FACULTAD DE INGENIERÍA

INGENIERÍA CIVIL INFORMÁTICA

Ciencia de Datos

Análisis y Preprocesamiento

https://github.com/JavierTallarin/proyecto-CreditoBancario-Ciencia-De-Datos

Jorge Ahumada Margarit

Javier Bravo Orellana

Cristóbal Olave Herrera

Luis Rodriguez Zamora

Índice

Introducción

Este conjunto de datos de UCI contiene cantidad crediticia, datos demográficos, historial de pagos y extractos de facturas de clientes de tarjetas de crédito en Taiwán desde abril de 2005 hasta septiembre de 2005. Finalmente esta investigación tiene como objetivo el caso de los pagos por incumplimiento de los clientes en Taiwán.

Objetivos

- Realizar estadística descriptiva básica para conocer los datos que se encuentran en el dataset y cuáles son sus características.
- Reconocer patrones presentes en el dataset que nos servirán para clasificar posteriormente con la ayuda de modelos de predicción y clasificadores probabilísticos.
- Finalmente la finalidad de este trabajo investigativo es estimar la probabilidad de incumplimiento de los clientes de Taiwán, de esta manera se podrá decidir si se le asigna el crédito a algún cliente del banco.

Descripción de los datos

El dataset default of credit card clients.csv contiene 24 variables, entre ellas datos demográficos e historial crediticio.
Las variables incluidas en el dataset son:

- ID: ID de cada cliente, variable numérica
- LIMIT_BAL: cantidad de crédito otorgado en dólares de taiwán (variable númerica)
- SEX: Sexo, variable categórica (1 masculino, 2 femenino)
- EDUCATION: Nivel máximo educacional, variable categórica (1 postgrado, 2 universidad, 3 bachillerato, 4 otros)
- MARRIAGE: Estado civil, variable categórica (1 casado, 2 soltero, 3 otros)
- AGE: edad, variable numérica
- PAY_X: Historial de pagos pasados, variable numérica
- BILL_AMTX: Monto del estado de la cuenta en dólares de taiwán, variable numérica
- PAY_AMTX: Monto del pago anterior en dólares de taiwán, variable numérica

Imagen 1.
En esta imagen se muestran los 10 primeros datos del dataset.
En este caso utilizamos la matriz traspuesta para visualizar
de mejor forma los datos.

	0	1	2	3	4	5	6	7	8	9
ID	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
LIMIT_BAL	20000.0	120000.0	90000.0	50000.0	50000.0	50000.0	500000.0	100000.0	140000.0	20000.0
SEX	2.0	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	1.0
EDUCATION	2.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	3.0	3.0
MARRIAGE	1.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	2.0
AGE	24.0	26.0	34.0	37.0	57.0	37.0	29.0	23.0	28.0	35.0
PAY_1	2.0	-1.0	0.0	0.0	-1.0	0.0	0.0	0.0	0.0	-2.0
PAY_2	2.0	2.0	0.0	0.0	0.0	0.0	0.0	-1.0	0.0	-2.0
PAY_3	-1.0	0.0	0.0	0.0	-1.0	0.0	0.0	-1.0	2.0	-2.0
PAY_4	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-2.0
PAY_5	-2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0
PAY_6	-2.0	2.0	0.0	0.0	0.0	0.0	0.0	-1.0	0.0	-1.0
BILL_AMT1	3913.0	2682.0	29239.0	46990.0	8617.0	64400.0	367965.0	11876.0	11285.0	0.0
BILL_AMT2	3102.0	1725.0	14027.0	48233.0	5670.0	57069.0	412023.0	380.0	14096.0	0.0
BILL_AMT3	689.0	2682.0	13559.0	49291.0	35835.0	57608.0	445007.0	601.0	12108.0	0.0
BILL_AMT4	0.0	3272.0	14331.0	28314.0	20940.0	19394.0	542653.0	221.0	12211.0	0.0
BILL_AMT5	0.0	3455.0	14948.0	28959.0	19146.0	19619.0	483003.0	-159.0	11793.0	13007.0
BILL_AMT6	0.0	3261.0	15549.0	29547.0	19131.0	20024.0	473944.0	567.0	3719.0	13912.0
PAY_AMT1	0.0	0.0	1518.0	2000.0	2000.0	2500.0	55000.0	380.0	3329.0	0.0
PAY_AMT2	689.0	1000.0	1500.0	2019.0	36681.0	1815.0	40000.0	601.0	0.0	0.0
PAY_AMT3	0.0	1000.0	1000.0	1200.0	10000.0	657.0	38000.0	0.0	432.0	0.0
PAY_AMT4	0.0	1000.0	1000.0	1100.0	9000.0	1000.0	20239.0	581.0	1000.0	13007.0
PAY_AMT5	0.0	0.0	1000.0	1069.0	689.0	1000.0	13750.0	1687.0	1000.0	1122.0
PAY_AMT6	0.0	2000.0	5000.0	1000.0	679.0	800.0	13770.0	1542.0	1000.0	0.0
default	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Análisis 1D y 2D de datos

