ÁRBOLES DE DECISIÓN

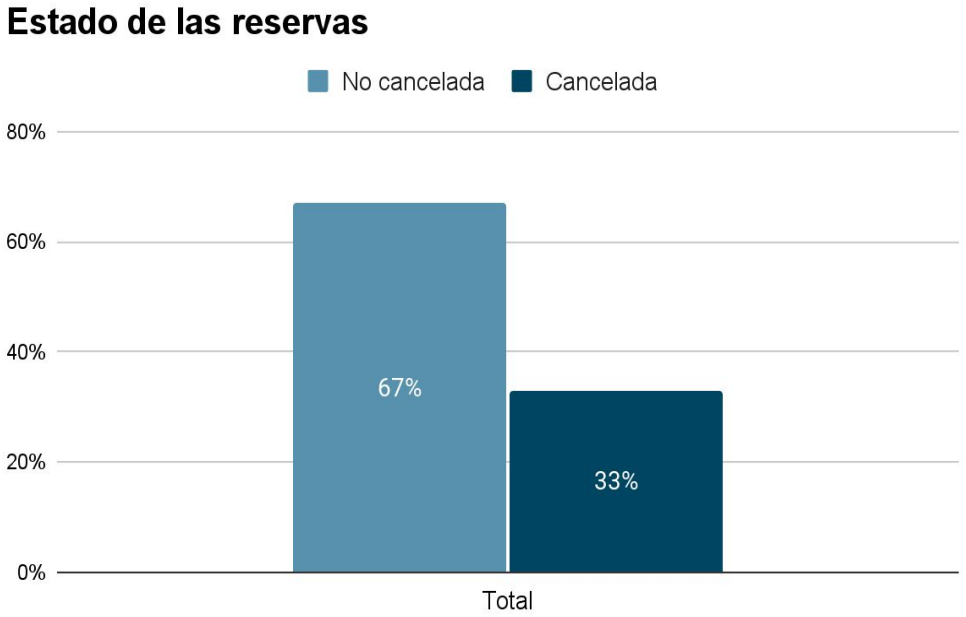
Grupo 5

Manuel García Plaza
José Miguel Ramírez Muñoz

Índice

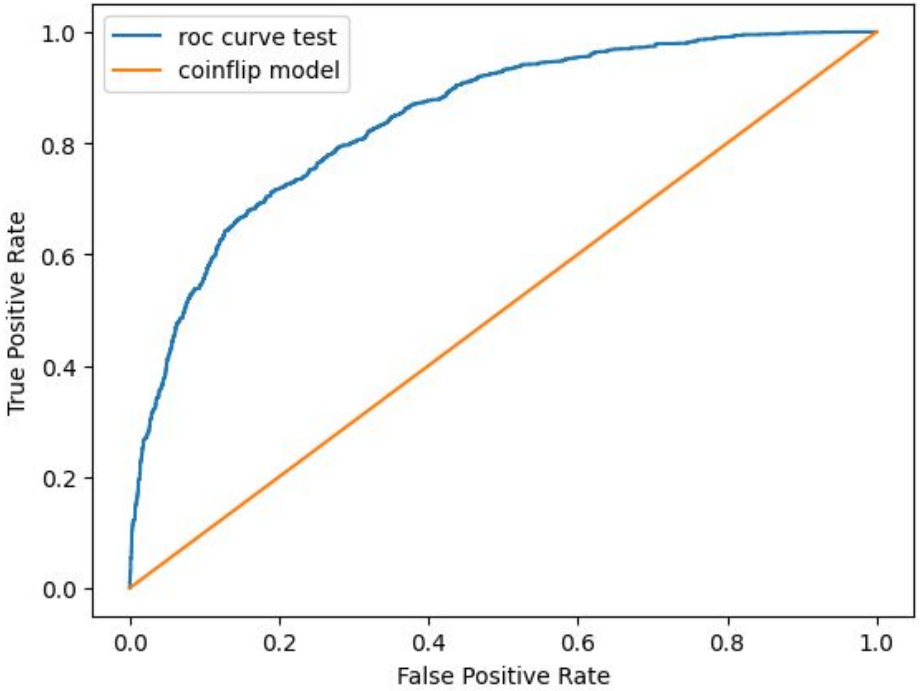
Target: reservas que van a ser canceladas.

