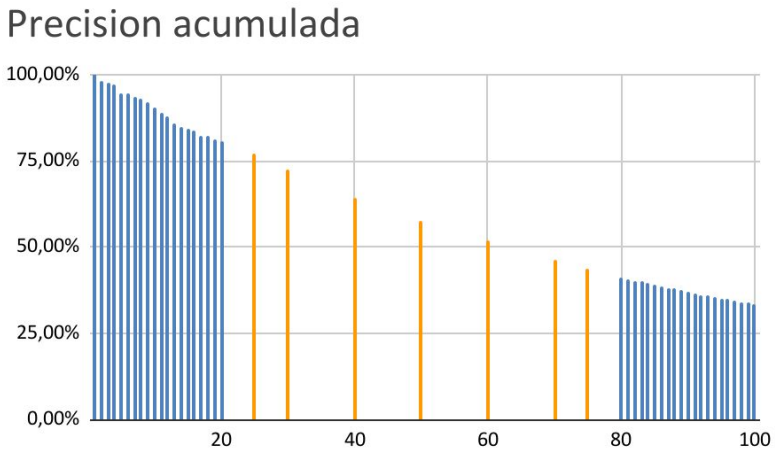


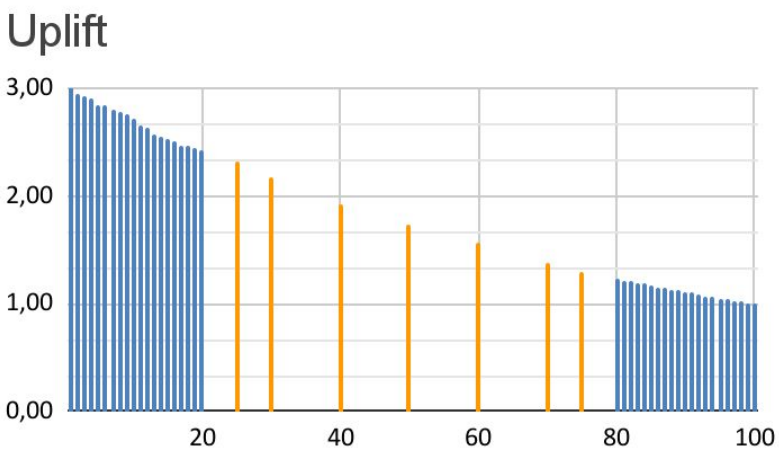


- Observamos que hay más cancelaciones hasta el top 60%. El **crecimiento** del número de
  No Cancelaciones es exponencial y el de cancelaciones es logarítmico.

- En el **top 20%** se encuentra el **50%** de los **verdaderos positivos**. En cambio, a partir del percentil 80 ya tenemos casi el 100% de ellos.

Métricas binarias (por percentiles top de scores)





- Observamos una precisión acumulada del **80%** para el percentil 20.

- Siempre existe una mejora con respecto a la predicción del modelo trivial. En concreto, para el **top 20%** es **entre 2 y 3 veces mejor**.

Diferencias entre los percentiles dados por los scores:

MEDIAS ENTRE EL TOP 20 PERCENTIL Y EL BOT 20 PERCENTIL			
TOP PERCENTILES	Nº DE INDIVIDUOS	no_of_adults	no_of_children
<= 20	7100	1,970	0,152
>= 80	7476	1,746	0,083
	Porcentaje de diferencia	12,83%	84,18%
		no_of_previous_cancellations	Repeate guest
		0,008	10
		0,085	759
	Porcentaje de diferencia	-90,37%	-98,68%
		no_of_weekend_nights	no_of_week_nights
		0,922	2,561
		0,741	2,006
	Porcentaje de diferencia	24,36%	27,67%
		avg_price_per_room	no_of_special_requests
		113,335	0,167
		92,688	1,140
	Porcentaje de diferencia	22,28%	-85,33%
		lead_time	scores
		191,966	0,816
		31,153	0,027
	Porcentaje de diferencia	516,21%	2891,50%
		required_car_parking_space	type_of_meal_plan (class 1)
		14	6.289
		650	6.560
	Porcentaje de diferencia	-97,85%	-4,13%

En **número de niños**, la media del top 20 % es un 84,18% mayor que la del 20% inferior.

En **número de solicitudes especiales**, la media del top 20% es un 85,33% menor que la del 20% de la cola.

Podríamos clasificar el **top 20%** como personas que reservan con mayor antelación, tienen menos solicitudes especiales y viajan menos en coche que el grupo 2.

Variables que disminuyen la propensión a cancelar:

Son las variables con coeficientes negativos:

- **(Intercept):**
  Provoca un factor independiente de 0,002 lo cual sesga hacia el 0 el modelo.

- **repeated_guest:**
  Ser cliente repetido disminuye la propensión de cancelar un **91%**.

- **no_of_special_requests:**
  Personalizar la reserva **(1 solicitud)** disminuye la propensión de cancelar un **78%**.

- **required_car_parking_space:**
  Solicitar parking reduce esta propensión un **79%**.

- **type_of_meal_plan:**
  Elegir algún menú baja la propensión de cancelar un **25%**.

Coeficientes

Variables que disminuyen la propensión a cancelar:

- **room_type_reserved:**

  Los clientes que reservan habitación distinta del tipo 1 son menos propensos a cancelar.

  - tipo 2: un **31%** menos.
  - tipo 4: un **21%** menos.
  - tipo 5: un **54%** menos.
  - tipo 6: un **63%** menos.
  - tipo 7: un **77%** menos.

