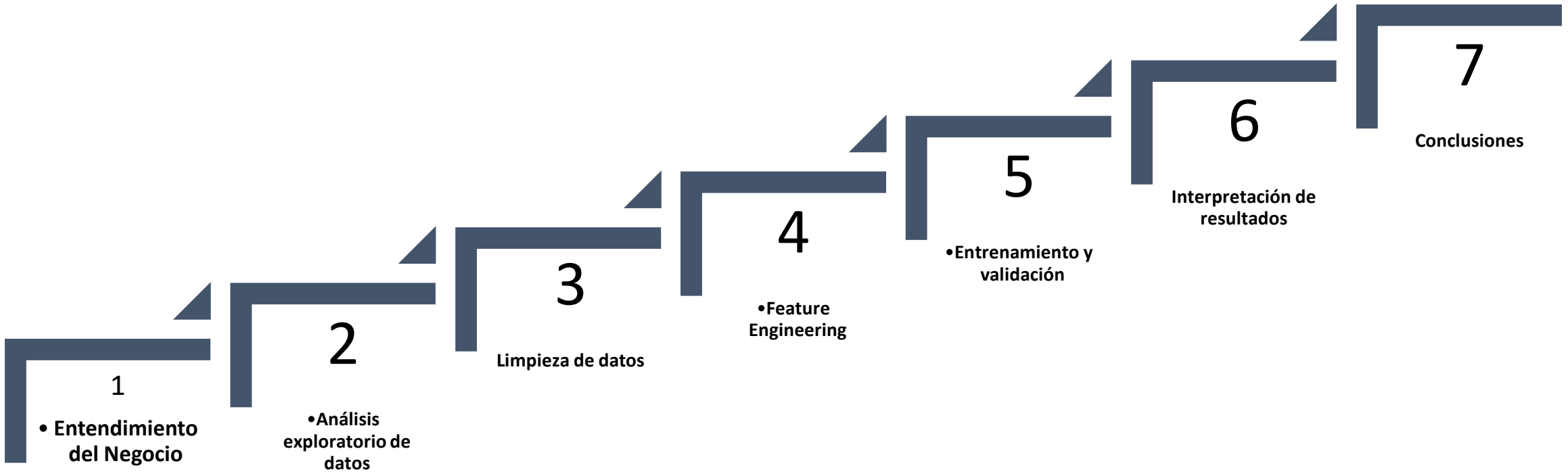
MACHINE LEARNING INMERSION

CREDITSCORING

PREDECIR LA PROPENSIÓN DE EXPERIMENTAR
DIFICULTADES FINANCIERAS EN LOS PRÓXIMOS 2 AÑOS

Silvia Ana Rodriguez Aguirre


1
• Entendimiento del Negocio
2
•Análisis exploratorio de datos
3
Limpieza de datos
4
•Feature Engineering
5
•Entrenamiento y validación
6
Interpretación de resultados
7
Conclusiones

ENTENDIMIENTO DEL NEGOCIO

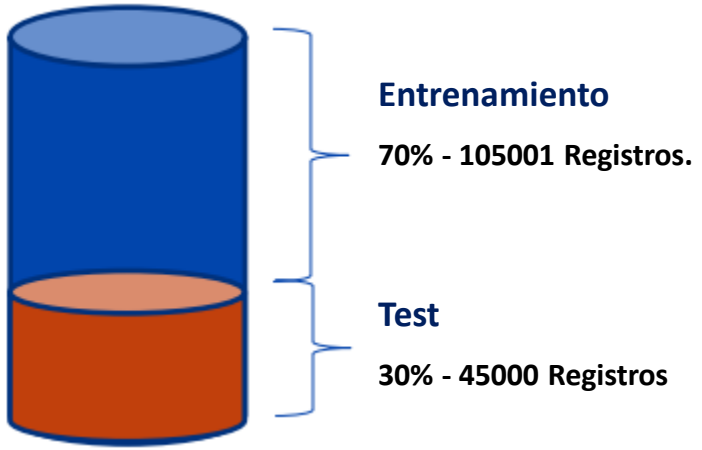
Solicitud de un crédito

Evaluar conducta financiera

Otorgar o no el crédito

Los bancos juegan un papel crucial en las economías de mercado. Deciden quién puede obtener financiación y en qué términos y pueden tomar o deshacer decisiones de inversión. Para que los mercados y la sociedad funcionen, los individuos y las empresas necesitan acceso al crédito. Los algoritmos de calificación crediticia, que adivinan la probabilidad de incumplimiento, son el método que utilizan los bancos para determinar si se debe otorgar o no un préstamo. No olvidemos que por cada cliente que nos deja de pagar, los bancos deben provisionar ese monto y por ende perder capital que podrían usar en otros negocios de inversión.

ANÁLISIS EXPLORATORIO DE DATOS



150001 registros

Variable	Definición	Tipo
RevolvingUtilizationOfUnsecuredLines	Saldo total de tarjetas de crédito y líneas de crédito personales, excepto bienes inmuebles y ninguna deuda a plazos, como préstamos para automóviles, dividido por la suma de los límites de crédito	**Numérica**
age	Edad del prestatario en años	**Numérica**
NumberOfTime30-59DaysPastDueNotWorse	Número de veces que el prestatario ha estado atrasado entre 30 y 59 días, pero no ha empeorado en los últimos 2 años.	**Numérica**
DebtRatio	Pagos mensuales de la deuda, pensión alimenticia, costos de vida divididos por el ingreso bruto mensual	**Numérica**
MonthlyIncome	Ingreso mensual	**Numérica**
NumberOfOpenCreditLinesAndLoans	Número de préstamos abiertos (cuotas como préstamos para automóviles o hipotecas) y líneas de crédito (por ejemplo, tarjetas de crédito)	**Numérica**
NumberOfTimes90DaysLate	Número de veces que el prestatario ha estado atrasado 90 días o más	**Numérica**
NumberRealEstateLoansOrLines	Número de préstamos hipotecarios e inmobiliarios, incluidas líneas de crédito sobre el valor neto de la vivienda	**Numérica**
NumberOfTime60-89DaysPastDueNotWorse	Número de veces que el prestatario ha vencido 60-89 días, pero no ha empeorado en los últimos 2 años.	**Numérica**
NumberOfDependents	Número de dependientes en la familia excluyéndose a sí mismos (cónyuge, hijos, etc.)	**Numérica**
Probability	PROPENSIÓN DE EXPERIMENTAR DIFICULTADES FINANCIERAS EN LOS PRÓXIMOS 2 AÑOS	**TARGET**

