Universidad del Desarrollo
Facultad de Ingeniería

# MODELO EMPRENDEDOR

Basado en Información Tributaria

POR: JUAN FRANCISCO JARA ABURTO

Tesis presentada a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Sr. LEO FERRES

Diciembre 2021

SANTIAGO

AGRADECIMIENTO

Este trabajo no habría sido posible sin los valiosos consejos de Leo Ferres (UDD) y Loreto Bravo (UDD), ni sin el apoyo incondicional de mi esposa, Nicole Graber.

# Resumen

Este trabajo busca entender si es posible identificar a aquellas empresas productivas que pueden llegar a convertirse en PyME (pequeñas y medianas empresas van desde 2.400 UF hasta 100.000 UF de venta anual) basándose en información tributaria disponible en Servicio de Impuestos Internos.

Esta información es útil para segmentar clientes empresas y ofrecerles desde un inicio una oferta de productos apropiada para apoyar el crecimiento, por ejemplo, desde la industria bancaria. La segmentación de una microempresa con potencial de convertirse en PyME y de una microempresa destinada a seguir como tal, es inherentemente distinta, tanto como en oferta de valor, modelo de atención, riesgo de default esperado en caso de otorgarles un crédito, etc. Poder identificarlas con antelación puede suponer una ventaja competitiva importante para una empresa B2B.

El mejor resultado obtenido fue mediante XGBoost preprocesando la data categórica con WoE Encoding y calibrando hiperparámetros con el algoritmo Subplex. Se obtuvo un AUC PR de un 45,43% en un set de test. Los resultados de estas predicciones fueron analizados mediante SHAP Values, donde la Actividad Económica de la empresa es la que más incide en la variabilidad final de la decisión del modelo. El modelo seleccionado posee un ECE de un 0,59%,

lo cual indica que las probabilidades obtenidas se encuentran bien calibradas y pueden ser utilizadas para estimar utilidades esperadas.

# Tabla de Contenidos

# Índice de Ilustraciones

# Índice de Tablas

# 1. Introducción

Las empresas constituyen un pilar importante para la generación de riquezas y empleos de la sociedad. Son gran fuente de innovación; y generan una enorme variedad de productos y servicios que aumentan la calidad de vida de las personas e incentivan el progreso.

Existen 1.294.136 empresas en el sistema chileno para el año comercial 2019 (ver Figura 1), las cuales se conforman en dos grandes grupos: las personas jurídicas y las personas naturales con giro (Servicio de Impuestos Internos, 2021).

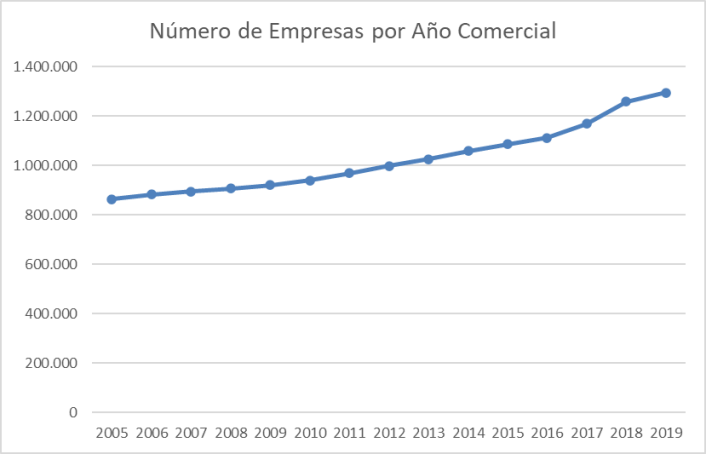



*Figura 1: Número de Empresas por Año Comercial*

Estas empresas pueden ser segmentadas por diversos factores, como actividad económica, ubicación geográfica, etc. Pero la forma más común de identificarlas es por el volumen de sus ventas. Esto no es un indicador de qué tan rentable es

la empresa (en términos de ingreso o ROE), pero si un indicador de tamaño transversalmente utilizado en industrias B2B, como la banca, para poder seleccionar elementos comerciales tales como: el modelo de atención, la oferta de valor, así como otros indicadores que apuntan a la solvencia económica o indicadores de riesgo. La gran mayoría de las empresas del sistema chileno son microempresas (59,48% al año comercial 2019, ver Figura 2). Muchas de ellas nacen como microempresas y se mantienen como tal. Algunas de ellas prosperan hasta convertirse en pequeñas, medianas o hasta grandes empresas con el pasar de los años. Esa barrera entre las microempresas y las PyME es un límite relativamente importante al momento de segmentar las empresas. El límite se sitúa en las UF 2.400 de venta anual, cerca de $ 5.800.000 de venta mensual.

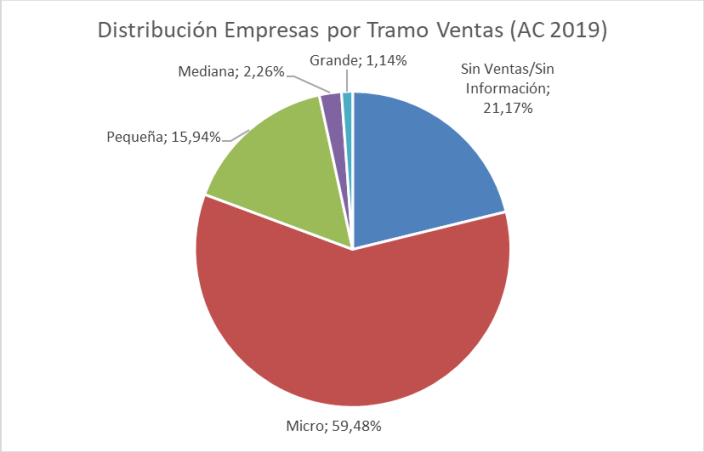



*Figura 2: Distribución de Empresas por Tramo de Ventas*

A medida que las empresas crecen en venta, son cada vez más las que van migrando de la figura de persona natural con giro en persona jurídica (ver Tabla

1). Existe un sesgo relevante en cuanto puede crecer una empresa persona natural con giro y es por esto en parte que este trabajo se basa principalmente en estudiar la evolución de las personas jurídicas.

| Empresas (#) | Femenino | Masculino | Persona Jurídica y otros | Total general |
|---|---|---|---|---|
| Sin Ventas/Sin Información | 11,27% | 19,73% | 69,00% | 100,00% |
| Micro | 26,19% | 34,58% | 39,23% | 100,00% |
| Pequeña | 9,74% | 20,61% | 69,65% | 100,00% |
| Mediana | 1,83% | 6,74% | 91,43% | 100,00% |
| Grande | 0,30% | 1,51% | 98,20% | 100,00% |
| **Total general** | **19,56%** | **28,21%** | **52,23%** | **100,00%** |

*Tabla 1: Distribución de Género por Tramo de Ventas*

Dentro del universo de personas jurídicas, el 44,67% de las empresas son microempresas y un 27,97% no poseen ventas o no tienen información suficiente aún para calcularlas (ver Figura 3).

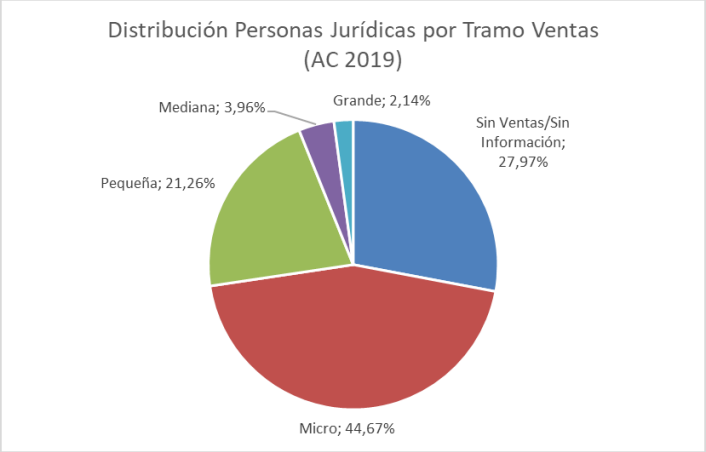



*Figura 3: Distribución de Personas Jurídicas por Tramo de Ventas*

Existen otros trabajos en la literatura para abordar proyecciones de venta (Cheriyan, 2018), éxito de empresas en el tiempo (Afolabi, 2019) o quiebra de empresas chilenas (Romani, 2002) que puede tomarse como un enfoque complementario al foco de este trabajo. Las diferencias con este proyecto se basan principalmente en la disponibilidad de la información (no existe la misma información disponible en otros países de forma pública) y la masividad del alcance (todas las empresas del sistema chileno pueden ser evaluadas con este enfoque). Los análisis elaborados en otros trabajos se basan en información más acuciosa de un grupo de empresas relativamente más pequeño. Estos análisis pueden servir de inspiración para complementar la información utilizada en este trabajo, con el cuidado de no sesgar las conclusiones a un subconjunto de las empresas analizadas.

# 2. Objetivos

## 2.1. Objetivos Generales

El objetivo general de este proyecto es tener una herramienta de discriminación entre las empresas que tienen una proyección de convertirse en PyME y las que no, con el objeto de ayudar a la segmentación comercial de forma temprana.

## 2.2. Objetivos Específicos

Como objetivos específicos, se encuentra:

- Aplicación de los diferentes pasos de la metodología KDD
- Flexibilizar el uso del modelo a diferentes escenarios en base a un enfoque de utilidad esperada.
- Inferir qué factores son los que más inciden en la decisión del modelo (*model explainability*).
- Utilización de algoritmos que sean posible poner en producción y mantener a bajo costo, como lo son la familia de los *gradient boosting trees* o logit.

# 3. Marco Conceptual

## 3.1. Información Disponible

Para realizar este análisis, se dispone de la base de contribuyentes personas jurídicas de todo Chile que dispone el Servicio de Impuestos Internos de forma anual, desde el año comercial 2005 al año comercial 2019 (Servicio de Impuestos Internos, 2021).

En esta base se encuentra información de:

a. **Año comercial**: año del que es representativo el dato. En contraste con el año tributario que es el año siguiente al comercial. El último dato disponible es año tributario 2020, que corresponde al año comercial 2019.

b. **Rut de la empresa**

c. **DV**: Dígito verificador del rut

d. **Razón Social**: Nombre de la empresa.

e. **Tramo de ventas**: Corresponde a 13 tramos de ventas anuales que dispone el SII (ver Tabla 2). El dato de venta se calcula en base a la información proporcionada en los formularios F22 (Declaración de Renta) y F29 (Declaración Mensual y Pago Simultáneo de Impuestos).

| Tramo de Venta | Mínimo Venta Anual en UF | Máximo Venta Anual en UF | Segmento |
|---|---|---|---|
| 1 | | | Sin Información |
| 2 | 0,01 | 200 | Micro Empresa |
| 3 | 200,01 | 600 | Micro Empresa |
| 4 | 600,01 | 2.400 | Micro Empresa |
| 5 | 2.400,01 | 5.000 | Pequeña Empresa |
| 6 | 5.000,01 | 10.000 | Pequeña Empresa |
| 7 | 10.000,01 | 25.000 | Pequeña Empresa |
| 8 | 25.000,01 | 50.000 | Mediana Empresa |
| 9 | 50.000,01 | 100.000 | Mediana Empresa |
| 10 | 100.000,01 | 200.000 | Gran Empresa |
| 11 | 200.000,01 | 600.000 | Gran Empresa |
| 12 | 600.000,01 | 1.000.000 | Gran Empresa |
| 13 | 1.000.000,01 | | Gran Empresa |

*Tabla 2: Tramos de Venta utilizados por el SII*

f. **Número de trabajadores dependientes informados**: Los trabajadores se cuentan por empleador. Aquellos con más de un empleo son contabilizados en cada empresa donde trabajan ese año.

g. **Fecha inicio de actividades vigente**: Fecha de inicio de la actividad económica vigente.

h. **Fecha término de giro**: Fecha del término de giro si es que lo hay.

i. **Fecha primera inscripción de actividades**: Fecha de la primera inscripción de actividades. Podría diferir de la fecha de actividades vigente si la primera actividad económica inscrita ya no se encuentra vigente.

j. **Tipo término de giro**: en caso de que haya un término de giro, este campo indica si es un término de giro de persona jurídica o un término de giro simplificado por la resolución exenta N°41 del año 2002 (es decir que el término de giro se ocasiona por 12 meses sin realizar actividades del giro o no declararlas en el F29). Para años más antiguos existen tipologías de Termino de Giro de forma genérica y término de giro persona natural con giro, pero son muy pocos casos y están en desuso desde el año comercial 2012.

k. **Tipo de contribuyente**: Existen las siguientes categorías de tipo y subtipo de contribuyente:

    a. Persona Jurídica Comercial
    b. Sin Persona Jurídica
    c. Organización Sin Fines de Lucro
    d. Sociedades Extranjeras
    e. Instituciones Fiscales
    f. Municipalidades

g. Organismos Internacionales

h. No Clasificados

l. **Subtipo de contribuyente**: son 52 subcategorías del punto anterior.

m. **Tramo capital propio positivo**: tanto como para el tramo de capital propio positivo o negativo, el SII utiliza codificaciones de tramos en UF (ver Tabla 3)

| Tramo de Capital Propio | Mínimo en UF | Máximo en UF |
|---|---|---|
| 1 | 0 | 10 |
| 2 | 10 | 25 |
| 3 | 25 | 50 |
| 4 | 50 | 100 |
| 5 | 100 | 250 |
| 6 | 250 | 500 |
| 7 | 500 | 1.000 |
| 8 | 1.000 | 2.000 |
| 9 | 2.000 | 10.000 |
| 10 | 10.000 | |

*Tabla 3: Tramos de capital propio positivo y negativo utilizados por el SII*

n. **Tramo capital propio negativo**: ver punto anterior.

o. **Rubro económico**: el Rubro, subrubro y actividad económica se encuentran detallados en (Servicio de Impuestos Internos, 2021). Hasta el año tributario 2018 se utiliza la información declarada en la operación renta u operación renta anterior. A partir del año tributario 2019, si el contribuyente posee una única actividad económica, se utiliza esa, si no, la que se declara explícitamente en el formulario 29 como actividad principal. En caso de que no se pueda determinar de esta forma, se procede como hasta el año tributario 2018. La actividad económica así

declarada o deducida podría no necesariamente ser la actividad económica principal del contribuyente. Existen 22 rubros económicos identificados por el SII en el año comercial 2019.

p. **Subrubro económico**: Existen 232 subrubros económicos identificados por el SII en el año comercial 2019.

q. **Actividad económica**: Existen 654 actividades económicas en uso identificadas por el SII en el año comercial 2019.

r. **Región**: Existen 16 regiones que conforman Chile más una categoría de "sin información". La ubicación geográfica aquí mencionada representa en general la ubicación de la casa matriz de la empresa. Dentro de una Región se encuentra varias Provincias y dentro de cada Provincia existen varias Comunas. La información se encuentra reprocesada hacia atrás, ya que, por ejemplo, la Región de Ñuble existe como tal desde septiembre 2018. No debería existir para años comerciales anteriores al 2018 y las comunas y provincias que actualmente la componen deberían aparecer en la región del Biobío, lo cual no es así. Esto genera una consistencia en la información (una comuna pertenece a solo una región), pero en caso de haber cierta influencia de la administración en la generación y prosperidad de empresas en estos sectores, podría ser más valioso tenerlo como realmente fue administrado y dejar que el modelo discrimine qué combinación resulta más provechosa.

s. **Provincia**: Existen 57 provincias en uso para el año comercial 2019.

t. **Comuna**: Existen 347 comunas en uso para el año comercial 2019.

Existe una segunda fuente de información de las actividades económicas encontrada en la página de SII (Servicio de Impuestos Internos, 2021) que contiene:

1. **Código Actividad Económica**
2. **Nombre Actividad Económica**
3. **Nombre Subrubro**
4. **Nombre Rubro**
5. **Afecto a IVA**: con categorías SI, NO y G (el cual se determina por características propias de la actividad)
6. **Categoría Tributaria**: con categorías 1, 2 y G (se determina por características propias de la actividad). Esto corresponde a actividades que tributan en primera categoría (rentas obtenidas del capital y de las empresas comerciales, industriales, mineras y otras) o segunda categoría (rentas del trabajo, asociadas a sueldos, gratificaciones, honorarios, etc.).
7. **Disponible Internet:** con categorías SI y NO.

## 3.2. KDD (Knowledge Discovery in Databases)

Esta metodología base es un proceso secuencial que permite extraer patrones de comportamiento de la información alocada en una base de datos (Chehab, 2020). El proceso KDD considera una serie de etapas (ver figura 4).

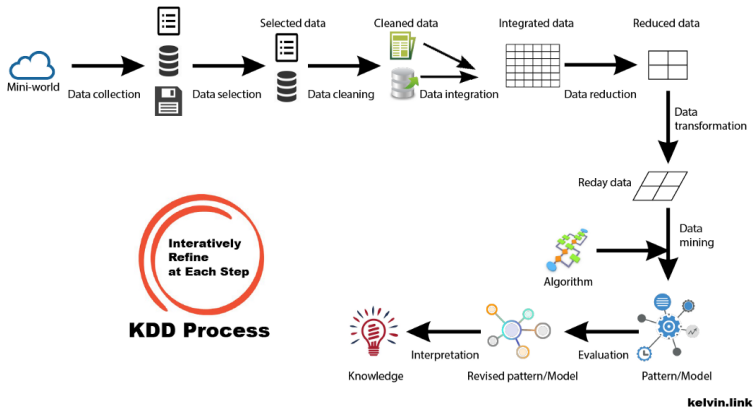



*Figura 4: Etapas Proceso KDD*

Existen otros marcos sobre los cuales se pueden desarrollar descubrimiento de patrones en data, como SEMMA o CRISP-DM. La metodología SEMMA es relativamente equivalente a la KDD, mientras que CRISP-DM está orientada al ciclo continuo de búsqueda, actualización y puesta en producción de modelos. El alcance de este trabajo no contempla la puesta en producción, se orienta más al potencial descubrimiento de patrones en una secuencia más lineal, similar a la propuesta en KDD o SEMMA.

Las etapas del proceso KDD se pueden resumir en:

1- Data selection: Comprensión del dominio en estudio, selección de data relevante para modelar la realidad que se busca analizar

2- Preprocessing: moldear la data en un formato que pueda ser asimilado por un modelo, estadística descriptiva básica, tratar *outliers*, tratar *missings values*, etc.

3- Transformation: transformar los atributos de modo que un modelo pueda hacer mejor uso de ellos: *data encoding*. También se incluye en esta sección técnicas de selección de atributos.

4- Data mining: evaluar diversos modelos (o una misma familia de modelos con diferentes hiperparámetros) y seleccionar el que mejor pueda explicar el fenómeno a analizar. Para entrenar se utiliza un set de training (calibración de parámetros) y para selección de modelo o calibración de hiperparámetros se utiliza un set de validación.

5- Evaluation: Considera la evaluación “out of sample” lo más objetiva posible del desempeño del modelo una vez puesto en producción. Se utiliza un set de test y debería reevaluarse el modelo de forma continua para asegurarse que los patrones encontrados siguen vigentes en el tiempo. También contempla el análisis e interpretación de los patrones encontrados, validación o rechazo de hipótesis previas a la construcción del modelo referente a los datos y limitantes del modelo. Eventualmente

se puede concluir que el modelo no explica suficientemente bien la realidad para la aplicación que se busca hacer de él.

## 3.3. WoE Encoding e Information Value

Para codificar la información categórica en números, existen diversas técnicas, cada una con ventajas y desventajas. En el diagrama de la Figura 5 (Mazzanti, Beyond One-Hot. 17 Ways of Transforming Categorical Features Into Numeric Features, 2020) se pueden apreciar algunas técnicas no supervisadas y supervisadas de encoding.

