Trabajo Práctico Equidad en el Aprendizaje Automático

1er cuatrimestre 2025 - Ezequiel Martinez Lopez, María Belén Quiñones, Lucia Vazquez

Introducción

En el siguiente informe desarrollamos los análisis, procedimientos y resultados que realizamos con el conjunto de datos "German Credit Data" y que volcamos en el repositorio. Buscamos formalizar lo detallado en el notebook realizado aunque toda la información que presentamos a continuación se encuentra duplicada ahí. Durante el trabajo nos enfocamos en analizar los datos, su distribución, sus posibles sesgos y armar un modelo que prediga si a una persona se le debe aprobar o no un préstamo bancario, todo esto desde la postura del banco y haciendo foco en la asignación de créditos a personas de distintos géneros.

En relación al conjunto de datos recopilamos la siguiente información:

Motivación

El conjunto de datos fue donado por el Profesor Dr. Hans Hofmann, del Institut für Statistik und Ökonometrie, Universität Hamburg, Alemania, como parte del proyecto europeo Statlog. Fue creado para clasificar a personas como buenos o malos riesgos crediticios en base a un conjunto de atributos. Su finalidad principal es apoyar tareas de clasificación en el ámbito financiero, específicamente para evaluar la solvencia crediticia de clientes.

No se indica una brecha específica que necesitaba ser cubierta con su creación pero sí se diseñó para facilitar el desarrollo y evaluación de modelos de predicción de riesgo crediticio.

Composición

El conjunto de datos está compuesto por 1000 instancias que representan a personas que solicitan un crédito bancario. De cada una de estas instancias se nos dan 20 atributos que incluyen tanto variables categóricas como numéricas, describiendo aspectos financieros y personales del solicitante, como por ejemplo el estado de la cuenta, la duración del crédito solicitado, el historial crediticio, el propósito, el monto del crédito, la edad de la persona, su sexo y su empleo , entre otros. No hay varios tipos de instancias, todas representan personas solicitantes de crédito.

Proceso de recopilación

Los datos provienen de registros bancarios reales, por lo que fueron observados directamente a partir de la información financiera y personal de los solicitantes. No se especifica que hayan sido informados por sujetos ni derivados indirectamente. En la bibliografía se menciona que el dataset es un muestreo estratificado de créditos reales, con 1000 créditos (300 malos y 700 buenos).La validación o verificación específica del conjunto no está detallada, aunque ha sido ampliamente utilizado y revisado.

Preprocesamiento/ limpieza / etiquetado

El dataset original tiene variables tanto numéricas como categóricas, y viene acompañado de un archivo (german.doc) que explica el significado de cada variable y sus categorías.

En general, se hizo lo siguiente:

- Se eliminaron algunas filas con datos faltantes para mantener la calidad.
- Se agruparon algunas categorías poco frecuentes para simplificar el análisis.
- Se renombraron algunas variables para que sean más claras.
- Se definieron algunas variables como ordinales (es decir, con un orden natural).
- Se creó una versión numérica del dataset para facilitar el trabajo con modelos computacionales.

No se hizo una limpieza profunda ni se eliminaron muchas filas, sino que principalmente se hizo codificación y organización para que los datos sean más fáciles de usar.

Usos

Este conjunto de datos es uno de los más usados para tareas de clasificación en machine learning, especialmente para modelos de riesgo crediticio y fairness en decisiones financieras. Existen múltiples repositorios en GitHub, así como competiciones y kernels en Kaggle que usan este dataset para análisis, modelado y evaluación de fairness.

Referencias

Comparative Testing and Evaluation of Statistical and Logical Learning Algorithms for Large-Scale Applications in Classification, Prediction and Control

Statlog (German Credit Data)

