COMPARACIÓN DE MODELOS GAM CON LIGHT GBM EN EL PRONÓSTICO DE FUGA EN CLIENTES DE UNA INSTITUCIÓN BANCARIA

Integrantes: Salcedo Guerra Nataly, Torres Achata Angie Fiorella

Docente: Manuel Valdivia Carbajal

Asignatura: Aprendizaje Estadístico

Escuela Profesional de Ingeniería Estadística

Facultad de Ingeniería Económica, Estadística y CCSS

Universidad Nacional de Ingeniería

RESUMEN

La fuga de clientes es un fenómeno que atañe a la gran mayoría de las instituciones financieras, siendo también un tema de intensivo estudio en contexto de pandemia. Este Proyecto presenta la comparación del modelo GAM bajo un paradigma estadístico con el modelo Light GBM de enfoque Machine Learning desarrollado para identificar los clientes con tendencias a la fuga en una institución financiera, primero realizamos un breve resumen de los Modelos. A continuación, introducimos el Modelo GAM y su estructura, caracterizada por el uso de funciones suaves sobre las variables explicativas que permite relaciones no lineales entre estas y la variable objetivo. Con el fin de obtener dichas funciones suaves, recopilamos diferentes técnicas. Seguidamente, introducimos el Modelo LightGBM con sus métricas de capacitacion y validacíon, parámetros, entrenamiento y optimización. Por último, haciendo uso del software Python, comprobamos si la implementación de las técnicas descritas por ambos modelos no lineales pueden ser beneficiosas a la hora de tomar decisiones relacionadas con el pronóstico de fuga de clientes en instituciones financieras.

SUMMARY

This Project focuses on the study and application of the comparison of the GAM model under a statistical paradigm with the Light GBM model under a Machine Learning approach for a financial institution, first we make a brief summary of the Models. Next, we introduce the GAM Model and its structure, characterized by the use of soft functions on the explanatory variables that allow non-linear relationships between them and the target variable. In order to obtain such smooth functions, we collected different techniques. Next, we introduce the LightGBM Model with its training and validation metrics, parameters, training and optimization. Finally, using Python software, we check whether the implementation of the techniques described by both non-linear models can be beneficial when making decisions related to the forecast of customer churn in banking institutions.

INTRODUCCIÓN

La cartera de clientes es uno de los activos más importantes para una institución financiera, ya que está estrechamente relacionada con las utilidades del negocio. el costo de adquirir un cliente nuevo es mucho mayor al retener, y la fuga significa una pérdida grande a la institución. Sin embargo, el problema de fuga de clientes en instituciones financieras ha sido relevante en contexto de pandemia por diversas causas, una de ellas es que el cliente ha sido afectado economicamente lo cual conlleva que busque menores tasas que ofrecen otras intituciones financieras, otra es el contexto de digitalizacion donde el cliente busca un banco con la mejor plataforma para hacer todo tipo de gestiones sin necesidad de acudir a una sucursal. La retención de clientes consiste en la identificación de los clientes con mayores tendencias a la fuga y en la determinación de las estrategias o procedimientos que aumenten el grado de fidelización y bajen los índices de fuga en la cartera. En el presente trabajo se presenta una comparación de modelos predictivos con enfoque estadístico y machine learning desarrollado para identificar los clientes con tendencias a la fuga en un banco. De esta forma es posible hacer más efectivas las políticas comerciales de retención

PRESENTACIÓN DEL PROBLEMA

Debido al contexto de pandemia, ha ido aumentando la fuga de clientes . Poniendo en una situación difícil a gran cantidad de intituciones financieras.

Se hará un analisis de la comparacion de dos modelos no lineales, modelo GAM bajo un paradigma estadístico con el modelo Light GBM de enfoque Machine Learning para identificar los clientes con tendencias a la fuga en un banco y así hacer más efectivas las políticas comerciales de retención de clientes**.**

OBJETIVOS

- Proporcionar una regla de clasificación de propensión de fuga de clientes para realizar acciones de retención de clientes.
- Plantear estrategias que aumenten el grado de fidelización y bajen los índices de fuga en la cartera.

