Clasificación

0. Análisis Exploratorio Inicial

Imprimimos todas las columnas originales de Data_2 para dejar evidencia de cómo venían nombradas en inglés antes de cualquier cambio.

Aplicamos un diccionario de renombrado a español solo para las columnas existentes, confirmamos el resultado imprimiendo las columnas ya estandarizadas y dejamos claro que la variable objetivo es loan_status, que ahora se llama estado_prestamo.

La tipología quedó corregida y consistente, separando correctamente variables numéricas y categóricas para preparar el modelado con estado_prestamo como target, donde id_prestamo, monto_prestamo y plazo_meses permanecen como int64 por ser conteos o montos enteros, mientras tasa_interes, cuota_mensual, ingreso_anual, ratio_deuda_ingreso, moras_ultimos_2_anios, utilizacion_credito_revolvente y bancarrotas_publicas quedan como float64 porque pueden tener decimales o venir con valores faltantes, y al mismo tiempo todas las variables de texto que definen segmentos o etiquetas se convierten a category para compactar memoria y asegurar codificación estable, incluyendo calificacion_crediticia, ocupacion, antiguedad_laboral, tenencia_vivienda, verificacion_ingreso, estado_prestamo, plan_pagos, proposito_prestamo, estado_residencia y tipo_solicitud.

1. Análisis Estadístico Inicial (Descriptivo)

1.1 Análisis Descriptivo Numérico

Excluimos id_prestamo del análisis porque es solo un identificador único (una llave) y no contiene información explicativa sobre el estado_prestamo.

Vemos una cartera con montos y plazos estandarizados porque monto_prestamo se concentra en valores redondos y plazo_meses casi solo toma 2 opciones, debe ser politica del banco, 36 meses con 74.6% y 60 meses con 25.4%. Esto sugiere productos predefinidos y deducimos que una parte del comportamiento de estado_prestamo puede estar explicado por el tipo de producto y no únicamente por el perfil del cliente.

Detectamos colas y outliers fuertes en variables clave. ingreso_anual tiene mediana 60,000 pero llega hasta 6,000,000 y utilizacion_credito_revolvente tiene p99 cerca de 98.5 pero un máximo de 5,829, lo que sugiere registros anómalos o escala inconsistente. Deducimos que el promedio queda inflado por pocos casos extremos y que para analizar estado_prestamo conviene usar transformaciones como log en ingreso y algún control de extremos en utilización para evitar que el modelo aprenda ruido.

Observamos que las variables de evento están dominadas por cero. moras_ultimos_2_anios tiene 89.31% en 0 y bancarrotas_publicas 94.66% en 0, así que su señal probablemente viene de diferenciar 0 contra mayor a 0 más que de los niveles finos.

Los faltantes son bajos pero no nulos en algunas numéricas, destacando bancarrotas_publicas con 1.61% y luego utilizacion_credito_revolvente con 0.14%, mientras ratio_deuda_ingreso y moras_ultimos_2_anios están por debajo de 0.06%.

1.2 Análisis Descriptivo de Variables Categóricas

Vemos que calificacion_crediticia está bien contenida y es usable tal cual para modelar, porque solo tiene 7 categorías y además la distribución tiene forma de escalera con mayor masa en B 30.47%, A 24.49% y C 20.53%. Deducimos que esta variable va a ser de las más informativas para predecir estado_prestamo, porque es un resumen directo del riesgo y además no tiene un long tail problemático. También notamos que las colas E 7.20%, F 2.71% y G 0.80% son pequeñas pero no inexistentes, así que el modelo debería poder aprender que los grados bajos son raros pero relevantes.

Mientras, ocupacion es el principal foco de limpieza y recodificación, porque su cardinalidad es 15048 y eso es típico de texto libre con miles de variantes. El top 1 ya es nan con 5.95% y después lo más frecuente apenas llega a 0.27%, lo que confirma que casi todas las categorías son extremadamente raras. En práctica esto no se puede one hot sin explotar el espacio y sin sobreajuste.

La antiguedad_laboral está bien estructurada con 12 categorías, con un patrón lógico donde 10+ years pesa 21.34% y luego hay escalones hasta 9 years, y un nan de 2.52% que sí hay que tratar.

