

exp_4

Tecnología Digital VI

Piñera - Saint-Nom - De Diego

2023-08-22

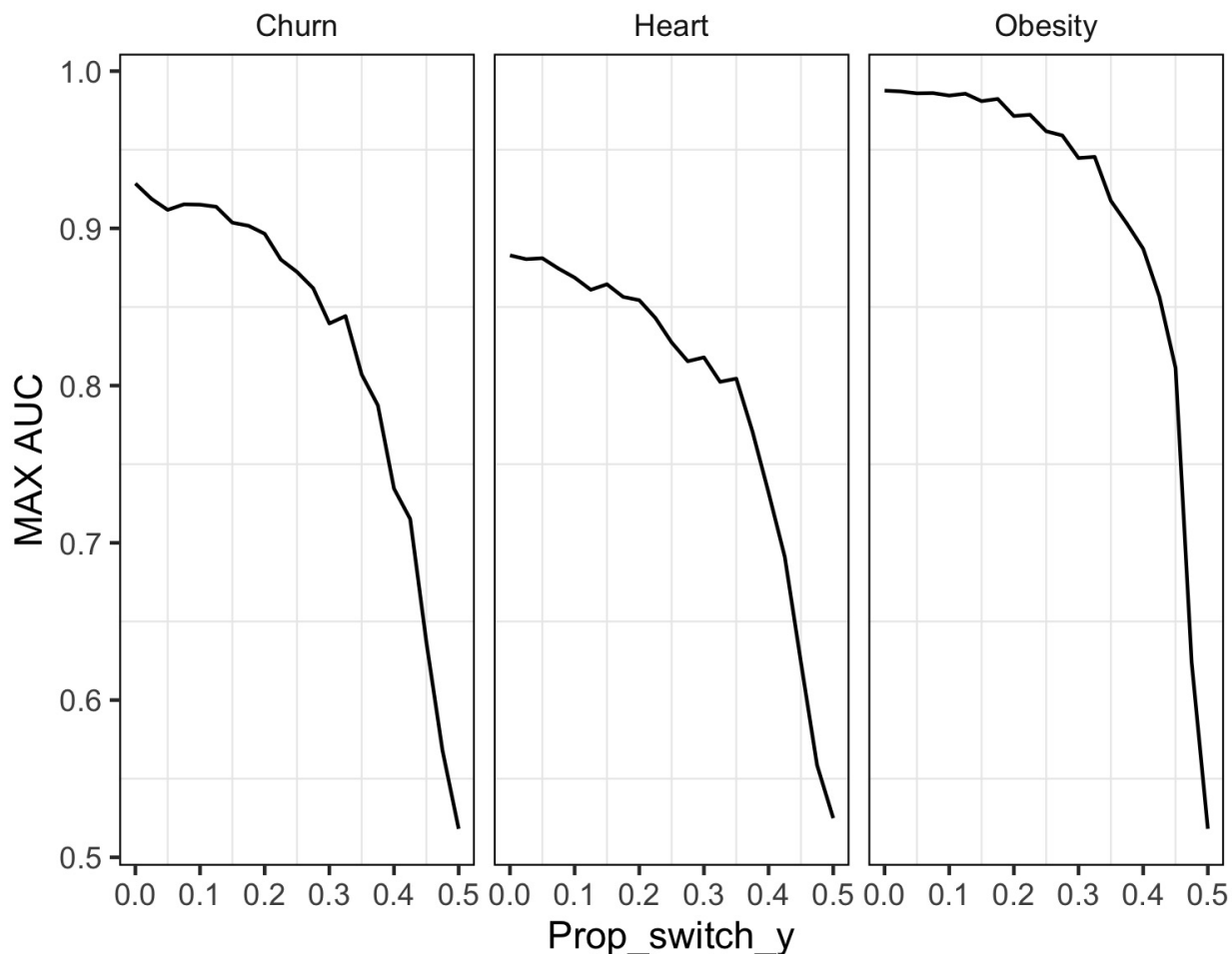
Ejercicio 2

Experimento 4

Introducción: En este experimento, buscamos investigar el impacto del ruido presente en el rendimiento de los árboles de decisión en diferentes conjuntos de datos y niveles de ruido. Este experimento es muy esencial, ya que el ruido es una característica común en muchos escenarios de modelado, lo que nos daría un entendimiento más profundo de cómo se comportan los modelos en situaciones como estas. Nuestro objetivo es obtener una perspectiva más completa sobre cómo el ruido puede afectar la capacidad predictiva de los árboles de decisión en contextos muy comunes en la práctica.

Hipótesis: Nuestra hipótesis plantea que a medida que aumenta el nivel de ruido en las variables objetivo, la precisión del rendimiento de los árboles de decisión disminuirán. Esperamos ver que el aumento del ruido introducirá mayor variabilidad en las predicciones, lo que resultará en una disminución en el valor de las métricas de evaluación del modelo, como el área bajo la curva (AUC).

Gráficos: A continuación se encuentra el gráfico con los máximos AUC variando los porcentajes de ruido en las variables objetivo de 0 a 0.5 con incrementos de 0.025. Estos valores se obtuvieron por la estimación de repetición validada, donde en cada iteración de probabilidad de ruido se corría el experimento. A modo de resumen, se agregó una tabla con los valores máximos de las áreas debajo de la curva para los valores 0, 0.10, 0.20, 0.30, 0.40 y 0.50, para así visualizar mejor el progreso y llegar a una conclusión sobre el experimento.



```
## dataset_name      0.1      0.2      0.3      0.4      0.5
## 1      Churn 0.9150741 0.8965479 0.8395076 0.7344885 0.5181130
## 2      Heart 0.8686744 0.8543093 0.8179391 0.7318702 0.5248849
## 3      Obesity 0.9844370 0.9714339 0.9447119 0.8870385 0.5181231
```

