Reporte Scorecard de Originación para créditos hipotecarios

Ángel Gustavo José Martínez

January 8, 2024

En este proyecto se desarrolla una Scorecard para la originación de créditos hipotecarios. En dicho desarrollo se usan técnicas de selección de características y remuestreo. Finalmente, el modelo se pone en producción de forma local. En el presente reporte se encuentran las especificaciones de la información usada y el proceso de modelado.

Palabras clave: Scorecard, Regresión Logística, Riesgo crédito, Regularización $L1$, Productivización.

1 Contexto

La implementación de Scorecards en la industria financiera lleva bastante tiempo en uso, pues a pesar de la existencia y posibilidad de implementar clasificadores más complejos como *Random Forest, SVM, AdaBoost, Gradient Boosting, XGboost* o incluso Redes Neuronales, no son de fácil interpretación (cajas negras), sobre todo para las áreas comerciales. Además, sulen consumir más recursos computacionales, lo cual se traduce en una facturación más costosa.

Ventajas de la Scorecard:

- Mayor intrpretabilidad
- Facilidad de implementación y menor gasto de recursos computacionales
- Los cambios o ajustes que se requieran en producción serán transparentes.

Desventajas:

- En algunos casos menor precisión que los algoritmos mencionados.

2 Fuentes de información

El conjunto de datos usado es público y se encuentra en: https://www.kaggle.com/competitions/home-credit-default-risk/data.
Se considera información relativa a:

- Solicitudes
- Solicitudes previas
- Comportamiento interno de pagos
- Buró de Crédito

El diccionario de columnas se encuentra en el repositorio correspondiente y el diagrama de relación es el siguiente:

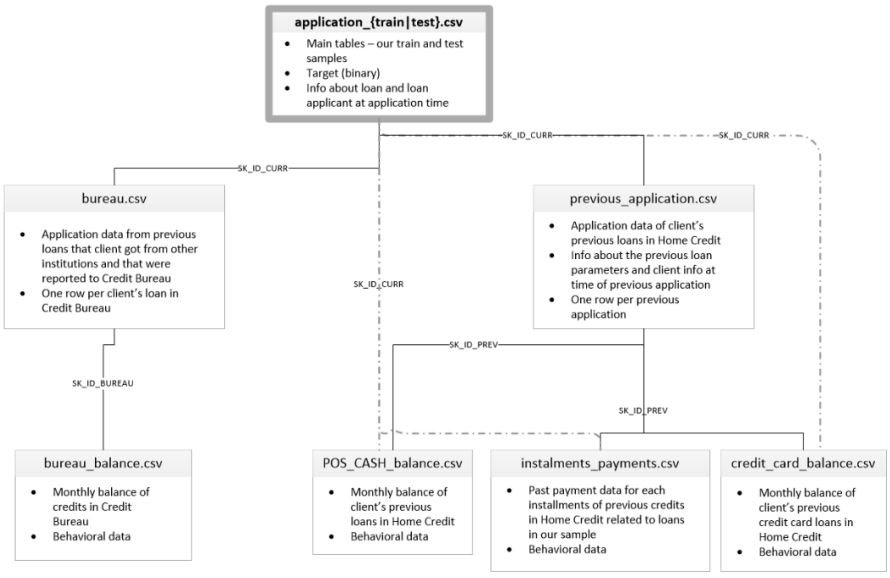

Figure 1: *Diagrama de relación entre tablas de información.*

2.1 Consideraciones

La target ya se encuentra definida y por tanto, la definición de cliente bueno y malo. Sin embargo, es importante mencionar que se deben definir mediante **matrices de transición y en consenso con las partes interesadas**. Además de la ventana de tiempo en la que se considere la información.