LIMPIEZA DE DATOS

Datos nulos

Existe una falta de datos respecto a dos columnas, ingresos y numero de dependientes.

Esta data faltante se reemplazó con valores fuera del rango, el valor de reemplazo es arbitrario. Con la finalidad de que el modelo pueda aprender de los valores faltantes.

Valores discordantes

Dentro del dataset tenemos valores discordantes notorios, por ejemplo una persona con 0 años de edad. Estos valores se dan en pocos casos y se consideró ignorarlos.

Se encontraron datos con valores repetitivos (96,98) dentro de 3 campos, es poco probable que se trate de un error, posiblemente es un código. Se consideraron para el entrenamiento porque en caso de que fueran incorrectos siguen siendo consistentes.

Coeficientes de correlación Pearson

0.992796182594 NumberOfTimes90DaysLate x NumberOfTime60-89DaysPastDueNotWorse

0.98700544748 NumberOfTime30-59DaysPastDueNotWorse x NumberOfTime60-89DaysPastDueNotWorse

0.983602681283 NumberOfTime30-59DaysPastDueNotWorse x NumberOfTimes90DaysLate

0.433958603056 NumberOfOpenCreditLinesAndLoans x NumberRealEstateLoansOrLines

-0.213302578045 age x NumberOfDependents

0.147705318271 age x NumberOfOpenCreditLinesAndLoans

0.125586964573 SeriousDlqin2yrs x NumberOfTime30-59DaysPastDueNotWorse

0.124958961095 MonthlyIncome x NumberRealEstateLoansOrLines

0.124684285213 NumberRealEstateLoansOrLines x NumberOfDependents

0.120046028125 DebtRatio x NumberRealEstateLoansOrLines

FEATURE ENGINEERING

☹

No se pudo diseñar una nueva característica que resulte útil.

Se intentó unificar las 3 características representativas de un pago retrasado, debido a que están linealmente relacionadas, sin embargo no representaron un cambio significativo dentro del modelo de predicción.

ENTRENAMIENTO Y VALIDACIÓN

Score

Se utilizó K-fold cross-validation en el conjunto de datos de entrenamiento para estimar el área bajo la curva ROC de los modelos.

Tipo de Modelo

Se utilizó un regresor de aumento de gradiente, debido a sus principales ventajas:

- Pocas suposiciones sobre los datos para encajar
- No es necesario seguir leyes de distribución específicas
- No hay necesidad de escalar o cambiar
- No hay problema con las funciones correlacionadas
- Aprende rápido

El área estimada bajo la curva ROC para el modelo inicial es 0.8640

Tuning

- Se ajustó el modelo con los siguientes hiper parámetros:
  - max_depth=4
  - n_estimators = 130

El área estimada bajo la curva ROC para el modelo inicial es 0.8649

Reducción de varianza

El aumento de gradiente reduce el sesgo de error, pero puede aumentar la varianza.

Para reducir la varianza, utilizamos la agregación bootstrap.