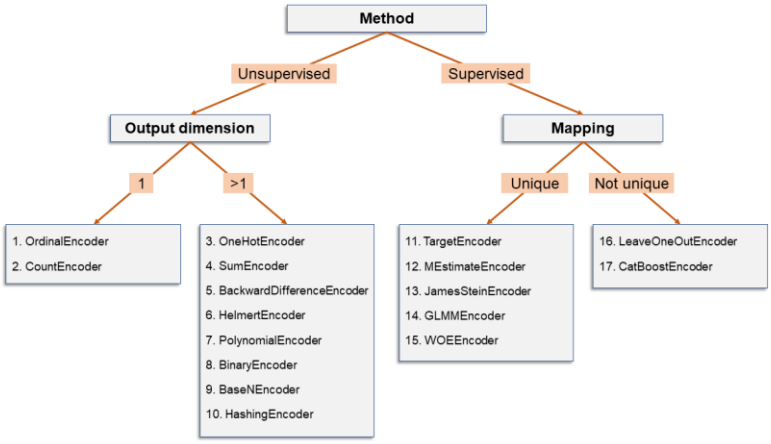



*Figura 5: Tipos de Encoding para variables Categóricas*

Técnicas populares como One Hot Encoding tienen la desventaja de incrementar mucho la dimensionalidad de la base, en particular en un dataset como este

donde existen muchas variables categóricas y cada una de esas variables tiene muchas categorías posibles.

Para este trabajo se utilizará Weight of Evidence (WoE) Encoding, el cual es un método supervisado para codificar las distintas categorías basado en la contribución de cada categoría en la variable a explicar (Bhalla, 2015). Es por esto que debe ser calibrado sobre el set de entrenamiento para evitar sobreajustes en los datos. La ventaja de este método es que permite asignar a cada categoría de una variable un único número representativo de su contribución. Esto implica que el número asignado debería guiar al modelo posterior en encontrar una diferenciación para la variable dependiente sin incrementar la dimensionalidad de la base.

El WoE para cada categoría t se calcula como:

$$WoE_t = ln\left(\frac{\%\ casos\ positivos_t}{\%\ casos\ negativos_t}\right)$$

Donde $\%\ casos\ positivos_t$ y $\%\ casos\ negativos_t$ son respectivamente la proporción de casos positivos y negativos de la categoría t sobre el total.

De esta forma, el WoE captura la el ratio entre los casos positivos y negativos de una categoría, permitiendo codificar dicha categoría en un número sin incrementar la dimensionalidad de la base.

El WoE se encuentra relacionado con otro indicador llamado Information Value (IV), y ambos se encuentran relacionados con la regresión logística. El IV tiene la siguiente fórmula:

$$IV_t = WoE_t * (\%\ casos\ positivos_t - \%\ casos\ negativos_t)$$

El IV ayuda a discriminar qué variables tienen un poder predictivo sobre determinado target. Existe la convención de que los valores entregados por el IV pueden ser interpretados como indica la Tabla 4 (Siddiqi, 2006).

| Information Value (IV) | Predictive Power |
|---|---|
| < 0.02 | useless for prediction |
| 0.02 to 0.1 | weak predictor |
| 0.1 to 0.3 | medium predictor |
| 0.3 to 0.5 | strong predictor |
| > 0.5 | suspicious or too good to be true |

*Tabla 4: Interpretación Information Value (Krishnan, 2018)*

De esta forma, el análisis de los atributos categóricos mediante Information Value nos puede servir para selección de atributos y el WoE para codificar aquellos atributos categóricos relevantes.

## 3.4. Selección de Atributos

La selección de atributos es importante para mejorar el desempeño de un modelo y también para la reducción del costo computacional (tiempo y almacenamiento).

Como se puede ver en la Figura 6, existen varias formas de selección de atributos (Brownlee, How to Choose a Feature Selection Method For Machine Learning, 2019):

- No Supervisado: no utiliza la variable dependiente a predecir. Por ejemplo, correlograma entre variables dependientes para remover multicolinealidad
- Supervisado: utiliza la variable dependiente para seleccionar qué atributos conservar. Existen de diferentes categorías:
    - Filtros: análisis univariado del cada atributo contra la variable predictora para determinar algún grado de relevancia. No considera interacción con otros atributos independientes para explicar el target. Por ejemplo: correlación de Pearson entre variable independiente y variable target, information value, etc.
    - *Wrapper Methods*: considera pasos iterativos entre la ejecución de un modelo y la selección de los atributos que hacen ese modelo tener cierto grado de precisión. Por ejemplo: se puede utilizar un Logit con todos los atributos inicialmente, ver el test estadístico de significancia de cada coeficiente, eliminar el que tenga el peor p-valor (bajo cierto umbral) e iterar hasta que todos los coeficientes den significativos.
    - Intrínseco: se refiere cuando la técnica de modelamiento tiene incorporada de forma inherente un método de selección de

atributos, como es el caso de los árboles de decisión: en cada división de una rama, se selecciona el mejor atributo en base a algún criterio predefinido (gain, entropía, test chi cuadrado, etc.)

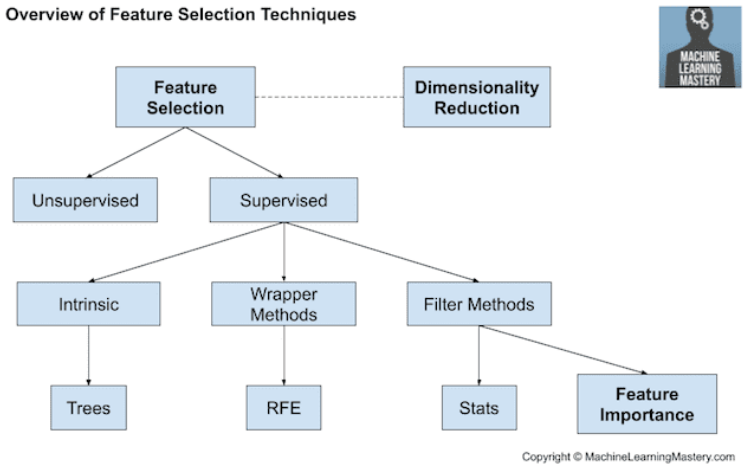



*Figura 6: Técnicas de Selección de Atributos (Brownlee, A Gentle Introduction to XGBoost for Applied Machine Learning, 2016)*

## 3.5. Curvas ROC y PR

Las curvas ROC (Receiver-Operating Characteristic) y PR (Precision-Recall) son curvas que se utilizan para evaluar modelos de clasificación binaria cuyas salidas son las probabilidades de pertenecer a la clase objetivo. En estas curvas se grafican los cambios en 2 indicadores, que usualmente presentan cierto grado de trade-off, para cada corte de probabilidad. Cada corte en probabilidad entrega una matriz de confusión (ver Figura 7) y estos indicadores de desempeño.

| | Reference | |
|---|---|---|
| Predicted | Event | No Event |
| Event | A | B |
| No Event | C | D |

$$Sensitivity = A/(A+C)$$

$$Specificity = D/(B+D)$$

$$Precision = A/(A+B)$$

$$Recall = A/(A+C)$$

*Figura 7: Matriz de Confusión e Indicadores de desempeño (Kuhn, 2020)*

La curva ROC (ver Figura 8) grafica el False Positive Rate (FPS o 1-Specificity) contra el Sensitivity o TPR (True Positive Rate) o Recall para cada corte de probabilidad (en el gráfico representado por una escala de colores). El óptimo se alcanza si algún corte en probabilidad otorga el 100% del TPR y el 0% del FPR, punto que se encuentra en la esquina superior izquierda del gráfico. Un modelo que no logra vencer la probabilidad aleatoria de pertenecer a cada clase para cada corte de probabilidad, es representado por una línea diagonal donde tanto FPR y TPR se incrementan a tasas idénticas.

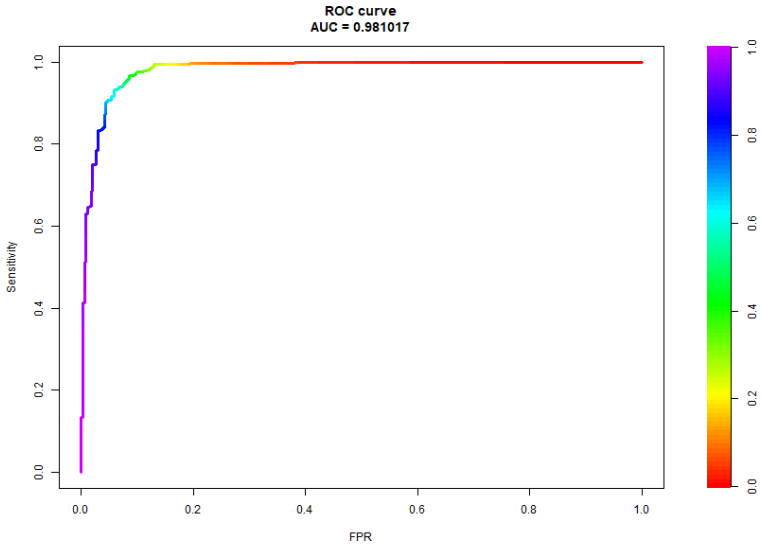



*Figura 8: Curva ROC*

Una medida global del desempeño del modelo, independiente del corte en probabilidad utilizado, es el área bajo la curva (AUC). El AUC de la curva ROC va desde la línea en diagonal que tiene un área de 0.5 al modelo perfecto que alcanza la esquina superior izquierda, la cual otorga al modelo un AUC de un 1.

La curva PR se construye de modo similar, pero graficando para cada corte en probabilidad el Precision y el Recall (ver Figura 9).

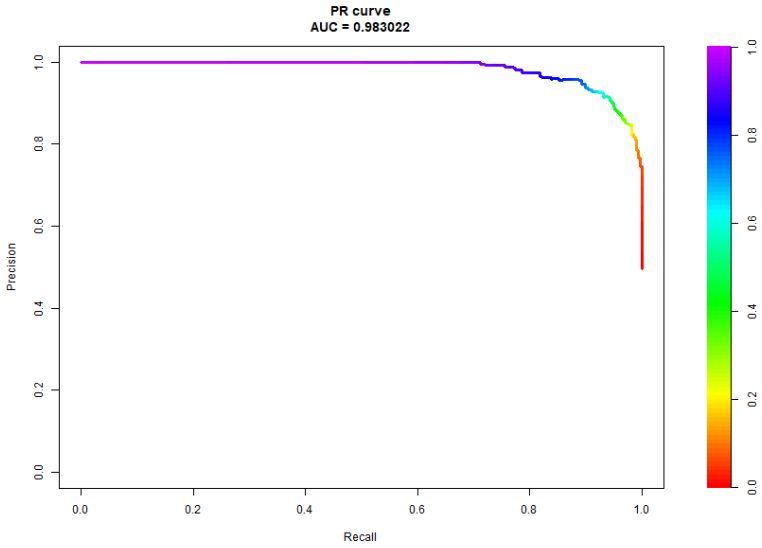



*Figura 9: Curva PR*

El óptimo se obtiene de un Recall y un Precision de 1 (esquina superior derecha del gráfico). Un modelo que no supera la aleatoriedad base es representado por una línea horizontal que corta el eje y en la probabilidad base de la clase minoritaria. De esta forma, el AUC PR de un modelo perfecto también es 1, al igual que en la curva ROC, pero un modelo que no supera la aleatoriedad base tiene un AUC PR igual a la probabilidad a priori de la clase minoritaria (= 1 – no information rate).

Se ha identificado que, en casos de clases desbalanceadas, el AUC PR es un mejor indicador de desempeño que el AUC ROC (Davis & Goadrich, 2006).

Como no todos los modelos y reglas de negocio entregan probabilidades (algunos entregan la clasificación ya determinada), es útil contar con una métrica para comparar modelos que se base en un corte ya definido. El F1-score es la

media armónica entre el indicador de Precision y Recall; y como tal, está relacionado a estrategias que buscan maximizar el AUC de la curva PR. El F1-score tiene valores reales entre 0 y 1.

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

Maximizar el F1-score implica asumir que el error tipo 1 y el error del tipo 2 tienen el mismo peso. Para casos donde esto no sucede, existe una generalización del F1-score que permite contabilizar ese desbalanceo.

Otra métrica interesante de desempeño global en clasificación binaria (que utilizan muchos algoritmos de forma tácita o predeterminada para calibrar) es el LogLoss.

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i * log_e(\hat{y}_i) + (1 - y_i) * log_e(1 - \hat{y}_i)\ ]$$

Con n la cantidad de registros, $y_i$ la clasificación real del registro (codificada como 0 o 1) y $\hat{y}_i$ la probabilidad estimada por el modelo de pertenecer a la clase 1. De esta forma LogLoss no solo penaliza una mala clasificación, si no que asigna mayor error si asignamos una probabilidad alta de que suceda el suceso equivocado.

## 3.6. Logit

El modelo Logit o de regresión logística es un modelo supervisado probabilístico por excelencia utilizado para clasificaciones binarias o extendido para clasificaciones múltiples. Se considera un modelo lineal generalizado y transforma la pertenencia o no a una clase (distribución de Bernoulli) a una probabilidad (distribución continua entre 0 y 1) mediante la función logística (ver Figura 10).

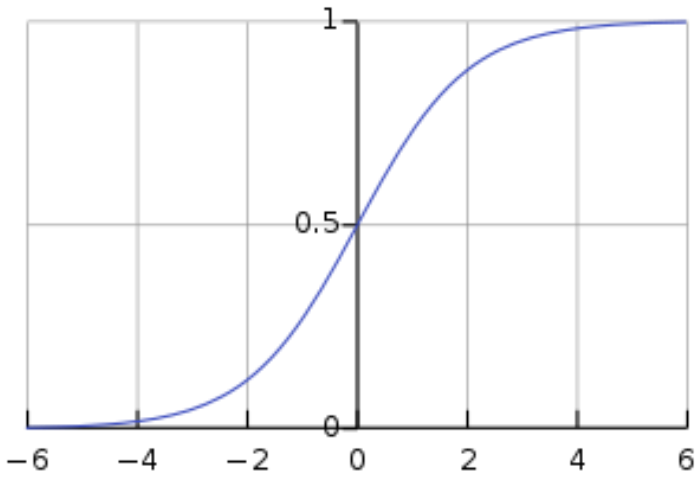



*Figura 10: Función logística (Wikipedia, s.f.)*

Un modelo logit se puede escribir de la forma

$$Logit: \mathbb{R} \longrightarrow (0,1)$$

$$Logit(\vec{x}) = \frac{1}{1 + e^{b_0 + \sum_i b_i * x_i}}$$

Con $x_i$ los atributos independientes y $b_i$ los coeficientes que debe encontrar el algoritmo para ajustar la data, típicamente maximizando el logaritmo de la verosimilitud mediante gradiente descendiente.

## 3.7. XGBoost

XGBoost o Extreme Gradient Boosting (Chen & Guestrin, 2016) es una técnica de modelamiento para modelos supervisados de clasificación y regresivos, basada en *gradient boosting* y que se caracteriza por su rápida convergencia y controles mediante regularización que evitan el sobreajuste (ver Figura 11). Desde su creación ha sido utilizado en competencias internacionales de Kaggle como *baseline model* debido a su rapidez y precisión.


*Figura 11: Características algoritmo XGBoost (Shikar, 2019)*

El *boosting* se basa en el principio de que muchos modelos simples y de rápido cómputo, típicamente árboles de decisión o regresiones lineales, utilizados en conjunto pueden generar un meta-modelo que reduce tanto el sesgo como la

varianza de los modelos originales, pudiendo competir con modelos mucho más complejos. A diferencia del *bagging*, que también se basa en varios modelos simples que aprenden de forma independiente (por ejemplo: random forest), el *boosting* se basa en un aprendizaje secuencial donde típicamente el próximo modelo trata de responder a los errores de los modelos anteriores, mejorando la predicción global.

El *gradient boosting* se basa en el principio de que el aprendizaje secuencial realizado por el algoritmo de *boosting* equivale a buscar la minimización del error mediante técnicas de gradiente descendiente o hessianos.

XGBoost posee varios hiperparámetros que deben ser ajustados para poder utilizarlo, dentro de los cuales destacan:

- Booster: si se utilizará árboles o regresiones lineales como modelo base.
- Objective: Indica la naturaleza del problema a calibrar. En el caso de clasificación binaria, se suele utilizar el parámetro “binary:logistic” para obtener directamente las probabilidades y no la clasificación de clases.
- Eval_metric: indicador objetivo a optimizar. Dependiendo de la naturaleza del problema, si es regresivo o de clasificación, se tienen múltiples alternativas. En clasificación binaria, se puede escoger entre logloss, error, auc (AUC ROC) y aucpr.

- Eta o Learning Rate: Posee valores entre 0 y 1. Representa un reescalamiento de la contribución al modelo global de cada modelo base. Valores más pequeños de eta implica que el meta-algoritmo debe ejecutar mayores iteraciones para poder converger.

- Subsample: Posee valores entre 0 y 1. Representa el porcentaje del dataset de training que es tomado para la calibración de cada modelo base. Este porcentaje es tomado de forma aleatoria en cada ejecución y permite no sobreajustarse a los datos.

- Colsample_bytree: Posee valores entre 0 y 1. Representa el porcentaje de los atributos del set de training que es utilizado en la calibración de cada modelo base. Al igual que Subsample, permite evitar el sobreajuste.

- Lambda: representa el coeficiente de regularización L2. Toma valores no negativos

- Alpha: representa el coeficiente de regularización L1. Toma valores no negativos

- Gamma: representa la mínima reducción de pérdida para hacer una partición más en una hoja de un árbol.

Si se utilizan árboles de decisión como modelo base, se puede evitar el paso de selección de tributos explícito, ya que los árboles tienen incorporados criterios de selección.

## 3.8. CatBoost

Desde la creación de XGBoost en 2016 han surgido diversos algoritmos basados en *boosting trees* y *gradient descent* que buscan competir en diversos aspectos con el algoritmo. LightGBM (Ke, 2017) busca tener aun mayor rapidez de ejecución sin sacrificar precisión y CatBoost (Prokhorenkova L. e., 2019) busca poder procesar variables categóricas de manera eficiente, evitando preprocesamiento de estas variables (XGBoost y LightGBM solo aceptan inputs numéricos). Existen numerosas comparaciones entre esos algoritmos tratando de justificar en qué contexto uno tiene ventajas sobre otros (Swalin, 2018). Lo cierto es que esa elección debe ser validada en base al problema a modelar.

CatBoost presenta una clara ventaja al momento de utilizarlo cuando la data posee atributos categóricos, ya que tiene incorporada una forma de encoding eficiente dividida en 2 partes:

- One Hot Encoding es evitado por el incremento de la dimensionalidad de la base, pero para atributos con pocas categorías puede ser más eficiente codificarlo con esta técnica. Existe como hiperparámetro el número máximo de categorías con que una variable será codificada como One Hot Encoding

- Para las categorías con alta cardinalidad, se procede a calcular el indicador numérico señalado en la Figura 12, en base a la variable target

y a los datos ya recorridos previamente por el algoritmo (para calcular el encoding del registro i-ésimo, se utiliza la data de las filas 1 a i-1).

$$\hat{x}_k^i = \frac{\sum_{\mathbf{x}_j \in \mathcal{D}_k} \mathbb{1}_{\{x_j^i = x_k^i\}} \cdot y_j + a\,p}{\sum_{\mathbf{x}_j \in \mathcal{D}_k} \mathbb{1}_{\{x_j^i = x_k^i\}} + a}$$

$$\mathcal{D}_k \subset \mathcal{D} \setminus \{\mathbf{x}_k\}$$

*Figura 12: Fórmula Encoding Catboost (Prokhorenkova L. e., 2019)*

Donde $\hat{x}_k^i$ es el encoding de la k-ésima categoría de la variable x en el registro i, el cual es calculado con $D_k$ la información de toda la data previamente recorrida que contiene la categoría k, excluyendo el registro actual que se quiere codificar. El parámetro “p” representa la probabilidad a priori del target y el parámetro “a” es un parámetro.

Una desventaja de esta forma de codificar es que una categoría de una variable puede tener diferentes valores de encoding en el mismo dataset, dificultando la interpretabilidad del resultado arrojado por el modelo.

## 3.9. Optimización de hiperparámetros

Existen varias formas de calibrar hiperparámetros de un modelo de machine learning. Algunas de éstas son (Ippolito, 2019):

1- Búsqueda manual (*Manual Search*): buscar valores de forma exploratoria manual y escoger la mejor combinación encontrada.

2- Búsqueda en grilla (*Grid Search*): fijar la región y la granularidad básica del espacio de los hiperparámetros donde se evaluará el modelo y realizar una búsqueda intensiva de todas las combinaciones posibles. Luego de esto, escoger la mejor combinación encontrada.

