



PROYECTO FINAL **REFERIDOS**

INTEGRANTES

Greycivier Espinoza

Maria Sol Pierozzi Luque



1.

**MOTIVACIÓN Y
AUDIENCIA**

2.

**PROBLEMA
COMERCIAL**

3.

PREGUNTAS/HIPÓTESIS

4.

OBJETIVO Y CONTEXTO

5.

**DESCRIPCIÓN DEL
DATASET**

6.

EDA

7.

**SELECCIÓN DEL
MODELO**


8.

CONCLUSIONES




1.

MOTIVACIÓN Y AUDIENCIA




El dataset seleccionado para analizar está compuesto por los referidos de un Sistema que otorga Créditos Personales a una población que no sería aprobada para un préstamo en una entidad bancaria.




Los datos son analizados por un Motor de Riesgos que determinará si finalmente se consideran aptos para recibir una oferta o no.

La política presente en el motor de decisión se basa en el análisis de distintos aspectos socioeconómicos de la persona, para luego asignarles a cada uno un puntaje o *score*.



De este puntaje depende la aprobación o el monto que se le otorgará al cliente.

Los referidos, se agrupan por su nivel socioeconómico y su situación actual (Mora de un cliente con otras entidades) y es posible identificar si su género es masculino o femenino.




Encontramos importante observar cómo se comporta el monto según situación actual, score, género y nivel socioeconómico; ya que estos valores nos permitirán atraer una mejor población para incrementar la otorgación y disminuir la mora.



2.

PROBLEMA COMERCIAL






El número de referidos rechazados es muy alto y analizar estos casos genera una pérdida, puesto que se utilizan otras entidades para tomar la decisión final.


Por este motivo, debemos buscar una mejor población tomando como referencia aquella que ha sido aprobada para obtener una mejor tasa de aprobación y disminuir los costos.



3. PREGUNTAS/HIPÓTESIS



- 
- ¿Existe algún factor común entre los referidos aprobados?
 - ¿Existe una diferencia en aprobados por su género?

- 
- ¿Existe algún sector socioeconómico que se destaque por poseer mayor/menor número de aprobados?
 - En base a los casos que analizamos diariamente, notamos que la mayoría de referidos aprobados son hombres en edades promedio entre 18 y 25 años. ¿Esto será correcto?



4.

DEFINICIÓN DEL OBJETIVO

-

CONTEXTO ANALÍTICO Y COMERCIAL



OBJETIVO

El objetivo principal de nuestro análisis es identificar los rasgos de los referidos que posiblemente serán aprobados por el motor para así poder tomar acciones por ej: campañas de publicidad para poder atraer a ese grupo poblacional.



CONTEXTO COMERCIAL

En el sistema, los clientes pueden provenir de dos fuentes: quienes inician una solicitud desde la web y los **Referidos**.

Estos últimos, son una fuente fundamental para el ingreso de clientes puesto que proveen información sobre clientes potenciales (pueden ser aprobados o rechazados). En este caso, nos centraremos únicamente en el funcionamiento de los referidos.



CONTEXTO ANALÍTICO

El equipo de riesgos ha recibido datos sobre los últimos 50.000 referidos recibidos a partir del 30 de Agosto de 2022.

Los mismos, identifican mediante la variable *approved* si han sido o no aprobados respectivamente. Debemos utilizar modelos de **agrupamiento** para abordar este problema de **aprendizaje no supervisado**.

5.

DESCRIPCIÓN DEL DATASET





DESCRIPCIÓN DE LAS COLUMNAS

- *id*: Número único que identifica a cada referido
- *gender*: Número que identifica el género (1 - masculino / 0 -femenino)
- *entity_code*: Número de 3 dígitos que identifica a una entidad bancaria
- *person_query_id*: Id relacionado con el análisis de riesgo creado.



