PRUEBA TÉCNICA APRENDICES EN ANALÍTICA DE RIESGOS

JORGE LUIS RODRIGUEZ LOPEZ
Estudiante de Ingeniería Industrial
Universidad de Antioquia

Resumen

En el presente informe se desarrollan casos prácticos sobre analítica de datos comprendidos desde una visión empresarial con características financieras. El primer caso abarca la elaboración de un modelo analítico predictivo, que a partir de una serie de características de los clientes, se busca evaluar su idoneidad como potenciales candidatos para adquirir un crédito con la compañía Nequi, este caso se llevó a cabo con programación en Python y algunas consultas externas con SQL. En el siguiente link se encuentra los archivos utilizados para el desarrollo planteado en este documento: https://github.com/Jorge-Roriguez/Prueba-Tecnica.git

1. CASO 1: MODELO ANALÍTICO DE CLASIFICACIÓN BINARIA

1.1 Caso de negocio

Para desarrollar un modelo analítico, primeramente es importante entender el problema de negocio para poder llevarlo a un problema analítico. En primera instancia, se dispone de una muestra de datos (características) la cual nos permite identificar un perfil transaccional de cliente, con el propósito de evaluar su idoneidad como potenciales candidatos para adquirir un crédito (objetivo).

Para llevar a cabo este problema de negocio, se utilizará un modelo analítico de clasificación binaria en el cual se espera obtener las probabilidades que un cliente se encuentre en default, es decir, que tan probable es que dicho cliente presente un estado de mora e incurra al incumplimiento de sus obligaciones con la compañía.

1.2 Análisis exploratorio

Para comprender de mejor manera los datos suministrados se analiza el comportamiento de algunas variables. Dado que los datos en su etapa inicial se encuentran normalizados, se dificulta el entendimiento de estos si se realiza un análisis profundo, por lo que sólo se analizarán las variables no normalizadas, las cuales son: 'Fechas_cruce', 'TransactionValue_PSE' y la variable objetivo 'target'.

Características de los datos: Se cuenta con 5000 registros cada uno con 203 características, de las cuales 200 están normalizadas. En la verificación de datos nulos no se encontraron evidencias de que existen, de igual manera tampoco se encontraron datos duplicados. Los datos presentan registros de cada fin de mes para todos los meses de los años 2022 y 2023.

Para un diagnóstico inicial, veamos a continuación en qué estado se encuentran los clientes respecto a sus obligaciones financieras con la empresa.

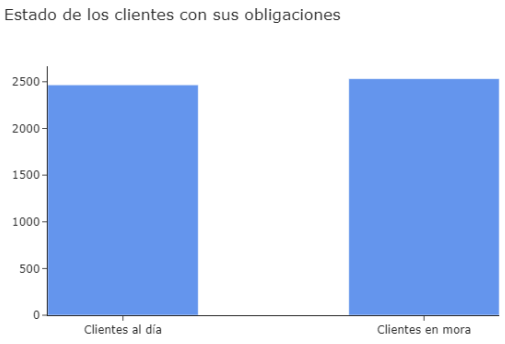

Gráfico 1. Estado de los clientes con sus obligaciones

Para el total de los registros en los datos, 2533 clientes se encuentran en estado de mora en sus obligaciones, lo que equivale un total del 50.66% de todos los clientes.

Gráfico 2. Estado de los clientes con sus obligaciones para 2022 y 2023

En el gráfico 2 se puede observar el comportamiento del estado en mora o al día de los clientes con sus obligaciones para los años 2022 y 2023, el comportamiento es muy similar sin embargo el cambio porcentual de clientes en mora pasa de ser del 51.15% para el 2022 al 50.18% para el 2023.

Para finalizar el análisis exploratorio veamos la distribución del valor de las transacciones vía PSE de clientes.

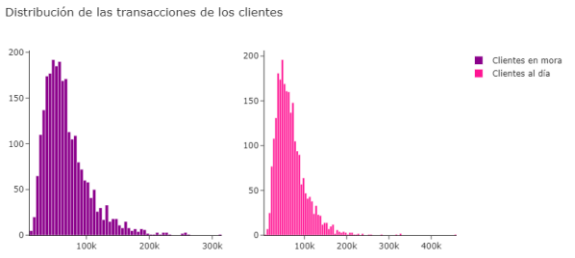

Gráfico 3. Distribución de las transacciones para los clientes.

Se observa que tanto para los clientes que se encuentran en estado de mora como para los que no, la distribución del valor de las transacciones realizadas se comporta en un rango muy similar, ambas sesgadas a la izquierda de la distribución, lo que indica que ambos tipos de clientes manejan valores transaccionales similar.

1.3 Preprocesamiento de datos

Para una correcta ejecución de los modelos analíticos, los datos con los que se entrenan deben estar normalizados, por lo que la variable 'TransactionValue_PSE' se escala y se agrega a los datos anteriormente normalizados. Adicionalmente se separa variables predictoras de la variable objetivo 'target'. Cabe resaltar que un supuesto a considerar dado que los datos fueron normalizados inicialmente, es que las variables no presentan problemas de multicolinealidad.