3- Búsqueda aleatoria (*Random Search*): buscar de forma aleatoria combinaciones de hiperparámetros y quedarse con la mejor combinación.

4- Búsqueda automática (*Automated Hyperparameter Tunning*): en esta familia se encuentran metodologías tales como Optimización Bayesiana de Hiperparámetros, Algoritmos Genéticos y métodos de optimización no lineal, tales como Nelder & Mead (Nelder & Mead, 1965).

La desventaja del primer método es una búsqueda no sistemática que no asegura explorar las regiones de hiperparámetros más prometedoras. Grid Search y Random Search son demandantes computacionalmente hablando y no aseguran tampoco explorar en más profundidad regiones de hiperparámetros con mejor desempeño, es más, no aprenden de las iteraciones pasadas por lo que pueden explorar regiones buenas y malas con la misma intensidad.

Sólo los métodos de búsqueda automática permiten indagar más en las regiones que van demostrando ser mejores. En este contexto, se utilizará una adaptación del método Nelder & Mead de optimización no lineal, el cual basa su búsqueda de hiperparámetros mediante un método simplex, minimizando el error en el espacio de hiperparámetros. La adaptación utilizada de este método se llama

Subplex (subspace-searching simplex method for the unconstrained optimization of general multivariate functions) (CRAN, 2020), la cual es una generalización del método Nelder & Mead para optimizar funciones ruidosas en altas dimensiones.

Un Simplex (Wikipedia, s.f.) en geometría es la generalización de lo que es un triángulo en un espacio $\mathbb{R}^n$, con n = 2, pero para cualquier $n \in \mathbb{N}$. Su nombre se deriva de que un Simplex es el más simple polítopo en cualquier espacio dado. Ejemplos:

- 0-simplex es un punto
- 1-simplex es un segmento
- 2-simplex es un triángulo
- 3-simplex es un tetraedro
- 4-simplex es un pentácoro (5-cell)

Algoritmo base (Singer & Nelder, 2009): En el contexto de calibración de hiperparámetros, es necesario que dichos hiperparámetros sean números reales. Si se está calibrando n hiperparámetros, el espacio dimensional de los hiperparámetros será $\mathbb{R}^n$. El método Nelder & Mead fue diseñado para minimizar una función no lineal $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ no restringida, sin intentar calcular el gradiente. En nuestro caso, la función f es la evaluación del error en un set de validación del modelo de machine learning calibrado en el set de entrenamiento. Y “n” es la cantidad de hiperparámetros a calibrar con este método.

El método se inicia con n+1 combinaciones de hiperparámetros distintas $x_0$,…, $x_n$ que representan un simplex en el espacio de hiperparámetros (ver ejemplo Figura 13).

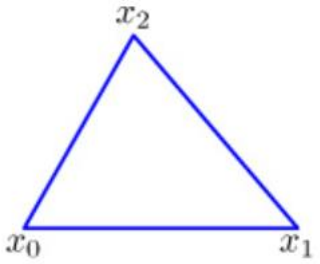



*Figura 13: Simplex en $\mathbb{R}^2$.*

El algoritmo consta de 3 etapas:

1- Orden: Evaluar cada vértice del simplex con la función f y ordenar los resultados obtenidos.

2- Centroides: calcular el centroide c del mejor lado del simplex (opuesto al peor vértice).

3- Transformación: Determina un nuevo Simplex tratando reemplazar el peor vértice del simplex anterior por un nuevo vértice mediante técnicas de reflexión (ver Figura 14), expansión (ver Figura 15) o contracción (ver Figura 16). Si lo logra, se acepta el nuevo punto como parte del simplex y se elimina el peor vértice anterior; y luego se itera sobre el nuevo simplex. Si no lo logra, se “encoge” (ver Figura 17) el simplex original, obteniendo también un nuevo simplex.

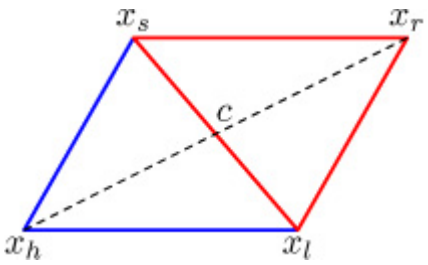



*Figura 14: Nelder and Mead Reflexión*

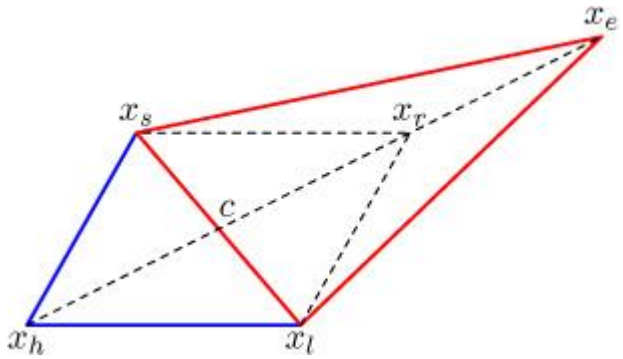



*Figura 15: Nelder and Mead Expansión*

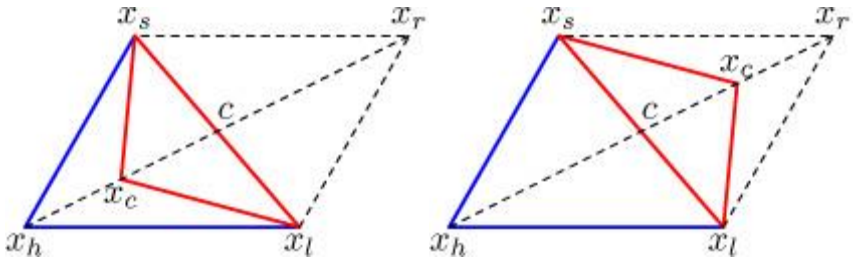



*Figura 16: Nelder and Mead Contracción Interna (izq.) y Contracción Externa (der.)*

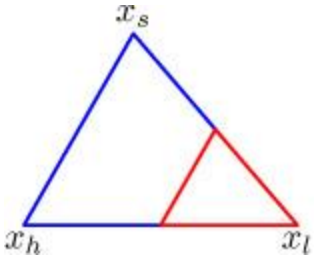



*Figura 17: Nelder and Mead Encogimiento.*

Existe un test para asegurar que el algoritmo converge en un número de iteraciones finitas. El óptimo se obtiene considerando el mejor vértice del simplex final una vez que el algoritmo converge.

## 3.10. Expected Calibration Error (ECE)

Existen diversas métricas para escoger un buen modelo de clasificación binaria. Independiente de la métrica y el modelo utilizado, existen formas de obtener las probabilidades registro a registro de pertenencia a cada una de las clases. Sin embargo, obtener como respuesta un número entre 0 y 1 no implica que se ha obtenido una probabilidad y en algunos casos es necesario tener una métrica adicional que represente qué tan bien calibradas están las probabilidades finales entregadas por el modelo, en especial si serán utilizadas como tal en la toma de decisiones posterior (Mazzanti, Python's «predict_proba» Doesn't Actually Predict Probabilities (and How to Fix It), 2021).

Una métrica útil es el Expected Calibration Error (ECE), el cual mide a nivel agregado qué tan lejos están las probabilidades entregadas por el modelo de las probabilidades observadas en un set de validación, ponderando por la cantidad de observaciones.

$$ECE = \frac{\sum_{b=1}^{B} |mean(y_b) - mean(proba_b)| \; x \; len(y_b)}{\sum_{b=1}^{B} len(y_b)}$$

Donde B son los bines en que se agrupan las predicciones. Típicamente estos bines son conformados por tramos equiespaciados de probabilidad (y por ende con cantidades de registros desbalanceados) o por tramos con igual cantidad de registros.

Un ECE cercano a cero implica que las probabilidades entregadas por el modelo son cercanas a las probabilidades reales, es decir que, si tomamos todos los registros que el modelo nos entrega con probabilidad en torno al 10%, 1 de cada 10 de esos registros presentarán la cualidad buscada. Un ECE bajo nos permite utilizar estas probabilidades como tal, en especial para calcular valores esperados de utilidad.

## 3.11. Shapley Values

Dentro de la línea de investigación de modelos, cada vez se han desarrollado modelos más complejos que permiten obtener mejores ajustes a los datos, permitiendo modelar relaciones tan complejas que la mente humana no siempre es capaz de asimilar. Dentro de este contexto, parece haber cierto compromiso entre un modelo preciso y uno interpretable. Hoy en día, un modelo preciso (CNN, XGBoost, etc.) no es fácil de interpretar, es decir, conocer cómo incide en la decisión del modelo cada uno de las variables independientes originalmente suministradas para la calibración. Por otro lado, modelos interpretables (árboles de decisión, regresiones lineales, etc.), donde podemos obtener directamente la

incidencia de cada atributo en la decisión final, suelen no ser tan precisos en casos complejos.

Cada vez es más necesario entender por qué un modelo toma una decisión. Esto va desde que la persona que está modelando la realidad pueda ver si el modelo toma decisiones razonables (podría indicar sesgo en la data las decisiones que toma), hasta romper la opacidad del modelo por temas de discriminación (que generalmente pueden estar presente en los datos históricos y verse perpetuados por un modelo “caja negra”).

Dentro de la línea de mejorar la interpretabilidad, se ha trabajado en 2 frentes: interpretación a posteriori de un modelo complejo, donde técnicas como LIME (Hulstaert, 2018) y SHAP Values (Lundberg & Lee, 2017) caen en estas categorías; y generar modelos de naturaleza interpretable que permitan ajustarse a problemas complejos (Kübler, 2021).

SHAP values (Shapley Additive Explanations) tiene su origen en la teoría de juegos y es la última tendencia en explicabilidad de modelos. Cabe destacar que lo entregado por SHAP values es una interpretación del modelo (cómo el modelo toma decisiones), no de la realidad representada en la data (no hace juicio de valor sobre si el modelo está bien o no ajustado a la realidad).

SHAP Values intenta explicar el impacto de tener cierto valor para algún atributo dado en la predicción final del modelo. Esto manteniendo 2 condiciones que en

general fallan los métodos tradicionales de *feature importance* (como indicadores de *weight*, *cover* o *gain*) (Lundberg S. , 2018):

1- Consistencia: si modificamos el modelo y el nuevo modelo se basa más fuertemente en un atributo particular, la importancia de ese atributo no debería decrecer.

2- Precisión: la suma de las importancias de todos los atributos debería sumar la importancia del modelo completo.

Para esto, SHAP se basa en el conjunto potencia de los atributos utilizados por un modelo calibrado (ver ejemplo en Figura 18). Se calibran los $2^n$ modelos parciales, donde n es el número de atributos del modelo final, conservando los mismos hiperparámetros (no se realizan ajustes de ningún tipo al modelo encontrado, es el mismo modelo reejecutado en subconjuntos distintos de los atributos originales).

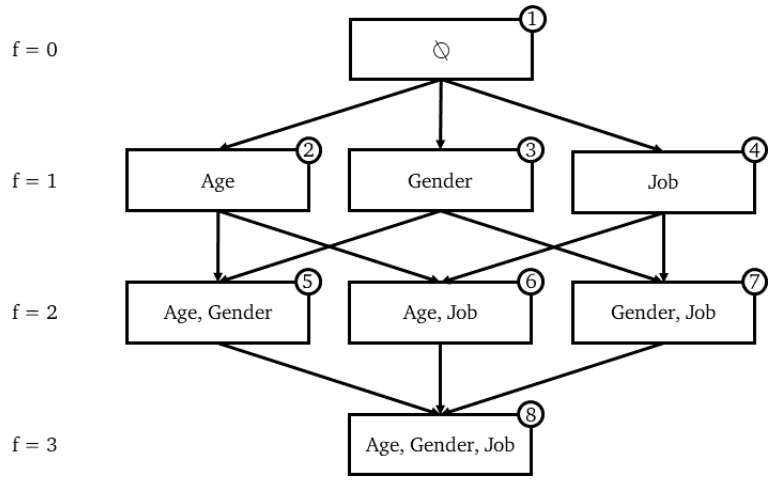
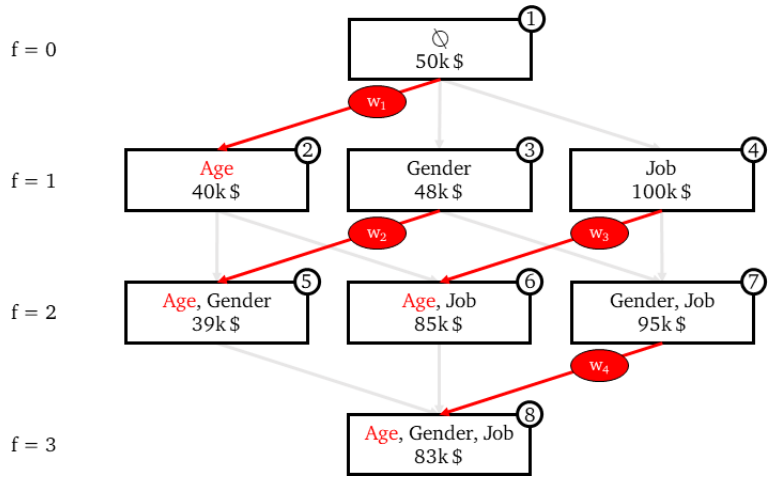



*Figura 18: Conjunto Potencia de los atributos utilizados en un modelo (Mazzanti, SHAP values explained exactly how you wished someone explained to you, 2020)*

De esta forma, para ver la relevancia de un atributo en particular, se pondera la participación de ese atributo en cada uno de los modelos encontrados en el conjunto potencia (ver Figura 19), de forma tal que:

- Todos los pesos asociados a un atributo en particular sumen 1
- Todos los pesos asociados a un atributo particular en el mismo nivel de conjunto potencia (con la misma cardinalidad de atributos) tengan igual peso.




$$SHAP_{Age}(x_0) = w_1 \times MC_{Age,\{Age\}}(x_0) + w_2 \times MC_{Age,\{Age,Gender\}}(x_0) + w_3 \times MC_{Age,\{Age,Job\}}(x_0) + w_4 \times MC_{Age,\{Age,Gender,Job\}}(x_0)$$

*Figura 19: Contribución Marginal de un Atributo en el Modelo Final (Mazzanti, SHAP values explained exactly how you wished someone explained to you, 2020)*

Esta metodología identifica la contribución marginal de cada atributo por sobre un modelo base (modelo sin atributos), respetando la consistencia y la precisión.

Esto permite identificar interpretabilidad global de un modelo (ver ejemplo en Figura 24), como además identificar interpretabilidad local o registro a registro (ver ejemplo en Figura 25).

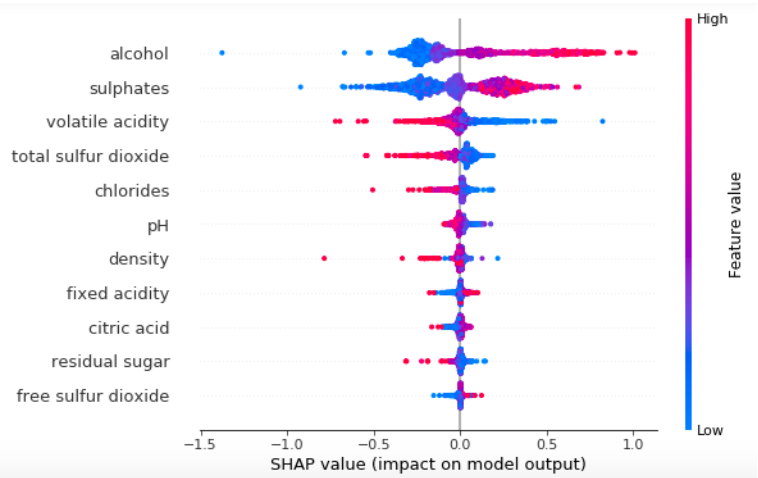



*Figura 20: SHAP Interpretabilidad Global (Dataman, 2019)*

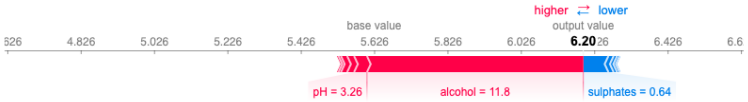



*Figura 21: SHAP Interpretabilidad Local (Dataman, 2019)*

# 4. Metodología

## 4.1. Metodología General

El desarrollo de este trabajo se enmarca en la metodología KDD como desarrollo secuencial en el descubrimiento de patrones en los datos de transformación de microempresas a PyME y posterior análisis explicativo de los patrones encontrados.

Se evaluarán 3 tipologías de modelos de clasificación binaria más una regla de negocio (ver Tabla 5):

1- **Logit**: para poder calibrarlo, será necesario realizar selección de atributos (para evitar multicolinealidad) y encoding de las variables categóricas (el algoritmo solo recibe inputs numéricos). No es necesario calibrar hiperparámetros (no posee hiperparámetros en su formulación básica).

2- **XGBoost**: para poder calibrarlo es necesario realizar un encoding de las variables categóricas (el algoritmo solo recibe inputs numéricos), pero los modelos de árboles basales son un método “wrapper” de selección de atributos, por lo que se pueden mantener los atributos iniciales. Para calibrarlo será necesario ajustar hiperparámetros.

**3-** **Catboost**: para poder calibrarlo, no es necesario seleccionar atributos (se basa en árboles y la selección de atributos se realiza de forma inherente al calibrar el modelo), ni tampoco realizar *encoding* (CatBoost maneja un *encoding* interno propio para variables categóricas). Es necesario calibrar hiperparámetros.

Estos modelos se contrastarán con una regla de negocio para determinar potenciales mejoras.

| Modelo | Selección Atributos | Encoding Variables Categóricas | Calibración Hiperparámetros | Precisión Esperada | Complejidad Modelo | Interpretabilidad Resultado |
|---|---|---|---|---|---|---|
| **Regla de Negocio** | Si, se filtra 1 atributo | No | No | Baja | Baja | Alta |
| **Logit** | Si, se filtran varios atributos | Si | No | Media | Media | Alta |
| **XGBoost** | No explicitamente. Método wrapper | Si | Si | Alta | Alta | Baja |
| **CatBoost** | No explicitamente. Método wrapper | No explícitamente. Calibración interna | Si | Alta | Alta | Muy Baja. Incluso con SHAP Values no es posible interpretar el encoding de forma sencilla. |

*Tabla 5: Cuadro Comparativo Estrategias de Modelamiento*

Para determinar el modelo más preciso se utilizará como indicador el área bajo la curva precisión-recall (AUC PR por sus siglas en inglés). Esta métrica nos permitirá seleccionar qué modelo tiene un mejor desempeño ajustándose a la data (selección de modelo y calibración de hiperparámetros). Para esto, el entregable del modelo deben ser las propensiones y no la clasificación.

Sin embargo, la regla de negocio al no contar con probabilidades, no se puede calcular la curva PR. Se definirá la ganancia en términos de F1-score del mejor modelo escogido contra la regla de negocio. Para determinar el F1-score de un modelo que entrega probabilidades, nos quedaremos con el mejor F1-score que pueda entregar el modelo en un set de validación.

Una vez se seleccione el mejor modelo, se debe determinar qué corte en propensión se ajusta mejor al desbalanceo en costos de la matriz de costo beneficio (donde tendremos posiblemente un peso distinto para el error de tipo I y el error de tipo II). El objetivo final es maximizar el beneficio esperado de

clasificar correctamente a un cliente versus los costos de clasificarlo incorrectamente. La matriz de costo beneficio depende de la aplicación que buscamos hacer del modelo. Un mismo modelo puede ser utilizado de formas muy distintas y tener de forma inherente costos asociados a sus beneficios y errores muy distintos. Esto modificará la probabilidad de corte sobre la cual se considerará que un cliente se convertirá en PyME a los 3 años, pero no cambiará los indicadores globales de desempeño (AUC PR). Por ejemplo, pensando en si el modelo determina de forma errónea que una empresa se convertirá en PyME en 3 años, es muy distinto el costo comercial de enviar un mail invitándolo a enviar sus datos para una evaluación comercial en el ingreso a un banco, que directamente hacer campaña con oferta asociada de preaprobado. Es por esto que esta última etapa quedará disociada de la elección del modelo ganador y se verá en un capítulo aparte.