3 Desafíos

- Se tiene un desbalance en la distribución del target (8% Ctes malos vs 92% ctes buenos).
  - **Solución**: Se aplican dos técnicas de remuestreo: SMOTE (Synthetic Minority Oversampling Technique) y Undersampling, ambas combinanadas en un pipeline:

```
Pipeline
Pipeline(steps=[('smote', SMOTE(random_state=11, sampling_strategy=0.2)),
                ('under', RandomUnderSampler(sampling_strategy=0.4)),
                ['model',
                 LogisticRegression(max_iter=1000, random_state=10,
                                    solver='liblinear')]])
```

Figure 2: *Pipeline remuestreo y ajuste del modelo.*

- El conjunto de datos después de realizar ingeniería de características y cruces entre tablas, **cuenta con 220 columnas** y 3 millones de registros, lo cual conlleva a tener un problema de dimensionalidad.
  - **Solución**: Se aplican filtros de valores faltantes, correlación, **regularización** $L1$ y Valor de la Información (IV).

  El resultado de aplicar estas técnicas, es la selección de las 10 características con más poder predictivo para el modelo.

4 Variables usadas

- EXT SOURCE 3: Score 1 Normalizado del cliente construido con información externa.

- EXT SOURCE 2: Score 2 Normalizado del cliente construido con información externa.
- EXT SOURCE 1: Score 3 Normalizado del cliente construido con información externa.
- REGION RATING CLIENT: Calificación interna del cliente.
- bd CREDITS ACTIVE sum: Suma de los créditos activos del cliente en Buró al momento de la solicitud.
- pre app APP REFUSED sum: Cuantas solicitudes previas a la actual le fueron rechazadas al cliente.
- bd CREDITS CLOSED sum: Suma de los créditos cerrados del cliente en Buró al momento de la solicitud.
- REG CITY NOT WORK CITY: Indica si la dirección del trabajo del cliente es la proporcionada en el contrato.
- SUM FLAGS DOCUMENTS: Cuantos documentos no obligatorios entregó el cliente al momento de la solicitud.
- REG CITY NOT LIVE CITY: Indica si la dirección del cliente es la proporcionada en el contrato.