1D

Imagen 2.
En esta segunda imagen generamos una tabla con las estadísticas unidimensionales, entre las cuales se encuentra:

- La media
- La desviación estándar
- El valor mínimo de cada columna
- El cuartil q1(25%), q2(50%), q3(75%)
- El valor máximo de cada columna

Con estos datos se puede armar una idea de que tan dispersos están los datos.

	count	mean	std	min	25%	50%	75%	max
ID	30000.0	15000.500000	8660.398374	1.0	7500.75	15000.5	22500.25	30000.0
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
SEX	30000.0	1.603733	0.489129	1.0	1.00	2.0	2.00	2.0
EDUCATION	30000.0	1.853133	0.790349	0.0	1.00	2.0	2.00	6.0
MARRIAGE	30000.0	1.551867	0.521970	0.0	1.00	2.0	2.00	3.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_1	30000.0	-0.016700	1.123802	-2.0	-1.00	0.0	0.00	8.0
PAY_2	30000.0	-0.133767	1.197186	-2.0	-1.00	0.0	0.00	8.0
PAY_3	30000.0	-0.166200	1.196868	-2.0	-1.00	0.0	0.00	8.0
PAY_4	30000.0	-0.220667	1.169139	-2.0	-1.00	0.0	0.00	8.0
PAY_5	30000.0	-0.266200	1.133187	-2.0	-1.00	0.0	0.00	8.0
PAY_6	30000.0	-0.291100	1.149988	-2.0	-1.00	0.0	0.00	8.0
BILL_AMT1	30000.0	51223.330900	73635.860576	-165580.0	3558.75	22381.5	67091.00	964511.0
BILL_AMT2	30000.0	49179.075167	71173.768783	-69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT3	30000.0	47013.154800	69349.387427	-157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT4	30000.0	43262.948967	64332.856134	-170000.0	2326.75	19052.0	54506.00	891586.0
BILL_AMT5	30000.0	40311.400967	60797.155770	-81334.0	1763.00	18104.5	50190.50	927171.0
BILL_AMT6	30000.0	38871.760400	59554.107537	-339603.0	1256.00	17071.0	49198.25	961664.0
PAY_AMT1	30000.0	5663.580500	16563.280354	0.0	1000.00	2100.0	5006.00	873552.0
PAY_AMT2	30000.0	5921.163500	23040.870402	0.0	833.00	2009.0	5000.00	1684259.0
PAY_AMT3	30000.0	5225.681500	17606.961470	0.0	390.00	1800.0	4505.00	896040.0
PAY_AMT4	30000.0	4826.076867	15666.159744	0.0	296.00	1500.0	4013.25	621000.0
PAY_AMT5	30000.0	4799.387633	15278.305679	0.0	252.50	1500.0	4031.50	426529.0
PAY_AMT6	30000.0	5215.502567	17777.465775	0.0	117.75	1500.0	4000.00	528666.0
default	30000.0	0.221200	0.415062	0.0	0.00	0.0	0.00	1.0

Imagen 3.

En esta imagen se hizo un gráfico de barras para visualizar el nivel de educación máximo que alcanzaron las personas del banco en Taiwán. De esta manera podremos identificar el nivel de educación que más predomina en el dataset.

Para este caso aproximadamente 14.000 personas tienen como nivel educacional máximo la universidad, por lo que la mayoría de las personas solamente hizo el pregrado.

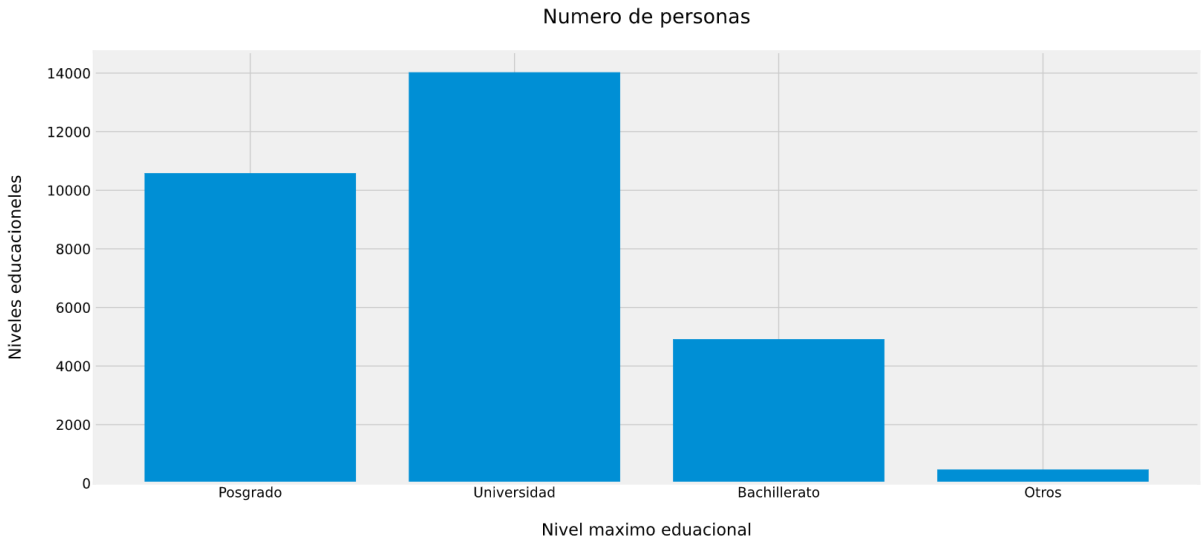


Imagen 4.

