Universidad Nacional de Ingeniería

Facultad de Ingeniería Económica, Estadística y Ciencias Sociales

TRABAJO DE SUFICIENCIA PROFESIONAL

DISEÑO DEL PROCESO DE VALIDACIÓN DE UN MODELO PREDICTIVO CON METODOLOGÍA XGBOOST PARA PREDICCIÓN DE INGRESOS DE PERSONAS NATURALES CON INFORMACIÓN LIMITADA EN EL SISTEMA FINANCIERO

Para obtener el grado de Ingeniero Estadístico.

Elaborado por

Isbella Merici Miranda Vásquez

0009-0004-9665-2544

Asesor

Dr. Luis Huamanchumo de la Cuba

0000-0002-2239-5301

LIMA – PERÚ

2022

Citar/How to cite	(Miranda, 2022)
Referencia/Reference Estilo/Style: APA (7ma ed.)	Miranda, I. (2022). *Diseño del proceso de validación de un modelo predictivo con metodología XGBoost para predicción de ingresos de personas naturales con información limitada en el sistema financiero*. [Trabajo de Suficiencia Profesional de pregrado, Universidad Nacional de Ingeniería]. Repositorio institucional Cybertesis UNI.

Dedicatoria

Con mucho amor a mi hijo Sam Winston por acompañarme durante toda la realización de este informe. A Reynaldo Miranda y Carmen Vásquez, por ser unos padres extraordinarios

Resumen

El presente informe se fundamenta en la necesidad de definir un proceso de validación robusta de modelos estadísticos que utilizan la técnica de aprendizaje XGBoost. Para este fin se presenta la propuesta del diseño de validación considerando una aplicación de la técnica en la predicción de ingresos de personas naturales que tienen información limitada en el sistema financiero. El objetivo de la validación de modelos es brindar la seguridad que todas las fases de construcción del modelo se han desarrollado de forma óptima considerando los criterios estadísticos necesarios para obtener un performance del modelo aceptable tal que el modelo pueda ser implementado en la gestión del negocio.

A lo largo del documento se presentará y revisará el esquema completo de validación, que incluye desde la definición de la metodología, análisis y extracción de datos, proceso de modelado hasta la revisión de la completitud de la documentación que facilitará la implementación del modelo para la puesta en producción. Cumplir con los requisitos mínimos de la validación es necesario para dar luz verde a la salida a producción del modelo. Finalmente, se podrá observar cómo contribuye este nuevo diseño de validación en el desempeño del modelo final y la correcta integración a la gestión.

Palabras claves: Machine learning, xgboost, aprendizaje supervisado, validación.

Abstract

This report is based on the needed to define a robust validation process for statistical models using the XGBoost learning technique. For this purpose, the validation design proposal is presented considering an application of the technique in the prediction of income of individuals who have limited information in the financial system. The objective of model validation is to provide assurance that all phases of model construction have been optimally developed considering the necessary statistical criteria to obtain acceptable model performance such that the model can be implemented in business management. Throughout the document, the complete validation scheme will be presented and reviewed, which includes everything from the definition of the methodology, data analysis and extraction, the modeling process to the review of the completeness of the documentation that will facilitate the implementation of the model for the put into production. Meeting the minimum validation requirements is necessary to give the green light for the model to go into production. Finally, is possible to observe how this new validation design contributes to the performance of the final model and the correct integration to management.

Keywords: Machine learning, xgboost, aprendizaje supervisado, validación.

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

INTRODUCCIÓN

Es conocido que el sistema financiero implementa diversos procesos para otorgar créditos, donde todos tienen como mínimo reglas de negocio hasta modelos estadísticos más elaborados, claramente mientras más robusta sea la solución implementada existirá una mejor gestión del riesgo de crédito.

Por ejemplo, la entidad financiera sobre la cual enfocaremos este estudio y denominaremos en lo sucesivo Banco, posee un portafolio de más de 100 mil clientes, los cuales han pasado por una etapa de originación dónde los modelos/procesos del banco han dado resultados aprobatorios para que estos puedan ser elegidos clientes. Si en alguna parte del proceso no se han colocado los suficientes controles para asegurar la legitimidad de los resultados estaríamos contaminando la cartera con clientes “malos”.

Si bien se definen un conjunto de controles durante el proceso de evaluación un aspecto importante que muchas veces pasa desapercibido es la importancia de validar la correcta implementación de las lógicas de admisión antes de su integración a la gestión. Estas lógicas involucran reglas de negocio definidas según el comportamiento del cliente hasta modelos estadísticos avanzados. Implementar un modelo estadístico sin considerar que a lo largo de su desarrollo pueden haberse omitido ciertas consideraciones que aseguren el resultado de su desempeño sumaria el error aleatorio intrínseco que tiene el modelo con un error sistemático que hubiese sido posible identificar preliminarmente. Es por ello en muchas empresas existe un equipo de Validación Interna que se encarga de revisar de inicio a fin cualquier iniciativa a implementar, generalmente esta validación sigue un esquema autoaprendido de acuerdo con los requerimientos de negocio y con un alcance acotado.

El avance de las técnicas estadísticas hace insuficiente que el esquema de validación auto aprendido asegure una correcta implementación del modelo. Por tanto, se identifica la necesidad de proponer un proceso de validación acorde a la necesidad, en el presente estudio se expone el caso de la integración a la gestión de un modelo de estimación de ingresos con metodología XGBoost y medir su efectividad en relación con la contribución del desempeño del modelo. Cabe resaltar que el contexto en el cual se desarrolla este modelo es pre-pandemia del Coronavirus y es recomendable su calibración dado que debido al confinamiento y que las empresas han adaptado sus trabajos a una forma remota no todos los empleos han sido adaptados al modo virtual sufriendo una variación importante la distribución de ingresos.

CAPITULO I

ANTECEDENTES

1.1 Datos de la empresa o institución

La institución donde se desarrolla la actividad es una entidad financiera del Perú con una trayectoria como Banco desde hace 81 años. El objeto social es el de favorecer el desarrollo de las actividades comerciales y productivas del país, que incluyen la actividad bancaria comercial y de ahorros.

Su misión es ser un líder financiero sostenible en Latinoamérica, guiado por un gran propósito, orientado al futuro y enfocado en crear valor superior para los colaboradores, clientes, accionistas.

1.2 Área donde se desarrolló la actividad

La actividad fue desarrollada dentro de un entorno ágil Scrum, esto es que se habilitó una mesa de trabajo multidisciplinaria donde intervenían diferentes equipos con el objetivo de entregar el producto, el modelo de estimación de riesgos, implementado. El fin de la habilitación de la mesa ágil fue entregar resultados cada sprint (2 semanas) resolviendo problemas, dudas y consultas dado la disponibilidad a tiempo completo del equipo, la mesa se encontraba dirigida por un denominado Product Owner perteneciente al área de Riesgos Personas y principal interesado que el producto en desarrollo cumpla con los criterios de aceptación definidos por todos los equipos. Cabe indicar que además del uso por primera vez de la metodología ML con técnica XGBoost también fue la primera vez que se habilitó esta metodología de trabajo ágil en el Banco.

1.3 Composición de la mesa de trabajo ágil

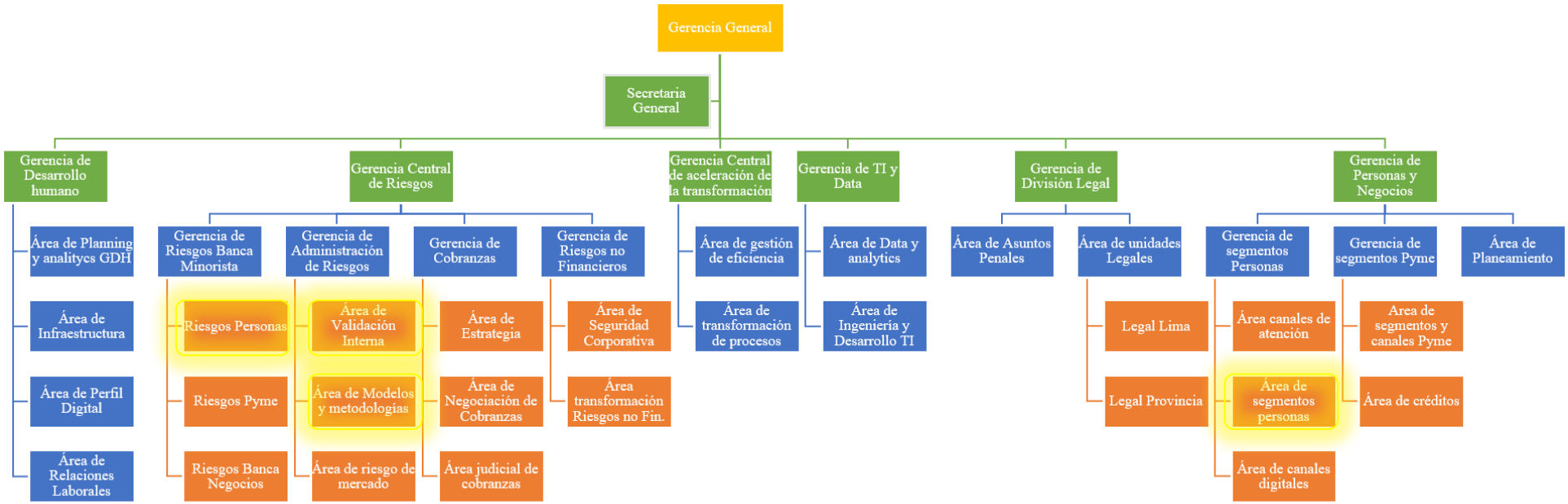
De acuerdo a lo especificado en 1.2 la mesa de trabajo ágil se compuso de la siguiente manera:

- 2 miembro del área RBM – Producto Owner (líder) y Analista
- 1 miembro del área de Modelos – Data scientist (estadístico)
- 1 miembro del área de Validación Interna – Data scientist (estadístico)
- 1 miembro del área de Data – Data Engineer

La función dentro de la organización de quien escribe fue representar al equipo de Validación Interna, quien por primera vez se encontraba participando de una mesa de trabajo ágil, y en una validación de un modelo con una metodología Machine Learning diferente a las antes desarrolladas en la institución.

Figura 1

Organigrama de la Entidad financiera



Nota: Se ha resaltado los equipos que interactuaron con el desarrollo de la actividad

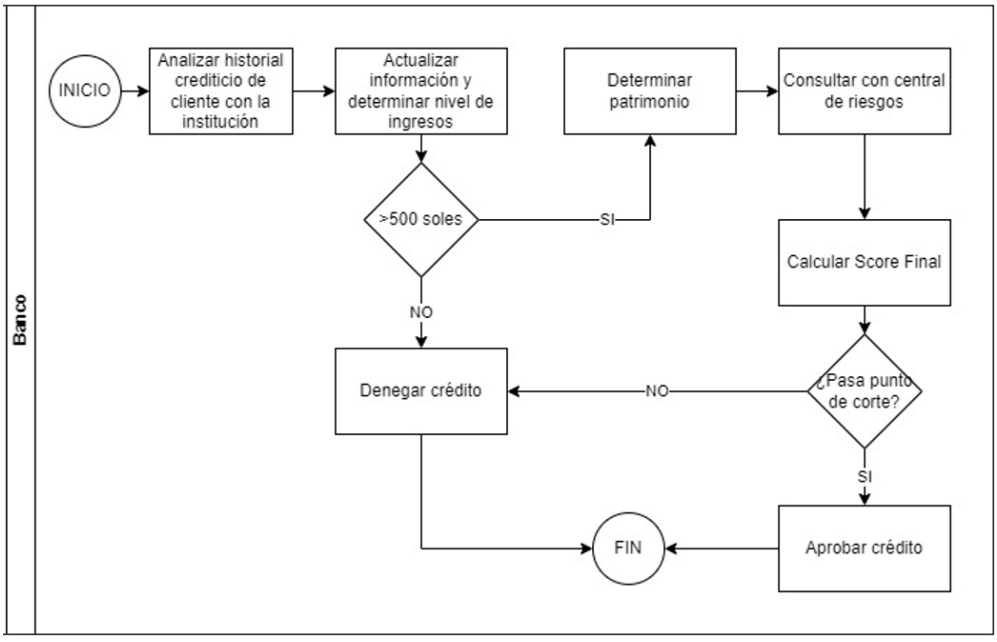
1.4 Proceso de admisión de clientes

El solicitante del producto puede ser un cliente del banco (cuenta con información e historial crediticio) o un no cliente (proveniente de otra entidad financiera o cliente nuevo en el sistema financiero). El proceso de admisión de clientes depende en primer lugar del perfil del cliente y del o los productos que éste desee adquirir en el Banco.

El perfil del cliente está basado principalmente en el nivel de ingresos y su nivel de deuda que pueda tener, el sector o rubro económico a que se dedica y su comportamiento en el sistema financiero. Y sea cliente o no cliente tiene la posibilidad de aplicar a la obtención de una tarjeta de crédito, un crédito efectivo, crédito vehicular, crédito hipotecario o crédito de estudios como facilidad financiera.

Figura 2

Proceso de evaluación de un cliente



Nota. La figura muestra el proceso de admisión de un cliente desde el punto de vista de la entidad financiera.

1.5 Gestión de Riesgos de crédito

En el contexto de este estudio indicaremos las áreas más relevantes que intervienen durante la integración a la gestión de un modelo estadístico utilizado para la evaluación de créditos del cliente. La estructura organizacional de la entidad financiera tiene las siguientes áreas:

Área de Segmentos personas: En adelante área de Productos, el objetivo de este equipo es la colocación de productos financieros. Para ello propone condiciones atractivas para su adquisición, evalúa paquetes alineados a sus necesidades con beneficios adicionales, ofrece campañas entre otras propuestas de negocio. Dado que la principal métrica de cumplimiento de este equipo es el nivel de ventas también son los principales decisores en las tasas y comisiones a cobrar a los clientes. Cabe resaltar que cada iniciativa realizada debe ser revisada y complementada por el área de Riesgos, cuyas principales funciones revisaremos a continuación. En el proceso de adquisición son los primeros en recibir la alerta por parte de la división comercial si alguna de las herramientas implementadas presenta sospecha de inadecuada construcción y/o implementación.

