Big Data - 1S 2025

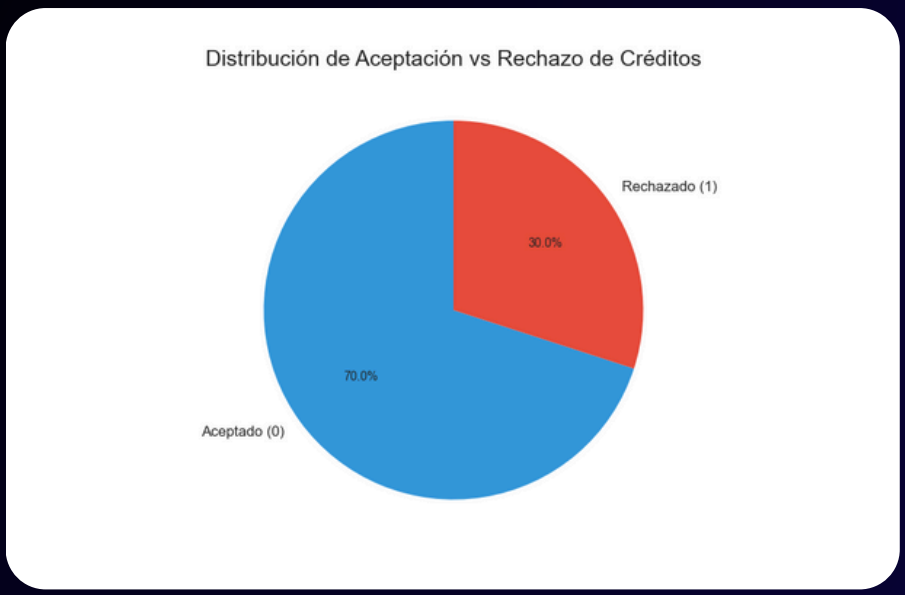
Trabajo Práctico 3

Credit Risk

Martín Quijano - Martina Coletto

Vamos a analizar el dataset de base clientes alemania. Nuestro objetivo es construir un modelo predictivo que reproduzca el criterio utilizado por el banco para aceptar o rechazar un credito.

Análisis exploratorio

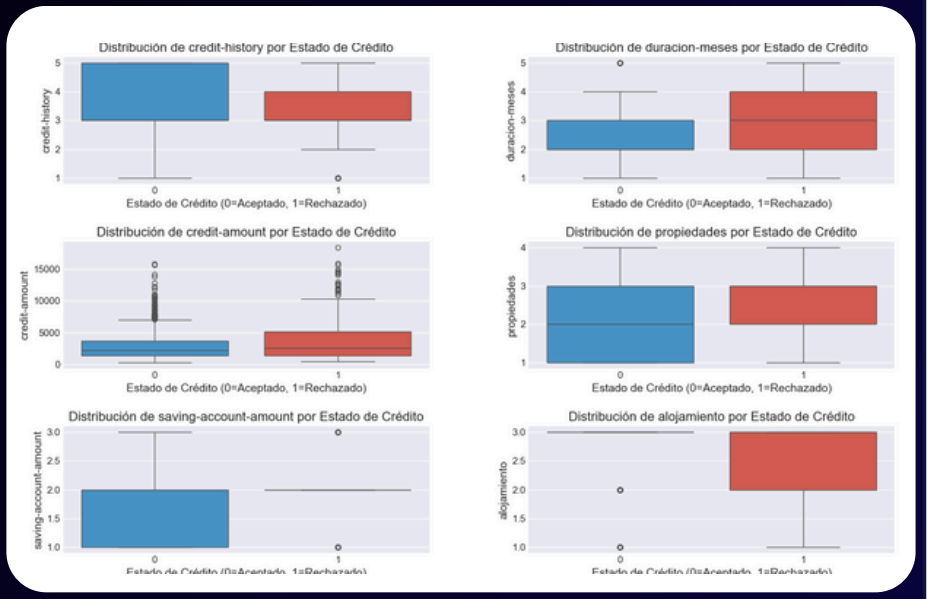


vemos que la muestra tiene un 70% de creditos que son aceptados, por lo que es una muestra desbalanceada.