MARCO TEORICO

MODELO GAM:

Los modelos aditivos generalizados (GAM) que se utilizan básicamente para la mejora del modelo lineal generalizado permitiendo formas funcionales no-paramétricas para cada una de las covariables mientras se mantiene la propiedad de aditividad..La estructura del modelo puede representarse de la siguiente forma:

$$g\big(E(Y)\big) = \beta + f_1(x_1) + f_2(x_2) + f_3(x_3) + \cdots + f_p(x_p)$$

.

- $g\big(E(Y)\big)$ es la función de enlace que vincula el valor esperado a las variables predictoras $x_1, x_2, x_3, \ldots., x_p$. Indica cómo el valor esperado de la respuesta se relaciona con las variables predictoras. GAM admite múltiples funciones de enlace.
- $f_1(x_1) + f_2(x_2) + f_3(x_3) + \cdots + f_p(x_p)$ es la forma funcional con una estructura aditiva que consiste en un número de términos $f_1(x_1), f_2(x_2), f_3(x_3), \ldots, f_p(x_p)$. Los términos denotan funciones suaves y no paramétricas.
- $Y \sim$ Distribución de la Familia Exponencial

El problema planteado se enfrentó con un enfoque de clasificación binaria. Este tipo de procedimiento se basa en la determinación de una función clasificadora que permite asignar a cada objeto a una de las dos clases definidas a priori. En nuestro caso, cada cliente será asignado a una de las clases “fuga” o “no fuga”.Podemos extender el usual modelo logítico.

$$Y \sim \text{Bernoulli}(\pi)$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta + f_1(x_1) + f_2(x_2) + f_3(x_3) + \cdots + f_p(x_p)$$

SUPUESTOS:

- La relación entre las características dependientes y las características independientes no es lineal.
- El modelo resulta significativo para un p-valor<0.05

VENTAJAS Y DESVENTAJAS:

- Los ajustes no lineales pueden pontencialmente llevarnos a mejores predicciones de la respuesta Y.
- Los modelos GAM permiten ajustar una función no lineal para cada covariable Xj de manera que podemos modelar automáticamente las relaciones no lineales que los modelos de regresión usuales no tomarían en cuenta.
- La principal limitación es la restricción que el modelo es aditivo. Con muchas variables, algunas interacciones importantes pueden ser no consideradas. Sin embargo, como en los modelos de regresión lineales, se pueden adicionar manualmente términos de interacción incluyendo predictores adicionales de la forma Xj × Xk en el modelo.

MODELO Light GBM:

LightGBM es un modelo no lineal, tiene un algoritmo de refuerzo (o también de potenciación) de gradientes (gradient boosting) basado en modelos de árboles de decisión. Puede ser utilizado para la categorización, clasificación y muchas otras tareas de aprendizaje automático, en las que es necesario maximizar o minimizar una función objetivo mediante la técnica de gradient boosting, que consiste en combinar clasificadores sencillos, como por ejemplo árboles de decisión de profundidad limitada.

PARAMETROS:

- **boosting_type:** 'gbdt', árbol de decisión de aumento de gradiente tradicional.
- **num_leaves:** Hojas de árbol máximas para los alumnos base**.**
- **max_depth:** Profundidad máxima del árbol para los alumnos base, <=0 significa que no hay límite.
- **learning_rate:** reducir / adaptar la tasa de aprendizaje en el entrenamiento utilizando la devolución de llamada.
- **n_estimators:** Número de árboles potenciados para encajar.
- **reg_alpha:** Término de regularización L1 sobre ponderaciones.
- **reg_lambda:** Término de regularización L2 sobre ponderaciones.
- **n_jobs:** Número de subprocesos paralelos a utilizar para el entrenamiento (se puede cambiar en el momento de la predicción).
- **importance_type:** El tipo de importancia de la función que se debe rellenar en . Si 'dividido', el resultado contiene números de veces que se utiliza la función en un modelo. Si 'ganancia', el resultado contiene ganancias totales de divisiones que utilizan la función.

VENTAJAS:

- Mayor velocidad de entrenamiento y mayor eficiencia.
- Menor uso de memoria.
- Mayor precision.
- Capacidad para manejar datos a gran escala.

OPTIMIZACIÓN:

El rendimiento de este modelo depende de los hiperparámetros. Un conjunto óptimo de parámetros puede ayudar a lograr una mayor precisión. Encontrar hiperparámetros manualmente es tedioso y computacionalmente costoso. Por lo tanto, la automatización de la sintonización de hiperparámetros es importante. GridSearchCV se utiliza generalmente para optimizar hiperparámetros.

