UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
TAREA 5

Maestría en Ciencia de Datos
ACT. ERICK ADRIÁN GARZA TAMEZ

Marzo 2023

Índice

1. Introducción

Este reporte busca generar un **aprendizaje no supervisado** para predecir de si un cliente va a caer en morosidad o no. Esta es una base de datos la cual contiene las siguientes características:

* *Monto del Préstamo*: Es el monto del pr´stamo que dispuso el cliente.

* *Term*: Es el plazo del préstamo.

* *Interest Rate*: Tasa de intéres del préstamo.

* *Grade*: Es el nivel de riesgo asignado por el banco, siendo 1 el mejor y 7 el peor.

* *Loan Status*: Esta es nuestra variable de ínteres y nos índica si el cliente es o no moroso.

El aprendizaje no supervisado es una técnica de aprendizaje automático en la que el modelo debe encontrar patrones y relaciones en los datos sin la ayuda de etiquetas o respuestas conocidas previamente. En otras palabras, el modelo se encarga de descubrir información útil a partir de datos sin saber de antemano cuáles son los resultados esperados.

En el contexto de predecir si los clientes de un banco pueden caer en morosidad, el aprendizaje no supervisado puede ser útil para identificar patrones en los datos de los clientes que puedan indicar un mayor riesgo de incumplimiento en los pagos. Por ejemplo, un modelo de clustering podría agrupar a los clientes en diferentes grupos según sus características, como su historial crediticio, sus ingresos, su edad, entre otros. Luego, se podría analizar si hay algún grupo que tiende a tener una mayor proporción de clientes en morosidad y, por lo tanto, tomar medidas para mitigar ese riesgo.

Para este caso en particular utilizaremos el ***método de Feature Scalling*** y el ***método del Codo*** para encontrar el mejor número de *clusters*.

Sin embargo, es importante tener en cuenta que el aprendizaje automático no es una solución infalible y siempre es necesario evaluar los resultados y tomar medidas adicionales para mitigar los riesgos identificados.

2. Feature Scalling

El Feature Scaling es una técnica utilizada en el preprocesamiento de datos que se utiliza para normalizar las características (features) de un conjunto de datos para que tengan una escala similar. En el contexto del aprendizaje no supervisado, el Feature Scaling puede ser útil para mejorar la precisión del modelo y reducir el tiempo de entrenamiento.

El motivo principal por el que se aplica Feature Scaling es porque los algoritmos de aprendizaje no supervisado funcionan mejor cuando las características tienen una escala similar. Si las características tienen diferentes escalas, los algoritmos pueden tener dificultades para encontrar patrones precisos en los datos y, en consecuencia, pueden generar resultados incorrectos. Al normalizar las características a una escala común, el modelo puede funcionar mejor y hacer predicciones más precisas.

Dicho esto, la siguiente tabla muestra las primeras 10 filas de nuestra base de datos:

Loan Amount	Term	Interest Rate	Grade	Loan Status
10000	59	11.135007	2	0
3609	59	12.237563	3	0
28276	59	12.545884	6	0
11170	59	16.731201	3	0
16890	59	15.008300	3	0
34631	36	17.246986	2	0
30844	59	10.731432	3	0
20744	58	13.993688	1	0
9299	59	11.178457	7	0
19232	58	5.520413	3	0

Aplicando nuestro método de Feature Scalling buscando una escala similar en los ditintos conjuntos de datos obtenemos lo siguiente:

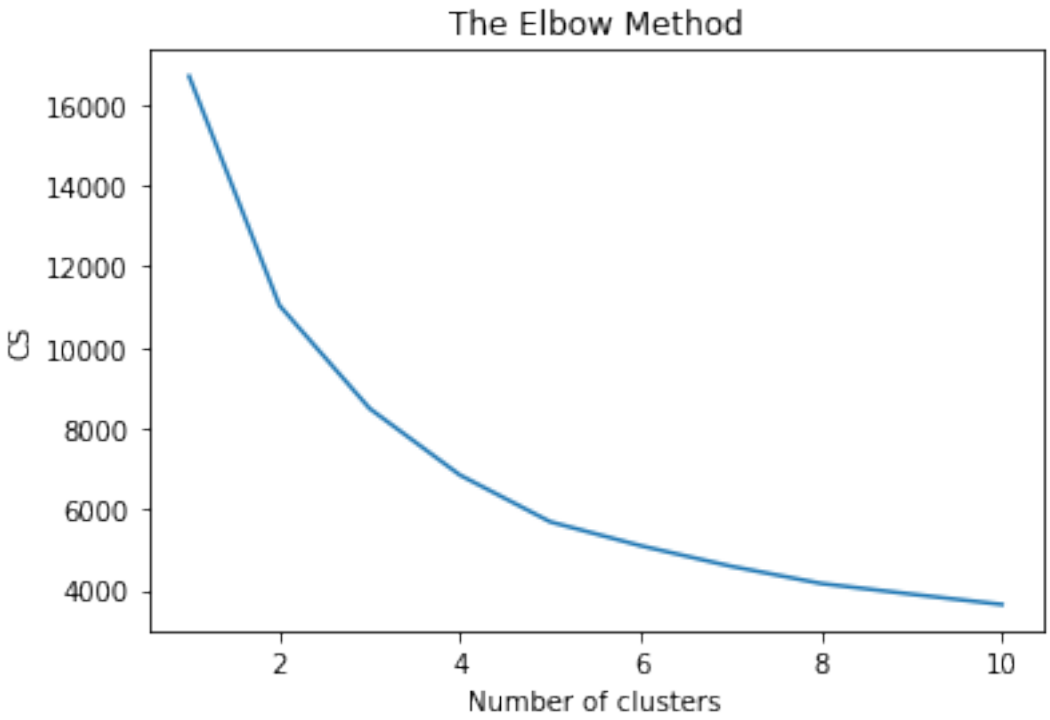
Loan Amount	Term	Interest Rate	Grade	Loan Status
0.264403	1.000000	0.265983	0.166667	0.0
0.076355	1.000000	0.316414	0.333333	0.0
0.802154	1.000000	0.330517	0.833333	0.0
0.298829	1.000000	0.521957	0.333333	0.0
0.467134	1.000000	0.443150	0.333333	0.0
0.989143	0.000000	0.545549	0.166667	0.0
0.877714	1.000000	0.247523	0.333333	0.0
0.580533	0.956522	0.396741	0.000000	0.0
0.243777	1.000000	0.267970	1.000000	0.0
0.536044	0.956522	0.009167	0.333333	0.0

3. Método del Codo

El método del codo es una técnica utilizada en el aprendizaje no supervisado para determinar el número óptimo de clusters (grupos) en un conjunto de datos. El método se llama así porque el gráfico de la variabilidad explicada por el número de clusters forma una curva que se parece a un codo.

Es útil porque puede ayudar a determinar el número óptimo de clusters en un conjunto de datos de manera objetiva, en lugar de tener que adivinar el número correcto de clusters. Además, puede ayudar a evitar el sobreajuste (overfitting) o el subajuste (underfitting) del modelo de clustering.

Dicho esto, nuestro modelo nos arrojó los siguientes resultados:



De la imágen anterior podemos llegar a la conclusión que la mayor distancia se encuentra entre 4 y 5 clusters, se hará la prueba para 4, 5 y 6 clusters para así identificar cual tiene el mejor rendimiento.

A continuación se muestran los resultados obtenidos.

* **4 Clusters** : Resultado: 27310 de 67463 muestras estan bien etiquetadas. Precisión: 0.40

* **5 Clusters** : Resultado: 19772 de 67463 muestras estan bien etiquetadas. Precisión: 0.29

* **6 Clusters** : Resultado: 25461 de 67463 muestras estan bien etiquetadas. Precisión: 0.38

4. Conclusiones

De lo anterior podemos concluir que nuestro modelo de aprendizaje no supervisado no fue muy preciso dado a que nuestro mayor rendimiento deacuerdo al numero de clusters fue de 40 %, tomando únicamente 4 de ellos.

Por lo tanto, se buscará una mayor precisión con un modelo de aprendizaje supervisado para así incrementar el rendimiento del modelo.