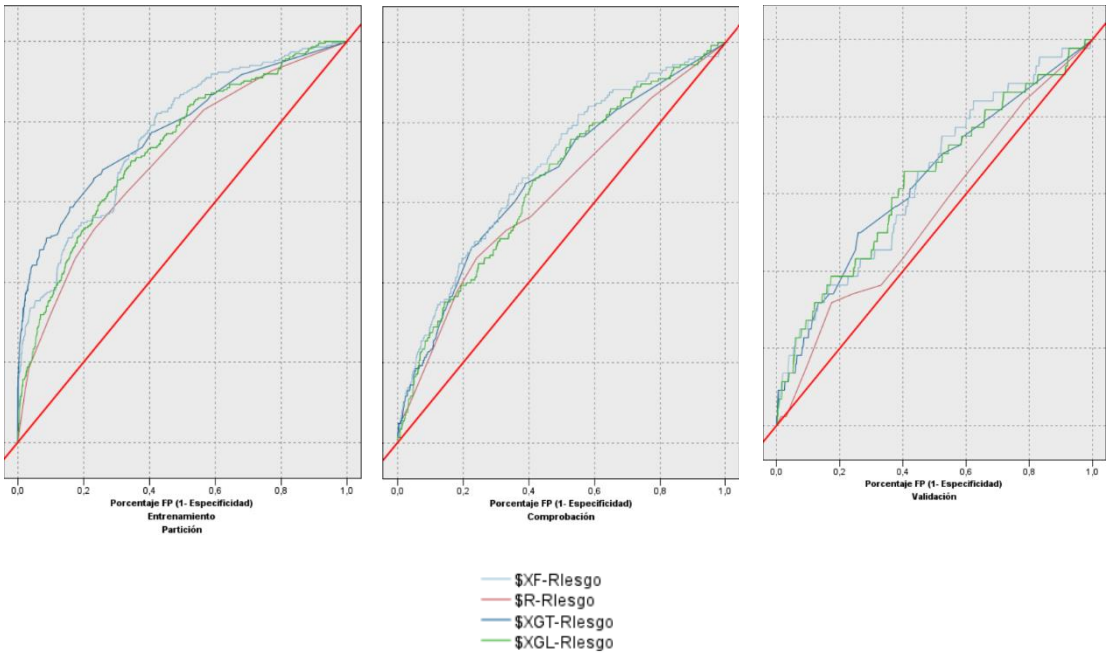
Evaluación del modelo

A continuación, procederemos evaluar los resultados obtenidos en los distintos modelos de XGBoost, XGBoost Lineal y Chaid, los 3 algoritmos que mayor área bajo la curva tuvieron en el análisis del clasificador automático (XGBoost, XGBoost Lineal y Chaid). Comenzaremos analizando las curvas ROC de las 3 muestras:

Entrenamiento Comprobación Validación



Métricas de evaluación

'Partición'	Comprobación		Entrenamiento		Validación	
Modelo	AUC	Gini	AUC	Gini	AUC	Gini
$XF-Riesgo	0,686	0,373	0,765	0,53	0,638	0,275
$R-Riesgo	0,62	0,24	0,706	0,412	0,55	0,101
$XGT-Riesgo	0,66	0,32	0,778	0,556	0,631	0,262
$XGL-Riesgo	0,652	0,303	0,734	0,469	0,633	0,266

La línea $XF-Riesgo representa al clasificador automático, $R-Riesgo el modelo CHAID, $XGT-Riesgo el árbol XGBoost, y $XGL-Riesgo el modelo XGBoost Lineal.

En nuestro caso observamos como el modelo XGBoost es el que mejor se ajusta a la muestra de entrenamiento con un área bajo la curva ROC de casi 0.78, sin embargo, el número no se mantiene constante para las otras 2 muestras, estando ambas en torno a 0.66. Analizando el resto de las modelos, parece que ninguno consigue mejorar los números de XGBoost, por lo que analizaremos más detalladamente los resultados de este modelo. La siguiente imagen muestra la matriz de confusión comprando las predicciones con los valores reales:

Matriz de coincidencias para $XGT-Riesgo (las filas muestran las reales)

'Partición' = Comprobación	False	True
False	1.576	7
True	139	6
'Partición' = Entrenamiento	**False**	**True**
False	3.299	3
True	201	34
'Partición' = Validación	**False**	**True**
False	555	1
True	43	1

La matriz confirma el sobreajuste del modelo sobre la muestra de entrenamiento. De los 235 malos reales, el modelo es capaz de identificar tan solo a 34 (14,46%). El porcentaje de identificación es todavía menor para las muestras de comprobación y validación (4,13% y 2,27% respectivamente).

Podemos concluir que el modelo por un lado no es preciso, ya que no es capaz de identificar correctamente a un porcentaje significativo de los peores créditos y tampoco es estable, al ser el porcentaje de identificación muy diferente en las 3 muestras. Como principal causa está que en todos los casos los modelos han predicho como False la gran mayoría de las observaciones. Esto puede deberse por un lado al gran desbalanceo entre observaciones de créditos buenos y malos, o por otro lado en que las variables utilizadas en los algoritmos no han sido capaces de ser los suficientemente explicativas a la hora de determinar si un crédito iba a ser malo.

Como posibles soluciones, se podría realizar una mejor selección de las variables iniciales o la mayor creación de nuevas columnas a partir de las que ya teníamos. Otra opción podría estar en dar mayor peso a aquellas observaciones de créditos malos (sobre muestreo/submuestreo) para evitar así el sesgo observado en las predicciones de los algoritmos estudiados.