Modelo predictivo para determinar el cumplimiento de pago de las tarjetas de crédito de clientes de un banco en Taiwan

Rosa Rodriguez, Luis Gerardo
Universidad Esan
Lima, Perú
18100115@ue.edu.pe

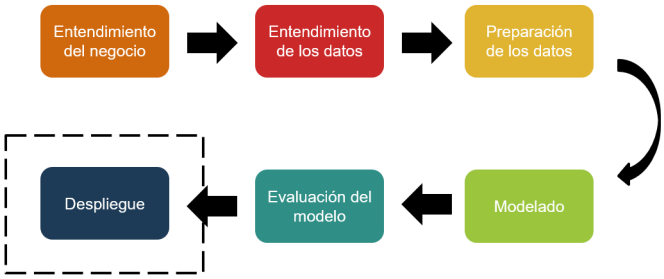
I. Introducción

En la actualidad, el manejo y uso de grandes cantidades de datos almacenados es una realidad para la gran mayoría de empresas que presentan datos de sus clientes, transacciones, procesos, etc., lo que permite mantener a estas empresas informadas acerca del estado de estos y de su comportamiento, Sin embargo, al momento de querer tomar decisiones necesitan entender y conocer los comportamientos de algún cliente, transacción o proceso que no se haya visto antes. Para ello, a modo de poder predecir estos comportamientos, se cuenta con el concepto de minería de dato, la cual se define como un proceso en el que se descubre nuevas y significativas relaciones, patrones, y/o tendencias cuando se examinan grandes cantidades de datos, por lo que hace posible el desarrollo de modelos que ayuden a predecir de forma precisa los comportamientos de diferentes objetos de estudio.

Por lo tanto, en esta investigación se tiene como propósito desarrollar un modelo de clasificación predictivo que permita determinar de la mejor manera posible y con los indicadores más adecuados el pago de las tarjetas de crédito de los clientes de un banco ubicado en Taiwán. De este modo, lo que busca el modelo es conocer si los clientes, dependiendo de distintos factores como su sexo, edad, nivel de educación, estado civil, y pagos pasados; sí pagarían su tarjeta a tiempo, o en caso contrario no. Para ello, se usará el programa R Studio, que permite el manejo de datos con fines estadísticos. El modelo será descrito con mayor precisión a medida que se desarrolle esta investigación, y para su interpretación se usarán los indicadores necesarios para determinar el mejor modelo de entre todos los presentados.

II. Metodología CRISP-DM

A modo de manejar un proceso de inicio a fin, y de mantener organizado y ordenado todos los procedimientos que se realicen para el desarrollo de estos modelos, se usará la metodología CRISP-DM, que en efecto, consiste en darle un sentido orientado a los proyectos de minería de datos. Esta metodología incluye de forma concisa la descripción por cada fase del proyecto, las tareas necesarias que se tengan que realizar en cada fase, y la explicación de las existentes relaciones entre cada una de estas fases. Las fases que conforman la metodología CRISP-DM son: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación del modelo, y despliegue. El proceso de esta investigación en relación las fases se describe en el siguiente diagrama:



III. Entendimiento del negocio

El entendimiento del negocio gira en torno a explicar el problema principal que se puede identificar en el contexto del negocio. En este caso, se tiene que gran cantidad de clientes de un banco en Taiwan que no pagan a tiempo sus tarjetas de crédito, lo cual afecta tanto al cliente, ya que daña su historial crediticio, así como al banco, que recibe quejas y reclamos por tomar acciones cuando un cliente no paga a tiempo su tarjeta. Este problema es tan grave que, de cada 100 personas, 78 no pagan a tiempo, mientras que solo 22 sí.

IV. Entendimiento de los datos

En relación al entendimiento de los datos, se busca describir las variables recolectadas de la empresa con la que se planea trabajar y determinar las unidades en las que se encuentran respectivas variables. En este caso, los datos con los que se trabajan pertenecen al Departamento de Gestión de la Información de la Universidad de Chung Hua, en Taiwán; y del Departamento de Ingeniería Civil de la misma universidad. El dataset contiene una lista de 23 variables predictoras, y una variable de salida. Estos se describen en el siguiente cuadro:

	Variables	Unidades
Y1	Realización del pago	1 = Si paga, 2 = No paga
X1	Monto del crédito	Dólares
X2	Género	1 = Masculino, 2 = Femenino
X3	Eduación	1 = Posgrado; 2 = Universidad; 3 = Secundaria; 4 = Otros
X4	Estado civil	1 = Casado; 2 = Soltero; 3 = Otros
X5	Edad	Numérico
X6-X11	Historial de los pagos pasados de septiembre a abril de 2005	-1 = Pago debido; 1 = Retraso de 1 mes; 2 = Retraso de 2 meses; . . .; 9 = Retraso de 9 meses
X12-X17	Monto del estado de cuenta de septiembre a abril de 2005	Dólares
X18-X24	Monto del pago anterior	Dólares