## 4.2. Limpieza de datos

El primer paso fue estandarizar las diversas categorías presentes en la base. Algunas de ellas, como los rubros o las comunas, presentan diferencias a través de los años en su formato, a pesar de ser la misma categoría. Se estandarizó la nomenclatura de los rubros, subrubros y actividades económicas homologándola a la información encontrada en la página de actividades económicas del SII (Servicio de Impuestos Internos, 2021), incorporando información adicional sobre si la empresa tributa en primera o segunda categoría y si está o no afecta a IVA

(en base a la actividad económica principal). También se homologaron los nombres de Comunas, Provincias y Regiones.

El segundo paso fue analizar fechas fuera de formato, las cuales fueron cotejadas mirando por rut distintos periodos y homologando a la más probable. La fecha importante a considerar es la de la primera inscripción de actividades para saber cuándo la empresa fue creada.

## 4.3. Estadística descriptiva

Como primer análisis descriptivo, se observó la incidencia de estar afecto o no a IVA (ver Tabla 6) y la Categoría Tributaria (ver Tabla 7) en el Tramo de Ventas. Para esto se observó la información al año comercial 2019.

| Empresas 2019 | | | | |
|---|---|---|---|---|
| | SI | NO | G | Total general |
| 1 | 98.361 | 42.885 | 47.512 | 188.758 |
| 2 | 70.137 | 7.617 | 25.187 | 102.941 |
| 3 | 48.188 | 6.450 | 20.158 | 74.796 |
| 4 | 78.629 | 12.196 | 31.814 | 122.639 |
| 5 | 40.612 | 6.812 | 14.052 | 61.476 |
| 6 | 31.255 | 4.558 | 8.489 | 44.302 |
| 7 | 28.597 | 3.304 | 5.807 | 37.708 |
| 8 | 12.952 | 1.551 | 2.204 | 16.707 |
| 9 | 7.632 | 1.046 | 1.356 | 10.034 |
| 10 | 4.555 | 660 | 849 | 6.064 |
| 11 | 3.705 | 578 | 695 | 4.978 |
| 12 | 888 | 176 | 127 | 1.191 |
| 13 | 1.748 | 336 | 169 | 2.253 |
| **Total general** | **427.259** | **88.169** | **158.419** | **673.847** |

*Tabla 6: Tramo de Ventas por Afecta a IVA, año comercial 2019*

| Empresas 2019 | | | | |
|---|---|---|---|---|
| | 1 | 2 | G | Total general |
| 1 | 170.510 | 4.845 | 13.403 | 188.758 |
| 2 | 91.903 | 1.393 | 9.645 | 102.941 |
| 3 | 66.560 | 1.634 | 6.602 | 74.796 |
| 4 | 108.390 | 4.684 | 9.565 | 122.639 |
| 5 | 54.369 | 3.324 | 3.783 | 61.476 |
| 6 | 39.877 | 2.103 | 2.322 | 44.302 |
| 7 | 35.052 | 1.138 | 1.518 | 37.708 |
| 8 | 15.785 | 382 | 540 | 16.707 |
| 9 | 9.557 | 210 | 267 | 10.034 |
| 10 | 5.838 | 87 | 139 | 6.064 |
| 11 | 4.812 | 63 | 103 | 4.978 |
| 12 | 1.163 | 9 | 19 | 1.191 |
| 13 | 2.216 | 18 | 19 | 2.253 |
| **Total general** | **606.032** | **19.890** | **47.925** | **673.847** |

*Tabla 7: Tramo de Ventas por Categoría Tributaria, año comercial 2019*

El Tramo de Ventas es un indicador de ingresos, no solo de ventas y no se ve afectado a simple vista por si la empresa está afecta a IVA o por la Categoría Tributaria. Algunos ejemplos de actividades económicas que no están afectas a IVA y que tributan en segunda categoría son: servicios médicos prestados de forma independiente, servicios de asesoramiento y representación jurídica, actividades de contabilidad, teneduría de libros y auditoría, consultoría fiscal, etc.

En términos del evolutivo de personas jurídicas, éstas se han incrementado a través de los últimos años, en especial en los segmentos más bajos de microempresas, ocasionando que la proporción de empresas Pyme y Grandes Empresas disminuya en la base total (ver Figura 22) hasta llegar a ser cerca de un 20% de la base de personas jurídicas.

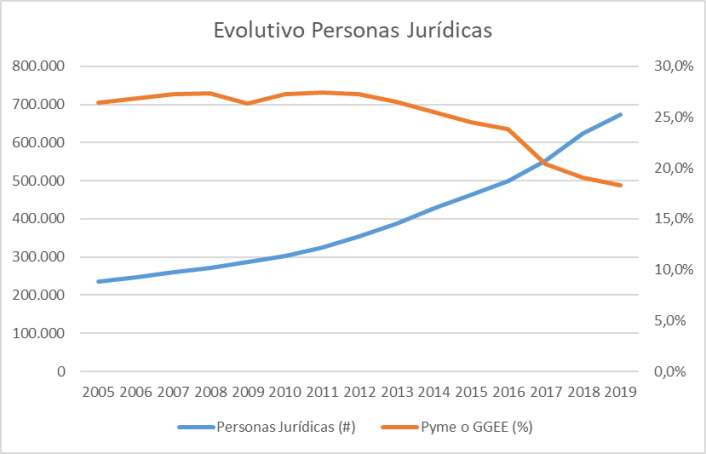



*Figura 22: Evolutivo Personas Jurídicas*

En términos de generación de empleo, se ve que los tramos de PyME y Grandes Empresas representan el 81,4% de los empleos a pesar de ser el 20% de la base (ver Figura 23).

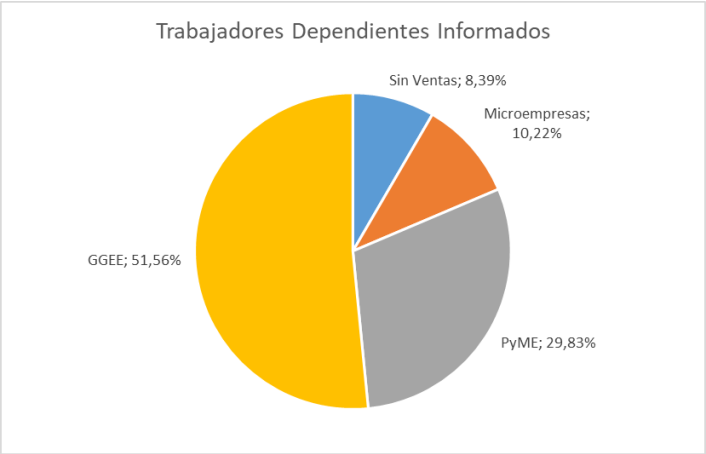



*Figura 23: Trabajadores dependientes Informados por Segmento, año comercial 2019*

El número de trabajadores dependiente pareciera ser a priori una buena variable a considerar para ver el tamaño de una empresa. Una microempresa del tramo de ventas alto tiene en promedio 5,8 trabajadores dependientes, mientras que el tramo más bajo en una pequeña empresa tiene en promedio 10,6 (ver Figura 24). Habrá que ver si esa relación de mayor empleo, mayor tamaño, se puede observar desde un inicio o es algo que se va ocasionando de forma paulatina a medida que evoluciona la empresa.

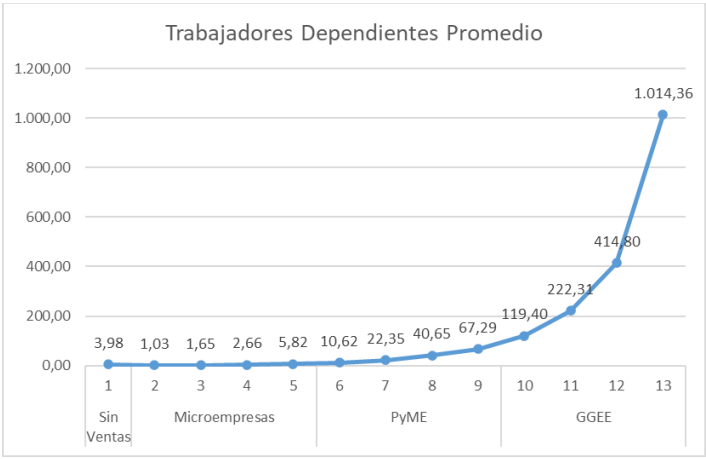



*Figura 24: Trabajadores Dependientes Promedio por Tramo de Ventas, año comercial 2019*

La Ubicación geográfica también parece indicar cierta segregación sobre donde se ubica una mayor concentración de empresas PyME o GGEE. Como se puede ver en la Figura 26, la región de Tarapacá (22,5%), la Metropolitana (19,7%) y la de Antofagasta (19,4%) concentran una alta tasa de empresas que pasan el umbral de las UF 2.400 de venta anual, mientras que otras regiones como la región de Aysén (11,8%) o la de los Ríos (13,5%) presentan una tasa menor. Nuevamente hay que evaluar si esta tendencia es algo establecido a priori o si la

geografía o las distintas administraciones regionales permiten una mayor o menor tasa de crecimiento en las empresas.

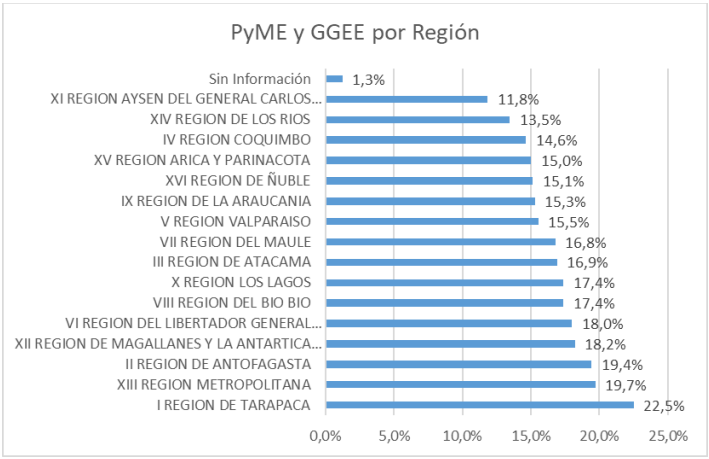



*Figura 25: Tasa de PyME y GGEE por Región*

Esta segregación espacial se puede observar también a nivel de provincia y comuna donde, por ejemplo, dentro de la Región Metropolitana tenemos comunas con tasas de PyME y GGEE de un 28%, como es el caso de Quilicura o Huechuraba, mientras que otras comunas, como Puente Alto y Lo Prado rondan en torno a un 10%.

En el caso de los Rubros, Subrubros y Actividades Económicas se puede apreciar un efecto similar al de la ubicación geográfica (ver Figura 27). Rubros como suministros de electricidad, gas, vapor y aire acondicionado presentan tasa de un 28,6%, mientras que otros como “actividades de los hogares” pueden llegar a 0%. En términos de actividades económicas, existen actividades con un 100%, tales como AFP, cajas de compensación y televisión de pago por cable; mientras

otras se quedan estancadas en un 0%, tales como pesca en agua dulce, extracción y procesamiento de litio y fabricación de tejidos de punto y ganchillo. Si bien es cierto que esta es una dispersión bastante amplia, no es todos los días que una nueva AFP o caja de compensación entra al mercado, por lo que parte de esta información podría no incidir necesariamente en las nuevas empresas que se estén generando como un atributo diferenciador.

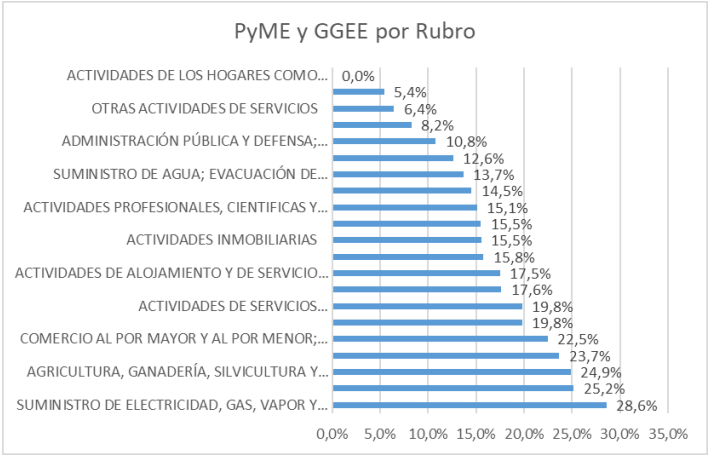



*Figura 26: Tasa de Pyme y GGEE por Rubro, año comercial 2019*

El tramo de capital propio parece también incidir en la tasa de PyME y GGEE. En este caso, la relación no es lineal, ya que altos tramos de capital propio negativo o positivo parecen estar relacionados con mayores tamaños de venta (ver Figura 27).

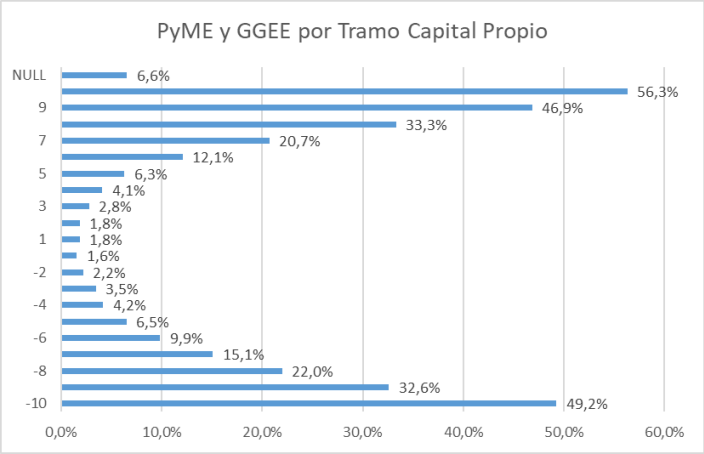



*Figura 27: Tasa de PyME y GGEE por Tramo de Capital Propio*

El tipo de contribuyente también parece incidir en la tasa de PyME y GGEE, pero dada la concentración de los contribuyentes en personas jurídicas comerciales, no parece ser un atributo muy decidor en la mayoría de los casos (ver Figura 28).

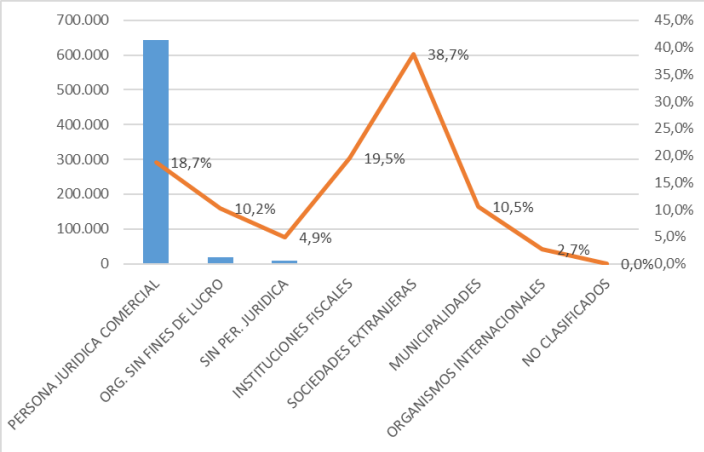



*Figura 28: Tasa de PyME y GGEE por Tipo de Contribuyente*

## 4.4. Planteamiento inicial del modelo

Se procede a estimar lo más cercano a la información de la generación de la empresa si ésta podrá alcanzar niveles de PyME o GGEE. Para esto se tomó la información disponible en el mismo año del primer inicio de actividades. Esto nos permitirá inferir con la primera información disponible la proyección de la empresa. En el año de inicio de actividades, parte de la información aún no se encuentra regularizada (por ejemplo, en algunos casos no está la información de la ubicación geográfica) y existe una inferencia en el tramo de ventas debido a la parcialidad de la información disponible al momento del cierre del año comercial. El tramo de ventas del primer año se puede ver fuertemente influenciado por si la empresa inició actividades en enero o en diciembre. El algoritmo del SII extrapola el tramo de ventas (no existe una diferencia significativa de la distribución de tramos en base al mes de inicio de actividades). Debido a este último punto, se incorporará el mes de inicio de actividades como una variable independiente. Potencialmente, esa variable podrá explicar diferencias en el cálculo del tramo de ventas, además de diferenciar parcialmente el horizonte de evaluación. Por ejemplo, una empresa evaluada a 5 años de su inicio de actividades y que apertura en enero, tiene en realidad 6 años de ventas. Si la misma empresa hubiera abierto en diciembre, tendría solo 5 años de ventas al cabo de 5 años.

Esta decisión posee ventajas y desventajas: por un lado, tenemos información más oportuna, pero por otro podríamos tener información incompleta o menos fidedigna.

Para entender el desfase, una empresa que inicia actividades entre enero a diciembre del 2015, aparecerá en la base año comercial 2015 (año tributario 2016) en el transcurso del 2016. La utilidad del modelo dependerá de cuanto demora una empresa en promedio en crecer y superar las UF 2.400 de venta anual.

De este modo, nuestra base tentativa para construir el modelo se reduce a 710.371 empresas que podemos tener la información en el año del primer inicio de actividades. De estas empresas, 36.459 ya son PyME o GGEE según el tramo de venta del primer año y no entrarían en el análisis a realizar. De las 673.912 empresas restantes, 552.741 (82,02%) nunca llegan a superar la barrera de las UF 2.400 en el horizonte observado. Las que si superan esa barrera un 81% logran convertirse en PyME o GGEE dentro de los primeros 3 años luego del inicio de actividades (ver Figura 29). Luego de eso, las tasas de conversión son cada vez menores.

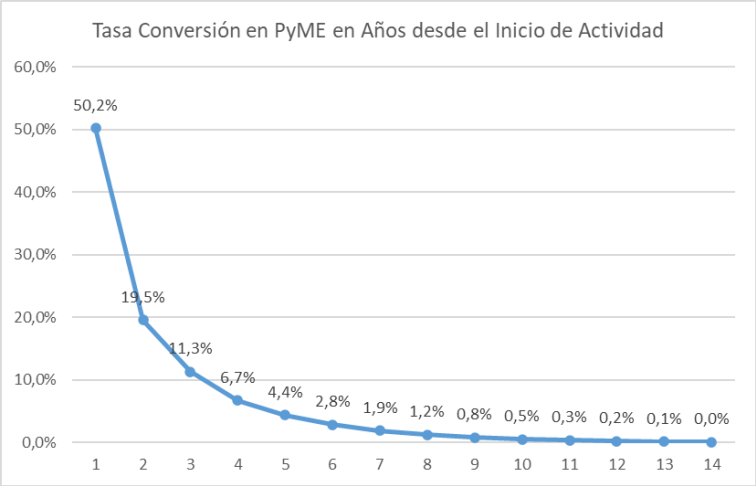



*Figura 29: Tasa Conversión en PyME en Años luego del inicio de actividades sobre el total de empresas que alguna vez se convierte en PyME.*

Se utilizará el tramo de venta del tercer año desde el inicio de actividades para construir la variable objetivo a predecir por el modelo. El modelo propuesto es un modelo de clasificación binaria (si al año 3 una empresa logra vender más de UF 2.400 anuales o no).

La elección de un horizonte más largo permitirá tener mayor certeza sobre el diagnóstico (más empresas lograrían eventualmente convertirse en PyME), pero tendríamos cada vez menos datos para calibrar (menos empresas se pueden evaluar en un horizonte más largo debido a que la ventana de tiempo de datos disponible es limitada) y esto puede ocasionar un peor modelo. Por ejemplo, si el horizonte utilizado fuera 9 años, podríamos considerar el 98,8% de las empresas que eventualmente evolucionan en PyME. Pero para evaluar 9 años de trayectoria, teniendo información disponible del 2005 al 2019, solo podría llegar

a evaluar a empresas con inicio de actividades entre el año 2005 y 2010, ya que evaluaríamos su evolución final al año 2019 para poder concluir. Este modelo, además de contar con menos data para calibrarse, tiene un desfase importante en considerar las nuevas tendencias en la generación de empresas (actividades económicas que tienen un mayor desarrollo en los últimos años no estarían analizadas en el modelo). Con horizontes largos, también se empieza a hacer presente un fenómeno que no se pretende abordar en este análisis: ¿qué sucede si en esos 9 años una empresa logra convertirse en PyME y luego quiebra? Al evaluar al año 9 la empresa ya no estará vendiendo, pero en el transcurso de los 9 años si logró superar la barrera de las UF 2.400 anuales. ¿Cómo se considera este caso? Como simplificación del análisis, al considerar el horizonte de 3 años, no consideraremos las oscilaciones de años dentro de la ventana de tiempo, solo se considerará la evaluación al año 3 del tramo de venta como un diagnóstico verídico y perdurable del cambio.

