

Checkpoint 4 - Grupo 07

Introducción

En esta etapa del proyecto se entrenaron diversas redes neuronales, siendo la más óptima, la cual usa 250 neuronas de entrada, activación '*relu*' y tipo de conexión densa, seguida de 4 capas ocultas de neuronas tipo densas, todas con activación '*relu*' (con 200, 150, 100 y 50 neuronas respectivamente). Por último, una capa de salida con activación *sigmoidal* y una sola neurona.

Por otro lado, a nivel datos, no se realizaron cambios destacables más allá de la estandarización de las variables numéricas cuantitativas.

Construcción del modelo

- Arquitectura escogida:
 - Capa entrada
 - neuronas de entrada: **250**
 - Función de activación: **'relu'**
 - Tipo de conexión: **'Dense'**
 - Capas ocultas
 - neuronas por capa
4 capas, cantidad de neuronas por capa (decreciente):
200, 150, 100, 50
 - Función de activación: **'relu'**
 - Tipo de conexión: **'Dense'**
 - Capa de salida
 - cantidad de neuronas: **1**
 - función de activación: **'Sigmoid'**
- Hiperparámetros optimizados
'Epochs' random.randint(10, 100, size=5) y 'batch_size': [16, 32, 64, 128]

¿Qué optimizador se utilizó?

SGD con learning rate de 0.15

- ¿Se utilizó alguna técnica de regularización? ¿Cuál?

Ninguna

- ¿Cuántos ciclos de entrenamiento utilizó?

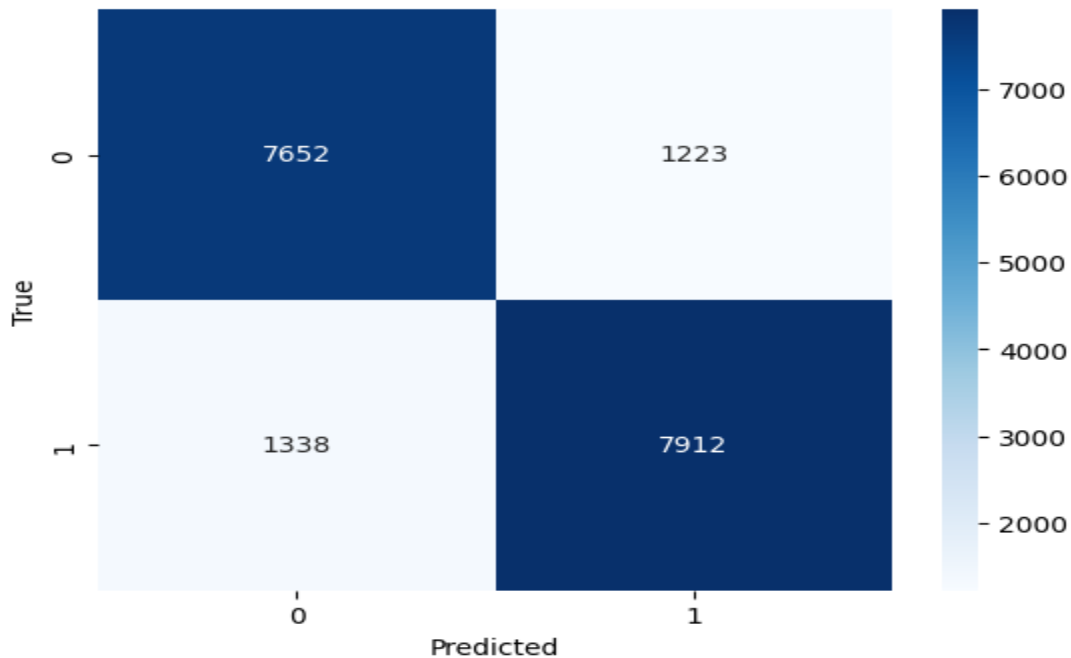
22

Cuadro de Resultados

Modelo	F1-Test	Precision Test	Recall Test	Kaggle
Modelo 1	0.84	0.84	0.84	0.836
Modelo 2	0.86	0.86	0.86	0.837
Modelo 3	0.86	0.86	0.86	0.853
Modelo 4 (ÓPTIMO)	0.86	0.86	0.86	0.860

Podemos observar por las métricas que todos los modelos tienen puntajes similares en todos los tests. También vemos que los colores de las pruebas de un mismo modelo son muy balanceados, tanto su Precisión como Recall son idénticos, lo que nos lleva a pensar que es un modelo equilibrado, ya que no se destaca por nada en ningún caso particular, ni para bien, ni para mal.

Matriz de confusión de modelo óptimo



Tareas Realizadas

Integrante	Tarea
Rafael Wu	Confección del reporte Creación de modelos
Mateo Suster	Confección del reporte Creación de modelos
Guido Menendez	Confección del reporte Creación de modelos

Conclusión TP1

Se ha confeccionado este apartado con el objetivo de concluir el primer trabajo práctico de la materia Organización de Datos.

La cátedra planteó resolver un problema real en el que se buscaba predecir, en base a un conjunto de datos de reservas de hoteles ya recopilado anteriormente, cuáles reservas tenían más chances de ser canceladas y cuáles no. El proyecto se dividió en 4 “*Checkpoints*”, cada uno con sus objetivos y fechas límite.

Al inicio de este proyecto se aplicaron técnicas de preprocesamiento, transformación y visualización de datos como One-Hot Encoding, tratamiento de Outliers y valores nulos, armado de diversos gráficos, análisis de covarianza y eliminación de variables despreciables para mejorar los datos proveídos por la cátedra.

Luego, se pasó a una etapa donde el foco se puso en el entrenamiento y evaluación constante de los modelos de predicción vistos a lo largo de la materia. Entre estos podemos mencionar ID3, KNN, SVM, Random Forest, XG Boost.

También se construyeron ensambles híbridos de tipo Voting y Stacking. Por último, y como objetivo del cuarto cuarto del trabajo práctico, se entrenó una red neuronal.

Uno y cada uno de los modelos mencionados anteriormente fue entrenado, optimizado y testeado utilizando técnicas como K-Fold Cross Validation, Grid Search, Random Search, entre otros.

Finalmente, cada uno de los modelos más óptimos de cada checkpoint fue utilizado para realizar una predicción, la cual posteriormente fue agregada a una competencia Kaggle.

En cuanto a la comparación de los resultados que se obtuvieron en Kaggle para cada modelo, podemos decir que el ensamble híbrido de tipo Stacking consiguió el mejor resultado (0.878 aprox.).

En el escalón de los resultados “promedio” podemos observar que se ubican la gran cantidad de otros modelos, desde ID3, XG Boost, ensamble de tipo Voting, etc., en un rango de entre 0.86 y 0.82.

En el otro extremo tenemos el caso particular de KNN, que nos fue imposible conseguir que puntuara por encima del 0.80 hasta que se decidió normalizar los datos y por fin se consiguió

un 0.82, manteniendo su puesto como modelo predictor menos óptimo entre todos los realizados.