5 Métricas de desempeño

ENTRENAMIENTO: ACCURACY: 0.877, ROC AUC: 0.721, GINI: 0.442, KS:0.32

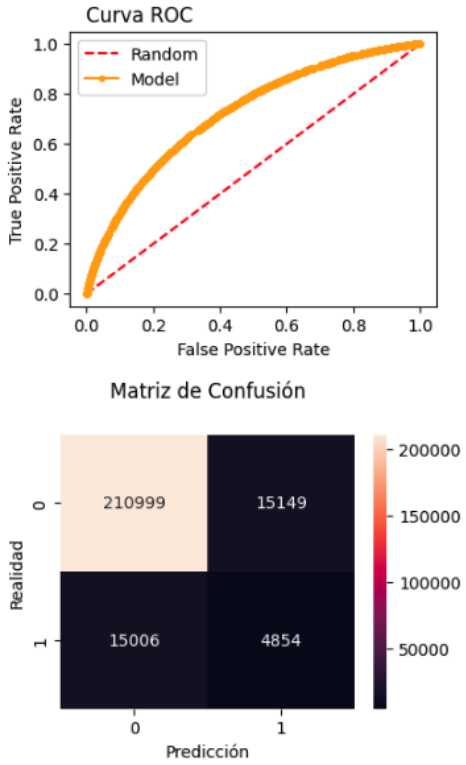

Figure 3: *Curva ROC y matriz de confusión en conjunto de entrenamiento.*

VALIDACIÓN: ACCURACY: 0.878, ROC AUC: 0.729, GINI: 0.457, KS:0.34

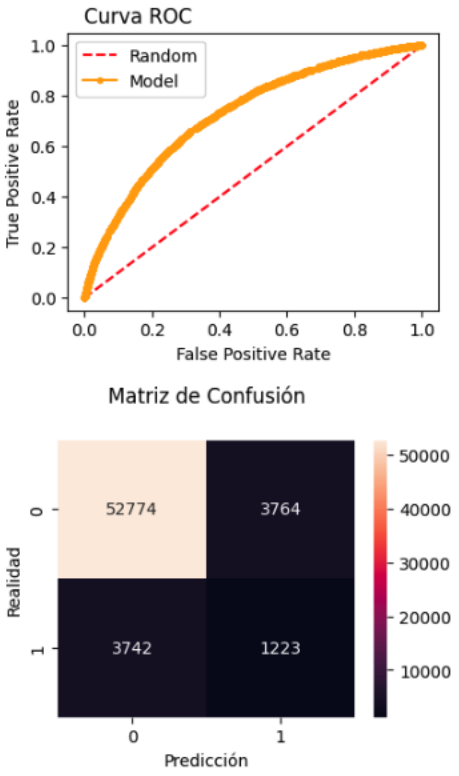

Figure 4: *Curva ROC y matriz de confusión en conjunto de validación.*

ENTRENAMIENTO:

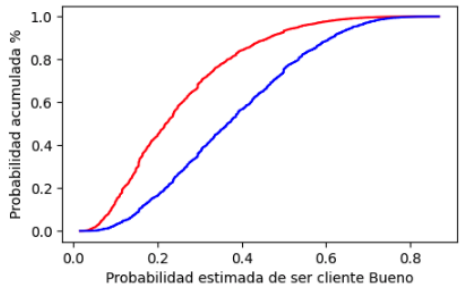

Figure 5: *Distribución acumulada de clientes buenos y malos.* Estadístico KS=0.32

VALIDACIÓN:

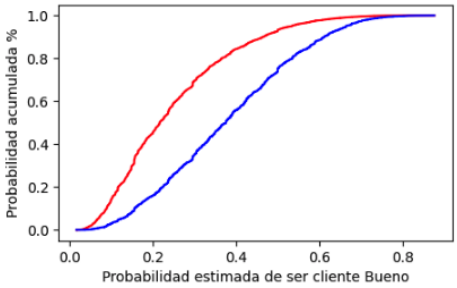

Figure 6: *Distribución acumulada clientes buenos y malos* Estadístico KS=0.34

6 Especificaciones Score

- Puntos base: 600
- Factor de escalamiento: 50
- Tabla de rangos y tasa de malos:

	Rango	Solicitudes	Ctes malos	Bad Rate
0	[506, 608)	6298	1438	0.228326
1	[608, 628)	6323	947	0.149771
2	[628, 643)	6378	695	0.108968
3	[643, 655)	6235	483	0.077466
4	[655, 666)	6267	413	0.065901
5	[666, 679)	5981	300	0.050159
6	[679, 689)	6267	254	0.040530
7	[689, 703)	5774	188	0.032560
8	[703, 720)	6414	164	0.025569

Figure 7: *Tabla de Rangos construidos en base a deciles del score.*

- Punto de corte: 655.

7 Vista aplicación

Figure 8: *Vista aplicación productiva.*

Las predicciones se calculan en lotes cargando la información en .csv y como resultado se obtienen las variables ingresadas, el score y el resultado.

Si el score> 655 entonces la solicitud es aceptada, en otro caso, rechazada.

Finalmente, se muestra un resumen con el número de solicitudes aceptadas, rechazadas, score promedio, tasa de aceptación y rechazo.

8 Comentarios finales

Debido a que el conjunto de datos está limitado solo a conjunto de entrenamiento y validación, no se pudieron realizar pruebas de *out of time* e inferencia de rechazados, no obstante debe tenerse en cuenta que ambos análisis son de gran utilidad en la construcción de Scorecards.

Respecto a la extracción de la información, se tendrá que construir un pipeline independiente para la ingesta de la información necesaria para el modelo y para productivizar la aplicación en línea, podemos usar Docker.

Contacto: angujoma@gmail.com