Análisis exploratorio

Boxplot

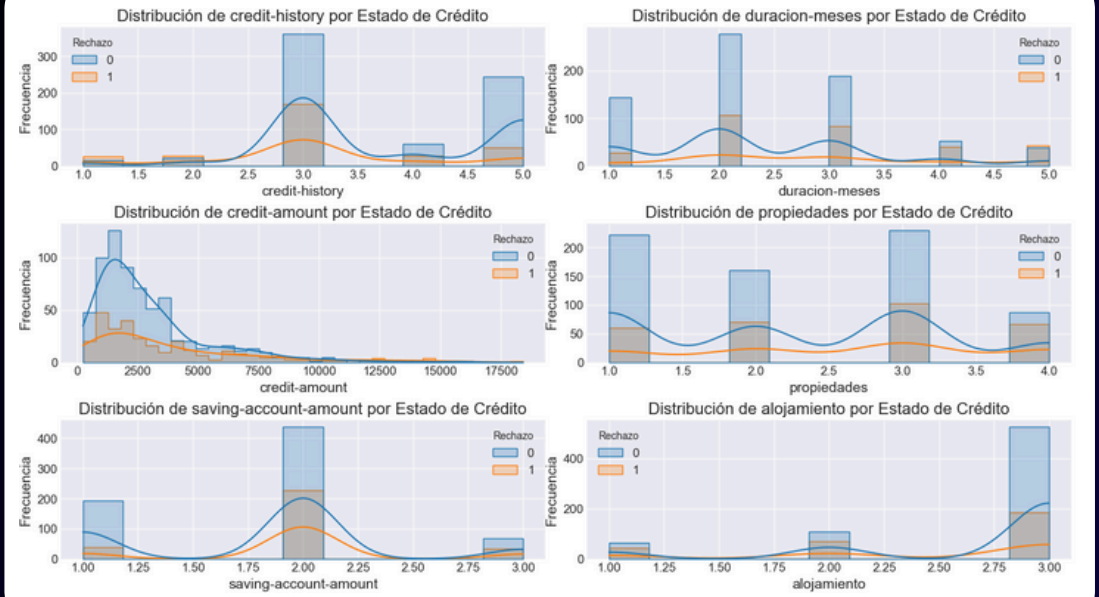
Para analizar la distribución de cada variable por churn.



usamos boxplots para analizar la distribución de cada variable por rechazo o no (1 es rechazo, 0 es aceptado). Para hacer esto, realizamos una serie de transformaciones tanto al nombre de las columnas como a los valores. Esto nos permitió identificar diferencias significativas en la dispersión de las variables entre creditos que se aceptan o no.

Vemos que en los creditos que se aceptan hay mas historial crediticio, que la duracion de los creditos es menor, que los montos son menores.

Análisis exploratorio



Aca quisimos hacer graficos exploratorios para ver la distribucion de algunas variables.

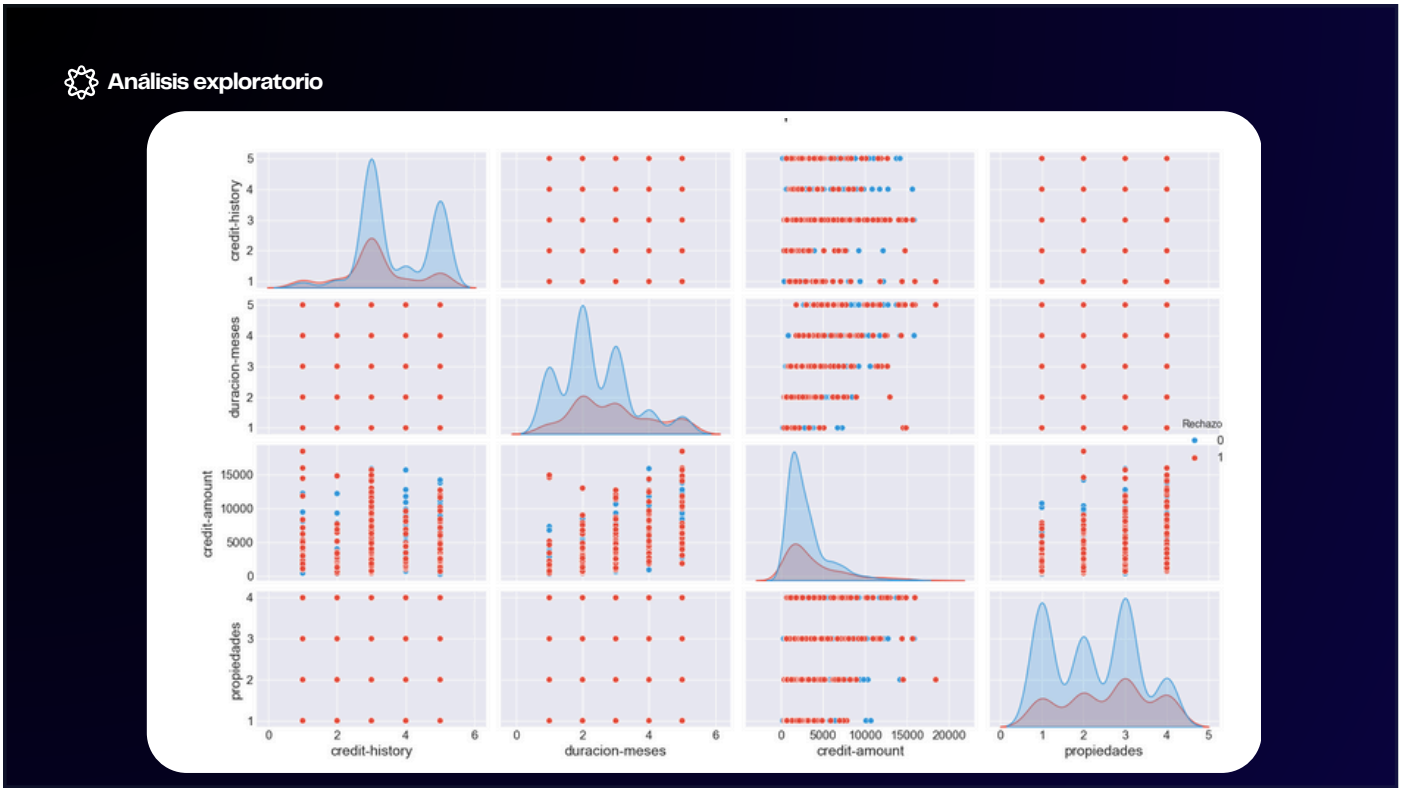
En credit-history vemos que la mayoria de clientes tenia o 3 o 5 creditos. Lo que sorprende es que a clientes con su quinto credito se le aceptan bastante sus creditos.