DESCRIPCIÓN DE LAS COLUMNAS

- *approved*: Número que clasifica a los aprobados/no aprobados por el motor (1 para aprobados, 0 no aprobados)
- *current_situation*: Número que representa la situación actual del individuo ante el BCRA (1, 2, 3, etc)
- *score_risk*: Double - Puntaje obtenido por el motor de riesgo.
- *score_nosis*: Double - Puntaje obtenido por el motor de nosis (analiza nivel crediticio)



DESCRIPCIÓN DE LAS COLUMNAS

- *socioeconomic_level*: String - Nivel socioeconómico del individuo en base al BCRA (A, B, C3, D2, etc)
- *cda*: String que representa la situación ante el BCRA (Aprobado, Observado, etc)
- *rate*: Tasa asignada por el motor.
- *term*: Numérico - Cantidad de cuotas asignadas por el motor,
- *min_term*: Mínima cantidad de cuotas asignadas.



DESCRIPCIÓN DE LAS COLUMNAS

- *max_amount*: Double - monto máximo asignado por el motor.
- *birthday*: Fecha de nacimiento de la persona
- *informed*: Indica si fue informado al referido luego de ser aprobado
- *tries*: Cantidad de intentos de un mismo individuo dentro de un período de tiempo.
- *created_at*: Fecha de creación
- *updated_at*: Fecha de última actualización



CANTIDAD DE REGISTROS

El dataset actual cuenta con **50.000** registros.

En un comienzo tenía 10.000, pero decidimos incrementarlo con el objetivo de mejorar el análisis.



6.

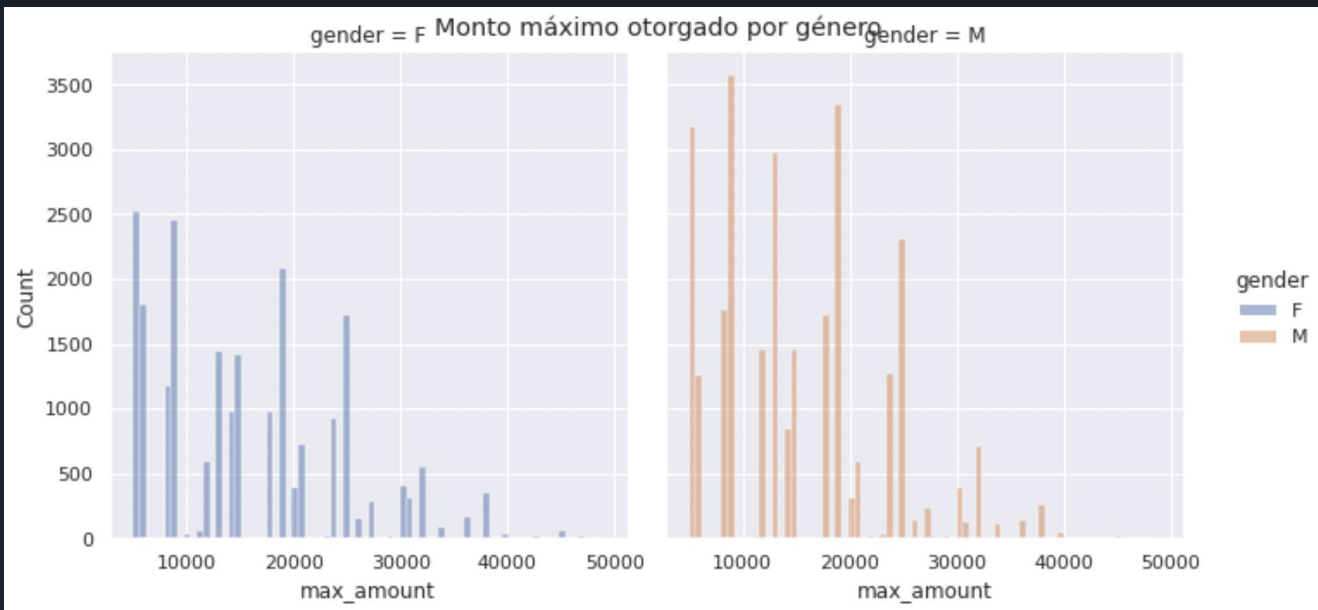
EDA

¿ Hay una diferencia significativa entre aprobados hombres y mujeres?