Las tenencia_vivienda tiene 6 categorías pero en realidad 3 dominan casi todo, RENT 55.08%, MORTGAGE 36.91% y OWN 7.77%, y el resto es ruido marginal OTHER 0.22%, nan 0.02% y NONE 0.01%. Deducimos que NONE es sospechoso y probablemente debe recodificarse a OTHER o a Missing, porque es un valor que suele venir de inconsistencias. También deducimos que verificacion_ingreso está limpia y bien balanceada, Not Verified 41.47%, Verified 32.47% y Source Verified 26.04%, lo cual es perfecto para capturar diferencias de riesgo sin colas raras.

Con respecto a la target, tiene como valores únicos Fully Paid 85.20% frente a Charged Off 14.78% y un nan residual 0.02% que debe imputarse o eliminarse porque no puede quedar en la variable objetivo.

2. Análisis de Calidad de los Datos (Diagnóstico)

2.1 Calidad de Variables Numéricas

Primero imputamos los faltantes de todas las columnas numéricas usando la mediana de cada variable. Esto es consistente con variables financieras con colas largas porque la mediana no se mueve tanto con extremos, y el resultado final confirma que el missing numérico quedó en 0% para todas las variables.

Después definimos una regla de winsorización por percentiles 1% y 99% solo para las variables continuas donde había colas claras, como monto_prestamo, tasa_interes, cuota_mensual, ingreso_anual, ratio_deuda_ingreso y utilizacion_credito_revolvente. A la vez, nosotros decidimos no winsorizar plazo_meses, moras_ultimos_2_anios y bancarrotas_publicas porque son variables discretas de conteo o con muchos ceros y un recorte por percentiles puede borrar señal real de eventos raros.

Vemos que la post limpieza efectivamente recorta las colas altas donde más dolía. ingreso_anual pasa a tener p99 cercano a 248,003 en lugar de dejar que el máximo de 6,000,000 domine, y utilizacion_credito_revolvente queda con p99 98.5 aunque el máximo original era 5,829. Deducimos que el modelo va a ser más estable porque estos extremos ya no van a forzar splits raros o coeficientes desproporcionados.

También revisamos outliers con el criterio IQR y vemos algo importante. Aun con winsorización, ingreso_anual sigue marcando 5.33% de outliers y monto_prestamo 3.30% y cuota_mensual 2.72%, lo cual es coherente porque el IQR detecta desviaciones incluso cuando el p99 ya fue cortado. En cambio, ratio_deuda_ingreso queda sin outliers por IQR, lo que sugiere que su rango ya era razonable y que el recorte por percentiles fue más una medida preventiva.

2.2 Calidad de Variables Categóricas

Los faltantes reales sí están bien identificados como NaN, por eso aparecen porcentajes claros en ocupacion 5.95% y antiguedad_laboral 2.52%, mientras que en otras columnas el missing es casi cero. En cambio, la búsqueda de missing “escrito” como texto no encontró nada, porque para todas las columnas salió False 19908, así que prácticamente no hay casos con "", "?", "NA", "null" dentro de estas categóricas, ya venían limpios y convertidos a NaN desde antes.

Concluimos que ocupacion es el gran problema: tiene 15048 valores distintos y además 5.95% de missing. Eso significa que casi todo son categorías súper raras y encima hay ruido por diferencias de escritura como US Army vs US ARMY vs us army o Self vs self vs Self Employed. Si hacemos one hot directo, el modelo puede memorizar nombres y

sobreajustar, por eso aquí toca normalizar texto, unificar variantes y agrupar la cola larga en Other con una regla simple como top N o mínimo 0.5% para conservar estabilidad.

Detectamos suciedad puntual en tenencia_vivienda y tipo_solicitud. tenencia_vivienda está dominada por RENT 55.08% y MORTGAGE 36.91%, pero aparecen categorías residuales como OTHER 0.22% y NONE 0.005% más 3 NaN, y NONE casi seguro es un error o valor mal codificado, así que conviene convertirlo a OTHER y Missing para no crear una dummy de 1 registro.

Nosotros confirmamos cómo queda el problema de modelado con la target estado_prestamo, está casi completa y está desbalanceada, 85.20% Fully Paid contra 14.78% Charged Off. Además, plan_pagos no sirve porque es constante 100% s y se elimina; proposito_prestamo tiene una categoría dominante debt_consolidation 47.15% y una muy rara renewable_energy 0.26%, así que las raras se pueden agrupar en Other; y en estado_residencia, los valores tipo 106xx y 951xx con 1 caso cada uno son claramente inconsistencias y deben pasar a Missing o corregirse con una regla para no contaminar el modelo.