Por otro lado, un horizonte más corto permite utilizar más data para calibrar, pero al tener muy poco tiempo para que una empresa evolucione en PyME puede erróneamente clasificar como una empresa que permanece como microempresa a alguien que simplemente no se le dio el tiempo suficiente para evolucionar. Otra desventaja de utilizar ventanas de tiempo muy cortas, es que el desfase inherente que posee la data hace rápidamente obsoleta la toma de decisiones. En el extremo, tomando solo 1 año de desfase, el 50,2% de las empresas que logran evolucionar a PyME serían detectadas. En el ejemplo de la empresa que inicia

actividades el 2015, su información se encontrará disponible en el transcurso del 2016; lograríamos predecir sobre ese 50,2% si se convierten en PyME al finalizar ese mismo año 2016, lo cual deja poco margen de acción comercial y una empresa con acceso más actualizado a información de ventas ni siquiera tiene la necesidad de realizar esa inferencia.

Un análisis complementario interesante sería ver de forma diferenciada si algún factor, como la actividad económica, incide en los años que demora una empresa en convertirse en PyME. Este es un modelo distinto al planteado, posiblemente abordable mediante modelos regresivos.

La información preprocesada para utilizarla en el modelo cuenta con 404.314 empresas con inicio de actividades entre el 2005 y el 2016; y con una evaluación en si la empresa se convierte en PyME o superior en un horizonte de 3 años. El límite superior en el año comercial 2016 existe porque debemos ser capaces de ver la evolución de las empresas 3 años posterior a la base inicial. En el extremo, las empresas con inicio de actividades 2016 debemos determinar si se convierten en PyME el 2019, que es la última información disponible. En dicha base, un 21,53% de las empresas logran convertirse en PyME al tercer año.

Las variables independientes utilizadas en el modelo son inicialmente:

- Número de trabajadores
- Tipo de Contribuyente

- Rubro
- Subrubro
- Actividad Económica
- Afecto a IVA
- Categoría Tributaria
- Región
- Provincia
- Comuna
- Tramo de Venta
- Tramo de Capital Propio
- Mes del inicio de actividades

Salvo el número de empleados, que es inherentemente un atributo numérico, todos los demás atributos son categorías ordenadas o no ordenadas.

El mes de inicio de actividades fue considerado debido a dos razones ya mencionadas anteriormente. La primera es que existe una diferencia en el primer año tributario cuando el inicio de actividades ocurre en enero o en el transcurso del año. Si la empresa fue creada en enero, el tramo de ventas corresponde a las ventas del año completo. Si la empresa fue creada en el transcurso del año (diciembre en un caso extremo), el tramo de ventas correspondería

presumiblemente a una mezcla entre información real e información extrapolada por el SII con criterio desconocido (esto por no haber diferencias significativas en la distribución de los tramos de ventas en base al mes de apertura). La segunda razón es que el horizonte de evaluación de 3 años para convertirse en PyME es muy distinto para una empresa que inició actividades en enero que en diciembre. En el segundo caso, el horizonte de evaluación es más cercano a los 2 años que a los 3, dejando menos tiempo para evaluar su crecimiento. Debido a esto, se consideró el mes de inicio de actividades como numérico (con el correlativo de que enero es el mes 1 y diciembre el mes 12) ya que solo lo incluimos como forma de discriminar la proporción de meses de existencia de la empresa en el año de inicio de actividades.

Como los modelos de machine learning en general necesitan de input valores numéricos, es necesario en general preprocesar esta información con un encoding apropiado que permita a estos modelos utilizar la data proporcionada.

## 4.5. Missing Values

El único atributo que presenta missings es el tramo de capital propio. A ser una variable categórica, se trabajará este missing value generando la categoría adicional “Sin Información” y dejando que la estrategia de encoding la maneje como una categoría más.

## 4.6. Outliers

Dado que la mayoría de los atributos son categóricos (salvo el número de empleados), la naturaleza de los outliers puede yacer en la combinación poco probable de varias categorías, no en el análisis univariado. El foco de este estudio se basará en encodings y técnicas de árboles que pueden manejar bien el concepto de outliers sin necesidad de realizar un tratamiento previo.

## 4.7. Split

La data fue dividida en 3 grupos: Training (60%), Validation (20%) y Test (20%). La división se hizo de forma estratificada de modo que cada set tenga la misma proporción de los casos positivos de la clasificación (21,53%).

## 4.8. Encoding y Selección de Atributos

El *encoding* de los atributos categóricos se realizó por medio del WoE Encoding calibrado en la base de entrenamiento. Este *encoding* será utilizado por el modelo Logit y los XGBoost (CatBoost posee un encoding propio).

La selección de atributos será utilizada solo para el modelo Logit, dado que las técnicas basadas en árboles tienen incorporada la selección de atributos.

El primer filtro de selección de atributos fue basado en Information Value (IV). Existen 4 atributos que no superan el mínimo asociado a su IV para ser considerado útil para predecir (ver Figura 30): Región, Provincia, Afecto a IVA y Categoría Tributaria. Cabe destacar que este es un análisis univariado y que no contempla las diversas interacciones que pueden existir entre los atributos para poder explicar la variable a predecir. Sin embargo, la variable Comuna está estrechamente relacionada con Provincia y Región (hay una dependencia jerárquica), por lo que es fácil asumir que portan información redundante. Del mismo modo, si está afecto o no a IVA y la categoría tributaria dependen de la Actividad Económica y se puede inferir en base a ésta. Por estas razones, se eliminarán estas 4 variables del modelo Logit.

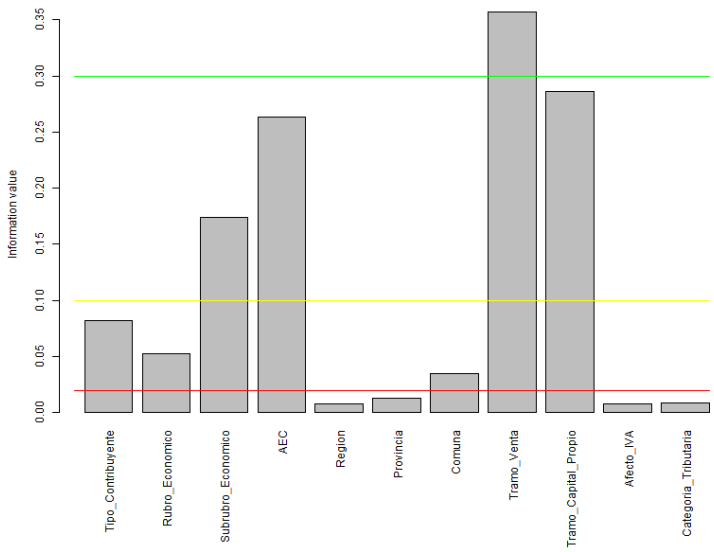



*Figura 30: Information Value variables Modelo Emprendedor*

Otra forma de ver relevancia, así como interdependencia (lineal) entre las variables dependientes e independientes, es mediante un correlograma (ver Figura 31).

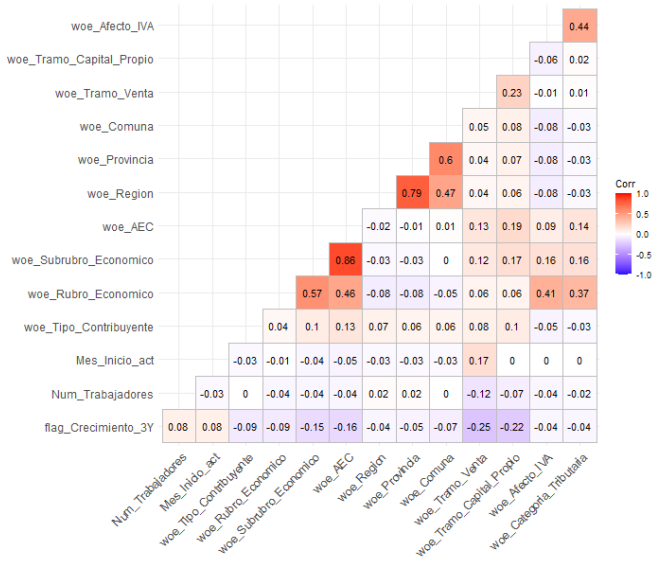



*Figura 31: Correlograma*

Se puede apreciar la estrecha relación entre Actividad Económica, Subrubro Económico y Rubro Económico, lo cual es lógico dada la dependencia jerárquica. Dependiendo de la técnica de modelamiento utilizada, esta dependencia puede ser un problema y debe ser manejada. Por ejemplo, con regresiones lineales la multicolinearidad implica que los coeficientes asociados a los atributos independientes no pueden ser interpretados como relevancia del atributo, y en algunos casos más graves, puede afectar el hallazgo de una solución al implicar diversas soluciones posibles. Técnicas como árboles de decisión son menos

sensibles a estas multicolinearidades. De las 3 variables mencionadas, la Actividad Económica es la con mayor IV y mayor correlación de Pearson con la variable dependiente. Dejaremos de estas 3 solo la Actividad Económica para el modelo Logit.

La variable independiente que más se relaciona con la variable objetivo es el tramo de ventas al inicio de los 3 años de medición (-0.25) seguida por el tramo de capital propio (-0.22) y la actividad económica (-0.16). Es esperable que el modelo que encontremos considere estos atributos con relevancias semejantes.

El que las correlaciones de la variable dependiente con las independientes codificadas con WoE sea negativo no implica necesariamente una relación inversa entre una y otra variable (para aquellas que hace sentido un orden en las categorías). Por ejemplo, el tramo de ventas tiene una correlación de -0.25, lo cual a priori podría insinuar una relación negativa entre el nivel inicial de ventas y la capacidad de la empresa en convertirse en PyME a los 3 años. Esto parece contraintuitivo, pero cuando vemos cómo fue codificado con el WoE, tenemos:

- Tramo 1 -> WoE 0.06025589
- Tramo 2 -> WoE 0.82827760
- Tramo 3 -> WoE 0.07085144
- Tramo 4 -> WoE -0.86343314

Dejando de lado el Tramo 1, que representa tanto empresas sin ventas como aquellas que aún no cuentan con información suficiente para su cálculo, los tramos 2, 3 y 4 tienen un WoE decreciente. Es decir, a mayor nivel de ventas inicial, menor codificación vía WoE. Una correlación negativa entre el WoE del tramo de ventas y la variable objetivo implica en verdad que, a mayor tramo de ventas inicial, mayor es la probabilidad de convertirse en PyME a los 3 años.

## 4.9. Modelo Base: Regla de Negocio

El modelo base a utilizar será una regla de negocio. Como el tramo de ventas al inicio de los 3 años es el atributo qué más incide en la tasa de conversión a PyME al final de esos 3 años (de acuerdo al correlograma de la Figura 31), cabe pensar que considerar a aquellas empresas que en el año de inicio de actividades ya se encuentran ya vendiendo en el tramo más alto de microempresas (entre 600,01 y 2400 UF de venta anual) tengan la mejor proyección de crecimiento. Dicho en otras palabras, se considerará el grupo con el tramo de venta 4 (ver Tabla 2) como el que tiene mayor probabilidad a pasar a tener un tramo de venta 5 o superior al cabo de los 3 años.

Esta regla de negocio tiene un F1 score de 38,7% (ver Tabla 8).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 53091 10753
         1 10355  6665

               Accuracy : 0.739
                 95% CI : (0.7359, 0.742)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 1.000000

                  Kappa : 0.2213

 Mcnemar's Test P-Value : 0.006285

            Sensitivity : 0.38265
            Specificity : 0.83679
         Pos Pred Value : 0.39160
         Neg Pred Value : 0.83157
              Precision : 0.39160
                 Recall : 0.38265
                     F1 : 0.38707
             Prevalence : 0.21540
         Detection Rate : 0.08242
   Detection Prevalence : 0.21048
      Balanced Accuracy : 0.60972

       'Positive' Class : 1
```

*Tabla 8: Matriz de Confusión e Indicadores Regla de Negocio*

Se puede apreciar a simple vista que la regla de negocio no tiene un mal accuracy (73,9%, principalmente debido al desbalanceo de clases), pero al ver el Sensitivity o Recall, solo logra detectar el 38,2 % de los casos que efectivamente se convierten en PyME. También posee un Precision de un 39,1%, lo cual indica que de los que la regla de negocio indica que se convertirán en PyME, solo un 39,1% realmente lo hace.

## 4.10. Modelo 1: Logit

Para poder calibrar el modelo Logit recurrimos a la selección de atributos vía filtros más el encoding vía WoE.

De esta forma, se obtuvo un modelo Logit con todos los regresores significativos (ver Tabla 9). El coeficiente con mayor peso fue efectivamente el Tramo de Ventas, seguido por la Actividad Económica (habría que estandarizar los regresores para poder tener una noción comparable de importancia y que no incidan factores de escala).

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.024551   0.013401 -151.08   <2e-16 ***
Num_Trabajadores          0.023196   0.001073   21.62   <2e-16 ***
Mes_Inicio_act            0.101685   0.001707   59.57   <2e-16 ***
woe_Tipo_Contribuyente   -0.733572   0.028210  -26.00   <2e-16 ***
woe_AEC                  -0.790734   0.012905  -61.27   <2e-16 ***
woe_Comuna               -0.742212   0.029212  -25.41   <2e-16 ***
woe_Tramo_Venta          -0.919965   0.009651  -95.32   <2e-16 ***
woe_Tramo_Capital_Propio -0.685257   0.010041  -68.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Tabla 9: Coeficientes Logit*

En el set de validación se obtuvo un AUC PR de un 42% (ver Figura 32) y un AUC ROC de un 73,5% (ver Figura 33). En el reliability plot (ver Figura 34) se puede observar que las probabilidades bajo el 50% se encuentran bien calibradas, mientras que probabilidades muy altas del modelo presentan sobreestimaciones.

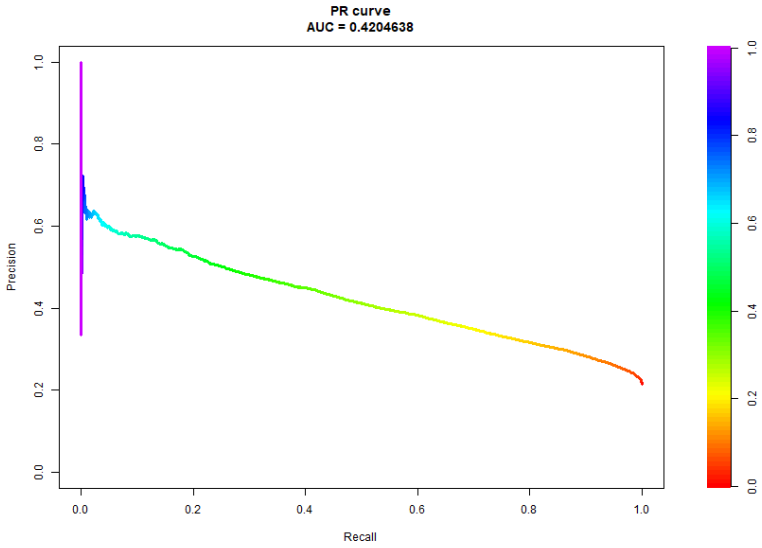



*Figura 32: Curva PR Logit*

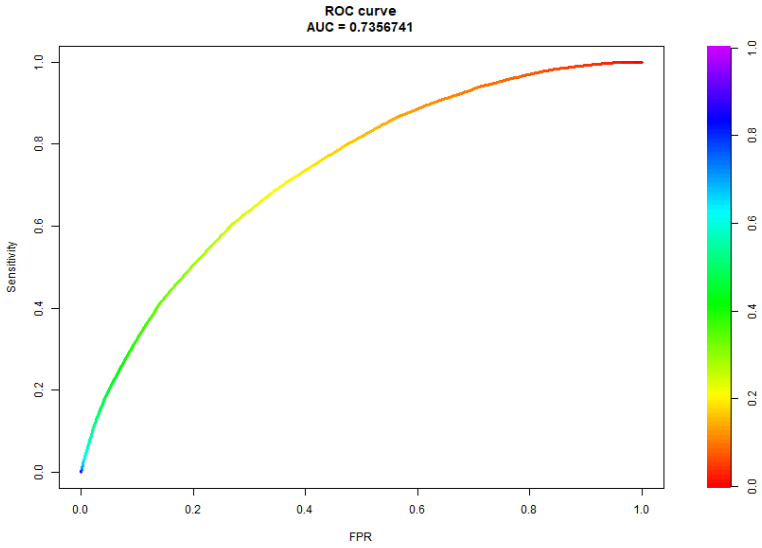



*Figura 33: Curva ROC Logit*

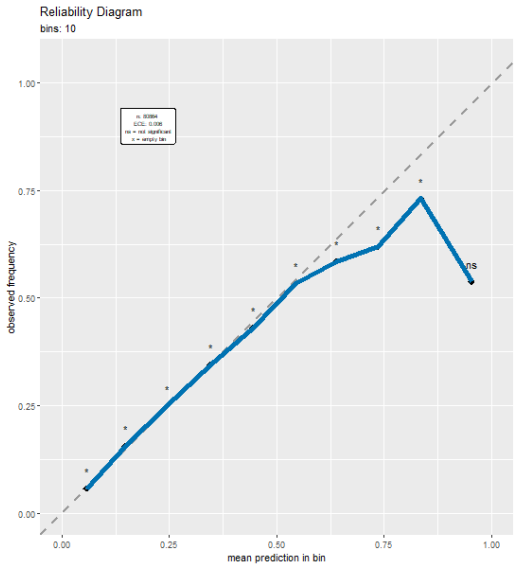



*Figura 34: Reliability Plot Logit*

Al maximizar el F1-score del modelo moviendo el corte de probabilidad, obtenemos que con un corte de un 24% en la probabilidad se obtiene el F1-score máximo de un 46,7% (ver Tabla 10).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 45140  6513
         1 18306 10905

               Accuracy : 0.6931
                 95% CI : (0.6899, 0.6963)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 1

                  Kappa : 0.271

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6261
            Specificity : 0.7115
         Pos Pred Value : 0.3733
         Neg Pred Value : 0.8739
              Precision : 0.3733
                 Recall : 0.6261
                     F1 : 0.4677
             Prevalence : 0.2154
         Detection Rate : 0.1349
   Detection Prevalence : 0.3612
      Balanced Accuracy : 0.6688

       'Positive' Class : 1
```

*Tabla 10: Matriz de Confusión e Indicadores Logit maximizando F1-score*

## 4.11. Modelo 2: XGBoost

Para el XGBoost se utilizó el encoding mediante WoE, pero no la selección de atributos (los árboles tienen métodos incorporados de selección de atributos). Adicional a esto, es necesario calibrar hiperparámetros. Para esto se utilizó el algoritmo Subplex y se calibraron de esta forma los siguientes hiperparámetros:

- Learning Rate
- Coeficiente de regularización L2
- Subsample

- Colsample by Tree
- Gamma

El óptimo fue encontrado con la métrica de evaluación LogLoss y los siguientes parámetros:

- Learning rate = 0.1
- Gamma = 0.035
- L2 regularización = 0.5
- Subsample = 0.9
- Colsample by tree = 1 (es decir, no utiliza subconjunto de atributos en la construcción de los árboles)

Con esto, se obtuvo un AUC PR de un 45.4% (ver Figura 35) y un AUC ROC de un 75,2% (ver Figura 36). El Reliability plot (ver Figura 37) se ve más preciso que el del Logit, aunque también con una ligera sobreestimación en los tramos altos de probabilidades.