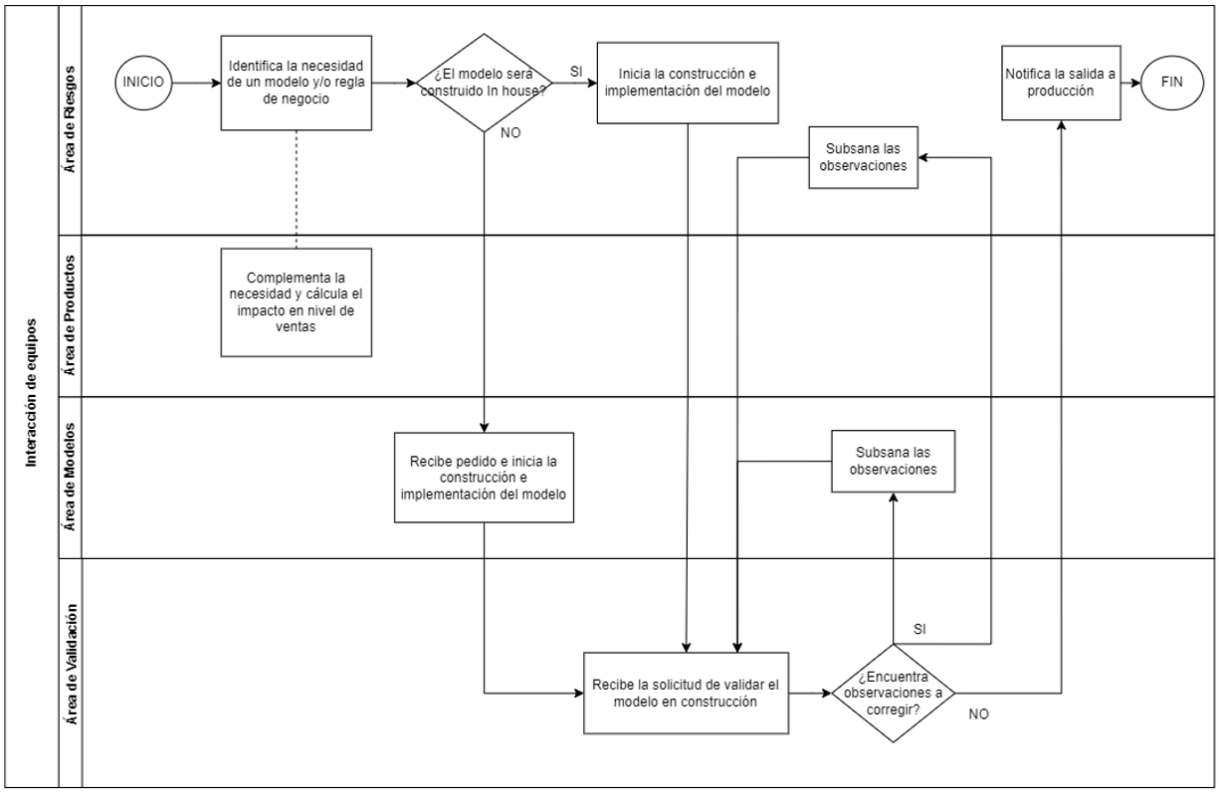
Área de Riesgos Personas: En adelante, área de Riesgos, esta área es la encargada de preservar la calidad de la cartera y de los clientes que son admitidos por tal proponen un conjunto de indicadores y/o reglas que permitan filtrar a clientes con mal comportamiento o con una alta probabilidad de incumplimiento. Asimismo, son responsables que el flujo crediticio implementado tenga los controles adecuados y suficientes en sus herramientas para asegurar el portafolio con clientes sanos siendo así la primera capa de **validación de implementación**. Recibe la alerta/consulta de parte de productos ante sospecha de inadecuada construcción y/o implementación.

Área de Modelos y metodologías: En adelante, área de Modelos, es el área encargada de la construcción de modelos estadísticos elaborados, es decir, para el proceso crediticio es posible crear reglas de negocio dentro del equipo de riesgos, pero si por la naturaleza se verifica que este modelo es un decisor clave en la gestión se decide convocar al equipo de modelos, éste último propondrá la planificación de desarrollo y tendrá reuniones ad hoc con los stakeholders (Productos, Riesgos y Validación). Además, en este caso este equipo asumiría la responsabilidad de ser la primera capa de validación. La decisión de involucrar al área de modelos es gobierno del área de riesgos.

Área de Validación Interna: Este equipo es formalmente la segunda capa de validación y se asegura que la primera capa haya implementado controles suficientes para lograr el objetivo deseado, en este caso un proceso adecuado de admisión de clientes que permita el ingreso de clientes sanos. Las dimensiones que evalúa el equipo de validación interna son amplias y dado la importancia del esquema de validación utilizado en la evaluación de modelos estadísticos, en el siguiente acápite se profundizará esta sección.

Figura 3

Proceso de construcción y puesta en producción de un modelo



Nota: Proceso basado en el esquema estándar secuencial de desarrollo

Para la decisión de la puesta en producción del modelo no debe haber objeciones por parte de las áreas involucradas. Los resultados y la decisión final es aprobado mediante un comité de Riesgos donde se encuentran los representantes de todos los equipos. El proceso actual sobre el cual es basado el esquema de validación es mostrado en la figura 3.

CAPÍTULO II

MARCO TEÓRICO

2.1 Esquema de validación:

Luego de aplicar el esquema de validación se resume en una nota o calificativo final, para ello se tiene en consideración la calificación por cada uno de los aspectos cualitativos y cuantitativos señalados por cada ámbito de validación (véase tabla 1) en función a la cantidad de observaciones encontradas, criterio experto del validador, además de una evaluación cuantitativa de materialidad, cuando sea posible y si el impacto-beneficio lo amerita.

La evaluación de cada ámbito se da en una escala porcentual cuya nota final tiene como máximo 100% de cumplimiento tras ponderación de cada dimensión.

Tabla 1

Ámbitos de evaluación del esquema de validación

Alcance	**Definición**
Documentación	El equipo de VI debe asegurar que la documentación sea completa y entendible por los stakeholders del negocio.
Gobierno	El equipo de VI debe asegurar que los actores encargados de realizar algún cambio, implementación y/o definiciones adicionales se encuentren correctamente definidos.
Calidad de datos	El equipo de VI debe asegurar que los datos utilizados sean creíbles, consistentes, completos y exactos, cumpliendo con la mayoría de las características de la norma ISO/IEC 25012.
Metodología	El equipo de VI debe asegurar que el estándar metodológico empleado por el equipo constructor del modelo/regla de negocio sea apropiado para cumplir con los objetivos deseados.
Integración y uso	El equipo de VI debe asegurar que la implementación del modelo/regla de negocio propuesta sea acorde a lo desarrollado metodológicamente y que su uso se circunscriba a la población definida.

El detalle de los controles cuantitativos y cualitativos por cada ámbito de validación se muestra en el siguiente resumen:

Tabla 2

Controles por ámbito de revisión

Ámbito	**Control cuantitativo y cualitativo**	**Peso**
Documentación	Revisión de Memoria de Construcción	5%
	Revisión de Memoria de Calibración	5%
	Revisión de Manual de Generación de datos	5%
Calidad de datos utilizados	Filtros de Universo, lógica de Target y Segmentación	5%
	Integridad y coherencia de variables	5%
	Reporte de incidencias, controles realizados	5%

Metodología	Definición y estabilidad de la Población	5%
	Definición y estabilidad del Target	5%
	Segmentación	5%
	Selección y tratamiento de variables	5%
	Bondad de ajuste del modelo	5%
	Análisis por subpoblaciones	5%
	Mix de riesgos, granulado y monotocidad del score.	5%
	Backtesting del modelo	5%
Gobierno	Aprobación y/o Modificación del modelo	2%
	Monitoreo y Seguimiento del modelo	2%
	Definición del Tiering del modelo	1%
Usos e Integración a la Gestión	Revisión de los usos del modelo	5%
	Medir desempeño del modelo post pauta de riesgo	5%
	Diagnostico e Idoneidad del Punto de Corte (en piloto)	5%
	Challenge entre modelos finalistas	5%
	Seguimiento del algoritmo de score en uso	5%
Benchmark (opcional)	Construcción de modelos Benchmark (retadores al modelo original)	4%

Fuente: Esquema de validación del modelo del área de Validación

Se puede observar que la sumatoria de los pesos da un score máximo del 100%, la unidad de validación tiene la opción de construir un modelo Benchmark si considera tener los recursos necesarios para su construcción.

2.2 Gestión de observaciones:

En cuanto se realice la ponderación de pesos por los diferentes ámbitos del modelo los resultados de la validación del proceso de modelamiento estarán comprendidos dentro de los siguientes niveles:

Tabla 3

Calificación final del modelo

Calificación	**Descripción**	**Acción**
Satisfactorio	El modelo no presenta deficiencias o su número es reducido y/o de poca relevancia.	Se recomienda su implementación
Aceptable	El proceso seguido satisface de manera general sus objetivos. Se han encontrado falencias que tienen un impacto menor.	Se recomienda su implementación. Sin embargo, se solicita un plan de acción a mediano plazo para las correcciones de las observaciones.
Regular	Necesariamente debe fijarse un plan de acción a corto plazo para su resolución.	Luego de superadas las correcciones se analizará la implementación.
Deficiente	El proceso presenta graves deficiencias. Por lo cual, genera una altísima exposición a riesgos.	No se recomienda su utilización.

Fuente: Esquema de calificación del modelo del área de Validación

Está puntuación final está sujeta a la cantidad y criticidad de observaciones que tiene el modelo por cada ámbito. En el informe de validación se presenta el detalle de las observaciones a subsanar. Dichos conceptos se clarifican a continuación:

- Las **observaciones** cumplen el propósito de alertar sobre alguna debilidad de los modelos y que requiera subsanación. Estas pueden ser agrupadas bajo su nivel de Riesgo Residual en Crítico, Alto, Relevante, Moderado, Bajo y Potenciales.

    - **Riesgo Crítico**

      Tienen el potencial de generar graves pérdidas financieras debido a errores encontrados en las fases evaluadas y por tanto poner en riesgo significativo la fortaleza financiera de la organización. Representan graves casos de incumplimiento de las leyes y regulaciones. Tienen el potencial de poner en grave riesgo la reputación del Banco. Tienen el potencial de poner en grave riesgo la autorización de funcionamiento del Banco u originar una intervención. Tienen el potencial de poner al Banco a Régimen de Vigilancia.

    - **Riesgo Relevante**

      Tienen el potencial de poner en grave riesgo la autorización de emplear la metodología de Modelos Internos ante el regulador. Tienen el potencial de generar importantes pérdidas financieras. Se considera que la falta de prevención o controles adecuados pueden generar pérdidas considerables. Representan incumplimientos de normas internas del Banco. Tienen el potencial de generar multas por parte de la

SBS u otras autoridades. Tienen el potencial de poner en significativo riesgo la reputación del Banco.

- **Riesgo Moderado**

  Tienen el potencial de generar moderadas pérdidas financieras. Tienen el potencial de exponer a la organización a amonestaciones de los entes reguladores.

  No aseguran un cumplimiento integral de normativas internas. No se cuenta con un adecuado sustento de las metodologías aplicadas.

- **Riesgo Bajo**

  Tienen el potencial de generar pérdidas menores para el Banco.

  Son considerados necesarios para impedir que en el mediano plazo la actuación o administración de las operaciones de la organización pierda su eficiencia.

- **Riesgo Potencial**

  Tienen el potencial de convertirse en riesgo Bajo. Sin embargo no son suficientes por su cuenta para determinar una perdida en el Banco.

- **Las oportunidades de mejora s**e formulan con el propósito de encaminar algún aspecto del modelo hacia el óptimo, en un contexto de mejora continua. Queda a potestad de la unidad validada el implementar la Oportunidad de Mejora sugerida o de incorporar una alternativa propia que será comunicada a la unidad de validación a fin de evaluar su idoneidad.

El **estado de las observaciones** se clasifica de la siguiente manera:

- **Pendiente:** La unidad validada no ha iniciado ninguna acción.
- **En proceso:** La unidad validada cuenta con un plan de acción y una fecha estimada de implementación.
- **Superado:** Observación ya implementada por la unidad validada y verificada por la Unidad de Validación Interna.

Es importante mencionar que toda **observación cuya mejora haya sido implementada** por la unidad validada, **deberá ser comunicada**, a fin de que se pueda emitir opinión sobre la idoneidad de la mejora de forma oportuna.

La ejecución del plan de acción relacionado con la implementación de las observaciones y/u oportunidades de mejora aceptadas, es responsabilidad del dueño del proceso o unidad validada.

Tabla 4

Nivel de observaciones detectadas en proyecto

	Estado de proyecto	
Nivel de Observación	**Abierto**	**Cerrado**
Bajo	75	188
Crítico	0	5
Moderado	30	96
Oportunidad de Mejora	28	48
Potencial	21	68
Relevante	12	31
Total general	**166**	**436**

Nota: Basado en proyectos realizados entre el 2016 y 2018

De manera mensual se informará en los comités correspondientes el monitoreo, la evolución y el nivel de cumplimiento en los plazos de implementación de las incidencias según su clasificación.

2.3 Importancia del nivel de ingresos para la admisión crediticia

Dentro de la evaluación del perfil del cliente el uso del ingreso del cliente juega un papel importante por las siguientes razones:

1. Existe una cota de ingreso mínimo para poder ser evaluado por cada producto financiero.

2. Es parte de la fórmula del CEM (Capacidad de Endeudamiento Máxima), el cual es el monto máximo expresado en soles por el cual un cliente puede endeudarse dado sus ingresos mensuales.

3. Según resolución SBS Nº6941-2008, si el nivel de endeudamiento es mayor a 50% el cliente debe ser considerando como sobre endeudado, donde entiéndase por nivel de endeudamiento al ratio de las cuotas que cancela mensualmente el cliente para pagar sus deudas totales del sistema financiero sobre los ingresos brutos que se tienen registrados en el banco.

