Victor Barbe 903715

Carlos Santiago Castillo 164426

Emanuel Ramírez Robles 166457

Diego Rodríguez 163036

**Reporte**

Para poder demostrar conocimientos adquiridos de estadística y programación en nuestra clase de Minería de datos, se nos pidió elaborar un reporte en el cual se propusiera vender un producto, de tal manera que el método para lograr esta venta sean datos, extrayéndolos de una base de datos de nuestra preferencia para posteriormente limpiarlos y filtrarlos y poder usarlos para conseguir nuestro objetivo. Nuestro proyecto consiste en suponer que un banco nos contrata para que podamos investigar y relacionar los tarjetahabientes a los cuales se le pueda ofrecer un aumento en su línea de crédito relacionada a su tarjeta. En este reporte expondremos la información obtenida del análisis exploratorio de la base de datos “BankChurners”, obtenida de Kaggle, la cual presenta los datos de usuarios con una tarjeta de crédito e información financiera relacionada, todo esto para lograr el objetivo que se mencionó previamente. No es posible incrementar el límite de crédito a todas las personas que lo solicitan dado que el banco deberá tener la seguridad de que al cliente que le estén incrementando su línea sea una persona confiable y no represente un riesgo para la institución. Por otra parte, si nosotros encontramos más personas posibles para poder ofrecerles el crédito el banco es capaz de ajustar el presupuesto para no perder clientes. En otras palabras, nuestra meta es poder analizar los datos de tal manera que nuestro cliente pueda ofrecer mejores créditos reduciendo el riesgo de que caigan en cartera vencida por sobregirar su tarjeta.

La base cuenta con 10,126 datos y 23 variables , de las cuales 13 son numéricas (cuyas últimas dos se recomienda ignorarlas por parte de la persona que subió la base de datos a la página) y 10 categóricas. Dado el objetivo del análisis hemos decidido trabajar únicamente con 6 variables, 3 categóricas y 3 numéricas, para denotarlas comúnmente se usa el símbolo X, acompañado de un subíndice para poder distinguirlas. Así que por lo tanto X1  Edad, X2  Card estatus, X3  Income, X4  C\_limit, X5  C\_category y X6  Utilization. Posiblemente los inputs estén con nombres distintos por comodidad y practicidad, pero estas son las variables que se utilizaran para el desarrollo del proyecto.

Edad:

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

La variable “Edad” nos muestra la edad de los usuarios con una tarjeta de crédito. Cuyo dato mínimo es 26 y máximo 73. Además, la edad promedio de los usuarios es de 46 años y una desviación estándar de 8.0168.

Gráfico, Gráfico de barras

Descripción generada automáticamenteCard\_estatus :

Nos muestra el estatus de la tarjeta (Activa o Inactiva). Para esta base las tarjetas Activas representan el 83.99% de los datos, y tan solo 16.01% de las tarjetas permanecen Inactivas.

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamenteIncome :

Se refiere a la categoría de ingreso anual del dueño de la tarjeta. Es claro que hay un mayor porcentaje de usuarios un sueldo menor a $40k al año, representando el 27.95%. Seguido por el rango de $40k – $60k con un 18.6%. Y los usuarios con un ingreso anual de más de $120k representan el 3.8%.

C\_limit :

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

Muestra el límite de crédito de cada tarjeta registrada. Siendo $1,438.3 el valor mínimo, los créditos entre $1,438.3 y $3,000 representan el 35% de los datos. Por otro lado, el valor máximo es de $34,516. El valor promedio de los límites de crédito es de $4,549, con una desviación estándar de $9,088.78.

Gráfico, Histograma

Descripción generada automáticamenteC\_category:

Nos muestra el tipo de tarjeta (Blue, Silver, Gold, Platinum). Es fácil ver que la mayor parte de las tarjetas son de tipo “Blue” representando el 93.18% de los datos, y el tipo de tarjeta menos común es el Platinum que representa tan solo el 0.20%.

Gráfico, Histograma

Descripción generada automáticamenteUtilization :

Se refiere al porcentaje promedio de uso del crédito. De igual forma vemos que la mayor concentración de los datos se encuentra entre el 0% y el 10% de uso, que representan el 49% de los datos. Además, el valor máximo de uso del crédito ha sido del 99.9%, pero con un promedio de uso de 27.49% y una desviación estándar de 27.57%.

En total, usando estos datos tenemos 105 usuarios duplicados, lo que representa el 1% de los datos.

Así mismo como parte del objetivo también se plantean preguntas importantes las cuales servirán a lo largo de la investigación:

1. ¿Existe una relación entre la edad y el ingreso?

Es importante saber si la edad tiene alguna correlación con el ingreso de una persona, para poder tomar una decisión más precisa. Si la evidencia llegara a mostrarse débil o no nos favorezca, se argumentará que no es necesario fijarse en la edad.

1. ¿Es un factor importante saber si la tarjeta esta activa o inactiva?

Creemos que la variable Card\_estatus es relevante para saber las razones de porque x personas dejaron de usar su tarjeta, para así poder ver a quienes les podemos dar el crédito.

1. ¿Cuántas personas están en el límite de crédito?

¿Depende el número de personas que estén a punto de terminarse el crédito para que sean posibles candidatas a un aumento de crédito? ¿Las personas no usan su crédito por no tener tanto disponible?

1. ¿Existe algún tipo de correlación entre las variables?

Tal vez solo es coincidencia que las personas que más usan su tarjeta son las que menos crédito tienen o incluso la edad no es un factor tan importante, entonces respondiendo esta pregunta finalmente se determinara si existe algún tipo de relación entre todas las variables.

Enlace de descarga de la base:

https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers