Predicción de aprobación de prestamos

Lucas Fortún Iñurrieta

Introducción y motivación

En el mundo financiero, la evaluación precisa de la elegibilidad para préstamos es muy importante tanto para los bancos como para los solicitantes. Esta evaluación implica un análisis de diversos factores financieros y personales para determinar si un individuo es apto para recibir un préstamo y sea capaz de reembolsarlo.

El objetivo principal es desarrollar modelos predictivos mediante la utilización de técnicas de aprendizaje automático, que puedan analizar características financieras y personales de los solicitantes para estimar con precisión la probabilidad de que un préstamo sea aprobado o rechazado.

Conjunto de datos

Particionamiento:

El conjunto de datos se ha dividido en un 80% para entrenamiento y validación, reservando el 20% restante para pruebas. Para encontrar los mejores hiperparámetros, he utilizado el método GridSearchCV, el cual incorpora la validación cruzada. Esto elimina la necesidad de separar manualmente el conjunto de entrenamiento del de validación, ya que GridSearchCV realiza esta división automáticamente. Específicamente, he configurado el parámetro cv = 5, lo que implica una validación cruzada de 5 pliegues. Esto divide el conjunto de datos en 5 partes iguales, entrenando y evaluando el modelo 5 veces. En cada iteración, se utiliza una porción diferente como conjunto de validación, mientras que el resto se emplea como conjunto de entrenamiento.

Análisis atributos:

Estadísticas Descriptivas:

	Media	Mediana	Desviación Estándar	Máximo	Mínimo
no_of_dependents	2.498712e+00	3.0	1.695910e+00	5	0
education	5.022253e-01	1.0	5.000536e-01	1	0
self_employed	5.036308e-01	1.0	5.000454e-01	1	0
income_annum	5.059124e+06	5100000.0	2.806840e+06	9900000	200000
loan_amount	1.513345e+07	14500000.0	9.043363e+06	39500000	300000
loan_term	1.090045e+01	10.0	5.709187e+00	20	2
cibil_score	5.999361e+02	600.0	1.724304e+02	900	300
residential_assets_value	7.472617e+06	5600000.0	6.503637e+06	29100000	-100000
commercial_assets_value	4.973155e+06	3700000.0	4.388966e+06	19400000	0
luxury_assets_value	1.512631e+07	14600000.0	9.103754e+06	39200000	300000
bank_asset_value	4.976692e+06	4600000.0	3.250185e+06	14700000	0

Balance de Clases:

	Cantidad
1	2656
0	1613

Marco experimental

Modelos descartados:

- **Regresión lineal:** La regresión lineal se utiliza para predecir valores continuos en función de variables independientes. Por ello no es apto para problemas de clasificación.
- **Clustering:** El clustering se utiliza para agrupar datos en diferentes grupos basados en la similitud de características, pero no realiza predicciones sobre etiquetas específicas o clases ya que es un algoritmo de aprendizaje no supervisado.

Modelos utilizados:

1. Regresión Logística: Calcula la probabilidad de que una instancia pertenezca a una clase particular y toma decisiones basadas en umbrales, empleando una función logística para estimar probabilidades.

2. Regresión Logística Polinómica: Extensión de la regresión logística que usa características polinómicas para capturar relaciones no lineales entre variables.

3. Naive Bayes: Un modelo que se basa en el teorema de Bayes y supone independencia entre características para la clasificación.

4. Redes Neuronales: Modelos basados en la estructura del cerebro humano que consisten en capas de nodos interconectados, útiles para capturar patrones complejos en datos.

Ensembles basados en variación de datos

Estos métodos se centran en introducir variaciones en los conjuntos de datos de entrenamiento para construir múltiples modelos base, donde cada modelo se entrena con un subconjunto diferente de los datos originales.

5. Árboles de Decisión: Utilizan un árbol invertido para tomar decisiones jerárquicas basadas en reglas de características, dividiendo los datos en ramas.

6. Bagging: Técnica que combina múltiples modelos para mejorar la precisión general, entrenando en diferentes conjuntos de datos y promediando sus predicciones.

