Factores de predisposición para otorgar créditos automatizados en línea

AUTOR: Arturo González Mendoza

Contenido

Contenido

CONTEXTO Y **AUDIENCIA**

Contexto

El presente trabajo tiene como objeto identificar las variables a recolector que influyen en el proceso de otorgamiento en la línea de crédito en línea. Hoy en día los créditos se pueden otorgar de manera inmediata y a través de un simple proceso en línea, sin embargo no hay suficientes recursos humanos que puedan interpretar la información en el momento para determinar si es posible otorgar un crédito. Es por esto que una posible alternativa, es el uso de machine learning en donde utilizando datos personales del usuario, y de su localidad, se puede determinar en el momento si se le otorga un credito, minimizando los riesgos de que el usuario al que se le otorgue el crédito no realice el pago de este.

Es por eso que este proyecto tiene como propósito identificar las variables importantes para el otorgamiento de un crédito y crear un modelo de machine learning que evalue si la persona es candidata o no al credito.

Audiencia

Este análisis intenta mostrar con evidencia, aquellos datos personales útiles para el otorgamiento de un crédito inmediato. Es por eso que este modelo está orientado a todas las personas que tienen un negocio de préstamos o a instituciones financieras que quieran agilizar sus procesos de otorgamientos de crédito.

Limitaciones

No se encontraron muchos datasets con información, y los disponibles es posible que sean analizados por separado ya que contienen información que no sea la misma. Al integrarse los datos al primer dataset nos puede agregar un error en el análisis.

PREGUNTAS DE **INTERÉS**

Preguntas principales o primarias

Hipótesis de interés

La empresa quiere automatizar el proceso de concesión de préstamos (en tiempo real) basándose en los datos que el cliente facilita al cumplimentar los formularios de solicitud en línea.

Hipotesis primarias

A) Se presume que distribución de ingresos es normal y no uniforme.

B) Se presume que los ingresos en los diferentes categorias de edad son diferentes

C) Se considera que los casados tienen más dinero que los solteros

D) Los datos demográficos no tienen relación con el cumplimiento de los créditos

E) Los datos Personales representan mayor información para la aprobacion de creditos

RESUMEN **METADATA**

21 **Variables, relacionadas con el usuario y el estado de su crédito**

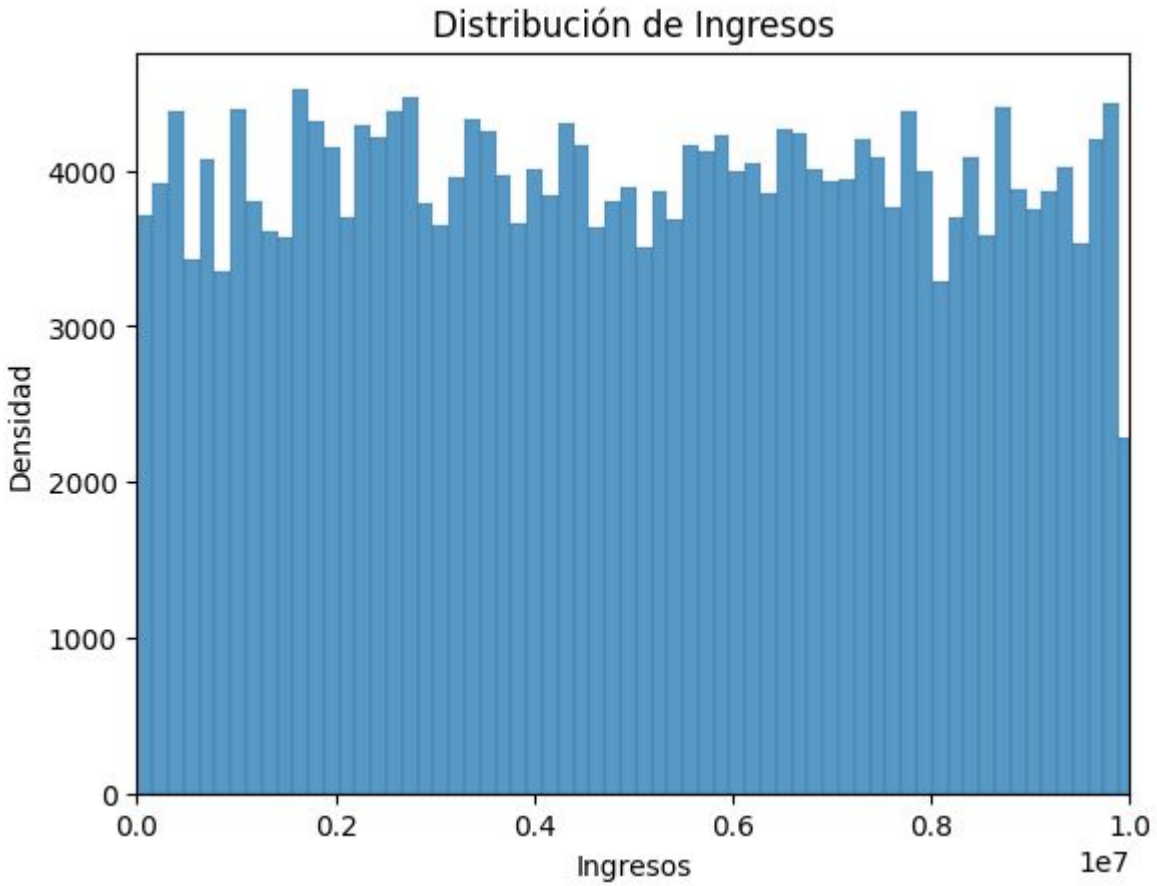
10 **Variables de datos personales del solicitante de crédito**