3. Preprocesamiento de Datos

3.1 Variables Numéricas

Esta etapa de imputación no cambia nada en la práctica porque ya no quedaban faltantes en numéricas antes de correr SimpleImputer, porque antes ya se había corregido. La tabla mean vs median solo nos sirve como diagnóstico de forma, vemos brechas grandes en ingreso_anual 69316 vs 60000 y en moras_ultimos_2_anios 0.14 vs 0.00, lo que indica colas y concentración en cero.

Aplicando log a ingreso_anual y la señal clave es que el sesgo cae de 1.772 a 0.012, o sea pasamos de una distribución muy asimétrica a casi simétrica, esto ayuda a que el modelo no sea dominado por pocos ingresos enormes y mejora estabilidad de coeficientes o splits.

3.2 Variables Categóricas

Partimos separando dos conceptos que antes se mezclaban y nos estaba rompiendo la limpieza UNKNOWN para falta de información real y OTHER para información válida pero demasiado rara. Por eso primero estandarizamos texto en todas las categóricas con strip y convertimos tokens típicos de missing como "", "?", "NA", "null" a NA real para que luego la regla sea consistente y todo missing termine en UNKNOWN sin contaminar el bucket de rarezas.

Resolvimos el problema clásico de duplicar significados con other vs OTHER. Como el dataset ya podía traer other como categoría legítima en proposito_prestamo, normalizamos cualquier variante de other en cualquier columna a un único OTHER usando una comparación case-insensitive, así evitamos tener dos “otros” compitiendo y duplicando señal en el modelo.

Tratamos la variable ocupacion como el foco crítico por alta cardinalidad y ruido de formato, y por eso aplicamos un bucketing controlado con top_k y min_pct más permisivos min_pct 0.003 y top_k 200 para no destruir señal. La lógica fue mantener explícitas las ocupaciones más frecuentes y colapsar el resto a OTHER, mientras que todo faltante va a UNKNOWN, con esto evitamos one hot masivo y también evitamos que ocupacion termine siendo 80% OTHER y nos damos cuenta que incluso así pierda utilidad.

Ordenamos variables que tienen estructura natural para que el modelo capture gradientes reales. calificacion_crediticia se convirtió en categórica ordenada de G a A, y antiguedad_laboral se normalizó a un formato consistente 0-1 year hasta 10+ year más UNKNOWN, corrigiendo variantes como < 1 year y years vs year.

Limpiamos variables con valores inesperados y dejamos reglas claras para estabilidad. tenencia_vivienda se llevó a mayúsculas y NONE se trató como missing real para caer a UNKNOWN, estado_residencia se validó con patrón de 2 letras y todo lo que no calza como 106xx se manda a UNKNOWN, tipo_solicitud corrigió el valor residual 1 a Individual para eliminar ruido de 0.01%. Finalmente el target estado_prestamo tenía 3 NA que aparecían como <NA> y nosotros los imputamos a UNKNOWN para no crear una “clase fantasma”.

4. Análisis Exploratorio de Datos (EDA Univariado) y Visualización

4.1 Variables Numéricas: Estadística Descriptiva y Visualización

Vemos productos bastante estandarizados en monto_prestamo y plazo_meses, el histograma de montos muestra picos en valores redondos como 10,000 que además es el valor más frecuente con 7.36% y el plazo es casi binario 36 meses 74.6% frente a 60 meses 25.4%. La deducción es que parte del riesgo en estado_prestamo puede venir del tipo de producto elegido y no solo del perfil, por eso estas variables capturan reglas comerciales reales y deben mantenerse tal cual.

Observamos que tasa_interes tiene dispersión relevante y una cola hacia valores altos, con p50 11.86 y p99 20.99, mientras cuota_mensual acompaña esa heterogeneidad con p50 285.78 y p99 933.13. ingreso_anual queda claramente sesgado a la derecha pero ya acotado por winsorización, p50 60,000 y p99 cerca de 248,003, y además el log reduce el sesgo de 1.772 a 0.012, así que la deducción es usar ingreso_anual_log para modelar sin que pocos altos dominen el ajuste.