En cuanto a la duracion, vemos que a los creditos con duracion de 1, 2 y 3 meses tienen tazas de aceptacion bastante altas, mientras que los de 4 y 5 tienen bastante taza de rechazo.

En cuanto a credit-amount, vemos qu en su mayoria son de montos menores a 5000, y que la tasa de aceptacion baja significativamente a mayor monto de credito.

Lo que tambien sorprende es que la mayoria de clientes tiene 3 propiedades. Y en este grafico tambien vemos que

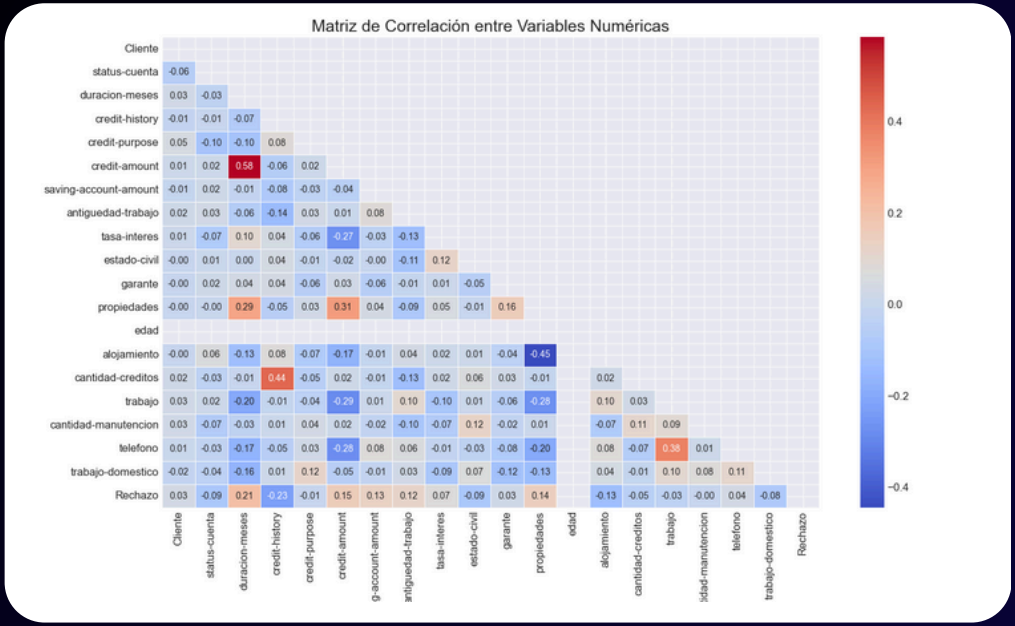
Tambien vemos que la mayoria de cuentas de ahorro se encuentran en el intervalo del medio, y teniendo tazas de aceptacion y rechazo parecida. Lo que nos sorprende es que a clientes del menor intervalo de dinero en su cuenta de ahorro tienen una taza de aceptacion bastante superior a la de rechazo.



Estos graficos nos sirven para ver frecuencias y relaciones de algunas variables.

Tambien pudimos ver como se relacionaban variables que ya sabiamos que se relacionaban. Por ejemplo a mayor duracion del credito, mayores son los montos permitidos. Y a mas propiedades mayor monto de credito.

