Modelos discriminantes

A partir de este apartado, se usará nuestra base de datos con el objetivo de predecir la variable target, en nuestro caso, el hecho de que un cliente se declare moroso. Para ello, se realizarán muchos modelos diferentes con el fin de predecir a cada uno de los clientes. Así pues, se comenzará por el más sencillo de todos: el LDA.

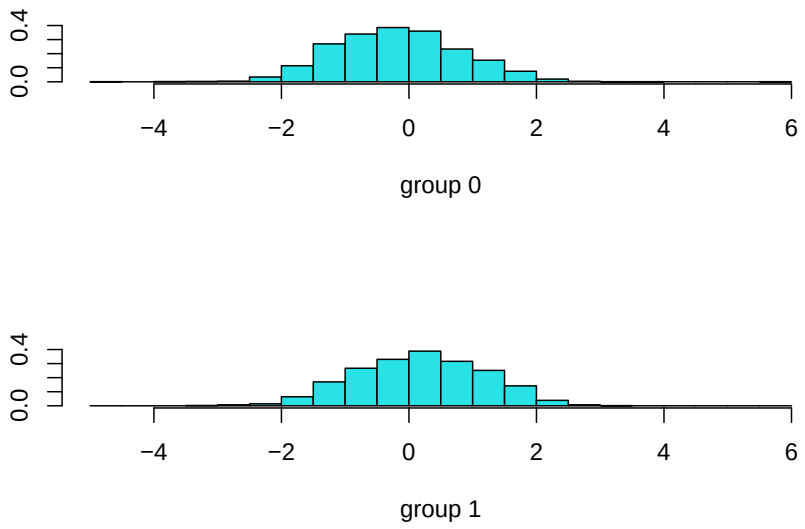
LDA (Linear Discriminant Analysis)

Para comenzar con los modelos discriminantes, se realizará en primer lugar un linear discriminant analysis (LDA) con el objetivo de intentar separar aquellos clientes que puedan tener dificultades de pago con aquellos solventes. Así pues, se procede a realizar dicho análisis discriminante.

Para ello, se recurrirá primero a un proceso de escalado de los datos a través de la función `scale()`, lo cual hará que todas las variables tengan un peso similar en la construcción del discriminante lineal. Una vez se ha realizado este proceso, el siguiente paso será realizar la partición de la base de datos disponible. Para ello, se realizará una partición clásica: el 80 % de los datos se destinarán a entrenar el modelo y el otro 20, a validarlo. Además, dentro de la partición del train se realizará un proceso 10-fold validation con el objetivo de reducir el overfitting y proporcionar un modelo robusto.

En el gráfico inferior se puede apreciar la proyección de cada observación sobre el discriminante:

Figura 122: Proyección de las observaciones sobre el discriminante para cada una de las clases LDA



Como se puede apreciar, los histogramas de las proyecciones se solapan entre ellos, lo cual da una idea que el LDA no es el modelo que mejor discrimina entre las clases. Sin embargo, se realizará más adelante la matriz de confusión.

Antes de analizar los resultados obtenidos por el LDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 5-fold cross validation ha sido del 0.59175, lo cual muestra unos resultados ciertamente pobres. Seguidamente, se ha validado el modelo contra el conjunto validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 46: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	502	316
Potencial moroso	71	111

Apreciando los resultados obtenidos, se puede ver que la precisión obtenida por el modelo ha sido del 61.3 %, algo baja en comparación con ejemplos en otras áreas. Si desglosamos por sensibilidad y especificidad, vemos que los resultados en estos dos indicadores han sido de 87.61 % , pero una especificidad del 26 %. Esta gran diferencia entre las dos métricas indica que el algoritmo tiene problemas para detectar la clase minoritaria, en este caso, los clientes morosos. Así pues, será necesario balancear nuestros datos para así conseguir resultados aceptables. Será necesario mejorar el dato de especificidad para así poder aceptar este algoritmo como válido. Adicionalmente, el valor del F-score es de 0.7213. Como otras métricas interesantes, se puede apreciar que el valor predictivo positivo es de 61.37 % y el valor predictivo negativo es de 60.99 %.