En esta imagen hicimos un gráfico de barras para visualizar el estado civil de las personas del banco en Taiwán.

Como resultado se obtuvo aproximadamente 16.000 personas solteras, por lo tanto podemos concluir que hay una cantidad considerable de personas solteras dentro del dataset.

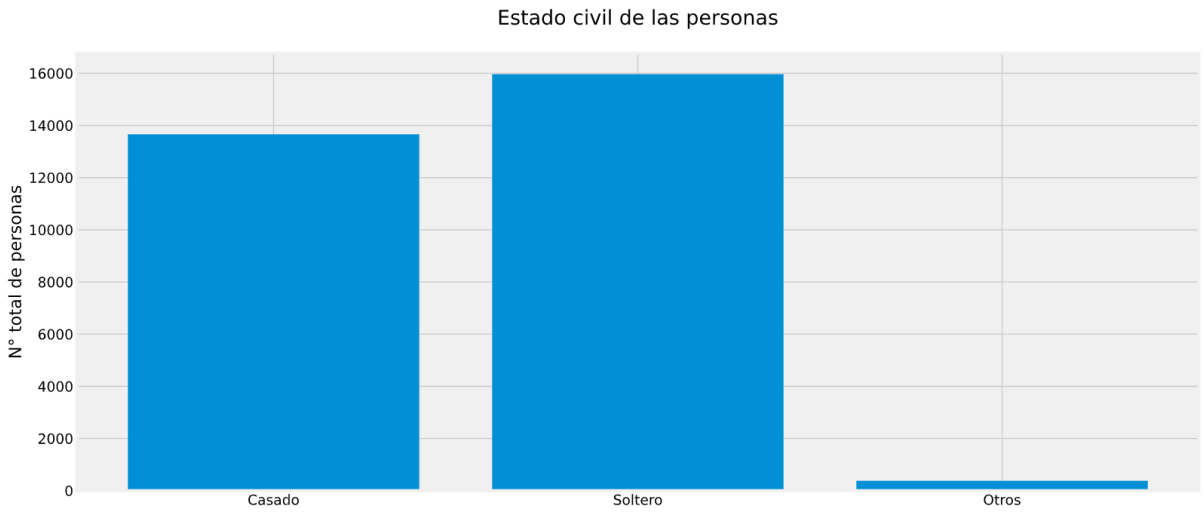


Imagen 5.

En esta imagen hicimos un gráfico de barras para visualizar cuál es el sexo que mas predomina el el dataset.

Claramente se visualiza que la mayoría de las personas en el dataset pertenecen al sexo femenino con mas de 17.500 personas.