Análisis exploratorio

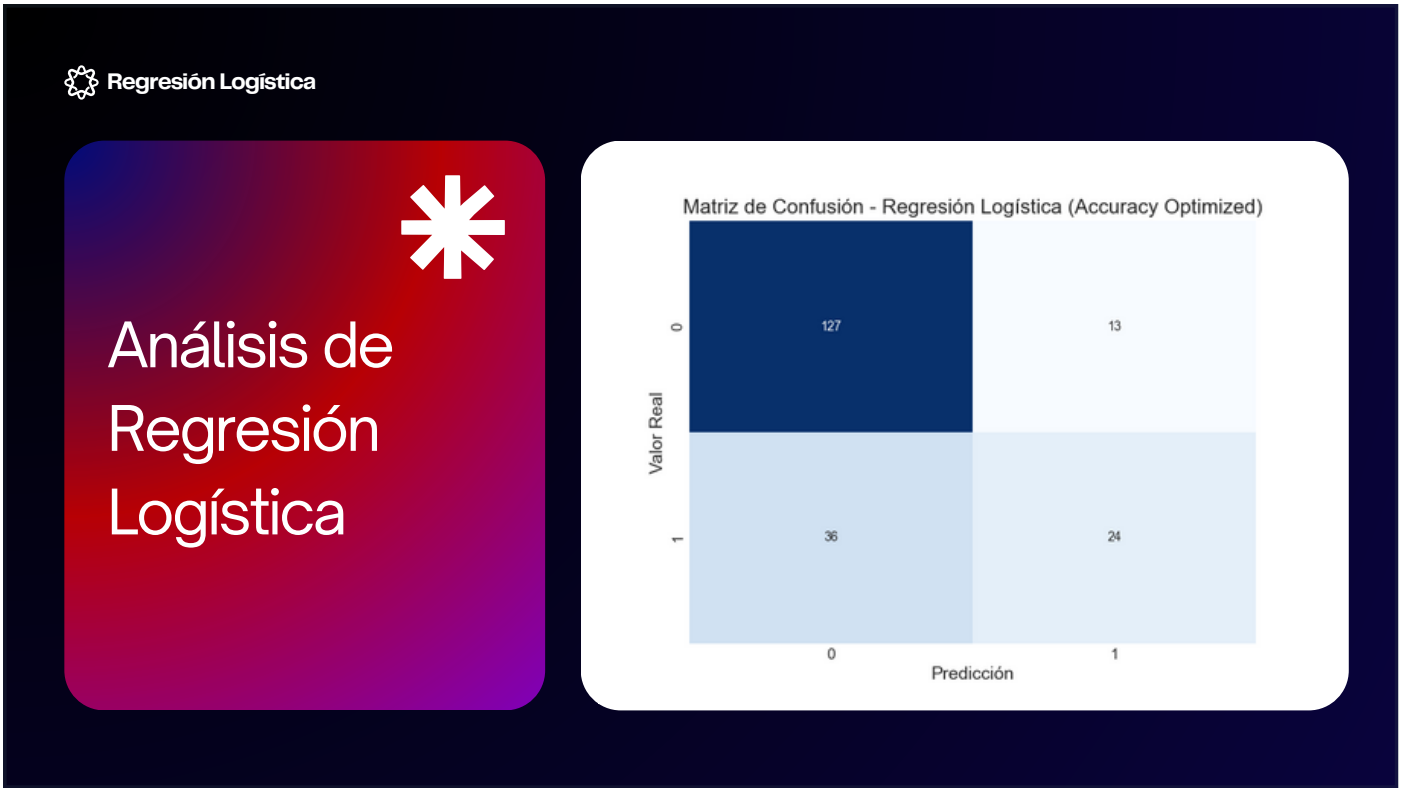


Con esta matriz de correlacion podemos ver que no hay variables con correlaciones tan fuertes.

Lo que si podemos destcar ess que la duracion del credito esta afectada positivente por el monto del credito, la cantidad de propiedades del que saca el credito, pero tambien vemos que a mayor duracion del credito mas probabilidad de rechazo.

Como es logico, vemos que la cantidad de creditos se relaciona con el historial crediticio.

Para el credit amount, vemos que la cantidad de propiedades vuelvle a ser un factor significativo, pepro vemos que la tasa de interes, el trabajo y si indican su telefono son variables que afectan negativamente a la cantidad de credito otorgado al cliente.

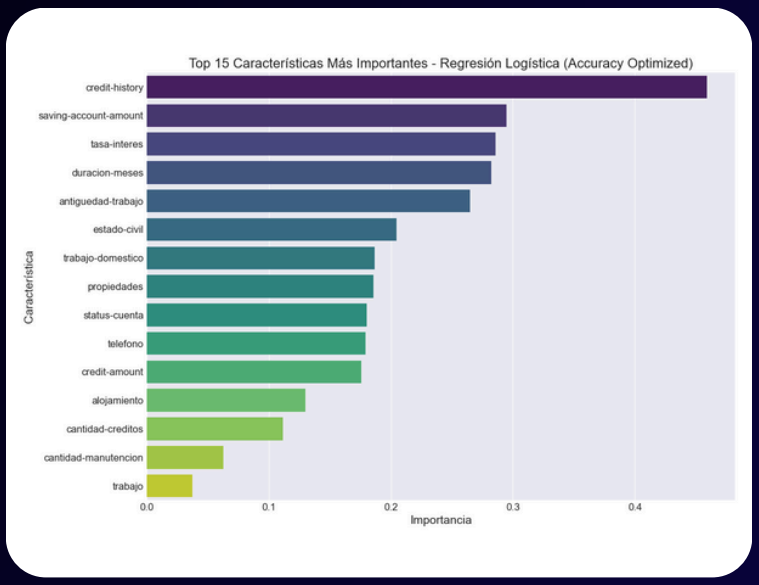


La accuracy fue de 0.755.

