REPORTE: DESAFÍO BCI - 2021

William Berrios
Departamento Ingeniería Mecánica
Universidad Nacional de Ingeniería
Lima, Peru
wberriosr@uni.pe

1 Objetivo

El desafío propuesto por el banco BCI busca estimar las ventas de empresas con el objetivo de ayudar en la toma de decisiones crediticias y generar oportunidades de negocio. El reporte presentado corresponde a la solución que ocupo el primer lugar en el aspecto cuantitativo y cualitativo en la competencia.

2 Problemática

Actualmente, el banco BCI cuenta posee problemas para estimar las ventas de los clientes del segmento empresa, esto debido a diversas razones como por ejemplo:

- La disposición de esta información es generalmente un proceso costoso, manual, y que toma tiempo, hoy crítico en la dinámica del negocio
- Algunos clientes, envían sus ventas totales al banco. Sin embargo, esta información está en la mayoría de los casos descontinuada.
- La data de ventas de los clientes empresa se tiene en muchos casos solo de manera parcial, ya que los clientes suelen operar con más de un banco.

3 Metodología

La metodología seguida para resolver el reto se expresa en la Fig1

Figure 1: Metodología utilizada en el reto BCI

3.1 Análisis Exploratorio

En esta sección se realiza la inspección de las variables categóricas, numéricas y la variable objetivo. El código asociado a esta sección se denomina como EDA

3.1.1 Variable Objetivo

La variable objetivo en estudio es el nivel de ventas de los clientes del segmento empresas del banco BCI. Durante el análisis de la variable objetivo se observa que esta tiene un valor medio mucho menor en los meses de Abril y Septiembre del año 2020. Esto se presume que es debido a los efectos del COVID-19 lo cual genero una reducción considerable en el nivel de ventas de las empresas.

Además se puede observar que durante los meses de diciembre, el nivel medio de las ventas aumenta por encima del promedio normal

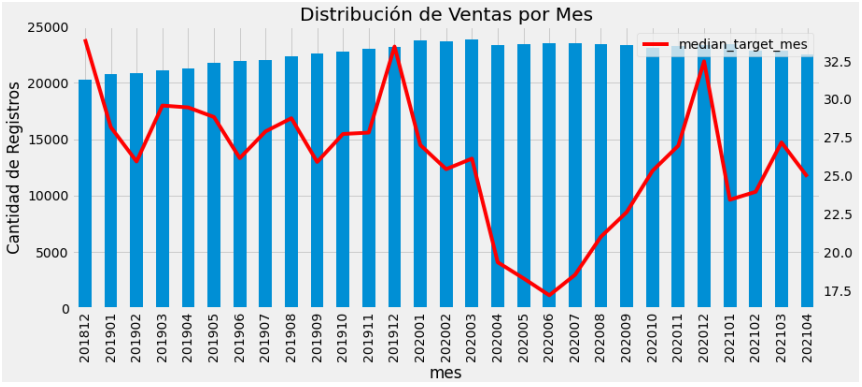


Figure 2: Distribución de la mediana de ventas a lo largo del horizonte de estudio

Adicionalmente se puede observar que un porcentaje considerable de empresas tienen valor 0 correspondiente a sus ventas. Esto se debe posiblemente a las causas mencionadas en la Sección 2

