Evaluación de modelos de predicción

Universidad Cenfotec
Data Analytics & Big Data

Grettel Rojas Conejo

Contenido

Introducción

El requerimiento de este análisis se basa en que se ha identificado que han aumentado las tasas de incumplimiento de los clientes, lo cual es algo negativo para Credit One, ya que su labor es aprobar los préstamos de acuerdo con información crediticia y sus respectivos análisis.

En el presente informe tiene como objetivo describir los pasos que se realizaron para generar y optimizar un modelo que permita predecir el pago de un cliente el próximo mes, si este pago se hará efectivo o si se incumplirá, esto con el fin de que la compañía pueda tomar decisiones y acciones ante posibles incumplimientos de sus clientes. De igual forma se presenta un análisis con los resultados obtenidos.

Análisis de datos y preprocesamiento

El dataset tiene las siguientes características:

- 30000 registros
- 25 columnas
- 24 columnas de predictores
- Variable a predecir: comportamiento del cliente ("Incumple" o "No Incumple“)

Se realizaron tareas de preprocesamiento, esto con el fin de disminuir los tiempos de procesamiento y el esfuerzo en el entrenamiento de los modelos. Las tareas que se realizaron se detallan a continuación

- **Se elimina la de columna de los ID**

Se determina que esta columna no agrega valor para el análisis, ya que se refiere solamente a un consecutivo de referencia y no está relacionado con el problema en cuestión.

- **Cambio de tipo de variables**

Todas las variables que poseía el dataset eran de tipo entero, por esta razón se procedió a cambiar las variables que debían ser de tipo categoría, que se detallan a continuación:

- SEX
- EDUCATION
- MARRIAGE
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5
- PAY_6
- default payment next month

Análisis de correlación

Se realizó una matriz de correlación entre las diferentes variables, con el fin de identificar variables que se pudieran eliminar, al estar relacionadas con otras. Con este análisis se obtuvo lo siguiente:

- Las variables: BILL_AMT1 y BILL_AMT2; BILL_AMT2 y BILL_AMT3; BILL_AMT3 y BILL_AMT4; BILL_AMT4; y BILL_AMT5; BILL_AMT5 y BILL_AMT6 fuertemente relacionadas entre sí.
- Se eliminan las variables BILL_AMT2, BILL_AMT4 y BILL_AMT6 ya que no agregan valor al estar relacionadas en gran medida con otras.

Entrenamiento de los modelos

Se utiliza un valor de aleatorización 42. Para la predicción de la variable DefaultPaymentNextMonth, se procede a entrenar los algoritmos de clasificación, con el fin de evaluar los resultados de calidad de la predicción.

El dataset a entrenar se dividió en datos de Train y de Test. A continuación se detallan los porcentajes y cantidad de registros para cada subset. Considerando un 80% para training y un 30% para testing.

Una vez realizado el preprocesamiento y la partición de los datos, se procede a entrenar el set de datos de train, con cada uno de los algoritmos que fueron seleccionados**.** En el siguiente cuadro se muestran los resultados obtenidos de los parámetros de calidad en el set de datos de test:

Modelo	Accuracy	Kappa
Logistic Regression	0,7812	0,0000
Random Forest Classifier	0,8155	0,3623
Support Vector Classifier	0,7812	0,0000
Gradient Boosting Classifier	**0,8207**	**0,3693**
K Neighbors Classifier	0,7498	0,1040
Gaussian Naïve Bayes	0,3802	0,0650

Escogencia del modelo

En la tabla anterior, con respecto a los parámetros de calidad obtenidos con los datos de test, se determina que el modelo que obtuvo mejores resultados de Accuracy y Kappa es el generado con el algoritmo GBC (Gradient Boosting Classifier), por lo que se considera como el más óptimo a utilizar para predecir los datos de cumplimiento de los pagos.

Con el objetivo de optimizar los resultados obtenidos, se realizaron varias iteraciones de entrenamiento variando los valores de n estimators, learning rate, max features, max depth, con este análisis no se pudo alcanzar un valor que fuese significativamente mayor al resultado del modelo.

Resultados

Value	Real	Predicted
No Incumple	4687	5275
Incumple	1313	725

En la tabla anterior se muestra un resumen de lo que el modelo logró predecir, con respecto a los valores reales. Se puede observar que se logró predecir correctamente el 82% de los datos del set de test.

El resultado de Accuracy es bastante bueno, 0,82, en embargo el Kappa está muy por debajo de lo esperado 0,36. Lo que indica esto es que a modo general y como un todo el modelo logró predecir los datos en gran medida, sin embargo en el detalle de cada uno de los casos es posible que existan falsos positivos que no se estén prediciendo correctamente. Esto es que, para cada uno de los clientes en el set de datos, con su respectiva información, no se está prediciendo el valor puntual que debería.

Para poder obtener mejores resultados de predicción es recomendable invertir un mayor tiempo y costo en este análisis, con el fin de indagar en otros algoritmos más complejos, los cuales a su vez requieren un mayor detalle en su parametrización y procesamiento.