Queremos **predecir** si según las características de la reserva, esta tiene **riesgo elevado** de acabar siendo **cancelada** para poder anticiparnos y **reducir pérdidas** de dinero.



El modelo logístico acierta siempre en el top 20% de predicciones

TRAIN	AUC	0.8570
	TOP20_ACCURACY	1.00
TEST	AUC	0.8437
	TOP20_ACCURACY	1.00



Variables usadas y coeficientes del modelo

Los dos mejores árboles en desempeño tienen máxima profundidad 10

Los mejores árboles:

n trees	max depth	max features	max leaf nodes	gini train	gini test	delta
DT 55	10	10	None	0.88	0.82	-0.07
DT 59	10	None	None	0.89	0.84	-0.06

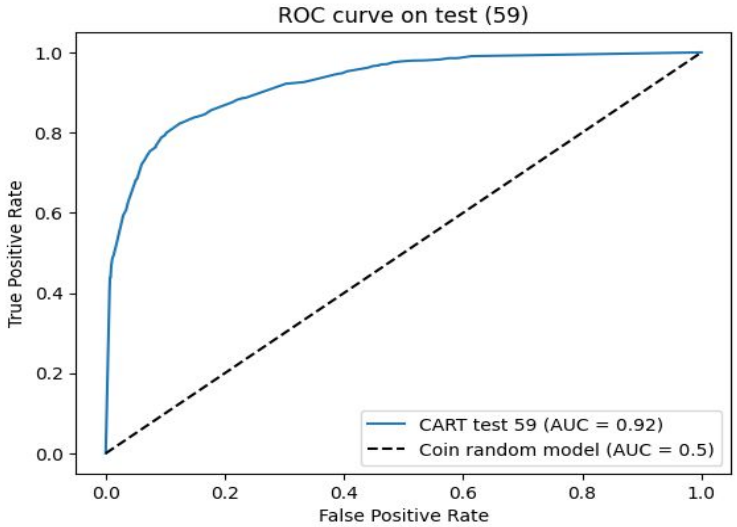
La profundidad es el parámetro más diferencial entre los diferentes árboles; el número máximo de hojas y el de variables no eran tan significativos.

Los árboles con **poca profundidad** pecan de **underfitting**; los de **mucha profundidad**, de **overfitting**.

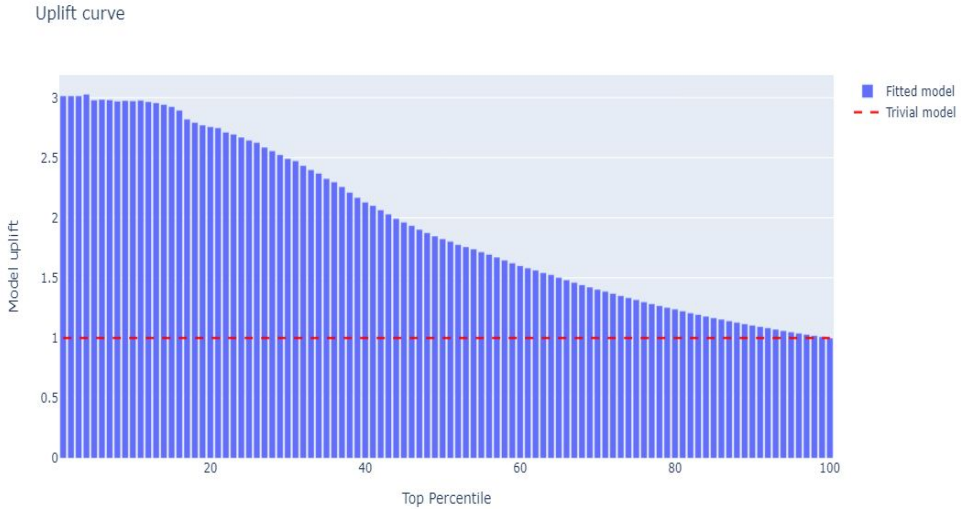
más métricas de entrenamiento/validación (ROC, precisión, sensibilidad, especificidad…)

El árbol sin límite de variables predice un poco mejor con nuevos datos de testeo:

Elegimos el árbol **59** frente al 55 porque en el **percentil 20** es un **1% más preciso**.