11 **Variables sociodemográficas del solicitante de crédito**

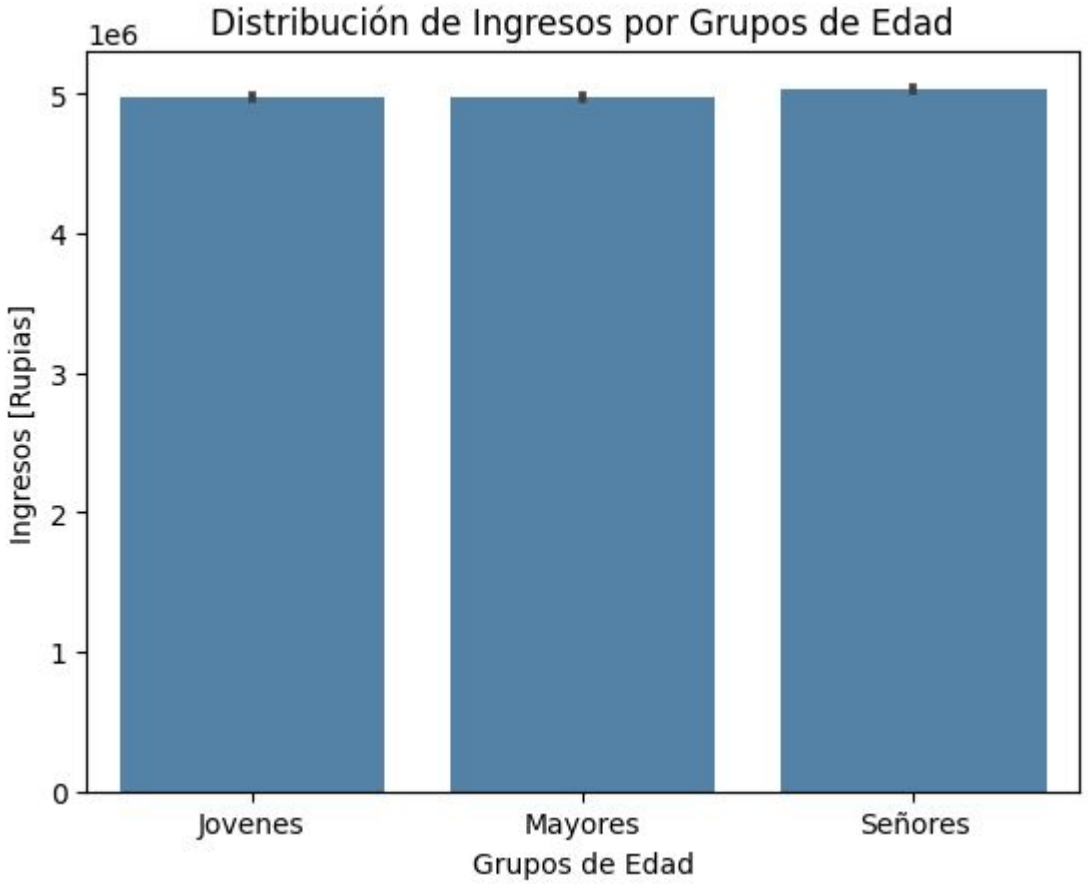
252981 **Datos en total**