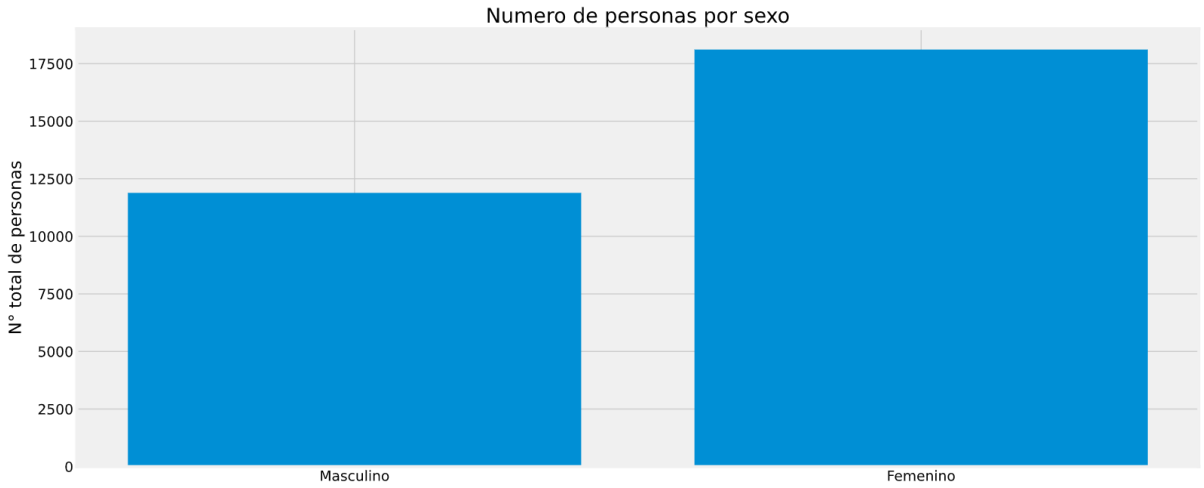
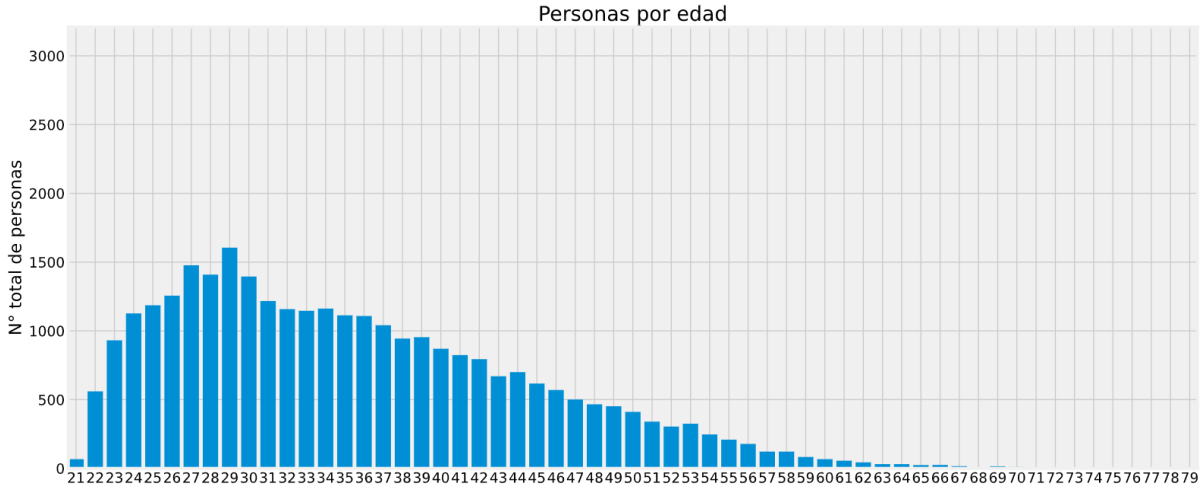


Imagen 6.

En esta imagen hicimos un gráfico para visualizar la distribución de las edades en el dataset.

Si hiciéramos una agrupación de edades en un rango de 5 años. tendríamos que la mayoría de los clientes del banco de Taiwán están entre los 25 y 30 años de edad, incluso la mediana es de 34 años.



2D

Imagen 7.

En esta imagen hicimos un gráfico Heatmap para visualizar la correlación entre las variables del dataset.

El objetivo es encontrar las variables que tengan relación lineal entre ellas o aquellas que tengan una relación lineal inversa.

