**Checkpoint 2 - Grupo 07**

**Introducción de las medidas tomadas**

* Se aplicó el mismo tratamiento de valores nulos del CHP1 al data set de test.
* Se agrego la columna Country Top 10 y high\_season\_month.
* Se realizaron pruebas con distintas particiones del train-test-split (70/30, 80/20) y semillas.
* Probamos distintos rangos de hiperparametros para el RSCV.
* Se aplicó One-Hot encoding a las columnas categóricas con el método get dummies.
* Probamos distintos tamaños de K-folder para el RSCV, en un rango de [5,15].

**Construcción del modelo**

* Optimizamos los hiperparametros: '*criterion*', *'min\_samples\_leaf'*, *'min\_samples\_split'* y *'max\_depth'*. Por otro lado, con el objetivo de buscar la mejor combinación de parámetros que mejore el f1\_score, usamos RSCV.
* Durante la búsqueda de los hiperparametros óptimos se utilizaron 15 folds.
* Consideramos que la mejor métrica para buscar los hiperparametros más eficientes es **F1 score.**
* Durante la primera submission a Kaggle, utilizando un modelo sin optimización de hiperparametros, se obtuvo un puntaje de [**0.83668** (se entrenó con 70/30)](https://www.kaggle.com/competitions/7506RTP1-2C2023/submissions).
* Durante la segunda submission a Kaggle, utilizando un modelo con optimización de parametros, se obtuvo un puntaje de [**0.84116** (se entrenó con 80 /20 y RSCV)](https://www.kaggle.com/competitions/7506RTP1-2C2023/submissions). Obtuvimos una mejora aproximadamente del **0.5 %**

Imagen que contiene Diagrama

Descripción generada automáticamente

[Notebook](https://colab.research.google.com/github/gmenendez0/7506R-2C2023-GRUPO07/blob/main/7506R_TP1_GRUPO07_CHP2_ENTREGA.ipynb)

**Podemos observar a partir del grafico que:**

**-** El nodo raíz evalúa con la variable *deposit\_type\_Non\_Refound,* donde vemos que la gente tiende a cancelar la reserva si esta tiene reembolso.

**-** El segundo nivel árbol se evalúa con las variables *required\_car\_parking\_spaces* y *Country\_top\_10*. Vemos que la gente tiende a no cancelar si la reserva incluye espacio para estacionar autos. En la segunda variable, vemos que la gente tiende a cancelar si no forma parte del top 10 países.

**Cuadro de Resultados**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | F1-Test | Precision Test | Recall Test | Metrica X | Kaggle |
| **Rafa (Entrega)** | 0.86008 | 0.83822 | 0.88310 | 0.85475 | 0.84471 |
| Mateo | 0.85682 | 0.84721 | 0.86665 | ¿? | 0.84292 |
| Guido | 0.83 | 0.85 | 0.81 | 0.84 | 0.84172 |

**Matriz de confusión**

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

*Matriz de confusión de nuestro modelo más optimo. Vemos que detecta mejor los casos de positivos (relación True positive y False negative). Esto se refleja en un alto score en la métrica Recall (TP / TP + FN).*

**Tareas Realizadas**

|  |  |
| --- | --- |
| Integrante | Tarea |
| Rafael Wu | Confección de informe y modelo |
| Mateo Riat Sapulia | Confección de informe y modelo |
| Guido Menendez | Confección de informe y modelo |