Préstamos garantizados por la SBA

Hernán Galletti

Mayo 2023

U.S. Small Business Administration

Contexto y problema comercial

- La Administración de Pequeñas Empresas de los Estados Unidos (SBA) se fundó en 1953 sobre el principio de promover y ayudar a las pequeñas empresas en el mercado crediticio de los Estados Unidos.

- La SBA ayuda a estas pequeñas empresas a través de un programa de garantía de préstamo, diseñado para alentar a los bancos a otorgarles préstamos. La SBA actúa como un proveedor de seguros al banco al asumir parte del riesgo garantizando una parte del préstamo. En el caso de que un préstamo entre en incumplimiento, la SBA cubre el monto garantizado.

Objetivo

- Implementar un modelo de aprendizaje automático que permita predecir con un margen de error razonable si la empresa solicitante va a poder cancelar el préstamo sin inconvenientes.

Descripción de los datos utilizado

El dataset proviene de la SBA y consta de una lista de datos de préstamos aprobados entre los años 1966 y 2014. Cada registro de esta lista corresponde a un préstamo y consta de los siguientes campos:

Campo	Descripción
LoanNr_Chk Dgt	Identificador-primary key 97%
Name	Nombre del prestatario
City	Ciudad del prestatario
State	Estado de EE.UU. del prestatario
Zip	Código postal del prestatario
Bank	Nombre del banco
BankState	Estado de origen del banco
NAICS	Codigo de clasificación de la industria de EE.UU.

Campo	Descripción
ApprovalDate	Fecha de emisión del compromiso emitido por SBA
ApprovalFY	Año fiscal del compromiso
Term	Plazo del préstamo en meses
NoEmp	Número de empleados de la empresa
NewExist	1=Negocio existente, 2=Nuevo negocio
CreateJob	Número de puestos de trabajos creados
RetainedJob	Número de puestos de trabajo retenidos

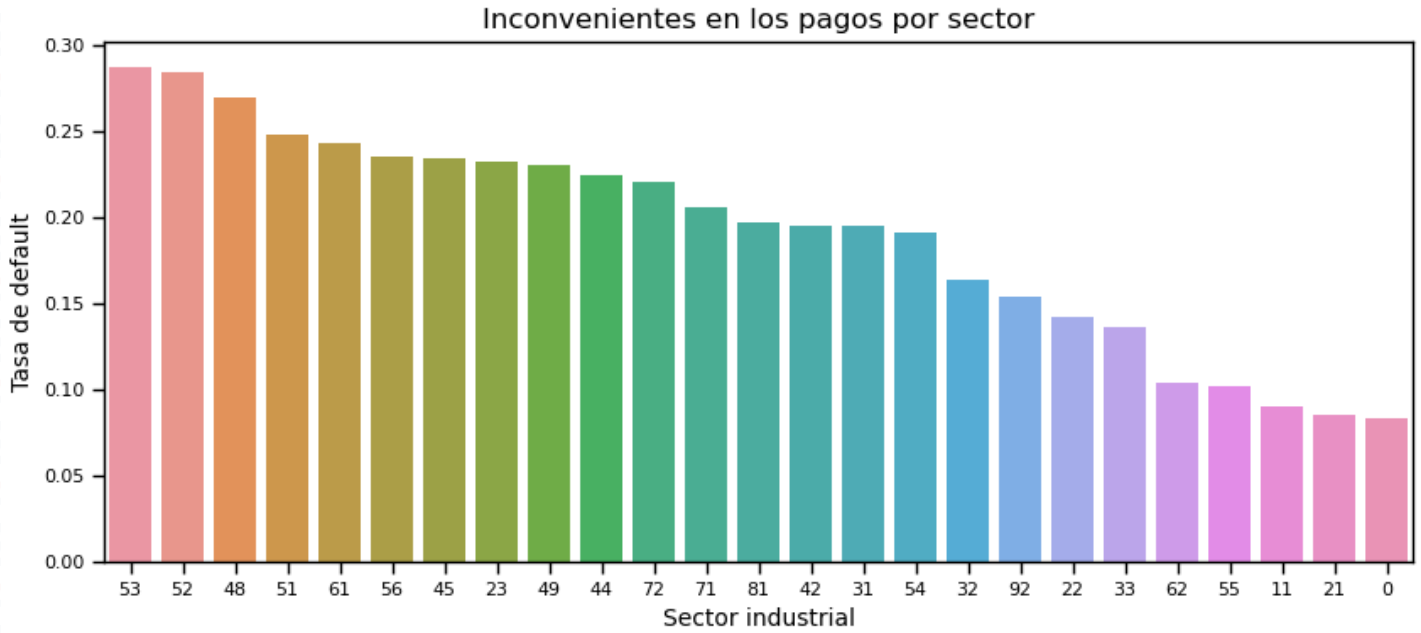
Campo	Descripción
FranchiseCode	Código de franquicia, (00000 o 00001)=Sin franquicia
UrbanRural	1=Urbano, 2=rural, 0=indefinido
RevLineCr	Linea de crédito renovable Y=Si, N=No.
LowDoc	Programa de préstamos LowDoc: Y=Si y N=No
ChgOffDate	Fecha en que se declara que un préstamo está en mora
DisbursementDate	Fecha de pago (entrega del préstamo)