ANÁLISIS **EXPLORATORIO**

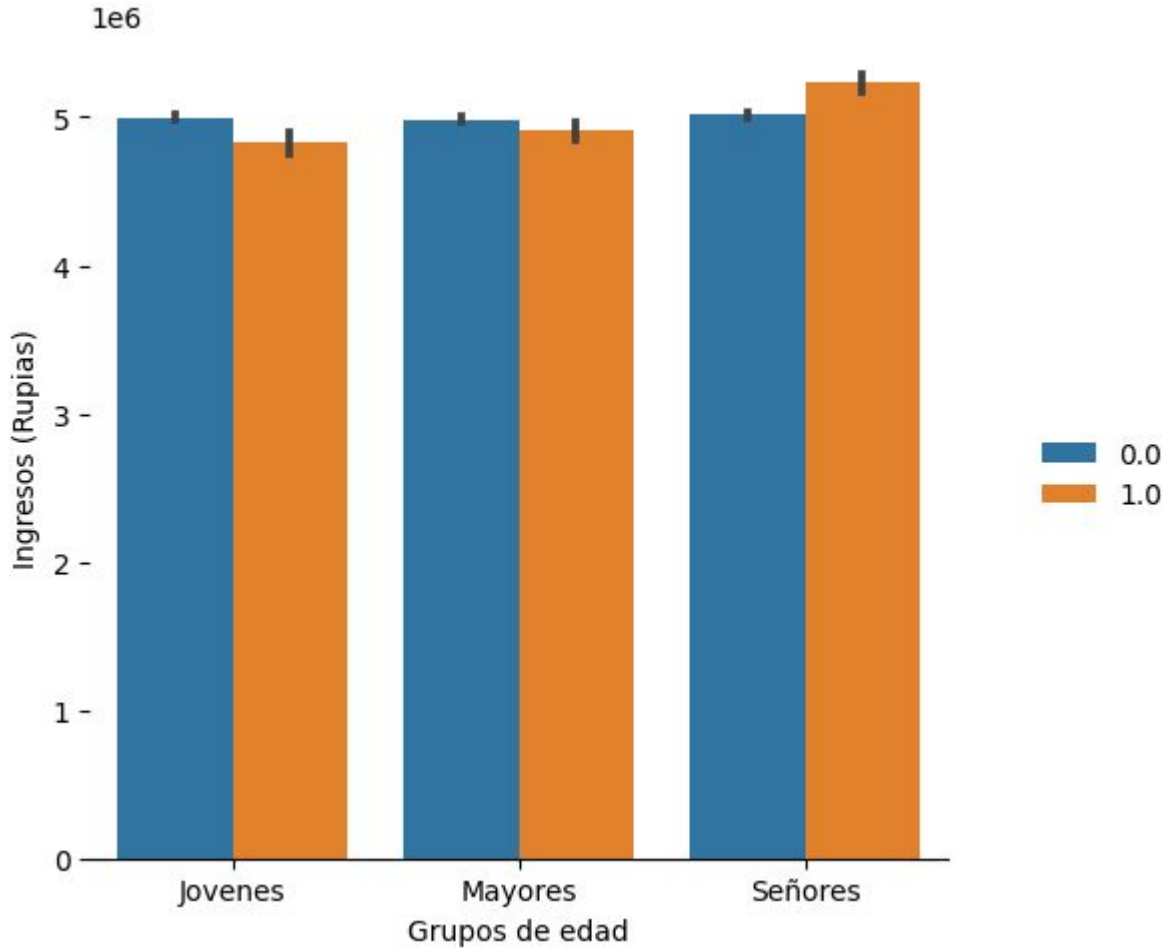
A) Se presume que distribución de ingresos es normal y no uniforme.



