Fundamentos de Machine y Deep Learning

Ejercicio N° 1 (Mejoras en el Desempeño de Modelos de Clasificación Resuelto)
02 / 09 / 2023

INTEGRANTES

- Camila Constanza Aguilera Bustamante
- Wladimir Richard Parada Rebolledo
- Néstor Patricio Rojas Ríos
- Ramiro Andrés Uribe Garrido

Pregunta 1

¿Cuáles columnas se eliminaron logrando mejorar el desempeño de los modelos?

Luego de múltiples iteraciones, el desempeño del modelo mejoró al eliminar las columnas ***NumEmpleados***, ***DíasAtrás***, ***ResultadoPrevio*** y ***Euribor3m*** (*Figura 1*, matriz de confusión al usar todas las variables; *Figura 2*, con las variables excluidas).

Si bien se reducen los valores de especificidad (desde ***96,1%*** antes al ***94,1%*** después) y de exactitud (***91,3%*** antes, ***90,5%*** después), la sensibilidad mejora más de 9 puntos porcentuales (***52,9%*** antes hasta ***62%*** después), por lo que, en general, este modelo tiene mejor desempeño.

Pregunta 2

¿Qué proporción (Probar: 0.8, 1.0, 1.2, 1.5, 1.8) de ejemplos NO/YES se eligió y por qué? Entonces, ¿de qué dimensiones quedó el dataset de entrenamiento modificado?

Considerando el dataset limpio obtenido en el punto anterior, es que se escoge la proporción ***1,5*** de ejemplos *NO*/*YES*. Se escogió esta proporción dado que mejora el valor de la especificidad obtenida con la proporción 1 a 1 (desde ***77,31%*** a ***85,77%***), dejándolo en un valor casi equitativo del valor de sensibilidad, que cae un poco (de ***92,95%*** a ***86,18%***) y mejorando el valor de la exactitud (de ***84,86%*** a ***85,93%***).

Finalmente, el dataset de entrenamiento queda constituido por ***6.270 registros*** al usar la nueva proporción. Por su parte, el dataset de validación contiene ***3.377 observaciones***. (*Figura 3*, matriz de confusión con la proporción 1:1 de ejemplos *YES*:*NO*; *Figura 4*, la matriz con la proporción 1:1,5).

Pregunta 3

Si se compara con el desempeño del Árbol de Decisión en el paso 3 anterior ¿Qué diferencia fundamental (ventaja/desventaja) se ve en los indicadores de desempeño y por qué se podría haber logrado esta diferencia?

Al aplicar el modelo de ***Random Forest*** sobre los datos originales del ejercicio, descartando sólo los registros con datos faltantes, se puede ver cambios importantes en todos los indicadores: la sensibilidad aumenta de un ***52,93%*** a un ***53,49%***; por otro lado, descienden en un margen de menos del **1%,** tanto la especificidad (de un ***96,12%*** a un ***96,08%***), como la exactitud (de un ***91,33%*** a un ***90,74%***, *Figura 5,* matriz de confusión al usar el método de Random Forest; *Figura 1,* la matriz original de la pregunta 1).

Para comprender correctamente por qué se genera esta alza en los parámetros es necesario entender bien ambos modelos: ***Decision Tree*** y ***Random Forest***. Entiéndase el algoritmo de ***Decision Tree*** que se utiliza en este caso para clasificación, donde básicamente se dividen iterativamente los datos en subgrupos más pequeños en base a las características de los datos buscando construir un árbol que permita tomar decisiones basada en las características de entrada. Un detalle a tener en cuenta sobre este algoritmo es que suele generarse sobreajuste, guiando a obtener un bajo rendimiento con los datos de validación.

Por otra parte, considerando el algoritmo de ***Random Forest*** como una técnica que se basa en la construcción de múltiples árboles de decisión durante el entrenamiento donde se combinan los resultados para obtener una predicción robusta y precisa. En simples palabras, cada árbol de decisión se entrena con un subconjunto aleatorio de los datos de entrenamiento para posteriormente combinarse por votación para obtener el resultado. Cabe destacar que al utilizar subconjuntos aleatorios permite evitar el sobreajuste generando un mayor rendimiento al momento de probar el modelo con el conjunto de validación.

Una vez descritos los modelos presentes en esta comparación es momento de aclarar por qué la variación en los parámetros fue tan leve. En este caso, se sustenta dada la ***posibilidad de errores en los datos*** y la ***falta de balanceo*** entre las categorías de la variable endógena, donde la categoría *NO* es aproximadamente ***7 veces mayor*** a la categoría *YES*.

Pregunta 4

Al comparar el desempeño del modelo con los datos de entrenamiento y con los datos de evaluación, se puede ver una diferencia importante. ¿Cómo se interpreta esta diferencia y, en teoría, cómo se podría resolver?