DESCRIPCION DE LA SOLUCION

La data utilizada corresponde a la cartera completa de clientes que poseía una institución bancaria. Esta data incluye todos los clientes vigentes durante el periodo y todos los clientes que se fugaron durante el mismo periodo, Se trabajo con 20 variables, las cuales son:

1. Edad (EDAD)
2. Sexo (SEXO)
3. Departamento (DEPARTAMENTO)
4. Ingreso bruto del mes anterior (INGRESO_BRUTO_MI)
5. Frecuencia de transacciones en agente (FREC_AGENTE)
6. Segmento del banco (SEGMENTO)
7. Adelanto sueldo, 1 mes antes de la campaña (FLG_ADEL_SUELDO_MI)
8. Convenios,1 mes antes de la campaña, en el sistema financiero (FLG_CONV_SF).
9. Cantidad promedio de transacciones con TC (PROM_CTD_TRX_6M)
10. Antiguedad del cliente (ANT_CLIENTE)
11. Si reclamaron elm es anterior (CTD_RECLAMOS_MI)
12. Frecuencia del uso de la Banca por internet (FREC_BPI_TD)
13. Recencia del uso del agente(REC_AGENTE_TD)
14. Si el cliente fuga (TARGET_MODEL2)

ANÁLISIS DESCRIPTIVO:

GRAFICO1

Representación del porcentaje de clientes de fuga y no fuga

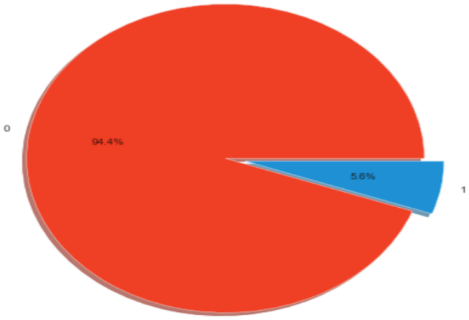
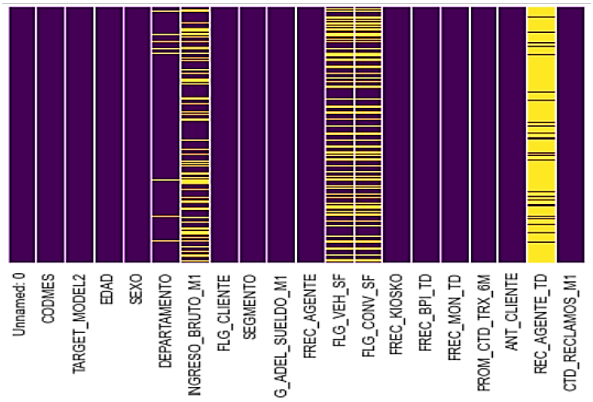


GRAFICO2

Representación los datos faltantes en cada variable.



En el **GRAFICO1** se aprecia que el porcentaje de los clientes de fuga de la institucion bancaria es el 5.5% y de no fuga es 94.5% de un total de 787494 clientes. En el **GRAFICO2** se aprecia hay variables con algunos datos faltantes, pero Tambien hay variables que la mayoria de sus datos son faltantes, es por eso que se realizará un tratamiento a las variables para tener un mejor análisis.

ETAPA DE TRATAMIENTO:

El objetivo de esta etapa es resolver el problema relacionado a los valores faltantes dentro de la base de datos, utilizar transformaciones de las variables originales para enriquecer la información contenida y elegir las variables que si aportan al modelo.

- IMPUTACIONES: Llenado de datos faltantes con promedios en variables numericas y modas en variables categóricas.
- DUMMIES: La transformación de las variables texto (categorías) a números.
- Para reducción de la dimension de las variables numericas se realizá ACP

ANÁLISIS DE COMPONENTES PRINCIPALES:

El análisis de componentes principales (ACP) es una técnica de análisis multivariante de reducción de datos. El método de componentes principales tiene como objetivo transformar un conjunto de variables originales, en un nuevo conjunto de variables (sin perder información), combinación lineal de las originales, denominadas componentes principales, componentes o factores. El ACP trata de hallar estos componentes o factores, los cuales se caracterizan por estar relacionadas para explicar el mayor porcentaje de la varianza total. Se tuliza para reducir la dimensión lineal utilizando Valor singular de descomposición (SVD) para proyectar un espacio de menor dimensión.