Campo	Descripción
DisbursementGross	Monto desembolsado/pagado bruto (por entrega préstamo)
BalanceGross	Cantidad bruta pendiente de devolución (al momento de confección del dataset)
MIS_Status	Estado del préstamo: Con inconvenientes = CHGOFF, pagado en su totalidad = PIF
ChgOffPrinGr	Importe cancelado (no devuelto)
GrAppv	Monto bruto del préstamo aprobado por el banco
SBA_Appv	Monto garantizado del préstamo aprobado por la SBA

El campo MIS_Status es el campo que se intenta predecir en el modelo y lo llamamos “variable objetivo”.

Preguntas

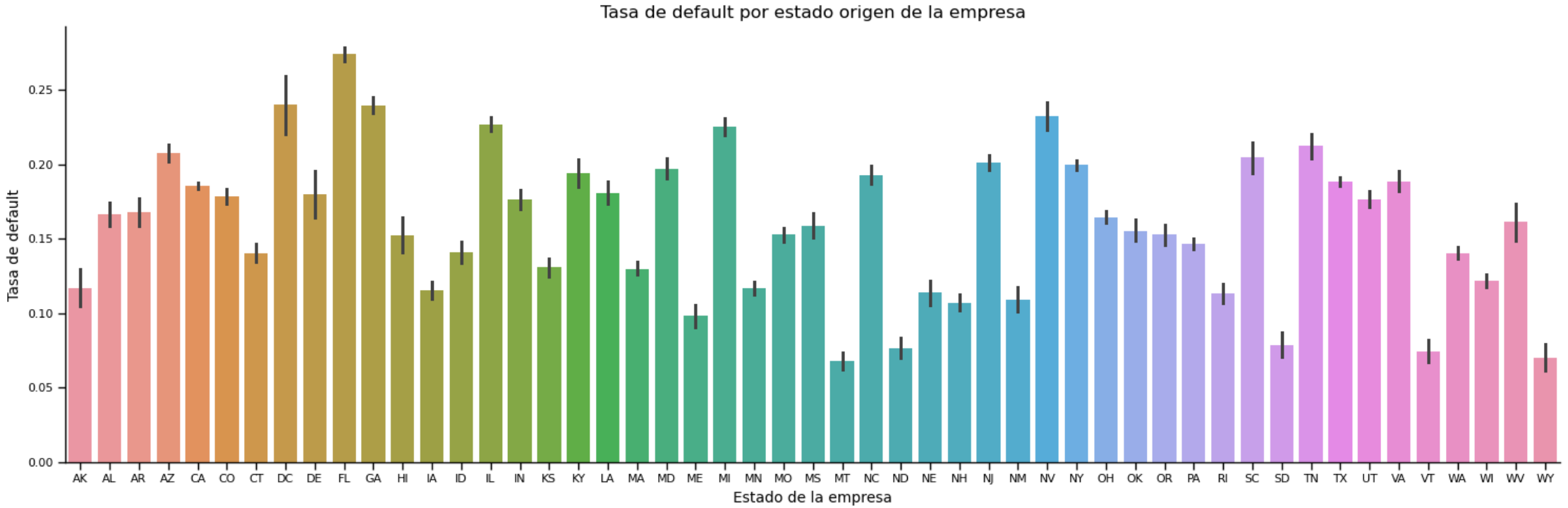
. ¿Cuáles son los sectores industriales a los que pertenecen las empresas que tienen más inconvenientes en devolver el préstamo? ¿Y los sectores en los cuales las empresas tienen menos inconvenientes?



