Checkpoint 2 - Grupo Alan Taylor

Introducción

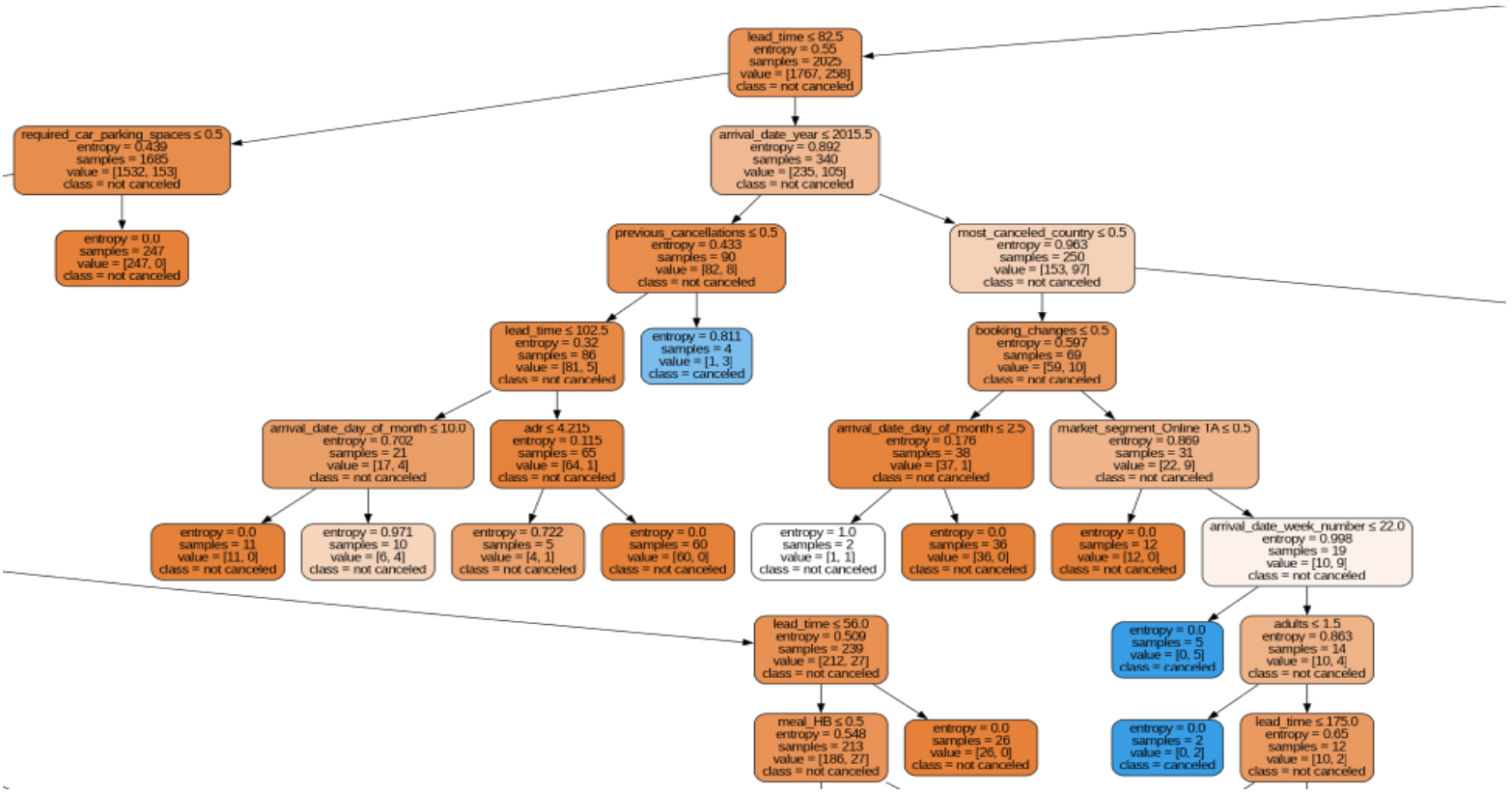
Estuvimos trabajando con diferentes modelos como `Decision Tree Classifier` y `Random Forest Classifier`, para probar nuevas combinaciones con el dataset volvimos a incorporar variables anteriormente eliminadas como agent, meal, arrival date month, etc.
Al incorporar nuevamente agent volvimos a tener el problema de los nulls en la variable, para solucionarlo intentamos imputar los datos con 0 y dividirla en rangos para luego hacer dummies, llegamos a la conclusión de que lo mejor era imputar valores en 3 y dejar la variable así.
A la variable meal le borramos los valores “Undefined” y le hicimos dummies.
Con arrival date month hicimos dummies.