CONSTRUCCIÓN DEL MODELO:

Se realizará la construcción de los dos modelos no lineales GAM de paradigma estadístico y Light GBM de enfoque Machine Learning, el cual cada modelo consta con dos etapas: entrenamiento y test. Para cada una de las etapas se considera un subconjunto del total de los objetos (clientes) a clasificar. Estos subconjuntos de objetos forman una partición del conjunto total de objetos y son llamados, conjunto de entrenamiento y conjunto de test, respectivamente.

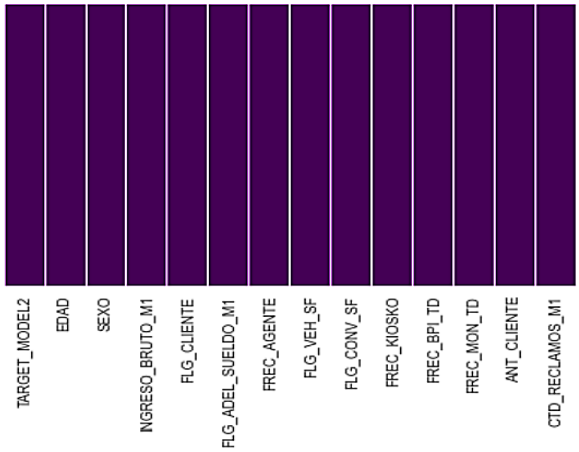
RESULTADOS

IMPUTACIONES:

GRAFICO3

Representación los datos faltantes en cada variable tratada.

.



Con nuestras variables ya tratadas podemos observar en el **GRAFICO3** que ya no hay datos faltantes y tambien que elegimos 14 variables para trabajar en nuestros modelos.

GENERAR DUMMIES:

Se genero dummies a las variables de texto en este caso las vatiables “SEXO” y “FLG_CLIENTE”.

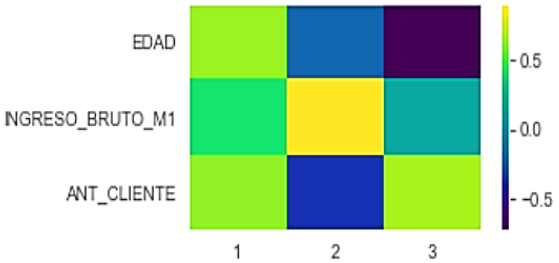
VARIABLES	DESCRIPCIÓN	CODIGO
SEXO_F	Sexo femenino	F(1),M(0)
SEXO_M	Sexo Masculino	F(0),M(1)
FLG_CLIENTE_CLIENTE	Es cliente	CLIENTE(1),NO CLIENTE(0)
FLG_CLIENTE_NO CLIENTE	No es cliente	CLIENTE(0),NO CLIENTE(1)

ACP:

Se trabajará con todas las variables numericas, en este caso con 'EDAD','INGRESO_BRUTO_M1' y 'ANT_CLIENTE'.

GRAFICO4

Matriz de correlaciones de las variables en cada componente.



En el **GRAFICO4** se aprecia que en la componente 1, la variable que más se relaciona es “EDAD”, en la componente 2 la varible que más se relaciona es “INGRESO_BRUTO_M1” y en la componente 3 la variable que más se relaciona es “EDAD”.

GRAFICO5

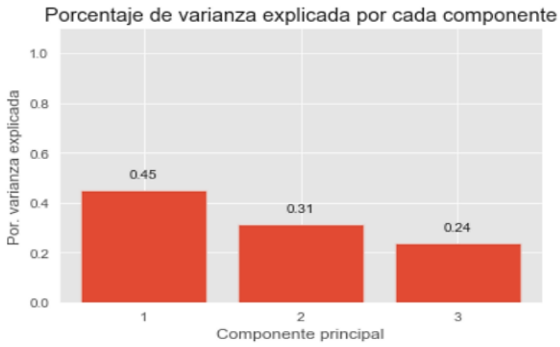
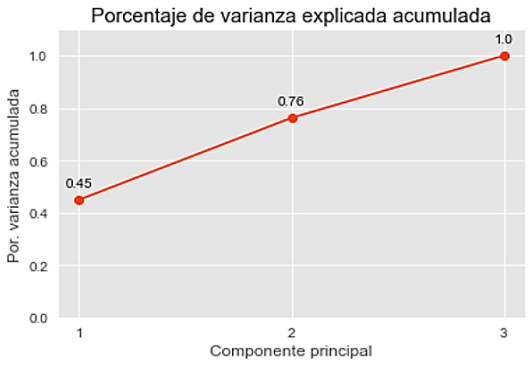


