Reporte P2: Credit Card Defaults

Contexto del Problema

Alrededor del año 2006, los emisores de tarjetas de crédito en Taiwán se enfrentaron a una crisis de efectivo y deudas de tarjetas de crédito. Con el fin de aumentar la cuota de mercado, los bancos emisores de tarjetas en Taiwán emitieron en exceso tarjetas de crédito y efectivo a solicitantes no calificados. Al mismo tiempo, la mayoría de los titulares de tarjetas, independientemente de su capacidad de pago, utilizaron en exceso las tarjetas de crédito para consumo y acumularon pesadas deudas de crédito y efectivo.

Usuario Objetivo y Preguntas de Negocio

Dado el contexto presentado, el usuario objetivo seleccionado para este proyecto es el área de ventas de una empresa de servicios financieros que ofrece productos de crédito a sus clientes. El objetivo principal es maximizar las ganancias de la empresa mediante la identificación de patrones en los datos que puedan ayudar a predecir el riesgo de default de los clientes y optimizar las estrategias de venta de productos de crédito.

Para orientar el análisis al área de ventas y maximizar las ganancias, podemos plantear las siguientes preguntas de negocio:

1. ¿Cuáles son las características demográficas más relevantes que afectan el riesgo de default?
2. ¿Cómo podemos utilizar el historial de pagos de los clientes en meses anteriores y su nivel de deuda para predecir qué clientes tienen mayor probabilidad de hacer default?

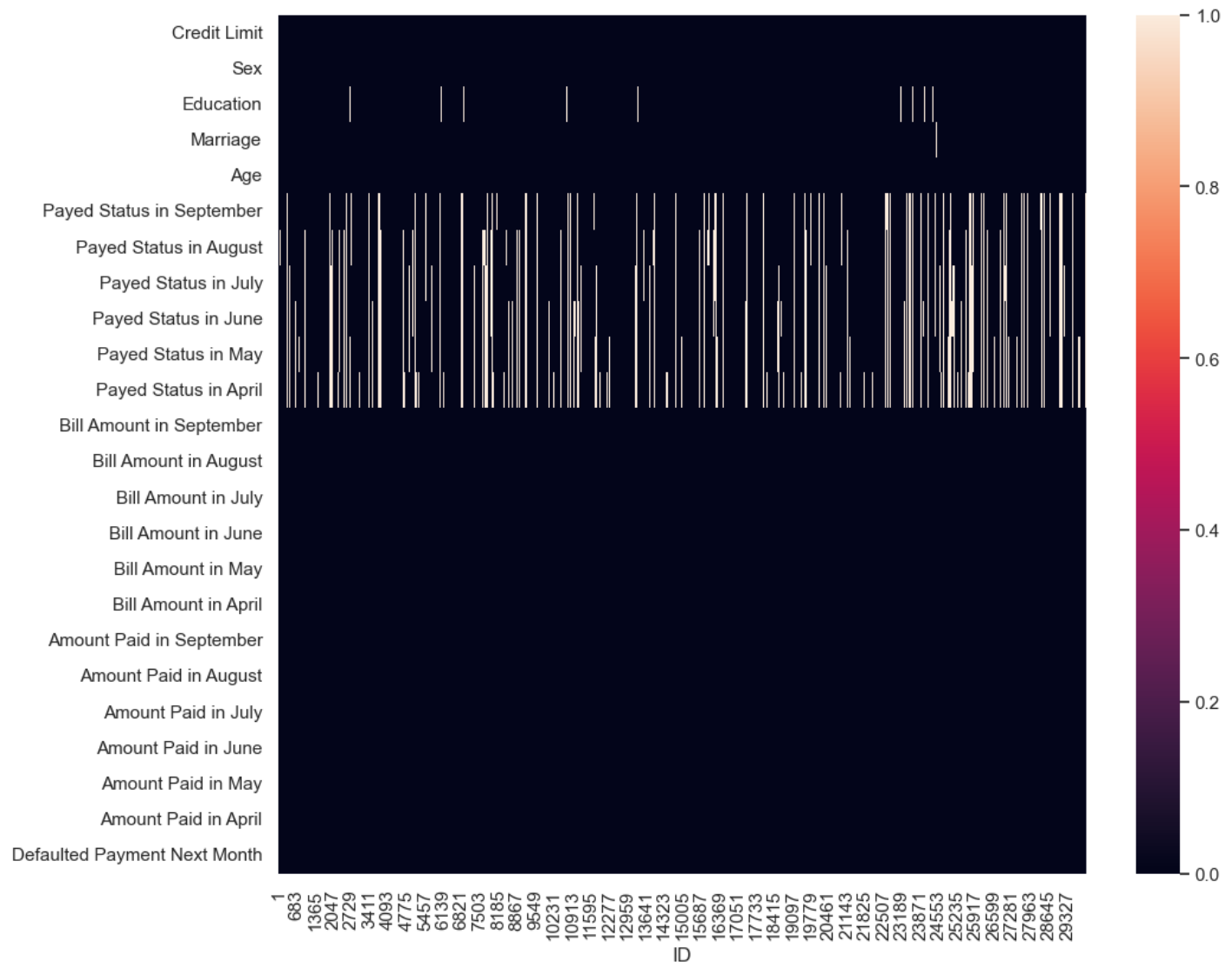
Datos Disponibles

Los datos disponibles para este análisis incluyen información demográfica de los clientes, como género, edad, nivel educativo y estado civil, así como historiales de pagos en meses anteriores y niveles de deuda de los clientes. También se dispone de información sobre el estado de los clientes en términos de default o no default en un período de tiempo específico (Octubre, 2005).