El área estimada bajo la curva ROC para el modelo inicial es 0.8654

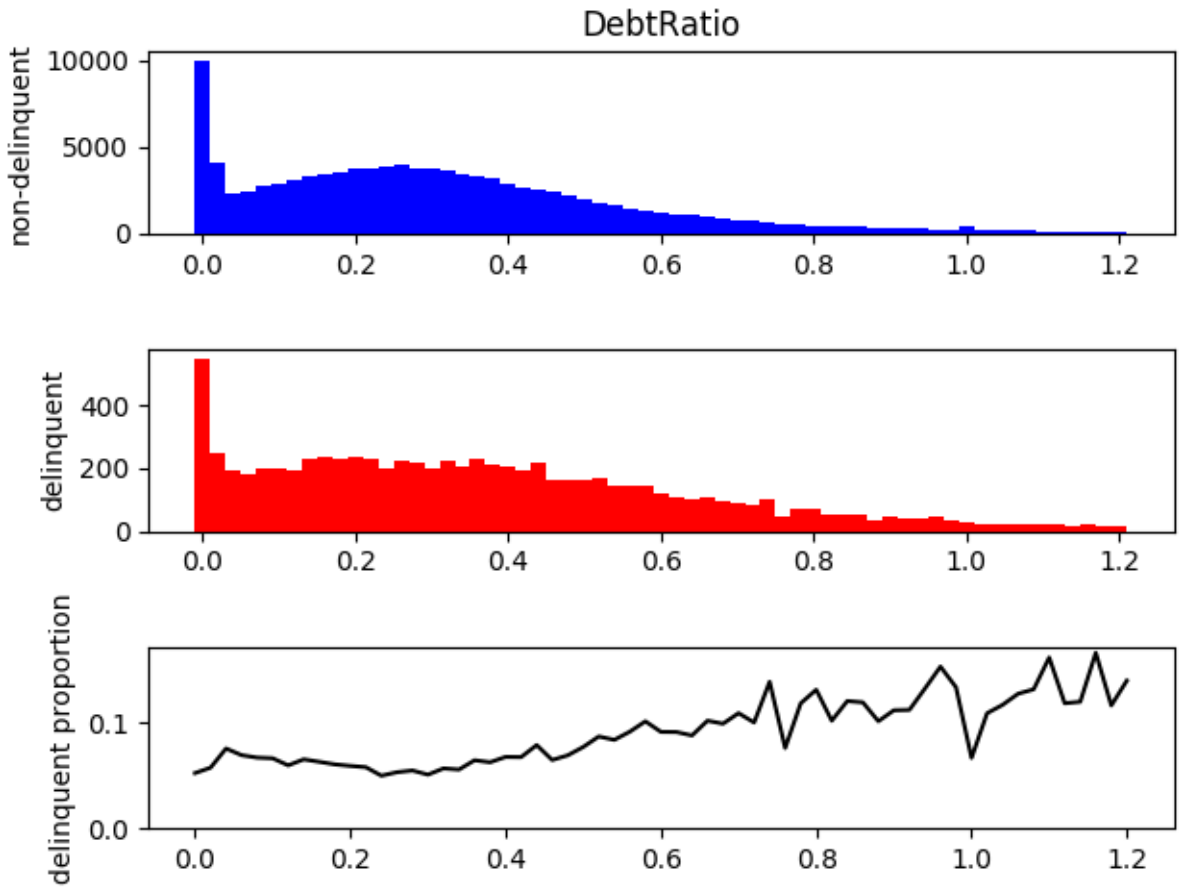
Tuning

- Se ajustó el modelo con los siguientes hiper parámetros:
  - n_estimators = 30

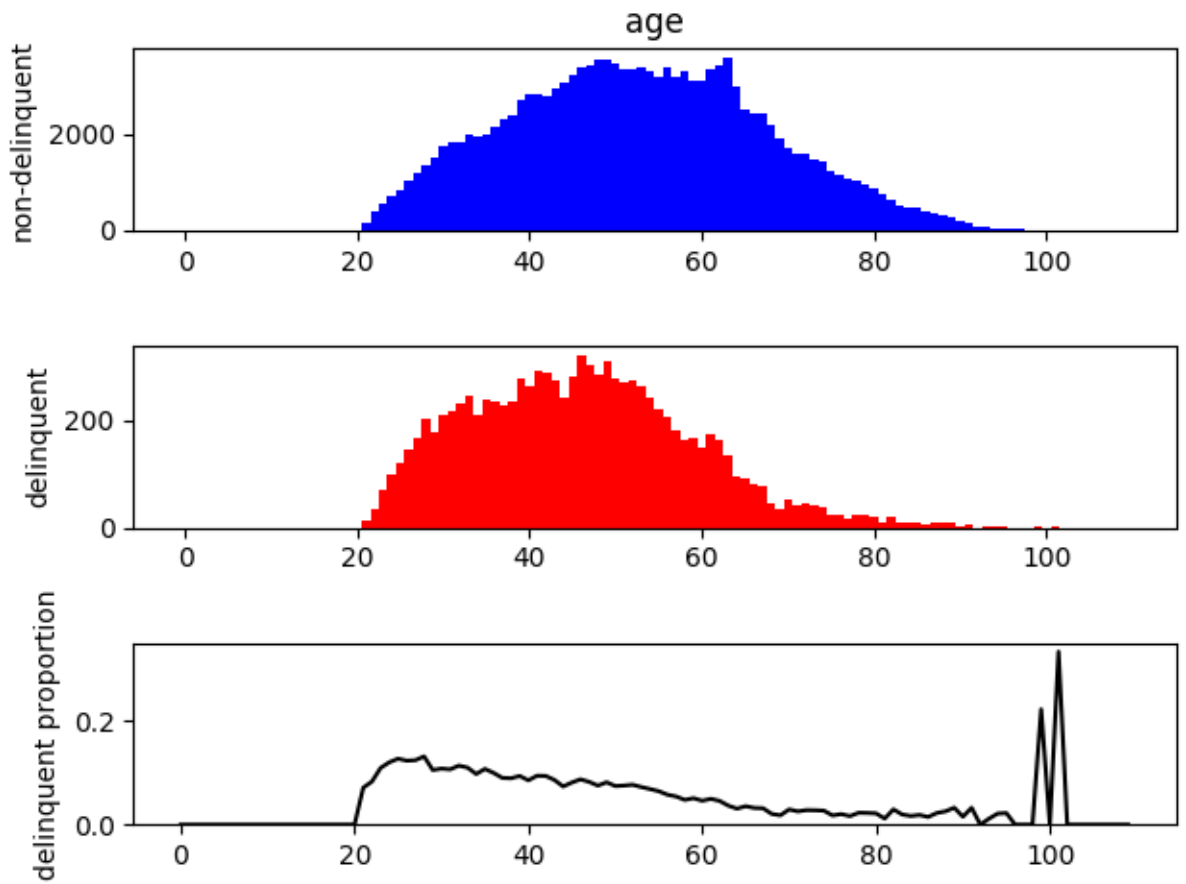
El área estimada bajo la curva ROC para el modelo inicial es 0.8656

