“Contraste de los clasificadores entrenados”

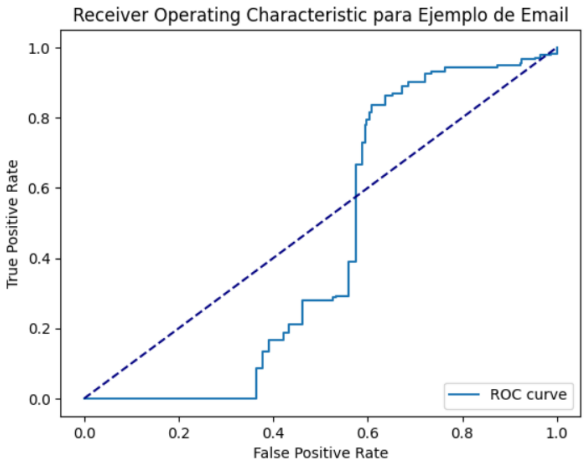
Autor: Rodolfo Jesús Cruz Rebollar **Matrícula:** A01368326

En primera instancia, en la presente actividad se generaron en total 4 modelos de clasificación: regresión logística, Support Vector Machine, Random Forest y Gradient Boosting, por lo que el primero de ellos, es decir el modelo de regresión logística resultó tener un desempeño mayormente favorable para clasificar los correos electrónicos que son spam y los que no lo son, siendo su valor de exactitud (accuracy) del 64.66%, lo cual significa que de todos los emails clasificados por el modelo, éste mismo logró clasificar de manera correcta el 64.66% de los mismos, por lo cual al tener en total 517,401 emails, el 64.66% de dicha cantidad equivale a haber clasificado correctamente 334,552 emails del total, mientras que al mismo tiempo, también se implementó un modelo clasificador de Support Vector Machine, mismo que resultó tener un desempeño para clasificar igualmente favorable en su mayoría, dado que el nivel de exactitud para el Support Vector Machine fue del 69.5%, indicando que dicho modelo fue capaz de clasificar correctamente el 69.5% de todas las instancias o emails, lo cual equivale a haber clasificado de forma acertada aproximadamente 359,594 emails del total mencionado anteriormente, motivo por el cual es posible notar que el modelo de Support Vector Machine tuvo un desempeño medianamente mejor que el modelo de regresión logística, dado que el Support Vector Machine logró clasificar correctamente 25,042 emails más que la regresión logística, lo cual a su vez representa una mejoría considerable en cuanto a la exactitud de las clasificaciones realizadas.

No obstante, de forma similar, también se implementó un tercer modelo de clasificación, particularmente de tipo Random Forest (Bosque Aleatorio), el cual, al igual que los 2 modelos previamente discutidos, también logró tener un desempeño mayormente favorable para clasificar los emails, esto principalmente debido a que éste modelo logró tener un grado de exactitud del 72.66%, lo cual equivale a que éste modelo consiguió clasificar adecuadamente un aproximado de 375,944 emails del total, lo cual también representa una mejoría considerable respecto al modelo de Support Vector Machine, dado que éste primer modelo de random forest consiguió clasificar de forma correcta 16,350 emails más que el modelo de Support Vector Machine, propiciando que la mejora en la clasificación de correos fraudulentos sea mucho más notoria que en el caso de los modelos anteriormente descritos, sin embargo, también es importante mencionar que posterior al primer modelo de Random Forest, se implementó otro modelo del mismo tipo, pero mejorado en el sentido de haber encontrado los mejores parámetros que maximizan el grado de exactitud de las predicciones de dicho modelo, esto se realizó por medio de la función GridSearchCV aplicada sobre los parámetros del modelo de Random Forest original, por lo cual, después de ejecutar dicha función para optimizar los hiper parámetros del modelo, se obtuvo finalmente como resultado que los mejores parámetros para maximizar la exactitud predictiva del modelo fueron los siguientes: “min_samples_leaf” (mínimo requerido de muestras para estar en una hoja o nodo de un árbol del bosque) con un valor de 3, además del parámetro “min_samples_split” (cantidad mínima de divisiones muestrales) con un valor de 10, aunado al parámetro “n_estimators” (cantidad de estimadores requerida para la optimización) con un

valor de 100, motivo por el cual, al poner a prueba nuevamente el modelo de Random Forest, pero esta vez empleando los valores anteriores como parámetros del modelo, se logra mejorar sutilmente el desempeño clasificatorio del modelo, dado que ahora el nivel de exactitud del mismo pasa a ser del 73%, lo cual resulta ser una exactitud ligeramente mejor que la del random forest inicial (72.66%), por lo cual, cabe mencionar que el random forest optimizado resulta ser sutilmente mejor (la mejoría en la exactitud es solo del 0.34%) para determinar si los emails son o no spam que su predecesor, además del hecho de que en general, el modelo de Random Forest optimizado resulta superar en términos de exactitud clasificatoria al random forest inicial (la exactitud mejora en un 0.34%), al support vector machine (la mejoría es del 3.5%) y también a la regresión logística (con una mejoría del 8.34%), motivo por el cual, el modelo optimizado de random forest es el mejor modelo clasificatorio para el spam en los emails hasta el momento. Adicionalmente, también cabe mencionar que finalmente se implementó también un modelo de clasificación de Gradient Boosting, el cual consiguió clasificar de forma correcta al 73.33% de todos los emails que se tienen, lo cual a su vez equivale a que dicho modelo clasificó acertadamente una cantidad aproximada de 379,411 emails de los 517,401 emails que se tienen en total, por lo tanto, es posible evidenciar que aquel modelo que tiene el mayor porcentaje de exactitud para clasificar a los emails como spam o no spam, es el clasificador de Gradient Boosting, el cual supera a todos los modelos antes entrenados, por lo cual, es posible concluir que nos quedaremos con el clasificador de Gradient Boosting como el mejor para clasificar los emails con el menor margen de error posible.

Por último, otro aspecto relevante consiste en realizar una curva ROC para observar cómo cambia la proporción de casos verdaderos positivos (emails que en realidad eran spam y el modelo los clasificó como tal) en función de los casos falsos positivos (emails que el modelo predijo que eran spam pero realmente no eran fraudulentos), por lo cual, tendiendo esto en cuenta, es posible notar que en la gráfica de la curva ROC, conforme la cantidad de casos falsos positivos aumenta, también incrementa la cantidad de casos verdaderos positivos que predice el modelo, es decir, que el clasificador, durante su etapa de entrenamiento, conforme más errores comente en términos de predecir de manera incorrecta la clasificación de un determinado email, mayor es el aprendizaje que adquiere el modelo mismo para realizar futuras clasificaciones y como resultado de dicho aprendizaje, el modelo predice correctamente la clasificación de cada vez más emails, lo cual provoca que la tasa de verdaderos positivos aumente de forma gradual, motivo por el cual, entre más grande sea la tasa de falsos positivos (que representa los errores que comente el modelo al momento de clasificar), la tasa de verdaderos positivos igualmente será mayor, reflejando de esa manera, la mejoría gradual del modelo para clasificar correctamente los emails que se le proporcionen, como consecuencia del aprendizaje que adquiere, derivado de sus errores pasados, por lo que al final del proceso, el modelo llega a tener una tasa de verdaderos positivos muy cercana a 1, que ocurre justamente cuando la tasa de falsos positivos también toma su máximo valor posible que también es 1, indicando que el mayor aprendizaje del modelo sucede cuando ha cometido la máxima cantidad de errores de clasificación.



Curva ROC de la mejoría clasificatoria del modelo en función de sus errores cometidos. Fuente: elaboración propia.