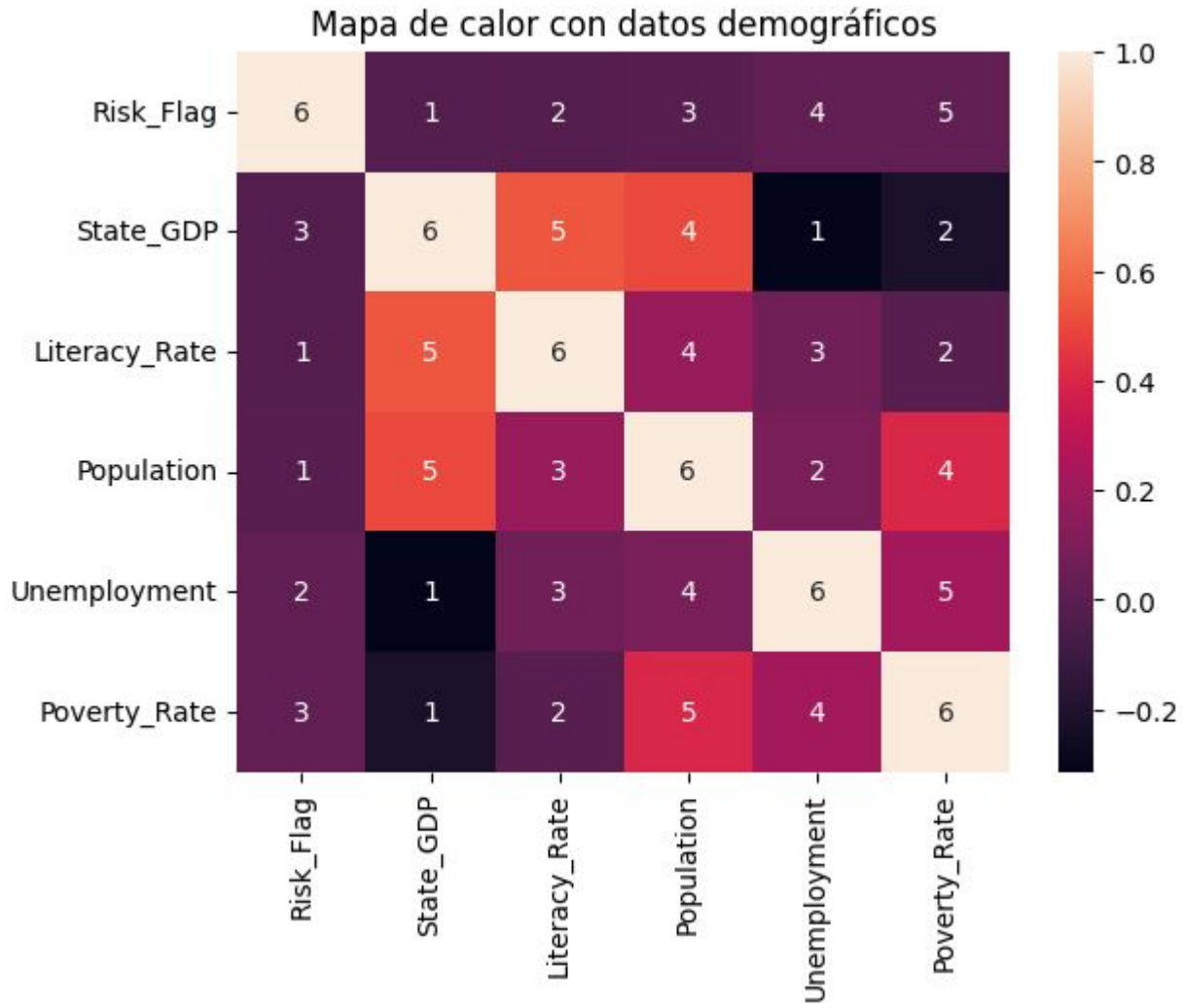
B) Se presume que los ingresos en los diferentes categorias de edad son diferentes