El ratio_deuda_ingreso tiene una forma más estable y continua, con centro alrededor de 13 y rango útil hasta p99 26.26, lo que lo hace buen predictor monotónico. En cambio moras_ultimos_2_anios y bancarrotas_publicas son eventos raros con masa en 0, 89.36% y 96.27% respectivamente, así que su señal está en separar 0 vs mayor que 0. utilizacion_credito_revolvente queda prácticamente en 0 a 100 con p99 98.5 y un pico en 0 de 2.29%, la deducción es que el capping al 99% fue clave para eliminar escalas imposibles y dejar una variable interpretable para estado_prestamo.

Vemos que monto_prestamo casi no explica la tasa_interes por sí solo, porque la nube es ancha y no hay una pendiente clara, con franjas verticales que reflejan montos “redondos” predefinidos. Deducción, la tasa está siendo determinada por otras variables de riesgo y por el producto, así que para explicar estado_prestamo conviene pensar en interacciones y no en una relación lineal simple monto → tasa.

Una relación mecánica muy fuerte entre monto_prestamo y cuota_mensual, con bandas diagonales bien marcadas. Deducción, esas bandas salen por combinaciones discretas de plazo_meses y tasa_interes, entonces cuota_mensual no es información independiente sino una función del mismo contrato, y meter las 3 juntas puede generar redundancia fuerte y multicolinealidad en modelos lineales o splits dominantes en árboles.

Observamos que monto_prestamo vs ingreso_anual, ratio_deuda_ingreso, moras_ultimos_2_anios y utilizacion_credito_revolvente no muestran un patrón limpio, pero sí una estructura discreta importante. En ingreso_anual hay más dispersión y un techo visible alrededor de 250,000 por el capping, en moras la mayoría está en 0 y en utilización el rango se concentra 0 a 100.

Vemos que cuota_mensual y ingreso_anual_log tienen una relación positiva clara, a mayor cuota la nube se desplaza hacia ingresos log más altos, pero con mucha dispersión vertical en cada nivel de cuota. Deducción, la cuota captura capacidad de pago pero no de forma determinística porque también depende de plazo_meses y tasa_interes, por eso aparecen “bandas” y una variabilidad grande para la misma cuota.

En moras_ultimos_2_anios se observa el patrón de evento raro, la mayoría está en 0 y los valores altos son pocos y dispersos, y al cruzarlo con ratio_deuda_ingreso no aparece una curva clara, solo una leve concentración de moras mayores cuando el ratio sube, lo que sugiere una señal débil pero útil como bandera más que como relación lineal.

4.2 Variables Categóricas: Frecuencias y Visualización

En tipo_solicitud ya quedó una sola categoría Individual, eso implica varianza casi nula y aporte predictivo cercano a cero, la deducción práctica es excluirla del modelo. Lo mismo con plan_pagos, si es 100% s entonces es constante y solo mete ruido al pipeline.

En estado_prestamo ya se ve el desbalance real del problema, domina Fully Paid y luego Charged Off, y aparece UNKNOWN solo por los registros faltantes que se imputaron. La lectura correcta es que UNKNOWN en el target no es “un tercer tipo de préstamo” sino falta de etiqueta, por eso sirve para control de calidad y para decidir si esos pocos casos se entrenan o se separan.

En proposito_prestamo se confirma una concentración fuerte en debt_consolidation y luego credit_card, y existe un bloque OTHER que agrupa rarezas sin confundirlo con UNKNOWN. Esto es sano porque UNKNOWN representa falta de información y OTHER representa información válida pero demasiado rara, la deducción es que el bucket evita explosión de dummies y mantiene señal de los motivos principales.

En estado_residencia la masa está en CA y NY y el resto cae en pocos estados, el UNKNOWN recoge valores no válidos o faltantes tras el filtro de 2 letras, eso limpia entradas tipo 106xx. En tenencia_vivienda domina RENT y MORTGAGE y OWN es menor, OTHER es pequeño y UNKNOWN casi nulo, y en antiguedad_laboral se ordenó el rango con 10+ year como mayor y UNKNOWN como faltante explícito, en calificacion_crediticia queda una escala clara donde B y A pesan más y G es marginal, todo esto queda listo para modelos lineales o árboles sin reventar cardinalidad.

Ocupacion no conviene incluirla porque, aun después del bucketing, sigue siendo una variable de altísima cardinalidad y muy sucia semánticamente, hay muchas etiquetas que representan lo mismo con distinto formato (mayúsculas, abreviaturas, typos) y una cola enorme de categorías con muy pocos casos, lo que obliga a crear muchas dummies o a colapsar masivamente en OTHER. En ambos escenarios el modelo pierde, si haces one hot se vuelve inestable y tiende a memorizar ruido, y si colapsas mucho terminas con una variable dominada por OTHER y UNKNOWN que casi no discrimina, además es una

variable dependiente del texto y no de un atributo económico estable, por lo que generaliza mal fuera de la muestra y puede introducir sesgos por empleadores específicos.

