Trabajo Práctico 1 - CHP3
Reservas de Hotel

[75.06/95.58] Organización de Datos
Primer cuatrimestre de 2023

ALUMNO	PADRON	CORREO
BARTOCCI, Camila	105781	cbartocci@fi.uba.ar
LOURENGO, Lucía	104880	llourengo@fi.uba.ar
PATIÑO, Franco	105126	fpatino@fi.uba.ar

Sobre los modelos construidos y sus métricas

Se construyeron varios modelos de distintos tipos y se realizaron iteraciones sobre los mismos, a saber:

- KNN. Para este modelo buscamos optimizar hiperparámetros mediante Random Search con el objetivo de mejorar el F1 Score. Hicimos distintas iteraciones separadas por algoritmo y por conjunto de métricas de distancia aceptadas por cada uno de ellos. Finalmente encontramos que los mejores hiperparámetros mejoran la métrica a un 82 % obteniendo un balance aceptable entre recall y precisión.
- SVM. Para entrenar este modelo optamos por normalizar el conjunto de datos y aplicar una reducción de dimensionalidad de a 20 componentes de las 35 totales, ya que con esta cantidad se explica casi el 90 % de la variabilidad de los datos. A partir de esto decidimos entrenar SVM con Kernel Lineal y con Kernel Radial. Mientras que para el primero obtuvimos métricas que rondan el 75 %, para el Kernel Radial obtuvimos un valor de Recall considerablemente más alto que el resto de las métricas (85 % aproximadamente).
- Random Forest. Realizamos varias iteraciones optimizando hiperparámetros con Grid Search. En las primeras buscamos optimizar F1 Score y en algunas probamos la optimización multimétrica. A partir de esta última pudimos elegir los mejores hiperparámetros. Logramos obtener las mejores métricas con este ensamble.
- XGBoost. Entrenamos el modelo optimizando hiperparámetros con Random Search. Las métricas obtenidas en distintas iteraciones rondan el 81 % y 85 %.
- Voting y Stacking. Construímos ambos ensambles híbridos por separado, realizando varias iteraciones. Utilizamos distintas combinaciones de modelos tanto para Voting como para los modelos base de Stacking. En ambos casos logramos métricas muy malas (todas por debajo del 52 %)

Sobre el mejor modelo obtenido

Como mencionamos en el apartado anterior, nuestro mejor predictor resultó ser el ensamble Random Forest, mejorando también las métricas obtenidas en el checkpoint anterior. Obtuvimos:

- 84 % de Accuracy
- 82 % de Recall
- 86 % de Precision
- 84 % de F1 Score

Algunos de los mejores hiperparámetros obtenidos son: Criterio de Entropía, 80 estimadores, una cantidad mínima de muestras por hoja de 1 y una cantidad mínima de muestras por split de 4. También analizamo el área bajo la curva ROC, siento de aproximadamente 92 %, lo que indica que el modelo tiene una capacidad de discriminación muy buena entre instancias positivas y negativas.