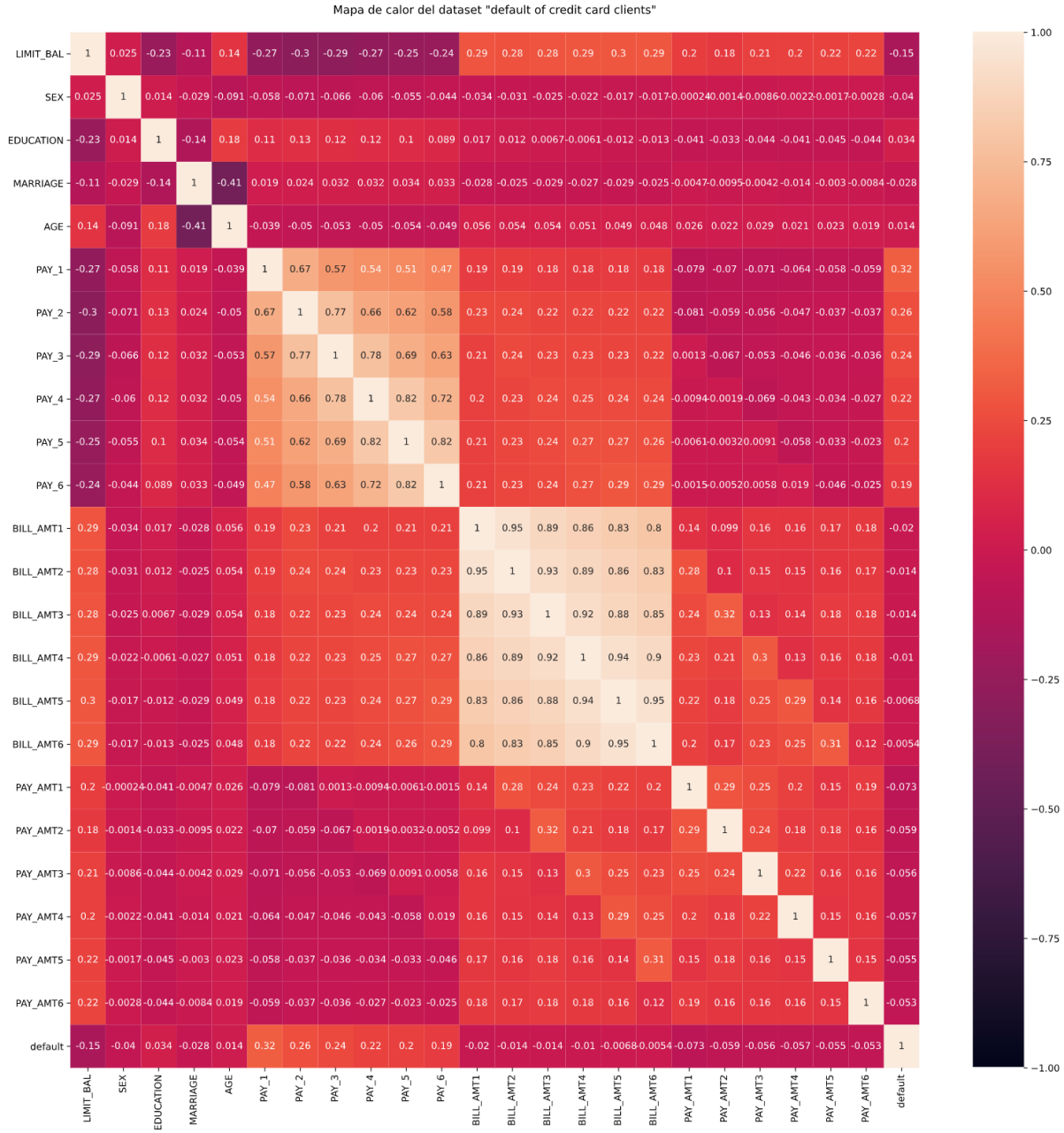
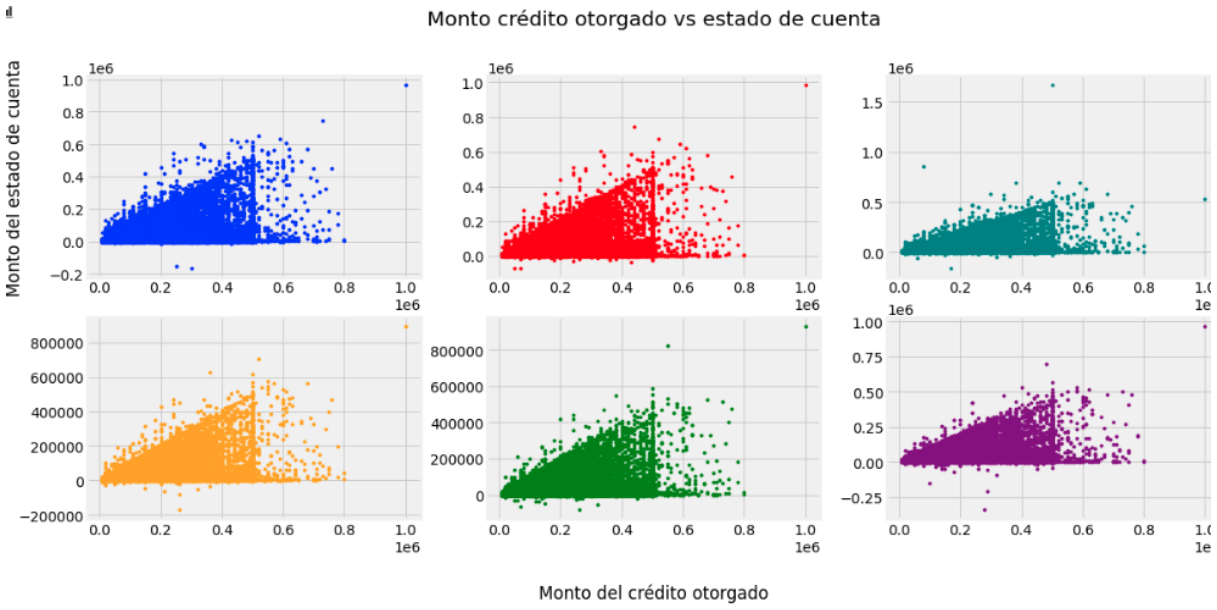


Imagen 8.

En esta imagen hicimos un diagrama de dispersión entre el monto del crédito otorgado y el estado de la cuenta.

En el Heatmap vimos que estas variables tenían un coeficiente de correlación de 0.3 aproximadamente, por lo que se puede decir que existe una relación leve entre estas.


Monto crédito otorgado vs estado de cuenta
Monto del estado de cuenta
Monto del crédito otorgado
1e6

Preprocesamiento de datos

Se parte bajo la condición que no existen valores nulos dentro del data set.
Tampoco hay valores faltantes, cada columna tiene sus respectivos 30000 datos.