GRAFICO6



Podemos observar en el **GRAFICO1** y **GRAFICO2** el porcentaje de varianza explicada en la componente 1 es 46%, de la componente 2 es 31% y el componente 3 es 24%. La varianza explicada acumulada para las dos primeras componente es 76%, lo cual nos explica casi toda la varianza explicada, es por eso, que nos quedaremos con las 2 primeras componente.

ENTRENAMIENTO Y PRUEBA:

Se divide el conjunto de datos en dos conjuntos: un conjunto de entrenamiento y un conjunto de pruebas. 70% para entrenamiento y 30% para pruebas.

- El modelo se entrena utilizando el conjunto de entrenamiento.
- El modelo se prueba mediante el conjunto de pruebas.

BALANCEO DE DATOS:

Se realiza el balanceo de datos para el Target entrenado.

Sin balanceo:

```
Counter({0: 520549, 1: 30697})
```

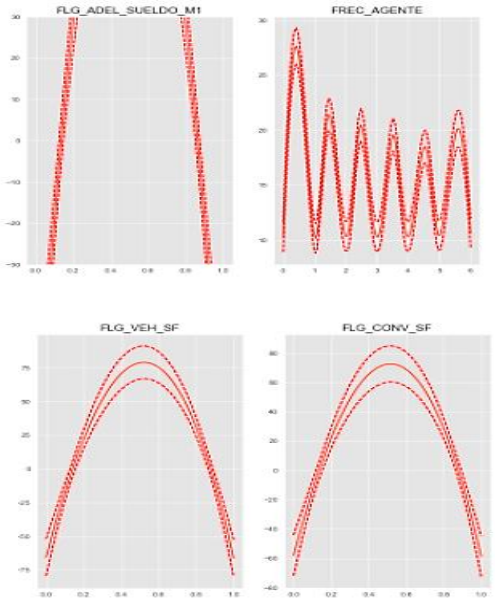
Con balanceo:

```
Counter({0: 520549, 1: 520549})
```

MODELO GAM:

GRAFICO7

Gráficos de dependencia parcial con intervalos de confianza



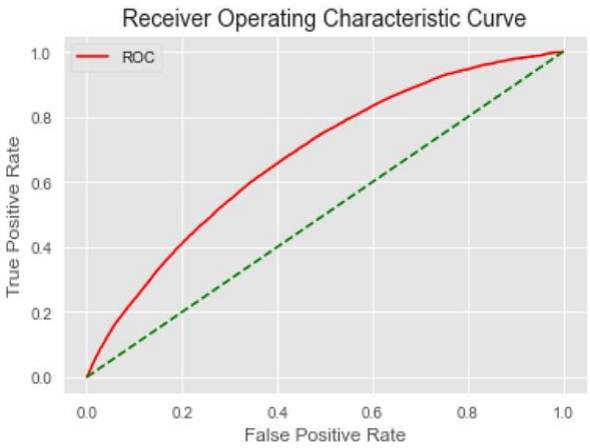
AUC:

```
auc o Roc on training in GAM data : 0.677
auc o Roc on testing in GAM data : 0.679
```

Significa que hay 67% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.

CURVA ROC:

GRAFICO8



MODELO LIGHT GBM:

IMPORTANCIA DE VARIABLES:

TABLA1

'FREC_AGENTE',
'FREC_KIOSKO',
'FREC_BPI_TD',
'FREC_MON_TD',
'SEXO_F',

En la TABLA1 se tiene que la primera variable más importante es “FREC_AGENTE” y la segunda variable más importante es “FREC_KIOSKO”. Esta table nos indica las 5 primeras variables más importantes del modelo LightGBM.