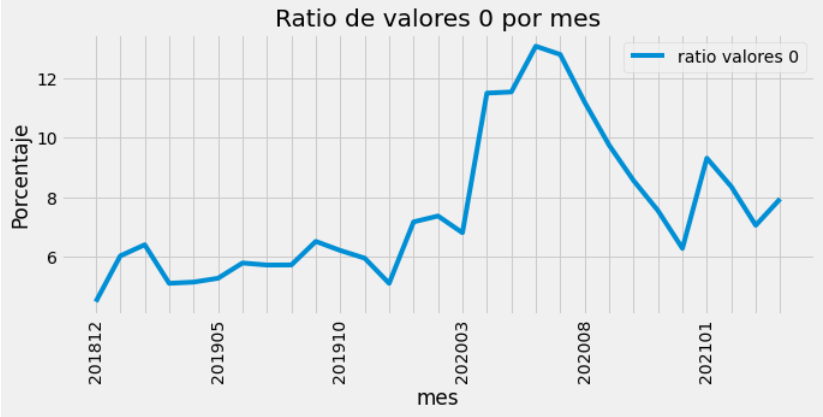


Figure 3: Distribución del ratio de zeros en el horizonte de estudio

Finalmente un aspecto útil durante la competencia fue el realizar la transformación del la variable objetivo. La transformación mas efectiva fue el de extraer la raíz cuadrada logrando suavizar la curva de distribución de la variable objetivo. Por otro lado la transformación logarítmica también fue utilizada para el ensamblado.

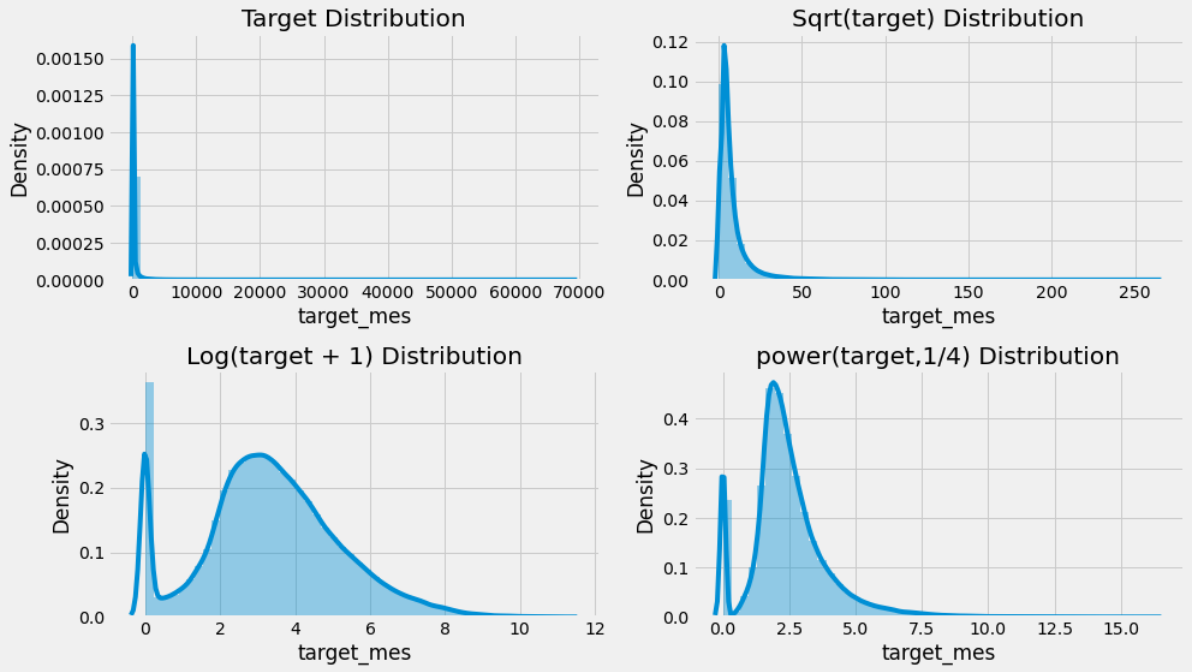


Figure 4: Transformaciones aplicadas a la variable objetivo

3.1.2 Variables Independientes

En esta sección se observa la distribución de ciertas variables que potencialmente pueden ser importantes en la realización de los modelos predictivos.

- Tipo Cli: Se observa que esta variable posee 2 categorías y logra diferenciar empresas con ventas altas y bajas.