Sector	Descripción
11	Agricultura, silvicultura, pesca y caza
21	Minería, canteras y extracción de petróleo y gas
22	Utilidades
23	Construcción
31–33	Fabricación
42	Comercio al por mayor
44–45	Comercio al por menor
48–49	Transporte y almacenamiento
51	Información
52	Finanzas y Seguros
53	Inmobiliaria y alquiler y arrendamiento
54	Servicios profesionales, científicos y técnicos.
55	Gestión de empresas y empresas.
56	Servicios administrativos y de apoyo y gestión y remediación de residuos
61	Servicios educativos
62	Asistencia sanitaria y asistencia social
71	Artes, entretenimiento y recreación
72	Servicios de alojamiento y alimentación
81	Otros servicios (excepto administración pública)
92	Administración Pública

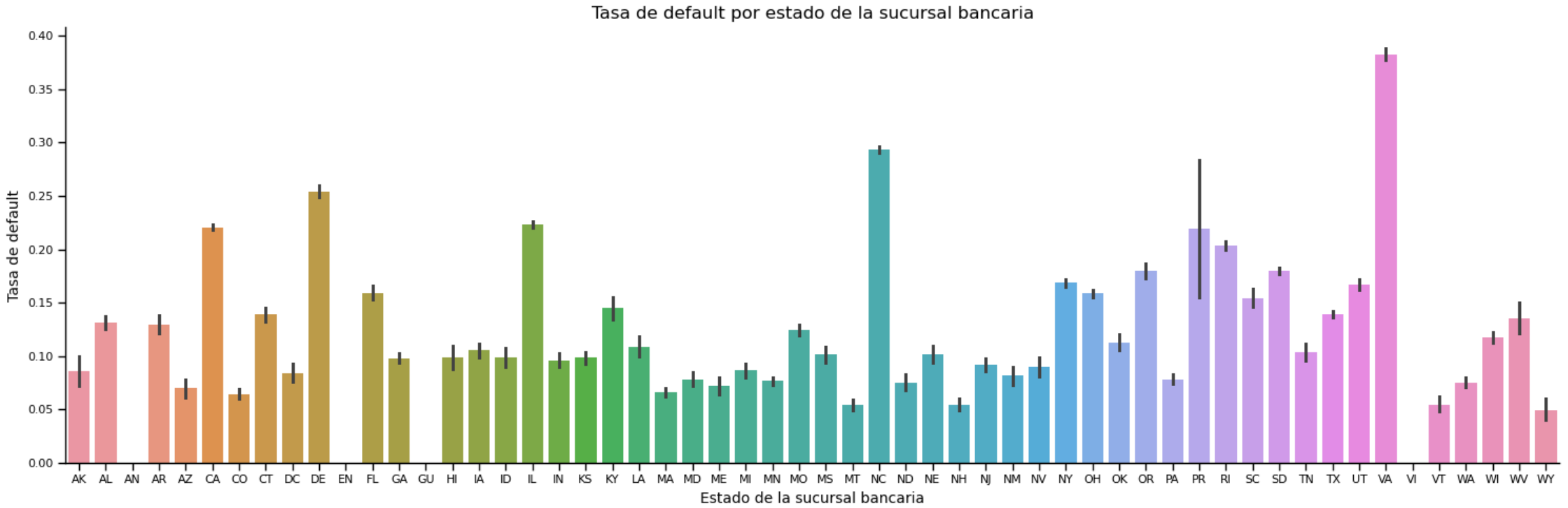
Los sectores "Finanzas y seguros" e "Inmobiliario y alquiler y leasing" tienen las tasas de default más altas, cerca del 30%, mientras que los sectores "Agricultura , silvicultura, pesca y caza" y "Minería, cantería y extracción de petróleo y gas", respectivamente) tienen la tasas más bajas, inferiores al 10%. ("0" representa a las empresas que no se clasificaron en ningún sector).