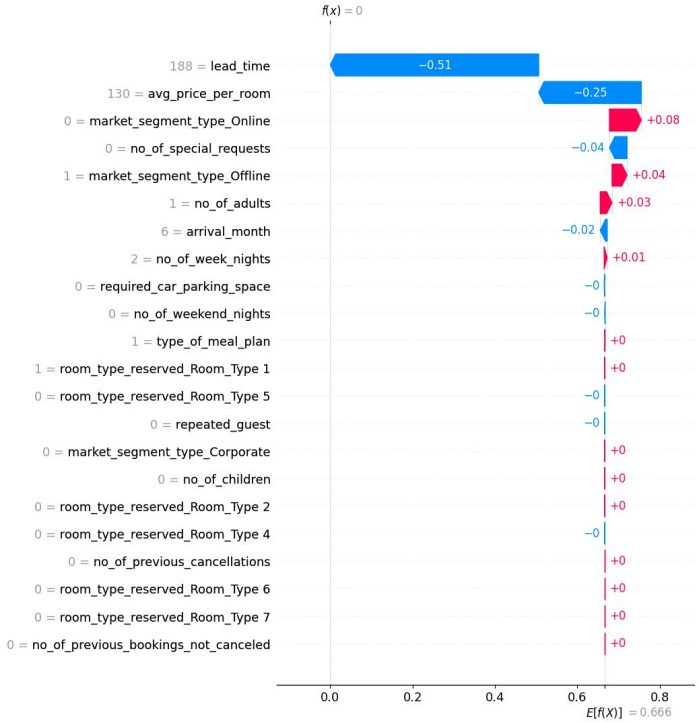
La ROC del 55 es prácticamente idéntica (mismo AUC)



El uplift del 55 es prácticamente igual (2% peor en percentil 20)

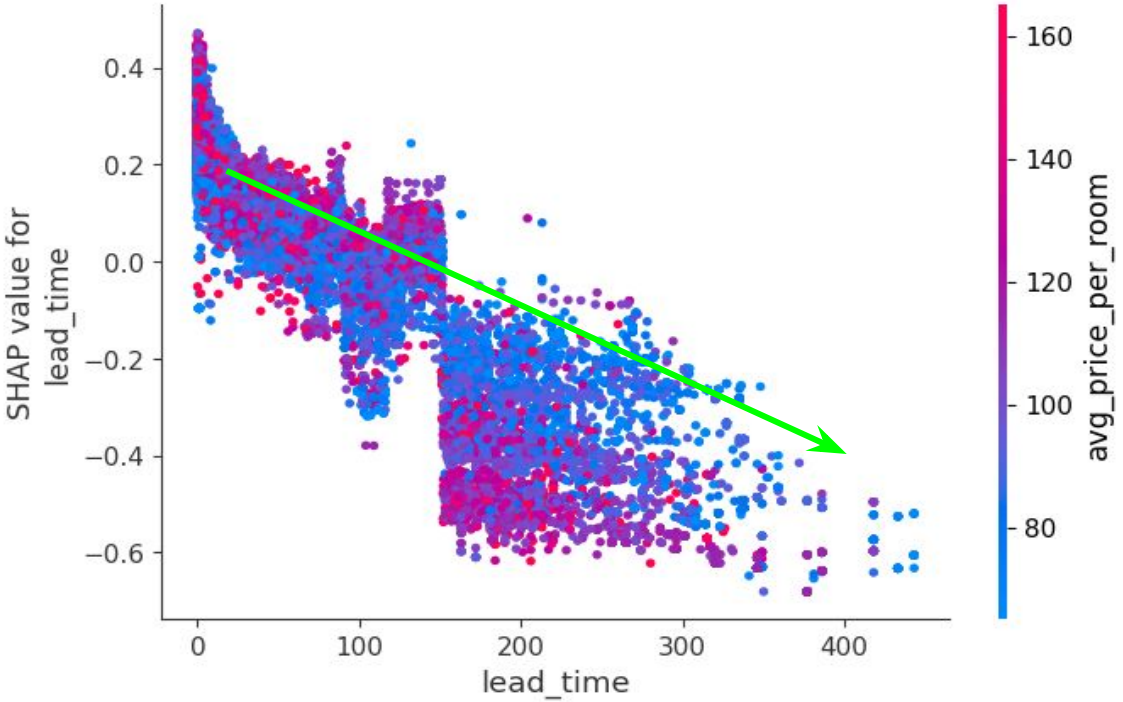
más métricas del test

Las variables más importantes son: el tiempo de antelación de la reserva, el precio de la habitación, el número de requisitos adicionales, y el tipo de reserva online.