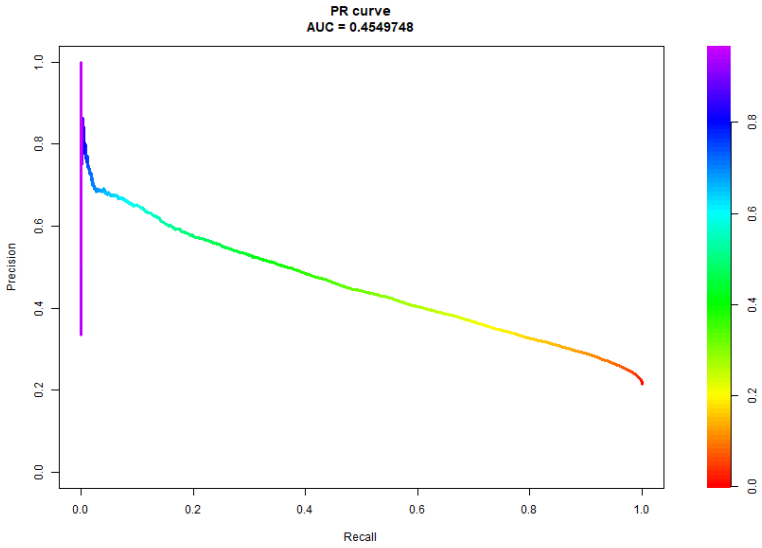



*Figura 35: Curva PR XGBoost*

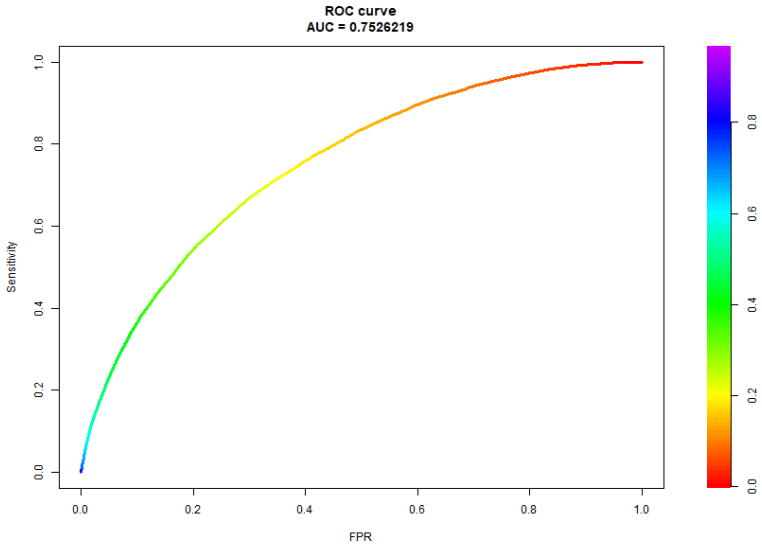



*Figura 36: Curva ROC XGBoost*

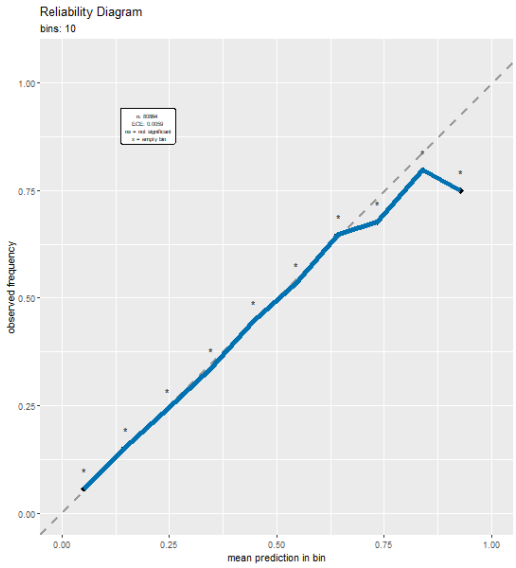



*Figura 37: Reliability Plot XGBoost*

Al maximizar el F1-score, se obtiene un 48,4% con un corte de probabilidad de un 24% (ver Tabla 11).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 45655  6159
         1 17791 11259

               Accuracy : 0.7038
                 95% CI : (0.7007, 0.707)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2946

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6464
            Specificity : 0.7196
         Pos Pred Value : 0.3876
         Neg Pred Value : 0.8811
              Precision : 0.3876
                 Recall : 0.6464
                     F1 : 0.4846
             Prevalence : 0.2154
         Detection Rate : 0.1392
   Detection Prevalence : 0.3592
      Balanced Accuracy : 0.6830

       'Positive' Class : 1
```

*Tabla 11: Matriz de Confusión e Indicadores XGBoost*

## 4.12. Modelo 3: CatBoost

Para calibrar el modleo CatBoost no se utilizó un encoding explícito ni selección de atributos previa a la calibración del modelo, ya que éste tiene incorporadas técnicas para manejar ambos tópicos.

Se utilizó el algoritmo Subplex para calibrar hiperparámetros propios del algoritmo, obteniendo los siguientes valores óptimos:

- Learning Rate = 0.137125
- Regularización L2 = 0.125

Con estos parámetros, se obtuvo un AUC PR de 46.2% (ver Figura 38) y un AUC ROC de un 75.6% (ver Figura 39). El reliability plot (ver Figura 40) muestra un buen balanceo de las probabilidades predichas con las tasas naturales de ocurrencia en los 10 bines analizados.

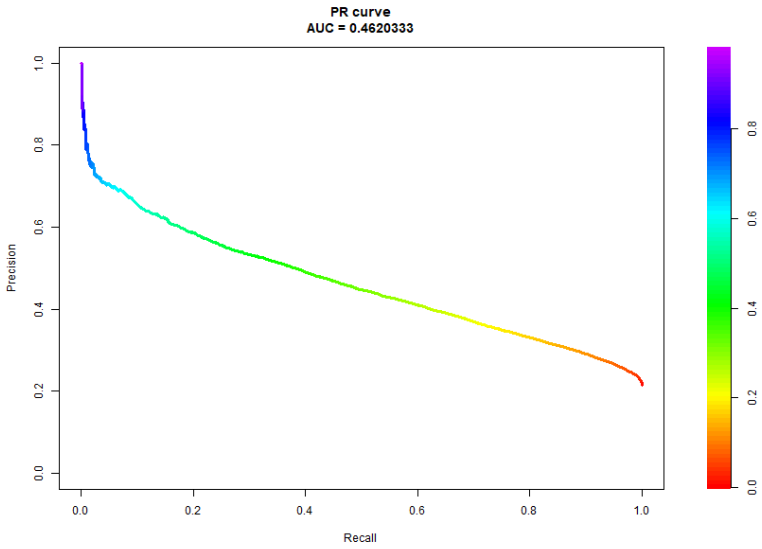



*Figura 38: Curva PR CatBoost*

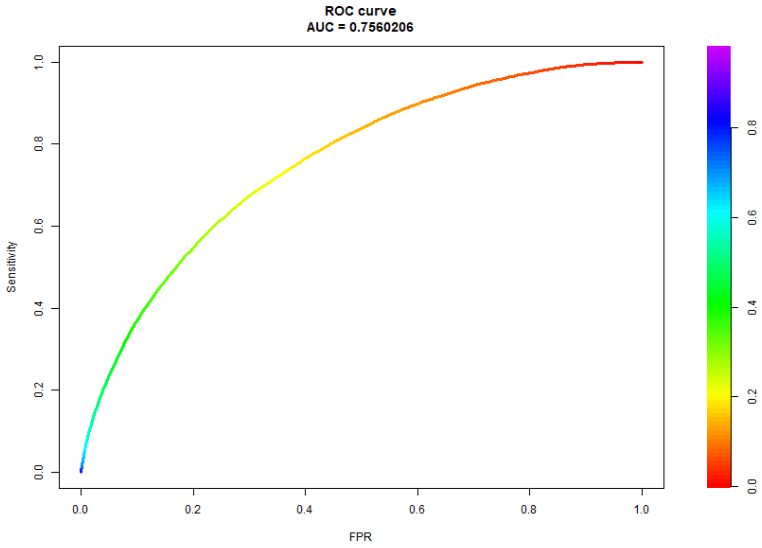



*Figura 39: Curva ROC CatBoost*

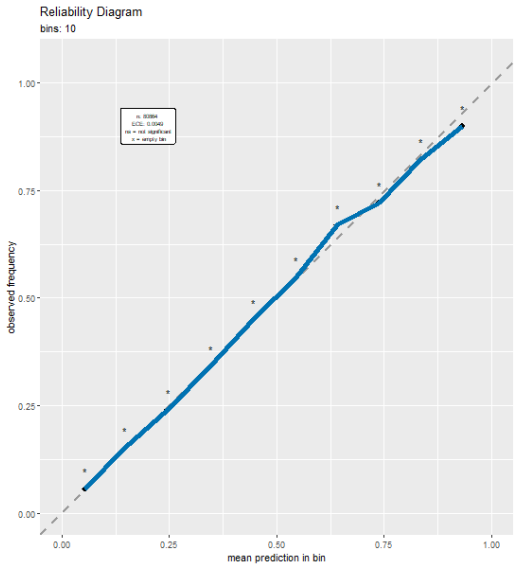



*Figura 40: Reliability Plot CatBoost*

Al determinar el punto de corte en probabilidad que maximiza el F1-score, tenemos que con un 24% de probabilidad se obtiene un F1-score de 48,8% (ver Tabla 12).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 45760  6086
         1 17686 11332

               Accuracy : 0.706
                 95% CI : (0.7029, 0.7092)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2995

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6506
            Specificity : 0.7212
         Pos Pred Value : 0.3905
         Neg Pred Value : 0.8826
              Precision : 0.3905
                 Recall : 0.6506
                     F1 : 0.4881
             Prevalence : 0.2154
         Detection Rate : 0.1401
   Detection Prevalence : 0.3588
      Balanced Accuracy : 0.6859

       'Positive' Class : 1
```

*Tabla 12: Matriz de Confusión e Indicadores CatBoost*

## 4.13. Resultados Obtenidos

Se puede ver un resumen de los indicadores obtenidos con las 4 estrategias de modelos en la Tabla 13.

| Modelo | AUC PR | AUC ROC | ECE | F1-score | Precision | Recall | Specificity |
|---|---|---|---|---|---|---|---|
| Regla de Negocio | - | - | - | 38,70% | 39,16% | 38,26% | 83,67% |
| Logit | 42,04% | 73,56% | 0,60% | 46,77% | 37,33% | 62,61% | 71,15% |
| XGBoost | 45,49% | 75,26% | 0,59% | 48,46% | 38,76% | 64,64% | 71,96% |
| CatBoost | 46,20% | 75,60% | 0,49% | 48,81% | 39,05% | 65,05% | 72,12% |

*Tabla 13: Resultados Modelos*

El modelo CatBoost presenta los mejores indicadores para la mayoría de las métricas. Existe cierto compromiso entre algunos indicadores al escoger el corte óptimo de probabilidad que hace que la regla de negocio sea muy buena en Precision y Specificity, pero esto es en desmedro de un recall extremadamente bajo. Tanto Logit, XGBoost y CatBoost podrían mejorar esa Precision y Specificity en desmedro de los otros indicadores. Precisamente esa falta de flexibilidad hace que sea poco atractivo adoptar una regla de negocio como estrategia general.

La diferencia en AUC PR entre Logit, XGBoost y CatBoost no es tan grande. La elección del mejor modelo es, por lo tanto, no tan obvia. Depende del caso de uso elegir entre el mejor modelo (CatBoost) a costa de sacrificar interpretabilidad (incluso con SHAP Values, donde podríamos estimar la relevancia global de cada atributo, no sería fácil entender el encoding de las variables categóricas) o un modelo directamente interpretable (Logit). El equilibrio entre desempeño e interpretabilidad (aunque no tan directa) es el XGBoost.

Tanto el Logit, XGBoost y CatBoost presentan un ECE relativamente bajo. Con esto, no es necesario recalibrar las probabilidades para utilizarlas, ni tampoco permite discriminar necesariamente qué modelo se ajusta mejor.

Adoptaremos el XGBoost como un modelo suficientemente sofisticado e interpretable mediante SHAP Values para una aplicación práctica.

## 4.14. Maximización de Utilidad

Una vez escogido el mejor modelo, podemos plantearnos la forma en que éste será utilizado y escoger un corte en probabilidad que maximice un retorno esperado.

Es conveniente disociar la elección del modelo de la maximización de la utilidad debido a:

- El modelo debería representar la realidad lo mejor posible, y luego al utilizarlo se debería considerar la estrategia que maximice el retorno con la información del modelo que representa mejor la realidad.
- Elegir el modelo que maximice el retorno esperado para algún corte puede ser una alternativa poco robusta si los costos están sujetos a cambios o hay imprecisiones en su calibración.
- Puede haber más de una aplicación práctica, con costos asociados distintos, para el mismo modelo.

Es importante destacar que más que el valor absoluto de los costos y beneficios, lo que va a hacer cambiar la decisión del corte óptimo son los valores relativos entre ellos.

Dentro de este contexto, para poder aplicar la maximización del retorno esperado y ver cómo varía con los distintos costos y beneficios, se plantean 3 escenarios hipotéticos de uso de este modelo. Los 3 escenarios serán ambientados desde

el posible uso del modelo por parte de una entidad financiera que busca perfilar clientes potenciales y nos centraremos en la relatividad de los beneficios y costos más que en sus valores absolutos.

### a. Escenario 1

La entidad financiera maneja una banca PyME y una Microempresas con costos de atención distintos. Atender un cliente en banca PyME tiene un costo anual de 3 unidades y en banca Microempresas de 1. La diferencia se debe al trato personalizado y a la menor carga de cartera que tienen los ejecutivos banca PyME. Un cliente PyME atendido en un ejecutivo PyME en general entrega al banco un beneficio bruto de 10, mientras que un cliente Microempresas atendido con un ejecutivo Microempresas solo tiene un beneficio bruto de 2, debido a la menor oferta de productos especializados, el menor volumen de monto, la atención menos especializada y el mayor riesgo de no pago.

El banco pretende carterizar inmediatamente los clientes nuevos en el sistema con potencial crecimiento a PyME en banca PyME en bases a las inferencias del modelo. Existen dos posibles errores en esta decisión:

- Carterizar como PyME a alguien que no tiene el potencial suficiente. En este caso se incurre en el costo de atención más elevado, pero no se logra obtener más rentabilidad bruta que la que tiene en promedio un cliente Microempresas.

- No carterizar como PyME a alguien que si tiene el potencial. En este caso se incurre en un costo de atención más reducido, pero probablemente el cliente será mejor atendido por la competencia, por lo que no dejará ingresos

La matriz de costo-beneficio neto de este escenario está representado en la Tabla 14.

| | | Realidad | |
|---|---|---|---|
| | | PyME | Micro |
| Predicción | PyME | 7 | -1 |
| | Micro | -1 | 1 |

*Tabla 14: Matriz Costos Beneficios Escenario 1*

Los costos de clasificar erróneamente un cliente son simétricos, pero el beneficio de clasificar bien a un cliente con potencial PyME es varias veces mayor que el costo de los errores y el beneficio de clasificar bien a un cliente con potencial Microempresas. ¿Cómo varía en base a esta información mi decisión de corte óptimo?

Es razonable pensar que, al haber un retorno más importante clasificando correctamente a los clientes con potencial PyME, el corte en probabilidad sea relativamente bajo, permitiendo encontrar más de estos clientes potenciales e incurriendo en más casos falsos positivos que tienen un costo relativo menor.

El corte en probabilidad que maximizaba el F1-score era un 24%. En la Figura 55 podemos observar que efectivamente el máximo retorno esperado para el escenario 1 es con un corte más bajo de un 19%, permitiéndonos obtener una utilidad esperada por caso de 1.25 (esto en unidades relativas a los costos y beneficios de la Figura 41).

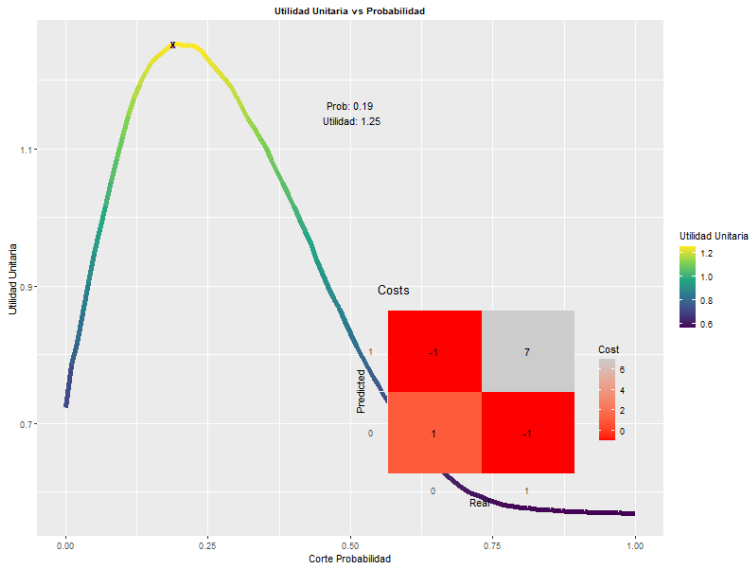



*Figura 41: Utilidad Esperada Escenario 1*

Con este corte en probabilidad, se mejora el Recall hasta un 74,8% empeorando el Precision a un 34,7% (ver Tabla 15) en comparación a la elección de corte que maximiza el F1-score (Recall de 64,6% y Precision de 38,7%). Esto debido a que es más rentable aumentar el Recall (es decir, encontrar más casos PyME a costa de equivocarse más en la tasa de falsos positivos).

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 38914  4380
         1 24532 13038

               Accuracy : 0.6425
                 95% CI : (0.6391, 0.6458)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2549

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7485
            Specificity : 0.6133
         Pos Pred Value : 0.3470
         Neg Pred Value : 0.8988
              Precision : 0.3470
                 Recall : 0.7485
                     F1 : 0.4742
             Prevalence : 0.2154
         Detection Rate : 0.1612
   Detection Prevalence : 0.4646
      Balanced Accuracy : 0.6809

       'Positive' Class : 1
```

*Tabla 15: Matriz de Confusión e Indicadores XGBoost Escenario 1*

## b. Escenario 2

Similar al escenario 1, pero en este caso el costo de clasificar a un Microempresario como PyME se incrementa al cuantificar el riesgo de otorgarle productos crediticios a un cliente con mayor probabilidad de default. La estimación se hace cerca del momento de inicio de actividades, por lo que no hay historia suficiente para estimar el default real que va a tener el cliente.

La matriz de costos ahora es la que se observa en la Tabla 16.

| | | Realidad | |
|---|---|---|---|
| | | PyME | Micro |
| Predicción | PyME | 7 | -5 |
| | Micro | -1 | 1 |

*Tabla 16: Matriz Costos Beneficios Escenario 2*

Ahora, los costos de equivocarse no son simétricos. El costo de clasificar erróneamente un cliente con potencial PyME es mucho mayor y es comparable al beneficio de la clasificación correcta del cliente con potencial PyME. Es esperable que el corte en probabilidad que maximiza el retorno esperado sea mayor para compensar este efecto.

Como se puede ver en la Figura 42, el nuevo corte se incrementó a 43%. Con este corte, la utilidad máxima esperada por cada unidad es de 0.74. Esta utilidad es menor a la obtenida en el escenario 1 debido a que incrementamos los costos, pero no los beneficios. También se puede observar que la forma de la curva de utilidad cambió, ahora no decae fuertemente sobre la probabilidad óptima. Esto se debe a que el elevado costo de considerar PyME a una empresa que no lo va a ser, hace poco atractivo (comparativamente hablando) arriesgarse en identificar más PyMEs verdaderos. Con un corte más alto, los costos se neutralizan con los beneficios. También se puede observar que, en la cola izquierda del gráfico, la utilidad ahora puede ser negativa. Una mala decisión puede hacer no solo que dejemos de ganar utilidad, si no que perdamos valor.

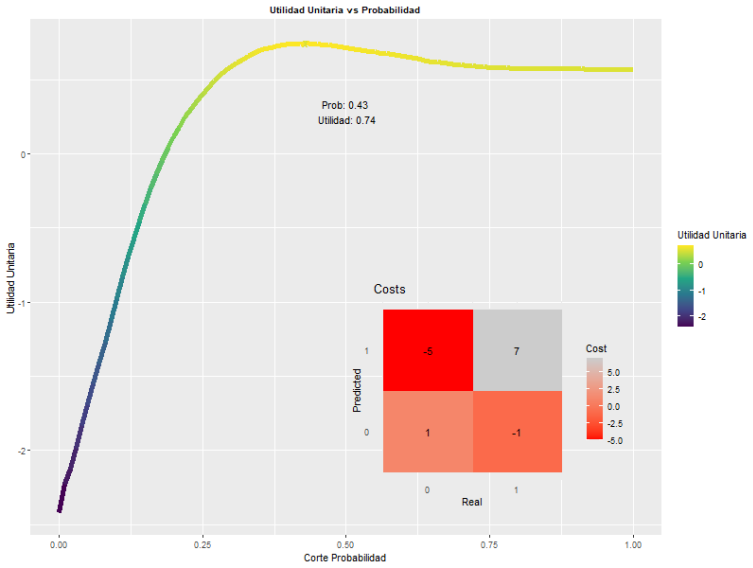



*Figura 42: Utilidad Esperada Escenario 2*

Con un corte en probabilidad de 43%, se obtiene un Recall de un 29,05% y una Precision de un 54,57% (ver Tabla 17). El Recall cae drásticamente con respecto al escenario 1 y la Precision aumenta significativamente. Esto debido a que Precision mide justamente cuántos casos de los predicho positivamente, son realmente positivos (es decir, busca ser eficiente en la predicción), lo que va de la mano con el incremento señalado en el costo. Como efecto colateral, cae el Recall, es decir, la tasa de detección de los casos originalmente positivos.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 59061 12357
         1  4385  5061

               Accuracy : 0.793
                 95% CI : (0.7902, 0.7957)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 3.216e-09

                  Kappa : 0.2655

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.29056
            Specificity : 0.93089
         Pos Pred Value : 0.53578
         Neg Pred Value : 0.82698
              Precision : 0.53578
                 Recall : 0.29056
                     F1 : 0.37679
             Prevalence : 0.21540
         Detection Rate : 0.06259
   Detection Prevalence : 0.11681
      Balanced Accuracy : 0.61072

       'Positive' Class : 1
```

*Tabla 17: Matriz de Confusión e Indicadores XGBoost Escenario 2*

### c. Escenario 3

El modelo será utilizado para sugerir capacitaciones tempranas para los socios de la empresa en términos de educación financiera y uso eficiente de caja. Existen 2 cursos, uno adaptado a las Microempresas en uso eficiente de caja y otro para PyME orientado a capacitar en el uso eficiente de productos bancario de comercio exterior, factoring, leasing, etc.

