Data Lending club

Lending Club es una de las empresas P2P más grandes, publica regularmente conjuntos de datos anonimizados de sus clientes que contienen características de los préstamos y los prestatarios. Esta información se puede utilizar para clasificar si un prestatario incumplirá o no con su préstamo.

Por ejemplo, los bancos suelen tener datos informativos para crear modelos que ayuden a decidir a quién conceder o denegar un préstamo. Es un problema de clasificación supervisado.

Definición del problema:

Se desea predecir, antes de conceder un préstamo a una persona, la probabilidad de que no se devuelva completamente. Por tanto, todas las variables del dataset que se empleen para el modelo, deben poderse utilizar en el momento de su llamada. Para ello, se empleará un algoritmo de clasificación supervisado.

El fichero *"Data_Dictionary.xls"* contiene toda las variables y sus descripción a utilizar en el modelo de clasificación.

El conjunto de datos se encuentra en el fichero *"pd_data_initial_preprocessing.csv"*

La variable objetivo a predecir, que define si un cliente pagará una hipoteca o no es ***loan_status***.

Puntos a resolver en la práctica:

Se pide realizar los siguientes puntos para resolver el problema de clasificación.

1. Hacer un pequeño análisis descriptivo de los datos. Un análisis sencillo que incluya simplemente: (0,5 pts)

1. Cantidad de valores nulos.
2. tipo de variables (cat, float, etc.)
3. En la variable objetivo, distribución de sus valores.
4. Matriz de correlación en variables continuas.

2. Tratar los valores missing tanto en las variables continuas como discretas de forma correcta. Recordad que si existen muchos valores missing no es adecuado eliminar dichas filas. Como guía, recordad que los valores missing: (0,5 pts)

1. En variables continuas se puede sustituir por la media o mediana, o por un valor muy distinto al resto de cantidades, que refleje que de algún modo que es un valor missing.
2. En variables discretas se pueden sustituir por la moda o una clase nueva que indique que no tienen valor, p.e. “SIN VALOR”.
3. Puede ser interesante, estudiar y evaluar algunos de los métodos de la librería sklearn.impute, como por ejemplo sklearn.impute.KNNimpute, que realiza una imputación mediante una regresión con KNN.

3. Transformar las variables continuas, si fuera necesario, y las categóricas correctamente. (0,5 pts)
4. Partir el conjunto de datos en un subconjunto de entrenamiento y otro de test. Recordad, que si la variable objetivo está desbalanceada, es fundamental, realizar una partición que conserve las proporciones originales en cada subconjunto. (0,5pt)
5. Vamos ha evaluar diferentes algoritmos de clasificación, para los cuales necesitaremos hacer los siguientes pasos:

1. Realizar algún tipo de transformación como escalado o normalización, si el método lo necesita.
2. Entrenar el modelo con el conjunto de datos de entrenamiento.
3. Mostrar cómo de bueno ha sido el entrenamiento.
4. Evaluar el modelo con el conjunto de datos de test. Para ellos se

pedirán las siguiente métricas:

- Accuracy
- Precision
- Recall
- Confusion Matrix
- F-SCORE
- Curva ROC
- Area bajo la curva.

Todos estos pasos serán necesarios evaluarlos para cada uno de los siguientes algoritmos:

A. Regresión logística (1 pt)

B. K-NN (1pt)

C. Arbol de decisión simple (1pt)

D. Bagging Classifier (1pt)

E. Random Forest (1pt)

F. GradientBoost (1pt)

G. XGBoost. (1pt)

6.- Cual de todas las métricas crees que es la más conveniente?. Ordena todos los métodos de mejor a peor según dichas métricas. (0,5pt)

7.- En esta práctica no hemos abordado todavía la importancia de hacer una correcta preselección y transformación de variables antes de entrenar un modelo. Sin embargo , los ensambladores nos pueden dar una buena pista, de que variables elegir a la hora de entrenar un modelo de clasificación o regresión. ¿Cómo medirías que importancia tiene cada variable a la hora de clasificar correctamente la variable

objetivo? Haciendo uso de alguno de los métodos propuestos, Muestra una gráfica que presente de forma ordenada la importancia que tiene cada variable en la clasificación (0,5pt)