Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101) Reporte sobre el desempeño del modelo

JESÚS YAIR RAMIREZ ISLAS A01275404[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey*

[1] *A01275404@TEC.MX*

Compiled September 14, 2023

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. ELECCIÓN DE LA DATA SET:

El dataset seleccionado tiene como nombre titanic.csv, en este se encuentra información ya clasificada sobre diversos pasajeros que estuvieron en el barco, es un conjunto de datos altamente conocido. Para trabajar con el mismo fue necesaria una serie de tratamientos en los datos, como la eliminación de columnas no relevantes como fue el caso de el id de cada pasajero y su nombre, el tratamiento de datos faltantes en algunos casos se eliminaron los registros si estos eran pocos o en casos como la edad se remplazaron estos valores vacíos por la mediana de los datos. Para el caso de variables categóricas se utilizó labelecconding para transformar estos datos a numéricos, finalmente se estandarizaron los datos para que fuera más fácil trabajar con ellos. El data set se eligió por que deseaba conocer las capacidades de las redes neuronales en un problema que implicara una cantidad considerable de variables, además al conocer un poco de este modelo y considerando que se deseaba predecir entre dos clases tenia una idea de que arquitectura podría ser util en este caso, si bien los datos no tenían en un inicio las características ideales para trabajar en un modelo que trabaja con números y se ve afecto por la variabilidad en los datos, tras el tratamiento adecuado antes mencionado considero que ya eran apropiados para el algoritmo y para demostrar que tan buen desempeño tiene el modelo.

2. ANÁLISIS DE BIAS, VARIANZA Y AJUSTE DEL MODELO

Cuando el modelo se implementó inicialmente, se utilizó la función *biasvariancedecomp* la cual nos da entre otras cosas el valor de sesgo y varianza que presenta el modelo a través de una serie de rondas de muestreo, el modelo obtuvo un valor de bias de .2224 y una varianza de .0002. Con esto en cuenta podríamos decir que el modelo presenta un sesgo alto en comparación de la varianza, la cual es extremadamente baja, esto nos podría estar hablado de underfiting o que el modelo es demasiado simple para los datos, si contrastamos esto con los valores presentes de accuracy en el conjunto de prueba y validación (imagen de los valores) podemos observar que, si bien hablan de un modelo aceptable, nunca llegan a superar el 85%, lo cual respaldaría nuestra teoría. Dados los requerimientos de este reporte y con el fin de observar si estas observaciones son correctas realizaremos un cross validation k-fold con 5 splits para el modelo, de esta forma no solo obtendremos los valores mencionados en un inicio si no que además podremos observar el comportamiento de otras métricas tanto para el conjunto de validación como el de prueba. Es importante considerar que se hicieron dos separaciones de datos, en un inicio la separación se realizo sobre todos los datos como se muestra en la imagen1 y la figura 2,

para el cross validation la data se dividió en la forma en la que se muestra en las imanges: 2 4

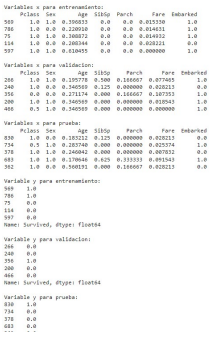

Fig. 1. Aqui observamos la forma en la que se dividio el dataset

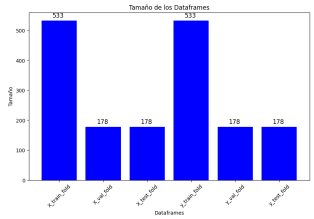

Fig. 2. Histograma de la division de data set

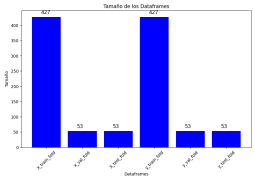

Fig. 3. Aqui observamos la forma en la que se dividio el dataset en el cross validation

Tras el cross validation podemos observar en la grafica 5 que el bais continua con valores bastante similares los de un inicio con cierto cambio en el flod 3 pero poco significativo. Sin embargo, en el conjunto de prueba el sesgo es menor y en el fold 3, lo cual podría decirnos que nuestro modelo se comporta mejor en datos con los que no haya trabajado. Lo mismo sucede con la varianza que si bien presenta un comportamiento mas variado los valores siguen siendo pequeños 6. En este caso vemos que el conjunto de prueba nuevamente tiene un comportamiento distinto, ya que parece que tener un valor estable a lo largo de los folds. Si observamos la varianza y sesgo promedio tiene valores similares a los de un inicio, con una reducción en la varianza