Como consecuencia, acertar en el diagnóstico temprano de quién se convertirá en PyME o quedará como Microempresas tiene un efecto de gratitud por parte de los potenciales clientes del banco, no directamente vinculable a convertirse

eventualmente cliente, pero si con una ligera propensión producto de la confianza en la certeza del diagnóstico. Equivocarse, por otro lado, tiene un efecto de suma cero. Un PyME que se le ofrece un curso muy básico y un Microempresario con cursos muy avanzados de productos bancarios que quizás nunca necesite pueden simplemente optar por no asistir. Estos costos y beneficios se pueden observar en la Tabla 18.

| | | Realidad | |
|---|---|---|---|
| | | PyME | Micro |
| Predicción | PyME | 1 | 0 |
| | Micro | 0 | 1 |

*Tabla 18: Matriz Costos Beneficios Escenario 3*

En este caso, la forma de la curva de utilidad es similar a la del escenario 2 (ver Figura 43). Esto, debido a que los costos y beneficios de predecir un PyME son similares. El corte en probabilidad óptimo es de un 48%, entregando una utilidad unitaria de 0.8. Esta utilidad unitaria es mayor a la que se obtuvo en el escenario 2 donde existían beneficios y costos más elevados. También es claro por la naturaleza de los costos de este escenario que no es posible obtener una utilidad negativa.

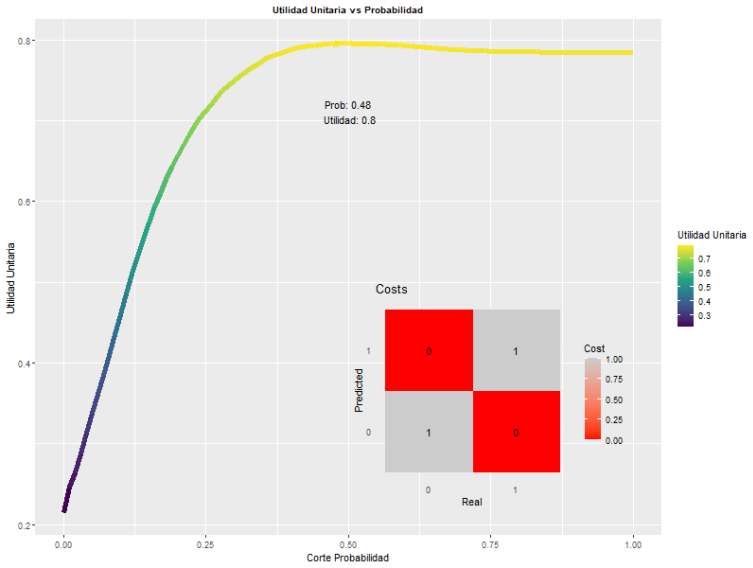



*Figura 43: Utilidad Esperada Escenario 3*

Con el corte óptimo, se obtiene un Recall de un 21,3% y una Precision de 57,1% (ver Tabla 19). Estos valores se asemejan a los obtenidos en el escenario 2 debido a la cercanía de los cortes en probabilidad óptimo y a que ambos están calculados sobre la misma curva PR.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 60654 13700
         1  2792  3718

               Accuracy : 0.7961
                 95% CI : (0.7933, 0.7988)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 8.046e-16

                  Kappa : 0.2193

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.21346
            Specificity : 0.95599
         Pos Pred Value : 0.57112
         Neg Pred Value : 0.81575
              Precision : 0.57112
                 Recall : 0.21346
                     F1 : 0.31077
             Prevalence : 0.21540
         Detection Rate : 0.04598
   Detection Prevalence : 0.08051
      Balanced Accuracy : 0.58473

       'Positive' Class : 1
```

*Tabla 19: Matriz de Confusión e Indicadores XGBoost Escenario 3*

En resumen, el mismo modelo que entrega probabilidades como salida puede ser utilizado para fines muy diversos. Esta flexibilidad no se tiene al trabajar con reglas de negocio o modelos cuya salida es directamente el valor de la categoría predicha. En el caso de modelos que entreguen directamente la categoría, existen transformaciones, como *Plat Scaling*, que pueden ayudar a transformarlas en probabilidades. Es muy importante adaptar la sugerencia del modelo al caso de uso específico que se pretende hacer de él.

Como todos esos usos distintos se apalancan en el desempeño original del modelo en términos de AUC PR, conviene utilizar el modelo que promete maximizar este indicador. Un mejor AUC PR debería mejorar los indicadores

generales de cada corte encontrado, mejorando la utilidad esperada de cada caso de uso.

## 4.15. Evaluación en set de Test

Hasta el momento, todas las calibraciones del modelo (*encoding*, selección de atributos y ajustes de los modelos) han sido hechas sobre el set de entrenamiento, y todas las decisiones adicionales, sobre el set de validación (elección de hiperparámetros, elección del mejor modelo, calibración de cortes óptimos de probabilidad, etc.).

Una vez tomadas estas decisiones, es necesario contrastar los resultados obtenidos con un set independiente de test que permita indagar el resultado real del modelo una vez puesto en producción. Sobre los resultados de este test no deben tomarse decisiones adicionales que modifiquen hiperparámetros o condiciones del modelo, ya que de hacerlo estaríamos ensuciando este diagnóstico. Si podemos observar qué hace el modelo con los datos y tratar de indagar de por qué toma esas decisiones.

En la Figura 44 podemos observar que el AUC PR de nuestro modelo en un set de test es de 45,43%, contra un 45,49% en el set de validación, lo cual es ligeramente menor pero muy similar. Es de esperar que la generalización del modelo encontrado se mantenga en nuevos datos si las condiciones del sistema se mantienen.

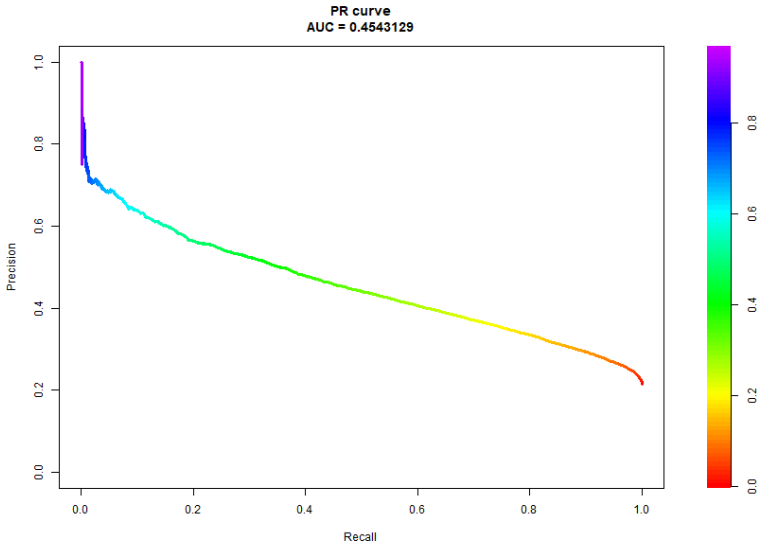



*Figura 44: Curva PR XGBoost en set de Test*

Utilizando los costes del escenario 3, podemos observar en la Figura 45 que el máximo corte no se encontró en 48%, si no en 52%, para alcanzar la misma utilidad esperada de 0.8. La diferencia no es mucha, pero habla de la falta de robustez en la elección de ese corte óptimo. Será necesario aleatorizar diferentes ejecuciones del modelo y quedarse con la opción que mejor se desempeñe sobre esos ligeramente distintos escenarios.

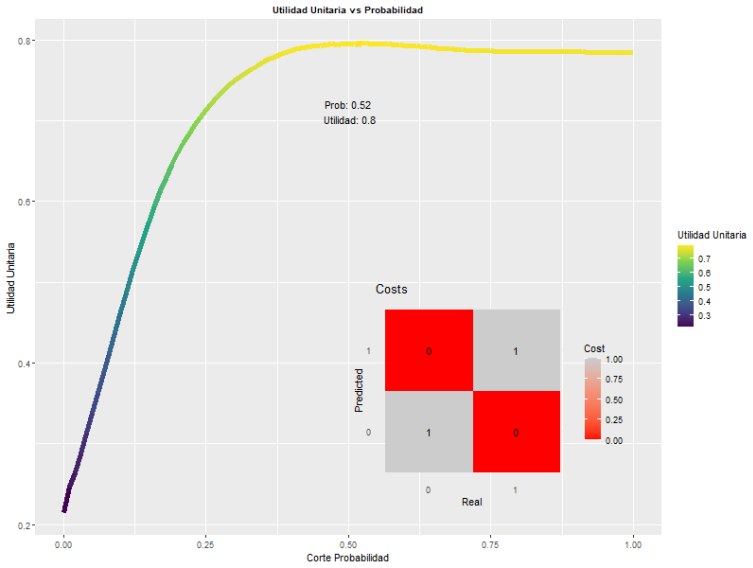



*Figura 45: Utilidad Esperada Escenario 3 set Test*

Con el corte en probabilidad de 48% (el escogido en el set de validación) y como se puede ver en la Tabla 20, se obtuvo un Recall de un 21,1% (Recall set validación fue un 21,3%) y un Precision de 55,8% (Precision del set de validación fue un 57,1%). Estos valores no son tan distintos a los esperados de acuerdo al set de validación y el modelo promete mantener las propiedades encontradas al momento de ponerlo en producción.

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 60534 13738
         1  2911  3679

               Accuracy : 0.7941
                 95% CI : (0.7913, 0.7969)
    No Information Rate : 0.7846
    P-Value [Acc > NIR] : 2.05e-11

                  Kappa : 0.2135

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.2112
            Specificity : 0.9541
         Pos Pred Value : 0.5583
         Neg Pred Value : 0.8150
              Precision : 0.5583
                 Recall : 0.2112
                     F1 : 0.3065
             Prevalence : 0.2154
         Detection Rate : 0.0455
   Detection Prevalence : 0.0815
      Balanced Accuracy : 0.5827

       'Positive' Class : 1
```

*Tabla 20: Matriz de Confusión e Indicadores XGBoost Escenario 3 set de Test*

## 4.16. Interpretación del Modelo

Una primera mirada a interpretar qué variables pesaron más en el modelo es analizar la importancia de atributos. Existen varias métricas utilizadas para determinar esta importancia, tales como ganancia (contribución del atributo al modelo en términos de las ganancias en las divisiones de ramas donde participa), frecuencia (cuantas veces el atributo está presente en los árboles) y cobertura (número de observaciones asociadas a un corte producido por el atributo evaluado).

En la Figura 46 se puede observar la importancia en términos de ganancia del modelo XGBoost. El atributo más significativo con esta métrica es el tramo de venta, seguido por el tramo de capital propio, como habíamos supuesto inicialmente del análisis de correlaciones.

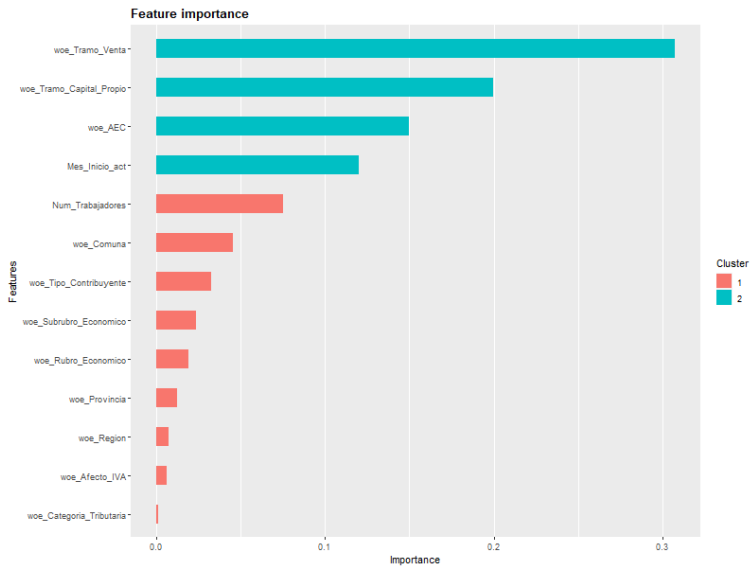



*Figura 46: Importancia de Atributos XGBoost por Ganancia*

En la Figura 47 se puede ver la importancia de atributos asociada a la cobertura. El atributo más valioso en base a esta métrica es la Actividad Económica seguido por el tramo de capital propio.

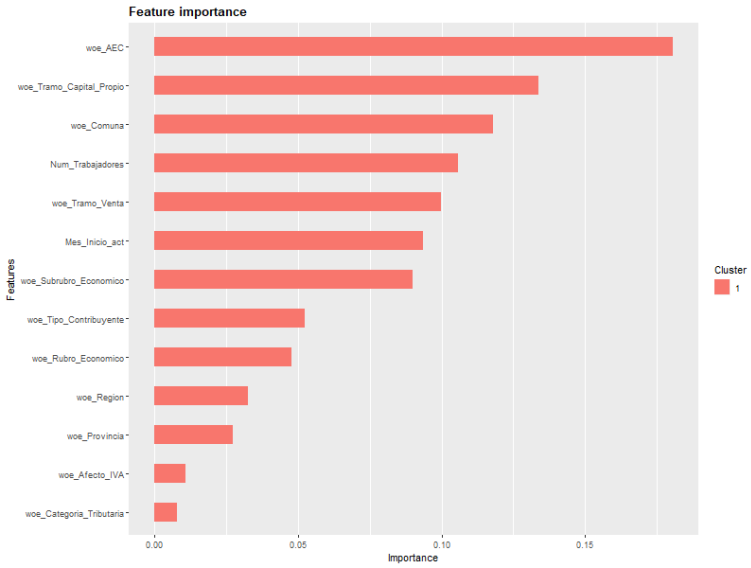



*Figura 47: Importancia de Atributos XGBoost por Cobertura*

En la Figura 48 se ve la importancia de atributos asociada a la Frecuencia. El atributo más importante vuelve a ser la actividad económica, pero esta vez seguido por la comuna.

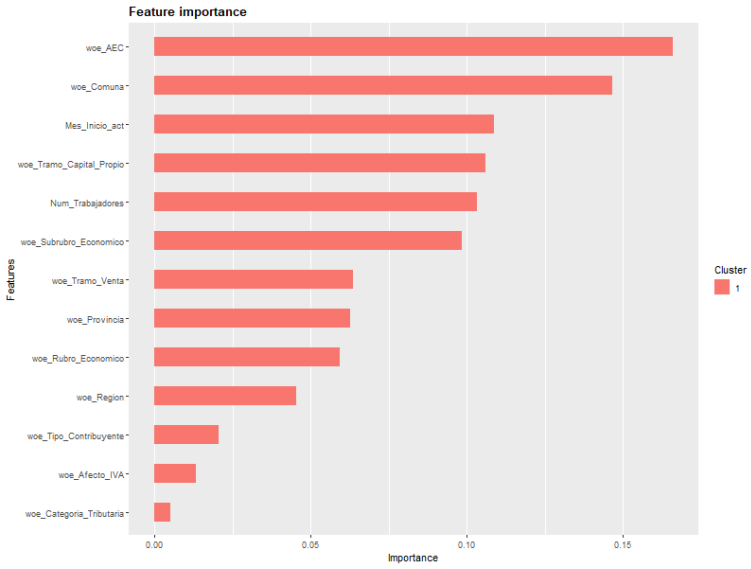



*Figura 48: Importancia de Atributos XGBoost por Frecuencia*

Estas distintas métricas no concuerdan en la noción de qué es importante al momento de discriminar la relevancia de un atributo en la predicción, pero como se describió en el capítulo 2.10., estas métricas no cumplen con los requisitos de consistencia ni precisión.

Es por esto que analizaremos los SHAP Values entregados por el modelo. Los SHAP Values se basan en determinar la contribución marginal de cada atributo por sobre un modelo base, donde la suma de la relevancia de los atributos debe dar una métrica relevante, en este caso, explica el cambio en probabilidad de la salida del modelo medido en log odds. Es decir, si $\mathbb{P}(x)$ es la probabilidad entregada por el modelo para el registro x, $SHAP_i(x)$ es el SHAP Value entregado para el registro x en el atributo i y BIAS es el sesgo o valor base del modelo (obtenido de la calibración del modelo sin atributos independientes), entonces:

$$\mathbb{P}(x) = \frac{1}{(1 + e^{BIAS + \sum_i SHAP_i(x)})}$$

En la Figura 49 podemos ver un resumen de la dispersión de los SHAP Values de cada atributo en el set de test. Aquellos atributos con más dispersión inciden más en las diferencias finales de la probabilidad predicha por el modelo. En este caso, la Actividad Económica, la Comuna y el Tipo de Contribuyente parecen destacar en dispersión. El color indica el verdadero valor del target: morado si efectivamente la empresa se convirtió en Pyme en 3 años y amarillo si no. Los valores junto a la variable indican el valor base asociado a ese atributo (calculado como el promedio de los $SHAP_i$. Por este valor se ordenan las variables

mostradas en el gráfico y se centra la dispersión en cero para visualizar las variables en la misma escala y poder compararlas. El Tramo de venta es el que tiene en promedio un SHAP value mayor.

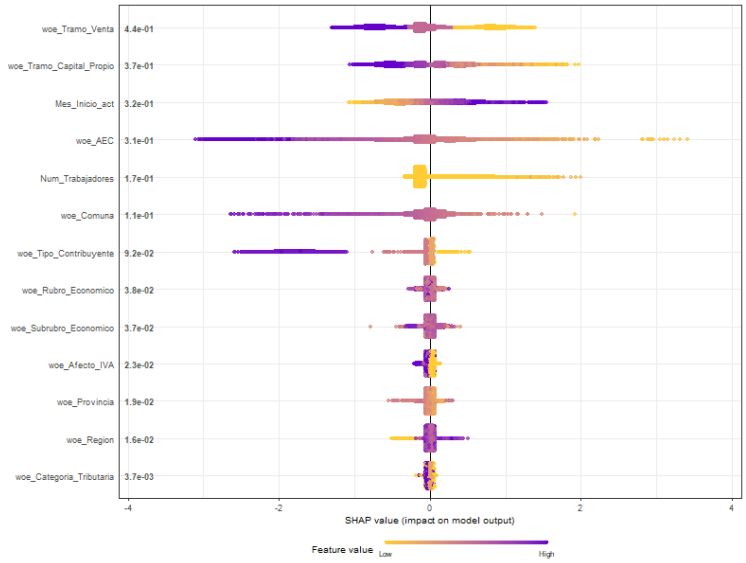



*Figura 49: SHAP Values Resumen por Atributos*

Adicional a este gráfico resumen que muestra las relevancias de cada atributo en la predicción, podemos ver la composición agregada por cada registro de los valores que conforman la predicción final de probabilidad (ver Figura 50).