Sin embargo, se sabe que el LDA puede presentar problemas en el momento en el que las variables no presentan normalidad o cuando las matrices de covarianzas son diferentes para cada grupo. Como ya se apreció en la descriptiva post-preprocessing, muchas de nuestras variables no presentaban normalidad, de forma que esto podría ser un problema de cara al uso del LDA. Es por eso por lo que se ha decidido realizar un QDA (Quadratic Discriminant Analysis) con el objetivo de corregir dichos problemas y mejorar la performance del LDA.

QDA (Quadratic Discriminant Analysis)

Así pues, repitiendo el procedimiento seguido anteriormente en el LDA, toca repetir los mismos pasos para este modelo. De esta forma, los resultados obtenidos son los siguientes:

Antes de analizar los resultados obtenidos por el QDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 5-fold cross validation ha sido del 0.59225, lo cual muestra unos resultados ciertamente pobres, pero mejores que LDA. Seguidamente, se ha validado el modelo contra el conjunto validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 47: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	344	217
Potencial moroso	229	210

Como se puede apreciar, los resultados obtenidos son bastante similares a los presentados en el discriminante lineal. De hecho, en este caso, la precisión ha sido del 55.4 %, algo peor que la del LDA. Si observamos sensibilidad y especificidad, apreciaremos que se ha obtenido una sensibilidad del 60.0349 %, pero una especificidad del 49.1803 %. Si observamos otras métricas disponibles, apreciaremos nuevamente valores altos en la tasa de valores positivos predecidos (61.3191 %) y valores bajos en la tasa de valores negativos predecidos (47.836 %). Sin embargo, la diferencia entre estos no es tan extrema como en LDA. Por último, podemos

apreciar que el valor del F-score es de 0.576244. Como se puede apreciar, las conclusiones que se extraen son las mismas que en LDA: es necesario balancear los datos.

Tras haber estudiado los resultados, se ha concluido que, si bien es cierto que los resultados obtenidos son mejorables, los datos presentan un ligero desbalanceo. Este hecho hace que las esimaciones proporcionadas puedan no ser del todo fiables, ya que es posible que los algoritmos tengan problemas en la detección de la clase minoritaria. Para ello, realizaremos un procedimiento undersampling con el objetivo de balancear nuestros datos evitando una mala calibración de los algoritmos. Así pues, repetimos el mismo proceso que el realizado anteriormente con los datos ahora balanceados.

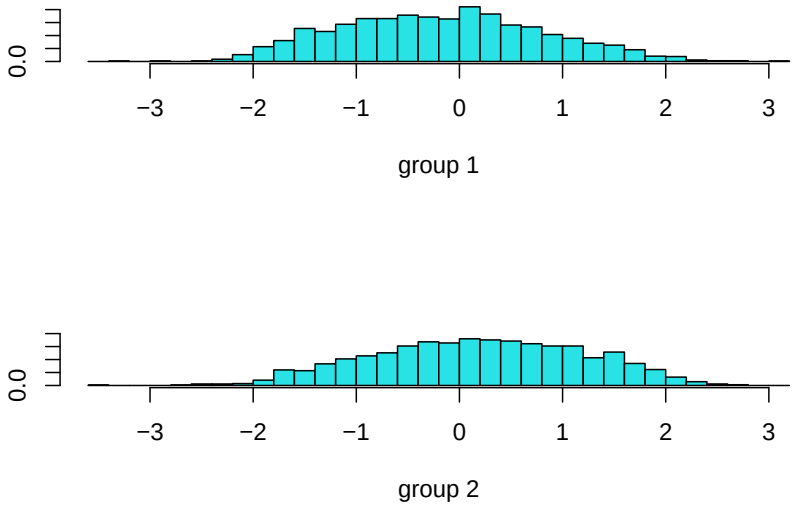
LDA (usando datos balanceados con undersampling)

