Modelo disruptivo de crédito

Juan Francisco Mandujano Reyes*, Leobardo Omar Jasso Romo

Angel Balderas Paredes, Salim Vargas Hernandez,

Mariana Menchero García

María del Carmen Ruíz Esquerra

September 8, 2019

1 Planteamiento del Problema

1.1 Problema

El crédito al consumo es uno de los productos preferidos por los clientes bancarios. Debido a esto, para las instituciones bancarias de la talla de BBVA, es muy importante brindar un trato personalizado tomando en cuenta las necesidades y contexto de los consumidores para determinar la colocación de crédito al consumo. Sin embargo, actualmente se realiza una estimación basada en variables macroeconómicas empleando un modelo genérico, dejando de lado la personalización de los procesos. Es por esto que nos interesa buscar fuentes externas de información para determinar cuales variables ayudan a predecir la contratación de crédito al consumo.

1.2 Hipótesis

Existen variables, típicamente no incluidas en los modelos usuales, que ayudan a predecir la colocación diaria de créditos al consumo (monto solicitado).

1.3 Preguntas Rectoras

¿Qué variables son potenciales indicadoras en la contratación de crédito al consumo? ¿Cómo podemos utilizar estas variables para realizar la creación de un

*Email address: juan.mandujano@cimat.mx

modelo estadístico o de aprendizaje máquina que nos permita predecir la colocación diaria de créditos al consumo?

2 Solución Propuesta

2.1 Objetivo

Desarrollar un modelo utilizando variables no convencionales que nos permita predecir la colocación diaria de créditos al consumo en BBVA (monto solicitado).

2.2 Métricas

Para la evaluación del modelo de pronóstico se optó por utilizar Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coeficiente de Determinación ajustado (EVS).

- **RMSE**
  Es la desviación estándar de los residuos (errores de predicción). Los residuos son una medida de qué tan lejos están los puntos de datos de la línea de regresión y RMSE es una medida de la dispersión de estos residuos.

  $$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}.$$

- **MAE**
  Mide la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Es el promedio sobre la muestra de prueba de las diferencias absolutas entre la predicción y la observación real donde todas las diferencias individuales tienen el mismo peso.

  $$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|.$$

- **EVS**

  Es una medida corregida de la bondad de ajuste (precisión del modelo) para modelos lineales. R^2 siempre aumenta a medida que se incluyen más variables explicativas en el modelo. R^2 ajustado intenta corregir esta sobreestimación. El R^2 ajustado puede disminuir si una variable específica no mejora el modelo.

  $$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1},$$

  donde n es el número de datos y k es el número de variables explicativas independientes.

2.3 Metodología

2.3.1 Obtención de Datos

Además utilizamos Google Trends para identificar el número de búsquedas de productos que se pueden englobar en una misma categoría de consumo y para los que usualmente se emplea crédito en la compra. Entre las categorías que se seleccionaron se encuentran electrodomésticos de larga duración, artículos tecnológicos y de entretenimiento. Asimismo, consideramos el número de búsquedas con temas relacionados al crédito y préstamos personales para cubrir el efecto que tiene la búsqueda de opciones de crédito.

Otra categoría importante, fue la de compras en línea, un fenómeno relativamente nuevo pero que ha ido cobrando relevancia.

Todas las variables anteriores además de no ser variables financieras son gratuitas y fáciles de obtener. Las muestras de hasta 3 meses se pueden obtener con precisión diaria y desagregadas a nivel estatal, lo cual facilita la adopción a la práctica del negocio y podría permitir en el futuro análisis más precisos.

Ligado a lo anterior, también se hizo web scrapping para twitter con el objetivo de abarcar el efecto que tienen ciertos momentos en la vida de los clientes como asaltos y robos. Tambien se realizo un análisis de sentimientos en los twitts con la intención de conocer la percepción de las personas sobre la situación social de su entorno. Lo anterior dado que existe literatura que relaciona la percepción del ambiente social con el crecimiento economico y el crédito.

Por último, consideramos que el efecto del calendario repercute fuertemente en la contratación de créditos por lo que creamos una variable dummie para el calendario de eventos que promueven el consumo.

Se descartaron las categorias de ropa y calzado, viajes, inicio de clases ya que incluían mucho ruido y no beneficiaban el poder predictivo del modelo.

Sabemos que el momento es un factor importante en la colocación de créditos por lo que también consideramos el número de nacimientos diarios, esta variable nos permite conocer un poco más sobre las necesidades del cliente y del momento que está viviendo, pero no la agregamos por la dificultad de obtener datos recientes diarios. Esto en un futuro, o para distintos mercados, podría no representar una dificultad.

2.3.2 Descripción del Modelo

A continuación describimos los detalles del plan de trabajo a seguir:

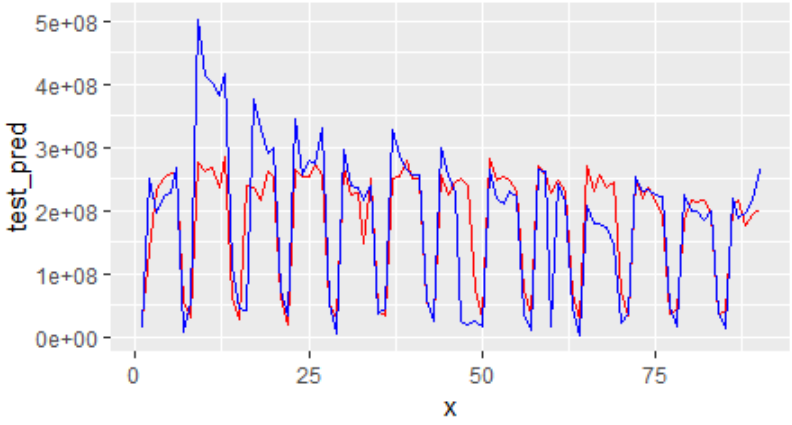
Decidimos usar el modelo de *xgboost* (eXtreme Gradient Boosting) cuyo propósito es la predicción. Para este problema usamos la variante de modelo lineal, desarrollando el código en R. Este tipo de algoritmos constituyen parte del estado del arte, en aprendizaje máquina, para la predicción de series de tiempo.

Los resultados obtenidos, considerando las variables predictoras antes mencionadas, se muestran a continuación:

2.3.3 Resultados

Notemos que usamos como conjunto de entrenamiento los datos desde el día 28/04/2017 al 28/02/2019. Constituyendo 674 registros. El conjunto de prueba va del 01/03/2019 al 31/05/2019, 90 días a predecir.

Hasta ahora los resultados lucen así:



3 Discusión y Conclusiones

Podemos notar que la elección de las variables no financieras fue satisfactoria. Obtuvimos un RMSE, MAE y EVS suficientemente bajos para considerar que el modelo está prediciendo correctamente 90 días a futuro. Las gráficas anexas confirman esta afirmación.

Los datos indican que existe una eventualidad en la segunda semana de marzo de 2019 que el modelo no pudo predecir. La demanda de creditos es multifactorial, por lo cual, consideramos que este evento está asociado a una variable que podría ser considerada en el futuro. Como posible solución consideramos las promociones de crédito "Días BBVA", antes llamadas "Días Bancomer".

4 Bibliografía

- Bishop, Christopher M. (2006). Pattern recognition and machine learning. New York :Springer.
- Christian Bjørnskov. (2012) How Does Social Trust Affect Economic Growth? Southern Economic Journal.
- Datos tomados de:
  - Google Trends
  - INEGI