South German Credit Data: Correcting a Widely Used Data Set

Github German Credit Data

Kaggle German Credit Data

Ejemplo uso en kaggle de los datos

Análisis exploratorio del conjunto de datos

Como ya mencionamos el conjunto de datos cuenta con 21 columnas y 1000 instancias. A continuación mostramos el nombre de las columnas y sus tipos.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   estado_cuenta_corriente  1000 non-null   object
 1   duracion_meses           1000 non-null   int64 
 2   historial_crediticio     1000 non-null   object
 3   proposito                1000 non-null   object
 4   monto_credito            1000 non-null   int64 
 5   cuenta_ahorro_bonos      1000 non-null   object
 6   años_empleo              1000 non-null   object
 7   porc_ingreso_disponible  1000 non-null   int64 
 8   estado_civil_sexo        1000 non-null   object
 9   otros_deudores_garantes  1000 non-null   object
 10  residencia_actual_anios  1000 non-null   int64 
 11  propiedad                1000 non-null   object
 12  edad                     1000 non-null   int64 
 13  otros_creditos           1000 non-null   object
 14  vivienda                 1000 non-null   object
 15  num_creditos_banco       1000 non-null   int64 
 16  ocupacion                1000 non-null   object
 17  num_personas_a_cargo     1000 non-null   int64 
 18  telefono                 1000 non-null   object
 19  trabajador_extranjero    1000 non-null   object
 20  riesgo_crediticio        1000 non-null   int64 
dtypes: int64(8), object(13)
memory usage: 164.2+ KB
```

A su vez es pertinente mencionar el significado de las claves internas, que corresponde a los valores únicos definidos en la documentación oficial, pero solo de variables categóricas o binarias. También, modificamos / traducimos algunos términos ya que va a usarse para gráficos.

estado_cuenta_corriente

- A11: negativo
- A12: 0-200 DM
- A13: + 200 DM / salario de 1 año
- A14: no tiene

historial_crediticio:

- A30 : no credits taken/ all credits paid back duly
- A31: all credits at this bank paid back duly
- A32: existing credits paid back duly till now
- A33: delay in paying off in the past
- A34: critical account/ other credits existing (not at this bank)

propósito:

- A40: car (new)
- A41: car (used)
- A42: furniture/equipment
- A43: radio/television
- A44: domestic appliances
- A45: repairs
- A46: education
- A47: (vacation - does not exist?)
- A48: retraining
- A49: business
- A410: others

cuenta_ahorro_bonos:

- A61: 0 - 100 DM
- A62: 100 - 500 DM
- A63: 500 - 1000 DM,
- A64: +1000 DM
- A65: desconocido/ no tiene

años_empleo :

- A71: unemployed
- A72: < 1 year
- A73: 1 - 3 years
- A74: 4 - 6 years
- A75: >= 7 years

estado_civil_sexo:

- A91: male : divorced/separated
- A92: female : divorced/separated/married
- A93: male : single
- A94: male : married/widowed
- A95 : female : single

otros_deudores_garantes:

- A101: none
- A102: co-applicant
- A103: guarantor

propiedad

- A121: real estate
- A122: building society savings agreement/ life insurance
- A123: car or other
- A124: unknown / no property

otros_creditos:

- A141: bank
- A142 stores
- A143: none

vivienda:

- A151 : rent
- A152: own
- A153: for free

ocupación:

- A171: unemployed/ unskilled - non-resident
- A172 : unskilled - resident
- A173: skilled employee / official
- A174: management/ self-employed/ highly qualified employee/ officer

teléfono

- A19: none
- A192 : yes, registered under the customer's name

trabajador_extranjero:

- A201: yes
- A202: no

En principio dividimos las variables en cuatro grupos:

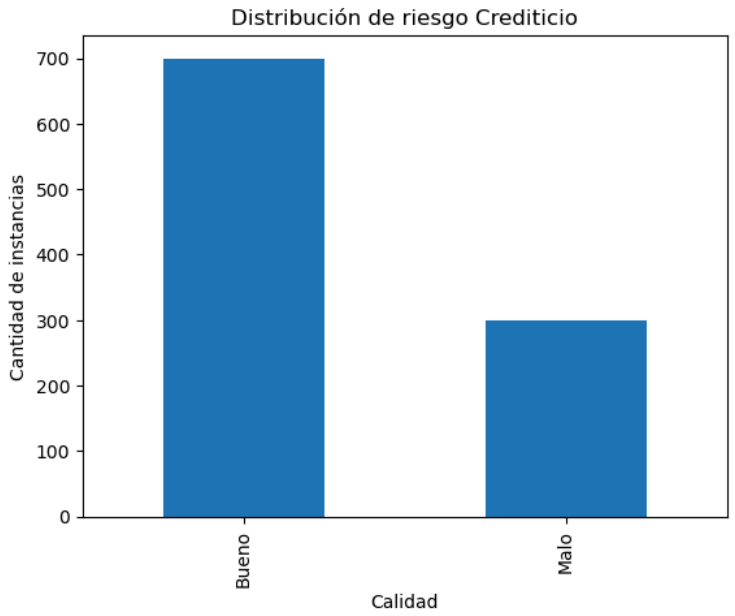
- **Target:** "riesgo_crediticio" la variable que nos va a interesar predecir
- **Relacionados a historial crediticio / situacion economica:** 'estado_cuenta_corriente', 'historial_crediticio', 'cuenta_ahorro_bonos', 'porc_ingreso_disponible', 'num_creditos_banco'

- **Relacionados al crédito que se está solicitando:** 'duracion_meses', 'proposito', 'otros_deudores_garantes', 'monto_credito'
- **Relacionados al cliente:** 'años_empleo', 'estado_civil_sexo', 'vivienda', 'residencia_actual_anios', 'propiedad', 'edad', 'ocupacion', 'num_personas_a_cargo', 'telefono', 'trabajador_extranjero'