En primer lugar, estandarizamos los datos usando la función de R `scale()` para así hacer que todos las variables tengan el mismo peso. Una vez los datos han sido normalizados y tienen todas las variables el mismo peso en el modelo, aplicamos undersampling:

Seguidamente, se realiza una partición del dataset entre train y validation con el objetivo de conseguir aproximaciones correctas que puedan ser usadas, evitando el over-fitting. Para ello, se realizará una partición 80-20 de la base de datos. Además, se realizará un 10-fold validation dentro del conjunto train, eliminando así cualquier problema de overfitting que pudiera existir.

En el gráfico inferior se puede apreciar la proyección de cada observación sobre el discriminante:

Figura 123: Proyección de las observaciones sobre el discriminante para cada una de las clases LDA (datos balanceados)



Como se puede apreciar, los histogramas de las proyecciones se solapan entre ellos, lo cual da una idea que el LDA no es el modelo que mejor discrimina entre las clases. Sin embargo, se realizará más adelante la matriz de confusión para acabar de aclarar este caso.

Antes de analizar los resultados obtenidos por el QDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 5-fold cross validation ha sido del 0.5635713, lo

cual muestra unos resultados ciertamente pobres, pero mejores que LDA. Seguidamente, se ha validado el modelo contra el conjunto validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 48: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	252	189
Potencial moroso	175	238

Esta vez, los resultados son más consistentes: los resultados obtenidos son muy diferentes a cuando los datos estaban desbalanceados. En este caso, la precisión obtenida ha sido del 57.38 %, dividido en una sensibilidad del 59.02 % y una especificidad del 55.74 %. Esta vez, se puede comprobar que los valores proporcionados por los datos de sensibilidad y especificidad están más balanceados, de forma que estos resultados parecen mucho más fiables.

Sin embargo, se sabe que el LDA puede presentar problemas en el momento en el que las variables no presentan normalidad o cuando las matrices de covarianzas son diferentes para cada grupo. Como ya se apreció en la descriptiva post-preprocessing, muchas de nuestras variables no presentaban normalidad, de forma que esto podría ser un problema de cara al uso del LDA. Es por eso por lo que se ha decidido realizar un QDA (Quadratic Discriminant Analysis) con el objetivo de corregir dichos problemas y mejorar la performance del LDA.

QDA (usando datos balanceados con undersampling)

Así pues, repitiendo el procedimiento seguido anteriormente en el LDA, toca repetir los mismos pasos para este modelo. De esta forma, los resultados obtenidos son los siguientes:

Antes de analizar los resultados obtenidos por el QDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 5-fold cross validation ha sido del 0.5805279, lo cual muestra unos resultados ciertamente pobres, pero mejores que LDA. Seguidamente, se ha validado el modelo contra el conjunto validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 49: Matriz de confusión del conjunto de validación

	Realidad	
	No moroso	Potencial moroso
Predicción		
No moroso	130	96
Potencial moroso	297	331

Por último, el modelo del QDA con datos balanceados presenta resultados peores al LDA. La precisión alcanzada por este modelo ha sido del 53.98 %, lo que retorna una sensibilidad del 30.44 % y una especificidad del 77.52 %. Este modelo corrige los problemas de desbalanceo que existían previamente, ya que ahora se puede apreciar cómo los resultados del modelo son más consistentes. Concretamente, el valor predictivo positivo y el valor predictivo negativo para este modelo son de 57.5221 % y 52.707 %, respectivamente. Para acabar, el F-score de este modelo es de 0.389324, algo más bajo que en LDA.

En resumen, observando los resultados obtenidos, se puede afirmar que los dos modelos discriminantes presentan resultados muy pobres: es probable que el hecho de añadir posteriormente las variables categóricas acabe de hacer que se mejore de forma clara los resultados conseguidos hasta ahora.