4. Permite prospectar el ingreso del cliente y anticiparse al ofrecimiento de un crédito.

   Para el sustento de ingresos, el cliente debe presentar los siguientes documentos que serán registrados en la base de datos del sistema financiero:

Tabla 5

Sustentos de Ingreso por tipo de renta

Tipo de renta	Descripción	Documentos para presentar
1	Alquiler de bienes muebles o inmuebles	Predio Urbano del año en curso del titular. 3 últimos pagos de Impuesto a la Renta.
2	Ganancias de valores mobiliarios, (acciones, bonos)	Declaración Jurada Anual de (IR) del último año.
3	Empresas, negocios	3 últimos formularios virtuales de SUNAT, Impuesto a la renta IGV de 3 últimos meses, Declaración jurada anual último año
4	Independientes	Todos los recibos por honorarios emitidos en últimos 3 meses o pdt.
5	Dependientes	Ingreso constante: 2 últimas boletas. Ingreso variable (pago con comisiones): 3 últimas boletas.

Fuente: Esquema de validación del modelo del área de Validación

El banco puede utilizar en el proceso de evaluación crediticia tanto fuentes confiables de ingresos (Ingresos confiables web, PDH, CTS) como Ingresos provenientes de herramientas analíticas (estimador de ingresos).

Sin embargo, En base a los ingresos declarados por clientes en distintas agencias y levantados por distintos funcionarios desde enero del 2017 a octubre 2018 se observó que al menos el 54% de los clientes poseen un mayor ingreso del que se tenía registrado en el banco.

Tabla 6

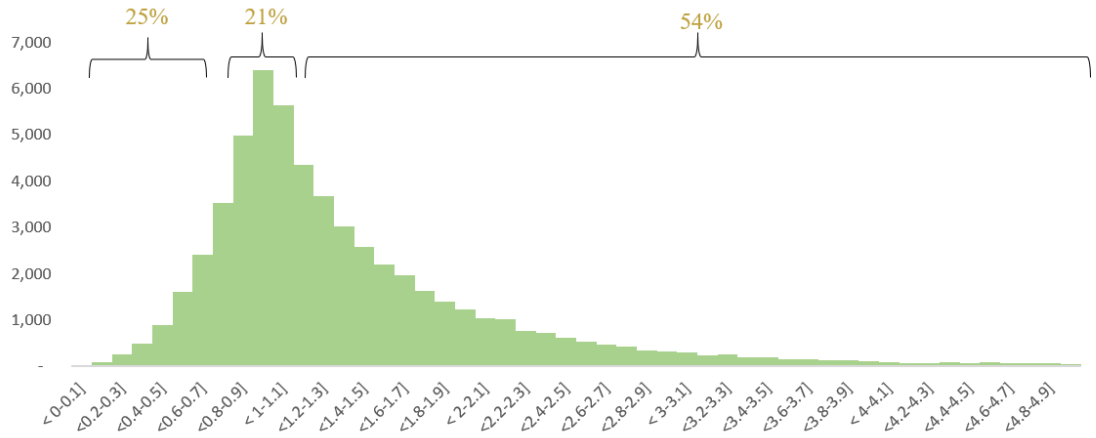
Tipos de Ingreso

Concepto	**Definición**
ICS	Ingresos confiables web
PDH0	Ingreso de pago de haberes. La diferencia está en la cantidad de meses que han tenido abonos en un lapso de 12 meses. Por ejemplo, PDH0 ha tenido abono el último mes y PDH2 ha tenido abono hasta hace 2 meses.
PDH1	
PDH2	
PDH3	
PDH4	
CTS2	CTS
CTS1	
Estimadores	Estimadores realizados el 2015
AEP1	Ingresos no confiables. En general no son tomados en cuenta

Fuente: Proceso de actualización de Ingresos del banco

Figura 4

Ratio Ingreso declarado/Ingreso confiable



Fuente: Descriptivos del área de Modelos

Debido al error en la medición del ingreso del cliente basado en los estimadores actuales, se ve conveniente construir un modelo que prediga de mejor forma el ingreso del cliente, de tal modo que sirva para determinar a qué clientes y en qué proporción, podemos confiar en el ingreso declarado por el cliente.

2.4 Metodología XGBoost

La metodología de gradient boosting utilizada en este modelamiento corresponde al XG Boost. En la metodología XGBoost (Chen, Tianqi; Guestrin, Carlos, 2016) se realiza la construcción de árboles de forma secuencial añadiendo en cada iteración el árbol que mejor compense los errores de los árboles existentes. El XG Boost tiene una ventaja comparativa en precisión con respecto a los modelos tradicionales, los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado donde cada árbol se obtiene mediante un proceso de dos etapas:

- Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que m < M (donde M = total de predictores).
- Cada árbol crece hasta su máxima extensión (Espinoza, 2020).

Según Eulufi (2020) el funcionamiento del algoritmo se puede resumir en iniciar el proceso con un árbol A0 para predecir la variable objetivo "y", al resultado obtenido se asocia con un residual (y – A0), de esta forma se obtiene un nuevo árbol que ajusta el error del proceso anterior y posteriormente se combina con A0 para obtener nuevo un nuevo árbol A1 realizando iteraciones hasta que el error cuadrático medio sea minimizado al máximo.

Espinoza (2020) señala que las principales ventajas del algoritmo son:

a) Pueden usarse para clasificación o predicción.
b) El modelo es más simple de entrenar en comparación con técnicas más complejas, pero con un rendimiento similar.

c) Tiene un desempeño muy eficiente y es una de las técnicas más certeras en bases de datos grandes.
d) Puede manejar cientos de predictores sin excluir ninguno y logra estimar cuáles son los predictores más importantes, es por ello que esta técnica también se utiliza para reducción de dimensionalidad.
e) Mantiene su precisión con proporciones grandes de datos perdidos.

Por otra parte, Espinoza (2020) señala que sus principales desventajas son las siguientes:

a) La visualización gráfica de los resultados puede ser difícil de interpretar.
b) Se deben ajustar correctamente los parámetros del algoritmo a fin de minimizar el error de precisión y evitar sobreajuste del modelo (lo que puede darse si se maneja un número muy grande de árboles).
c) Puede consumir muchos recursos computacionales en grandes bases de datos, por lo que se recomienda antes de aplicar esta técnica en bases de este tipo, determinar cuáles son las variables que aportarán más información a fin de considerar solo dichas variables en la obtención del modelo.
d) Se tiene poco control sobre lo que hace el modelo (en cierto sentido es como una caja negra).

CAPÍTULO III

FORMULACIÓN DEL PROBLEMA

3.1 Descripción de la Problemática

Las entidades financieras están obligadas de informar a la Superintendencia de Banca y Seguros las deudas que adquieren los clientes con las mismas y éste a su vez hace de conocimiento público estas deudas del Sistema Financiero con el fin de que todas las entidades se encuentren alineadas y puedan tomar previsiones en cuanto esta información antes de brindar un crédito. Sin embargo, no se solicita información y mucho menos se comparte con otras entidades los estados de cuentas de ahorro, información de ingresos entre otros productos que no representan una obligación financiera con la entidad. Es por ello, que el conocimiento del nivel de ingresos de una persona para un Banco es muy valioso, principalmente en el contexto del país donde el nivel de informalidad es bien elevado.

Dado que no se tiene la disposición del nivel de ingresos de todos los ciudadanos del Perú, representa un limitante para diversas iniciativas a realizar, ofrecer campañas, habilitar productos, verificar la veracidad de los ingresos declarados, entre otros que se detallaron en el capitulo de importancia del nivel de ingresos. Debido al error en la medición del ingreso del cliente, se ve conveniente incluir en el cálculo del ingreso final la información de ingreso declarado proporcionado por el cliente y para tal fin es necesario controlar y/o monitorear la información de ingreso declarada y así evitar fraudes.

Por ello, la primera iniciativa de estimar ingresos se realizó en el 2016 utilizando una metodología de regresión lineal la cual a la fecha poseía poca capacidad de desempeño de los predictores existentes para la determinación de ingresos y por ello se plantea la

necesidad de desarrollar herramientas estadísticas más precisas y actuales; que mejoren no solo la exactitud en la estimación, sino que permita adecuar tales herramientas a las necesidades específicas del banco y a los productos y fines que este persigue. Además, los estimadores de ingresos utilizados hasta esa fecha no puntúan a poblaciones que se han clasificado como ingresos bajos. El nuevo estimador levantará esta restricción para coberturar y puntuar al total del universo de personas.

Se espera que el nuevo modelo permita mejorar la precisión en la medición del ingreso de los clientes mediante la incorporación de la información de ingreso recogida en el momento de la solicitud y por tanto con ello incrementar el nivel de colocaciones con el mismo nivel de riesgo o reducir el riesgo manteniendo un nivel de colocaciones similar.

Al identificarse la necesidad del modelo de estimación de ingresos se decide habilitar una mesa de trabajo ágil para acelerar el proceso de construcción, validación y puesta en producción. Asimismo, se decide innovar con la utilización de la metodología Machine Learning con XGBoost, es ahí donde se detecta la principal problemática, que no existe un proceso metodológico ni un esquema de validación establecido que permitan abordar de forma eficiente la revisión de un modelo de este tipo.

Desde el punto de vista de validación, se decide utilizar de forma inicial el esquema de validación de metodologías tradicionales bajo la condición de realizar propuestas de mejora a este esquema. Los puntos más resaltantes del esquema de trabajo actual hasta antes de este proyecto era la metodología secuencial de trabajo que implicaba que el área de modelos termine todo su análisis para que el equipo de validación inicie con su revisión, algo que ocasionaba muchos reprocesos y dista de la forma de trabajo de una mesa ágil. Otro punto, era el uso de la técnica Machine Learning y conocer que otros criterios de validación se deberían considerar.

3.2 Problema General

El esquema de validación actual no permite identificar observaciones en el modelo de ingreso Machine Learning desarrollado con metodología XGBoost.

3.3 Problemas Específicos

- El esquema de validación actual no permite identificar observaciones del modelo de estimación de ingresos en el ámbito de calidad de datos.
- El esquema de validación actual no permite identificar observaciones metodológicas propias del modelo ML de estimación de ingresos desarrollado con XGBoost.

OBJETIVOS

3.4 Objetivo General

Establecer un esquema de validación que permita identificar observaciones en el modelo de ingreso Machine Learning desarrollado con metodología XGBoost.

3.5 Objetivos Específicos

- Establecer un esquema de validación que permita identificar observaciones del modelo de estimación de ingresos en el ámbito de calidad de datos.
- Establecer un esquema de validación que permita identificar observaciones metodológicas propias del modelo ML de estimación de ingresos desarrollado con XGBoost.

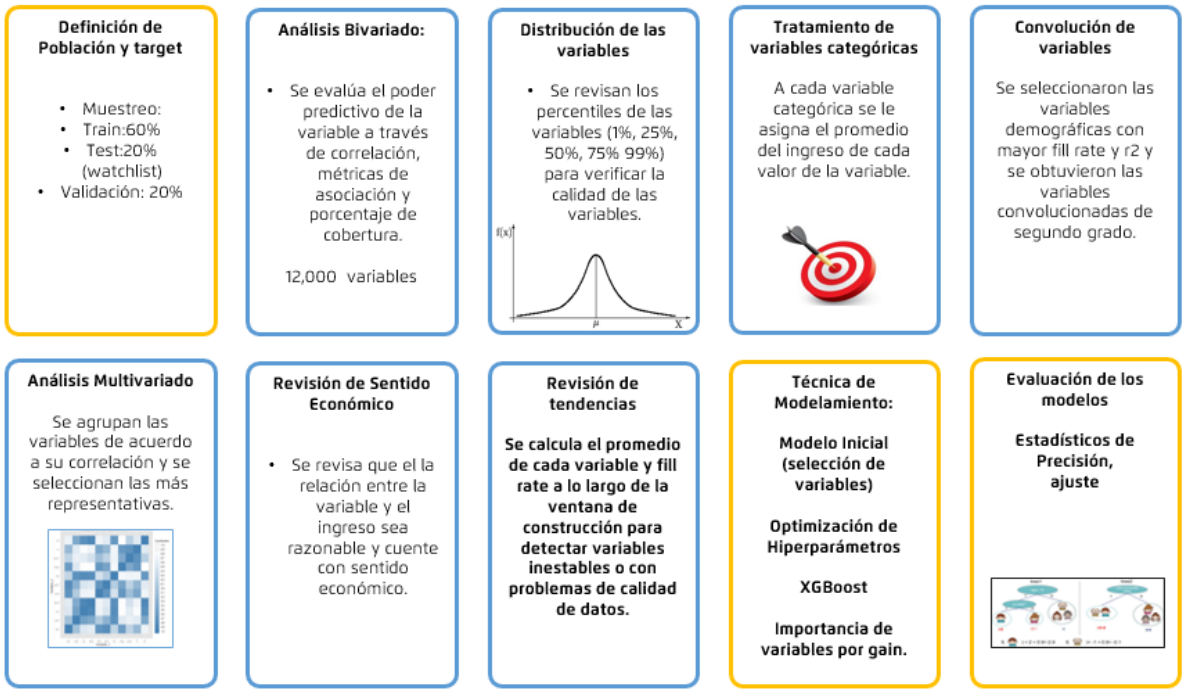
CAPÍTULO IV
MÉTODOS Y MATERIALES

4.1 Desarrollo del modelo

A continuación, se muestra el resumen del proceso de modelamiento propuesto y empleado por el área de Modelos:

Figura 5

Proceso de desarrollo del modelo



Fuente: Proceso del área de Modelos

En los próximos acápites se hará revisión de las fases de este proceso haciendo énfasis en las fases donde se detectaron las observaciones (resaltados en amarillo en la Figura 5).

4.2 Definición de la población

Compuesto por clientes con ingresos declarados en todas las solicitudes aprobadas y denegadas de los productos crédito efectivo y tarjetas de crédito cuyos valores son mayores a S/. 930 y menores a S/. 50,000 desde setiembre 2016 hasta agosto 2017. Se decide esta ventana de modelación dado la confiabilidad de los ingresos declarados en ese periodo.