V. Preparación de los datos

Para poder realizar el modelado, se tiene que realizar un paso previo, que hace referencia a preparar los datos para su uso, Este proceso se le llama, preprocesamiento de los datos y para este caso, con la herramienta R Studio, se pudo preparar los datos de forma correcta para su modelado, realizando lo siguiente:

- Lectura de datos: El R Studio carga la base de datos que está en formato Excel.
- Eliminar variable ID: La variable ID no se considera ni predictora ni variable de salida, por lo que se elimina del modelo.
- Imputación de valores perdidos: En algunas ocasiones, no se logra almacenar, o se hace de forma incorrecta, el valor de un dato para una variable, lo cual puede afectar a los resultados del modelo. Para ello se realiza una imputación de los datos faltantes usando el algoritmo KNN.
- Selección de variables: Una gran cantidad de variables no es sinonimo de un mejor modelo, ya que no siempre todas las variables ayudan a predecirlo, algunas pueden ser innecesarias o incluso perjudiciales. Por ello, se realiza una selección de que variables usando el algoritmo Boruta. Los resultados indican que se deben eliminar las variables Sexo y Educación, teniendo así un dataset con 21 variables predictoras.
- Selección de muestra de entrenamiento y de evaluación: A fin de que el modelo aprenda a predecir, se requiere previamente que el modelo sea entrenado, para así luego poder compararlo con un conjunto de datos reales. Por lo tanto, se selecciona una muestra de entrenamiento, que será el 70% de todo el dataset, y el 30% restante de prueba.
- Preprocesamiento general: En esta parte se realiza un preprocesamiento general, el cual incluye la eliminación de variables que presenten poca correlación y variancia con la variable de salida, Asimismo, se estandarizan las variables númericas para que sus valores tengan un rango de 0 a 1, y se crean las variables dummy de las variables categóricas. Los resultados son un dataset que pasa de tener 21 variables a 66 variables predictoras.
- Balanceo de datos: Algunos dataset presentan un desbalanceo de datos, lo que significa que en la variable de salida se tendrá una gran diferencia entre las categorías que tenga, lo cual no permitiría que se realize un buen modelado. En este caso, se requiere hacer un balanceo de datos ya que como se mencionó antes, un 78% no pagan a tiempo su tarjeta de credito y un 22% sí, habiendo una diferencia de más de 50%. Para evadir este problema, se realiza un balanceo de datos utilizando el algoritmo Smote, teniendo como resultados de 50%-50%, donde 13938 personas no pagan a tiempo, y 13938 sí.

VI. Modelado

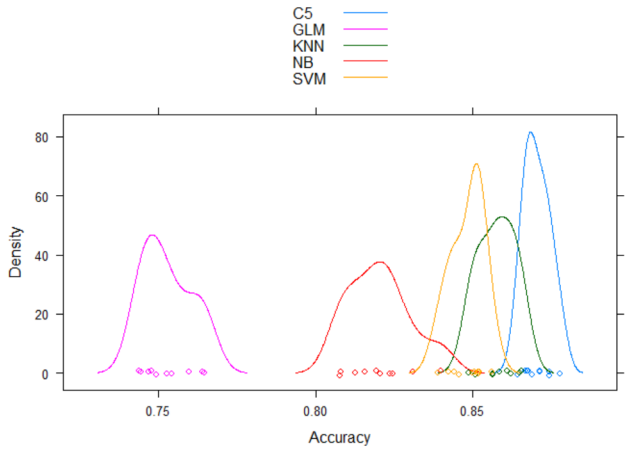
Con los datos ya preprocesados se puede proceder a realizar el modelado, y para ello, se hará uso de 5 algoritmos de clasificación y 1 técnica de stacking observados en clase. Cada uno de los modelos descritos a continuación presentan parámetros que, cuando se ajustan, permiten obtener mejores resultados que usar sus parámetros por defecto; en este caso, se usaron los mejores parámetros para cada modelo. De la misma forma, a cada modelo se le puede aplicar un umbral óptimo el cual busca mejorar la predicción del modelo, mas no siempre lo mejora ya que incluso en algunas ocasiones puede disminuir la precisión de la predicción, pero de todas formas se usará en el proceso y se podrá apreciar en un cuadro comparativo final, el cual se detallará en la fase de evaluación del modelo. Por otro lado, cabe destacar que el modelado se aplicará a la data de entrenamiento, de modo que el algoritmo aprenda con esa data y así posteriormente aplicarlo en la data de prueba a modo que prediga y lo compare con esos mismos datos.

