

Proyecto: Modelo de regresión logística múltiple

Facultad de Ciencias Exactas UJED

Maestría en Estadística Aplicada

Curso: Modelos lineales

Maestrando: Alexis Rangel Calvo

Presentado: 6 de Mayo del 2023

ALCANCES, OBJETIVOS Y LIMITACIONES:

El presente proyecto pretende aplicar, con un ejercicio práctico y generalizable con aplicación en finanzas, los conocimientos aprendidos en el modulo de Modelos Lineales de la Maestría en Estadística Aplicada. En dicho proyecto se ha tratado de abordar todas las áreas que tiene que cubrir un proyecto real, por lo que es pertinente mencionar que por cuestión de tiempo no se ha podido profundizar en el análisis multivariado y PCA referentes al EDA, por lo que queda como labor pendiente para mejorar el proyecto y los resultados.

INTRODUCCIÓN:

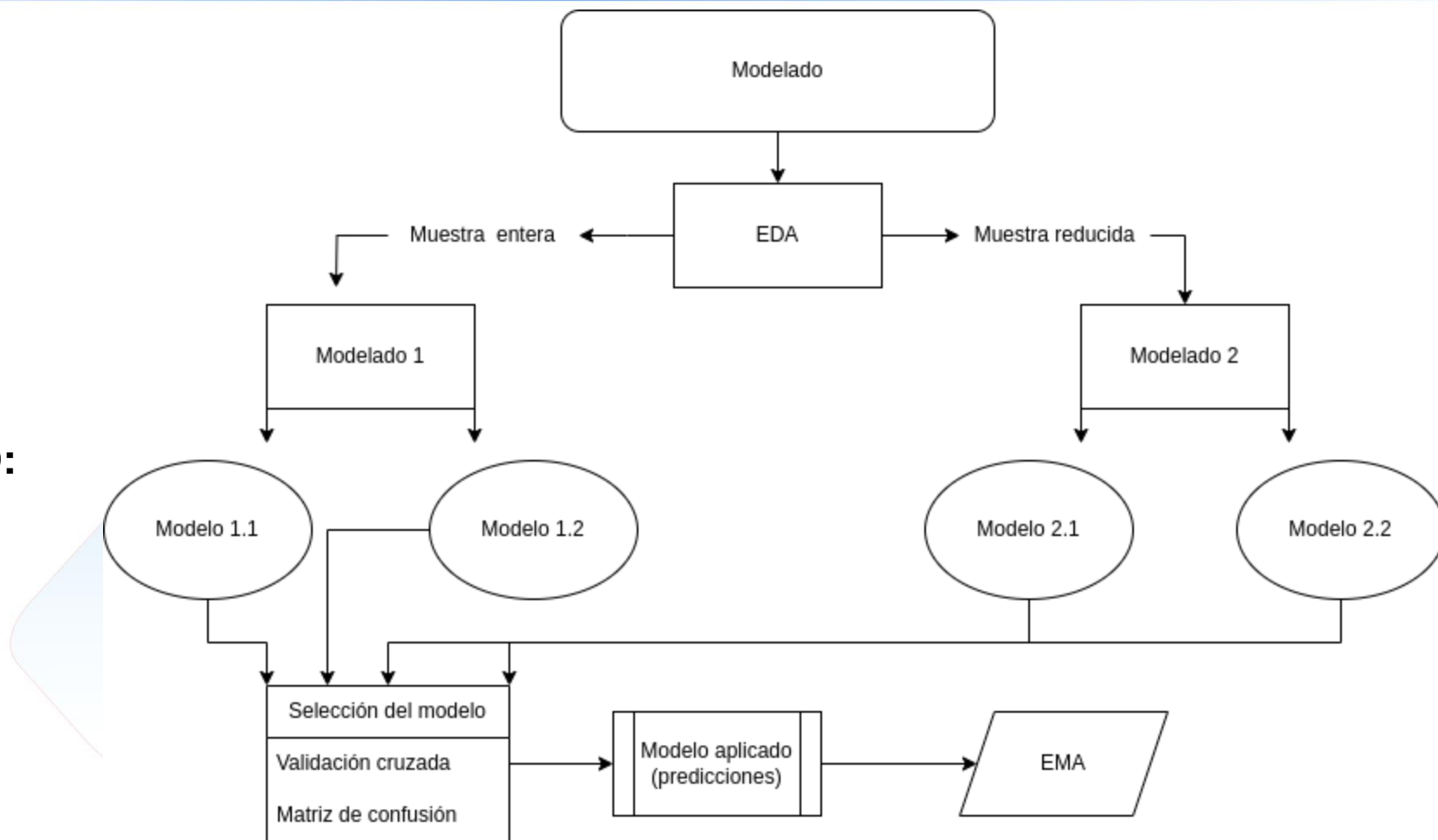
La regresión logística es un método estadístico que trata de modelar la probabilidad de una variable binaria (dos posibles valores) en función de una o más variables independientes. La principal aplicación de la regresión logística es la creación de modelos de clasificación binaria.

