Universidad de Buenos Aires
Facultad de Ingeniería

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06/95.58 – Organización de Datos

Curso 02
1C 2023

Trabajo Práctico 1: Reservas de Hotel

Checkpoint 3

Grupo 14

Integrantes:

- *BOTOSHANSKY, ALAN* *108933*
- *BOTOSHANSKY, IVAN* *108897*
- *NIEVA, ULISES TADEO* *107227*

Corrector: Nacho

Fecha de entrega: *11* de *mayo* del 2023

Universidad de
Buenos Aires
Facultad de
Ingeniería

Resumen

En este reporte describiremos los pasos ejecutados para la obtención de los mejores clasificadores, las características de cada uno y los resultados obtenidos.

Desarrollo

Las tareas iniciales fueron las mismas realizadas en la etapa anterior: importar los datasets (de entrenamiento y evaluación), preprocesarlos de la misma forma, aplicarles One Hot Encoding y quedarse con las variables dummies comunes a ambos. Para desarrollar las optimizaciones elegimos como métrica F1 Score y 5 folds para la validación cruzada.

Para los primeros cuatro modelos (KNN, SVM, Random Forest y XGBoost) procedimos de manera similar: creamos un clasificador con sus parámetros por defecto, lo entrenamos y realizamos una predicción. El rendimiento de éste era la vara a superar en el siguiente paso, optimizando los hiperparámetros de cada modelo mediante K-Fold Cross Validation.

Para KNN variamos el número de vecinos, el tipo de peso, el algoritmo, y las métricas de distancia. Los valores máximos de F1 Score conseguidos rondaban 0.78, y aunque manejaban cantidades de vecinos y algoritmos distintos, coincidían en emplear *distance* como función de peso y *manhattan* como métrica. El valor máximo de F1 Score obtenido fue 0.7812, usando 17 vecinos y *ball tree* como algoritmo.

En SVM, además de crear un modelo por defecto, probamos qué rendimiento alcanzaba sin normalizar los datos. Luego, tras aplicar Escalado estándar, apreciamos una mejora significativa. Después, realizamos una optimización de hiperparámetros particular para cada tipo de kernel, obteniendo todos los valores de F1 Score cercanos a 0.83. Curiosamente, el que mejor resultado dio, 0.8302, fue el SVM cuyos parámetros coinciden con los valores por defecto: un kernel radial, un parámetro de regularización igual a 1 y *scale* como coeficiente gamma.

A partir de Random Forest empezamos a ver una gran mejora en el rendimiento de los clasificadores, dando el modelo con parámetros por defecto un F1 Score de 0.8758. Los resultados tras la optimización no variaron mucho respecto de ese número, siendo el máximo valor encontrado 0.8785. Dicho RF tiene *gini* como criterio de elección, 140 árboles y toma como cantidad máxima de variables para cada árbol del RF la raíz cuadrada del total de variables.

XGBoost mostró resultados semejantes, aunque se pudo apreciar una leve mejora, dando varios modelos rendimientos alrededor de 0.88. El mejor de ellos tuvo un F1 Score de 0.8819, y sus parámetros fueron 180 estimadores, profundidad máxima de 24, tasa de aprendizaje igual a 0.1, un umbral gamma de 0.1, y una regularización alpha y lambda igual a 0.8 y 1, respectivamente.

Dada la superioridad de rendimiento de Random Forest y XGBoost, elegimos utilizarlos para la conformación de los ensambles híbridos de tipo Voting y Stacking. La cantidad de estimadores que establecimos para cada uno es 3. Para la validación cruzada, consideramos todas las combinaciones posibles de a 3 de los 7 modelos importados (35 combinaciones en total). Además, elegimos como estimador final para Stacking un XGBoost. Ambos ensambles consiguieron sus mejores resultados utilizando tres XGBoost, siendo 0.8852 el valor de F1 Score del mejor Stacking, y 0.8861 el del mejor Voting, siendo este último el mejor rendimiento conseguido a lo largo del trabajo.

Conclusión

Queda en evidencia la mejora progresiva conforme los clasificadores aumentan su complejidad, aunque esta

característica también viene acompañada de un incremento en la cantidad de recursos necesarios para la optimización y ejecución de estos modelos.