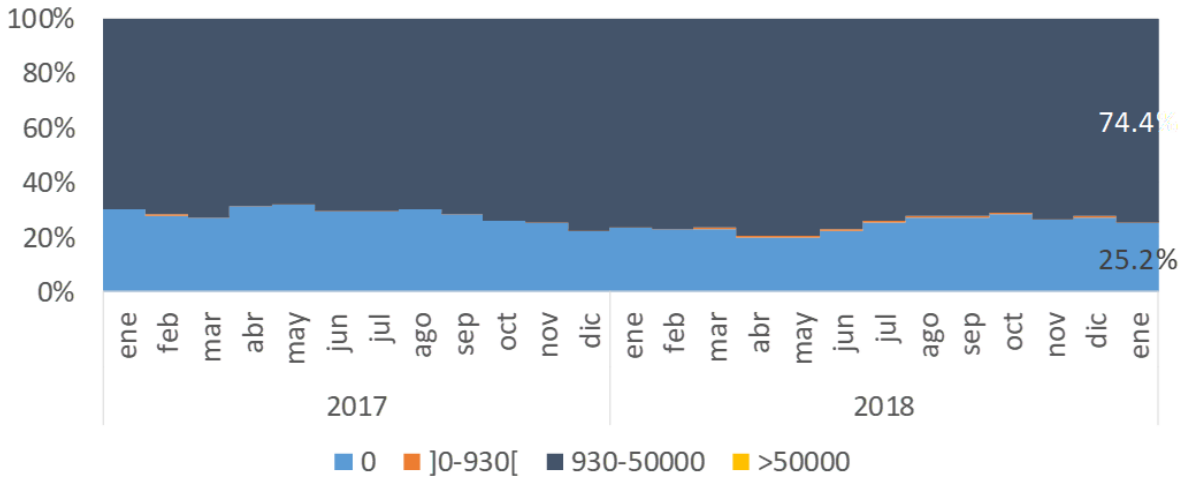
Se excluyen registros de solicitudes si en dicho mes ya se cuenta con una solicitud previa del mismo producto.

En la ventana de modelación se cuenta con 1,357,389 personas con ingreso declarado entre 930 y 50,000 soles.

Se observa una alta concentración de ingresos declarados como 0, sin embargo, casi el 75% de los ingresos serán parte de nuestra muestra de modelamiento según las definiciones de negocio acordadas.

Figura 6

Materialidad de solicitudes según rango de Ingreso declarado



Fuente: Descriptivos presentados por el área de Modelos

De la población definida se divide en 3 secciones necesarias para el entrenamiento del modelo:

- 60% Para el entrenamiento del modelo (TRAIN)
- 20% De muestra de test (TEST)
- 20% De fuera de muestra (VAL)

Estas tres muestras se estratifican de acuerdo con el rango de ingresos de los clientes y según el segmento del modelo.

4.3 Definición del target

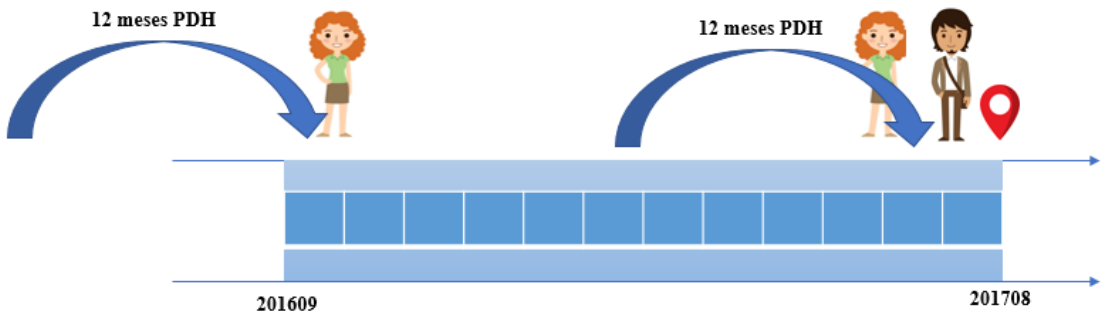
Para de la definición del target se consideran ingresos confiables aquellos con las siguientes etiquetas: ICS, PDH0, PDH1, PDH2, PDH3, PDH4, CTS1, CTS2 (véase tabla 5). Se define como target de modelamiento "Y" a la sumatoria de todos los ingresos percibidos en 1 año entre 14 (dado las gratificaciones).

$$Y = \frac{\sum_1^{12} Ingresos\ confiables}{14}$$

Por la naturaleza de este, la muestra se acota a clientes que hayan percibido un ingreso confiable por 12 meses consecutivos. Con ello la cantidad de registros utilizados para el modelamiento asciende a 7,909,526. Cabe resaltar que un cliente puede aparecer varias veces dentro de la ventana de modelamiento (setiembre 2016 hasta agosto 2017) si es que cumple la condición de tener ingreso confiable por 12 meses consecutivos. Además, a diferencia de los modelos anteriores que estimaba el ingreso en los próximos 12 meses este target estima el ingreso actual.

Figura 7

Definición del target del modelo



Nota: Esquema de definición del target considerando el periodo de construcción del modelo

4.4 Segmentación por nivel de información

El equipo constructor del modelo definió 5 segmentos de modelación excluyentes según disponibilidad de información de deuda en el Banco, deuda en el sistema financiero y tenencia de pasivos con el Banco, estos son los siguientes:

1. Segmento con información de Activos
2. Segmento con información de Pasivos y RCC.
3. Segmento con información de RCC.
4. Segmento con información de Pasivos.
5. Segmento con información Demográfica.

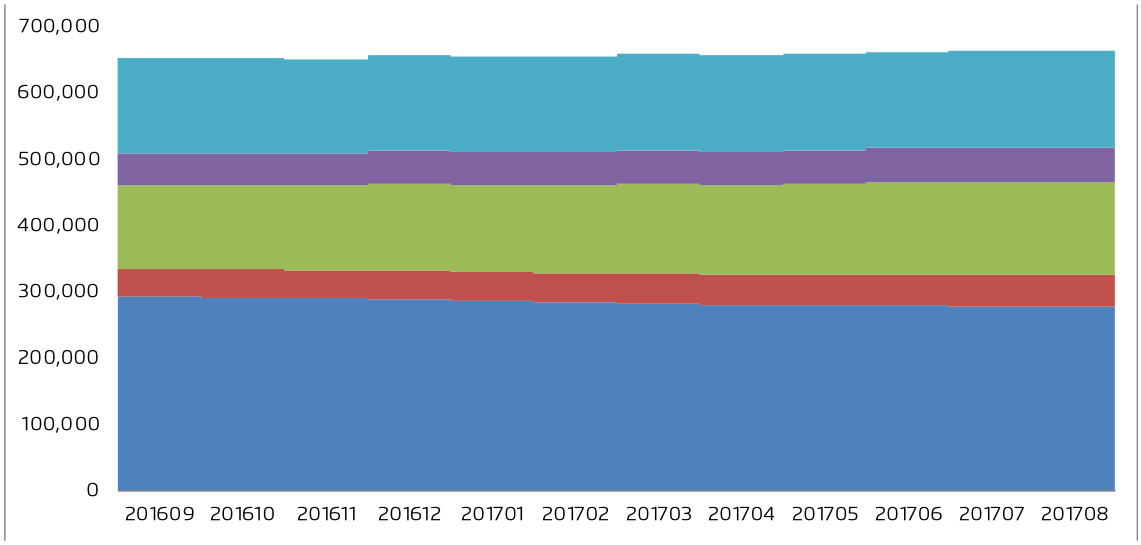
Las condiciones de tenencia de información estarán alineadas a la ventana de observación del ingreso, estas son:

- Activos: Suma de los activos promedio de mes mayor a 0 en los últimos 12 meses.
- Pasivos: Suma de los pasivos promedio de mes mayor a 0 en los últimos 12 meses (sin incluir cuenta sueldo).
- RCC: Máxima deuda total en los últimos 12 meses mayor a 0.

Con esta definición se optimiza el uso de las 12,000 variables en aquellos clientes que realmente tienen poblados dichos campos por lo que el segmento con mayor cobertura de variables es el segmento de activos y el menor el segmento de demográficos.

Figura 8

Distribución de clientes por segmento



Fuente: Descriptivos presentados por el área de Modelos

Se confirmó que, bajo este enfoque, se cumple con tener materialidad tanto en la población de construcción como en la población de aplicación y cada segmento tiene un nivel de ingreso diferenciado.

4.5 Análisis Bivariado

En esta etapa se identificaron aquellas variables explicativas que tienen una relación más fuerte con el ingreso observado, de tal manera que se puedan descartar aquellas variables que tienen menos probabilidades de ser seleccionadas en el modelo final.

Los indicadores de correlación se muestran a continuación:

- **Correlación Lineal:** Se utiliza el coeficiente de correlación de Pearson.

- **Correlación No Lineal Monótona** Se utiliza el coeficiente de correlación de Spearman

- **Correlación No Lineal** Se ajusta una función mediante splines y se calcula el coeficiente de determinación (R2) de dicha función.

4.6 Distribución de las variables

En esta sección se revisan los estadísticos más relevantes por variables con el fin de detectar valores anómalos en la distribución de estos. La revisión se realiza por segmento del modelo. En las tablas resumen se muestra representativamente sólo 5 variables por cada segmento, elegidas según sentido de negocio.

Tabla 7

Univariados del segmento Activos

VARIABLE	%MISS	Media	Mediana	Mínimo	Máximo
Monto promedio de transacciones monetarias por Uso de Banca por Internet (12m).	0	2042.67	0	0	89861127
Cantidad de transacciones del establecimiento ingreso de los clientes de categoría a cuenta en los últimos 12.	14.96	0.06	0	0	12
Cantidad de operaciones en cuentas de ahorros mínimo 12 meses.	0.42	11.03	8	0	501
Ticket promedio de cargos en cuentas de ahorros mínimo 9 meses.	0.42	231.12	167,143	0	36485.42
Monto de compras promedio en los últimos 24 meses.	11.89	1225.17	168,586	-435,107	316179.95

Tabla 8

Univariados del segmento Pasivos

VARIABLE	%MISS	Media	Mediana	Minimo	Máximo
Monto total de operaciones monetarias en atm - promedio últimos 12 meses	0	1834.45	1312.5	0	57234.64
Ticket promedio de operaciones en Uso de Banca por Internet - promedio últimos 12 meses.	0	83.32	0	0	30277.38
Ticket promedio de operaciones en ventanilla - promedio últimos 12 meses.	0	989.19	129.75	0	324940.64
Ticket promedio de cargos en cuentas de ahorros - promedio 12 meses	0.31	513.84	257.85	0	138484.12
Máximo en los últimos 12 meses monto máximo.	31.6	1504	740.36	0	270566.63

Tabla 9

Univariados del segmento Pasivos + RCC

VARIABLE	%MISS	Media	Mediana	Minimo	Máximo
Monto total de operaciones monetarias en atm - promedio últimos 12 meses	0	1645.2	1320.5	0	552314.64
Ticket promedio de operaciones en ventanilla - promedio últimos 12 meses.	0	990.3	142	0	324948.95
Ticket promedio de operaciones en Uso de Banca por Internet - promedio últimos 12 meses.	0.1	85.2	0	0	30299.8
Ticket promedio de cargos en cuentas de ahorros - promedio 12 meses	32	514.4	251	0	138400.22
Máximo en los últimos 12 meses monto máximo.	32.5	1602	690	0	290666.42

Tabla 10

Univariados del segmento RCC

VARIABLE	%MISS	Media	Mediana	Minimo	Máximo
Número promedio de empresas que reportan al cliente en los últimos 24 meses.	0	1.91	1.63	0.042	10
Nivel educacional + actividad económica / mediana.	4.22	1538.92	1355.08	504,464	6148.09
Línea máxima con tarjetas de crédito en los últimos 24 meses.	0	11702.57	3264.4	0	572263.72
Deuda total (directa más indirecta calculada para pj - mto total deuda indirecta_pj) m máxima a cliente en bs últimos 24 meses.	0	25825.4	9605.76	0.01	4213243.1
Nivel educacional + actividad económica /promedio.	4.22	1887.14	1529.66	532,526	6812.28

Tabla 11

Univariados del segmento Demográfico

VARIABLE	%MISS	Media	Mediana	Minimo	Máximo
Edad de la persona	6.53	42.59	40	22	103
Distrito + genero	9.65	1355.3	1420.27	601,675	7246.7
Nivel educacional + actividad económica	3.85	1613.02	1503	460,358	6193.43
actividad económica + relación laboral	0.37	1635.69	1654.92	909,779	4621.15
Distrito + estado civil	12.61	1657.5	1578.45	605,931	6802.24

4.7 Tratamiento de variables categóricas

En esta etapa se asignan valores numéricos a aquellas variables que son del tipo categórica, la transformación que se ha utilizado es asignar el promedio del ingreso de los registros que caen en dicha categoría en la muestra de entrenamiento. Por ejemplo:

Tabla 12

Ejemplo de tratamiento de variables categóricas

Cliente	Ingreso Obs	Cat_zon	Cate_zon_trans
1	**1,200**	CUZCO	**1,400**
2	**1,400**	CUZCO	**1,400**
3	**1,600**	CUZCO	**1,400**
4	2,000	LIMA	2,500
5	2,500	LIMA	2,500
6	3,000	LIMA	2,500

A los clientes 1,2 y 3 se les asigna el mismo valor de la categoría zona numérica el cual representa el promedio de los ingresos de esos clientes al pertenecer a una misma zona de Cuzco.

4.8 Convolución de variables

En esta etapa se crean variables que dependen de 2 o más variables categóricas con el fin de crear una nueva variable que recoja los efectos no lineales sobre la variable dependiente. Las variables elegidas para la convolución son categóricas elegidas en base a sentido

económico y que se encuentren bien pobladas tanto en la muestra de construcción como en la muestra de aplicación.

Para una mejor explicación podríamos decir que es una forma de hacer tratamiento a la interacción de niveles de 2 variables categóricas a las cuales se les asigna el valor promedio o la mediana del ingreso de los individuos con esa categoría, tal como se le hacia el tratamiento con variables categóricas.

Con ello se incluyeron 72 variables adicionales, 36 considerando el promedio y 36 considerando la mediana del ingreso.