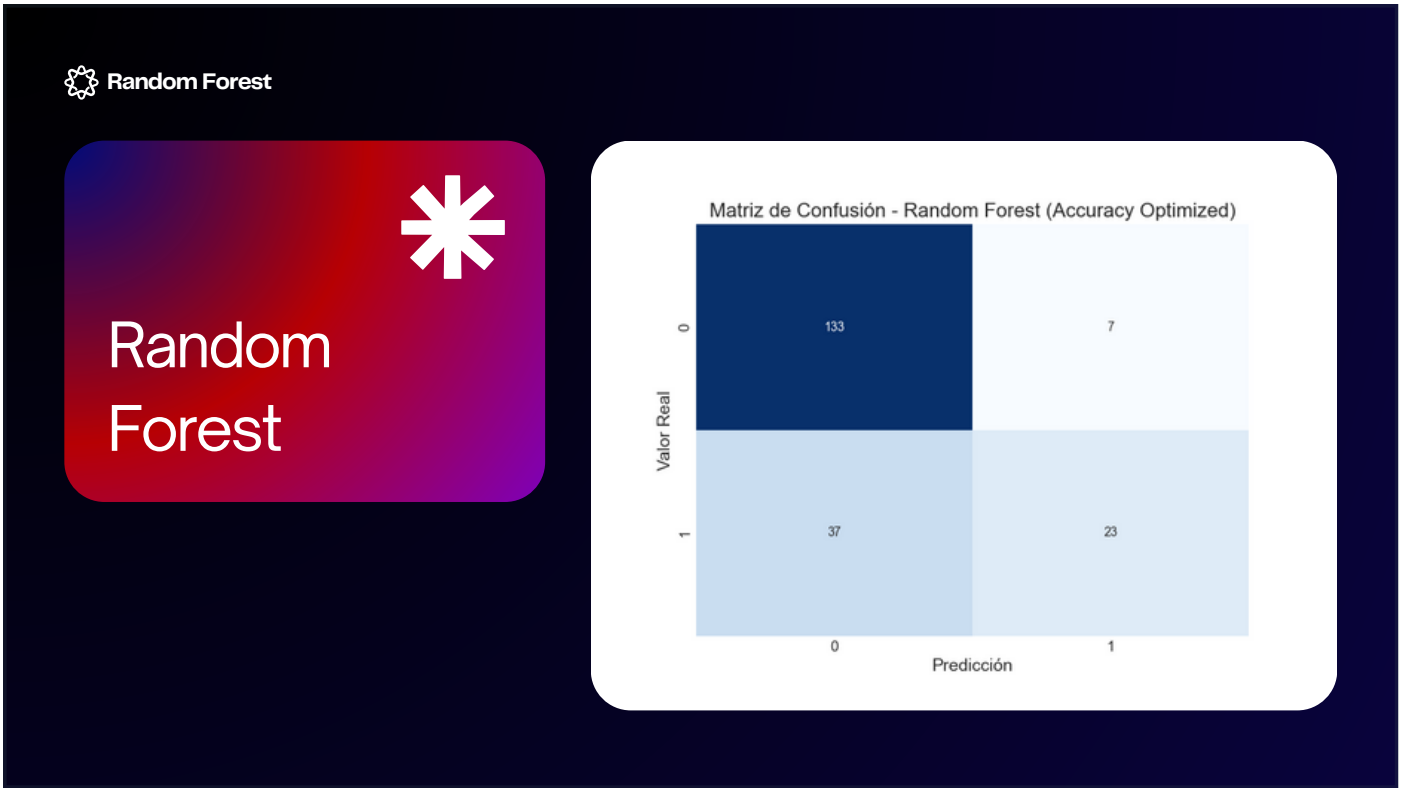
Vemos que por lo desbalanceada de la muestra la mayoria de los valores se encuentran en el cuadrante superior izquierdo. En este caso, la cantidad de errores no esta distribuida, y se encuentran mas errores en aceptacion predicha pero rechazo real. Esto implica que este modelo predice casos de aceptacion que en realidad serian rechazados.