1.4 Selección de variables y evaluación de desempeño de modelos

Ya que el objetivo de estudio es realizar un modelo predictivo que permita saber la probabilidad de que un cliente se encuentre en estado de default, se proponen los siguientes algoritmos para la solución del problema analítico:

1. Regresión logística (LR)
2. Clasificador de bosques aleatorios (RF)
3. Clasificador XGBoost (XGB)

Estos algoritmos presentan las cualidades adecuadas para abordar el problema en cuestión, dado que son buenos clasificadores binarios, son fáciles de interpretar, poseen gran capacidad de generalización en los datos y son eficientes al trabajar con grandes cantidades de información.

Dada la cantidad de variables en los datos, es importante realizar un método de selección de variables, con el propósito inicialmente de encontrar las variables que expliquen de mayor manera la variabilidad de las predicciones y por otro lado ahorrar capacidad computacional. La selección de variables se realizó mediante la técnica de eliminación hacia atrás (Forward). El resultado obtenido indica las características más adecuadas que mejoran el rendimiento de los algoritmos. En total fueron 42 variables significativas para los modelos.

La medición del desempeño de los modelos se realizó tanto con todas las variables a disposición como con las variables seleccionadas, con el objetivo de saber si los modelos tienen mejor rendimiento con menos variables y si estas son capaces de predecir la variabilidad de la variable objetivo. Cabe aclarar que la métrica de desempeño seleccionada para evaluar los modelos es 'Accuracy', ya que en la gráfica 1 vemos que los datos se encuentran balanceados y es una métrica simple de comprender ya que califica la proporción de casos que el modelo clasificó correctamente.

En la tabla 1, podemos observar el desempeño de cada uno de los modelos para cuatro iteraciones en los datos, donde A hace referencia a la métrica con todas las variables y B hace referencia el valor de la métrica con las variables seleccionadas.

LR		RF		XGB	
A	B	A	B	A	B
0.497	0.520	0.482	0.533	0.488	0.536
0.471	0.561	0.492	0.532	0.510	0.547
0.487	0.503	0.500	0.525	0.484	0.520
0.492	0.561	0.506	0.538	0.487	0.542

Tabla 1. Desempeño de modelos con distintas cantidades de variables.

Se puede observar que las métricas mejoran en una pequeña proporción al evaluar los modelos con las variables seleccionadas, aunque el cambio porcentual en métrica sea poco, el cambio en datos procesados es grande, ya que solo el 21% de las variables explican de mejor manera las predicciones que todas las 200 variables. Con estos resultados podemos concluir que es mejor trabajar con las variables seleccionadas.

Ahora veamos un gráfico comparativo para el desempeño de los modelos con estas variables.

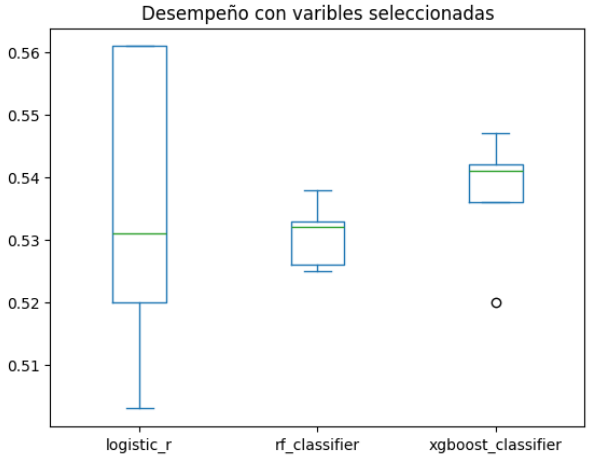

Gráfico 4. Distribución del desempeño de los modelos.

En el gráfico 4 vemos que el modelo de la regresión logística suele alcanzar un desempeño más alto que los otros dos modelos, sin embargo este no es tan estable dado que también alcanza los valores más bajos. El desempeño del modelo XGBoost es mucho más estable y en promedio este modelo suele tener un mejor desempeño que los otros, por esta razón ahora podemos concluir que la mejor combinación para realizar las predicciones es hacerlas con un modelo de clasificación XGBoost junto a las variables seleccionadas anteriormente.

1.5 Afinamiento de hiperparámetros para modelo XGBoost con variables seleccionadas

Los parámetros por considerar son los siguientes:

1. **Max_depth:** Para controlar la profundidad del árbol y evitar sobre ajustes en los datos.
2. **Eta:** Reduce las ponderaciones de las características para potenciar el proceso.
3. **Subsample:** Varia la proporción de la submuestra de la instancia de capacitación.

Los resultados obtenidos se muestran en la siguiente tabla.

Params	Mean test score
Subsample:0.4, Max_Depth:5, eta: 0.09	0.5484
Subsample:0.4, Max_Depth:3, eta: 0.2	0.5371
Subsample:0.5, Max_Depth:4, eta: 0.4	0.5358
Subsample:0.5, Max_Depth:5, eta: 0.4	0.5292

Tabla 2. Desempeño del modelo con afinamiento de hiperparámetros.