. ¿Importa el estado de donde proviene la empresa para evaluar el riesgo de impago?



Vemos que hay estados donde el riesgo de impago es significativamente mayor que en otros. Esto puede deberse a las leyes laborales, la riqueza del estado y factores culturales.

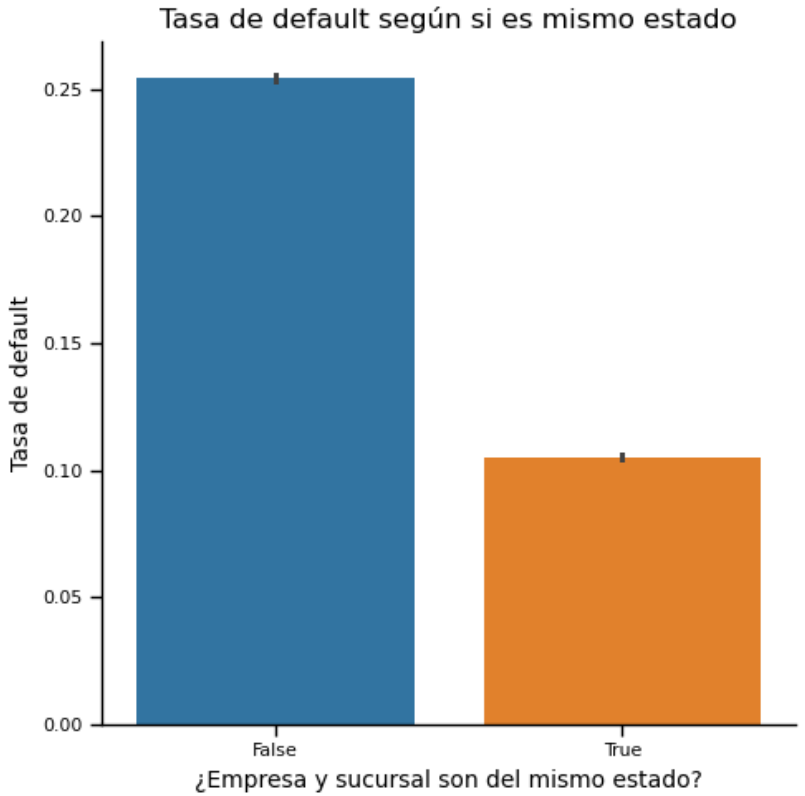
. ¿Pasará lo mismo con el estado de la sucursal bancaria?



Es interesante ver que las diferencias en la tasa de morosidad con respecto al estado de la sucursal bancaria son incluso mayores que las diferencias con respecto al estado de la empresa. Por ejemplo, si el banco es de VA (Virginia) los inconvenientes rondan el 40%. Un valor realmente alto.

