Experimento 4

Tomas Glauberman, Natalia Massri y Juan Silvestri

Experimento 4

Se realizó un experimento en el que se analiza el impacto que tiene el ruido en la variable a predecir en 3 datasets distintos con un modelo de Árbol de decisión. Para ello, se entrenaron múltiples modelos para cada uno de los datasets, lo que permite estimar y evaluar la performance de ellos.

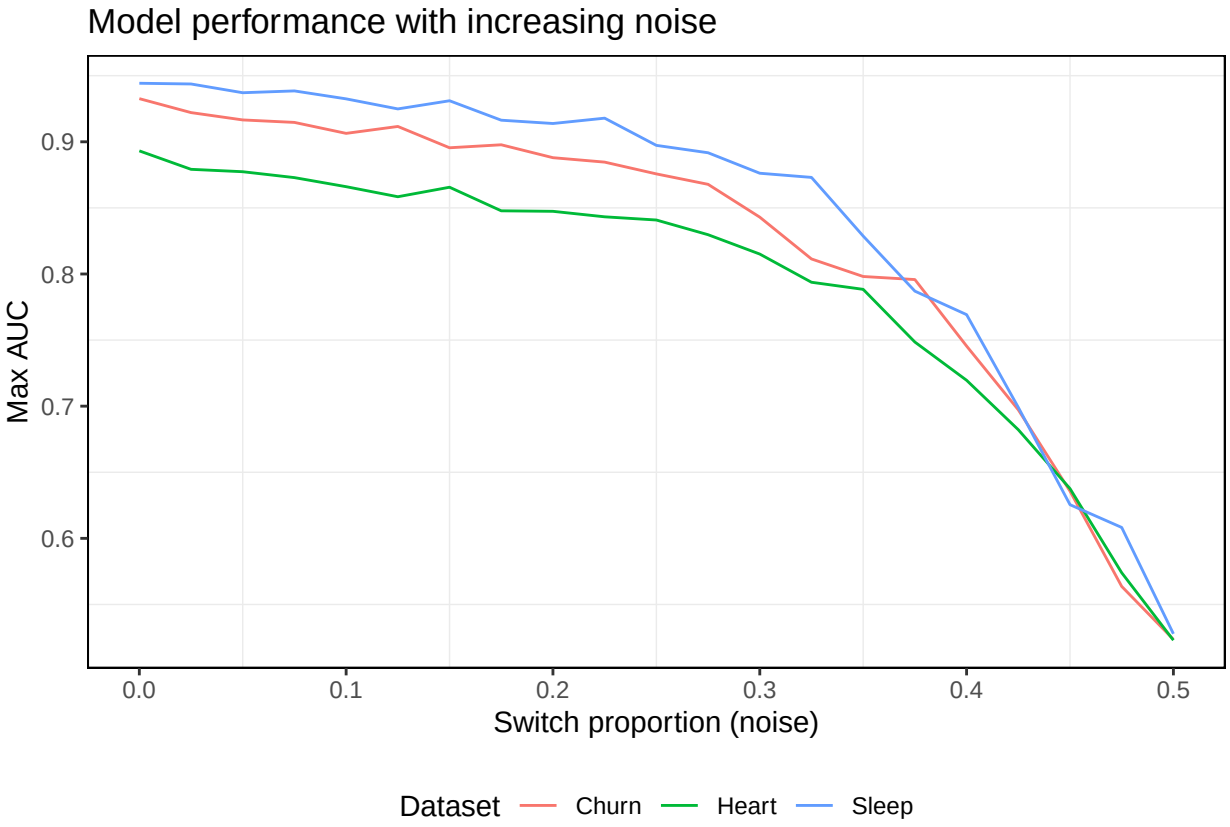
Las variables independientes del experimento son:

- Profundidad del árbol (1-30) - `maxdepth`
- Proporción de switch en la variable a predecir (0-1) - `PROP_SWITCH_Y`

Así, se entrena un modelo para cada profundidad y con diferentes valores para la proporción de switch. Para un rango de valores de `PROP_SWITCH_Y` y cada dataset, se obtiene la estimación de AUC del modelo, y se guarda sólo el modelo con mejor performance de todas las profundidades.

En los 3 datasets se obtiene resultados similares que indican que la performance de los modelos seleccionados disminuye a medida que se agrega ruido a la variable a predecir. La peor performance se alcanza con ruido máximo en `PROP_SWITCH_Y=0.5` (ya que al aumentar el`PROP_SWITCH_Y` mejora, pues para AUC menores a 0.5 se invierten los labels pues el modelo está prediciendo al reves).

```
source("exp_4.R")
```



Los resultados obtenidos son los esperados. Si durante el entrenamiento el 50% de los datos están invertidos, es natural pensar que la performance luego sea cercana al 0,5. El modelo aprende bien de los datos pero éstos son incorrectos. Sin embargo, algo interesante para notar es que ésta no es una relación lineal. Con un 30% de los labels invertidos la performance todavía es relativamente buena, cercana a un AUC=0,9. Con proporciones mayores de ruido la performance disminuye drásticamente.