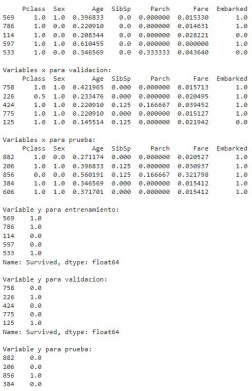

Fig. 4. Histograma de la division de data set en el cross validation

para el caso del conjunto de prueba, por lo que parece que nuestro modelo tiene las mismas limitantes aun haciendo cambios en los conjuntos de datos implementados. Como lo mencione el objetivo de este proceso era también poder evaluar las métricas en los diferentes folds por lo que si observamos las gráficas 12, las métricas 14 y la desviación estándar de las mismas 13. Podemos observar que el valor promedio de la precisión (accuracy) en el conjunto de validación es aproximadamente 0.80, lo que sugiere que el modelo tiende a predecir correctamente alrededor del 80% de las instancias. Sin embargo, al examinar los valores de precisión, recall y F1-score en diferentes folds de validación, notamos que hay variabilidad en estas métricas. Por ejemplo, el recall varía de 0.40 a 0.76 en diferentes folds. Esto indica que el modelo puede tener un sesgo variable en diferentes conjuntos de datos de validación. Pues un valor bajo de recall (0.4) en un fold sugiere que el modelo no está identificando adecuadamente algunas instancias positivas, lo que podría considerarse un sesgo hacia las instancias negativas. Por lo tanto, el sesgo del modelo no se limita a la precisión, sino que se extiende a otras métricas como recall y F1-score. Por otro lado, en el conjunto de prueba el valor promedio de accuracy es un poco mejor .82, asi mismo los valores de recall y f1- score no presentan valores tan bajos como el conjunto de validación, esta mejora también se observa en la desviación estándar pues en el caso del conjunto de prueba las métricas tienen una desviación menor, lo cual habla de un mejor desempeño en el conjunto de prueba. Para el caso de la varianza vemos que la desviación estándar del accuracy en la validación es de aproximadamente 0.034, lo que indica una varianza moderada en el rendimiento del modelo en diferentes

folds de validación. Esto podría deberse a que el modelo es sensible a la composición específica de los datos en cada fold, lo que sugiere cierta inestabilidad. Para el caso del conjunto de prueba vemos nuevamente una mejoría con un valor de 0.047. La varianza no solo se aplica a la accuracy, sino que también se refleja en las desviaciones estándar de otras métricas como precision, recall y F1-score. Por ejemplo, una desviación estándar de 0.11 en la precisión de validación indica que el modelo puede tener un rendimiento variable en diferentes folds en términos de identificar correctamente las instancias positivas y para este caso no existe mejoría pues en el conjunto de prueba el valor es de 0.096. La estabilidad y consistencia del modelo se reflejan en las desviaciones estándar de las métricas. En nuestro caso podemos concluir que el modelo tiene un ajuste menos robusto a diferentes conjuntos de datos. Es decir, tras las distintas observaciones podemos observar cierto underfitting Esto sugiere que el modelo podría beneficiarse de una mayor regularización o de una selección más cuidadosa de hiperparámetros para reducir la variabilidad en su rendimiento. Es importante destacar que no hay evidencia clara de overfitting en el modelo. Pues el sobreajuste generalmente se caracteriza por un rendimiento deficiente en el conjunto de prueba en comparación con la validación. En este caso, las métricas en el conjunto de prueba no son significativamente peores que las de validación, lo que sugiere que el modelo generaliza de manera razonable.

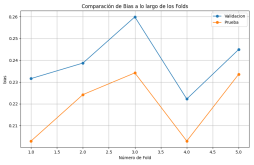

Fig. 5. Aqui observamos el comportamiento del bias a lo largo de los folds

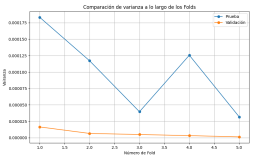

Fig. 6. Aqui observamos el comportamiento del bias a lo largo de los folds

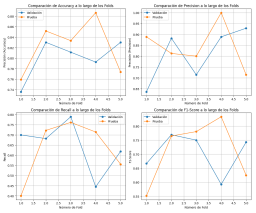

Fig. 7. Metricas del primer modelo en los folds

Fig. 8. Desviaciones estandar de las metricas

3. AJUSTE DE PARÁMETROS PARA MEJORAR EL DESEMPEÑO

Como ya lo observamos el modelo parece ser simple para el problema que estamos abordando por lo que pasaremos de la arquitectura que teníamos compuesta por: 4 capas con 64, 32,16 y 1 neuronas, funciones de activación relu para las primeras 3 y sigmoide para la última, una función de perdida de *binarycrossentropy* optimizador de Adam, 50 épocas y batch size de 32. A uno más complejo con 5 capas con 128, 64, 32, 16 y 1 neuronas, de la misma manera todas las capas menos la ultima tiene función relu, utilizamos la misma función de perdida, optimizador de Adam pero en este caso con un learning rate de 0.01 y aumentamos las épocas a 70. Con estas nuevas funcionalidades nuestro modelo debería ser mas robusto y obtener mejores resultados.

Tras realizar la evaluacion podemos observar una clara mejoria en las metricas del modelo 1 14 y las del modelo 2 10 si bien en algunas la mejoria es pequeña como el caso del Accurcy en la que las pruebas con el conjunto de validacion aumentaron un .037 y para el conjunto de prueba el aumento fue .008. Las mejorias se notan en la grafica 11

Asi mismo observamos que la varianza parece estar comportandose de manera distinta pues como se muestra en la grafica **??**, en los ulitimos folds la varianza entre el conjunto de prueba y entrenamiendo empieza a paracerse, en comparacion al comportamiento anterior6. Sin embargo en el Bias no parace existir un cambio en el comportamiento**??**

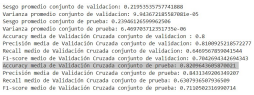

Fig. 9. Metricas del primer modelo

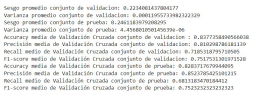

Fig. 10. Metricas del modelo 2

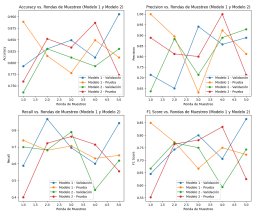

Fig. 11. Aqui el comportamiento de los 2 modelos para sus conjuntos en los folds

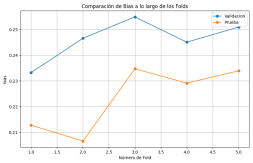

Fig. 12. sesgo en el modelo 2

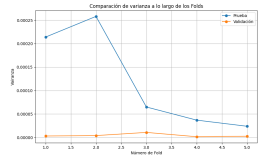

Fig. 13. Variacion del modelo 2 en los folds