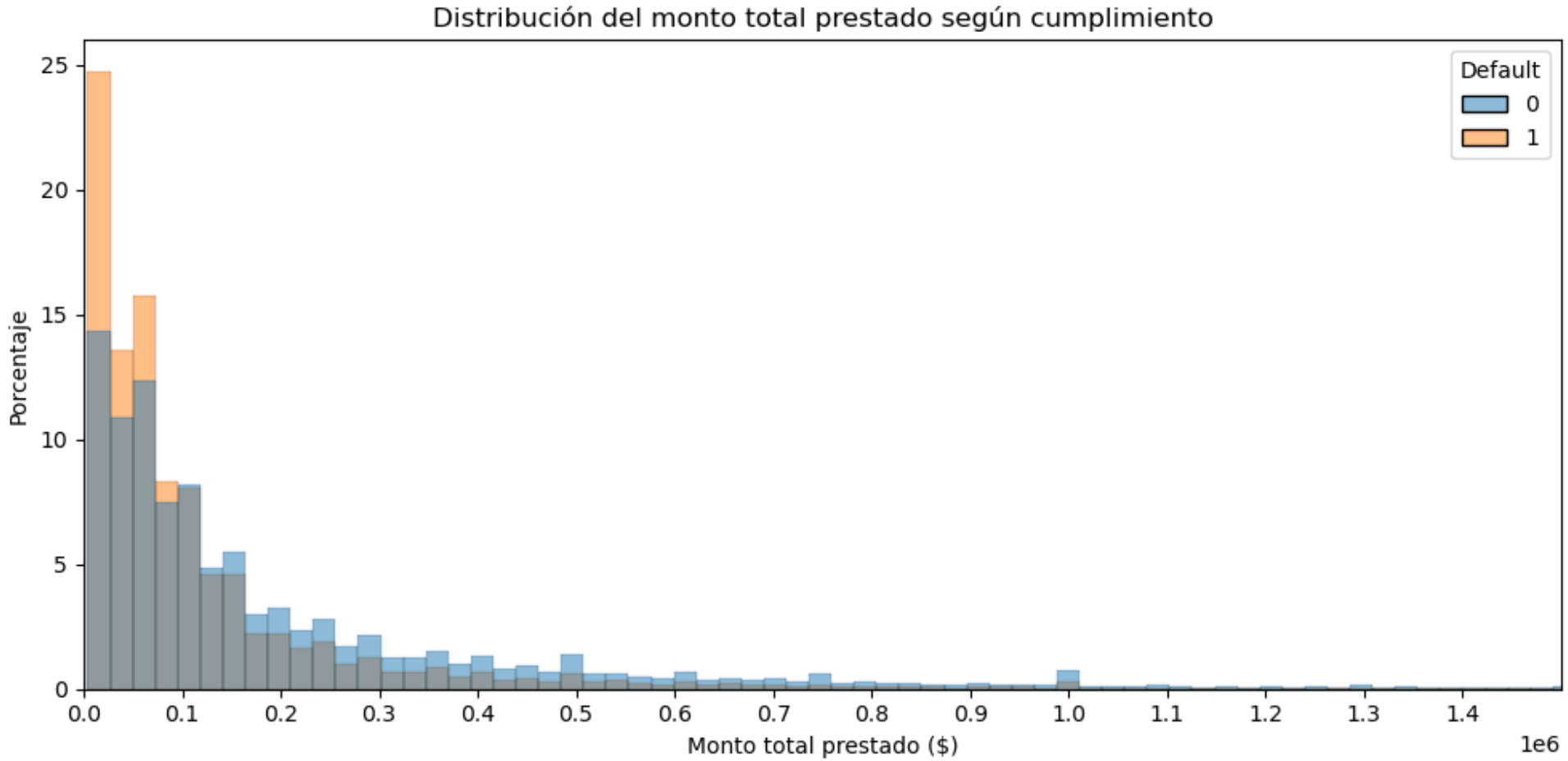
Pero lo que resulta más interesante es ver que sucede si miramos los préstamos donde la empresa y la sucursal radican en el mismo estado, y los que no.

¡La tasa de morosidad cuando el estado de la sucursal bancaria es diferente al de la empresa es 2,5 veces mayor!