4.9 Análisis Multivariado

En esta fase se busca seleccionar la mejor variable de las familias de variables construidas, esto dependiendo de la correlación que se tenga con el target. Para ello primero se construye la matriz de correlaciones y se reordena de acuerdo con la correlación entre las distintas variables. Como segundo paso, se calcula la correlación lineal (Pearson) de cada variable con el target y de cada grupo de variable se selecciona la variable que tenga mayor correlación (en valor absoluto) para ser candidata en el modelo final. Cabe resaltar que en esta técnica basada en arboles XGBoost no es necesario cumplir con los supuestos tradicionales de la no multicolinealidad, pero por motivos de parsimonia y a fin de llevar un orden en el proceso de modelamiento se vio recomendable realizarlo.

4.10 Revisión de Sentido económico y tendencias

En esta etapa se descartan aquellas variables que: (1) no cuenten con un sentido económico claro, es decir, no se considere que la variable es relevante para predecir el ingreso, o (2) una cuya correlación lineal sea contraria al sentido esperado, dicha revisión se realiza para cada segmento del modelo y se realiza en conjunto por las áreas involucradas (área de Modelos, Validación Interna, Riesgos y Productos).

4.11 Técnica de Modelamiento

4.11.1 Entrenamiento del modelo

Para el presente trabajo de investigación se utilizaron las siguientes herramientas:

- R Studio para el entrenamiento del modelo
- Excel para los reportes
- PLSQL para la extracción de la data

De cara a utilizar el algoritmo XGBoost en su implementación en R para el entrenamiento de un modelo predictivo, según se menciona en XGBoost tutorials para R, la entrada y salida de datos así como los input de la función de entrenamiento del modelo se deben definir por los siguientes parámetros:

- Data: objeto xgb.DMatrix que contiene la tabla maestra con la información de entrenamiento del modelo. Las variables que se encuentran en este objeto serán numéricas.

- nrounds: define el máximo número de iteraciones que entrenará el modelo.

- early_stopping_rounds: Total de rondas seguidas que se permite al algoritmo no mejorar la "performance" antes de parar. Si se establece en el parámetro el valor 5, el algoritmo validará con el "**validation set**" que el performance sigue mejorando y cuando empeore durante 5 iteraciones seguidas parará.

- watchlist: es utilizado para especificar el "validation set" durante el entrenamiento del modelo. Por ejemplo, se puede especificar en la watchlist de la siguiente forma: "watchlist = list(train = DTrain, dev = DDev)" para evaluar el performance de cada ronda del modelo en el **Train set** y **Development set**.

- verbose: indica si se quiere mostrar la información en consola relativa al performance del modelo.

- num_class: establece el número total de clases que se quieren predecir. Es un parámetro que solamente aplica en modelos multi-clase y que en este caso toma el valor por default.

Por tanto, los valores ingresados por el área de modelos para la parametrización y entrenamiento del modelo en cada segmento son como se muestra en la tabla 6.

Tabla 13

Parámetros por cada segmento

	Activos	**Pas+RCC**	**RCC**	**Pasivos**	**Demo**
Depth	11	11	13	15	13
Eta	0.05	0.05	0.05	0.1	0.1
subsample	1	1	0.6	1	0.6
colsample	0.6	0.8	0.8	0.6	1
min child	360	360	360	360	360
gamma	0	0	0	0	0
Lambda	0	0	0	0	0
Rmse	2413.055	1530.083	1103.112	1227.007	1055.276

Fuente: Descriptivos presentados por el área de Modelos

Luego de haber realizado la selección de variables a través de todas las fases del modelo y de ingresar los parámetros que necesita el algoritmo para el entrenamiento se logran obtener la cantidad de variables mostradas en la tabla 7, el modelo con menos cantidad de variables y menor Rmse lo presenta el demográfico. Es presumible dado que es el segmento con menor cantidad de variables iniciales y más difícil de explicar por tener información acotada.

Tabla 14

Resumen de cantidad de variables del modelo XGBoost por cada segmento

Modelo	Variables Univariado	Variables Multivariado	Variables con Sent. Económico	Cantidad de variables finales
Activos	7,786	935	440	200
Pasivos + RCC	6,371	720	479	200
RCC	932	171	178	178
Pasivos	5,179	560	395	200
Demográfico	395	138	139	139

Fuente: Descriptivos presentados por el área de Modelos

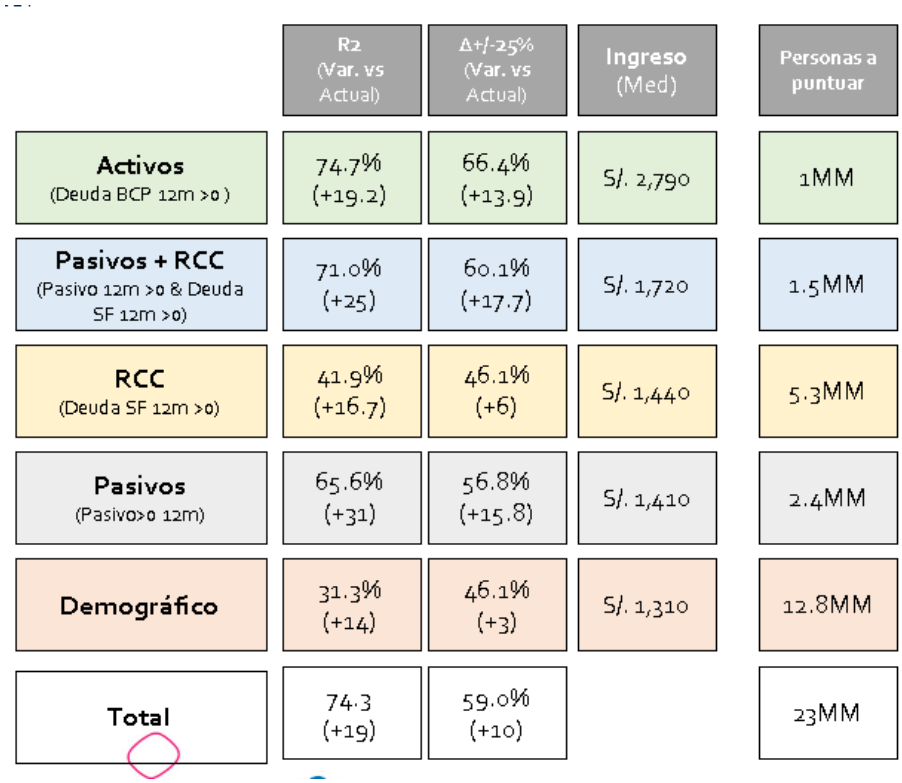
Para la elección del mejor modelo se consideran 2 indicadores, la precisión en ingresos altos, medido como el porcentaje de observaciones cuyo ingreso estimado está en el rango del 75% y 125% del ingreso observado (mayor a 5,000 y mayor a 12,000) y el coeficiente de determinación (R2). Ambos indicadores son medidos en la muestra de test.

4.11.2 Evaluación de los modelos

En la figura 9 se muestran los principales estadísticos de precisión y ajuste por cada segmento del modelo sobre el universo de entrenamiento.

Figura 9

Estadísticos sobre universo de aplicación del modelo



	R2 (Var. vs Actual)	Δ+/-25% (Var. vs Actual)	Ingreso (Med)	Personas a puntuar
Activos (Deuda BCP 12m >0)	74.7% (+19.2)	66.4% (+13.9)	S/. 2,790	1MM
Pasivos + RCC (Pasivo 12m >0 & Deuda SF 12m >0)	71.0% (+25)	60.1% (+17.7)	S/. 1,720	1.5MM
RCC (Deuda SF 12m >0)	41.9% (+16.7)	46.1% (+6)	S/. 1,440	5.3MM
Pasivos (Pasivo>0 12m)	65.6% (+31)	56.8% (+15.8)	S/. 1,410	2.4MM
Demográfico	31.3% (+14)	46.1% (+3)	S/. 1,310	12.8MM
Total	74.3 (+19)	59.0% (+10)		23MM

Fuente: Descriptivos presentados por el área de Modelos

Al tener mayor cantidad de información para la data de entrenamiento, el segmento activo es el que posee mejor R2. El segmento demográfico solo fue modelado con variables sociodemográficas por lo que tiene el menor R2 sin embargo es el que tiene mayor cobertura.

Finalmente, los modelos resultantes para los estimadores de ingresos personas se implementan mediante los objetos rds de R.

4.12 Validación del modelo

Dado que es la primera vez que se utiliza un modelo machine Learning con metodología XGBoost para estimar ingresos no existía un estándar especifico, sin embargo, se respetó el ya definido para modelos tradicionales citado en la tabla 2.

Se corroboraron los resultados obtenidos por el área de modelos asegurando que habían realizado los cálculos correctos sobre las definiciones que habían considerado. Sin embargo, a pesar de ello, se detectaron observaciones en base a la definición inicial de sus análisis.

Validación Interna calculó los estadísticos de ajuste sobre la muestra de construcción, puede verse en la tabla 8, separando data de Train, Test y Validación, es importante realizar este ejercicio para poder observar el desempeño del modelo en el subconjunto de validación que es usado para la selección del mejor modelo y no para el entrenamiento en sí.

Tabla 15

Resumen de resultados del modelo XGBoost por cada data de entrenamiento

201609-201708	**Train (60%)**		**Test (20%)**		**Validación (20%)**	
Segmento	**N**	**R2**	**N**	**R2**	**N**	**R2**
Activos	2,052,852	75.6	682,856	75.7	682,652	74.4
Pasivos+RCC	327,009	73.2	107,956	70.9	108,363	69.5
Rcc	960,778	63.2	319,069	50.1	320,282	50
Pasivos	364,423	70.6	121,428	70.6	121,008	68.2
Demográficos	1,044,441	36.8	348,452	33.8	347,957	33.3

Fuente: Descriptivos calculados por el área de Validación

Si bien se realizaron los cálculos sobre los diferentes segmentos de la muestra de entrenamiento, esto no es suficiente porque estos datos han sido utilizados para la selección del mejor modelo. Tomar una muestra independiente fuera de tiempo (OOT) nos dará mejor visibilidad del ajuste del modelo sobre la aplicación en datos independientes.

Tabla 16

Resumen de resultados del modelo XGBoost para la data fuera de tiempo (OOT)

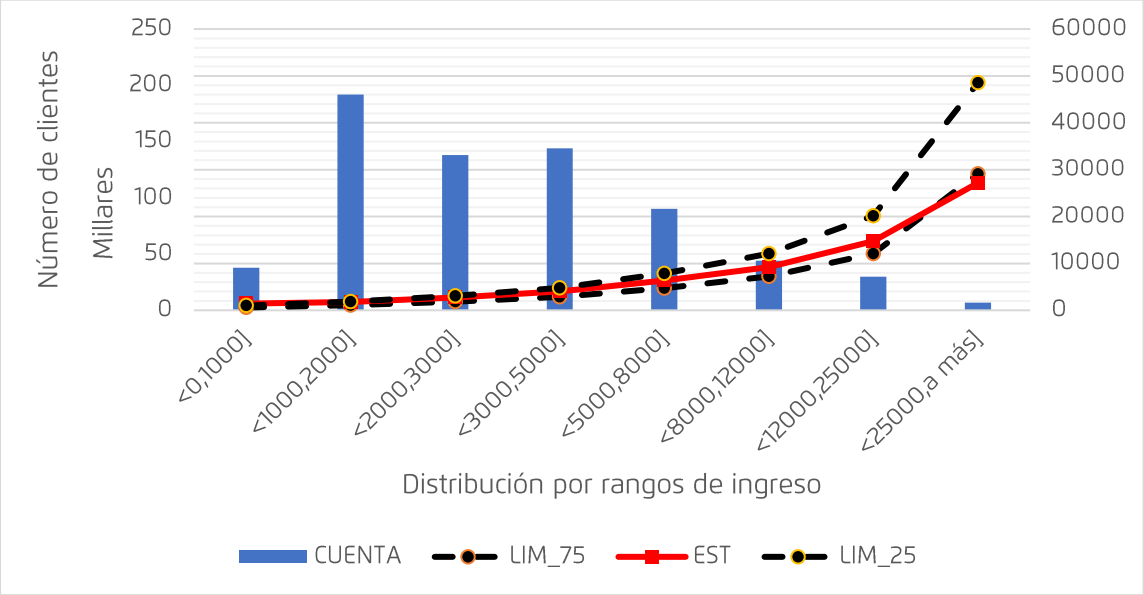
201709-201712	**OOT (60%)**	
Segmento	**N**	**R2**
Activos	1,072,811	66.3
Pasivos+RCC	201,177	68.6
Rcc	573,276	49.8
Pasivos	213,546	60.6
Demográficos	582,493	35.8

Fuente: Descriptivos calculados por el área de Validación

Asimismo, se revisó la precisión del modelo por cada segmento del modelo y rango de ingreso con el fin de notar los intervalos donde sub o sobreestimaba el modelo resultante, se toma como referencia los intervalos +-25% del ingreso real para identificar observaciones en el modelo.

Figura 10

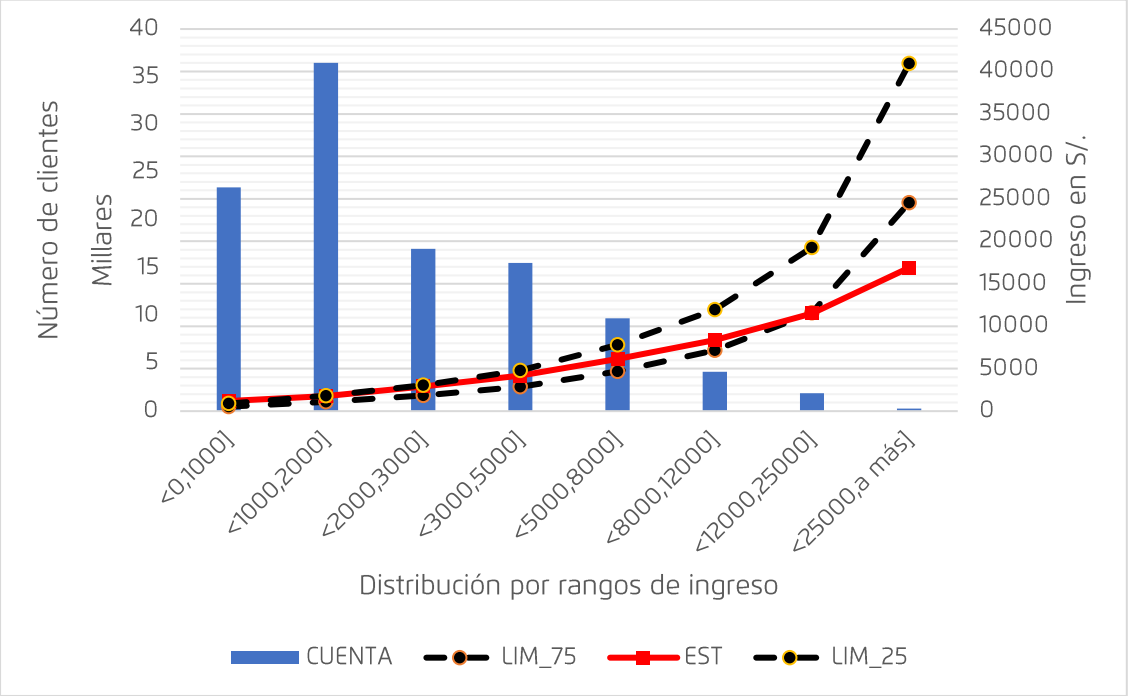
Estado de sub y sobreestimación por rango de ingreso en el segmento Activos



Para el segmento Activos se observa ligera subestimación en ingresos mayores a 25mil soles.