Decidimos analizarlo de esta forma porque los grupos refieren al tipo de atributos que representan sobre la instancia. Los relacionados al cliente son los atributos personales, o proxys de ellos, y por ende los atributos sensibles que pueden originar un sesgo. Por ejemplo, si bien los años de empleo o la propiedad no son sensibles y tiene sentido que sean evaluados en un análisis económico, una combinación de todo esto puede significar clase social, y por ende, podría buscarse evitar que se le niegue un préstamo a alguien, en base a esto.

Por otro lado, consideramos que tiene sentido que el banco evalúe el riesgo crediticio basándose en el comportamiento o la historia que tenga el cliente en una institución financiera. A nuestro parecer, más allá de un posible sesgo histórico o institucional que pueda haberlos influido, no debería haber sesgos en esta parte de los datos. De igual manera, vamos a revisar estas columnas para entenderlas mejor.

En primer lugar, revisamos la columna target 1 significa buen riesgo crediticio y 2 significa un mal riesgo crediticio, queremos tener una idea de que tan balanceado está el dataset.

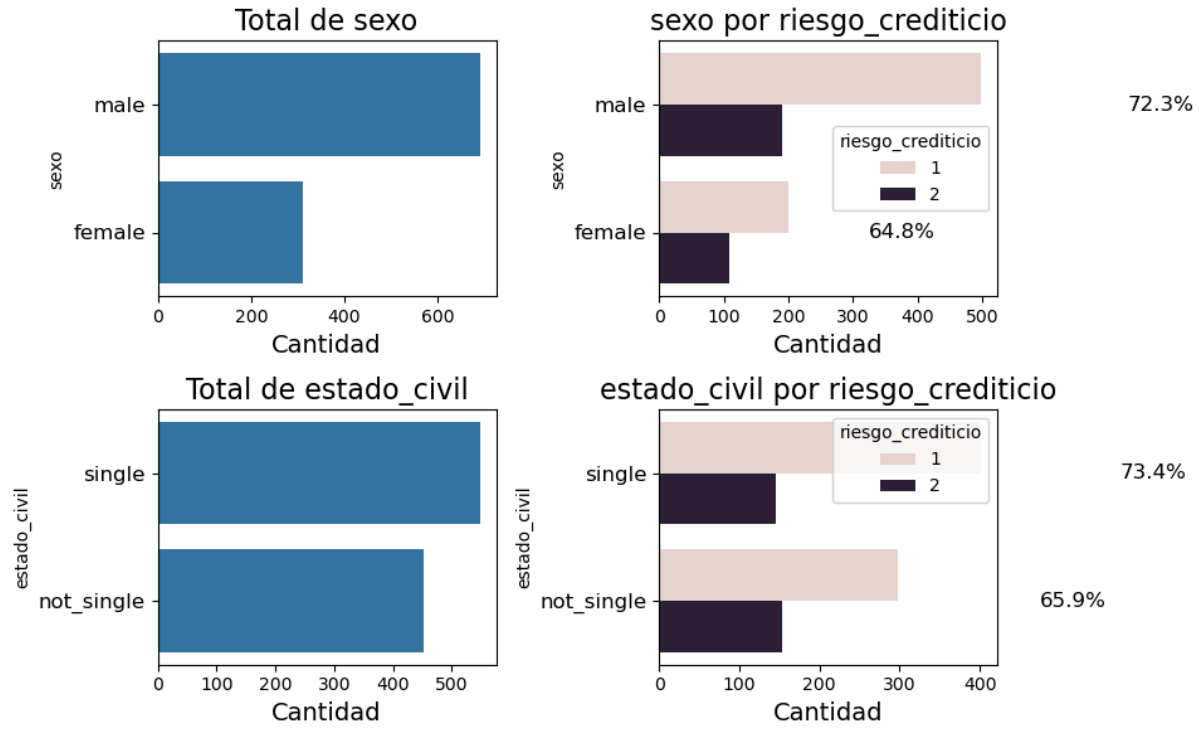


Vemos que hay un desbalance, donde la gran mayoría de las muestras son positivas. Por esto, nosotros buscaremos en el análisis exploratorio, ver que esta diferencia se mantenga, al menos de manera relativa, dentro de la mayoría de las

categorías de cada atributo. De lo contrario, podríamos analizar si es posible que haya un sesgo en ese caso.

Como segundo análisis vemos las columnas referidas al cliente y sus características. El objetivo es poder realizar la limpieza necesaria, comprender un poco más la distribución y analizar posibles sesgos. Creemos que esta manera va a permitirnos comprender mejor los datos con los que estamos trabajando. Para eso comenzamos analizando si tratar o no dos de las variables: 'estado_civil_sexo' y 'ocupacion'.

Para estado_civil_sexo, la columna presenta dos tipos de sesgo. En primer lugar tenemos posibles sesgos historicos relacionados con el sexo o el estado civil, que perpetuan direfencias en las oportunidades que tienen ciertos grupos sociales. Por otro lado, también hay un problema con la codificación de datos. Mas allá de que mezclar sexo y estado civil no es una buena practiva, para el sexo femenino tenemos dos categorias de estado civil, soltera y casada/divorcida/separada, mientras que para hombre tenemos tres, soltero, casado/viudo y divorciado/separado. Las categorias sexo-estado_civil deberían tener todas las combinaciones.



Para el modelado, tomamos la decicion de separarla en dos columnas, sexo (sobre todo porque genero crearia un sesgo de cohorte, ya que estariamos usando categorias por defecto) y estado civil. Esta última vamos a separarla en soltero / no soltero para que haya categorías comunes a ambos sexos y no sea proxy del mismo. Somos conscientes de que se puede llegar a estar introduciendo un sesgo de etiqueta por una interpretación errónea del significado de las columnas, pero

por nuestra investigación previa creemos que esto no es el caso. Por otro lado, estamos introduciendo un sesgo de codificación de datos al manipular las categorías, pero nos parece que vale la pena y es necesario para poder lograr que sean independientes.

Para ocupacion, vemos que unskilled está separado por residente y no residente. Idealmente, podríamos juntar estas variables en un mismo grupo, creando una variable separada para el estado de residencia. El problema, es que para el resto de las categorías no tenemos esta información. En un comienzo teníamos la teoría de que la residencia podría estar relacionada con la variable trabajador extranjero, y por ende esta información estaría en el modelo en otra columna y podríamos juntar los grupos con 'unskilled', pero esto no fue así. Como sabemos que el estado de residencia puede ser una fuente de sesgo, además de que puede ser un proxy de extranjero, no queremos eliminar esta información del modelo, ya que podríamos estar introduciendo un sesgo por codificación de datos. Para evitar esto, decidimos dejarla tal cual, y existe la posibilidad de un análisis posterior de si efectivamente el estado de residencia está sesgando el nivel de riesgo crediticio.