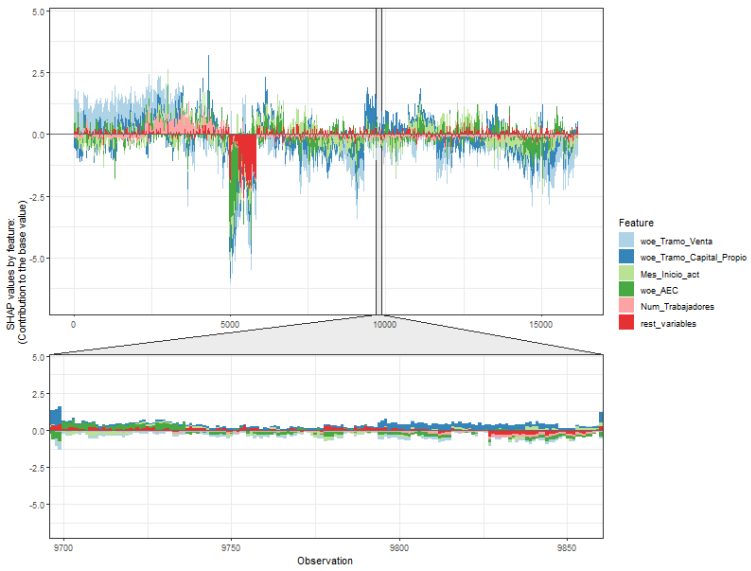



*Figura 50: SHAP Value Resumen por Registro*

Esta descomposición se puede agrupar en clusters para una mejor visualización (ver Figura 51).

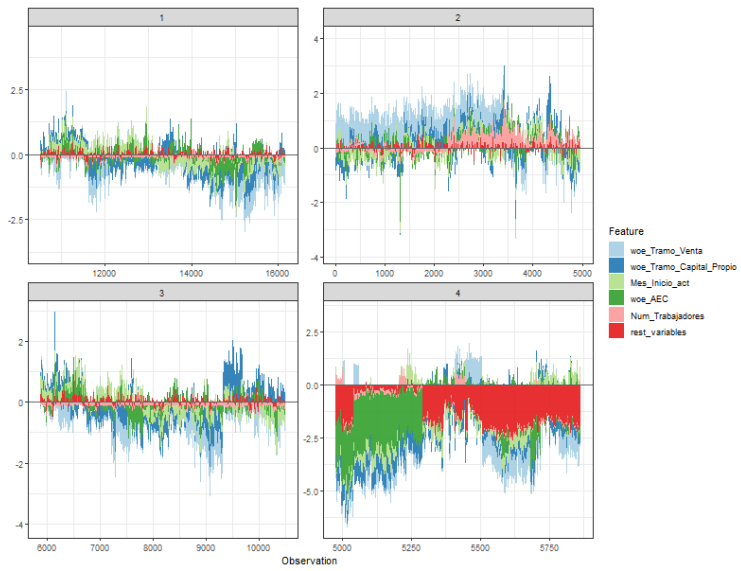



*Figura 51: SHAP Value Resumen por Registro por Segmentos*

También es posible, a un mayor costo computacional, visualizar las interacciones entre diversas combinaciones de SHAP Values. En la Figura 52 se puede visualizar la interacción del SHAP value del WoE de Tramo de Ventas (color), el WoE de la Actividad Económica (eje x) y el WoE del Tramo de Capital Propio (eje y). Se pueden observar Actividades económicas aisladas del resto en términos de SHAP values, con una gran diversidad de dispersión en términos de capital propio (algunas concentran ciertos valores, otras tienen una amplia gama de valores de capital propio) y en muchas de ellas la codificación del tramo de ventas no es lineal. Más allá de las posibles interpretaciones puntuales que se puede hacer de este gráfico, éste habla de las complejas interacciones con las que se calcula finalmente la probabilidad final en base al modelo escogido.

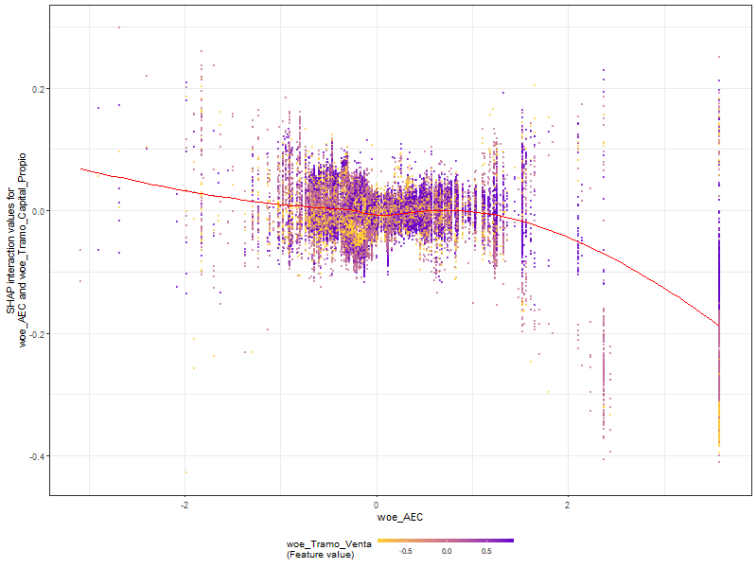



*Figura 52: SHAP Value Interacción Actividad Económica, Tramo de Capital Propio y Tramo Venta*

Otra de las ventajas de usar SHAP values, es que es posible interpretar qué factores inciden en la predicción final de cada registro en particular. En el ejemplo 1 visto en la Tabla 21, se escogió un registro con una alta probabilidad predicha por el modelo. Esta empresa efectivamente se convirtió en PyME al cabo de los 3 años.

| | X256576 |
|---|---|
| flag_Crecimiento_3Y | 1 |
| Num_Trabajadores | 3 |
| Tipo_Contribuyente | persona juridica comercial |
| Rubro_Economico | industria manufacturera |
| Subrubro_Economico | fabricacion de otros productos elaborados de metal; actividades de servicios de trabajo de metales |
| AEC | fabricacion de articulos de cuchilleria, herramientas de mano y articulos de ferreteria |
| Region | viii region del bio bio |
| Provincia | concepcion |
| Comuna | coronel |
| Tramo_Venta | 4 |
| Tramo_Capital_Propio | 7 |
| Afecto_IVA | SI |
| Categoria_Tributaria | 1 |
| Mes_Inicio_act | 8 |

*Tabla 21: Ejemplo 1 registro con alta probabilidad predicha*

En la Figura 53 se puede ver qué factores incidieron en la alta probabilidad predicha. La Actividad Económica “Artículos de Cuchillería, Herramientas de Mano y Artículos de Ferretería” fue el factor que más destacó en incrementar la probabilidad estimada por el modelo.

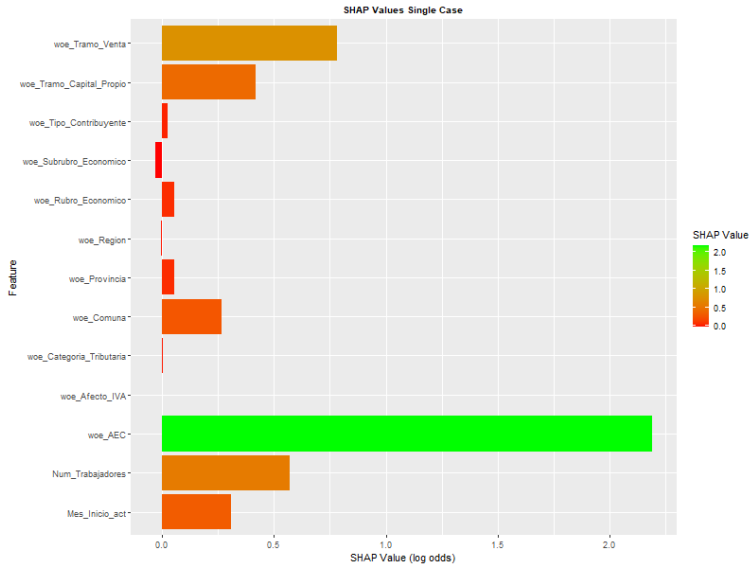



*Figura 53: SHAP Values Ejemplo 1*

En la Tabla 22, se puede ver un segundo ejemplo escogido con una probabilidad predicha baja. Este caso efectivamente no logró convertirse en PyME al cabo de 3 años.

| | X240093 |
|---|---|
| flag_Crecimiento_3Y | 0 |
| Num_Trabajadores | 0 |
| Tipo_Contribuyente | sin per. juridica |
| Rubro_Economico | actividades profesionales, cientificas y tecnicas |
| Subrubro_Economico | otras actividades profesionales, cientificas y tecnicas n.c.p. |
| AEC | actividades de agencias y agentes de representacion de actores, deportistas y otras figuras publicas |
| Region | vi region del libertador general bernardo ohiggins |
| Provincia | cachapoal |
| Comuna | rancagua |
| Tramo_Venta | 2 |
| Tramo_Capital_Propio | -100 |
| Afecto_IVA | SI |
| Categoria_Tributaria | 1 |
| Mes_Inicio_act | 2 |

*Tabla 22: Ejemplo 2 registro con baja probabilidad predicha*

La Actividad Económica de “Actividades de Agencias y Agentes de Representación de Actores, Deportistas y Otras Figuras Públicas” incidió negativamente en la probabilidad predicha, seguida por el Tipo de Contribuyente “Sin Persona Jurídica” (ver Figura 54).

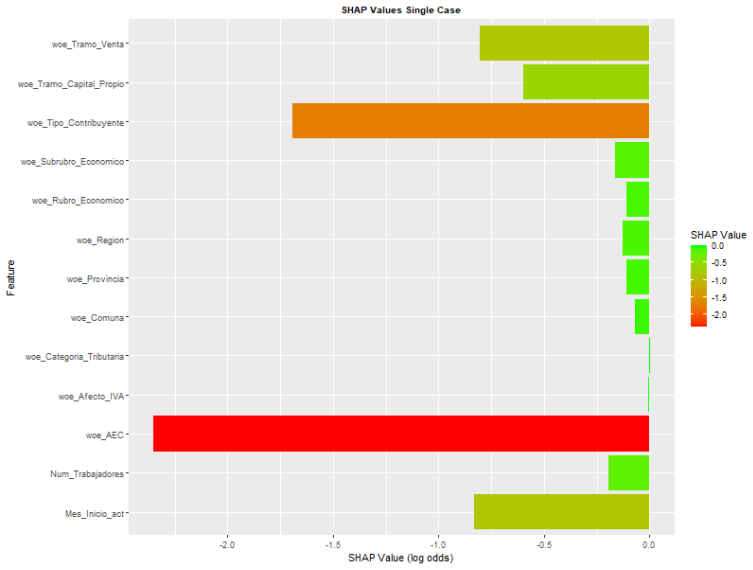



*Figura 54: SHAP Value Ejemplo 2*

# 5. Conclusiones

Es posible generar un modelo con indicadores de desempeño aceptables para predecir el crecimiento de una empresa a un horizonte de 3 años utilizando información pública. En particular, se analizó el resultado de un XGBoost calibrado con un preprocesamiento importante de las variables categóricas (mediante WoE Encoding) y una selección de hiperparámetros mediante el algoritmo Subplex. Este algoritmo alcanzó un AUC PR de un 45,43% en un set de Test.

La aplicación de este modelo a una finalidad comercial depende de los costos inherentes de esa aplicación en los aciertos y errores del modelo. En el capítulo 3.14 se ejemplificaron 3 aplicaciones posibles del modelo con resultados muy distintos de corte óptimo en probabilidad y utilidad esperada. Es posible utilizar directamente las probabilidades entregadas por el modelo debido al bajo valor ECE de un 0,59%. De no haber tenido un valor bajo, habría sido necesario recalibrar estas probabilidades antes de utilizarlas.

Los SHAP Values permiten identificar de forma global y local explicaciones prácticas sobre cómo el modelo está tomando las decisiones que inciden en la probabilidad final predicha. Esto cumpliendo con los requisitos de consistencia y precisión descritos en el capítulo 2.10. La actividad económica es el atributo que más incide en las predicciones del modelo escogido, seguidos por la Comuna y el Tipo de Contribuyente.

# 6. Trabajo Futuro

El actual trabajo se basa exclusivamente en información pública tributaria de empresas del sistema chileno. Es posible enriquecer la información aquí encontrada mediante el cruce con otro tipo de bases como, por ejemplo: información de la malla societaria que conforma la persona jurídica e indagaciones acerca de dichos socios (historial crediticio, participación en otras sociedades, información del nivel de ventas y rubros de esas sociedades, etc.), bancarización temprana de la empresa (banco que apertura cuenta, nivel de endeudamiento, garantías, etc.), información de balances, información de los empleados de esas empresas, en especial de las personas en cargos de tomas de decisiones, información de indicadores macroeconómicos, entre otras.

No se indagó en otras técnicas de modelamiento, como redes neuronales, Tensorflow, SVM u otras técnicas recientes y populares de gradient boosting, como LightGBM. Sería interesante comparar mediante SHAP Values, cómo estos otros modelos inferirían las respuestas que nos dan sobre la realidad modelada y qué tan distintas estas inferencias pueden ser.

# 7. Referencias


Afolabi, I. e. (2019). A Model for Business Success Prediction using Machine Learning Algorithms. *Journal of Physics Conference Series*, págs. http://dx.doi.org/10.1088/1742-6596/1299/1/012050.

Bhalla, D. (2015). *Listen Data*. Obtenido de https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

Brownlee, J. (2016). *A Gentle Introduction to XGBoost for Applied Machine Learning*. Obtenido de https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Brownlee, J. (2019). *How to Choose a Feature Selection Method For Machine Learning*. Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

*Catboost*. (2021). Obtenido de https://catboost.ai/

Chehab, M. (2020). *Knowledge Discovery Data (KDD)*. Obtenido de Medium: https://medium.com/analytics-vidhya/knowledge-discovery-data-kdd-a8b41509bff9

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD: Knowledge Discovery and Data Mining.* https://arxiv.org/abs/1603.02754.

Cheriyan, S. e. (2018). Intelligent Sales Prediction Using Machine Learning Techniques. *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE).* http://dx.doi.org/10.1109/iCCECOME.2018.8659115.

CRAN. (2020). *Package Subplex*. Obtenido de https://cran.r-project.org/web/packages/subplex/subplex.pdf

Dataman. (2019). *Explain Your Model with the SHAP Values*. Obtenido de Towards Data Science: https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. http://dx.doi.org/10.1145/1143844.1143874.

Hulstaert, L. (2018). *Understanding model predictions with LIME*. Obtenido de Towards Data Science: https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b#:~:text=LIME%20provides%20local%20model%20interpretability,the%20output%20of%20a%20model.

Ippolito, P. (2019). *Hyperparameters Optimization*. Obtenido de Towards Data Science: https://towardsdatascience.com/hyperparameters-optimization-526348bb8e2d

Ke, G. e. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems (NIPS)*, https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Krishnan, S. (2018). *Weight of evidence and Information Value using Python*. Obtenido de Medium: https://sundarstyles89.medium.com/weight-of-evidence-and-information-value-using-python-6f05072e83eb

Kübler, R. (2021). *The Explainable Boosting Machine*. Obtenido de Towards Data Science: https://towardsdatascience.com/the-explainable-boosting-machine-f24152509ebb

Kuhn, M. (2020). *confusionMatrix (caret) Documentation*. Obtenido de rdocumentation: https://www.rdocumentation.org/packages/caret/versions/6.0-86/topics/confusionMatrix

Lin, A., & Hsieh, T.-Y. (2014). Expanding the Use of Weight of Evidence and Information Value to Continuous Dependent Variables for Variable Reduction and Scorecard Development.

Lund, B. (2016). Weight of Evidence Coding and Binning of Predictors in Logistic Regression.

Lundberg, S. (2018). *Interpretable Machine Learning with XGBoost*. Obtenido de Towards Data Science: https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

Lundberg, S. e. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Lundberg, S. M. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. https://arxiv.org/abs/1905.04610.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Conference on Neural Information Processing Systems.* https://arxiv.org/abs/1705.07874.

Mazzanti, S. (2020). *Beyond One-Hot. 17 Ways of Transforming Categorical Features Into Numeric Features*. Obtenido de Towards Data Science: https://towardsdatascience.com/beyond-one-hot-17-ways-of-transforming-categorical-features-into-numeric-features-57f54f199ea4

Mazzanti, S. (2020). *SHAP values explained exactly how you wished someone explained to you*. Obtenido de Towards Data Science: https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

Mazzanti, S. (Mayo de 2021). *Python's «predict_proba» Doesn't Actually Predict Probabilities (and How to Fix It)*. Obtenido de Towards Data Science:

https://towardsdatascience.com/pythons-predict-proba-doesn-t-actually-predict-probabilities-and-how-to-fix-it-f582c21d63fc

Messalas, A. e. (2019). Model-Agnostic Interpretability with Shapley Values. http://dx.doi.org/10.1109/IISA.2019.8900669.

Nelder, J., & Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, https://doi.org/10.1093/comjnl/7.4.308.

Nori, H. e. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *https://arxiv.org/abs/1909.09223*.

Ozaki, Y. e. (2017). Effective hyperparameter optimization using Nelder-Mead method in deep learning. *IPSJ Transactions on Computer Vision and Applications*, http://dx.doi.org/10.1186/s41074-017-0030-7.

Ozenne, B. e. (2015). The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of Clinical Epidemiology*, https://doi.org/10.1016/j.jclinepi.2015.02.010.

Prokhorenkova, L. e. (2019). CatBoost: unbiased boosting with categorical features. https://arxiv.org/abs/1706.09516.

Ribeiro, M. T. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *https://arxiv.org/abs/1602.04938*.

Romani, G. A. (2002). Modelos de clasificación y predicción de quiebra de empresas: Una aplicación a empresas chilenas. pág. http://dx.doi.org/10.33801/fe.v7i1.3371.

Servicio de Impuestos Internos. (Marzo de 2021). *Estadísticas de Empresa*. Obtenido de https://www.sii.cl/sobre_el_sii/estadisticas_de_empresas.html

Servicio de Impuestos Internos. (Marzo de 2021). *Personas Jurídicas y Empresas*. Obtenido de https://www.sii.cl/sobre_el_sii/nominapersonasjuridicas.html

Servicio de Impuestos Internos. (Marzo de 2021). *Todos los códigos de actividad económica*. Obtenido de https://www.sii.cl/ayudas/ayudas_por_servicios/1956-codigos-1959.html

Shikar. (2019). *The recent Queen of ML Algorithms: XGBoost, and it's future*. Obtenido de Medium: https://medium.com/analytics-vidhya/the-recent-queen-of-ai-algos-xgboost-and-its-future-22d6df3cd206

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.*

Singer, S., & Nelder, J. (2009). *Nelder Mead Algorithm*. Obtenido de Scholarpedia: http://www.scholarpedia.org/article/Nelder-Mead_algorithm

Swalin, A. (2018). *CatBoost vs. Light GBM vs. XGBoost*. Obtenido de KDnuggets: https://www.kdnuggets.com/2018/03/catboost-vs-light-gbm-vs-xgboost.html

Wikipedia. (s.f.). *Gradient Boosting*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Gradient_boosting#:~:text=Gradient%20boosting%20is%20a%20machine,prediction%20models%2C%20typically%20decision%20trees.

Wikipedia. (s.f.). *Logistic regression*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Logistic_regression#:~:text=In%20statistics%2C%20the%20logistic%20model,%2Fdead%20or%20healthy%2Fsick.

Wikipedia. (s.f.). *Simplex*. Obtenido de https://en.wikipedia.org/wiki/Simplex