#	Column	Non-Null Count	Dtype
0	ID	30000 non-null	int64
1	LIMIT_BAL	30000 non-null	int64
2	SEX	30000 non-null	int64
3	EDUCATION	30000 non-null	int64
4	MARRIAGE	30000 non-null	int64
5	AGE	30000 non-null	int64
6	PAY_1	30000 non-null	int64
7	PAY_2	30000 non-null	int64
8	PAY_3	30000 non-null	int64
9	PAY_4	30000 non-null	int64
10	PAY_5	30000 non-null	int64
11	PAY_6	30000 non-null	int64
12	BILL_AMT1	30000 non-null	int64
13	BILL_AMT2	30000 non-null	int64
14	BILL_AMT3	30000 non-null	int64
15	BILL_AMT4	30000 non-null	int64
16	BILL_AMT5	30000 non-null	int64
17	BILL_AMT6	30000 non-null	int64
18	PAY_AMT1	30000 non-null	int64
19	PAY_AMT2	30000 non-null	int64
20	PAY_AMT3	30000 non-null	int64
21	PAY_AMT4	30000 non-null	int64
22	PAY_AMT5	30000 non-null	int64
23	PAY_AMT6	30000 non-null	int64
24	default	30000 non-null	int64

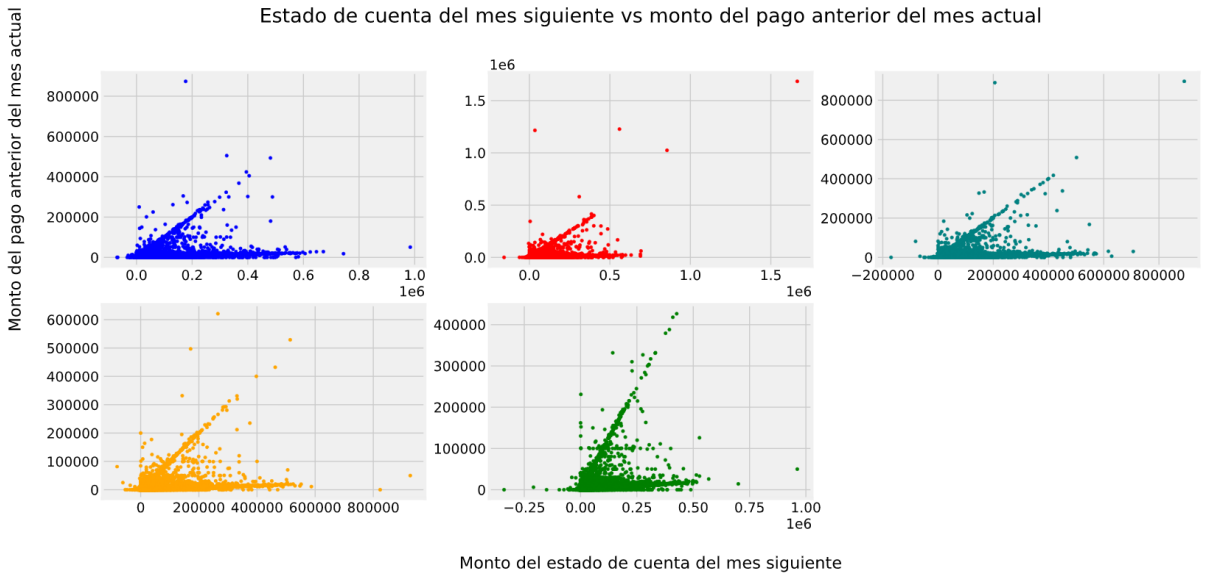
ID: se eliminó esta columna.
Educación: se tomó los valores que no eran parte de las categorías(0, 5 y 6) según la documentación y se cambió por 4(otros).
Estado civil: se cambió los valores que no pertenecen a las categorías definidas en la documentación (0 en este caso) y se cambió por 3.
Sexo: no fue necesario hacer un preprocesamiento adicional

Las decisiones que se tomaron fue considerando que las variables tenían dentro de sus categorías una variable “otros” por lo tanto los datos no pertenecientes a las categorías documentadas podrían ser asignadas a dicha categoría, además los datos no categorizados dentro de las variables era muy bajo con respecto al conjunto total de datos por su respectiva columna, es más la variable educación tenía dentro de sus datos solo un 1.15% de datos no categorizados según documentación.
Otras posibles soluciones con respecto al preprocesamiento de los datos pudo haber sido la eliminación de las tuplas completas en donde alguna de sus columnas no estaban bien

categorizadas o reemplazar dichos valores por la media de la columna debido a que esos valores no categorizados no eran los suficientes para afectar considerablemente la media o la mediana.

Diseño de experimentos

Se considero el experimento de relacionar 2 variables las cuales aparentemente tenían una correlación bastantante fuerte, las cuales no podrían ayudar a nuestros algoritmos a predecir con mayor exactitud
en los cuales se encontró un patrón entre el estado de cuenta del mes siguiente y el monto del pago anterior.



Revisión métodos relacionados

Primero

Predictive Analysis of Credit Score for Credit Card Defaulters
Nupura Torvekar, Pravin S. Game - Enero 2019

https://www.researchgate.net/profile/Pravin-Game/publication/332557433_Predictive_analysis_of_credit_score_for_credit_card_defaulters/links/5f44ec63458515b7294fc74c/Predictive-analysis-of-credit-score-for-credit-card-defaulters.pdf

Que proponen:

Los autores le dan énfasis al contexto en el que se encuentra el sector bancario afirmando que es uno de los más volátiles y vulnerables en el mundo con sus factores de riesgo cada vez mayores.
El riesgo de crédito continúa siendo un factor integral para que las instituciones bancarias sufran pérdidas del orden de cientos de millones de dólares debido a la imposibilidad de recuperar el dinero concedido a los clientes.