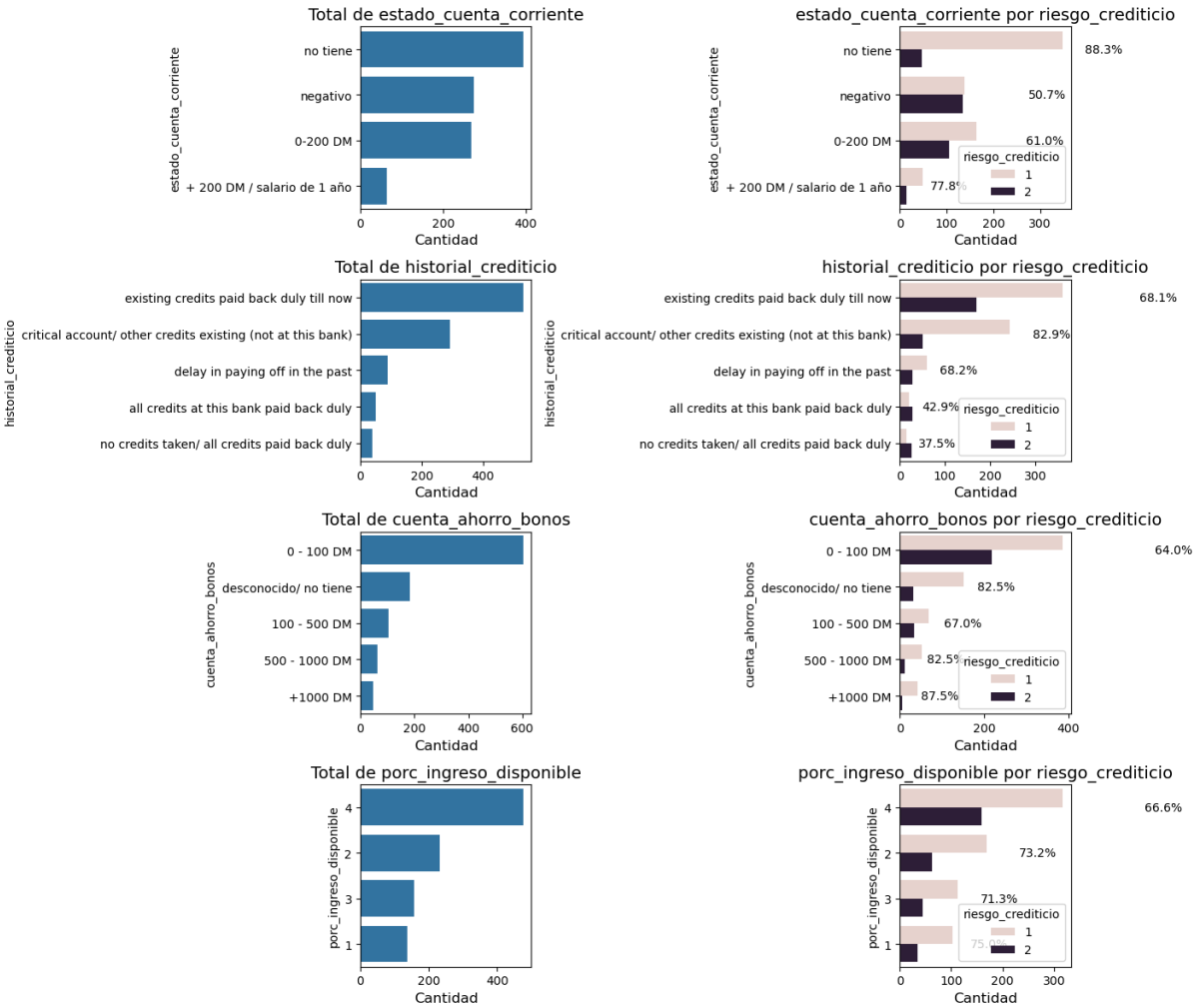
Por otro lado, la columna trabajador_extranjero podría indicar que hay o puede existir un sesgo por discriminaciones sistemáticas. El racismo / xenofobia podría verse integrado en las prácticas sociales, llevando a que una persona extranjera no tenga la posibilidad de acceder a un préstamo. También, por prácticas históricas, su historial crediticio podría verse afectado.

Una vez concluido el análisis de las variables categóricas procedimos a analizar las variables numéricas y encontramos dos posibles tipos de sesgos:

- Propiedad / vivienda / num_personas_a_cargo / años_empleo: Proxy de clase social. Estas variables podrían llevar a que el modelo infiera información sobre la situación socioeconómica de la persona, perpetuando el sesgo histórico hacia la clase alta. Si bien tiene sentido que sea parte del análisis, es bueno analizar si el modelo no genera daños de asignación, quitando oportunidades a personas en base a su condición socioeconómica.
- Edad: El modelo podría estar generando diferencias a partir de la edad de una persona. Si bien una mayor edad podría darle características deseables como historia crediticia y altos cargos o propiedades, existen casos donde se niegan préstamos a personas en base a su edad.

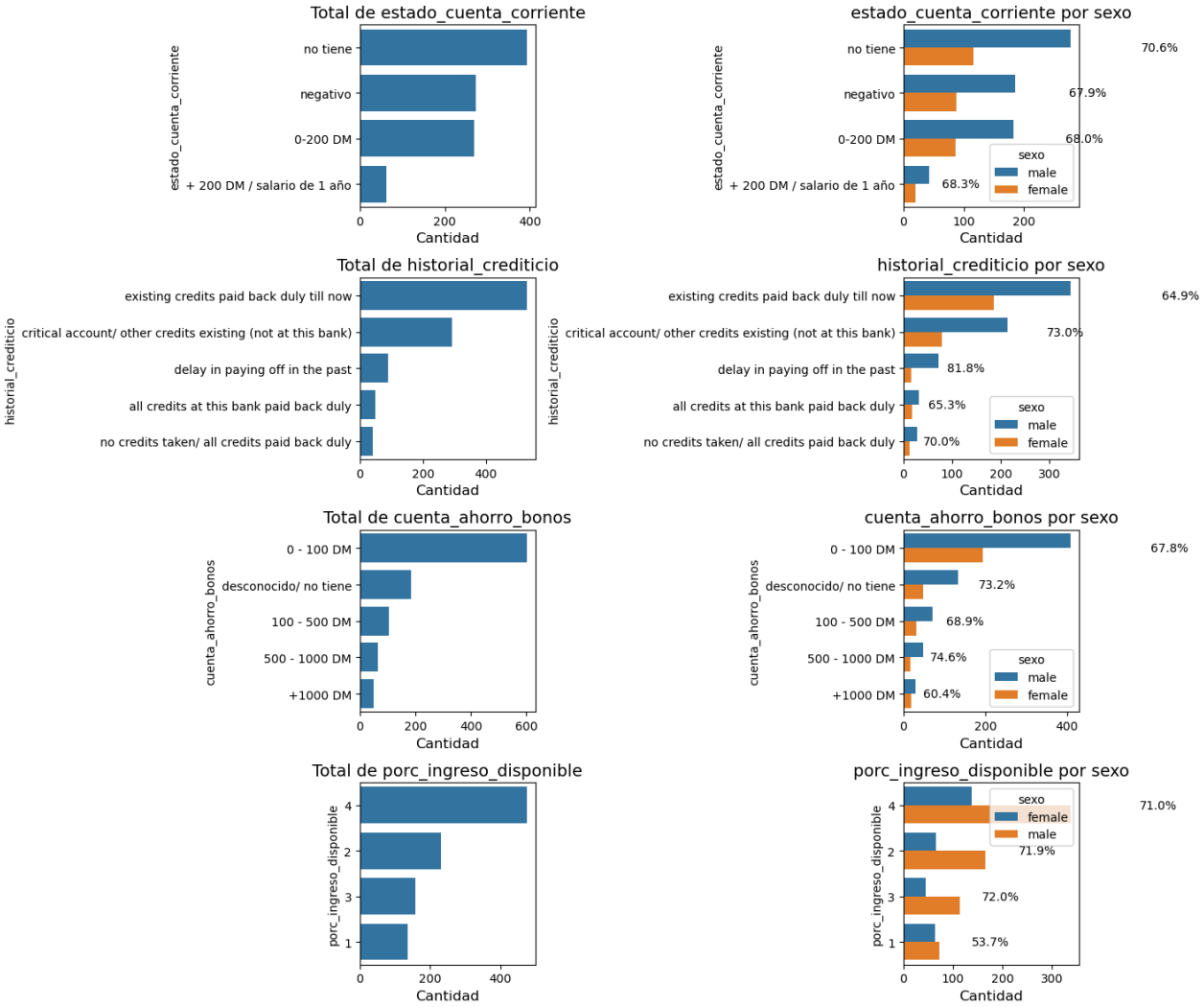
En relación a la edad se nos había ocurrido que quizá podíamos separar a las edades en distintas categorías, pero luego nos pareció que estaríamos introduciendo un sesgo innecesario, además de que los modelos de clasificación trabajan mejor con variables numéricas, por lo que al hacerla categórica, nos estábamos perjudicando sin razón.

Siguiendo con el análisis analizamos los atributos relacionados al historial crediticio con el objetivo de cerciorarnos, o al menos no encontrar de manera burda, algo que indique que estas columnas son muy dispares para un atributo sensible, como sexo.



Lo que podemos ver a primera vista, es que no parece haber una relación entre estas variables y el riesgo crediticio. Por ejemplo, cuanto más tiene en la cuenta de ahorros, más probable debería ser que le den un crédito, pero, si bien en algunas categorías es así, hay mas % de instancias positivas entre los que no tienen cuentas de ahorro, que entre los que tienen poco dinero en la misma. Dado que aunque nuestro rol es pensar como el banco pero somos un banco, no podemos saber cuál de los dos casos es más riesgoso y cual es la justificación, pero notamos que es algo que sucede en las otras categorías también.

Por ende, veamos tambien como se ve esto si lo separamos por sexo, el atributo sensible por excelencia.



Luego de ver los gráficos consideramos que no hay una relación directa que se pueda observar a simple vista, porque la mayoría de las instancias son masculinas y por ende tiene sentido que sean un porcentaje mayor de cada categoría. Es simplemente algo a notar dado que no es suficiente para probar una relación. Habrá que investigar si efectivamente hay un sesgo en el modelo a diseñar.

Más allá de eso, podría existir un sesgo institucional o social por el que las mujeres no acceden o piden préstamos, lo que afecta su historia crediticia. Habría que ver cómo se mide este punto en el modelo con su historia completa.

De este primer análisis podemos resaltar que las variables historial_crediticio y/o num_creditos_banco podrían estar integrando un sesgo institucional, ya que el atributo representa una interacción previa con el sistema financiero. Si una persona nunca accedió a un crédito, no puede tener un historial positivo, pero esta falta de acceso puede deberse a una discriminacion previa. Además, aunque tuviera un préstamo previo, su comportamiento ante el pago del mismo también puede estar sesgado. Dependiendo de si tuvo mejores condiciones a partir de algún sesgo, puede haberle resultado más fácil o difícil cumplir lo acordado, por lo