Exploración de los Datos

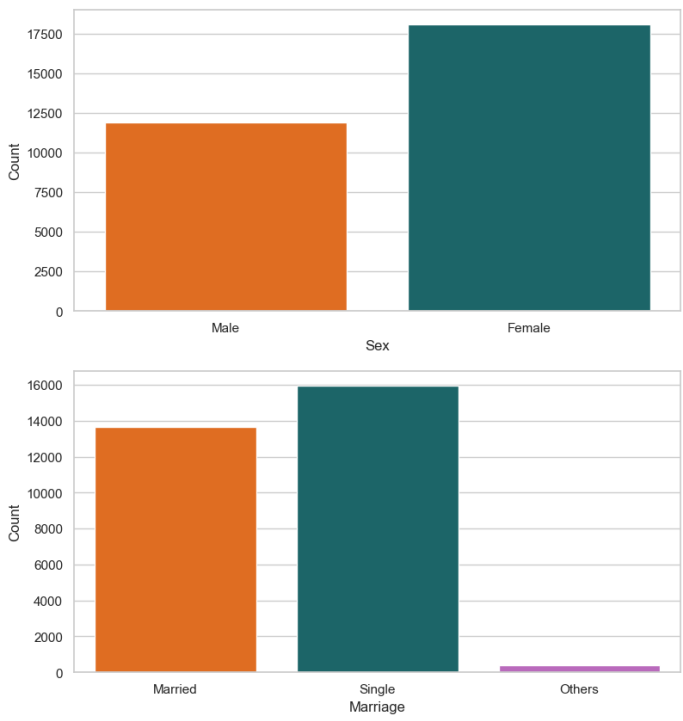
Valores Nulos o Inconsistentes

Los datos no contienen valores nulos, pero sí valores inconsistentes de acuerdo al diccionario de datos. Estos valores inconsistentes se encuentran en las variables de educación, estado marital y estado del pago en los últimos 6 meses. Nótese también que los valores nulos se presentan en los mismos registros (líneas verticales en la gráfica siguiente):