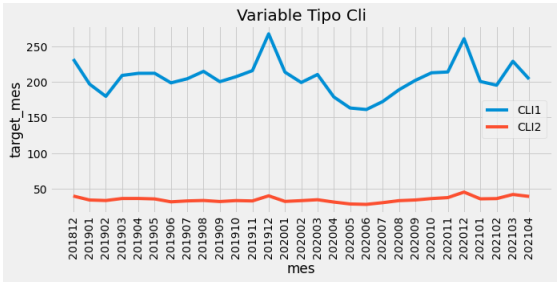


Figure 5: Relación entre la variable tipo cli y la variable dependiente

- Tipo Ban: Esta variable posee 4 categorías las cuales logran diferencias el valor promedio de las ventas.

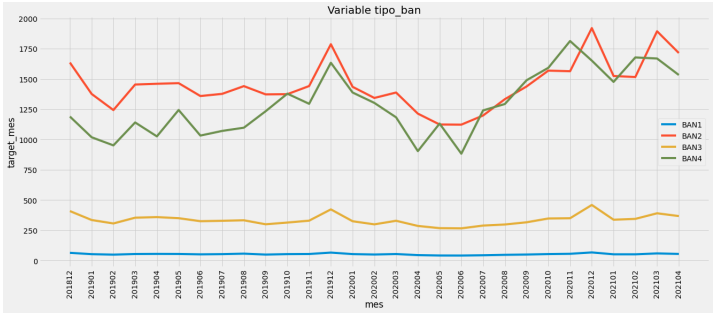


Figure 6: Relación entre la variable tipo ban y la variable dependiente

- Tipo Com: Esta variable posee 5 categorías y presentan una relación interesante con la variable objetivo. Se observa que las categorías COM1, COM2, COM3 posee ventas medias mucho menores que las categorías COM4 y COM5.

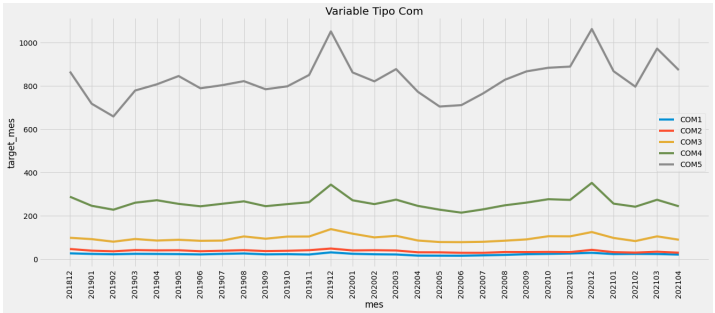


Figure 7: Relación entre la variable tipo com y la variable dependiente

Se observa que a diferencia de las variables numéricas, las variables categóricas poseen propiedades que segmentan muy bien la variable ventas en segmentos de alto, medio y bajo nivel promedio.

3.2 Tratamiento de los Datos

El tratamiento de los datos consistió principalmente en la imputación de valores nulos y transformación de las variables categóricas.

- Imputación de Datos: Se evaluaron distintas técnicas de imputación obteniendo como mejor resultado, la imputación con ceros a las variables numéricas.
- Transformación Variables Categóricas: Adicionalmente a las variables categóricas se les realiza el proceso de label encoding, es decir convertirlas en valores numéricos a partir de 0. Esto con el objetivo de que puedan ser utilizados directamente en los modelos de gradient boosting.

3.3 Esquema de Validación

Durante la competencia se observaron aspectos importantes en la distribución del test, tal como se puede apreciar en la Fig8:

- Durante los periodos entre 201812 y 202104, los clientes que aparecen en la data de test no poseen ningún registro en la data de entrenamiento. Este grupo de clientes representan el 20 % de todos los clientes presentes en cada mes de observación.
- Durante los periodos entre 202105 y 202109, existe un porcentaje de clientes que poseen algún registro de ventas entre los periodos 201812 y 202104. Estos representa un 65% de clientes con respecto al total entre los periodos 202105 y 202109 en la muestra de test.
- Finalmente, el 35% de clientes entre los periodos 202105 y 202109 y que aparecen en la muestra de test no poseen ningún registro de ventas en los meses previos.