que puede haber un sesgo de decisiones históricas de las instituciones que afecta o modifica su riesgo real.

Esta es una observación más teórica y de análisis del tipo de atributo, pero que también puede generar la disparidad en las muestras. Aun teniendo en cuenta esta posibilidad, elegimos "dar por perdida" la batalla contra lo histórico, y trabajar en base a que esto es fiel representación del accionar crediticio de una persona/instancia.
Para concluir este primer análisis exploratorio revisamos las columnas relacionadas al préstamo a solicitar con la idea de entender o ver si había algún tipo de préstamo más relacionado a un buen crédito o mal crédito, o a alguna variable sensible y llegamos a la conclusión de que más allá de las diferencias que tengan estas categorías en los targets (que no son muchas) nosotros consideramos que el crédito que solicitan debería influir un modelo que predice el riesgo crediticio. Si el banco históricamente aprueba más de un tipo de crédito que otro, esto podría afectar. Por eso tomamos la decisión de sacar estas columnas del dataset, porque las consideramos leaks.

Nuevamente, estamos incorporando un sesgo de codificación de datos y mas allá de las columnas, hay otros tipos de sesgos que creemos pueden estar involucrados:

- Sesgo de representación: El conjunto representa una muestra poblacional de personas que solicitaron un crédito en una institución particular. Por ende, la población objetivo puede no estar debidamente representada, impidiendo que el modelo pueda generalizarse. La diferencia en la cantidad de instancias entre hombres y mujeres sugiere que los datos no son representativos de la población general. Al encontrarnos con menos instancias de mujeres, es probable que un modelo entrenado con estos datos presente un comportamiento injusto para las mujeres.

- Sesgo cronológico: Los datos son de 1994, por ende, no necesariamente se adecuan a las características poblacionales y a los criterios actuales de quien es o no un buen acreedor.

- Sesgo de selección: Es posible que el momento histórico y el contexto donde se generaron los datos, haga que cierto conjunto de atributos esté más presente que otro, interfiriendo en los resultados.

- Sesgo de confusión: En el dataset hay varios grupos de variables que podrían considerarse proxies de atributos sensibles, por lo que no solo hay relaciones directas, sino que estamos usando varias variables proxy que aluden al mismo atributo. Como hay mucha influencia, pueden distorsionar los resultados.

- Sesgo poblacional: Los datos del modelo no necesariamente representan la población objetiva, por ende,si bien el modelo puede tener buenos resultados en el entrenamiento y el testing, no necesariamente va a modelar correctamente la población.

El tamaño de la muestra también es algo a considerar, ya que más allá de las distribuciones desiguales de la cantidad de instancias por grupos de edades y por género, consideramos que hay pocos datos.

Armado del primer modelo

En primer lugar, transformamos el dataset utilizando one-hot encoding para que el clasificador pueda procesarlo correctamente. En un principio, consideramos la posibilidad de interpretar algunas variables como ordinales para reducir la cantidad de columnas generadas, pero finalmente decidimos no hacerlo. Aunque muchas categorías tienen un orden lógico y sus descripciones reflejan los límites entre ellas, el clasificador tiende a interpretar estas categorías ordinales como si unas fueran mejores que otras.

Como queríamos evitar introducir un sesgo de etiqueta artificial en el modelo, preferimos tratarlas como variables categóricas sin orden implícito. De esta manera, no se asume que una categoría es superior o inferior a otra. También guardamos la información separada por sexo, para después poder analizar si el modelo esta sesgado por ese atributo y cambiamos el valor de alto riesgo crediticio a 0 para que sea consistente con que el bajo riesgo que queremos predecir es la clase positiva identificada con un 1.

Realizamos dos modelos distintos con el fin de ver cual presentaba mejor performance. Como primer intento realizamos un Random Forest y los resultados fueron los siguientes:

```
----- Matriz de Confusión -----

               | Pred. Positiva  | Pred. Negativa  |
 Real Positiva  |       630       |       70        |
 Real Negativa  |       181       |       119       |

----- Métricas de Evaluación -----

Precisión: 0.78 - 0.02
Recall: 0.9 - 0.02
Accuracy: 0.76 - 0.02
F1-score: 0.84 - 0.01
True positive rate (TPR): 0.9 - 0.02
True negative rate (TNR): 0.42 -  0.06
Positive predictive value (PPV): 0.78 - 0.02
```

Luego armamos un segundo modelo, pero esta vez utilizamos un regresor logístico, para analizar cual nos da mejores métricas de entrada y así analizar con cuál nos quedamos como modelo final para analizar la equidad. Los resultados fueron los siguientes:

```
----- Matriz de Confusión -----

               | Pred. Positiva  | Pred. Negativa  |
 Real Positiva |       618       |       82        |
 Real Negativa |       171       |       129       |

----- Métricas de Evaluación -----

Precisión: 0.78 - 0.02
Recall: 0.88 - 0.04
Accuracy: 0.75 - 0.02
F1-score: 0.83 - 0.02
True positive rate (TPR): 0.88 - 0.04
True negative rate (TNR): 0.43 -  0.07
Positive predictive value (PPV): 0.78 - 0.02
```

Para definir qué modelo nos quedamos , nos quedamos con las métricas que consideramos que se adaptan mejor a la casuística particular. Dado que el banco que otorga los créditos tiene el objetivo institucional de maximizar la cantidad de personas que efectivamente pagarán el préstamo, el error más perjudicial sería el falso positivo (clasificar como buen riesgo crediticio a alguien que no pagaría). Esto se debe a que dar dinero que no vuelve, representa una pérdida económica directa para la institución.

Aunque los falsos negativos pueden implicar perder posibles buenos clientes, los falsos positivos suponen un riesgo financiero más importante. Si comparamos la potencial ganancia ante una perdida real creemos que una perdida directa es mas grave.

En este contexto consideramos que si lo más importante es controlar los falsos positivos, entonces una métrica muy importante para el modelo es el PPV(Positive Predictive Value), ya que un valor cercano a 1 indica que las predicciones positivas son mayormente correctas. En este caso, el valor para ambos modelos es muy similar. Como esto sucede en la mayoría de las métricas, la decisión que tomamos se basa en la más distinta, el True Negative Rate. Esta métrica nos indica, de las muestras negativas, qué proporción se predice correctamente. Siendo que es más alta en el modelo del regresor lineal, ese es el que nos quedaremos para continuar. Somos conscientes de que puede no ser el mejor modelo, sobre todo porque la predicción sobre la clase negativa no es buena, pero reconocemos que no es el objetivo del trabajo. Queremos quedarnos con el modelo, y poder ver que si es malo, que sea igual de malo para todos los grupos.

En el contexto de la asignación de préstamos, después de ver los datos, consideramos como grupos de interés para buscar la equidad a las mujeres y los hombres, aunque también se podría considerar un grupo cada franja de edades. Siendo estos los grupos que nos interesan, en este contexto interpretamos los criterios de fairness de la siguiente manera:

- **Statistical Parity**

Este criterio se cumple si para todos los grupos la tasa de predicciones positivas es la misma. En este contexto, eso significa que tanto para hombres como para mujeres se debería tener la misma probabilidad de conseguir que el modelo los clasifique como un buen crédito y en consecuencia el banco les de un crédito.

- **Equalized Odds**

En el caso de este criterio se busca que la probabilidad de que se apruebe el crédito dado que era un buen crédito y la probabilidad de que se apruebe el crédito siendo que en la realidad era un mal crédito sean las mismas entre los grupos. Esto se traduce en que se busca que para hombres y mujeres exista la misma probabilidad de conseguir un préstamo dado que realmente es un buen crédito, y la misma probabilidad de obtenerlo dado que no es un buen crédito. Se espera que para las asignaciones de préstamos el modelo acierte y se equivoque con la misma tasa entre grupos.

- **Equal Opportunity**

Para este contexto implica que para hombres y mujeres exista la misma probabilidad de conseguir un préstamo dado que realmente es un buen crédito. Solo mira los True Positive Rate.

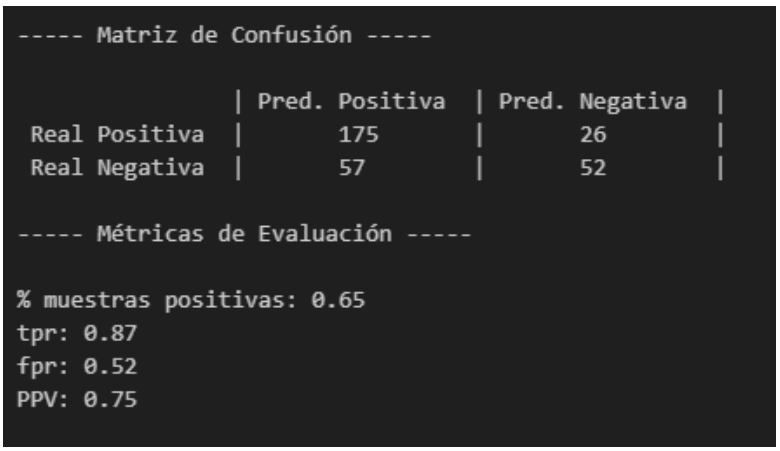
- **Predictive Parity**

Este criterio se cumple cuando entre los grupos de interés la probabilidad de que en la realidad sea positivo dado que la predicción es positiva es la misma o similar. En el contexto de los préstamos se traduce buscando que si el modelo aprueba un crédito, la probabilidad de que realmente sea un buen crédito sea la misma tanto para mujeres como para hombres, es decir que la tasa de préstamos aprobados correctamente sea similar.

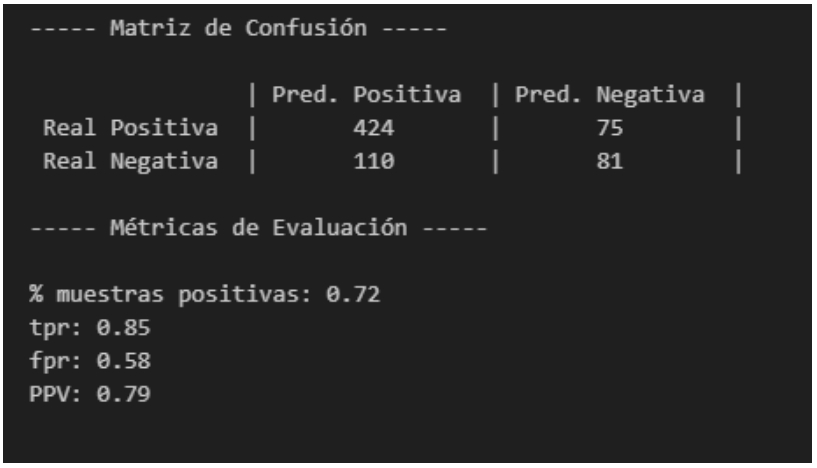
Analisis de fairness para los distintos grupos

Para ambos sexos armamos una matriz de confusión con el modelo seleccionado ya entrenado, y analizamos las distintas metricas y como estas se traducen a un criterio de fairness.

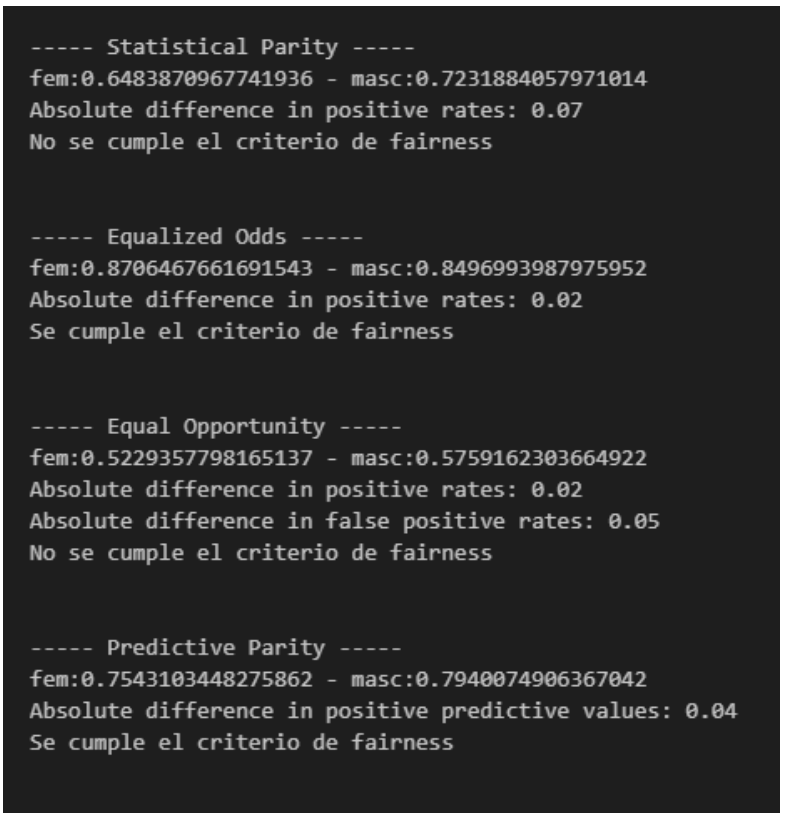