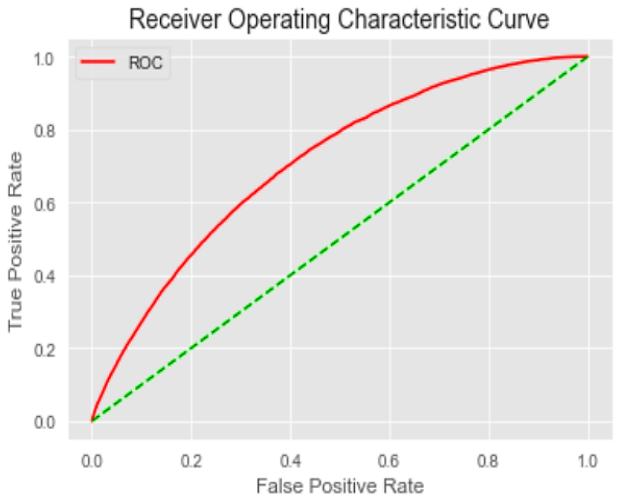
AUC:

```
auc o Roc on training in LGBMClassifier data : 0.731
auc o Roc on testing in LGBMClassifier data : 0.710
```

Significa que hay 71% de probabilidad de que el modelo pueda distinguir entre clase positiva y clase negativa.

CURVA ROC:

GRAFICO9



CONCLUSIONES

Los modelos de predicciones de fugas son una herramienta importante de apoyo a la hora de decidir cuáles de los clientes de la cartera poseen una mayor tendencia a la fuga. Al identificar de mejor forma a estos clientes es posible:

- Aumentar las utilidades y la rentabilidad del negocio, ya que da la posibilidad de retener clientes fugitivos y mantenerlos como clientes captando sus flujos futuros.
- Focalizar los recursos productivos sobre los segmentos que realmente necesitan de acciones de retenci´on fuertes aumentando la eficiencia y efectividad de las políticas de retención.
- Generar un mejoramiento y fortalecimiento de la relación entre el cliente y la institución, al anticiparse a sus necesidades, así como hacer menos sensible al cliente frente a las campañas de marketing de la competencia.
- Disminuir el riesgo total de la cartera, al reducir la necesidad de atraer nuevos clientes potencialmente más riesgosos.

Respecto a la técnica utilizada para la modelación, se reafirma su fortaleza en términos de predicción al modelo Light GBM alcanzando una mayor probabilidad de distingir clases positivas y negativas. Este modelo se podrá utilizar solo para la Cartera de clientes con tendencia a fuga, cada cierto tiempo se tiene que ir revisando ya que los clientes van cambiando.
Se recomienda elaborar un modelo para cada cartera o portafolio ya que los clientes tienen diferente comportamiento u otras variables adicionales para analizar.

BIBLIOGRAFIA

[1] C. Cortés and V. Vapnik. Support vector networks. *Journal of Machine Learning, 20: 273–297, 1995.*

[2] E. Ramusson. Complaints can build relationships. *Sales and Marketing Management, 151(9):89–90, 1999*.

[3] F. Reichheld and E. Sasser. Zero defections: Quality comes to services. *Harvard Business Review, 1990:105–111, September–October, 2000.*

[4] V. Vapnik. *The nature of statistical learning theory*. Springer–Verlag, New York, 1995.

ANEXOS

Se adjunta el código del software utilizado para dejar constancia de los resultados obtenidos. En este Proyecto que se trabajó con PYTHON.

```
pip install imblearn


# In[45]:


from imblearn.over_sampling import SMOTE


# In[46]:


y=fuga['TARGET_MODEL2']
X=fuga.drop(['TARGET_MODEL2'],axis=1)


# In[47]:


from sklearn.model_selection import train_test_split


# In[48]:


X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.30,random_state=101,stratify=y)


# - Contamos:

# In[49]:


from collections import Counter
counter = Counter(y_train)
print(counter)


# - Balanceo:

# In[50]:


os = SMOTE(random_state = 0)

# In[51]:


os_data_X, os_data_y = os.fit_resample(X_train.values, y_train.values)


# In[52]:


counter = Counter(os_data_y)
print(counter)


# - Modelos de Machine Learning

# In[53]:


from sklearn import preprocessing
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import train_test_split
```