Regresión Logística

Análisis de Regresión Logística

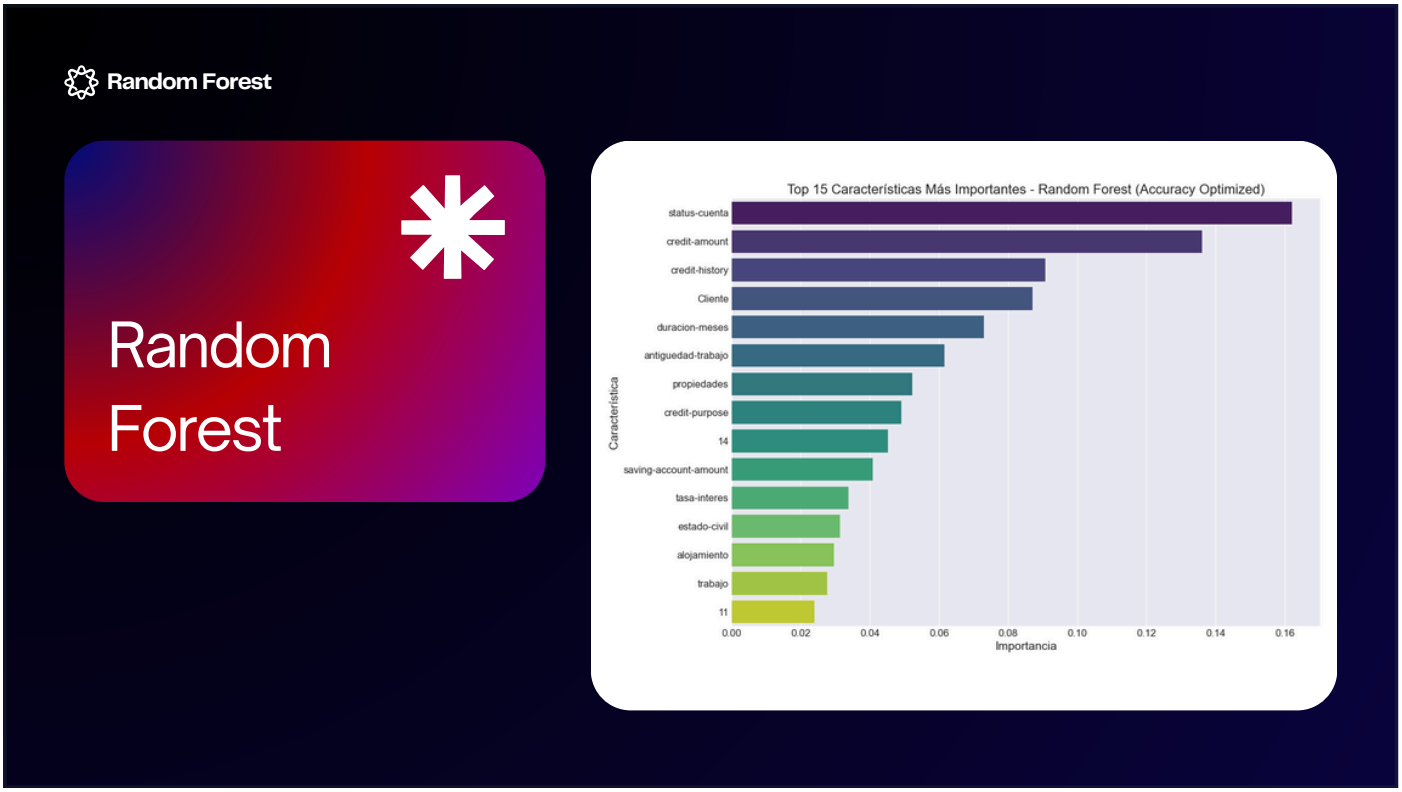


Vemos que las variables que mas influyen son credit-history, saving-account-amount, la tasa de interes, la duracion del credito y la antiguedad en el trabajo



La accuracy obtenida con este modelo fue del 0.780, ligeramente superior a la de regresion logistica