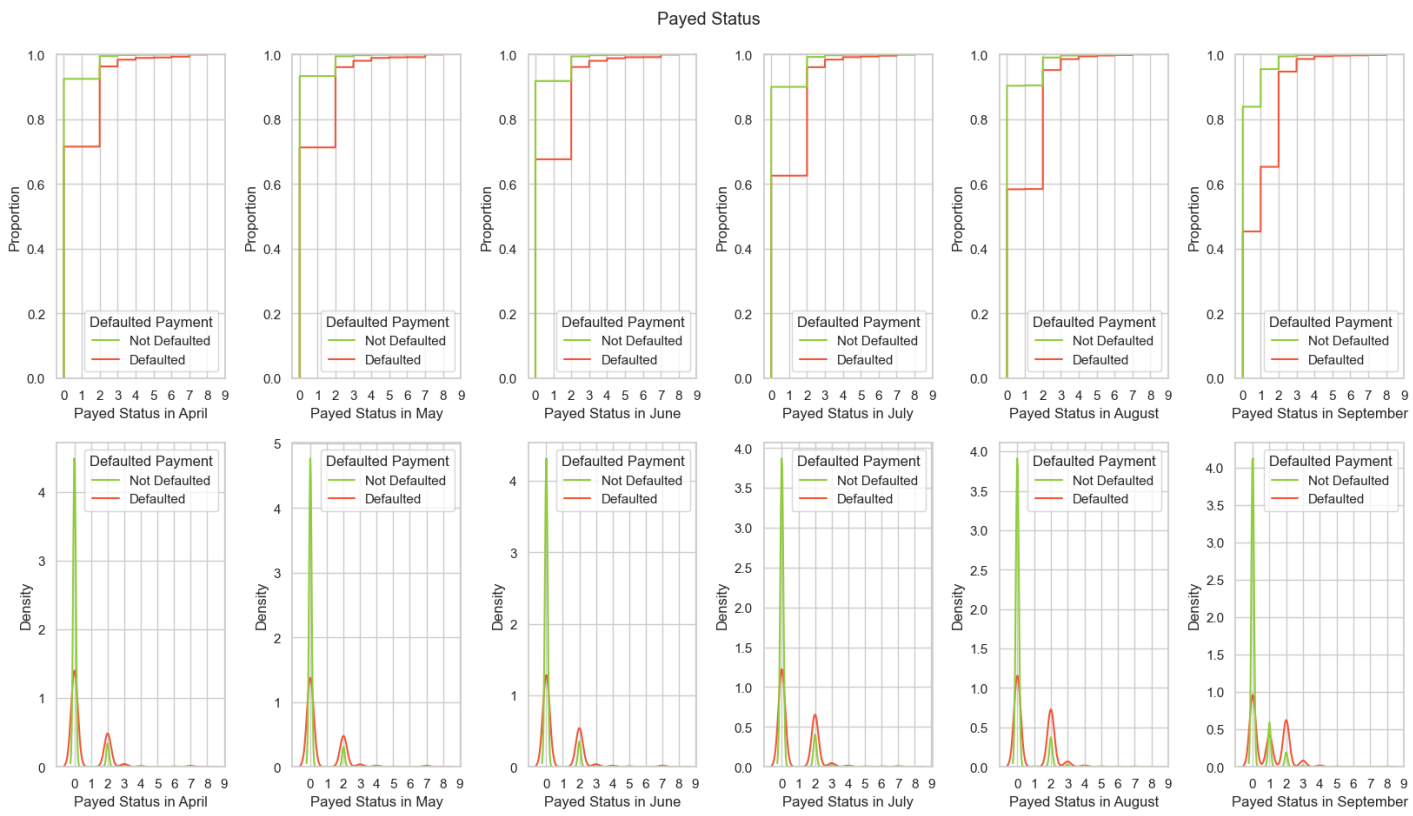


Variables Categóricas

Existe un gran desbalance entre las dos clases de la variables objetivo. Alrededor del 77% de los datos corresponden a clientes que no incumplieron sus pagos, mientras que un 22% corresponde a aquellos que sí. La distribución de las variables de sexo, educación y estado civil no se ve afectada al ser condicionada por la variable objetivo. En cuanto a las variables de estado de pago en los meses anteriores, se observa que la mayoría de los valores corresponden a pagos a tiempo. No existe gran cantidad de valores por encima de los 2 meses.