Observaciones: Se evaluó el máximo AUC hasta donde la probabilidad de intercambiar los valores era del 50%, es decir, un 0.5 de ruido en los datasets. En el conjunto de datos Churn, iniciamos con un máximo AUC de 0,93, pero a partir del 20% de ruido, este valor decreció por debajo de 0,9, llegando finalmente a un nivel de desempeño de 0,52. De manera similar, en el conjunto de datos Heart, el máximo AUC inicial de 0,89 se redujo rápidamente hasta alcanzar un valor mínimo de 0,51 con mayores niveles de ruido. Respecto al dataset Obesity, el modelo comenzó con un máximo AUC de 0,99 en condiciones de ausencia de ruido (0%), y a partir del 20% de ruido, disminuyó poco a un valor de 0,97. En todos los casos, tienen probabilidad de ser intercambiados del 50%, el máximo AUC se encuentra en 0,5 aproximadamente.

La performance del dataset Obesity muestra ser más consistente al principio del experimento, a diferencia de los otros dos datasets. Intuitivamente, esto puede darse porque el modelo tiene mayor cantidad de atributos a comparación de los demás, logrando predecir mejor apesar del ruido.

Conclusión: Nuestros resultados confirmaron nuestras expectativas. A medida que incrementamos el porcentaje de ruido, observamos una tendencia clara de disminución en la métrica AUC en los tres datasets.

Esta tendencia se da porque la presencia de ruido indica una mayor proporción de valores en la variable objetivo que se intercambian entre 0 y 1.

Como se puede observar, cuando tenemos un 50% de ruido en los datos, es probable que el modelo no pueda aprender ninguna pauta significativa y esté prediciendo al azar. Esto resulta en una métrica AUC de 0.5, ya que el modelo no puede diferenciar entre las clases debido a la presencia grande de ruido en los datos.

Estos hallazgos resaltan la influencia negativa del ruido en las variables objetivo sobre la capacidad predictiva de los modelos de árboles de decisión, validando nuestra hipótesis inicial.