Figure 8: Distribución mensual de las muestras de entrenamiento y test

Por las razones previamente mencionados, se decidió adoptar 2 esquemas de validación:

- En el primero, al que denominaremos clientes sin historia, se realiza una validación con 5 folds en los cuales se distribuyen los clientes mediante el método "Group-kfold" perteneciente a la libreria scikit-learn. Este método permite entrenar con registros que pertencen a los mismos clientes y validar con clientes que nunca se han visto en la muestra de entrenamiento. Claramente esto refleja la distribución del test durante los periodos 201812-202104 y los clientes sin historia en los periodos 202105-202109
- El segundo esquema representa a los clientes de los cuales ya se registra alguna historia con respecto a sus ventas en meses anteriores. Es por ello que para este grupo se realiza una validación con 5 folds en los cuales se utiliza el método "StatifiedKfold". Esto permite reflejar la distribución del 65% de clientes que pertenecen a la data de test durante los periodos 202105 y 202109

Esto se puede apreciar en la distribución de clientes combinando la data de entrenamiento y test durante los periodos 202001 - 202104. Ver Fig9

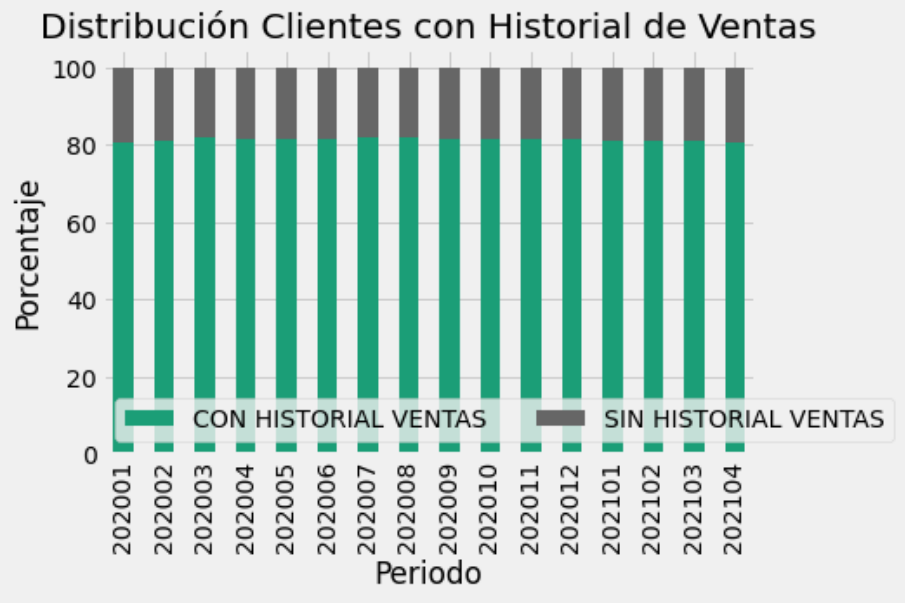


Figure 9: Distribución del total de clientes con respecto a su historial previo de ventas

El código asociado con esta sección se encuentra en el notebook $02.GenerateDataset$.

3.4 Feature Engineering

En esta sección se generaron diversas variables tanto categóricas como numéricas. Para ello se concatenaron los datasets de entrenamiento y test. Además para los clientes con historial de ventas, se generaron agregaciones de la variable dependiente. Este proceso se resumen en la siguiente tabla:

Durante el desarrollo de la competencia se utilizó 2 sets de variables. El primero es generado utilizando el archivo $generate\_feature\_v1.py$ y el segundo utilizando el archivo $generate\_features\_v2.py$

Table 1: Ingeniería de variables

Variables	Descripción
Numéricas	- Agregaciones (Media, Máximo, Desviación Estándar) de las variables numéricas en los últimos N meses - Variación de las variables numéricas en los últimos N meses - Media de las variables numéricas por cada categoría - Media de las variables numéricas por cada categoría en cada mes - Flag Covid durante los periodos 202004-202009
Categóricas	- Combinación de variables categóricas hasta grado 4 como máximo - Cantidad de valores únicos de cada variable categórica en los últimos N meses
Variable Objetivo	- Agregaciones (Media, Máximo,cantidad de valores 0, kurtosis, skewness) de la variable ventas en los últimos N meses hasta T_{Obs-5} - Pendiente de la recta que estima las ventas en los últimos N meses hasta T_{Obs-5}

3.5 Modelamiento

La solución planteada consta de 3 componentes principales:

- Modelo en el Pasado: Este modelo busca estimar las ventas para la ventana de tiempo entre los periodos 201812 - 202104. Se utiliza el esquema de validación "Group-Kfold" y se añade la variable mes como categoría para poder capturar los efectos de las variaciones en la media de las ventas mensuales.
- Modelo de Comportamiento: Este modelo busca estimar las ventas de clientes que ya cuentan con algún registro de ventas en el pasado y que se encuentran presentes entre los periodos 202105-202109. Se utiliza el esquema de validación "StratifiedKfold" ya que se busca predecir sobre clientes de los cuales se tiene historia. Cabe resaltar que este modelo utiliza registros de entrenamiento desde 202004, esto para poder capturar comportamientos mas recientes a causa del Covid-19
- Modelado Nuevos Clientes : Este modelo busca estimar las ventas de clientes nuevos que aparecen entre los periodos 202105-202109. Cabe resaltar que este modelo utiliza registros de entrenamiento desde 202004, esto para poder capturar comportamientos mas recientes a causa del Covid-19

A continuación se menciona los archivos de entrenamiento de los modelos de Machine Learning utilizados para cada componente de la solución.

Table 2: Modelos de Machine Learning Entrenados

Nombre	Archivo
Modelo en el Pasado	- 04_1 Model 1 - New Clients Model-Boosting-Target-sqrt.ipynb • Lightgbm - Target raiz cuadrada • Catboost - Target raiz cuadrada - 04_1 Model 1 - New Clients Model-Boosting-Target-log.ipynb • Lightgbm - Target logaritmo
Modelo de Comportamiento	- 04_2 Model 2 - Behavior Model.ipynb • Lightgbm - Target raiz cuadrada
Modelo Nuevos Clientes	- 04_3 Model 3 - Future Model-Boosting-Target-sqrt.ipynb • Lightgbm - Target raiz cuadrada • Catboost - Target raiz cuadrada

3.5.1 Ensamblaje

Como último paso, se realizó el ensamblaje de los mejores modelos desarrollados a lo largo de la competencia. El ensamblaje final se realizo de la siguiente manera:

- Modelo en el Pasado: Para ensamblar las predicciones de estos modelos se utilizó la media geométrica, esto para asegurar de que si existe un valor 0 o muy cercano a 0, la predicción del ensamblaje también sea muy pequeña.
- Modelo de Comportamiento: En este caso solo se obtuvo un modelo así que no se realizó ensamblaje.
- Modelo Nuevos Clientes: Al igual que en el caso de los modelos en el pasado, se utilizó la media geométrica para el ensamblaje.

El código asociado a esta sección se encuentra en el notebook 03.Final_Ensembling.ipynb

4 Impacto en el Negocio

4.1 Efectividad del Modelo por Rango de Ventas

Para determinar las métricas cualitativas del modelo se analizó el último año de información en la data de entrenamiento (202005 - 202104). Para ello se define los rangos de sobre y sub-estimación con un +/- 30% de error acorde a los criterios del Banco BCI. Además la definición de los rangos de ventas se establece de acuerdo. Ministerio de Economía de Chile.