Figura 11

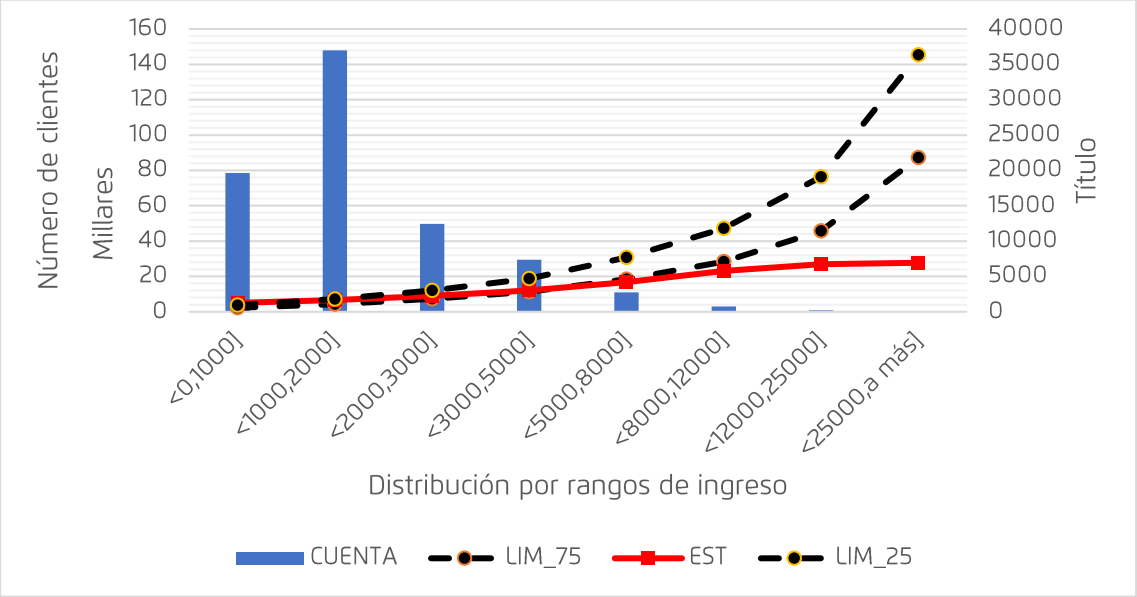
Estado de sub y sobreestimación por rango de ingreso en el segmento Pasivos+RCC



Para el segmento Pasivos+RCC se observa ligera subestimación en ingresos mayores a 25mil soles.

Figura 12

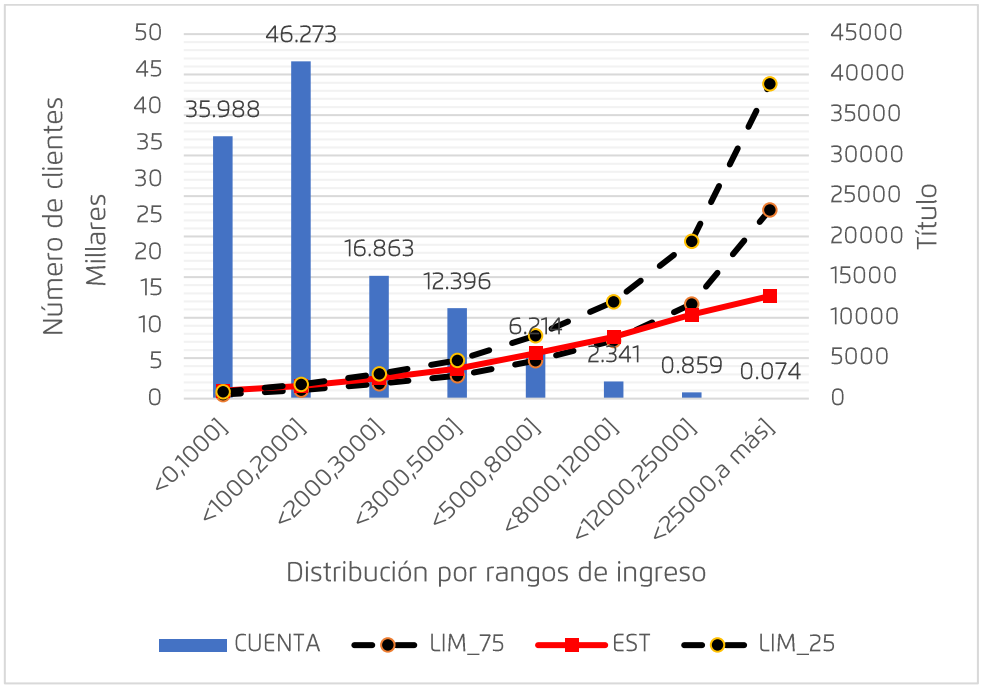
Estado de sub y sobreestimación por rango de ingreso en el segmento RCC



Para el segmento RCC se observa ligera subestimación en ingresos mayores a 8mil soles.

Figura 13

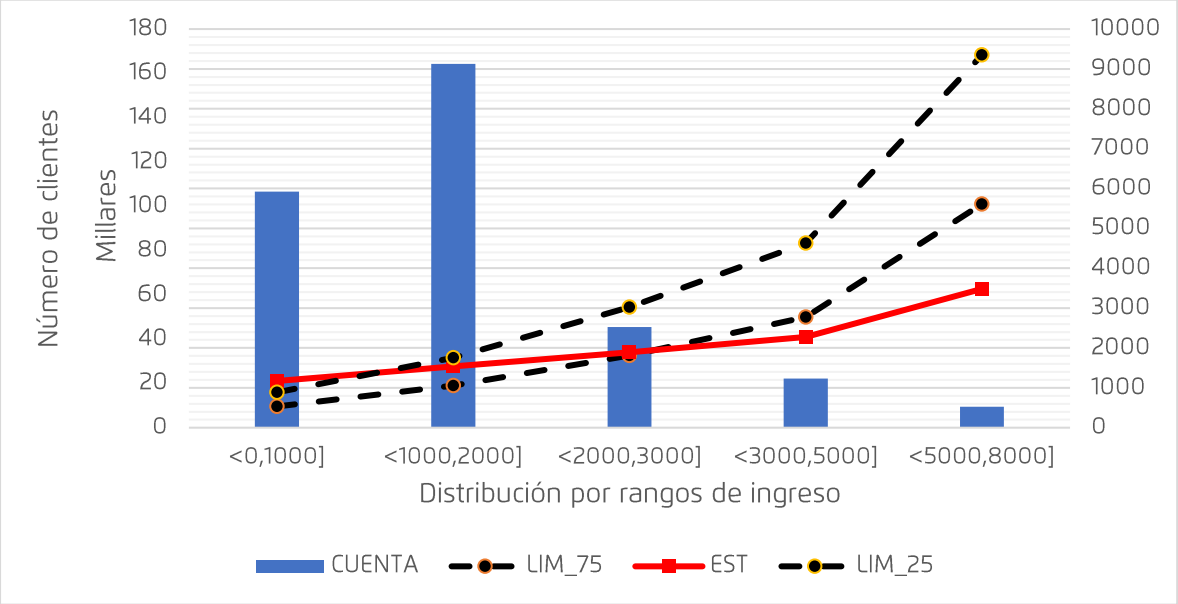
Estado de sub y sobreestimación por rango de ingreso en el segmento Pasivos



Para el segmento Pasivos se observa subestimación en ingresos mayores a 8 mil soles.

Figura 14

Estado de sub y sobreestimación por rango de ingreso en el segmento Demográfico



Para el segmento demográfico se observa ligera subestimación en ingresos mayores a 2 mil soles.

En conclusión, se observa que para todos los segmentos en los ingresos más altos estamos subestimando (25% menos del ingreso real) y en los ingresos bajos estamos sobreestimando más del límite establecido (25% más del ingreso real). Esto sugiere una revisión mas profunda de la muestra considerada para el entrenamiento del modelo.

Si bien, el detalle de los resultados de la validación se verá en el siguiente capítulo se adelanta que las observaciones encontradas generaron un reentrenamiento del modelo para así poder tener una opinión favorable del modelo.

CAPÍTULO V

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

5.1 Principales observaciones del modelo

El objetivo de la validación es asegurar que los modelos de Estimadores de Ingresos se encuentran en condiciones de ser utilizado en la gestión. La validación ha contemplado los siguientes ámbitos:

- **Documentación**

  No se encontraron observaciones relevantes.

- **Calidad de Datos**

  - No se realizó un correcto control de primera línea sobre la calidad de datos de las variables:

    - Se encontró una tendencia inversa al sentido esperado al negocio según caída en el ratio de utilización de banca por internet que afectó a los modelos de Activos, Pasivos y Pasivos+RCC (70% clientes).

      Evidencia:

      Dado que el estándar de validación aplicado no incluía la revisión del porcentaje de missings en el tiempo sino en datos con cortes transversal no se logró evidenciar que los missing de las variables de utilización de banca por internet se incrementaban en cada mes. Este comportamiento es inverso al esperado por el negocio y en general a la adaptación de la digitalización de la banca a través del tiempo.

Tabla 17

Porcentaje de utilización de la banca por internet a través del tiempo

Codmes	Promedio monto de utilización	Porcentaje de utilización
201609	1099.285	29%
201610	1076.592	29%
201611	1066.848	29%
201612	1046.431	28%
201701	1022.284	28%
201702	1007.428	28%
201703	1006.425	28%
201704	989.8662	28%
201705	946.8171	27%
201706	923.3587	27%
201707	896.7077	27%
201708	865.2643	26%
201709	822.2907	26%
201710	789.23	25%
201711	782.0348	25%
201712	756.7606	25%

Fuente: Elaboración propia.

Se identificó que la fuente de extracción de estos datos había cambiado y si bien el universo de clientes se mantenía estos registraban como “0” la cantidad de transacciones diarias en la utilización de la banca por internet por el cambio de tabla fuente.

Cabe indicar que tras el hallazgo el equipo de Validación incluyó esta revisión de datos a nivel longitudinal para las variables dentro de su estándar.

- En el set de variables de uso de servicios, las cuales fueron evaluadas para su inclusión en el modelo, las variables presentaban proporciones mayores a uno.

  Evidencia:

Las variables de uso de servicio identificaban que proporción de viviendas del total contaban con servicios de agua u otros, o qué proporción de viviendas respecto del total eran alquiladas. Se logró identificar que en la mitad de las variables finalistas el máximo de sus valores presentaba proporciones mayores a 1, ello reflejaba la inconsistencia de la creación de estas variables por lo cual fueron eliminadas del modelo. Cabe indicar que estas variables tenían poca relevancia en el modelo ML.

Tabla 18

Estadísticos univariados de la población en evaluación

Descripción	PCTL_5	PCTL_95	MISSING_PERCEN	MEDIAN	MEAN	MODE	STD_DEV	MIN_VALUE	MAX_VALUE
PROPORCIÓN DE PERSONAS AFILIADAS A ESSALUD EN LA MANZANA	0.09	0.55	51.86	0.33	0.32	0	0.14	0	1
PROPORCIÓN DE PERSONAS AFILIADAS A ESSALUD EN LA MANZANA	0.16	0.76	51.86	0.47	0.5	0.5	0.19	0	1
PROPORCIÓN DE PERSONAS AFILIADAS A NINGÙN SEGURO DE SALUD EN LA MANZANA	0.01	0.56	51.86	0.13	0.19	0	0.17	0	1
PROPORCIÓN DE PERSONAS DE PERSONAS AFILIADAS AL SIS EN LA MANZANA	0.00	0.16	51.86	0.04	0.05	0	0.06	0	1
PROPORCIÓN DE PERSONAS AFILIADAS A ESSALUD, SIS Y A OTROS, EN LA MANZANA	0.24	0.96	51.86	0.54	0.57	0.5	0.22	0	2
PROPORCIÓN DE VIVIENDAS CON SERVICIO DE AGUA DIARIO POR MANZANA	0.41	1.24	51.90	0.96	0.93	1	0.25	0	4
PROPORCIÓN DE VIVIENDAS EN LA MANZANA CON TENENCIA DE LA VIVIENDA:	0.58	1.20	51.90	0.93	0.92	1	0.21	0	2.25
PROPORCIÓN DE VIVIENDAS EN LA MANZANA CON TENENCIA DE LA VIVIENDA: PROPIA PAGADA	0.00	0.30	51.90	0.04	0.08	0	0.11	0	1.5
PROPORCIÓN DE VIVIENDAS EN LA MANZANA CON TENENCIA DE LA VIVIENDA: PROPIA TOTAL	0.20	1.04	51.90	0.62	0.63	1	0.25	0	2

Nota: Univariados calculados sobre todos los segmentos

Se logra observar en la muestra de variables de la tabla 18 seleccionado en rojo las variables de Uso de Servicio cuyo calculo de ratios que por definición deben ser menores a 1 no cumplían con ese criterio. Se resuelve retirar toda esa familia de variables.

- **<u>Metodológica</u>**

Estas observaciones están enfocado a corregir la desventaja del uso de modelos de Machine Learning respecto al sobreajuste.

- No se evaluó las diferentes posibilidades de configuración de parámetros. VI realizó este ejercicio.