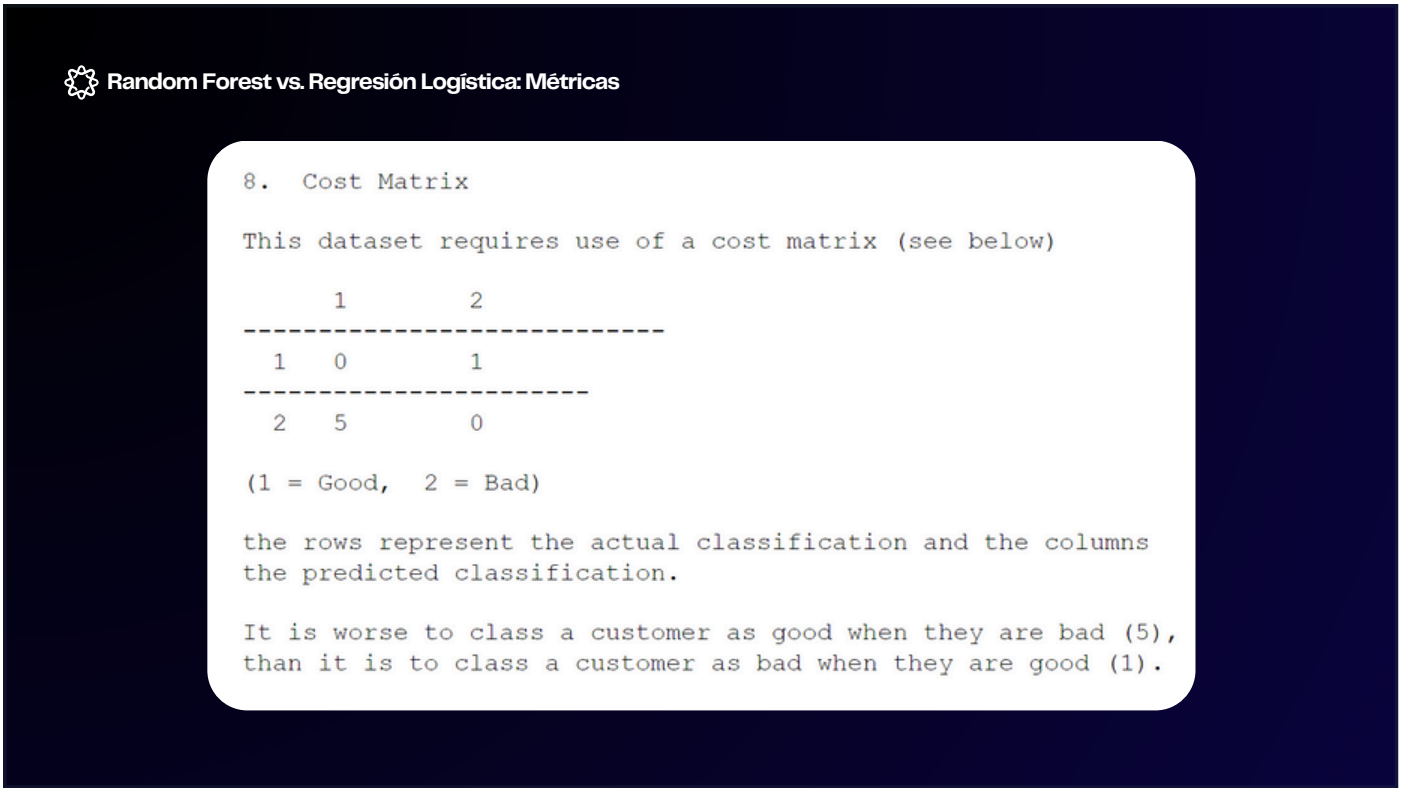
Vemos que por lo desbalanceada de la muestra la mayoria de los valores se encuentran en el cuadrante superior izquierdo. Igualmente, la cantidad de errores esta menos distribuida aun que en regresion logistica.



Vemos que las variables que mas influyen son el status de la cuenta, el monto del credito, el historial crediticio.

Algo que nos dimos cuenta aca fue que dejamos la variable de cliente, cuando deberiamos haber sacado esta variable del dataframe

Aca comparamos los dos modelos, y vemos que en todo menos Recall el modelo de random forest es mejor