- **market_segment (type Offline):**

  Aquellos que hacen la reserva **en persona** son un **63%** menos propensos a cancelar.

Coeficientes

Variables que aumentan la propensión a cancelar:

Son las variables con coeficientes positivos:

- **avg_price_per_room:**
  Un aumento de **1€** incrementa esta propensión un **2%**.

- **lead_time:**
  Cada día de antelación aumenta la propensión de cancelar un **1,7%**.

- **market_segment (type Online):**
  La gente que reserva **online** es un **140%** más propensa a cancelar.

- **no_of_week_nights:**
  Cada noche entre semana aumenta la propensión a cancelar un **4,5%**.

- **no_of_weekend_nights:**
  Cada noche de fin de semana aumenta la propensión a cancelar un **15%**.

Coeficientes

Variables que aumentan la propensión a cancelar:

- **arrival_month:**
  Las reservas con mes de llegada que **no** es **enero** son más propensas a ser canceladas.
  - febrero: un **1538%** más.
  - marzo: un **1075%** más.
  - abril: un **836%** más.
  - mayo: un **518%** más.
  - junio: un **722%** más.
  - julio: un **601%** más.
  - agosto: un **563%** más.
  - septiembre: un **494%** más.
  - octubre: un **690%** más.
  - noviembre: un **1086%** más.
  - diciembre: un **74%** más.

- **no_of_previous_cancellations:**
  Cada cancelación previa incrementa un **32%** esta propensión.
- **no_of_adults:**
  Cada adulto más aumenta el ratio un **12%**.
- **no_of_children:**
  Cada niño incrementa el ratio un **23%**.

Coeficientes

Backup

Coeficientes del modelo final:

(Intercept)	-5.9107917
no_of_adults	0.1118634
no_of_children	0.2081663
repeated_guest	-2.4334715
no_of_previous_cancellations	0.2787310
avg_price_per_room	0.0194522
no_of_week_nights	0.0437110
no_of_weekend_nights	0.1428774
type_of_meal_plan	-0.2928763
no_of_special_requests	-1.5191488
required_car_parking_space	-1.5811593
room_type_reservedRoom_Type 2	-0.3784563
room_type_reservedRoom_Type 4	-0.2306438
room_type_reservedRoom_Type 5	-0.7769555
room_type_reservedRoom_Type 6	-1.0005850
room_type_reservedRoom_Type 7	-1.4711227
market_segment_typeOffline	-0.9943034
market_segment_typeOnline	0.8756189
as.factor(arrival_month)2	2.7959038
as.factor(arrival_month)3	2.4639190
as.factor(arrival_month)4	2.2365525
as.factor(arrival_month)5	1.8216967
as.factor(arrival_month)6	2.1065163
as.factor(arrival_month)7	1.9475972
as.factor(arrival_month)8	1.8925531
as.factor(arrival_month)9	1.7822127
as.factor(arrival_month)10	2.0674740
as.factor(arrival_month)11	2.4733842
as.factor(arrival_month)12	0.5535651
lead_time	0.0167256

p-valores ANOVA (FDR):

data: newdata$booking_status and as.factor(newdata$arrival_month)

	1	2	3	4	5	6	7	8	9	10	11
2	< 2e-16	-	-	-	-	-	-	-	-	-	-
3	< 2e-16	0.00387	-	-	-	-	-	-	-	-	-
4	< 2e-16	1.8e-14	3.1e-07	-	-	-	-	-	-	-	-
5	< 2e-16	2.1e-13	1.6e-06	0.79835	-	-	-	-	-	-	-
6	< 2e-16	< 2e-16	< 2e-16	0.00238	0.00109	-	-	-	-	-	-
7	< 2e-16	< 2e-16	< 2e-16	1.0e-11	2.9e-12	7.7e-05	-	-	-	-	-
8	< 2e-16	< 2e-16	1.4e-14	0.02264	0.01156	0.36786	4.8e-07	-	-	-	-
9	< 2e-16	1.2e-09	0.00165	0.00875	0.02259	4.7e-10	< 2e-16	3.5e-08	-	-	-
10	< 2e-16	5.6e-15	4.2e-07	0.44212	0.65224	1.2e-05	< 2e-16	0.00038	0.02561	-	-
11	< 2e-16	0.00392	0.87509	2.4e-08	1.8e-07	< 2e-16	< 2e-16	< 2e-16	0.00037	1.8e-08	-
12	8.8e-11	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16	< 2e-16

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
required_car_parking_space	1	58	58.35	264.6	<2e-16
Residuals	35526	7835	0.22		

Summary modelo por tipo habitación:

- tipo 1: **R² =** 0,34 ; **AIC =** 22924.
- tipo 2: **R² =** 0,4 ; **AIC =** 553.
- tipo 4: **R² =** 0,33 ; **AIC =** 5146.
- tipo 5: **R² =** 0,53 ; **AIC =** 193.
- tipo 6: **R² =** 0,46 ; **AIC =** 743.
- tipo 7: **R² =** 0,71 ; **AIC =** 86.

Summary modelo pocas variables:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.3737405	0.0534242	-63.15	<2e-16
newdata$lead_time	0.0129749	0.0001781	72.87	<2e-16
newdata$avg_price_per_room	0.0188885	0.0004344	43.48	<2e-16
newdata$no_of_special_requests	-1.0393790	0.0211890	-49.05	<2e-16