Una diferencia significativa se ve en el desempeño del ***Random Forest*** entre los datos de entrenamiento y los datos de validación: existe disminución en la precisión (***99,64%*** a ***90,74%***), en la especificidad (***99,98%*** a ***96,08%***), pero más considerable en la sensibilidad (***97,33%*** a ***53,49%***). Estás diferencias son explicables porque ambos conjuntos de datos tienen características diferencias en sus variables (el desbalance entre *YES* y *NO*, además de la posible presencia de datos erróneos o registros incompletos), generando *overfitting* (sobre ajustamiento) del modelo con los datos con los que fue entrenado *(Figura 6*, matriz de confusión de entrenamiento; *Figura 5*, la matriz del dataset de validación). Las estrategias para corregir esto pueden ser:

- Limpieza de datos: Ya sea agregando información faltante, corrigiendo aquellos que están erróneos o directamente eliminando registros *outliers* o erróneos.
- Balance de datos: Se puede generar un balance en los datos, disminuyendo la cantidad de registros de la categoría más representada en la variable endógena.
- Redistribución de datos: Cambiar los conjuntos de datos de entrenamiento y validación, por ejemplo, usando la estrategia de *cross-validation*.

Anexos

```
****Desempeño Decision Tree en conjunto de evaluación****
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  14080   860
       yes   569   967

               Accuracy : 0.9133
                 95% CI : (0.9089, 0.9175)
    No Information Rate : 0.8891
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5272

 Mcnemar's Test P-Value : 1.7e-14

            Sensitivity : 0.52928
            Specificity : 0.96116
         Pos Pred Value : 0.62956
         Neg Pred Value : 0.94244
             Prevalence : 0.11089
         Detection Rate : 0.05869
   Detection Prevalence : 0.09323
      Balanced Accuracy : 0.74522

       'Positive' Class : yes
```

Figura 1. Matriz Decision Tree con conjunto de evaluación incluyendo todas las variables.

```
****Desempeño Decision Tree en conjunto de evaluación 2****
*Sin NumEmpleados, DíasAtrás, ResultadoPrevio ni Euribor3m*
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  13784   695
       yes   865  1132

               Accuracy : 0.9053
                 95% CI : (0.9007, 0.9097)
    No Information Rate : 0.8891
    P-Value [Acc > NIR] : 6.929e-12

                  Kappa : 0.5386

 Mcnemar's Test P-Value : 1.879e-05

            Sensitivity : 0.61959
            Specificity : 0.94095
         Pos Pred Value : 0.56685
         Neg Pred Value : 0.95200
             Prevalence : 0.11089
         Detection Rate : 0.06871
   Detection Prevalence : 0.12121
      Balanced Accuracy : 0.78027

       'Positive' Class : yes
```

Figura 2. Matriz Decision Tree con conjunto de evaluación excluyendo las variables elegidas.

```
****Desempeño Decision Tree en conjunto de entrenamiento 2****
****Balance entre SÍ y NO EN 1****
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  1080   92
       yes  317 1213

               Accuracy : 0.8486
                 95% CI : (0.8346, 0.8619)
    No Information Rate : 0.517
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6986

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9295
            Specificity : 0.7731
         Pos Pred Value : 0.7928
         Neg Pred Value : 0.9215
             Prevalence : 0.4830
         Detection Rate : 0.4489
   Detection Prevalence : 0.5662
      Balanced Accuracy : 0.8513

       'Positive' Class : yes
```

Figura 3. Matriz Decision Tree con conjunto de entrenamiento limpio con proporción 1,0.

```
****Desempeño Decision Tree en conjunto de entrenamiento 2****
****Balance entre SÍ y NO EN 1,5****
Confusion Matrix and Statistics

          Reference
Prediction   no  yes
       no  1730  188
       yes  287 1172

               Accuracy : 0.8593
                 95% CI : (0.8472, 0.8709)
    No Information Rate : 0.5973
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.711

 Mcnemar's Test P-Value : 6.907e-06

            Sensitivity : 0.8618
            Specificity : 0.8577
         Pos Pred Value : 0.8033
         Neg Pred Value : 0.9020
             Prevalence : 0.4027
         Detection Rate : 0.3471
   Detection Prevalence : 0.4320
      Balanced Accuracy : 0.8597

       'Positive' Class : yes
```

Figura 4. Matriz Decision Tree con conjunto de entrenamiento limpio con proporción 1,5.

```
****Desempeño Random Forest en conjunto de evaluación****
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  10276   714
       yes   419   821

               Accuracy : 0.9074
                 95% CI : (0.9021, 0.9124)
    No Information Rate : 0.8745
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5401

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.53485
            Specificity : 0.96082
         Pos Pred Value : 0.66210
         Neg Pred Value : 0.93503
             Prevalence : 0.12551
         Detection Rate : 0.06713
   Detection Prevalence : 0.10139
      Balanced Accuracy : 0.74784

       'Positive' Class : yes
```

Figura 5. Matriz Random Forest con conjunto de validación limpio.

```
****Desempeño Random Forest en conjunto de entrenamiento****
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  15931    62
       yes     3  2262

               Accuracy : 0.9964
                 95% CI : (0.9955, 0.9973)
    No Information Rate : 0.8727
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9838

 Mcnemar's Test P-Value : 6.291e-13

            Sensitivity : 0.9733
            Specificity : 0.9998
         Pos Pred Value : 0.9987
         Neg Pred Value : 0.9961
             Prevalence : 0.1273
         Detection Rate : 0.1239
   Detection Prevalence : 0.1241
      Balanced Accuracy : 0.9866

       'Positive' Class : yes
```

Figura 6. Matriz Random Forest con conjunto de entrenamiento limpio.