INTERPRETACIÓN DE RESULTADOS

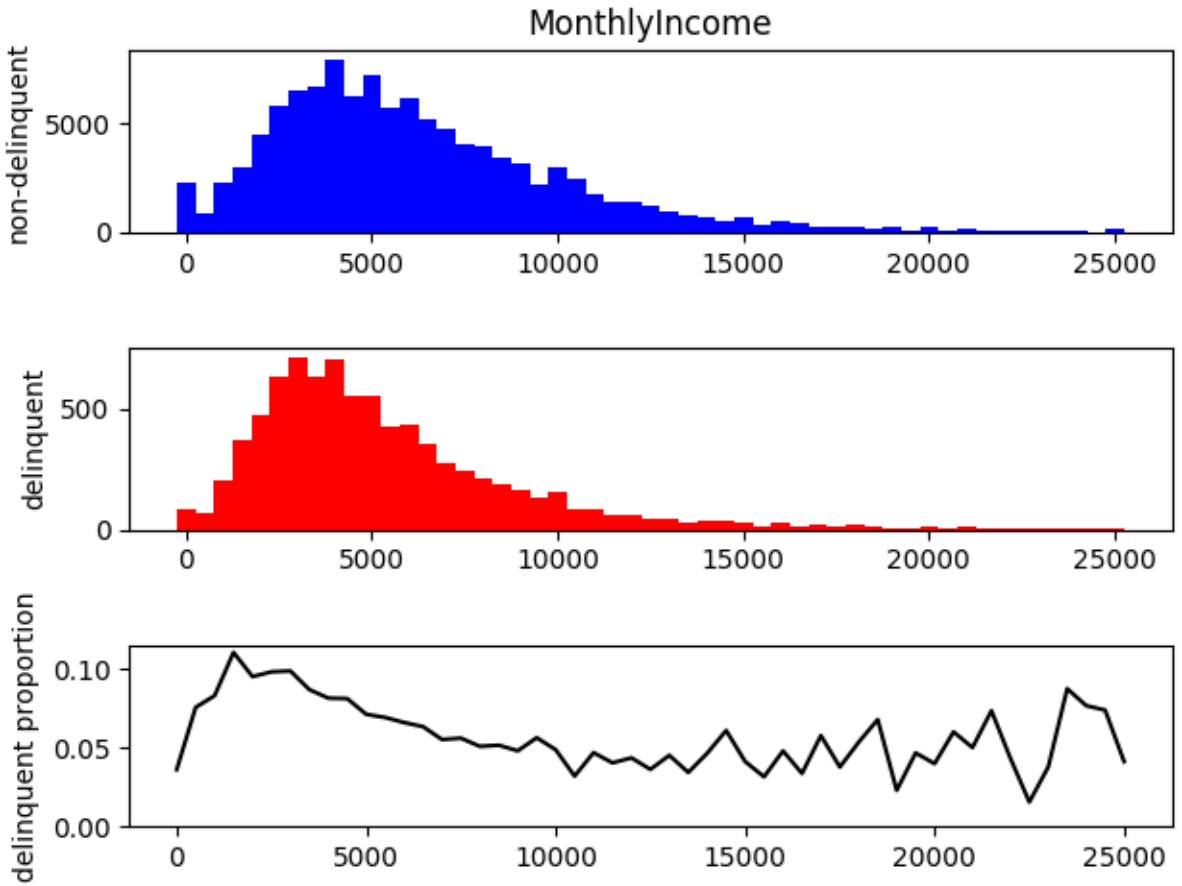
Morosidad x Ratio de deuda



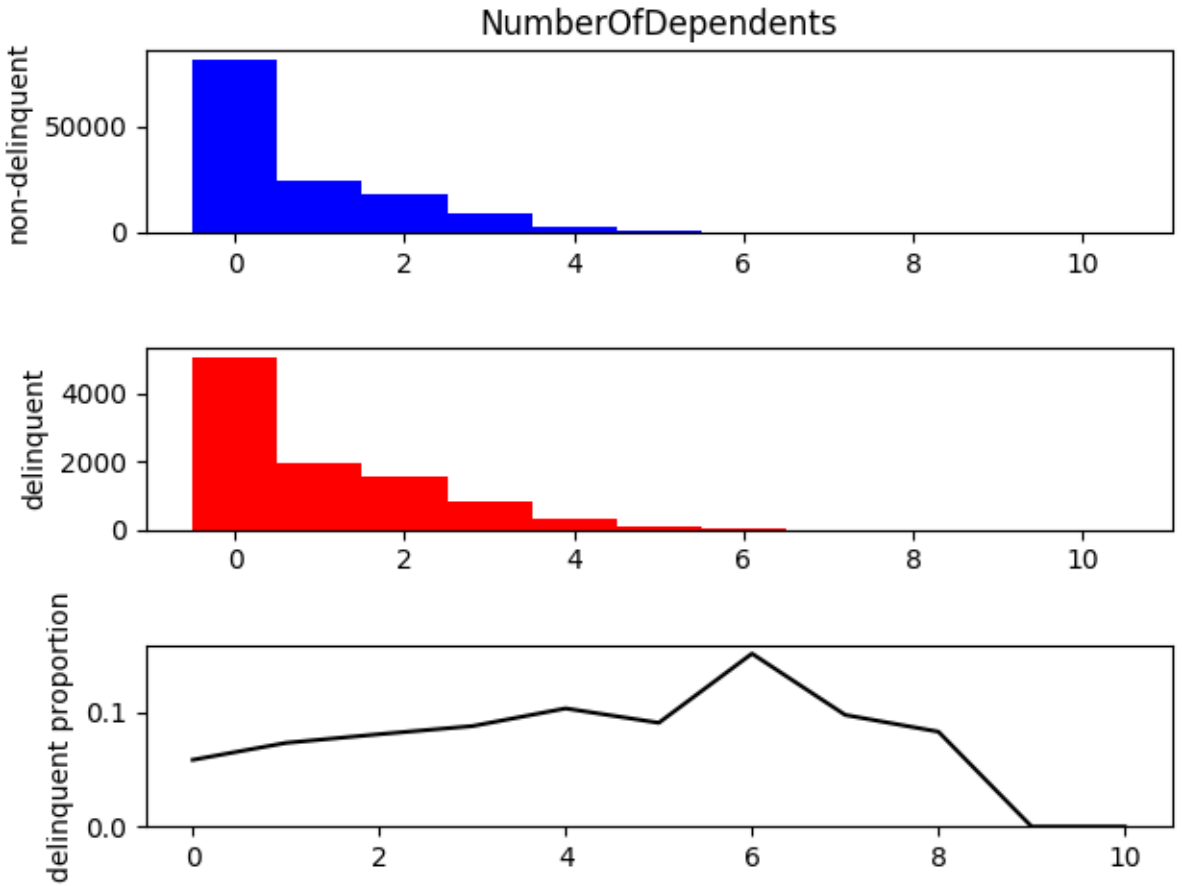
