# Modelos discriminantes

A partir de este apartado, se usará nuestra base de datos con el objetivo de predecir la variable target a partir de los nuevos datos, 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.

Como se preeverá, será necesario usar las dos bases de datos: la desbalanceada y la balanceada. Desde el grupo se es consciente que los resultados que se mostrarán contra la base de datos desbalanceada serán malos, ya que los modelos serán incapaces de detectar la clase minoritaria. Sin embargo, este paso es necesario para justificar que se balancea la base de datos. Así pues, se empezará con el modelo más sencillo, el Linear Discriminant Analysis (LDA).

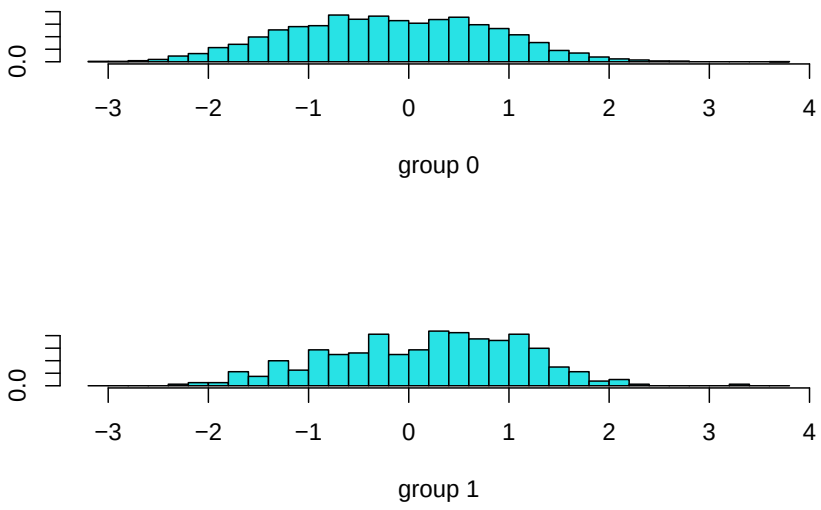
## LDA (Linear Discriminant Analysis)

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.

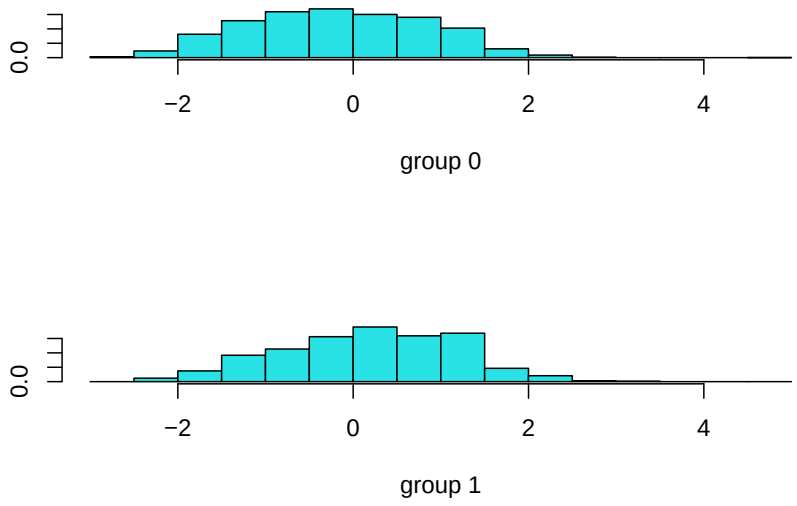
Antes de analizar los resultados obtenidos por el LDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 10-fold cross validation ha sido del 0.9198, lo cual muestra unos resultados ciertamente pobres. Seguidamente, se ha validado el modelo 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 | 919 | 80 |
| Potencial moroso | 0 | 0 |

Como se puede apreciar, los resultados son los esperados: al utilizar datos desbalanceados, el modelo no detecta bien la clase minoritaria, de forma que todas las predicciones de los datos llevaban a predecir todo como clientes no morosos. Así pues, se ha decidido aplicar este algoritmo a los datos ya balanceados (usando oversampling y undersampling a la vez):

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



Cuadro 47: Matriz de confusión del conjunto de validación

| | Realidad | |
|---|---|---|
| | No moroso | Potencial moroso |
| **Predicción** | | |
| No moroso | 284 | 203 |
| Potencial moroso | 220 | 292 |

Como se puede apreciar en los resultados de la matriz de confusión están bastante más balanceados. Apreciando los resultados obtenidos, se puede ver que la precisión obtenida por el modelo ha sido del 57.66 %, 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 58.99 % y una especificidad del 56.35 %. Así pues, el modelo ha sido capaz de detectar correctamente el NA % de clientes potencialmente morosos, lo cual puede ser un resultado bajo, pero asumible. Adicionalmente, el valor del F-score es de 0.57. Como otras métricas interesantes, se puede apreciar que el valor predictivo positivo es de 57.03 % y el valor predictivo negativo es de 58.32 %.

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 0.5871029, 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 | 326 | 221 |
| Potencial moroso | 178 | 274 |

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 60.0601 %, algo mejor que la del LDA. Si observamos sensibilidad y especificidad, apreciaremos que se ha obtenido una sensibilidad del 55.3535 % (peor que LDA), pero una especificidad del 64.6825 % (algo peor que el LDA). Si observamos otras métricas disponibles, apreciaremos una tasa de valores positivos predecidos de 60.6195 % y una tasa de valores negativos predecidos de 59.5978 %. Este hecho implica que que al predecir una clase, la probabilidad de que ésta sea clasificada correctamente es de entorno al 60 %. Por último, podemos apreciar que el valor del F-score es de 0.6228565,

métricada perjudicada por el bajo valor de la sensibilidad. Así pues, se podría decir que el modelo más útil entre estos dos es el LDA, ya que detecta de forma más consistentes el número de morosos.

En resumen, observando los resultados obtenidos, balanceando los datos se obtienen resultados más interesantes: el modelo es capaz de predecir e idenntificar las dos clases por igual. Sin embargo, 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.