Universidad Simón Bolívar
Departamento de Cómputo Científico y Estadística
CO-6612 Introducción a las Redes Neuronales y sus Aplicaciones
Prof. Minaya Villasana de Armas

Proyecto

Autor
José Barrera **15-10123**

Sartenejas, julio de 2020

Resumen

En el presente proyecto se trabaja con el conjunto de datos denominado, *Default of Credit Card Clients* [1], disponible en la página web del *UC Irvine Machine Learning Repository*. Con el objetivo de generar una red que logre clasificar lo mejor posible los datos, utilizando los conocimientos aprendidos en el curso, particularmente perceptrones multicapa y boosting [2].

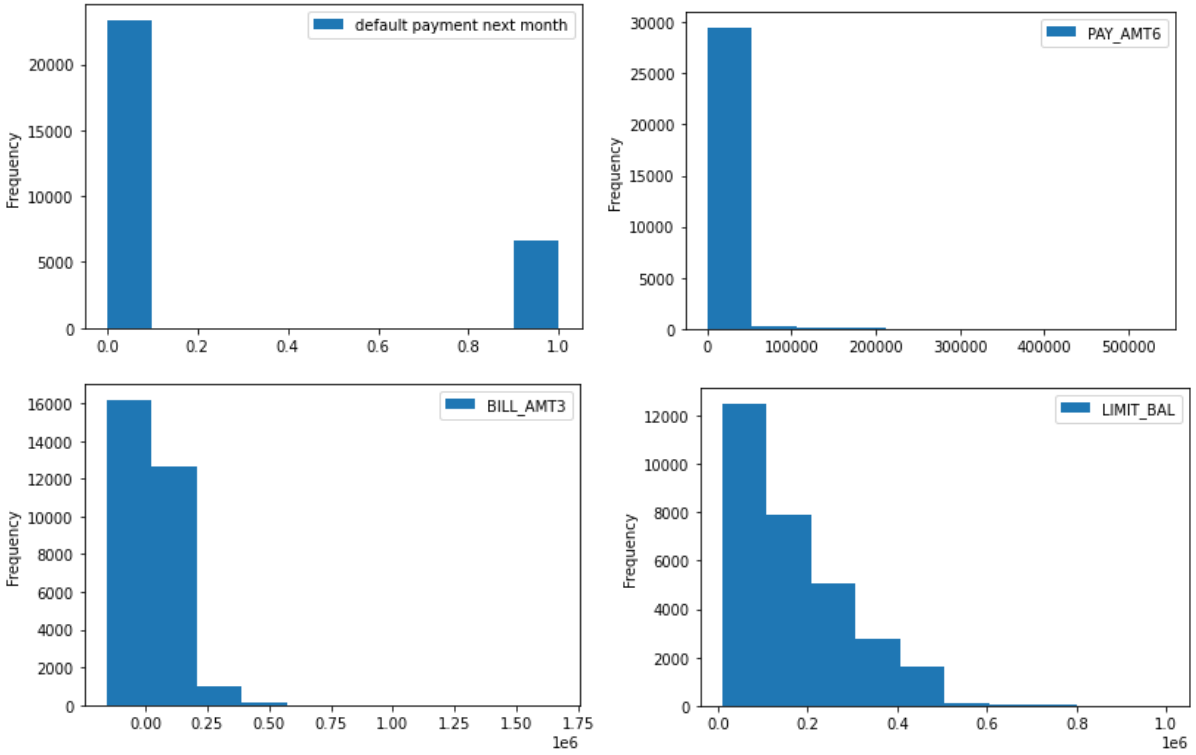
Pre-Procesamiento de los Datos

Análisis: El conjunto de datos consiste en unas 30000 entradas cada una con un total de 23 atributos expresados en números reales. Estos representan información bancaria de los clientes, tales como: edad, estado civil, nivel de instrucción, género, estado de cuenta para ciertos meses, tiempo de retraso en pagos, entre otros. Y por último el dato de clasificación, si el cliente cometió impago o no, que será la información que la red aprenderá a clasificar y predecir.

Tipificación: Ya que el rango de los distintos atributos posee una variedad importante, por ejemplo: las edades y la cantidad de crédito. Se consideró necesario hacer una tipificación para que todos se encuentren en el rango [0, 1] siguiendo la siguiente fórmula:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Histogramas De Atributos: Se realizaron histogramas de frecuencia para cada uno de los atributos incluyendo la clase a la que pertenecen, se encontró que para el atributo de clasificación existe aproximadamente 4 veces más miembros de la clase 0 que de la clase 1. También se observó que para los atributos de tipo PAY_AMT, BILL_AMT y el caso de LIMIT_BAL existen un serie de valores atípicos, por ejemplo: para PAY_AMT6, el valor más alto es 528666, para BILL_AMT3 es 1664089, para LIMIT_BAL es 1000000, la gravedad de esto se visualiza con la siguiente gráfica.

