Caso Contactabilidad – Examen Final AML

Se comparte la siguiente información:

1. data_selec_entre: data de entrenamiento
2. data_selec_test: data de validación
3. Diccionario de Variables (40 variables)
4. Ppt Modelo Contactabilidad: hay más variables en la ppt, en la data se disminuyeron las variables y observaciones

Los valores -999 en la data son valores nulos

Se pide realizar:

1. Análisis descriptivo
2. Ingeniería de variables (Transformaciones, imputación de nulos, etc.)
3. Modelos ensamblados (Boosting, Stacking, etc.) y Modelos Estadísticos (Regresión lineal, GAM, GAMLSS, etc.)
4. Optimización
5. Resultados

El día del examen final se presenta el caso en un script de Python con las iniciales de su nombre, seguida por un guion, su sección y las letras CD.

Por ejemplo, para Manuel Valdivia Carbajal de la primera sección sería: MVC_1CD

Recomendación: comentar el código.

Saludos

Manuel – Profesor de la Maestría en Ciencia de Datos

Modelos Vía Ensamble

XGBoost o LightGBM para especificar los parámetros para el modelo. Algunos de los parámetros importantes son:

- **objetivo:** La función de pérdida que se va a minimizar. Para la clasificación binaria, podemos usar binary:logistic para XGBoost o binary para LightGBM.

- **eval_metric:** La métrica que se utilizará para la evaluación. Para la clasificación binaria, podemos usar "auc" para el área bajo la curva ROC o "logloss" para la pérdida logarítmica.

- **n_estimators:** El número de árboles a construir. Podemos usar un número grande y confiar en detenernos temprano para encontrar el número óptimo.

- **max_ Depth:** La profundidad máxima de cada árbol. Un valor mayor puede capturar patrones más complejos, pero también aumentar el riesgo de sobreajuste.

- **learning_rate:** El factor de contracción para cada árbol. Un valor menor puede reducir la varianza, pero también aumentar el sesgo y el tiempo de entrenamiento.

- **submuestra:** La fracción de muestras que se utilizarán para cada árbol. Un valor menor puede reducir la correlación entre los árboles, pero también aumentar la varianza.

- **colsample_bytree:** La fracción de características que se utilizarán para cada árbol. Un valor menor puede reducir la correlación entre las características, pero también aumentar el sesgo.

- **reg_alpha:** El término de regularización L1 para los pesos. Un valor mayor puede reducir la complejidad y el sobreajuste.

- **reg_lambda:** El término de regularización L2 para los pesos. Un valor mayor puede reducir la complejidad y el sobreajuste.