Experimento Propio

Zoe Borrone Luca Mazzarello Ignacio Pardo

2023-08-21

Introducción y Metodología

Para desarrollar este experimento se planteo testear el comportamiento de los modelos al aplicarle ruido a los valores de los atributos numericos en distintas proporciones, viendo si esto afecta el desempeño de los modelos.

Para esto se modificó el script de ejemplo `sample_exp.R` para aplicar ruido a cada dataset pasado a la función `run_experiment` en distintas proporciones. El ruido se aplica de la siguiente manera:

A través de las distintas proporciones de ruido a aplicar: `prop_noise` $\in 0, 0.2, 0.4, 0.6, 0.8$.

- Para cada dataset dado: se obtienen las columnas de atributos numericos.
- Para cada columna de atributos numericos: se obtiene el valor minimos y el valor maximo.
- Para cada proporción `prop_noise`,
- Para cada valor de cada atributo numerico,
- Se genera un valor aleatorio r entre el minimo y el maximo con una distribución uniforme.
- Se genera un segundo valor aleatorio uniforme p entre 0 y 1.
- Si \$p $<$ \$ `prop_noise`, se reemplaza el valor original por r.

Luego, cada valor número va a haber sido reemplazado por un ruido entre los valores conocidos de la columna, con una probabilidad `prop_noise`.

Algo que cabe detallar, optamos por aplicar este ruido luego de haber imputado datos faltantes. En este sentido el ruido que simulamos podría tratarse a de ruido a la hora de registrar los datos y no una vez que el dataset ya se armó. Esto lo hicimos para poder comparar con los resultados de los experimentos anteriores.

Como notamos en el experimento anterior (opción 3), los datasets cuentan con distintas proporciones de variables numéricas, el dataset de *Heart* cuenta con 6 variables numéricas de 11 atributos totales, el dataset de *Churn* cuenta con 10 variables numéricas de 13 atributos totales y el dataset de *Student* cuenta con 13 variables numéricas de 30 atributos totales. Si los modelos priorizan los atributos numéricos, esperamos que el ruido afecte más a los modelos de *Heart* y *Churn* que al de *Student*.

Resultados

Churn

Para el dataset de *Iranian Churn* que cuenta con mayoría de atributos numéricos, se puede observar que la performance del modelo disminuye a medida que crece el ruido porcentual aplicado a los atributos numéricos. Podemos ver en la Tabla 1 como el máximo AUC promedio a traves de los distintos `max_depths` de arboles disminuye a medida que aumenta el ruido.

Ademas, tanto para cuando se imputan datos faltantes como para cuando no, el `max_depth` que maximiza el AUC promedio decrece a medida que aumenta el ruido.