Para el sexo femenino:



Mientras que para el sexo masculino:



Decidimos que el criterio a utilizar para determinar si es fair o no, es que el módulo de la diferencia en las métricas sea menor a 0.05. La razón de elegir este valor es que era uno que permitía que no se cumpla algún criterio de equidad. Nuestra idea original había sido usar 0.15 como umbral, ya que es el valor que usamos a lo largo de la cursada y por ende entendemos que es un valor razonable, pero como la intención de la materia es mejorar estas inequidades, decidimos ponernos más estrictos.



Este resultado representó buenas noticias para nuestra idea inicial de que como banco, queremos que sea importante el Positive Predictive Value. Esta métrica se relaciona directamente con Predictive parity, ya que representa la cantidad de predicciones positivas que fueron correctas. Siendo que el modelo es fair bajo este criterio, podríamos decir que nuestro objetivo principal se cumple desde el modelo inicial.

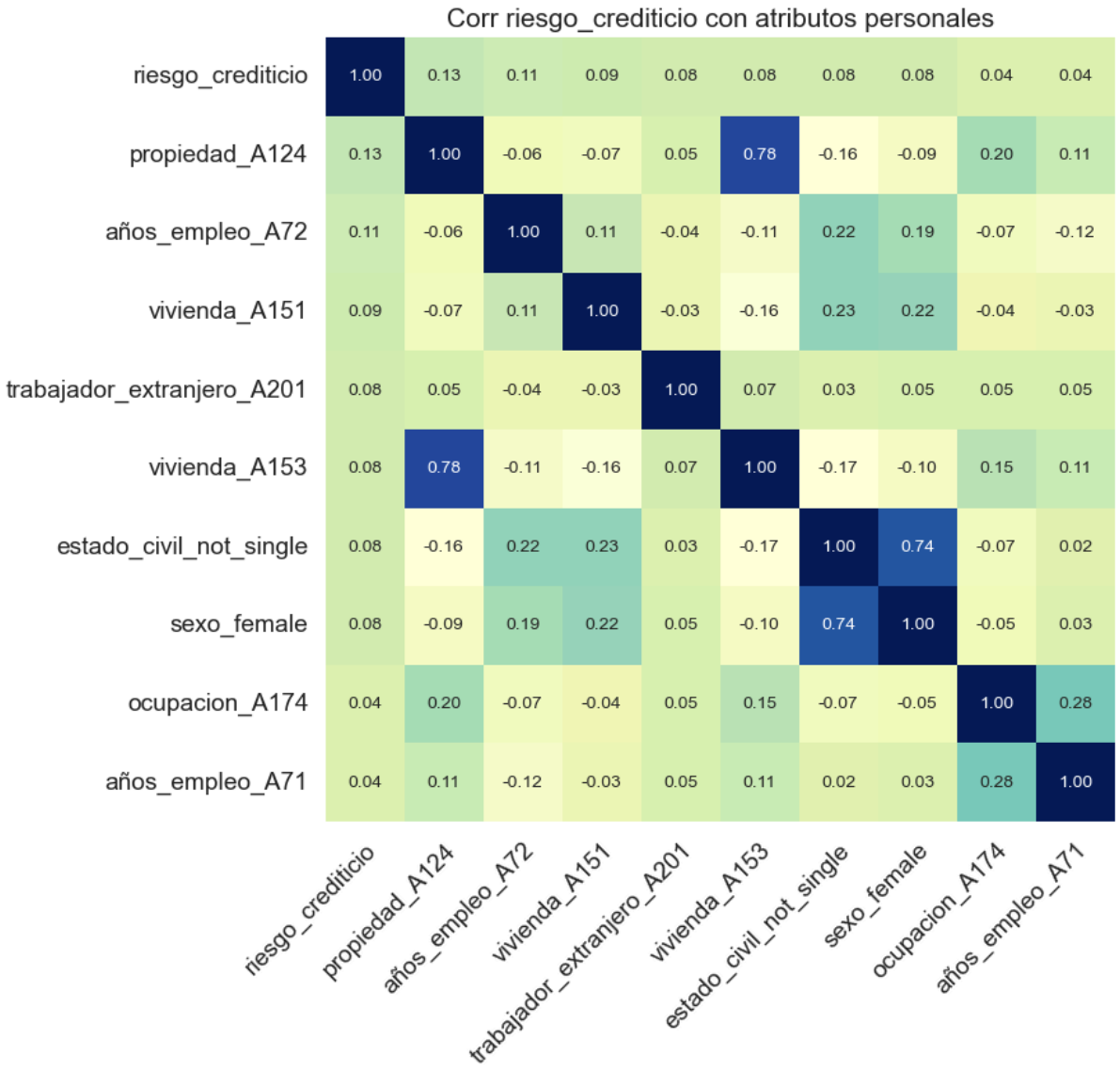
Por otro lado, siendo que nosotros notamos que nuestro modelo erraba mucho con la clase negativa, y teníamos la esperanza de que fuera igual de malo para todos, podemos ver que no pareciera ser así. La métrica relacionada con la clasificación de la clase negativa es el False Positive Rate, que indica la proporción de casos negativos que fueron clarificados de forma incorrecta. Siendo que es alto, buscábamos que fuera igual de alto para ambos grupos. En este caso, por el criterio equal opportunity, siendo este el que se relaciona con esta metrica, vemos que no se cumple con el umbral previamente definido. Por suerte, se cumple equalized odds, y en consecuencia, para la clase positiva, funciona lo suficientemente igual de bien para ambos.

Mitigación de sesgos

En este apartado exploraremos técnicas para mejorar la equidad del modelo creado. A lo largo del informe se presentaron diversas posibles causas de sesgos por lo que en primer lugar evaluaremos con la librería holistic la performance del modelo que elegimos en relación a la equidad.

Como ya se vio en gráficos anteriores se puede intuir una disparidad en la representación en el conjunto de datos. Esto podría afectar, todavía no lo sabemos a la imparcialidad y la equidad en los modelos que entrenamos y en su precisión. También vemos más de cerca la relación entre los atributos y la variable objetivo, para ver si hay algún patrón o alguna tendencia presente en los datos.

Utilizando la separación de los atributos que mencionamos al principio (relacionados al crédito, personales y económicos) vimos más en detalle la matriz de correlación de los atributos y el target.



En relación a las correlaciones entre los atributos, vemos que en el caso de los que consideramos personales, los que tienen más relevancia son los asociados a estabilidad económica, como puede ser la vivienda y el tiempo que llevan trabajando. Igualmente se ve que de estos datos hay más correlación con quienes trabajan hace menos de un año, no tienen bienes propios o son desconocidos y

alquilan su vivienda. También hay correlación con el atributo de no soltero, en particular cuando se asocia al sexo de la instancia, puede deberse a que para la gente que esta casada o en pareja al dividirse los gastos puede aumentar la estabilidad economica.

Siguiendo esa idea, se ve que el sexo, tanto femenino como masculino tiene una correlación con el riesgo_crediticio. Aunque es un valor bajo siguen siendo atributos personales que nos gustaría que para el riesgo crediticio tengan una correlación lo más cercana a cero posible. Lo mismo sucede con el atributo de trabajador extranjero, que está asociado con la nacionalidad de la persona que solicita el crédito por lo que para evitar discriminaciones étnicas o raciales estaría bueno que también sea lo más cercano a cero posible.

Elegimos probar dos técnicas de mitigación que se utilizan en el preprocesamiento, reponderación y Correlation Remover. Decidimos que estas serían nuestras herramientas para intentar mitigar el sesgo dado que fueron las que más trabajamos antes de realizar el trabajo y vemos que los problemas que consideramos más fuertes en los datos, en relación al sesgo, son el desbalance entre instancias de los distintos grupos y la correlación entre los atributos que consideramos sensibles con la variable target. Con esto en mente realizamos pruebas para ver qué técnica mejoraba mejor el rendimiento del modelo en relación a la equidad y las evaluaciones que mostramos a continuación llegamos a una conclusión.

Metric	Baseline	Preprocessing Mitigator Reweighting	Preprocessing Mitigator Correlation Remover	Reference
Statistical Parity	0.194296	0.083333	0.046346	0
Disparate Impact	1.293603	1.111111	1.059463	1
Four Fifths Rule	0.773035	0.900000	0.943875	1
Cohen D	0.489944	0.211436	0.118341	0
2SD Rule	3.195706	1.409061	0.791439	0
Equality of Opportunity Difference	0.188859	0.091486	0.069746	0
False Positive Rate Difference	0.176768	0.030303	-0.042929	0
Average Odds Difference	0.182813	0.060894	0.013409	0
Accuracy Difference	0.102496	0.080660	0.088235	0