Evidencia:

Dado que el XGBoost corrige el error de las predicciones en cada iteración, ello tiende a sobreajustar los datos, por tanto, para evitarlo, se recomienda regularizar los valores de los parámetros en la función objetivo. El equipo de validación realizó el ejercicio sin embargo no se encontraron diferencias significativas en el cambio de las principales variables y estadísticos del modelo.

- Definir un mejor criterio de segmentación y target para clientes con un perfil diferente al Banco.

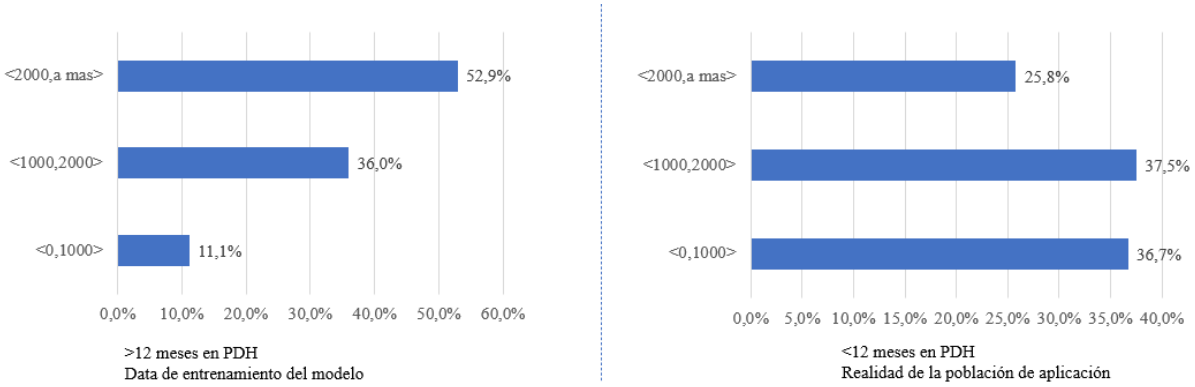
Evidencia:

La población de la entidad financiera en estudio tiene una distribución de ingresos diferente a la población de aplicación (peruanos con ingresos mayores a 900 soles y menores a 50 mil soles). Los clientes bancarizados en esta entidad financiera generalmente poseen ingresos más elevados que la población general por lo que se presume existirá una tendencia de sobreestimar el ingreso para la población no bancarizada al tomarse esta muestra como la data de modelamiento. Además, dado que el target del modelo tiene como premisa que el cliente haya tenido un ingreso

confiable durante 12 meses obliga a que la población de entrenamiento tenga un ingreso estable, algo que dista bastante de la realidad económica del país.

Figura 15

Comparación de distribuciones según continuidad de ingresos formales

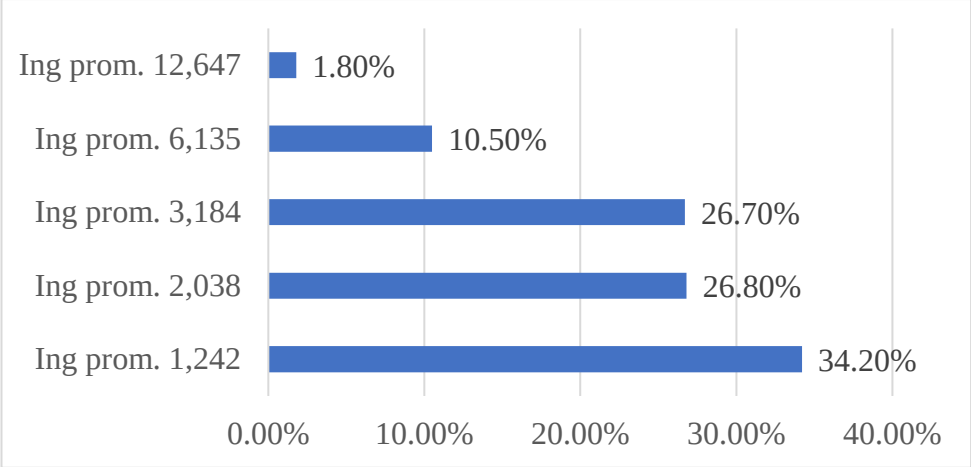


Fuente: Elaboración propia

En la figura 10 se puede observar que al forzar que un cliente tenga al menos 12 meses consecutivos de ingreso formal se considere mas de un 50% con ingresos superiores a 2000 soles. En la figura 11 se muestra la distribución real de ingresos a nivel nacional, donde se corrobora que al forzar los 12 meses consecutivos se invierte la forma de la distribución real de ingresos a nivel nacional.

Figura 16

Distribución de Ingresos en hogares 2019



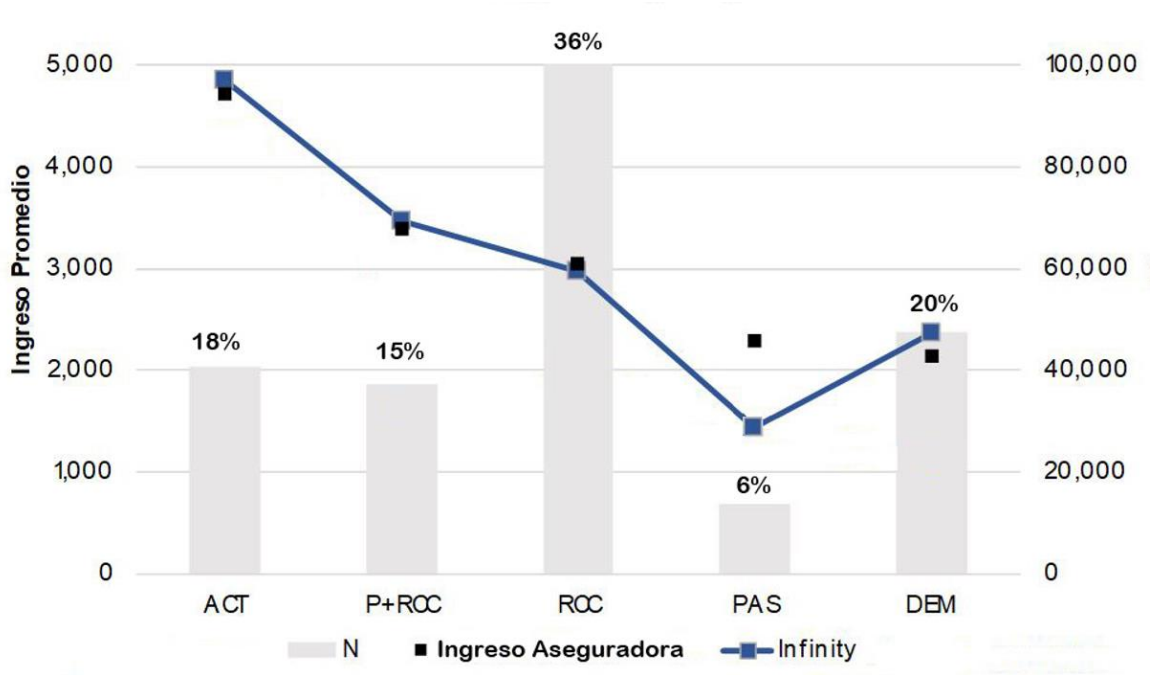
Fuente: Ipsos 2021

Comparando con la distribución de ingresos real a nivel nacional se puede presumir que es necesario cambiar el supuesto del target. Tras esta observación el área de modelos contrastó el ajuste del modelo sobre una muestra externa.

Para poder contrastar el ajuste del modelo se logró conseguir el apoyo de una Empresa de Seguros, la cual puntuó a nivel de segmentos el ingreso que ellos tenían registrado y este se comparó con el resultado del modelo.

Figura 17

Comparación entre los promedios de ingresos de la aseguradora vs el estimador de Ingresos



Fuente: Resultados de la aseguradora

Se puede identificar que en el segmento pasivos (PAS) de la población puntuada existe una subestimación de ingresos del modelo y en el segmento demográfico (DEM) una sobreestimación. Esto es razonable dado que el segmento pasivos está definido por la población que tiene ahorros en el sistema financiero y la aseguradora levanta dicha información en función a los pagos que realizan los asegurados, a diferencia de la relevancia que se da en el banco a levantar información de los egresos del cliente para el control de la morosidad. En el segmento demográfico también es explicable la sobreestimación debido a que son clientes no bancarizados y al estimar el modelo con población bancarizada se tiende a estimar ingresos mas altos de los que le corresponde.

Para corregir este problema se realizó un reentrenamiento del modelo con un target que no exigía tener 12 meses de ingresos continuos sino solo 6 meses de ingresos confiables dentro de una ventana de 8 meses a fin de representar el nivel de informalidad que existe en el país.

- **Gobierno:**

  Observación de riesgo potencial:

  - La información presentada de los indicadores siempre debe corresponder a data totalmente independiente de la modelización a fin de sincerar los resultados. Esta observación se debe a que los primeros resultados se mostraron sobre la **muestra de aplicación del modelo** y el sinceramiento de resultados por cada muestra de entrenamiento, validación y test fue calculado por el equipo de validación**.**

- **Integración a la gestión**

  No se encontró un estándar definido para un modelo machine Learning con metodología XGBoost.

5.2 Reformulación del esquema de validación

Dado que es la primera vez que se utilizaba un modelo machine Learning con metodología XGBoost para estimar ingresos no existía un estándar ad hoc para comparar métricas. Por tanto, se planteó el siguiente esquema de seguimiento la cual ayudaba a brindar visibilidad la necesidad de calibración del modelo.

Tabla 19

Semáforos asignados por rango de R2

R2	Rango de R2
<40%	Rojo
<40%-50%]	Aceptable
<50%-70%]	Bueno
>70%	Muy Bueno

Nota: Semáforo definido en coordinación con el área de Modelos y aceptado como criterio para calibración

En la tabla 11 se muestran los semáforos de R2 definidos para un modelo XGBoost, el R2 calculado.

Tabla 20

Semáforos asignados por precisión dentro del intervalo +-25%

% Dentro del intervalo	+- 25%
<35%	Rojo
<35%-45%]	Aceptable
<45%-55%]	Bueno
>55%	Muy Bueno

Nota: Semáforo definido en coordinación con el área de Modelos y aceptado como criterio para calibración

En la tabla 12 se muestran los semáforos de precisión definidos para un modelo XGBoost. Esto quiere decir que si al menos un 35% de la población se encuentran dentro del intervalo de ingreso de +-25% el modelo tendrá calificación de aceptable.

Tabla 21

Semáforos asignados para la correlación entre variables del modelo

	Correlación entre variables
>95%	Rojo
<95%-90%]	Evaluar su uso

Nota: Semáforo definido en coordinación con el área de Modelos y aceptado como criterio para calibración

Es conocido que la multicolinealidad es un problema de los modelos de regresión lineal dado que esto afecta al método de mínimos cuadrados ordinarios. Sin embargo, dado que el modelo utilizado es un ML con metodología XGBoost, esto no nos debería importar tanto, pero a fin de que no haya duplicidad de conceptos de variables y un proceso de implementación redundante se decide establecer un tope en el uso de las variables del modelo en función a su correlación.

Tabla 22

Semáforos asignados sobre la cantidad de variables del modelo

	Cantidad de variables del modelo
>200	Rojo (aceptable previa justificación)
<200	Aceptable

Nota: Semáforo definido en coordinación con el área de Modelos y aceptado como criterio para calibración

Si bien la cantidad de variables de un modelo ML puede incrementar la aceptación de los estadísticos de ajuste, esto puede ser contraproducente durante la implementación y seguimiento del modelo. Por tanto, se decide proponer el semáforo mostrado en la tabla 14.

Tabla 23

Semáforos asignados por concentración de gain por variable

	Gain por variable (Concentración)
>30%	No aceptable
<20%-30%]	Evaluar y justificar su uso, establecer controles de seguimiento sobre la variable

Nota: Semáforo definido en coordinación con el área de Modelos y aceptado como criterio para calibración

Se podría decir que el gain por variable en un modelo de ML reemplaza el p-valor de un modelo tradicional, haciendo un símil con un modelo de regresión. En ML el gain refleja el % de la importancia de la variable en el modelo y a fines de no tener una alta concentración solo en algunas de ellas se decide colocar este tope.

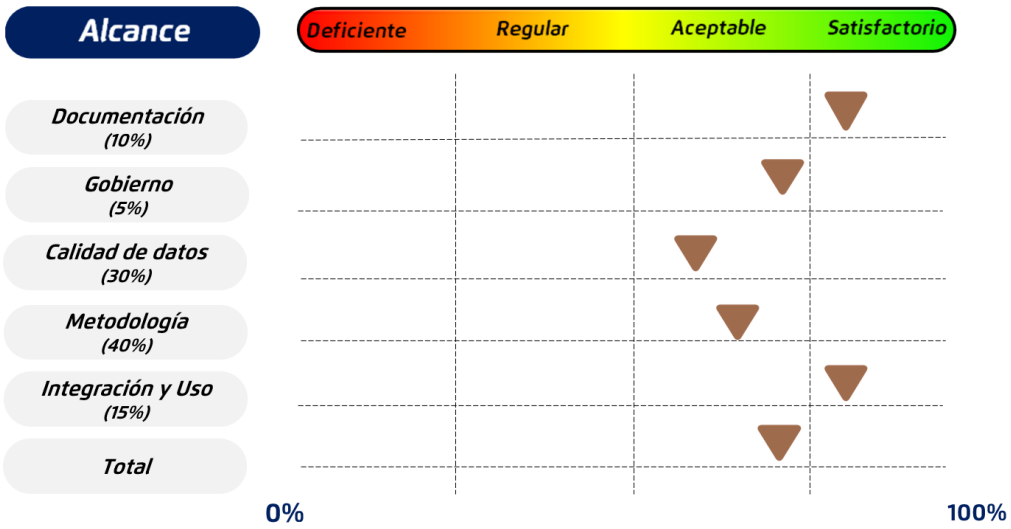
Es importante señalar que las métricas definidas anteriormente se realizaron en consenso con el área de Modelos, el área de Riesgos y el área de Validación a fin de llevar un mejor control de modelos de este tipo.