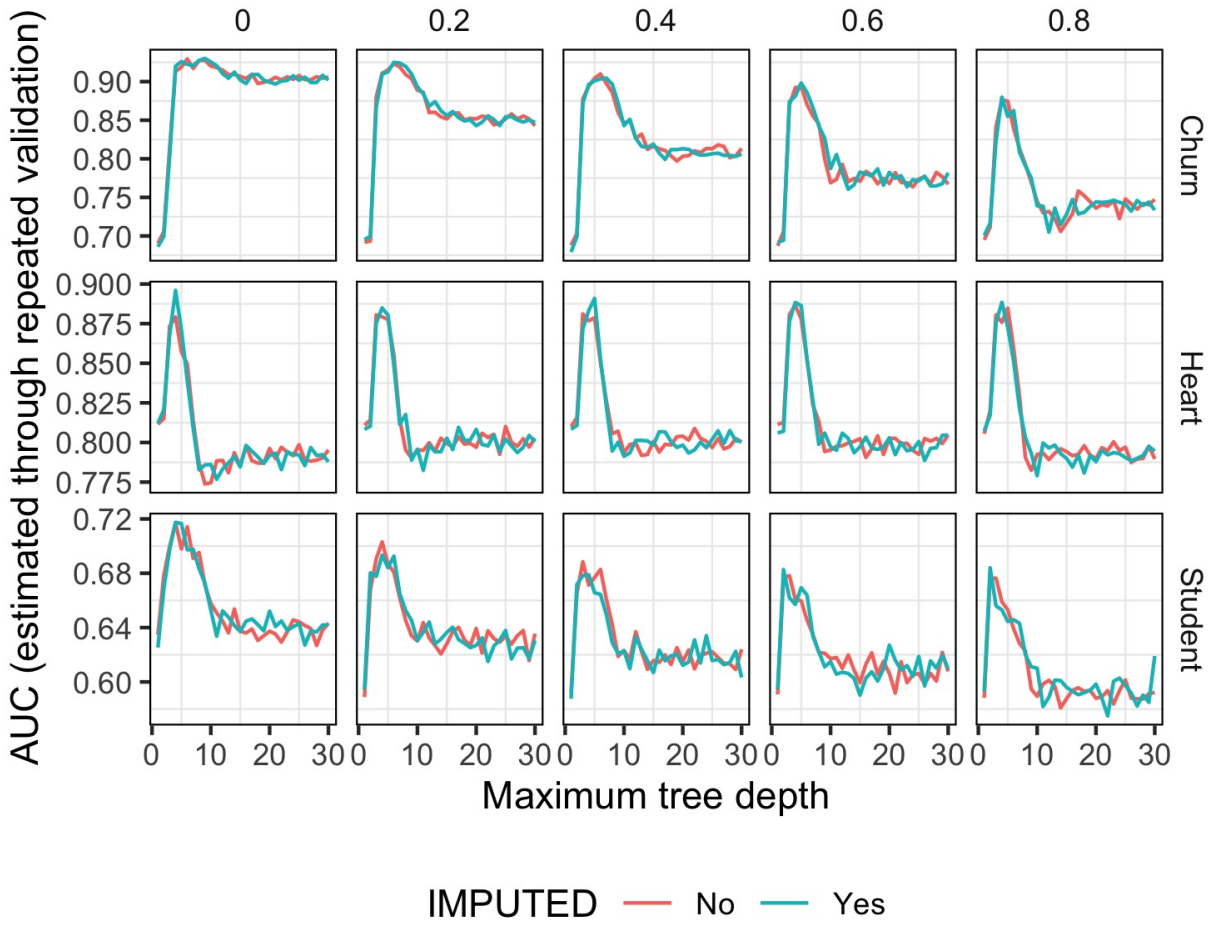

Figure 1: Resultados `prop_noise` $\in 0, 0.2, 0.4, 0.6, 0.8$

Table 1: Max mean AUC a traves de profundidades de arboles maximas en Iranian Churn y metodos de imputacion

maxdepth	IMPUTED	mean_auc	noise
6	No	0.9294095	0.0
6	No	0.9230376	0.2
6	No	0.9099543	0.4
5	No	0.8938293	0.6
4	No	0.8765402	0.8
9	Yes	0.9300399	0.0
6	Yes	0.9249940	0.2
7	Yes	0.9045746	0.4
5	Yes	0.8982098	0.6
4	Yes	0.8797933	0.8

Algo interesante que notamos, es que para profundidades de árboles mayores, a mayor ruido la performance empeora mas que para profundidades de árboles menores. Esto se puede ver en la Figura 1 para los gráficos de Churn, donde se puede observar que para profundidades de árboles mayores, el AUC promedio decrece más a medida que aumenta el ruido. Esto es porque el modelo se vuelve mas flexible y se ajusta mas a los datos de entrenamiento que están alterados por el ruido.

Table 2: Min mean AUC a traves de profundidades de arboles maximas en Iranian Churn y metodos de imputacion

maxdepth	IMPUTED	mean_auc	noise
18	No	0.8976336	0.0
30	No	0.8428871	0.2
19	No	0.7971215	0.4
24	No	0.7636765	0.6
14	No	0.7058096	0.8
21	Yes	0.8969152	0.0
20	Yes	0.8429643	0.2
17	Yes	0.7993087	0.4
13	Yes	0.7606050	0.6
12	Yes	0.7049549	0.8

En contraste con la Tabla 1 donde podemos ver que la performance máxima promedio del modelo a medida que crece el ruido decrece en a lo sumo 5 puntos, en la Tabla 2 podemos ver que la performance mínima promedio del modelo a medida que crece el ruido decrece en a lo sumo 20 puntos. Afirmando lo que observabamos en la Figura 1 para los gráficos de Churn, que para profundidades de árboles mayores, el AUC promedio decrece más a medida que aumenta el ruido.

Heart

Para el dataset de *Heart Disease* que cuenta con mayoría de atributos categóricos, se puede observar que la performance del modelo no se demasiado ve afectada por el ruido aplicado a los atributos numéricos. Podemos ver en la Tabla 3 como el máximo AUC promedio a traves de los distintos `max_depths` de arboles se mantiene constante a medida que aumenta el ruido.

Ademas, tanto para cuando se imputan datos faltantes como para cuando no, el `max_depth` que maximiza el AUC promedio se mantiene constante a medida que aumenta el ruido.

Table 3: Max mean AUC a traves de profundidades de arboles maximas en Heart Disease y metodos de imputacion

maxdepth	IMPUTED	mean_auc	noise
4	No	0.8791533	0.0
3	No	0.8805077	0.2
3	No	0.8811366	0.4
4	No	0.8866049	0.6
5	No	0.8847389	0.8
4	Yes	0.8959828	0.0
4	Yes	0.8848785	0.2
5	Yes	0.8909271	0.4
4	Yes	0.8883593	0.6
4	Yes	0.8884267	0.8

Student

Para el dataset de *Student Performance* sucede algo similar a lo que sucede con el dataset de *Iranian Churn*, la performance del modelo disminuye a medida que crece el ruido porcentual aplicado a los atributos numéricos. Podemos ver en la Tabla 4 como el máximo AUC promedio a traves de los distintos `max_depths` de arboles disminuye a medida que aumenta el ruido.

Ademas, tanto para cuando se imputan datos faltantes como para cuando no, el `max_depth` que maximiza el AUC promedio decrece a medida que aumenta el ruido. Aunque inicialmente ya la profundidad máxima que maximiza el AUC promedio es menor que en los otros datasets.

Table 4: Max mean AUC a traves de profundidades de arboles maximas en Student Performance y metodos de imputacion

maxdepth	IMPUTED	mean_auc	noise
4	No	0.7170410	0.0
4	No	0.7030489	0.2
3	No	0.6883808	0.4
3	No	0.6782525	0.6
3	No	0.6763430	0.8
4	Yes	0.7174910	0.0
4	Yes	0.6933784	0.2
4	Yes	0.6788769	0.4
2	Yes	0.6826902	0.6
2	Yes	0.6840113	0.8

Conclusiones

Como se puede observar en los resultados, aplicar ruido a los atributos numéricos afecta la performance de los modelos, pero no de la misma manera para todos los datasets.

Para el dataset de *Heart Disease*, que cuenta iguales partes de atributos numéricos y categóricos, la performance del modelo no se vió afectada.

Para los datasets de *Iranian Churn* y *Student Performance* que cuentan con gran cantidad de atributos numéricos, la performance del modelo disminuyó a medida que crece el ruido porcentual aplicado a los atributos numéricos. Esto se debe a que el modelo se vuelve mas flexible y se ajustó mas a los datos de entrenamiento que están alterados por el ruido.