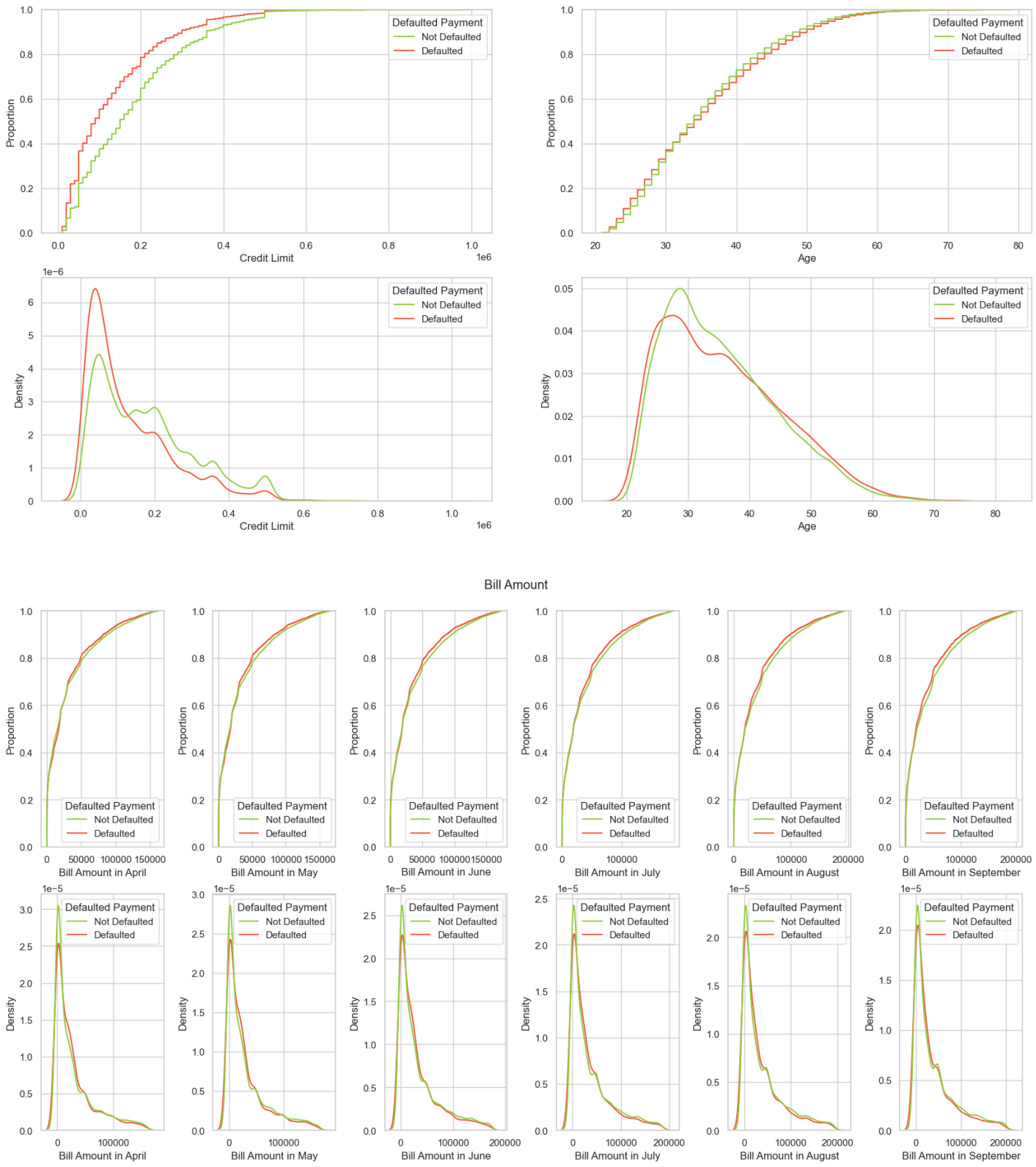


Nótese que la variable del estado de pago muestra una tendencia clara en los primeros dos meses de retraso, que delatan en gran manera a los clientes con mal comportamiento de pago. Este indicio es uno de los más fuertes en la construcción del modelo.



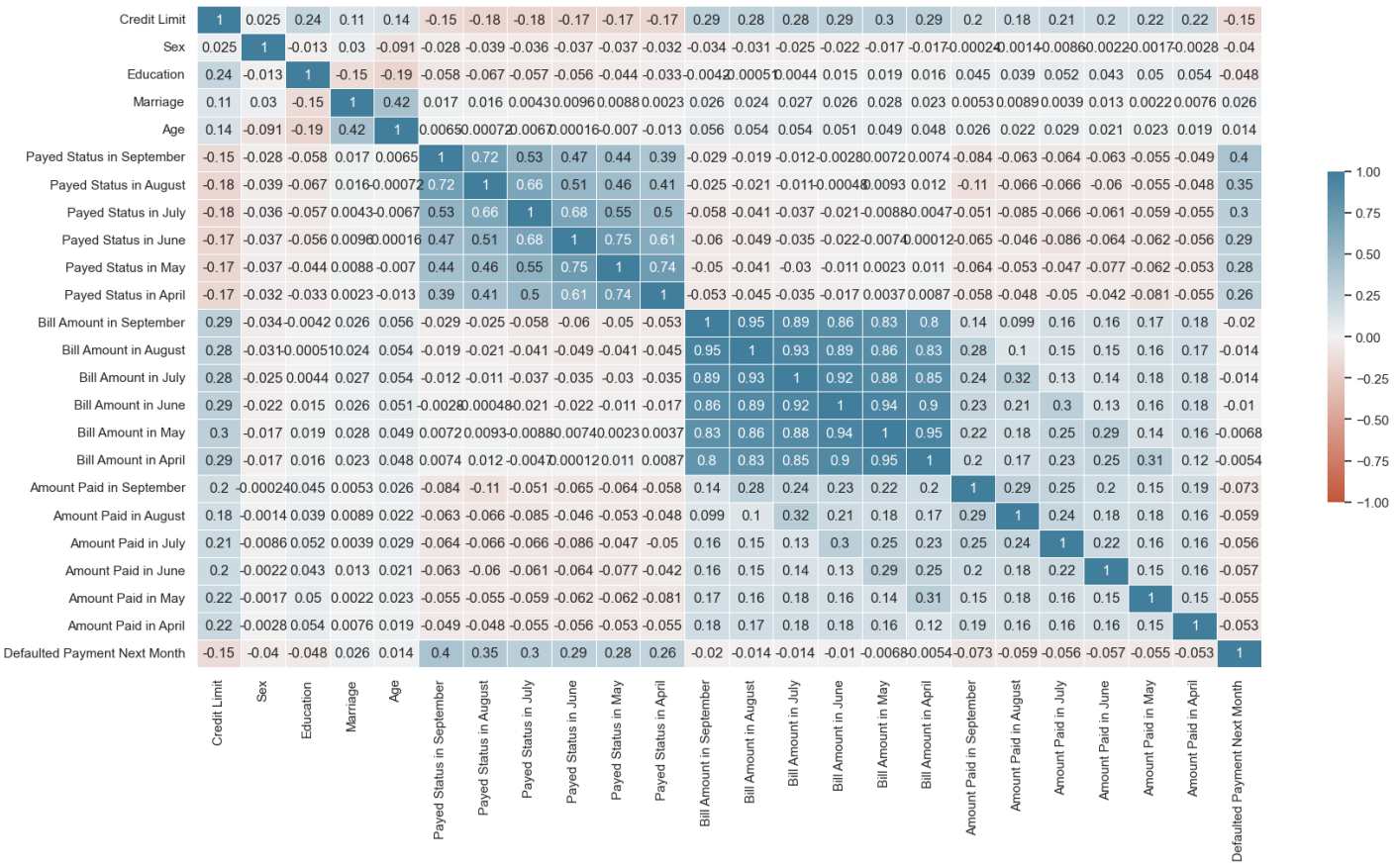
Variables Numéricas

Variables como la edad y el límite de crédito muestran ligeras diferencias al ser condicionadas por la variable objetivo. Por su parte, las variables de Bill Amount y Amount Paid muestran una distribución similar a la exponencial.