Las tablas de shap value son de dos individuos tomados del train.

Poco tiempo de antelación propicia más cancelaciones, en cambio, mucho tiempo baja esta propensión.



Con precios bajos la relación tiende a ser casi lineal, pero los precios elevados hacen que el modelo sea más radical en su predicción.

En RandomForest encontramos modelos buenos; ExtraTrees no proporciona nada interesante.

n trees,max depth	gini train	gini test	delta
RF 10, 12	0.899	0.840	-0.066
RF 15, 12	0.902	0.842	-0.067
RF 30, 12	0.911	0.854	-0.063
RF 50, 12	0.911	0.853	-0.064
RF 100, 12	0.913	0.857	-0.062
RF 250, 12	0.913	0.857	-0.061

Los mejores RandomForest son de profundidad máxima 12 (los mejores en accuracy/delta). A partir de 30 árboles no hay mejora significativa.

Los modelos ExtraTrees con delta menor que un 7% no sobrepasa un gini de 0.75 y aquellos con buen accuracy tienen delta alto (overfittean). En conclusión, no mejoran los modelos anteriores.

n trees, max depth	gini train	gini test	delta
XGB 3, 12	0.914	0.856	-0.064
XGB 5, 12	0.921	0.859	-0.067
XGB 10, 12	0.931	0.869	-0.066
XGB 30, 8	0.907	0.868	-0.044
XGB 50, 8	0.917	0.873	-0.048
XGB 100, 8	0.936	0.880	-0.060
LGBM 30, 8	0.874	0.844	-0.034
LGBM 50, 12	0.895	0.859	-0.040
LGBM 100, 8	0.916	0.870	-0.050
LGBM 100, 16	0.922	0.874	-0.052

Los modelos AdaBoost no mejoran al resto; con XGBoost y LightGBM obtenemos muy buenos resultados.

Los AdaBoost con learning rate muy bajo tienden al underfitting; con learning rate 0.5 o 1 ajustan mejor pero no tan bien como los otros algoritmos.

Los parámetros de regularización y learning rate de XGB y LGBM son por defecto para no demorar mucho los cálculos.

Conclusiones y next steps

- Podemos predecir con la regresión logística para captar cancelaciones casi seguras, a las cuales asignar fianzas elevadas.

- Podemos usar los algoritmos de boosting o Random Forest para clasificar con alta precisión cualquier reserva, no solo las de probabilidad alta, por ejemplo, para poder aplicar fianzas proporcionales siempre.

- Para afinar más si cabe los modelos, se pueden crear nuevas variables.

- En pos de alargar la vida útil de los modelos, hay que recoger más datos a lo largo del tiempo y añadir la componente temporal para predecir a futuro.

Backup

Regresión Logística

```
X_train.columns
```

```
Index(['no_of_adults', 'no_of_children', 'no_of_weekend_nights',
       'no_of_week_nights', 'type_of_meal_plan', 'required_car_parking_space',
       'lead_time', 'arrival_month', 'repeated_guest',
       'no_of_previous_cancellations', 'no_of_previous_bookings_not_canceled',
       'avg_price_per_room', 'no_of_special_requests',
       'room_type_reserved_Room_Type 1', 'room_type_reserved_Room_Type 2',
       'room_type_reserved_Room_Type 4', 'room_type_reserved_Room_Type 5',
       'room_type_reserved_Room_Type 6', 'room_type_reserved_Room_Type 7',
       'market_segment_type_Corporate', 'market_segment_type_Offline',
       'market_segment_type_Online'],
      dtype='object')
```

```
lr.intercept_, lr.coef_
```

```
(array([-1.21417626]),
 array([[-0.08548249, -0.14859586,  0.20064385, -0.04699523, -0.76345682,
         -0.57907282,  0.01652447, -0.09350219, -0.17165259, -0.07666378,
         -0.47527212,  0.0148628 , -1.48735517, -0.33655602, -0.24988093,
         -0.36405473, -0.09492839, -0.14263551, -0.02307823, -0.43545054,
         -1.28274188,  0.50705861]]))
```

Árboles y más:

Ver notebooks en:

https://github.com/mgp165/trabajo_tema_5/tree/main