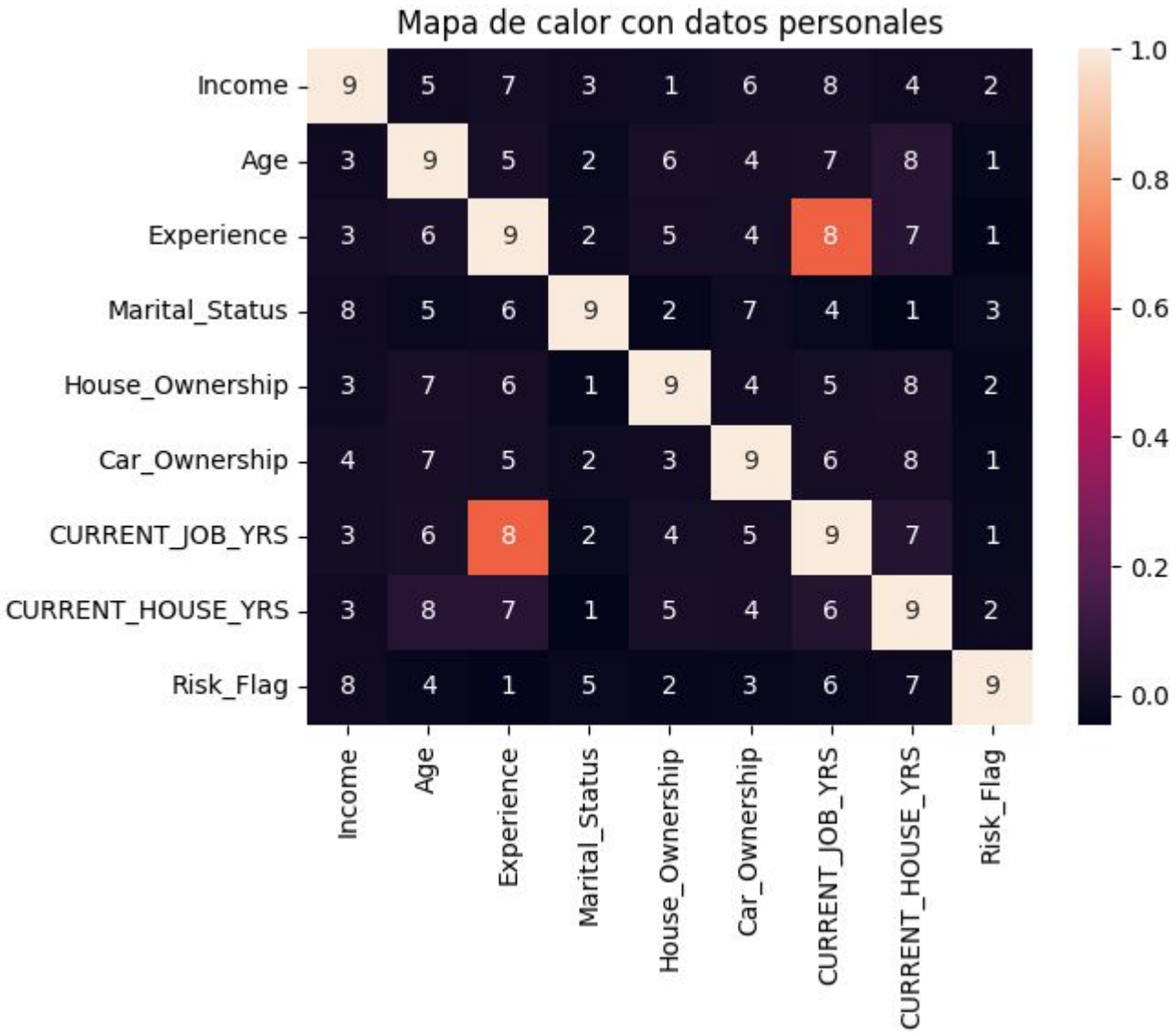
C) Se considera que los casados tienen más dinero que los solteros