La métrica de desempeño al afinar los hiperparámetros no se vio tan afectada significativamente, a pesar de varias iteraciones y ampliar el rango de cada valor de los parámetros el mejor desempeño que se obtuvo fue que el modelo predijera correctamente el 56.72% de los datos.

1.6 Predicciones

Al realizara las predicciones de la probabilidad de que un cliente se encuentre al día con sus obligaciones con la compañía, se obtiene los siguientes resultados:

Clientes potenciales adquirir un crédito			
Cliente	Probabilidad	Cliente	Probabilidad
4354	92.65 %	3718	88.99 %
717	90.61 %	706	88.93 %
4010	89.52 %	4004	88.51 %
4919	89.51 %	509	88.30 %
1336	89.45 %	1480	88.12 %

Tabla 3. Clientes potenciales adquirir un crédito.

La tabla 3 muestra los mejores 10 clientes, los cuales tiene una probabilidad muy alta de que se encuentren al día con sus obligaciones financieras y por lo tanto sean clientes potenciales para adquirir un crédito con la compañía con mayor confianza.

Ahora bien, para tener resultados más concretos se puede realizar un análisis por intervalos de confianza para las probabilidades de que los clientes se encuentren al día con sus obligaciones, con el objetivo de saber las proporciones de clientes para cada segmento. Con esta premisa se realiza la siguiente segmentación:

A. Probabilidad alta de otorgar un crédito: Clientes con probabilidad mayor o igual al 80% de estar al día con sus obligaciones.
B. Probabilidad media de otorgar un crédito: Clientes con probabilidad de mayor o igual al 70% y menor al 80% de estar al día con sus obligaciones.
C. Probabilidad baja de otorgar un crédito: Clientes con probabilidad menor al 70% de estar al día con sus obligaciones

A partir de la segmentación anterior se obtiene los siguientes resultados:

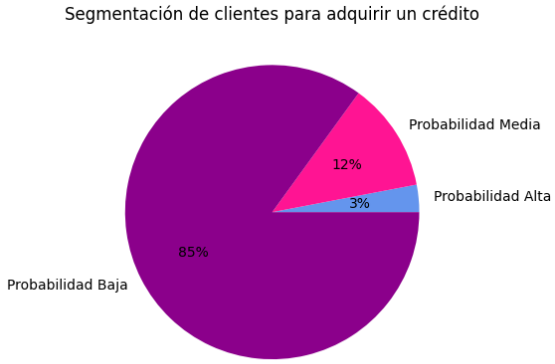

Gráfico 5. Segmentación de clientes para adquirir un crédito.

En el gráfico 5 vemos que tan solo el 3% del total de los clientes tiene alta capacidad de pago con sus obligaciones, por lo que estos clientes son potenciales para adquirir un crédito con la compañía. Por otro lado, vemos que el 85% de los clientes tienen una probabilidad baja de recibir un crédito.

1.7 CONCLUSIONES

- Tanto las predicciones obtenidas como la segmentación de clientes realizada pueden llegar a ser vital para la toma de decisiones estratégicas para la compañía, ya que permiten identificar clientes altamente potenciales que puedan adquirir un crédito con mayor confianza y que la compañía esté segura de que estos clientes no quedarán mal con sus obligaciones financieras.

- Es necesario evaluar las políticas que se tiene actualmente en la compañía respecto a los clientes en estado de mora, ya que al tener tantos clientes en mora con sus créditos implica que no se están obteniendo los pagos esperados, lo que afecta la liquidez y rentabilidad del banco, dado que se deben asociar costos adicionales para la gestión de morosidad como lo son las cobranzas.

- Para mitigar los efectos negativos mencionados en el anterior inciso, además de otras estrategias implementadas, es fundamental realizar análisis modelos analíticos para poder evaluar el riesgo de los créditos antes de ser otorgados.

- Dada la métrica de desempeño con los hiperparámetros afinados del modelo final, hay que tomar decisiones considerando que la métrica no se encuentra en su mejor punto y que esta pude llegar a ser mejorada.

- A partir de la métrica de desempeño obtenida hay que tener algunas consideraciones a mejorar. Alguna de ellas son: Realizar una búsqueda más exhaustiva de los parámetros que mejoran el desempeño del modelo, analizar de manera más profunda las variables antes de normalizarlas y validar supuestos de multicolinealidad. Es fundamental que el modelo analítico se encuentre bien entrenado ya que esto permitirá obtener resultados confiables y poder tomar decisiones con mayor precisión.

2. CASO 2: MODELO DE OPTIMIZACIÓN

Para el caso número 2 intenté formularlo desde distintas herramientas como programación lineal y optimización estocástica en Python. Sin embargo no me fue posible encontrar una forma para desarrollar el problema desde el conocimiento que actualmente tengo. Pido disculpas por no haber propuesto algún resultado para este caso. Si pudiesen brindame una retroalimentación para poder desarrollar este punto estaría gratamente agradecido.