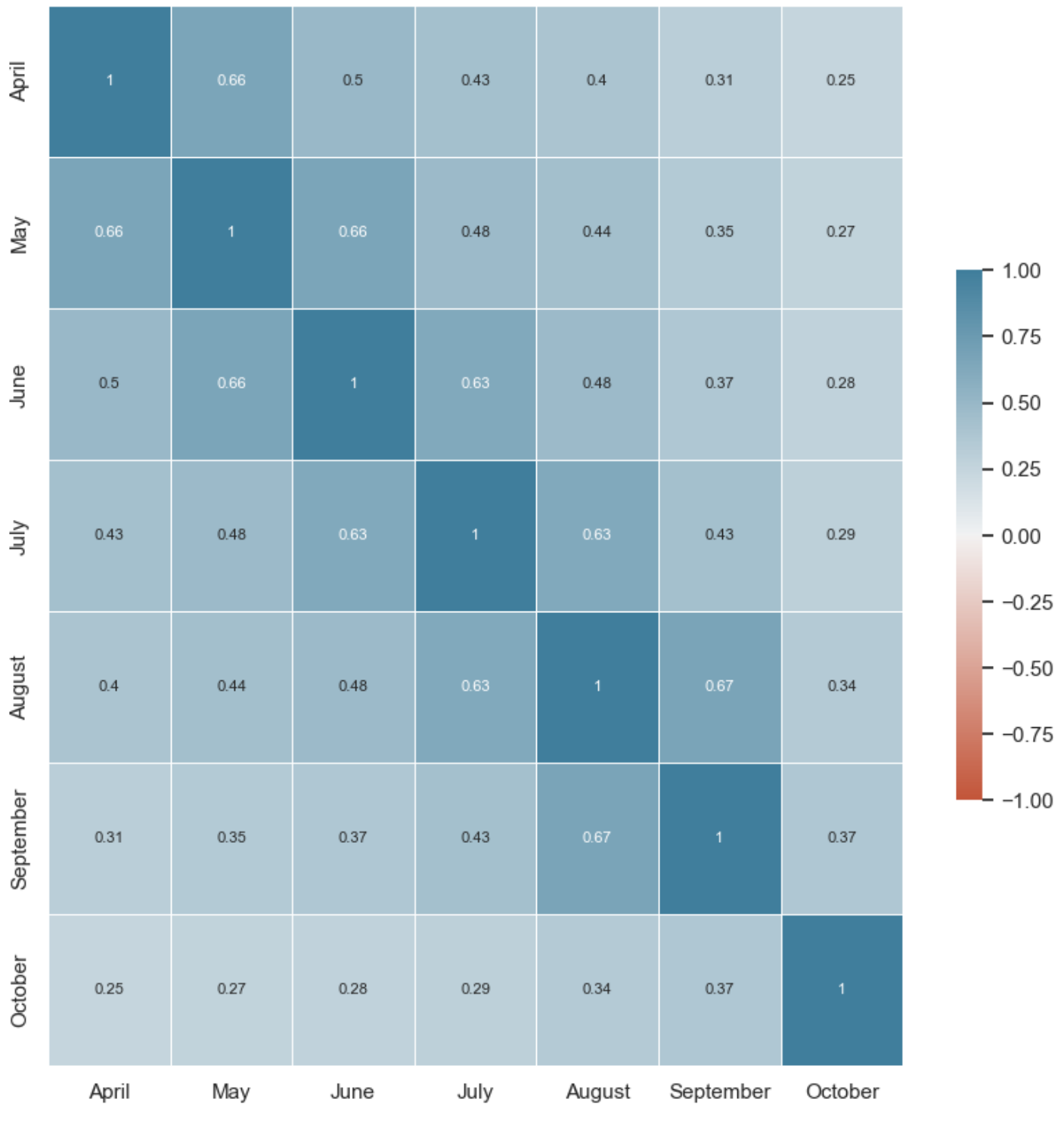


Correlaciones entre Variables

Existen clústeres de variables muy correlacionadas entre sí, en particular la variable Bill Amount y la variable Payed Status a lo largo de los meses. Un análisis más detallado demuestra que se puede predecir con gran exactitud el valor de una de estas variables usando el valor del mes anterior. Esto justifica la decisión de reducir estas variables a el promedio durante los 6 meses.