E) Los datos Personales representan mayor información para la aprobacion de creditos



D) Los datos demográficos no tienen relación con el cumplimiento de los créditos



INSIGHTS & **RECOMENDACIONES**

INSIGHTS & **RECOMENDACIONES**

Insights

- A. En la gráfica de distribución de ingresos podemos observar un comportamiento uniforme. Debido a que la distribución de salarios es uniforme, no podemos asumir que hay un grupo de salarios que podamos identificar para enfocarnos en este segmento para créditos.
- B. En la gráfica de "Distribución de Ingresos por Grupos de Edad", podemos ver como el salario no se inclina por un grupo de edad, por lo que podemos determinar que no es un factor determinante para la entrega de un salario.
- C. En la gráfica de grupos de edad por ingresos y estado civil, podemos observar que los señores casados, presentan un notable mayor ingreso que los otros grupos por lo que este si puede ser un factor importante para la evaluación de un crédito.
- D. En el mapa de calor de variables socioeconómicas podemos observar que entre ellas hay relaciones lineales. Sin embargo es importante señalar que con respecto a la aprobación de crédito todas variables sociodemográficas se encuentran por un valor menor a 0.2, por lo que podemos decir que no hay una correlación alta en alguna de estas variables.
- E. Al igual que en los datos personales no hay correlaciones mayores a 0.2 con respecto a riskflak. Es decir los datos no se relacionan de una manera directa.

Recomendaciones

- La determinación del crédito no puede hacerse por relaciones simples entre datos, sino que tienen que hacerse búsquedas más complejas para buscar conjuntos de datos que nos den más información para la entrega de créditos
- Un ejemplo de esto puede ser utilizando algoritmos tipo stepwise, o métodos de componentes principales. Aplicando un algoritmo stepwise.backward elimination and forward selection, se recomienda el siguiente set de variables para la determinación de un crédito a traves de estos datos.

```
['Experience', 'Per_Capita_Income', 'House_Ownership', 'Car_Ownership', 'Age', 'Marital_Status', 'Region',
'Unemployment', 'CURRENT_JOB_YRS', 'State_GDP', 'Income_Category', 'CURRENT_HOUSE_YRS']
```

Evaluación de **modelos**

Evaluación de modelos

Ordenados de menor a mayor

Modelo	**Precisión**
KNN	0.93938316
SVM	0.93938316
KNN Características seleccionadas	0.93938316
SVM con características seleccionadas	0.93938316
Árbol de decisión	0.93993656
Árbol de decisión con Características seleccionadas	0.93999585
Random Forest	0.94694297
Random Forest características seleccionadas	**0.94704179**

Optimización de modelado

Modelo: Random Forest

Mejores hiperparametros :

1.'max_depth': None,
2. 'min_samples_split': 5,
3. 'n_estimators': 200

Precisión en pruebas :

0.9495

Selección de **modelos**

Selección de modelos

Árbol de decisión

Average expected loss: 0.060
Average bias: 0.051
Average variance: 0.030

Random Forest

Average expected loss: 0.053
Average bias: 0.053
Average variance: 0.001

Conclusiones

Conclusiones

En este caso podemos observar que el porcentaje de clasificacion en los diferentes modelos de entrenamiento entre los features completos y los features seleccionados es muy parecido entre los modelos.

El mejor dataset que brinda mejores resultados es el Random Forest con un % de precisión de 94% tanto para el set de datos completo y el reducido. Con la nueva columna de datos sintéticos se aumento el % a 94% de precisión.

El mejor modelo con precisión es el Random Forest y con una menor varianza por lo que se recomienda utilizar.

Se puede decir que es únicamente con las variables:

['Experience', 'Per_Capita_Income', 'House_Ownership', 'Car_Ownership', 'Age', 'Marital_Status', 'Region', 'Unemployment', 'CURRENT_JOB_YRS', 'State_GDP', 'Income_Category', 'CURRENT_HOUSE_YRS'], solicitadas en un formulario, se podrían agilizar los trámites de préstamo bancario.