```
COMPARACION METRICAS
              precision    recall  f1-score   support

           0       0.48      0.34      0.40        58
           1       0.76      0.85      0.80       142

    accuracy                           0.70       200
   macro avg       0.62      0.59      0.60       200
weighted avg       0.68      0.70      0.68       200

              precision    recall  f1-score   support

           0       0.53      0.34      0.42        58
           1       0.77      0.87      0.82       142

    accuracy                           0.72       200
   macro avg       0.65      0.61      0.62       200
weighted avg       0.70      0.72      0.70       200

              precision    recall  f1-score   support

           0       0.51      0.34      0.41        58
           1       0.76      0.87      0.81       142

    accuracy                           0.71       200
   macro avg       0.64      0.61      0.61       200
weighted avg       0.69      0.71      0.70       200
```

```
COMPARACION ACCURACY
Accuracy sin mitigacion: 0.7
Accuracy con Correlation Remover: 0.72
Accuracy con reponderacion: 0.715
```

```
True Positive Rate:
Baseline : 0.845
Reponderado : 0.866
Correlation Remover : 0.873

False Positive Rate:
Baseline: 0.655
Reponderado: 0.655
Correlation Remover : 0.655

Positive Predictive Value
Baseline: 0.759
Reponderado: 0.764
Correlation Remover: 0.765
```

```
True Positive Rate:
Baseline female: 0.455 male: 0.278
Reponderado female: 0.364 male: 0.333
Correlation Remover female: 0.318 male: 0.361

False Positive Rate:
Baseline female: 0.283 male: 0.094
Reponderado female: 0.196 male: 0.104
Correlation Remover female: 0.174 male: 0.104

Positive Predictive Value
Baseline female: 0.435 male: 0.526
Reponderado female: 0.471 male: 0.545
Correlation Remover female: 0.467 male: 0.565
```

Implementamos dos técnicas de mitigación de sesgo, Reweighin o reponderacion y Correlation Remover, ambas de preprocesamiento. A pesar de ser técnicas de preprocesamiento las utilizamos al final con el objetivo de poder comparar los resultados después de haber trabajado sin intentar mitigar los sesgos que pensamos podrían estar influenciando el modelo. En primer lugar utilizamos reponderación porque vimos el desbalance entre clases y grupos sensibles. Luego utilizamos con fines comparativos Correlation Remover porque vimos que en las matrices de correlación las variables sensibles tenían cierta relación con la variable objetivo

Usando un modelo de regresión logística evaluamos el desempeño de las métricas clásicas y las métricas de fairness en el caso base (modelo sin mitigación) y los modelos mitigados. Si pensamos en los aciertos generales, podríamos buscar la versión con por ejemplo mejor accuracy, sin embargo en nuestro caso "como banco", como ya dijimos antes, nos es más costoso un falso positivo porque implica que se considere darle un crédito a alguien que no puede pagarlo. Considerando esto, queremos la versión del modelo que tiene mejor

Predictive Parity, que a su vez presenta menores diferencias entre las métricas entre clases y que cumpla mejor con los criterios de fairness relevantes para la casuística de estudio.

```
----- Predictive Parity -----
Baseline: fem:0.43478260869565216 - masc:0.5263157894736842
Reponderado: fem:0.47058823529411764 - masc:0.5454545454545454
Correlation Remover: fem:0.4666666666666667 - masc:0.5652173913043478
Diferencia absoluta en PPV para baseline: 0.09
Diferencia absoluta en PPV para reponderacion: 0.07
Diferencia absoluta en PPV para Correlation Remover: 0.1
No se cumple el criterio de fairness para Baseline

No se cumple el criterio de fairness para Reponderado

No se cumple el criterio de fairness para Correlation Remover
```

Es por eso que viendo los resultados consideramos que la mejor versión es la de reponderación, porque aunque en algunas cosas no es mejor que Correlation Remover, si es mejor que el baseline y es mejor según los criterios que pusimos. Además no nos arriesgamos a que Correlation Remover nos lleve a cero correlaciones que son importantes.

Conclusiones Generales

A pesar de que los modelos de aprendizaje automático en algún punto nacieron con la intención de limitar la subjetividad en la toma de decisiones y de “medir a todos con la misma vara”, después de muchos años de estudio, uso y varios casos de ejemplo podemos ver que el sesgo humano, que tanto se quería evitar, de alguna forma llega a los modelos a través de los datos y a través de quienes los entrenan.

Por esta razón, es importante que esta intención de justicia y objetividad no se pierda, y si bien hay herramientas para reducir los sesgos, es necesario que quienes desarrollen e implementen estos modelos sean conscientes y cuidadosos al momento de hacerlo. Un modelo no solo repite, sino que amplifica, por lo que el conocimiento de campo, y el buen criterio al momento de tomar decisiones, es la herramienta más certera con la que cuenta un investigador.