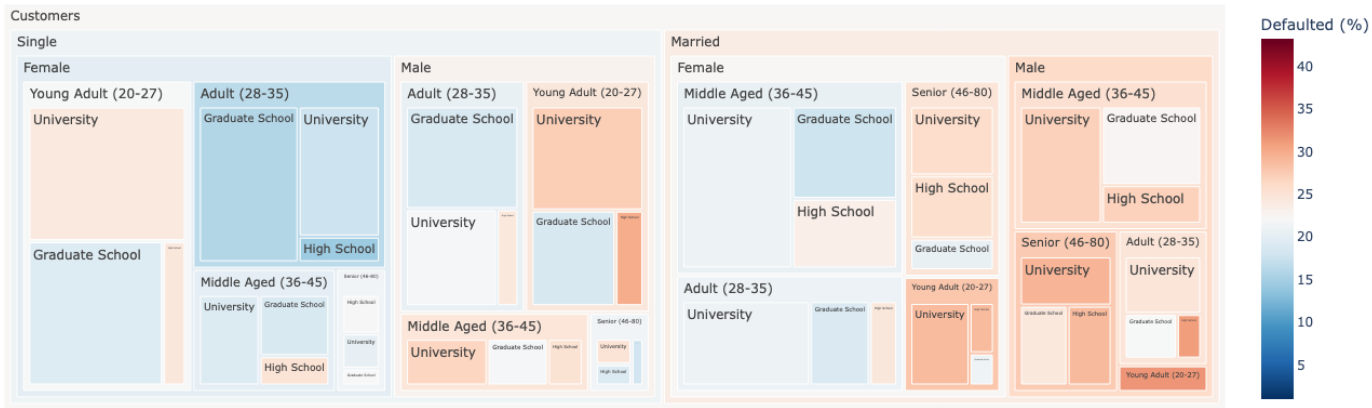


Al revisar en detalle la variable de Payed Status, se observa que entre mayor sea la diferencia entre meses, menor es la capacidad predictiva de la variable. Se observan franjas de intensidad al calcular las correlaciones de estas variables a lo largo del tiempo.

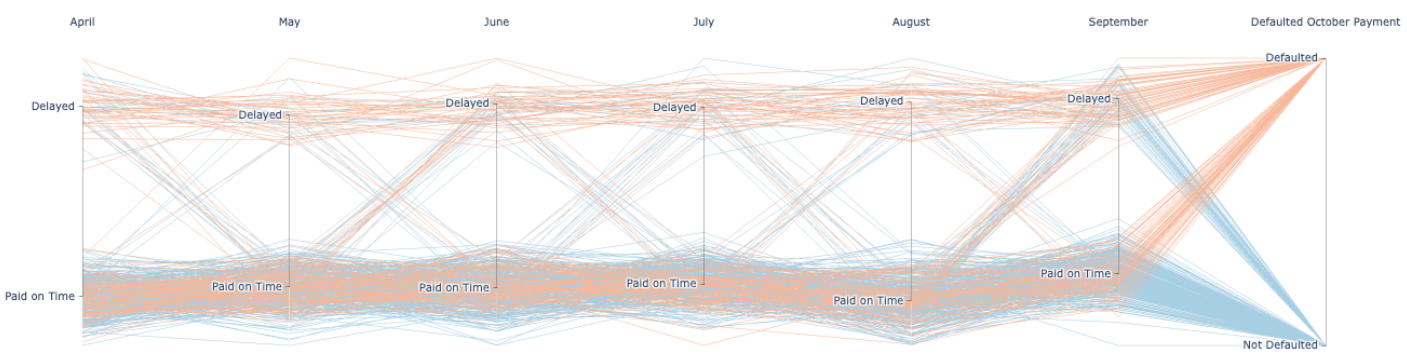


Visualizaciones Interactivas

Con el fin de explorar fácilmente los datos tabulares de la demografía, se utiliza un Treemap que permite expandir sobre un grupo poblacional en particular. La jerarquía se ordena según la relevancia de la variable para predecir el incumplimiento (variable objetivo).



Por su parte, se utiliza un gráfico de ejes paralelos que permite visualizar en detalle la historia de los distintos clientes, permitiendo observar quiénes pasaron de un estado de mora a un estado de pago a tiempo, y viceversa, a lo largo de los meses. Con esto, se puede observar la tendencia de las personas que pagan a tiempo de seguir sobre la misma línea.



Limpieza

Se convierte la columna 'SEX' en un formato binario, donde 1 representa masculino y 0 representa femenino.b. Las columnas 'EDUCATION' y 'MARRIAGE' se mapean a categorías más descriptivas y se convierten en variables categóricas. Los valores faltantes se llenan con 'Others'.c. La columna 'AGE' se agrupa en rangos (20-27, 28-35, 36-45, 46-80) y se etiqueta con etiquetas descriptivas.d. Las columnas de historial de pago (PAY_0, PAY_2, ..., PAY_6) se convierten a un formato binario donde los valores negativos y 0 se consideran pagos puntuales (0), y los valores positivos se consideran pagos atrasados (1).e. Se calcula el valor mediano de las columnas 'BILL_AMT1' a 'BILL_AMT6' y 'PAY_AMT1' a 'PAY_AMT6' y se crean dos nuevas columnas 'Average Bill Amount' y 'Average Pay Amount' respectivamente. Luego, se eliminan las columnas originales.f. Se renombran algunas columnas para hacerlas más descriptivas.g. Se crea variables ficticias (one-hot encoding) para las variables categóricas restantes ('EDUCATION', 'MARRIAGE', 'AGE') y se almacenan como variables binarias.