Construcción del modelo

- Optimizamos hiper parámetros y nos dimos cuenta que los mejores para cada árbol eran:
  - Decision Tree Classifier: 0.8515 - Random Forest Classifier: 0.8846

criterion="entropy",	n estimators=550,
max depth = 15,	bootstrap=True,
max features=None,	max depth=25,
min samples_split=14,	min samples split=2,
min samples leaf=2,	min samples leaf=1,
random state=45	random state=70

- Utilizamos cross validation en ambos casos, para el Decision Tree Classifier utilizamos 10 folds.
- Comenzamos con valores vistos en clase, nosotros fuimos probando otros e íbamos viendo cuánto nos daba el score, una vez conseguido los mejores por nosotros hicimos Cross validation en un rango cercano a los conseguidos y conseguimos mejores resultados.
- El primer modelo subido a kaggle fue un Decision Tree Classifier con parámetros vistos en clase, el cual nos devolvió una predicción del 81.7% de acierto, el último modelo subido a kaggle fue el de un Random Forest Classifier con parámetros encontrados a través de Cross Validation que nos devuelve un porcentaje de acierto del 88%.
- Imagen del árbol generado completa.



Para mejor entendimiento mostramos una parte representativa del árbol.

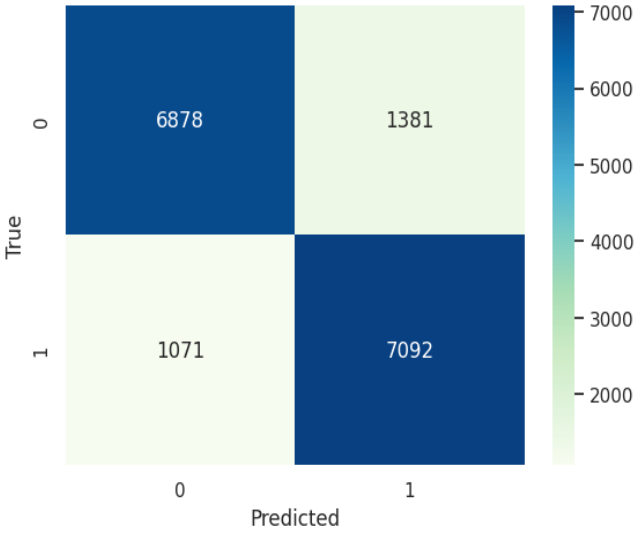
Cuadro de Resultados

Modelo	F1-Test	Precision Test	Recall Test	Accuracy	Kaggle
Decision Tree Classifier	0.8533	0.8381	0.8691	0.8515	0.8487
Decision Tree Classifier	0.8557	0.8474	0.8642	0.8533	0.8484
Random Fores Classifier	0.8847	0.8897	0.8797	0.8846	0.88002

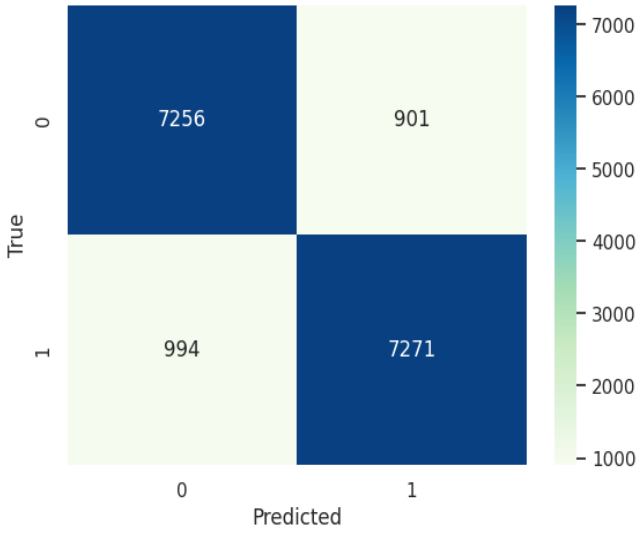
Aclaración: El modelo actualmente subido a kaggle con el mejor score es el random forest

Matriz de Confusión

- Decision Tree Classifier:



- Random Forest Classifier:



Tareas Realizadas

Indicar brevemente en qué tarea trabajo cada integrante del equipo, si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte).

Integrante	Tarea
Santiago Trezeguet	Análisis de variables predictoras, Gráficos, Submits en Kaggle, Búsqueda de Hiperparametros, Armado de reporte
Estefano Polizzi	Análisis de variables predictoras, Gráficos, Submits en Kaggle, Búsqueda de Hiperparametros, Armado de reporte
Ignacio Oviedo	Análisis de variables predictoras, Gráficos, Submits en Kaggle, Búsqueda de Hiperparametros, Armado de reporte