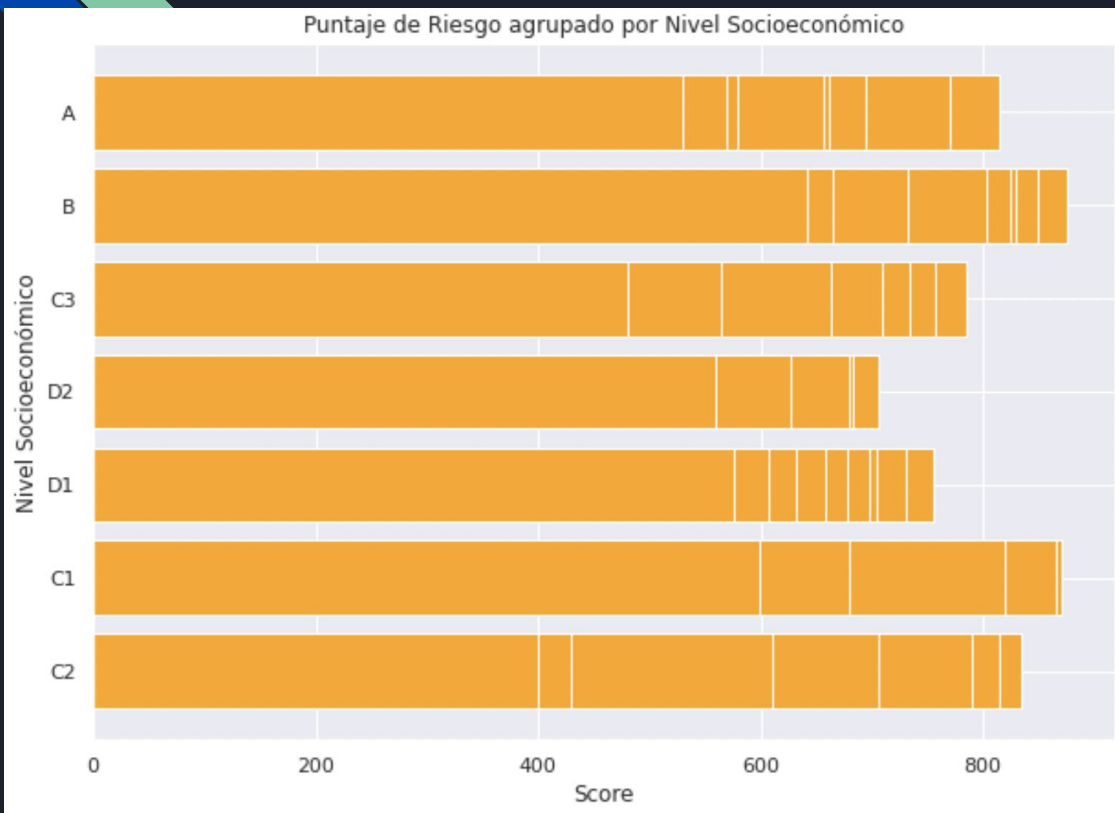
Los aprobados en su mayoría son hombres, superando casi en un 20% al porcentaje femenino. Sin embargo, no consideramos que el género sea un factor determinante.

¿A quiénes se les otorgan mayores montos?



Se les otorgan mayores montos a los hombres que a las mujeres, siendo la cantidad más otorgada de \$10.000 en hombres y \$5.000 en mujeres.