5. Análisis Exploratorio Bivariado (EDA): Variables Explicativas vs Target

5.1 Variables Numéricas vs Target

El monto no separa el riesgo de forma limpia ni lineal, parece capturar mezcla de producto y selección, por eso conviene tratarlo como señal secundaria y no como el predictor principal.

En plazo_meses sí hay separación clara porque casi todo cae en dos valores, 36 y 60. Fully Paid se concentra en 36, mientras Charged Off tiene mucho más peso en 60, lo que eleva su promedio frente a Fully Paid. Deducción, el plazo está actuando como variable de segmentación del tipo de préstamo o política de originación, y por eso empuja fuerte la probabilidad de Charged Off cuando es 60.

En tasa_interes la separación es consistente, Charged Off aparece desplazado hacia arriba con mediana y media mayores, lo que confirma una relación direccional clara, a mayor tasa mayor riesgo de terminar en Charged Off. En cuota_mensual vuelve el solapamiento, la diferencia promedio es pequeña y no hay corte nítido entre clases, porque la cuota es un resultado de monto_prestamo, plazo_meses y tasa_interes.

En ingreso_anual se ve un desplazamiento claro entre clases del estado_prestamo. Fully Paid tiene una mediana y también una media más altas, mientras Charged Off se concentra en valores de ingreso más bajos y con menor centro de distribución. Hay solapamiento porque los ingresos altos también aparecen en Charged Off, pero la separación por nivel central es consistente , a mayor ingreso menor probabilidad de caer en Charged Off.

En ratio_deuda_ingreso la señal va en la dirección esperada pero es más sutil. Charged Off tiende a mostrar valores centrales algo más altos y una dispersión grande, lo que sugiere que un mayor esfuerzo de deuda está asociado a mayor riesgo, pero no separa tan limpiamente como la tasa de interés o el plazo.

En moras_ultimos_2_anios la variable está fuertemente concentrada en cero para ambas clases, por eso el boxplot casi no se mueve y la señal está en la cola. Aun así, los valores positivos y altos aparecen con más frecuencia relativa en Charged Off y eso la vuelve útil como indicador binario o por cortes, cero frente a mayor que cero y luego niveles más altos.

En utilizacion_credito_revolvente sí hay separación más visible, Charged Off tiene una mediana más alta y más masa en niveles elevados, lo que es coherente con estrés de liquidez, y el histograma confirma una distribución amplia con mediana alrededor de 50.3, por lo que su contribución al modelo será por direccionalidad creciente del riesgo cuando la utilización sube.

La relación entre estado_prestamo y cada variable numérica, donde evento positivo es Charged Off y las tasas que ves son el % de préstamos que terminan en Charged Off dentro de cada bucket. Así, cuando decimos 0.255 en tasa_interes significa 25.5% de estado_prestamo igual a Charged Off en ese grupo.

Vemos que monto_prestamo tiene señal pero no es lineal, en quintiles baja de 14.8% a 12.5% y luego sube hasta 18.8% en el tramo alto. El corte por mediana confirma una señal moderada, por encima de 10,000 el Charged Off sube de 13.9% a 16.0%, útil pero más estable si se modela con buckets o con interacción con plazo_meses y tasa_interes.

Concluimos que plazo_meses es de las señales más fuertes para estado_prestamo porque en la práctica es casi una variable binaria. A 36 meses el Charged Off es 11.6% y a 60 meses salta a 24.0%, diferencia grande y estructural. Los préstamos largos concentran mucho más riesgo o vienen de un segmento distinto, por eso plazo separa fuerte, pero también puede dominar el modelo y tapar señal de ingreso.

La tasa_interes es la más direccional y estable contra estado_prestamo. En quintiles sube de 5.5% a 25.5% Charged Off y por mediana pasa de 9.5% bajo 11.86 a 20.1% sobre 11.86. Esto calza con pricing de riesgo, a mayor tasa el banco ya está cobrando más por un perfil más riesgoso y eso se refleja en más Charged Off, por eso es una variable clave.