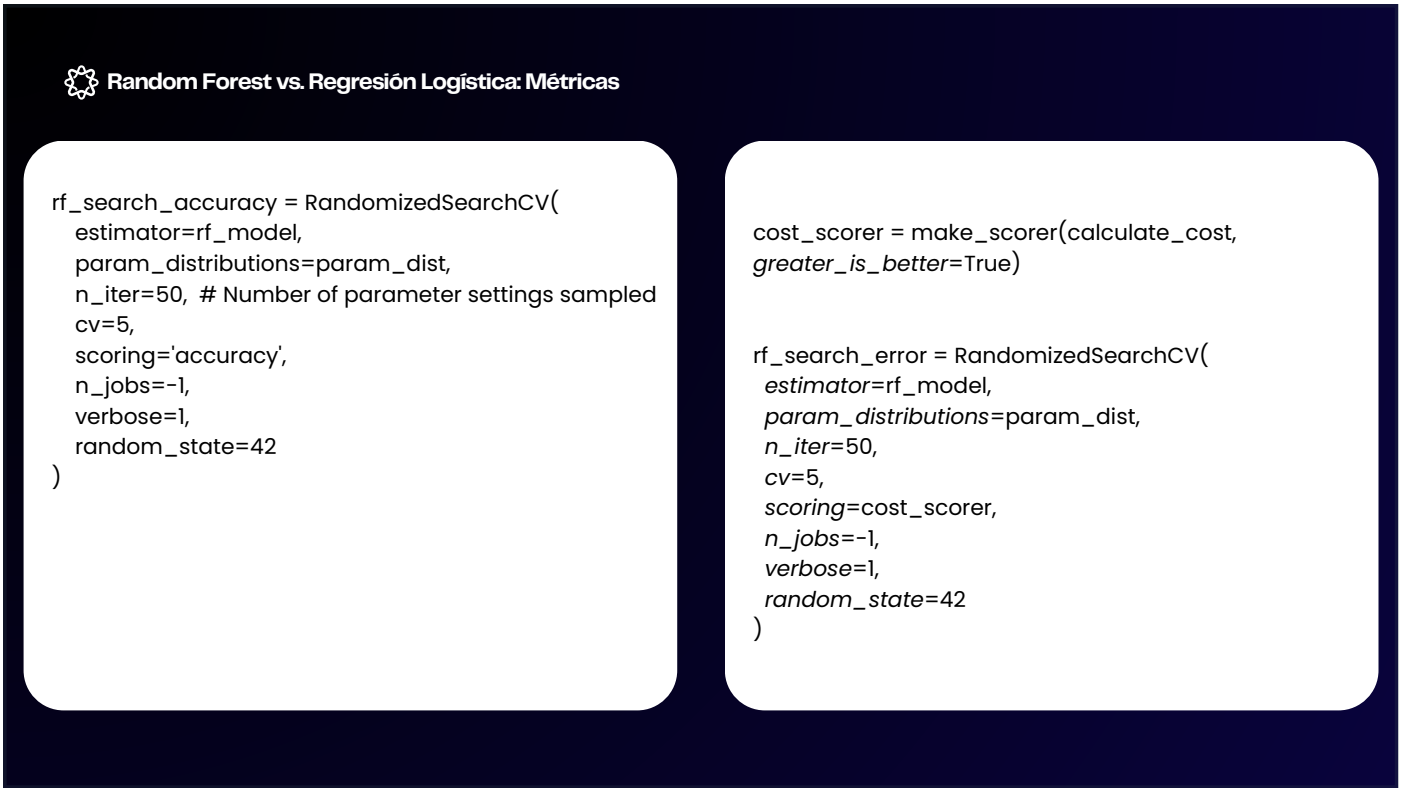


Lo que hicimos fue considerar la cost matrix que usa el banco. Esto implica que cada persona a la que le damos un credito pero no es capaz de pagarlo, lo consideramos como un costo 5. Y a la gente que no le damos un credito cuando deberiamos haberle dado un credito es un costo de 1. Esto implica que el mejor modelo es un modelo mas conservador, que reduzca el falso positivo.

Random Forest vs. Regresión Logística: Métricas

```
def calculate_cost(y_true, y_pred):

    cm = confusion_matrix(y_true, y_pred)

    tn, fp, fn, tp = cm.ravel()

    false_positive_cost = fp * 1
    false_negative_cost = fn * 5

    total_cost = false_positive_cost + false_negative_cost
    return -total_cost
```

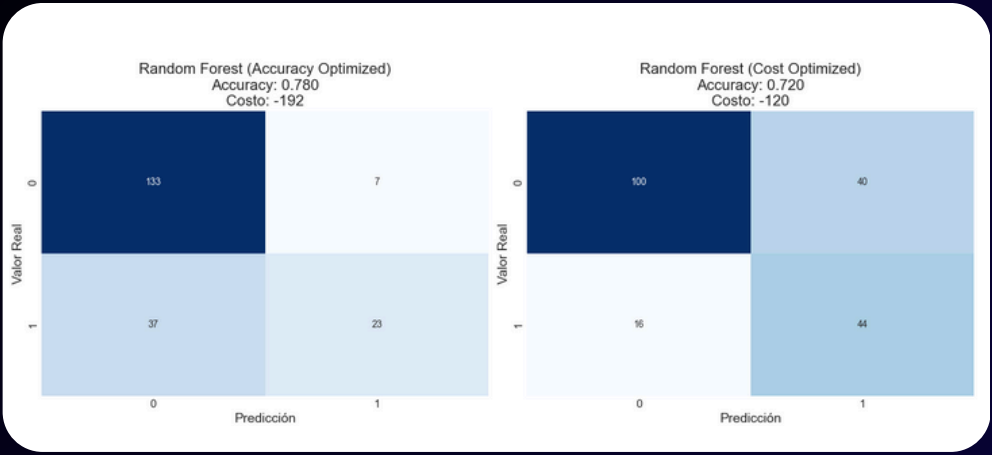
Lo que hicimos entonces fue armar una funcion con la que evaluamos el costo de cada modelo que generamos. Decimios usar unicamente un modelo de Random Forest ya que tenian mayor accuracy que regresion logistica.



Al hacer el grid search para definir hyperparametros, hicimos que guarde el modelo con mas accuracy y el modelo con menor costo.

Como la funcion devuelve un valor negativo, ponemos greater_is_better=True

Random Forest: Accuracy vs. Costo



Aca lo que vemos son los dos modelos resultantes de random forest. Uno con mayor accuracy, y el otro con menor accuracy, pero reduciendo el falso positivo en pos de menos aciertos y mas falsos negativos. Esto lo que hace es reducir los costos de morosidad del banco.

Gracias

Martín Quijano - Martina Coletto

-