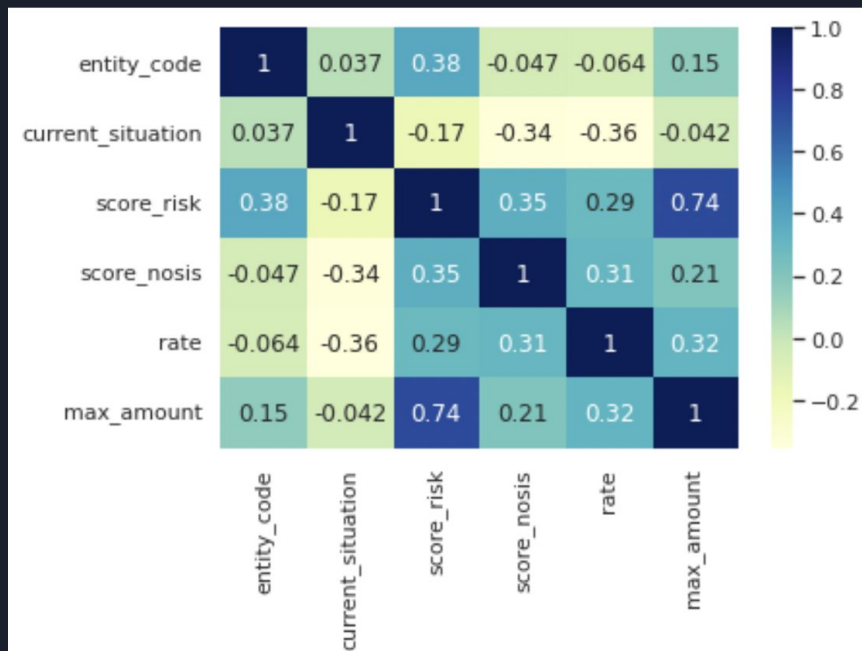
PUNTAJE DE RISK AGRUPADO POR NIVEL SOCIOECONÓMICO



No existe una variación muy marcada entre el score otorgado por el motor de riesgo en cuanto al nivel socioeconómico de la persona.

El nivel que más puntaje obtuvo es el Nivel Socioeconómico **C1**, casi a la par del nivel **B** que corresponde a profesionales independientes mientras que la clase **A** representa a la clase socioeconómica más alta dentro de nuestro problema.

CORRELACIÓN ENTRE LAS VARIABLES



Podemos observar que las columnas *score_risk* y *max_amount* se encuentran en esta situación se encuentran más cercanas a 1.

Para nuestro análisis, no es indispensable conocer el monto máximo otorgado puesto que nuestro objetivo es incrementar los clientes aprobados independientemente del monto.

Por este motivo, optamos por **eliminar la columna *max_amount*** del análisis.

CORRELACIÓN ENTRE LAS VARIABLES

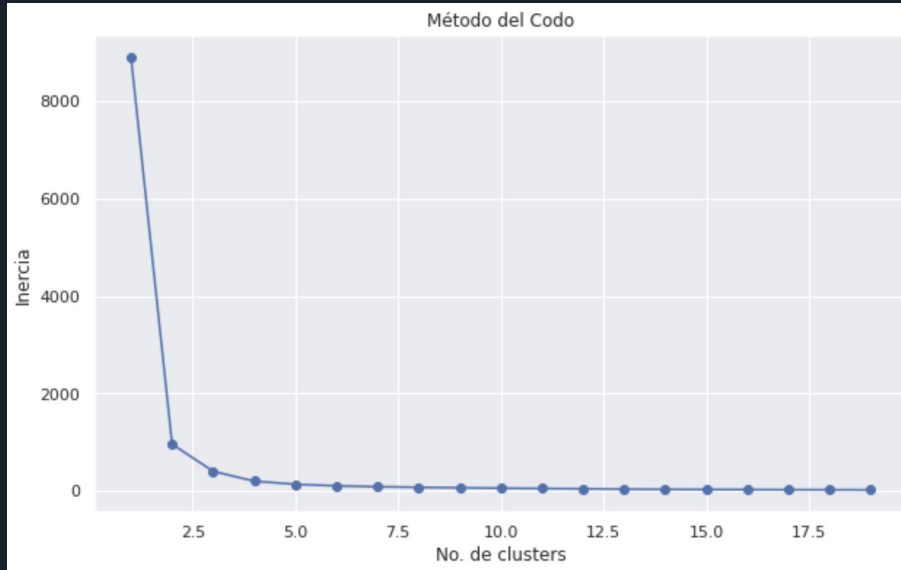


Con la columna *max_amount* eliminada, podemos afirmar que los demás valores no de la tabla no resultarán un "riesgo" para el modelo puesto que se encuentran lejanos al valor 1