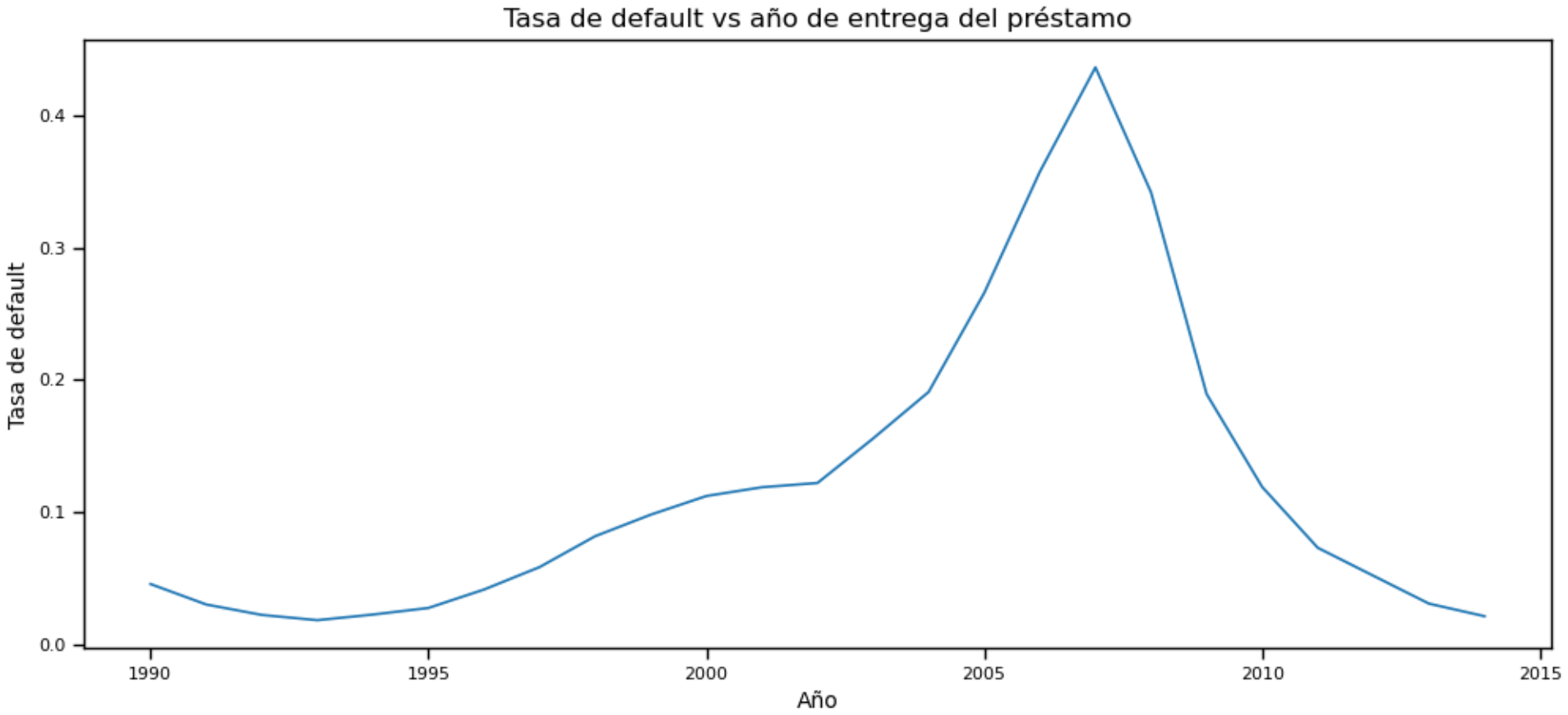
. Otra pregunta interesante para hacerse es si la tasa de default depende del monto total prestado o no.

La respuesta es que sí, pero quizá no de la manera que uno esperaría:



Este gráfico nos muestra que los desembolsos más chicos representaron un porcentaje mayor para los préstamos que entraron en default con respecto a los que no.

.¿Influye el contexto económico?



Claramente hay un pico en los inconvenientes de pago de los préstamos durante la gran recesión.

Modelado del problema

- Es un problema de clasificación: Decidir si un préstamo tuvo problemas en su devolución.
- Una vez obtenido el modelo, se utiliza para decidir la aprobación de un préstamo por la SBA.

Se propusieron los siguientes modelos de aprendizaje automático (supervisado):

- Árbol de decisión.
- Regresión logística.
- K-Nearest Neighbors.
- Random Forest.
- XGBoost.

Feature engineering

- Hay variables que se descartan porque sus valores no se tienen en el momento de la aprobación del préstamo, o porque no tienen mucho valor predictivo. Por ejemplo: 'Name', 'City', 'Zip', 'ApprovalDate', 'ApprovalFY', 'ChgOffDate', 'DisbursementDate', 'DisbursementGross', 'BalanceGross', etc.
- Se crearon nuevas variables como posibles variables a formar parte del modelo:

Campo	Descripción
Is_franchise	Indica si la empresa es una franquicia.
Real_Estate	Indica si el préstamo estado respaldado por bienes raíces (corresponden a los préstamos a plazos iguales o mayores a 20 años).
SameState	Indica si el estado de origen de la empresa y el de la sucursal del banco son el mismo.
BankClass	Es una clasificación de los bancos según el nivel de riesgo. La clasificación se basa en la tasa de default.

- Se hizo una segunda versión del set de datos formada solo por los registros a partir de 1990 con el fin de ver si se logra una mejor performance con los préstamos más recientes. Lllamamos 'all years' y 'last years' para distinguirlos.