Morosidad x Edad



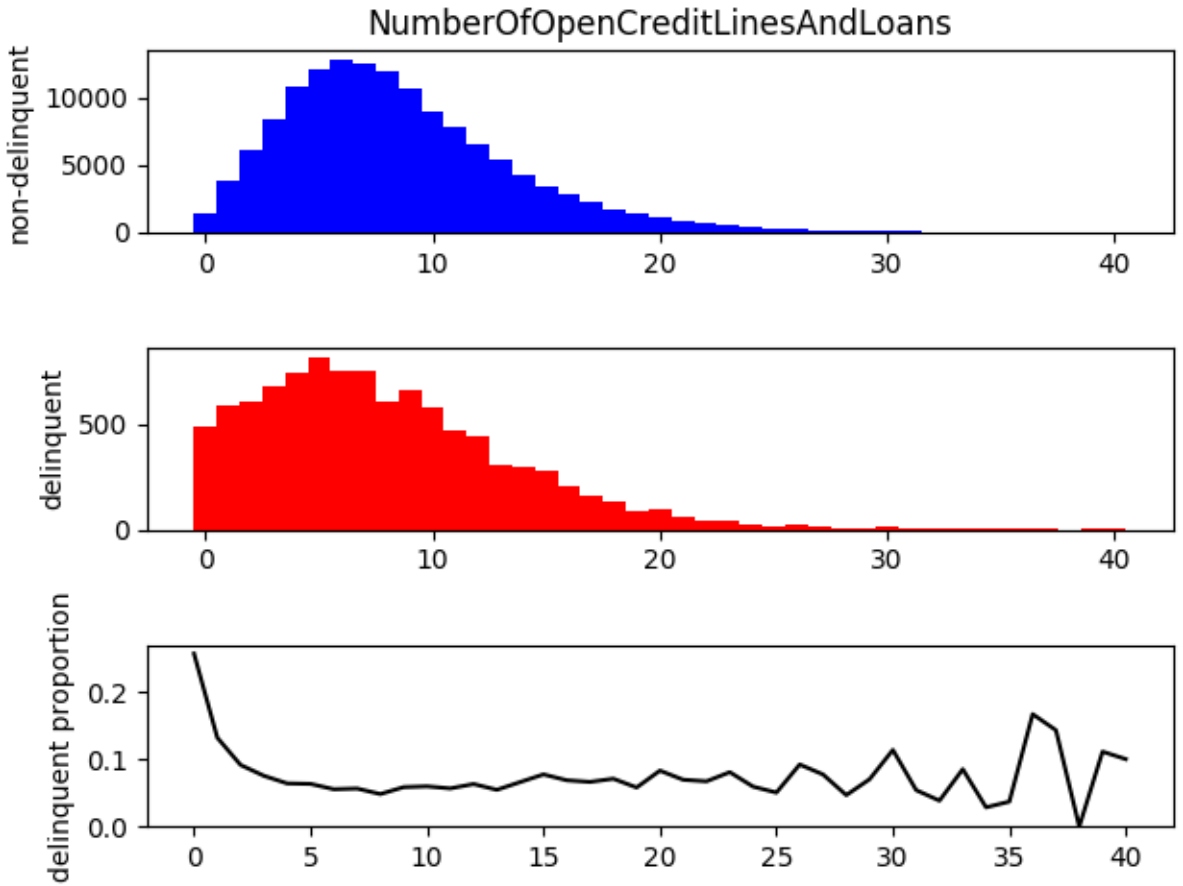
Morosidad x Ingreso mensual



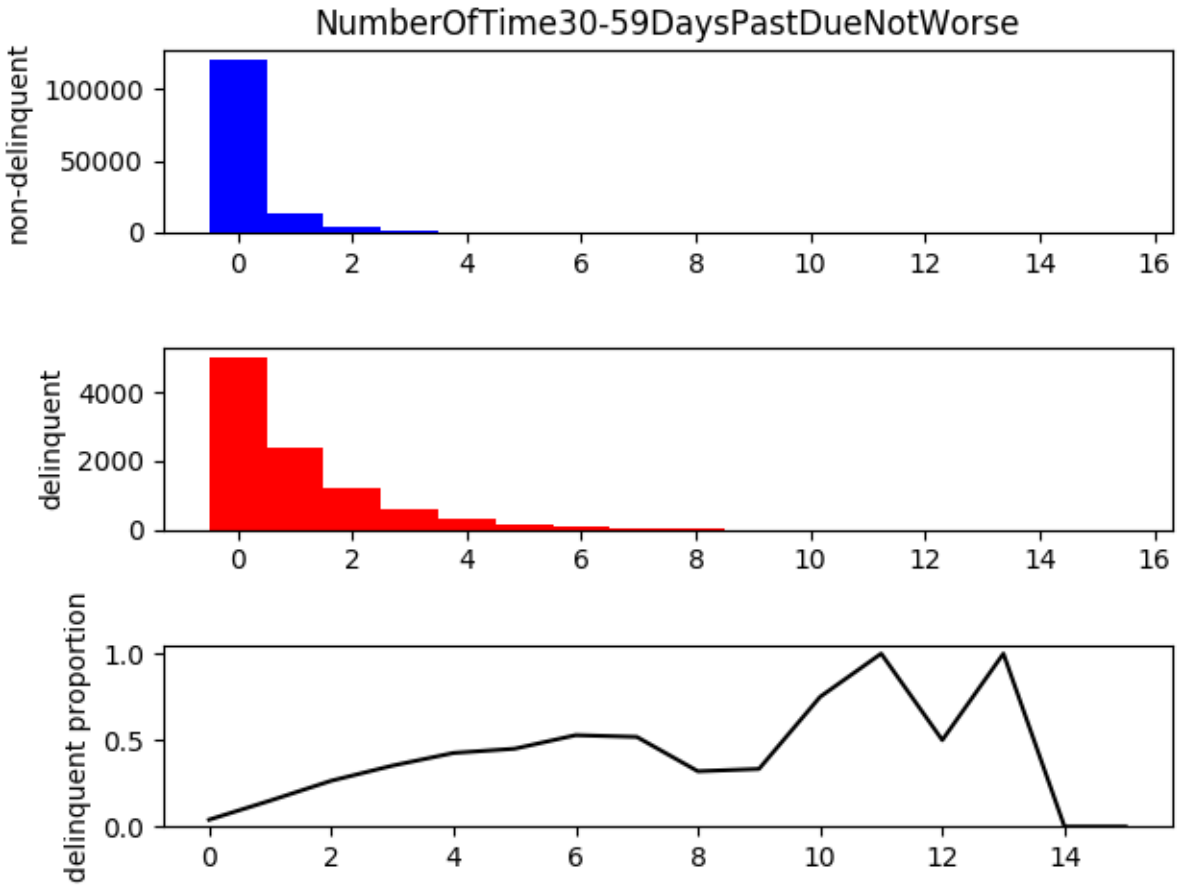