Lacuota_mensual aporta menos como predictor independiente. Los quintiles se mueven en un rango estrecho de 13.7% a 17.2% Charged Off y el split por mediana apenas sube de 14.3% a 15.2%. La deducción es que cuota está casi determinada por monto_prestamo, plazo_meses y tasa_interes, entonces suele repetir información y puede traer colinealidad, se usa pero no debería ser el motor principal.

Encontramos una direccionalidad clara en ingreso_anual y en ingreso_anual_log frente a estado_prestamo. A mayor ingreso baja el Charged Off, de 18.5% en el quintil bajo a 11.2% en el alto, y por mediana baja de 17.0% a 12.4%. El log no cambia la historia, solo vuelve la forma más “suave” para el modelo, por eso es mejor para regresión lineal o árboles sin que los extremos manden.

Vemos que ratio_deuda_ingreso y utilizacion_credito_revolvente se comportan como medidas de estrés y suben el riesgo de estado_prestamo. En DTI la tasa sube de 13.4% a 17.1% en quintiles y por mediana de 13.5% a 16.1%, tendencia general creciente aunque con algo de ruido en el medio. En utilización la señal es más fuerte y ordenada, sube de 10.8% a 19.8% y por mediana de 12.0% a 17.6%, lectura directa, más utilización de deuda implica más presión de liquidez y más Charged Off.

Tratamos moras_ultimos_2_anios y bancarrotas_publicas como variables de eventos raros contra estado_prestamo, donde lo importante es 0 vs mayor que 0. Con moras, pasar de 0 a >0 sube Charged Off de 14.7% a 15.9%, efecto real pero moderado porque el grupo con moras es pequeño. Con bancarrotas el salto es grande, de 14.5% a 21.5% cuando hay al menos 1, señal fuerte pero concentrada en pocos casos, así que es útil y a la vez sensible a errores de imputación o limpieza.

5.2 Variables Categóricas vs Target

En calificacion_crediticia la señal contra estado_prestamo es la más limpia y direccional. La tasa de Charged Off sube casi en escalera desde A que es la más baja hasta G que es la más alta, lo que es coherente con una variable que ya resume riesgo. Además, los grupos con más datos son B, A y C, así que esa parte del patrón es estable y confiable para el modelo, mientras que F y G tienen pocos registros y su tasa alta puede ser más volátil.

En antiguedad_laboral la relación existe pero es más suave y con ruido. UNKNOWN aparece con una tasa de Charged Off alta, lo que suele indicar falta de información asociada a mayor riesgo o a perfiles menos trazables.

Entre los años reportados no hay una pendiente perfecta, pero en promedio los extremos tienden a peor desempeño que los tramos intermedios, así que esta variable sirve más como segmentador por rangos que como una escala continua.

En tenencia_vivienda hay diferencias moderadas entre categorías grandes y una alerta clara con categorías pequeñas. MORTGAGE luce con menor tasa de Charged Off que RENT y OWN, mientras que OTHER sale alto pero con prácticamente nada de datos, por lo que no conviene tomarlo como conclusión fuerte y debería agruparse o tratarse con cuidado.

En verificacion_ingreso, la tasa de Charged Off es más alta en Verified, alrededor de 0.17, seguida por Source Verified cerca de 0.15, y más baja en Not Verified alrededor de 0.13. No es una relación “intuitiva” de control y riesgo, suena más a selección y pricing, a quién termina en cada canal de verificación según perfil y producto.

En plan_pagos no hay señal utilizable porque solo aparece una categoría, por lo tanto no separa riesgo y conviene excluirla del modelo.

En proposito_prestamo sí hay diferencias claras por tipo de uso. small_business tiene la tasa más alta, cerca de 0.28, pero con pocos registros, así que es señal fuerte pero menos estable.

Debt_consolidation domina el volumen y tiene una tasa intermedia alrededor de 0.16, por eso pesa mucho en el promedio general, mientras credit_card y car se ven más bajos cerca de 0.11 a 0.12. Conclusión práctica, esta variable sirve para segmentar riesgo por producto, pero hay que vigilar categorías raras y mantener un bucket OTHER bien definido.

En estado_residencia hay variación moderada, no extrema. FL sale con la tasa más alta cerca de 0.18, CA y NJ quedan en la parte alta alrededor de 0.15 a 0.16, y TX aparece más bajo cerca de 0.12, con NY e IL en el medio alrededor de 0.13. Como CA tiene el mayor número de registros, su tasa importa mucho para el comportamiento global.