- Se utilizó proceso SFS para seleccionar las variables que mejor predicen en cada algoritmo. A partir de ello se utilizaron dos estrategias de selección de variables:

1. Basado en un sistema de votación. Las variables más elegidas son consideradas las más importantes y las únicas que forman parte del conjunto de entrenamiento de los algoritmos.
2. Se guardan las variables elegidas por cada algoritmo y para cada una de las dos versiones del set de datos. Este sistema es mejor al anterior en cuanto a performance pero es más complicado de implementar.

Selección de hiperparámetros

- Con las variables obtenidas en cada estrategia, se hizo la búsqueda de los hiperparámetros de cada algoritmo con mejor rendimiento.

Aclaración: Tanto el algoritmo de selección de variables como el de búsqueda de los mejores hiperparámetros se aplicaron a una muestra de los datos. Esto es debido a que necesitan mucho procesamiento.

Comparación entre los modelos

Los resultados obtenidos son los siguientes:

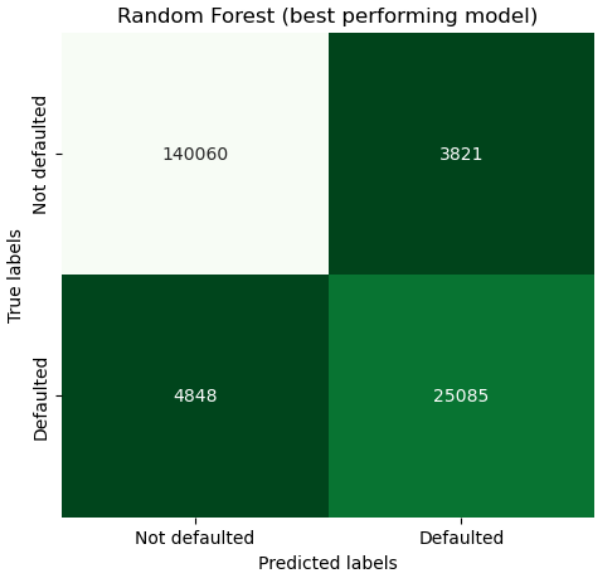
	data set version	variable selection strategy	hyperparameters	f1 (test)	overfitting rate	elapsed time
RandomForest Classifier	last years	2	found in strategy 2	0.852666	13.10 %	1043.92 seg.
XGBClassifier	last years	2	default with fixed seed	0.849161	0.97 %	22.83 seg.
XGBClassifier	last years	none	default	0.847756	0.91 %	30.37 seg.
XGBClassifier	last years	1	default with fixed seed	0.844535	0.67 %	26.94 seg.
XGBClassifier	last years	2	found in strategy 2	0.842547	0.36%	67.77 seg.

```
RandomForestClassifier
RandomForestClassifier(n_estimators=OrderedDict([('criterion', 'log_loss'),
                                                 ('max_depth', 21),
                                                 ('max_features', 0.7),
                                                 ('n_estimators', 350),
                                                 ('random_state', 4865)]))
```

Con las siguientes variables:

'Term', 'NewExist', 'CreateJob', 'RetainedJob', 'UrbanRural', 'RevLineCr', 'LowDoc', 'Is_franchise', 'Real_Estate', 'SameState_False', 'SameState_True', 'SameState_Unknown', 'State', 'BankState', 'Sector', 'BankClass'

- Tiene mejor tasa de positividad (recall): 83,8% (contra 82,8%).
- Tiene la mejor performance con la métrica elegida (F1-score): 0,853 (contra 0,849).
- Tiene considerable overfitting: 13,10%.




```
XGBClassifier
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=None, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=None, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=100, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=4865, ...)
```

Con las siguientes variables:

'Term', 'NoEmp', 'NewExist', 'CreateJob', 'RetainedJob', 'UrbanRural', 'RevLineCr', 'LowDoc', 'Is_franchise', 'SameState_False', 'State', 'BankState', 'Sector', 'BankClass'

- Tiene mejor precisión: 87,13% (contra 86,78%).
- Necesita mucho menos tiempo de procesamiento: 22,83 seg. (contra 1043,92 seg.). Es decir, 45,73 veces menos.
- No presenta prácticamente overfitting: 0,97%.