Morosidad x Número de dependientes



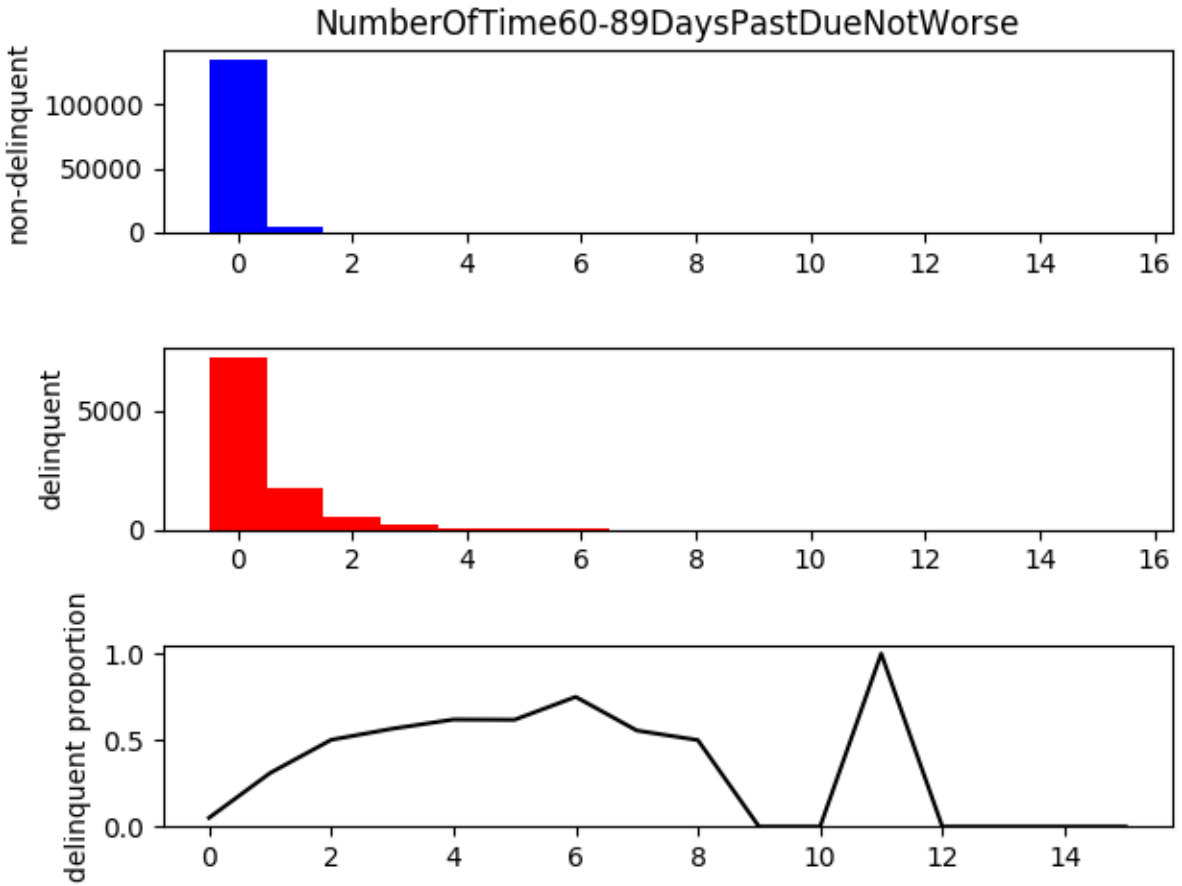
Morosidad x Número de líneas de crédito abiertas y prestamos