A partir del gráfico se puede concluir:

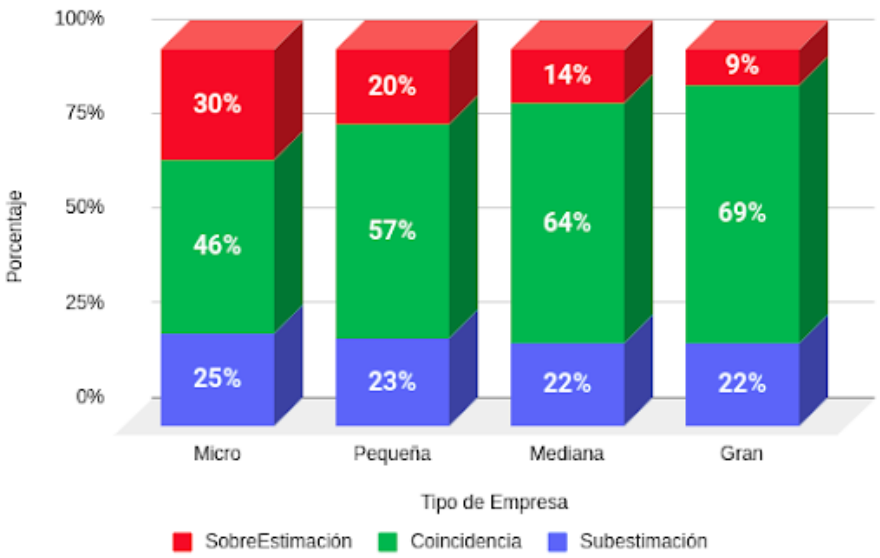


Figure 10: Medición cualitativa de la solución

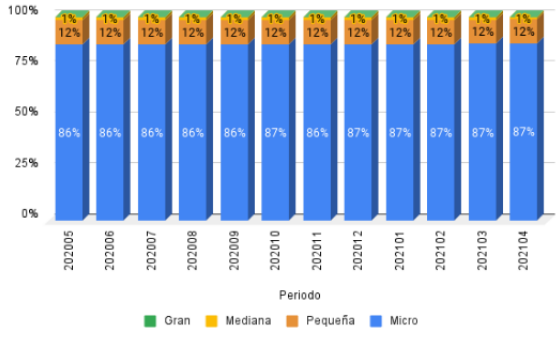
- La coincidencia supera el 45 en todos los rangos, lo que asegura una solución efectiva
- La sobre-estimación no supera el 30%
- El segmento que posee una linea de crédito mayor (Gran Empresa) se predice con mayor exactitud (Coincidencia 69%)

4.2 Distribución del Portafolio

La distribución del portafolio en los últimos 12 Meses se muestra en los siguientes gráficos.

Tipo Empresa	Venta Promedio Anual (uF)	Distribución
Micro	452	86.5%
Pequeña	6913	12%
Mediana	44251	1.3%
Gran	202335	0.2%

(a) Distribución promedio de Ventas



(b) Distribución del portafolio mensualmente

Figure 11: Distribución del Portafolio durante los periodos 202005-202104

Se observa que el segmento Micro Empresa contribuye con el 86% al portafolio del banco BCI. Además los segmentos Mediana y Gran empresa son aquellos que poseen un mayor promedio anual de ventas.

5 Recomendaciones

Debido al tiempo límite del desafío y la anonimización de variables, se debe mencionar que es posible mejorar el performance y la interpretabilidad de los modelos desarrollados para una futura implementación. Las recomendaciones son las siguientes:

- Revisar las variables creadas de forma automática y definir si cuentan con sentido de negocio.
- Realizar un proceso de backward selection para poder obtener las mejores variables. Esto debido a que por la duración de la competencia, se crearon variables y se seleccionaron solamente de acuerdo a la importancia de los modelos de boosting. Sin embargo existe la posibilidad que variables aparezcan importantes pero que al quitarlas, el perfomance del modelo aumente.
- Una vez teniendo las mejores variables, seleccionar una cantidad adecuada para asegurar la interpretabilidad que el negocio requiere y facilitar la implementación.
- Implementar el mejor modelo para cada segmento (Con y Sin Historia). Esto debido a que si bien es cierto que el ensamblaje de modelos es bastante beneficioso durante una competencia. En un entorno real, esto puede complicar la implementación y solo generar una ganancia mínima en la métrica a evaluar.