Modelo de Predicción de Incumplimiento (Default)

Estado del Arte

Los precedentes de modelos para la predicción de incumplimientos crediticios han sido objeto de estudio en diversas investigaciones. Yeh y Lien (2009) se enfocaron en el caso de los pagos incumplidos de clientes de tarjetas de crédito en Taiwán, comparando la precisión predictiva de la probabilidad de incumplimiento entre seis métodos de minería de datos. Desde la perspectiva de la gestión de riesgos, destacaron que la precisión predictiva de la probabilidad estimada de incumplimiento es más valiosa que el resultado binario de clasificación de clientes como creíbles o no creíbles. Para abordar la dificultad de estimar la verdadera probabilidad de incumplimiento, presentaron el novedoso "Método de Suavizado de Clasificación", que utiliza regresión lineal simple para relacionar la probabilidad de incumplimiento real con la probabilidad de incumplimiento predictiva.

Descubrieron que el modelo de predicción generado por redes neuronales artificiales fue el único capaz de estimar con precisión la probabilidad real de incumplimiento.

Por otro lado, Islam, Eberle y Ghafoor (2018) abordaron el desafío de predecir cuentas de incumplimiento crediticio potenciales con anticipación. Reconocieron que las técnicas estadísticas tradicionales tienen dificultades para manejar grandes cantidades de datos y la naturaleza dinámica del fraude y el comportamiento humano. Para superar esta limitación, presentaron y validaron un enfoque heurístico para extraer cuentas de incumplimiento potenciales con anticipación. Este enfoque combina la precomputación de la probabilidad de riesgo con un enfoque de aprendizaje automático recientemente propuesto. Demostraron que estos enfoques aplicados superan a las técnicas existentes de vanguardia en términos de precisión predictiva.

Pre procesamiento de datos

Se estandariza los datos, lo que implica ajustar la distribución de cada característica para que tenga una media de 0 y una desviación estándar de 1 asegurando que todas las características de los datos tengan la misma escala. En general, esto ayuda al entrenamiento de los modelos.

Arquitectura de Redes Neuronales

Entradas de Datos:

La red neuronal tiene tres entradas de datos distintas: una para los estados de pago (6 características), otra para las características de crédito (3 características) y la última para las características personales y demográficas (12 características).

Procesamiento de los Estados de Pago:

- Los estados de pago pasan por una capa de convolución 1D con 32 filtros, seguida de una capa de MaxPooling1D para reducir la dimensionalidad.
- Luego, se aplican otra capa de convolución 1D con 16 filtros y otra capa de MaxPooling1D.
- Finalmente, se aplanan los datos y se aplica una capa de Dropout para evitar el sobreajuste.
- Se elige este enfoque para aprovechar la naturaleza temporal de los datos.

Procesamiento de las Características de Crédito y Personales:

Las características de crédito y personales pasan por capas densas con activación ReLU y capas de Dropout para procesar la información y prevenir el sobreajuste.

Concatenación de las Salidas de las Tres Subredes:

Las salidas de las subredes de estados de pago, características de crédito y características personales se concatenan para combinar la información extraída de estas fuentes diferentes.

Capas Densas Finales:

- La salida concatenada se procesa a través de varias capas densas adicionales con activación ReLU y capas de dropout.
- La última capa es una capa densa con activación sigmoidea, adecuada para realizar una clasificación binaria.

Compilación del Modelo:

El modelo se compila utilizando el optimizador Adam y la función de pérdida de entropía cruzada binaria, adecuada para problemas de clasificación binaria.

Entrenamiento y Validación

Durante el entrenamiento del modelo, se emplean dos técnicas clave para optimizar el proceso y mejorar su desempeño: el Model Checkpoint y el Early Stopping. El Model Checkpoint, mediante el uso del callback ModelCheckpoint, asegura que el mejor modelo se guarde según la pérdida de validación, evitando la pérdida del progreso obtenido durante el entrenamiento. Por otro lado, el Early Stopping, implementado a través del callback del mismo nombre, monitorea la pérdida de validación y detiene el entrenamiento si no mejora después de un número específico de épocas. Este mecanismo restaura los pesos del modelo al mejor estado cuando se interrumpe el entrenamiento. Durante el proceso de entrenamiento, los datos de entrada se dividen en tres partes: estados de pago, características de crédito y características personales, adaptándose a las entradas definidas en el modelo. Este se entrena a lo largo de 500 épocas, actualizando los pesos del modelo después de procesar

30000 muestras de entrenamiento en cada lote. Además, el 20% de los datos se reserva para la validación del modelo.