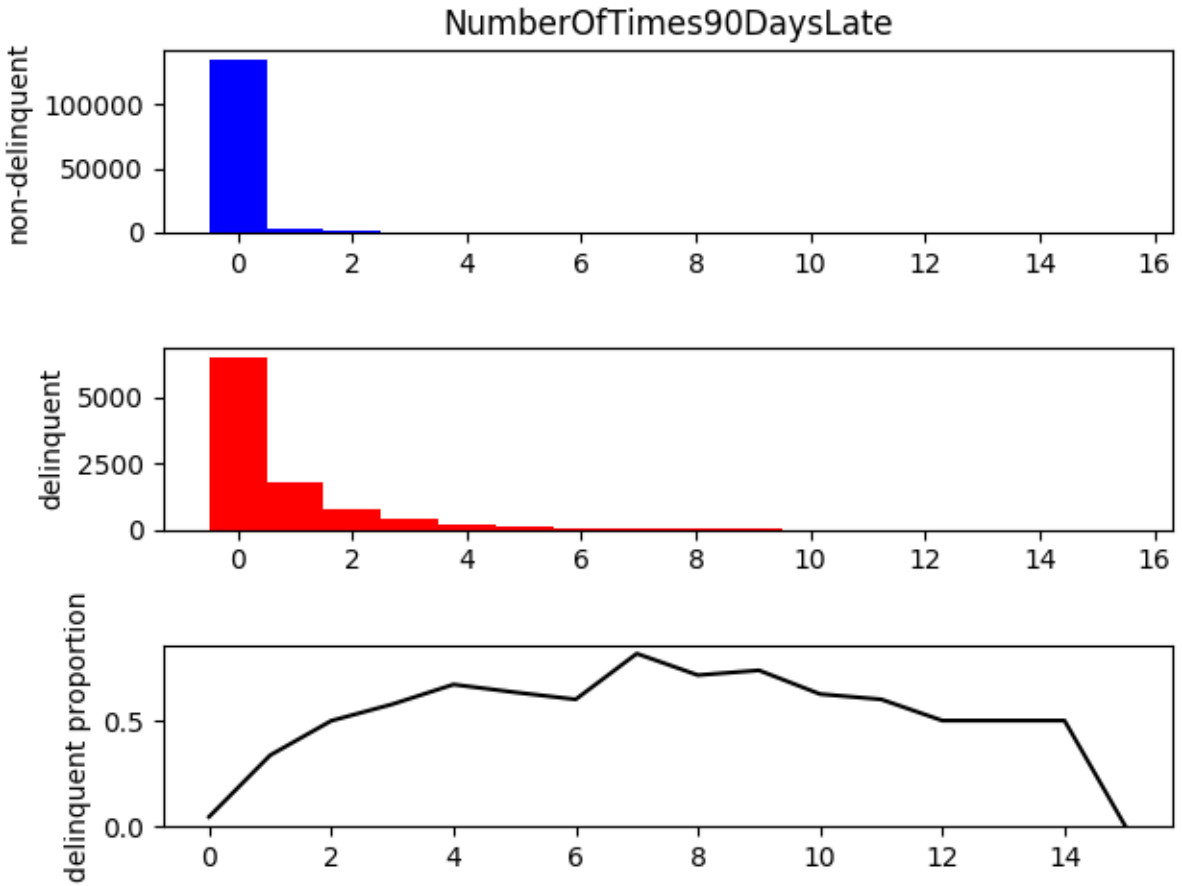
Morosidad x Número de veces de 30-59 días vencidos



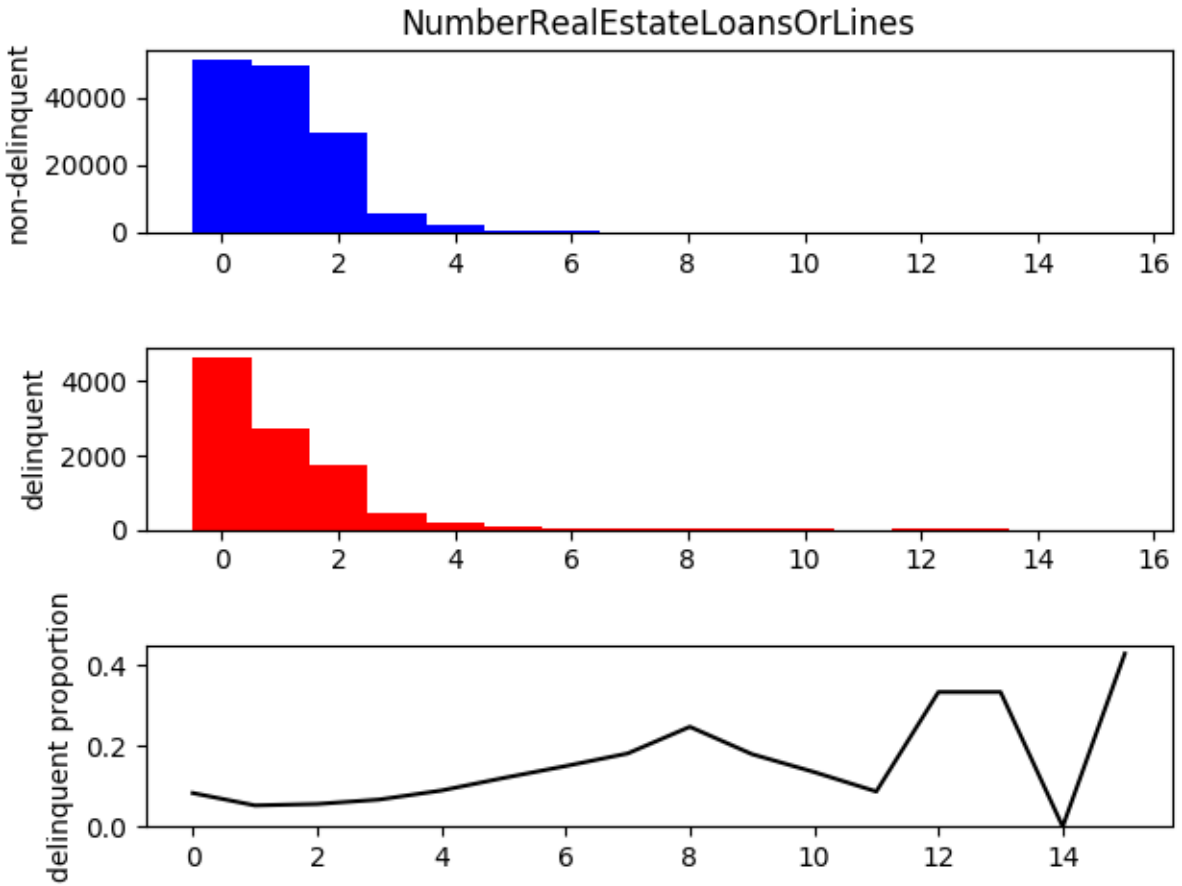
Morosidad x Número de veces de 60-89 días vencidos