7. Boosting: Similar al bagging, pero enfocado en mejorar iterativamente el rendimiento dando más peso a las predicciones incorrectas.

8. Random Forest: Extensión de árboles de decisión que construye múltiples árboles y combina sus predicciones para obtener mayor precisión y reducir el sobreajuste.

Ensembles basados en descomposición

Estos métodos se enfocan en descomponer el problema de aprendizaje en subproblemas más pequeños, cada uno resuelto por un modelo base.

9. OneVsAll (OVA): Estrategia para problemas de clasificación multiclase donde se entrena un clasificador por clase, tratando cada clase frente a las demás.

10. OneVsOne (OVO): Estrategia para problemas de clasificación multiclase donde se entrena un clasificador por cada par de clases, comparando cada par de clases entre sí.

Hiperparámetros:

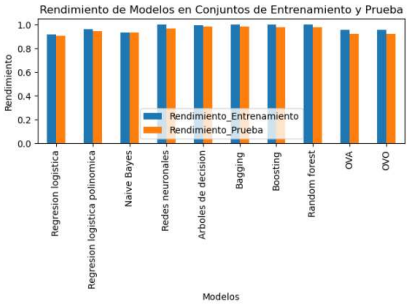
Para la implementación de estos modelos se han probado diversos hiperparámetros, de los cuales se han escogido aquellos que mejor resultado dan.

Hiperparámetros obtenidos:

Modelo	Parametro regularizacion	Grado polinomio	Numero capas ocultas	Numero de neuronas	Profundidad	Numero de estimadores
Regresión logística	C = 0.097	-	-	-	-	-
Regresión logística con características polinómicas	C = 100	2	-	-	-	-
Naive bayes	-	-	-	-	-	
Redes neuronales	Alfa = 1/1000	-	2	50	-	-
Árboles de decisión	-	-	-	-	15	-
Bagging	-	-	-	-	-	50
Boosting (Adaboost)	-	-	-	-	-	10
Random forest	-	-	-	-	-	250
OVA	-	-	-	-	-	-
OVO	-	-	-	-	-	-

Estudio experimental

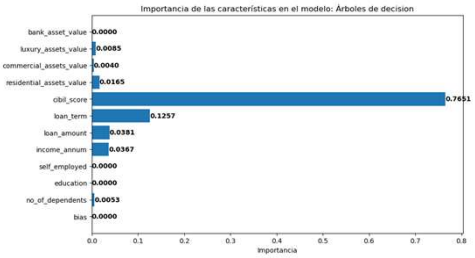
Resultados obtenidos:



Como se observa en la gráfica, la mayoría de los modelos dan un gran rendimiento y generalización. Sin embargo, el modelo que mejor ha clasificado ha sido el modelo de Árboles de decisión. El rendimiento de este modelo ha sido del 99,6% para el conjunto de entrenamiento del 98,47% para el de test.

Aunque hay modelos, que han obtenido un 100% de aciertos en la clasificación del conjunto de entrenamiento, el modelo elegido ha sido este porque es el que mejor generalización ha obtenido, es decir, el que mejor ha clasificado los ejemplos de test, ejemplos no utilizados en el entrenamiento.

Por otro lado, la característica del problema que mayor importancia ha tomado en la clasificación con el modelo de Árboles de decisión han sido la puntuación financiera del ciudadano con un 76,5%.



Conclusiones y líneas futuras

En este estudio sobre evaluación de préstamos mediante Aprendizaje automático, se han explorado diversos modelos para predecir la aprobación o rechazo de préstamos. En este proceso ha destacado el modelo de Árboles de Decisión que logra una gran capacidad de generalización, alcanzando un rendimiento del 98,47% en el conjunto de test.
Para poder mejorar el resultado sería adecuado mejorar el balance de clases ya que el número de ejemplos de préstamos concedidos es casi el doble que el de no concedidos.

upna

Lucas Fortún Iñurrieta
Universidad pública de Navarra