Que utilizaron:

Naive Bayes, Regresión logística, máquinas de vectores de soporte y bosque aleatorio. Los algoritmos anteriores se evalúan utilizando entorno weka para aprendizaje automático y la minería de datos, además se usa KNIME que es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual.

Que encontraron:

Los autores afirman que el uso de técnicas de aprendizaje automático para la predicción de los morosos de tarjetas de crédito es esencial para la identificación de riesgo crediticio. Esto puede ayudar a las instituciones financieras a diseñar sus estrategias futuras.
En cuanto a los algoritmos utilizados los autores evaluaron los clasificadores utilizando la precisión de su predicción. Encontraron que el clasificador con la mayor precisión es el bosque aleatorio.

Segundo

Machine Learning Approaches to Predict Default of Credit Card Clients
Ruilin Liu - 2018.
https://www.scirp.org/html/7-7201927_88577.htm?pagespeed=noscript

Que proponen:
El autor afirma que la red neuronal puede explorar la relación entre las características de entrada y las etiquetas correspondientes, por lo que es adecuada para problemas complejos de aprendizaje automático. Por otro lado, otros modelos de aprendizaje automático como la regresión lineal o máquinas vectores de soporte pueden resolver problemas más simples de manera más eficiente. Por tanto, después de analizar problemas específicos, se debe responder a la pregunta de "¿es realmente necesaria la red neuronal en este caso?Además el autor menciona que existen varios tipos de redes neuronales.

Que utilizaron:
El autor utiliza varios modelos entre ellos k-vecinos cercanos, máquinas vector de soporte, árbol de decisión, bosques aleatorios, redes neuronales y redes neuronales recurrentes.

Que encontraron:
El autor afirma que los modelos tradicionales de aprendizaje automático solo pueden lograr una precisión de 0,8040, que se logra con SVM es decir mayor a los demás. La mayor precisión de la red neuronal es 0,8246, por lo tanto el autor concluye que las redes neuronales superan a los modelos tradicionales, excepto en situaciones en las que la investigación se centra fuertemente en predicciones positivas.

Tercero

Real Time Credit Card Default Classification Using Adaptive Boosting-Based Online Learning Algorithm
Hongya Lu, Haifeng Wang and Sang Won Yoon Department of Systems Science and Industrial Engineering State University of New York at Binghamton, Binghamton, NY- 2017

https://www.researchgate.net/profile/Haifeng-Wang-25/publication/319689046_Real_Time_Credit_Card_Default_Classification_Using_Adaptive_Boosting-Based_Online_Learning_Algorithm/links/59b991e1458515bb9c48a3f8/Real-Time-Credit-Card-Default-Classification-Using-Adaptive-Boosting-Based-Online-Learning-Algorithm.pdf

Que proponen:
Los investigadores proponen una aplicación de aprendizaje en línea al sistema de detección de incumplimiento de tarjetas de crédito que logra un ajuste del modelo en tiempo real con un mínimo esfuerzo computacional. Para los emisores de tarjetas de crédito, el

número de clientes de tarjetas de crédito, la cantidad de consumo y las tasas de incumplimiento son factores que influyen en la participación de los bancos en el mercado. Los autores además afirman que el banco sufrirá pérdidas debido a una mala gestión de la administración de tarjetas de créditos.

Que utilizaron:
Los autores utilizaron una técnica que se llama aprendizaje. Los investigadores explican que el aprendizaje en línea representa una familia de algoritmos eficientes y escalables en comparación con el aprendizaje por lotes tradicional, y aborda los problemas de la memoria. costes de consumo y reciclaje con nuevos datos entrantes.
Las técnicas exactas que utilizan los investigadores son métodos de máquina de aprendizaje extremo secuencial en línea (OS-ELM) y aumento de adaptación secuencial en línea (OS-AdaBoost). El OS-ELM está adaptado de la Extreme Learning Machine básica (ELM) para permitir que el modelo aprenda uno a uno fragmento a fragmento . El ELM básico propone una red feedforward con una única capa oculta generalizada que no funciona estrictamente como neuronas.

Que encontraron:
Los autores mencionan que con las técnicas en línea a otras técnicas con determinadas medidas de rendimiento, se obtienen resultados experimentales en comparación a modelos tradicionales de aprendizaje automático, además los autores afirman que el aprendizaje en línea tiene un gran potencial para los problemas de identificación en tiempo real y que esa idea genera futuras discusiones y un gran potencial investigativo.