Morosidad x Número de veces +90 días vencidos



Morosidad x Número de prestamos o líneas

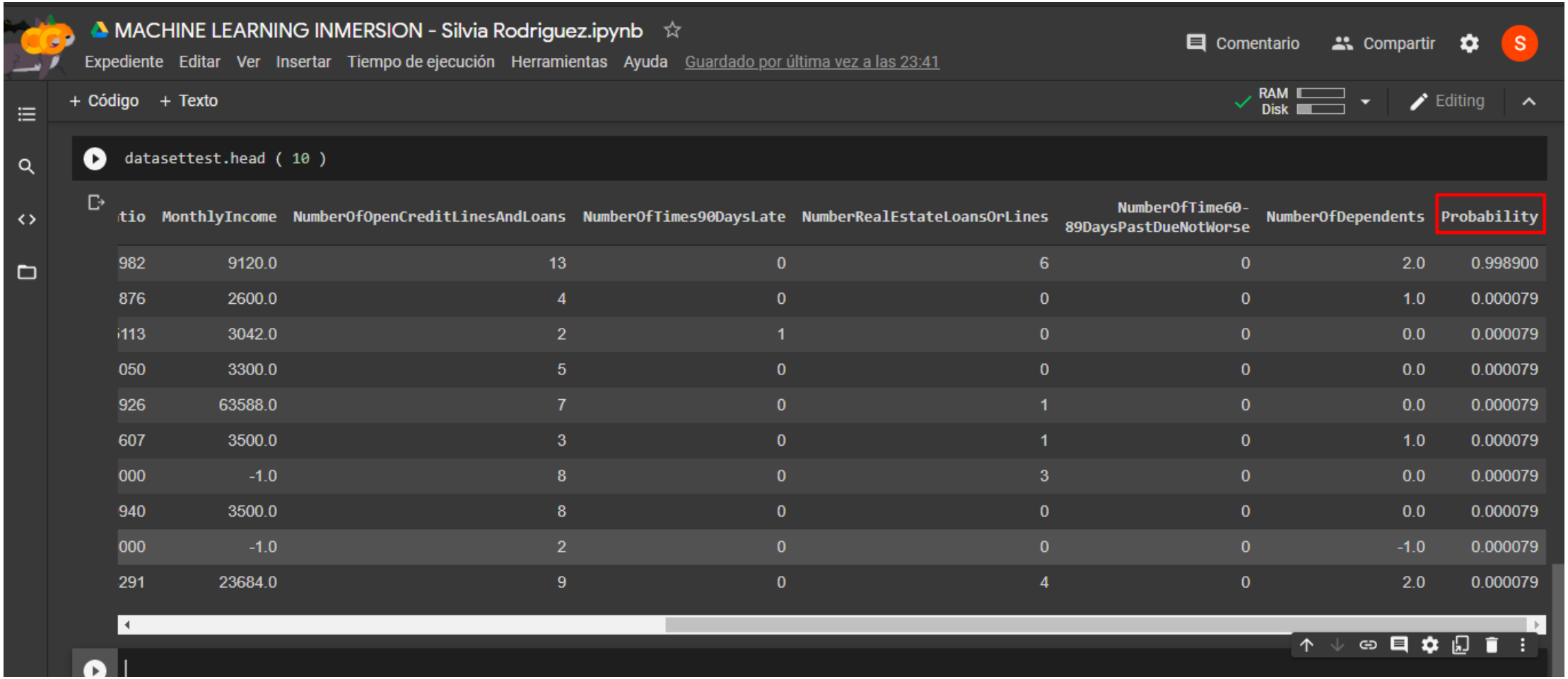


CONCLUSIONES

No se pudo diseñar una característica importante que afecte de manera considerable los resultados después del entrenamiento

A pesar de contar con datos que no conocemos el significado, 96 y 98 igualmente fueron utilizados para el entrenamiento, debido a que se puede tratar de un código o también de un valor incorrecto pero consistente.

Dataset de prueba 10 filas



tio	MonthlyIncome	NumberOfOpenCreditLinesAndLoans	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	NumberOfTime60-89DaysPastDueNotWorse	NumberOfDependents	Probability
982	9120.0	13	0	6	0	2.0	0.998900
876	2600.0	4	0	0	0	1.0	0.000079
113	3042.0	2	1	0	0	0.0	0.000079
050	3300.0	5	0	0	0	0.0	0.000079
926	63588.0	7	0	1	0	0.0	0.000079
607	3500.0	3	0	1	0	1.0	0.000079
000	-1.0	8	0	3	0	0.0	0.000079
940	3500.0	8	0	0	0	0.0	0.000079
000	-1.0	2	0	0	0	-1.0	0.000079
291	23684.0	9	0	4	0	2.0	0.000079