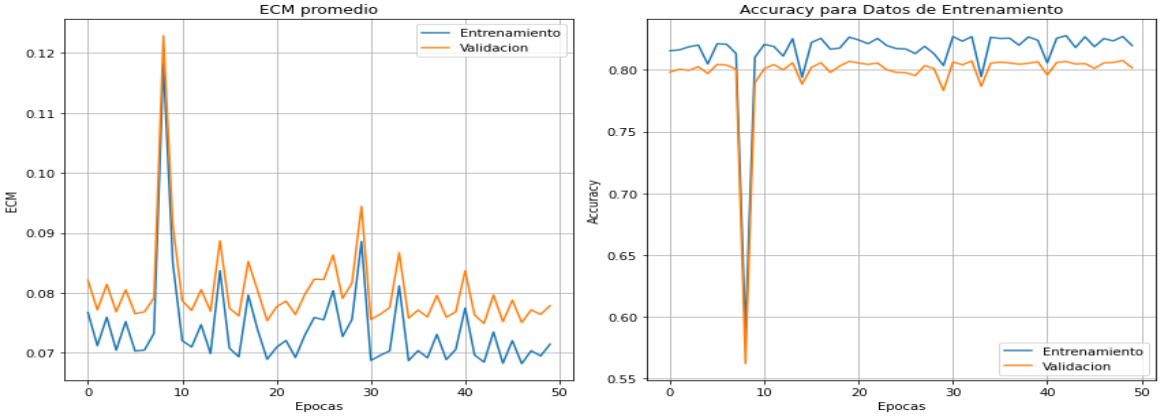


Más aún, si intentamos limpiar los datos de tal manera que nos quedamos con la parte más significativa de cada atributo, se terminaría tomando menos del 80% de los datos originales.

Particionamiento: Se particionó los datos tomando ⅔ de los datos para entrenamiento y ⅓ para validación, tomando el primer tercio según el orden en el que vienen dados en el archivo, como datos de entrenamiento. Esto dado que la distribución de los datos era razonablemente uniforme entre ambos conjuntos. Quedando 20000 datos de entrenamiento y 10000 de validación.

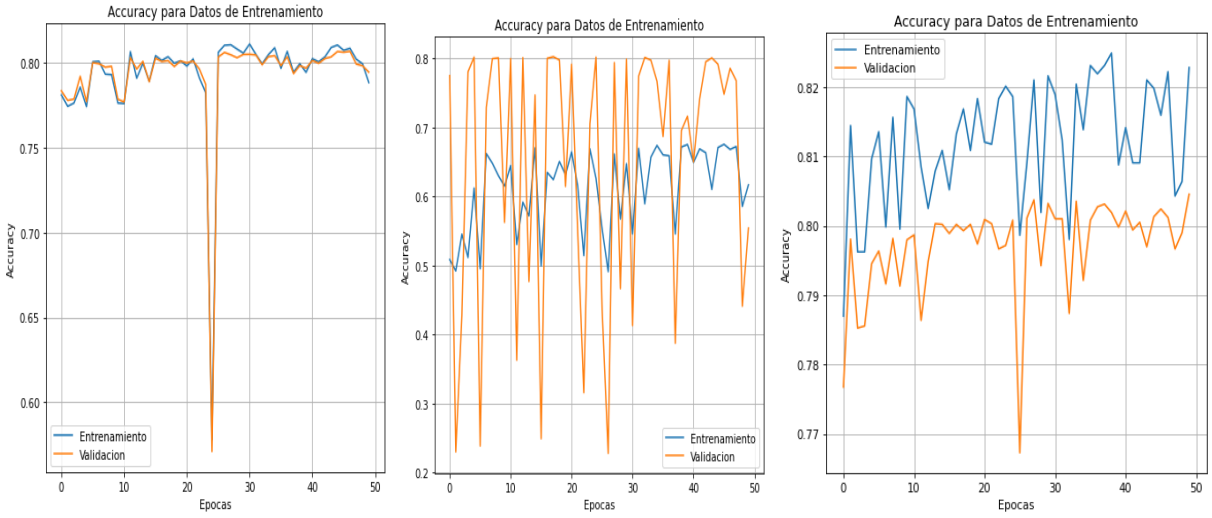
MLP

Como primera red de clasificación se empleó un perceptrón multicapas o MLP. La mejor arquitectura encontrada fue usando la implementación de la tarea 4, con 1 capa oculta de 46 neuronas, con función de activación Logística y 1 neurona en la capa de salida. Una tasa de aprendizaje de 0.1, parámetro de momentum igual a 0.5. Con estas configuraciones encontramos una mejor precisión para datos de validación de 80.17%. Se intentó aumentar el número de capas ocultas, variar el momentum, la tasa de aprendizaje y el número de épocas, sin embargo la ganancia en complejidad no producía un aumento significativo de la precisión final.



Boosting usando MLP

Se aplicó el método descrito por la profesora Minaya Villasana en [2]. Para esto se particionaron los datos de entrenamiento en 3, para cada uno de los expertos E1, E2 y E3, y mantenemos los datos de validación anteriores. Cada experto es un MLP con las mismas configuraciones anteriores. La precisión conseguida con esta estrategia fue de 80.4%, apenas 0.13% mejor en cuanto a la encontrada con MLP.



De izquierda a derecha la evolución de la precisión de los expertos E1, E2, E3 a través de sus épocas de entrenamiento.

Conclusiones

Desde el estudio inicial de los datos, se observó que el conjunto escogido era bastante problemático y se decidió afrontarlo sin realizar ninguna modificación a los datos más allá del escalamiento. El Boosting con MLP dio mejores resultados que el MLP individual, como era esperado, sin embargo la mejora no fue muy significativa. Aunque el mejor porcentaje de acierto encontrado, 80.4%, en términos absolutos no es bueno, considerando las grandes variaciones e irregularidades presentados en los datos, se estima que la red hizo un buen trabajo intentando generalizar el comportamiento.

Bibliografía

- [1] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.
- [2] Villasana, Minaya (2020). Perceptron Multicapas (Backpropagation). Lámina 81.