Proyecto Final
Segunda Entrega
Ciencia de Datos Aplicada

Universidad de los Andes, Bogotá, Colombia

Ivan Saavedra Villamil, Salomón Novoa Montenegro, Julián Sanabria Mejía

{ir.saavedra232, s.novoa485, jj.sanabria}@uniandes.edu.co

Limpieza y preparación de datos

Objetivo: Calcular la propensión de que las empresas queden en mora

- Reunión con el director de riesgo de Sempli
  - Variables descartadas
    - IDs, variables semejantes, nulos, alto nivel de detalle, mora de cada mes
  - Créditos descartados
    - covid
  - Valores imputados
    - Edad empresarios, año de constitución
  - Valores atípicos
    - Outliers ignorados

Selección del modelo

Hiper-parámetro	Random Forest
random_stat	0
n_estimators	1000
class_weight	balanced

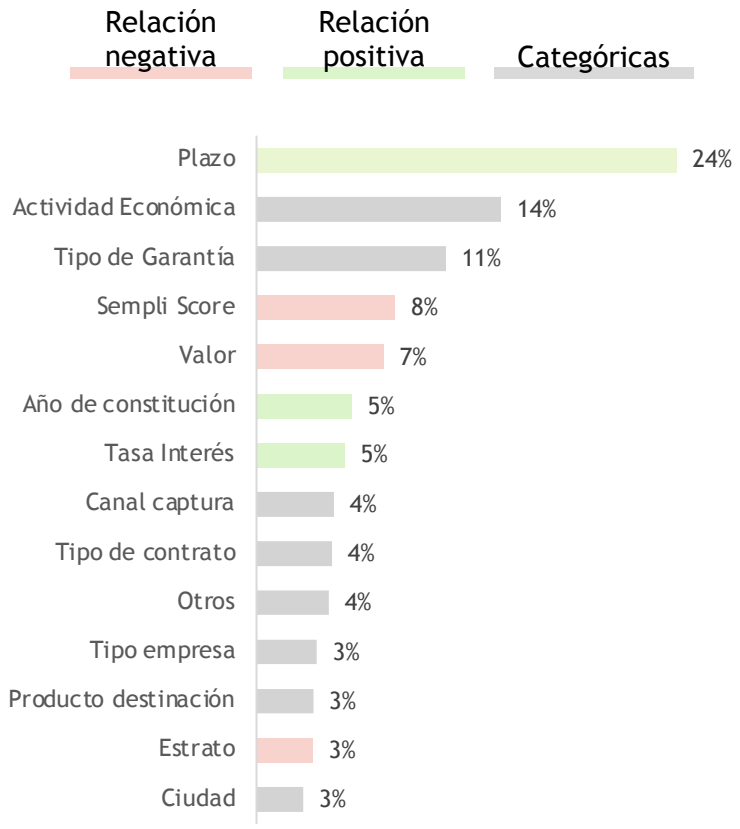
Hiper-parámetro	SVC
kernel	linear
class_weight	balanced

Hiper-parámetro	Logistic Regression
penalty	l2
solver	sag
class_weight	balanced
random_state	80
max_iter	23

- Para class_weight, dado que las clases estaban desbalanceadas (70% - 30%), se utilizó el valor de balanced.

- Se probaron hasta nivel 4 de polinomialidad, pero en lugar de mejorar el resultado del Recall, disminuyó considerablemente por lo cual se utilizó el grado 1.

Modelo – Importancia de Variables



Algoritmo	ENTRENAMIENTO			PRUEBAS		
	Precisión	Recall	F1 Score	Precisión	Recall	F1 Score
Random Forest	0.4752	0.7023	0.5582	0.4344	0.6154	0.5093
SVC	0.4532	0.6804	0.5440	0.4148	0.6090	0.4935
Logistic Regression	0.3868	**0.8099**	0.5236	0.3664	**0.7564**	0.4937

Recall igual a 0.756 muestra que que el modelo identifica de manera correcta el 75.56% de los clientes que realmente caen en mora

Las variables qué más discriminan la propensión de quedar en mora son:

1. Plazo con 24%
2. Actividad económica 14%
3. Tipo Garantía 11%

Conclusiones

- Si bien la data que nos compartió Sempli era abundante en variables y registros, al hacer el EDA encontramos que buena parte del dataset se resumía tanto en créditos como en variables.
- A pesar de que la regresión logística es un considerado uno de los algoritmos más sencillos dentro del *Machine Learning*, es una herramienta poderosa que genera buenos resultados
- Perfil de cliente:
  - Los créditos con plazo más largos tienen mayor propensión a no pagar.
  - Las actividades económicas Commerce, Retail, Merchants tienen menor propensión a incurrir en mora. Mientras que Tourism, Hotels and Restaurants tienen mayor propensión a no pagar.
  - Los créditos con tipos de garantía Prenda Maquinaria, Inventarios o Equipos tienen mayor propensión a no pagar. Mientras que los créditos con tipo de garantía Pagaré tiene menor propensión a incurrir en mora.
  - Los clientes con valores de Sempli Score más bajo tienen mayor propensión a no pagar.