GRÁFICOS EN BASE A LA COLUMNA SOCIOECONOMIC_LEVEL

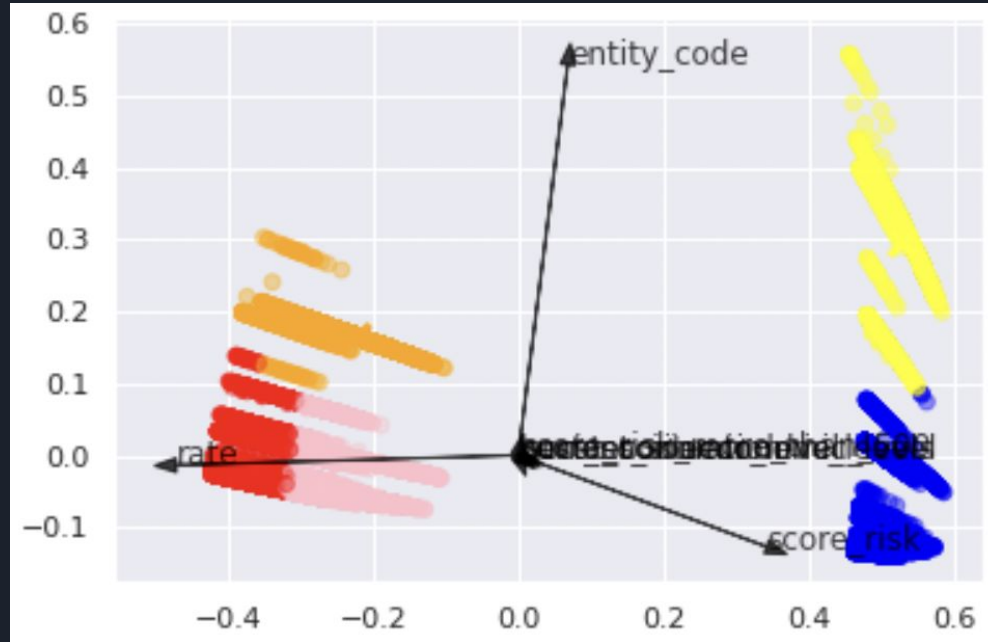


MÉTODO DEL CODO

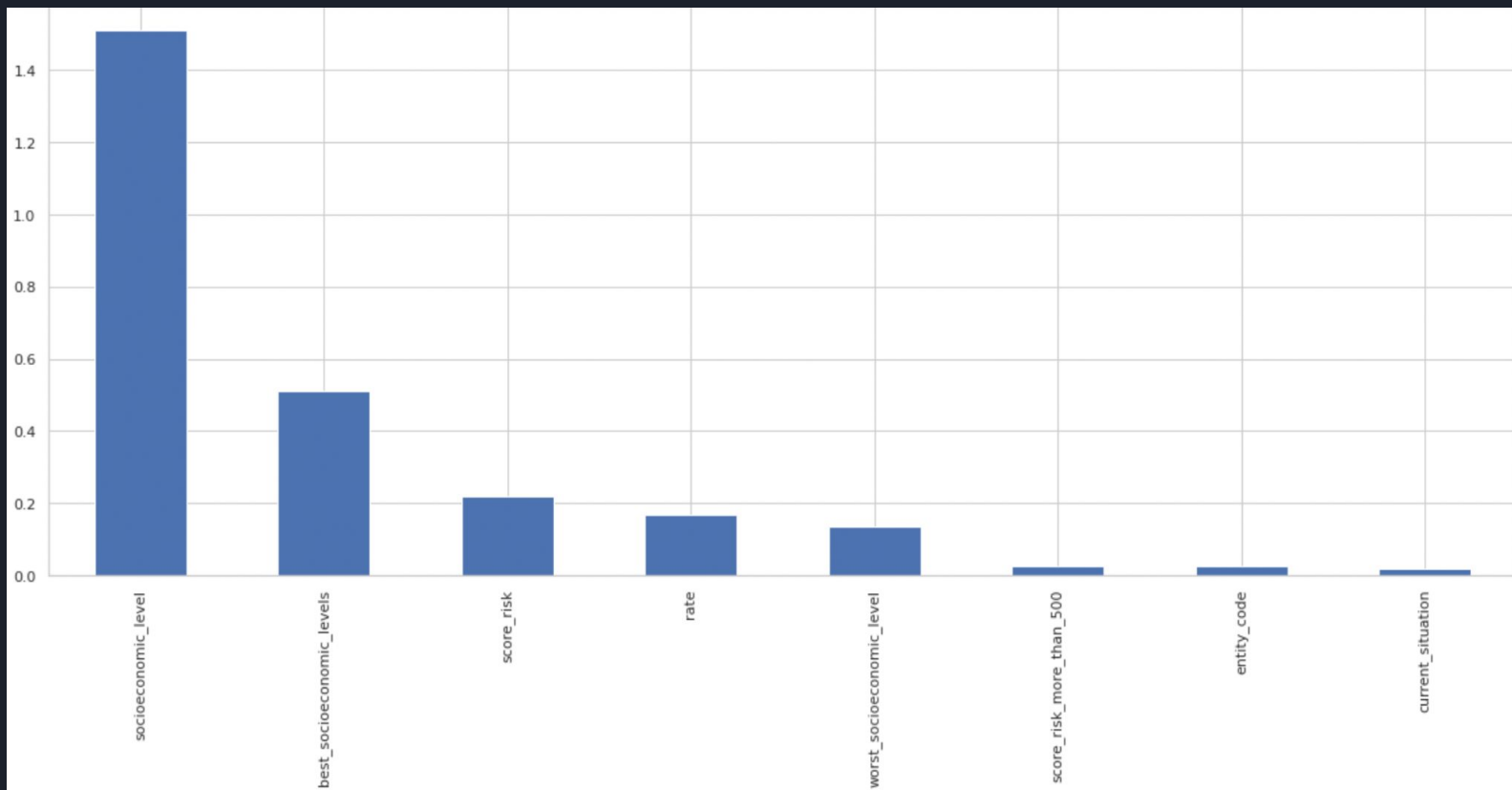


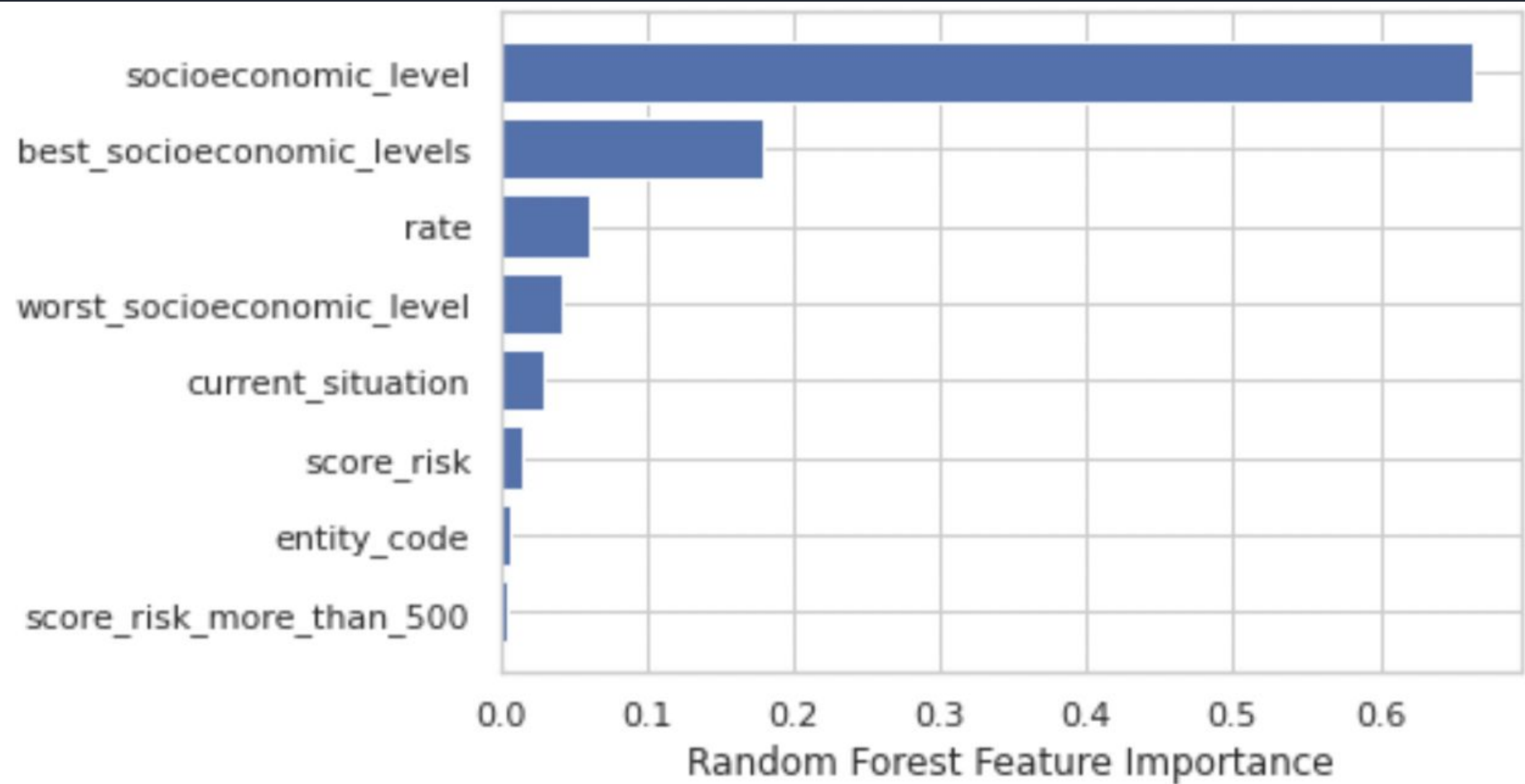
A partir del método del codo realizado, decidimos analizar 5 clusters.

GRÁFICO DE LOS CLUSTERS OBTENIDOS UTILIZANDO K-MEANS



ANÁLISIS DE LA IMPORTANCIA DE CADA VARIABLE







CONCLUSIONES

- Identificamos cuáles son los features más importantes (*score_risk*, *socioeconomic_level*, *rate*, *best_socioeconomic_levels*, *worst_socioeconomic_level*)
- Identificamos cuáles columnas poseían correlaciones altas y decidimos eliminar la columna más afectada: *max_amount*. De esta manera, logramos observar un cambio positivo en el gráfico de correlaciones.
- Observamos que tanto *xTest* como *xTrain* no poseen valores no deseados como *missing_values* o valores no numéricos.



CONCLUSIONES

- En cuanto al análisis de género, no encontramos diferencias muy marcadas, por lo que descartamos la hipótesis inicial de que se aprueban más ofertas del género masculino. Sin embargo, los montos otorgados sí son mayores.
- Concluimos en que la variable que más influye en la aprobación es el nivel socioeconómico, puesto que los que poseen niveles más altos son los más aprobados. Por lo que se podría aplicar en futuro una estrategia de marketing para atraer a ese sector, con el objetivo de mejorar las ventas.
- Decidimos utilizar clustering (K-means) para identificar a los distintos grupos.