En primer lugar, se debe definir una validación cruzada de 10 iteraciones, lo que quiere decir que el conjunto de data de entrenamiento se dividirá en 10 subconjuntos, así cada modelo cuando use la data de entrenamiento, la evaluará por partes, hasta completar las 10 totales. Los algoritmos de clasificación son los siguientes:

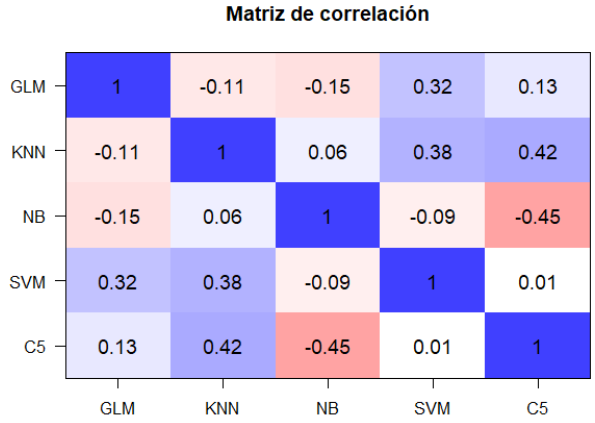
- Regresión Logística (GLM)
- K-Nearest Neighbors (KNN)
- Naive Bayes (NB)
- Support Vector Machine (SVM)
- Árboles de decisión C5.0 (C5)

La técnica de stacking dependerá de escoger los mejores 3 algoritmos de clasificación y ensamblarlos en un mismo modelo, y a ese modelo aplicarle otro algoritmo de clasificación, el cual será, el algoritmo Clasification And Regression Tree (CART).

De este modo, se realiza el respectivo modelado por cada algoritmo. Para interpretar los datos se hará uso de un gráfico de densidad y una matriz de correlaciones, ya que ayudan a entender, de forma resumida, cuales son los mejores algoritmos. El gráfico de densidad se muestra a continuación:



Para poder determinar cuales son los mejores modelos y así usarlos en el modelo de stacking, se tiene que tener en cuenta que tengan el mayor accuracy, y su gráfica presente una forma de campana alargada hacia arriba (a esto se denomina leptocúrtica). En este gráfico de densidad se descarta directamente el algoritmo GLM ya que presenta un accuracy mucho menor. Por otro lado, los tres modelos con el mayor accuracy serían C5, KNN y SVM, y se podría tomar en cuenta que NB también es bueno pero su gráfica no tiene una forma leptocúrtica como las otras, por lo que los algoritmos escogidos momentáneamente serían C5, SVM y KNN. Ahora se procede a comparar los algoritmos en una matriz de correlaciones, la cual es la siguiente:



Para entender el gráfico, se escogen dos algoritmos, uno por cada eje, y nos ubicamos en la intersección. Una buena correlación se aprecia cuando el índice se encuentra entre -0.5 y 0.5, en caso sea mayor, existe mucha correlación entre los algoritmos, por lo que no se deberían usar en conjunto para el ensamble de stacking. En este caso, nos fijamos que KNN y C5 presentan una correlación de 0.42, lo cual se encuentra dentro del rango. Del mismo modo KNN y SVM, y SVM y C5, presentan correlaciones dentro del rango óptimo, siendo de 0.38 y 0.01.

Con esta información recolectada se llega a la conclusión de que los algoritmos que se procederán a ensamblar a modo de stacking son C5, SVM y KNN. De esta forma, se realiza el respectivo cálculo, tomando en cuenta sus respectivos mejores parámetros ajustados, y se aplica el algoritmo CART al modelo ya ensamblado. Posteriormente a esto, se realiza el cálculo nuevamente pero esta vez con el umbrál óptimo, así como los modelos de clasificación.

VII. Evaluación del modelo

Con los resultados de los modelos de clasificación y el modelo de ensamle con stacking, se procede a identificar cual de estos logra predecir con mayor precisión los datos de la variable de respuesta. Para ello, se harán uso de dos indicadores llamados Sensibilidad, y Accuracy Balanceado ya que nuestro dataset necesitó realizarle un balanceo previo como se menciona en la fase de preprocesamiento. Por un lado, se entiende que la sensibilidad es la proporción de la variable de respuesta, en este caso la realización del pago, que se identifica correctamente cuando sale positivo, mientras que por el otro lado, el accuracy balanceado, o también llamado precisión balanceada, mide la precisión con la que se realizó la predicción. Los resultados se pueden observar en el siguiente cuadro comparativo:

Umbral óptimo	Algoritmo	Sensibilidad	Accuracy Balanceado
No	Regresión Lineal	0,54673	0,6888
Si	Regresión Lineal	0,57337	0,6916
No	KNN	0,51407	0,6243
Si	KNN	0,5201	0,6264
No	Naive Bayes	0,02714	0,5117
Si	Naive Bayes	0,65477	0,674
No	SVM	0,41759	0,6716
Si	SVM	0,55628	0,6934
No	Arbol C5	0,43568	0,6756
Si	Arbol C5	0,43568	0,7053
No	Stacking CART	0,48342	0,6773
Si	Stacking CART	0,50402	0,6854

En términos de sensibilidad, se puede observar que el que presenta el mayor índice es Naive Bayes con umbral óptimo con 0.65477, seguido de la Regresión Lineal con umbral óptimo con 0.57337 y en tercer lugar SVM con umbral óptimo con 0.55628. Esto quiere decir que un 65.5%, 57.3% y 55.6% de los casos donde sí realizan el pago se lograron predecir de forma correcta usando Naive Bayes con umbral óptimo,

Regresión Lineal con umbral óptimo y SVM con umbral óptimo.

Por otro lado, en términos de accuracy balanceado, se puede determinar que el mayor indice corresponde a Arbol C5.0 con umbral óptimo con 0.7053, luego está SVM con umbral óptimo con 0.6934, y después Regresión Lineal con umbral óptimo de 0.6916. Esto significa que un 70.5%, 69.3% y 69.2% de los casos se logró predecir si una persona pagaba o no a tiempo correctamente, usando Arbol C5.0 con umbral óptimo, SVM con umbral óptimo y Regresión Lineal con umbral óptimo.

En base a esta información se obtiene que los mejores modelos predictivos son el modelo de Regresión Lineal con umbral óptimo, y el modelo de SVM con umbral óptimo. Sin embargo, se debe escoger solamente un modelo, para ello nos fijamos nuevamente en los resultados.

Umbral óptimo	Algoritmo	Sensibilidad	Accuracy Balanceado
Si	Regresión Lineal	0,57337	0,6916
Si	SVM	0,55628	0,6934

La diferencia de los accuracy balanceado de ambos modelos es de 0.0018, con ventaja en SVM, mientras que la diferencia de los índices de sensibilidad es de 0.0171, con ventaja en Regresión Lineal. Con esto se entiende que la ventaja que tiene la Regresión Lineal con umbral óptimo en la sensibilidad es mayor que la ventaja que tiene SVM con umbral óptimo en el accuracy balanceado, por lo que finalmente se puede determinar que el mejor modelo predictivo para determinar el cumplimiento de pago de las tarjetas de crédito de clientes en un banco de Taiwan es el modelo de Regresión Lineal con umbral óptimo, el cual logra predecir con una exactitud de 69.2%.

VIII. Despliegue

El despliegue del modelo se podría desarrollar de forma que el banco de Taiwan que presenta estos datos podría usar el modelo para predecir con anticipación si un cliente pagará su tarjeta de crédito o no a tiempo y así brindarles opciones de pagos distintas a las que ya conoce, o tal vez ofrecerle un servicio diferente en caso sea un cliente que paga con frecuencia y así darle más facilidades de pago, o también en caso sea un cliente que no paga con frecuencia, darle opción a que pueda realizar los pagos en periodos más largos, ya sea el caso que se realizen de forma mensual, pasarlo a un periodo bimestral, donde pague cada dos meses, así se les brindaría más tiempo para recaudar el dinero y paguen a tiempo.

IX. Conclusiones

A lo largo de esta investigación, se concluye que las técnicas de machine learning para el desarrollo de modelos predictivos son de suma importancia para las empresas, ya que les permitiría realizar predicciones con el fin de llegar a conocer mayormente a sus clientes, para así brindarles un mejor servicio, ya sea personalizado, o que se ajuste mejor a la situación de cada persona. Asimismo, ayudaría mucho a la toma de decisiones ya que estas tecnicas permiten conocer el mercado y quienes lo integran, sería muy útiles para saber como afrontar las debilidades o amenazas empresariales.

Por otro lado, en relación al modelo trabajado en esta investigación, los resultados que se mostraron fueron óptimos, ya que el accuracy balanceado que se obtuvo, se encuentra en un rango perfecto, ya que no es ni demasiado bajo como para causar underfitting, ni muy alto como para causar overfitting. Por otra parte, este trabajo me permitió entender más a profundidad los algoritmos de clasificación, a fin de saber como manejarlos en un futuro, y así aplicar el conocimiento y experiencias que brindan estos y la minería de datos en la vida real.

References

[1] C. Pérez Lopez, D. Santín Gonzales. “Minería de datos. Técnicas y herramientas: técnicas y herramientas”, 1st ed., Thomson. 2008

[2] I. Cheng Yeh. “Default of credit card clients Data Set”, 1st ed., UCI Machine Learning Repository. 2009