Métricas de Desempeño

Para evaluar el desempeño del modelo de predicción, se utilizaron varias métricas que proporcionan una visión holística de su rendimiento. Estas métricas ayudan a comprender la capacidad predictiva del modelo y su capacidad para generalizar a datos no vistos. A continuación, se presentan las métricas que se calcularon:

- **Accuracy (Precisión):** Mide la proporción de predicciones correctas sobre el total de predicciones realizadas por el modelo. Es una métrica general que indica cuán efectivo es el modelo en clasificar correctamente las muestras.
- **Recall (Sensibilidad):** Mide la proporción de verdaderos positivos clasificados correctamente sobre el total de verdaderos positivos en el conjunto de datos. Es útil cuando el costo de los falsos negativos es alto. En este contexto, es de particular importancia ya que el costo de que un cliente no pague sus créditos es bastante alto.
- **F1-Score:** Es la media armónica de precision y recall. Proporciona un equilibrio entre precision y recall, lo que lo hace útil cuando hay un desequilibrio entre las clases en el conjunto de datos.
- **Matriz de Confusión:** Es una tabla que muestra las clasificaciones correctas e incorrectas realizadas por un clasificador. Permite una visualización más detallada del desempeño del modelo en términos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.
- **Curva ROC (Receiver Operating Characteristic):** Es una representación gráfica de la sensibilidad (recall) frente a la tasa de falsos positivos. La AUC (Área Bajo la Curva) ROC proporciona una medida del desempeño general del modelo en todos los umbrales de clasificación.

Resultados Financieros

Con el fin de evaluar la viabilidad financiera del modelo implementado, se estimó el cambio en las utilidades generadas gracias a la capacidad de identificar clientes morosos de forma preventiva. El proceso para calcular el revenue implica varios pasos y consideraciones clave:

1. **Margen de Ganancia (p):** Este parámetro representa el margen de ganancia que la empresa obtiene por cada dólar prestado a los clientes. Se asume que este margen se obtiene durante un período de tiempo determinado, en este caso, se considera un período de 6 meses en los cuales factura intereses corrientes. Se define un valor de ***p*** del 3%.
2. **Tasa de Recuperación (r):** La tasa de recuperación representa el porcentaje de pérdidas que la empresa puede recuperar de los clientes que incumplen con sus obligaciones financieras. Se asume que, a pesar del incumplimiento inicial, la empresa puede recuperar un pequeño porcentaje de estas pérdidas. Se define un valor de ***r*** del 30%.

El proceso para calcular el revenue se basa en la diferencia entre los ingresos generados por los clientes que cumplen con sus obligaciones financieras y las pérdidas incurridas por los clientes que no cumplen. A continuación, se detalla la lógica detrás de las ecuaciones utilizadas:

- **Income (Ingresos):** Se calcula multiplicando el promedio del monto de la factura por el margen de ganancia (p) por la cantidad de clientes que pagan a tiempo y por 6 meses. Esto representa los ingresos generados por los clientes que cumplen con sus obligaciones financieras durante el período de interés especificado.
- **Loses (Pérdidas):** Se calcula multiplicando el promedio de la deuda (diferencia entre el monto de la factura y el monto del pago) por la tasa de incumplimiento (1 - tasa de recuperación) y por el número de clientes que no cumplen.
- **Revenue (Ingresos Netos):** Se calcula restando las pérdidas de los ingresos. Esto proporciona una medida del beneficio neto que la empresa obtiene de sus actividades crediticias después de tener en cuenta las pérdidas por incumplimiento.

Después de aplicar el modelo de clasificación de incumplimiento crediticio y calcular los ingresos obtenidos, se observa una mejora significativa en las ganancias de la empresa. Inicialmente, con la base de datos original, se calculó un revenue total de -4 millones de dólares de Taiwán, lo que indica pérdidas. Sin embargo, tras aplicar el modelo de clasificación, se mejoró el resultado.

Income	Loses	Final Revenue
$150,813,993.06	$62,379,399.60	$88,434,593.46

Estos resultados indican que a través del modelo de clasificación, la empresa puede mejorar sus ganancias en un monto significativo. La diferencia entre los ingresos y las pérdidas es positiva, lo que demuestra la efectividad del modelo en la gestión del riesgo crediticio y la maximización de las ganancias para la compañía.

El proceso de búsqueda del mejor revenue implica variar el umbral de clasificación utilizado en el modelo y calcular el revenue correspondiente para cada umbral. Se elige el umbral que maximiza el revenue, lo que indica la confianza necesaria para clasificar a un cliente como riesgoso. En este caso, se encontró que el mejor umbral es del 38% de confianza, lo que maximiza el revenue obtenido. Esto permite a la empresa ajustar su estrategia de clasificación para optimizar sus ganancias y gestionar de manera más efectiva el riesgo crediticio.

Estos resultados respaldan la eficacia del modelo de clasificación en la identificación de clientes con mayor riesgo de incumplimiento crediticio y en la mejora de las ganancias de la empresa. Con este enfoque, la empresa puede tomar decisiones más informadas y estratégicas en la concesión de crédito y la gestión del riesgo, lo que contribuye a su éxito financiero a largo plazo.