Cuarto

Estimation of Credit Card Customers Payment Status by Using kNN and MLP
Murat KOKLU, Kadir SABANCI - 2016
https://www.ijisae.org/IJISAE/article/view/969/546

Que proponen:
Los autores afirman que para los bancos, lo más importante durante la comercialización de tarjetas de crédito es la capacidad de pago de los clientes. Los investigadores en el estudio proponen una estimación del estado de pago para los clientes de tarjetas de crédito. Para ello han utilizado algoritmos de minería de datos. Los autores describen la minería de datos como un proceso computacional que revela patrones en conjuntos de datos utilizando métodos como inteligencia artificial, aprendizaje automático, estadísticas, etc. Los métodos utilizados en la minería de datos se investigan en dos grupos: predictivos y descriptivos

Que utilizaron:
Los autores utilizaron entorno Weka para el análisis de aprendizaje automático y recurrieron al uso de modelos predictivos como k-nn, mlp(percepción multicapa)

Que encontraron:
Utilizando el algoritmo knn, los autores obtuvieron tasas de éxito de la estimación de pago para diferentes valores de k. Se han alcanzado el error medio absoluto (MAE) y el error cuadrático medio (RMSE), el éxito de la estimación del método k-nn (en porcentaje), además en el estudio mencionan el impacto basado en MAE y RMSE según cuantos k vecinos se elija.

Descripción de algoritmos básicos

Árbol de decisión

Es un modelo analítico que a través de una representación esquemática de las alternativas disponible facilita la toma de mejores decisiones para obtener un resultado. Su nombre deriva de la apariencia del modelo parecido a un árbol y su uso es amplio en el ámbito de la toma de decisiones bajo incertidumbre (Teoría de Decisiones).
Este modelo sirve para resolver problemas de clasificación.

La división entre valores de entrenamiento se hizo utilizando la técnica hold-out separando 30 test y 70 de entrenamiento.
En el problema que se busca solucionar se utilizó como criterio la entropía y de máxima profundidad un valor de 10.

Naive Bayes

En un sentido muy general, el modelo naive bayes es una clase especial de algoritmos de clasificación de Aprendizaje Automático el cual se basan en el teorema de bayes.
En este tipo de modelo se asume que las variables predictoras son independientes entre sí. En otras palabras, la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con cualquier otra característica.

La división entre valores de entrenamiento se hizo utilizando la técnica hold-out separando 20 test y 80 de entrenamiento.
Para implementar el algoritmo se usó como parámetro binarización de 0.5.

Reporte de resultados

Después de que se hayan implementado ambos algoritmos el que obtuvo mejores resultados con una diferencia mínima y tomando en consideración el accuracy fue naive bayes con una métrica accuracy de 0.793 frente a la métrica del árbol de decisión que obtuvo 0.776.

Conclusión

Finalmente podemos considerar que existen 3 posibles caminos si hablamos de preprocesamiento los cuales eran eliminar, reemplazar por media/mediana y re asignar los valores mal categorizados a la categoría que correspondan (otros).
En cuanto a las correlaciones que se pudieron encontrar en los diagramas 2D fueron el monto crédito otorgado vs el estado de cuenta donde existía una correlación positiva un tanto débil pero considerable, entre otras variables con relaciones interesantes. Cabe destacar que en esta primera etapa se puede evidenciar que existen relaciones entre las variables y que esto nos permitirá a futuro entrenar modelos predictivos de manera idónea.
Otro punto importante que podemos concluir en base a los paper mencionados en este trabajo, es que existen múltiples caminos para dar solución a este problema, algunos caminos priorizan la eficiencia en función del tiempo para eventualmente poder implementar la solución en entornos de tiempo real, otras formas utilizan algoritmos más tradicionales los cuales prácticamente son un estándar cuando de clasificar hablamos, sin embargo el método que se considera como mejor solución actual en condiciones normales es el uso de redes neuronales en otras palabras deep learning, este método es catalogado como la mejor solución en dos de los cuatro paper, y en la página oficial de ics uci.
Actualmente el deep learning, nlp y la inteligencia artificial nos ha llevado a avanzar a pasos agigantados en múltiples problemas que anteriormente la solución no era la mejor o simplemente no era óptima.

Bibliografía

- https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
- https://pandas.pydata.org/docs/
- https://matplotlib.org/stable/tutorials/introductory/pyplot.html#sphx-glr-tutorials-introductory-pyplot-py
- https://matplotlib.org/stable/gallery/images_contours_and_fields/image_annotated_heatmap.html

- https://www.researchgate.net/profile/Pravin-Game/publication/332557433_Predictive_analysis_of_credit_score_for_credit_card_defaulters/links/5f44ec63458515b7294fc74c/Predictive-analysis-of-credit-score-for-credit-card-defaulters.pdf
- https://www.ijisae.org/IJISAE/article/view/969/546
- https://www.researchgate.net/profile/Haifeng-Wang-25/publication/319689046_Real_Time_Credit_Card_Default_Classification_Using_Adaptive_Boosting-Based_Online_Learning_Algorithm/links/59b991e1458515bb9c48a3f8/Real-Time-Credit-Card-Default-Classification-Using-Adaptive-Boosting-Based-Online-Learning-Algorithm.pdf
- https://www.scirp.org/html/7-7201927_88577.htm?pagespeed=noscript
- https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html
- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html