5.3 Reentrenamiento del modelo

Tras la subsanación de las observaciones planteadas por el equipo de Validación el modelo obtuvo una calificación de Aceptable. Finalmente, la nota de validación por cada ámbito del modelo de estimador de ingresos con las correcciones de las observaciones realizadas por el equipo de Modelos se muestra en la figura 11.

Figura 18

Resumen de resultados de la validación



Nota: Esquema de resumen de validación presentado en comité

Finalmente, el detalle del estado de las observaciones por cada ámbito se muestra a continuación:

Calidad de Datos

- No se realizó un correcto control de primera línea sobre la calidad de datos de las variables:
  - Se encontró una tendencia inversa al sentido esperado al negocio según caída en el ratio de utilización de la banca por internet. que afectó a los modelos de Activos, Pasivos y Pasivos+RCC (70% clientes). **Se retiraron del modelo**
  - En el set de variables de uso de servicios, las cuales fueron evaluadas para su inclusión en el modelo, las variables presentaban proporciones mayores a uno. **Se retiraron del modelo**

Metodológica

- Ante la salida de las variables de utilización de la banca por internet (dado la observación de calidad de datos antes citada), la variable más relevante por gain "Monto de consumo promedio de los últimos 12 meses" asume mayor relevancia (gain >25%) **Se asume el riesgo**
- No se evaluó las diferentes posibilidades de configuración de parámetros. VI realizó este ejercicio. **Subsanado al ser calculado por Validación**
- Definir un mejor criterio de segmentación y target para clientes con un perfil diferente al banco. **Subsanado**

Gobierno:

Observación de riesgo potencial:

- La información presentada de los indicadores siempre debe corresponder a data totalmente independiente de la modelización a fin de sincerar los resultados.

Esta observación se debe a que los primeros resultados se mostraron sobre la **muestra de aplicación del modelo** y el sinceramiento de resultados por cada muestra de entrenamiento, validación y test fue calculado por el equipo de validación**. Subsanado al ser calculado por validación**

- **<u>Integración a la gestión</u>**

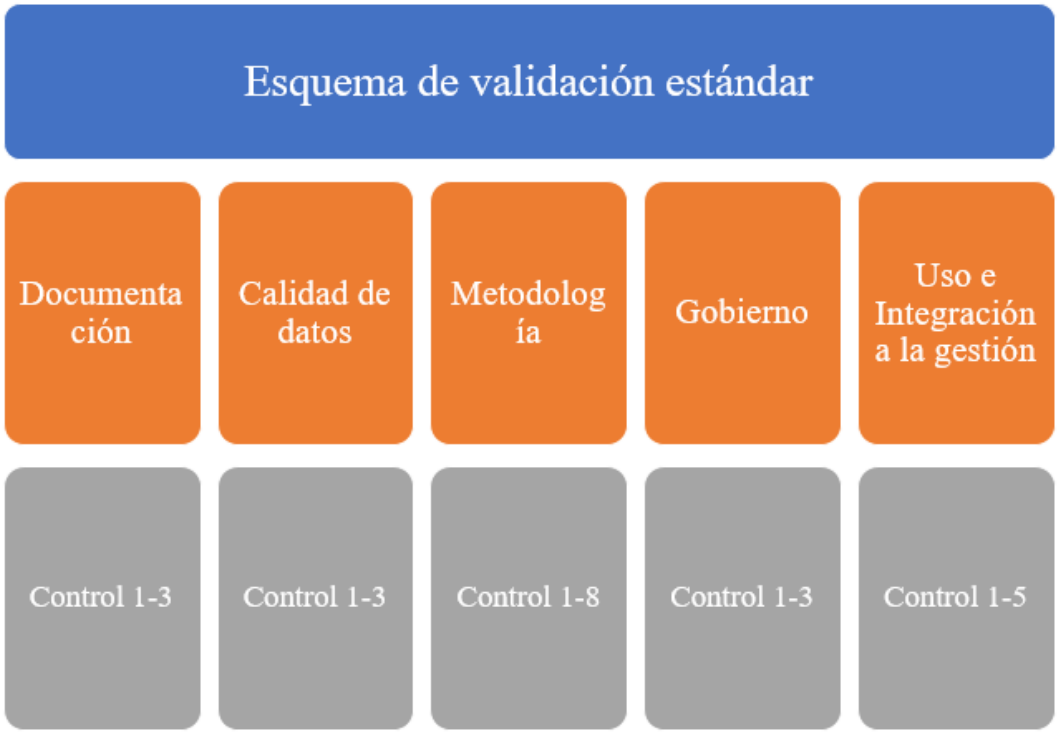
  No se encontró un estándar definido para un modelo machine Learning con metodología XGBoost. **Subsanado dado la propuesta de Validación y acuerdo con el área de Modelos.**

Al finalizar la validación del modelo el área correspondiente brindó el conforme con el visto bueno de las áreas involucradas (Riesgos, Productos, Validación, Modelos) en el proceso de creación e implementación del modelo.

Con los hallazgos encontrados durante la validación se logró complementar el esquema de validación actual estándar incluyendo aspectos de revisión dentro de una metodología ágil y los controles adicionales a considerar en una metodología Machine Learning. Los cambios más relevantes se pueden observar en las figuras 19 y 20 los cuales evidencian la comparación de diferencias, nótese la ausencia de niveles de revisión en el esquema estándar.

Figura 19

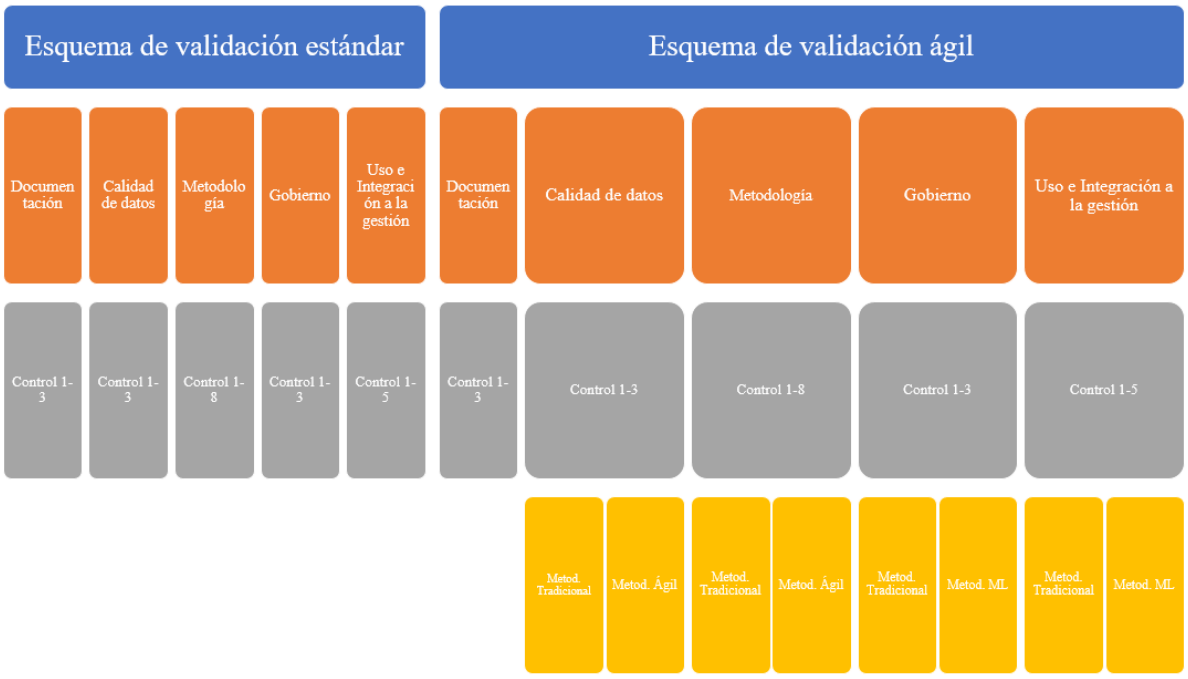
Esquema de validación estándar



Nota: Estructura del esquema de validación previo al uso de metodología ágil y de Machine Learning

Figura 20

Esquema de validación potenciado



Nota: Estructura del esquema de validación posterior a la revisión del modelo dentro de un esquema ágil y de metodología Machine Learning.

Con esto se puede concluir el aporte significativo que tuvo este tipo de validación y que queda como precedente para los próximos desarrollos y validaciones de este tipo.

CONCLUSIONES

- El esquema de validación no fue suficiente para identificar observaciones en el modelo de ingreso Machine Learning desarrollado con metodología XGBoost. Por lo que complementar el estándar ayudó a subsanar y mejorar significativamente el modelo para que sea aceptado su puesta en producción. Con ello se puede concluir que los beneficios e importancia de los modelos construidos mediante el método de Machine Learning con metodología XGBoost si bien mejoran el nivel de predicción de los modelos tradicionales, estos deben tener en cuenta las desventajas de sobreajuste, optimización de parámetros, guías referenciales para su validación y considerarse la complejidad que este tipo de modelos puede tener para su implementación.

  Además, se debe considerar que el esquema de validación del modelo es cambiante a la par con la actualización de las nuevas metodologías existentes, el no adaptarse a ello puede conllevar a un uso inadecuado del modelo, a una implementación errónea y con ello un fallo total del proceso. Por ello es importante el cuestionamiento y reto ante nuevas metodologías emergentes. Es importante que la empresa invierta en cursos, talleres u otros para la actualización de conocimientos en todos los equipos multidisciplinarios y así se logre potenciar rápidamente las propuestas y los análisis a realizar, ello ahorraría tiempos en investigación y revisión de documentación.

- El esquema de validación **no fue suficiente en el ámbito de calidad de datos** para identificar observaciones en el modelo de ingreso Machine Learning desarrollado con metodología XGBoost. Complementar el esquema de validación proponiendo en lo sucesivo una revisión de los datos a nivel longitudinal permitió alertar errores en las fuentes de extracción de los datos tal como sucedió en las variables relacionadas al uso de la banca por internet. En este caso se reforzó el esquema de validación con una vista diferente a los datos, sin embargo, ello no implica que se tenga un documento guía cerrado para las próximas validaciones sino más bien evidencia la necesidad de complementar constantemente el esquema y que no todos los controles van a resultar necesariamente en una observación. Por ende, es importante evaluar cuales controles adicionales que se retan en un modelo especifico van a pasar a ser parte del esquema.

- El esquema de validación **no fue suficiente en el ámbito metodológico** para identificar observaciones en el modelo de ingreso Machine Learning desarrollado con metodología XGBoost. La asertiva decisión de revisar la deficiencia del sobreajuste que tiene la metodología XGBoost permitió la aplicación del modelo para población con características diferentes a la del modelamiento por lo que se incrementó la admisión de clientes. Hay que considerar que también es posible probar alternativas como LightGBM o CatBoost, bajo la consideración que ello necesitaría evaluar complementar el estándar de validación presentado. Además, se debe tener en cuenta que las técnicas de gradient boosting al ser iterativas toman un tiempo considerable en su ejecución y corren el riesgo de no llegar a la convergencia, lo cual no podría asegurar que siempre exista un modelo (con las variables independientes ingresadas) que puedan explicar la variable dependiente.

RECOMENDACIONES

- Es importante mantenerse actualizado respecto a la aparición de las técnicas computacionales sin dejar de lado el conocimiento estadístico sobre metodologías tradicionales. El conocimiento combinado de ambos te permite tener un mejor desempeño en el desarrollo y evaluación del modelo propuesto. Asimismo, esto está ligado al uso adecuado del software y hardware que van a permitir el entrenamiento eficiente del modelo.

- Muchas veces un ingeniero estadístico especializado en el desarrollo de modelos se enfoca en lograr resultados de ajustes más precisos, descuidando los supuestos iniciales del modelo. Es importante reconocer que todas las fases, desde la originación de la idea, generación de la data hasta la implementación, son importantes e igual de relevantes.

- Es importante desarrollar la habilidad de proponer mejoras en los procesos dentro de una mesa de trabajo multidisciplinaria, puesto que los puntos de vista de otros profesionales complementarán las ideas y así se retará de mejor forma la propuesta.

- Dentro de la carrera de Ingeniería estadística es recomendable potenciar los conocimientos estadísticos con los de informática para poder implementar en diferentes herramientas (R, Python, Julia, entre otros) los nuevos algoritmos de aprendizaje que son utilizados en el Machine Learning los cuales generalmente iterativos y requieren de gran uso de memoria y CPU.

- También se recomienda sumar la necesidad de complementar la malla curricular con cursos de actualización sobre metodologías Machine Learning que cada vez se convierte en un mundo más amplio y que muchas veces es usado indiscriminadamente sin tener consideración de nociones básicas de estadística.

REFERENCIAS BIBLIOGRÁFICAS

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Espinosa-Zúñiga, J.J., (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería Investigación y Tecnología, Recuperado de https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Eulufi, C., (2020). Modelo de predicción de operaciones de crédito con posible default financiero. Ingeniería de la Universidad del Desarrollo, Recuperado de https://repositorio.udd.cl/bitstream/handle/11447/4042/Evitemos%20el%20castigo,%20Modelo%20de%20predicci%C3%B3n%20de%20operaciones%20de%20cr%C3%A9dito%20con%20posible%20default%20financiero.pdf?sequence=1

Pesantez-Narvaez, J., Guillen, M., Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Revista Española Risks*. Recuperado de https://www.mdpi.com/2227-9091/7/2/70

XGBoost Tutorials (2021). Recuperado de https://xgboost.readthedocs.io/en/stable/tutorials/

INEI. (2020). La informalidad y la fuerza del trabajo. INEI. Recuperado de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1764/cap04.pdf

Banco de crédito. (2020). *Memoria Integrada 2020.* La Molina, Lima.

IPSOS (2021). *Perfiles Socioeconómicos del Perú 2021*. Recuperado de: https://www.ipsos.com/es-pe/perfiles-socioeconomicos-del-peru-2021

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

Luckner, M., Topolski, B. & Mazurek, M. (2017). Application of XGBoost algorithm in fingerprinting localisation task. 16th IFIP TC8 International Conference, CISIM 2017 661-671. Bialystok, Poland: CISIM. https://doi.org/10.1007/978-3-319-59105-6_57.

Mitchell, T. M. (1997). Does machine learning really work? AI magazine, 18(3), 11-11.