DATOS:

Se nos proporciona un archivo .Rdata, el cual se puede encontrar en el directorio de 'brief' cargado en el repositorio del proyecto, en el que se almacena un conjunto de datos para entrenamiento y otro para pruebas. El conjunto de entrenamiento cuenta con 34,209 registros con 17 columnas, mientras que el de pruebas contiene 15,422 con el mismo número de columnas.

Ambos conjuntos presentan 4 variables nominales (Term, Years.in.current.job, Home.Ownership y Purpose) y 12 variables continuas (Current.Loan.Amount, Credit.Score, Annual.Income, Monthly.Debt, Years.of.Credit.History, Months.since.last.delinquent, Number.of.Open.Accounts, Number.of.Credit.Problems, Current.Credit.Balance, Maximum.Open.Credit, Bankruptcies, Tax.Liens) y una variable respuesta 'Loan.Status'. Esta última no se encuentra en el conjunto de prueba dado que es la variable a predecir una vez ajustado el modelo de regresión logística.

DIAGRAMA DEL PROCESO:



JUSTIFICACIÓN DEL PROCESO:

Una vez terminado el Análisis exploratorio, que tuvo la finalidad estudiar la distribución de cada variable así como la detección de relaciones relevantes entre las mismas y encontrar la lógica del fenómeno estudiado, se concluye que del conjunto de prueba se podría obtener un subconjunto. A dicho subconjunto se le aplicarían filtros y límites por variable en donde se hayan detectado patrones del fenómeno y/o valores extremos que puedan perjudicar la generalización de la regresión. Los anteriores con el objetivo de poder ajustar más de un modelo que pretenda clasificar de la mejor forma las observaciones nuevas y nunca vistas.

JUSTIFICACIÓN DEL PROCESO:

Para cada subconjunto se decidió ajustar dos modelos los cuales en una primera se compararon, mediante validación cruzada por k pliegues y bootstrapping, entre si para preseleccionar el mejor modelo de cada subconjunto. Posteriormente la selección final del modelo que se empleará para las predicciones sobre el conjunto de prueba se hizo con base en la exactitud de la clasificación por matriz de confusión de cada uno de los modelos preseleccionados con anterioridad. De esta forma habremos podido ajustar el mejor modelo que generalice el caso de estudio.

Para información más a detalle de la justificación del proceso y el ajuste de dos modelos distintos con conjuntos de entrenamiento de diferente tamaño, por favor revisar el Análisis exploratorio realizado, publicaod en formato rmarkdown en rpubs: https://rpubs.com/alexisrangel/EDA_pf_ml

PUNTOS DEL PROCESO:

1. Análisis Exploratorio
2. Modelados con dos conjuntos de entrenamiento diferentes (muestras)
 - Muestra entera y muestra reducida a criterio, una vez efectuado el EDA.
 - Para cada modelado se generan dos modelos.

En el modelado se emplean las siguientes técnicas:

- a. Selección de variables:
 - Por criterio de Akaike information criterion (AIC) y observación del Mean squared prediction error (MSPE).
- b. Validación cruzada:
 - Validación cruzada interna (Train y Test interno)
 - Validación cruzada por k-pliegues.
 - Bootstrapping
- c. Performance del modelo:
 - Matriz de confusión, ROC y ODDS Ratios.

PUNTOS DEL PROCESO:

3. Selección del modelo.
 - A criterio con base en la exactitud por matriz de confusión.
4. Modelo aplicado al conjunto de datos 'Test'
5. Exploratory Model Analysis (EMA).

Cada uno de los procesos del ajuste del modelo de regresión logística se encuentra en archivos markdown y se pueden consultar a detalle en el repositorio del proyecto, del cual se cuenta con un link en anexos.

ANÁLISIS EXPLORATORIO

En este apartado se persiguió estudiar la relación univariada y bivariada del conjunto de entrenamiento, con la finalidad de dar un primer acercamiento al modelo y para comprender la lógica del fenómeno estudiado.

Algunos de los principales hallazgos fueron:

- Un conjunto de entrenamiento desbalanceado para la variable de respuesta (la mayoría de las observaciones tenían la variable de respuesta 'Fully Paid').
- Variables altamente correlacionadas.
- Patrones de cumplimiento en algunas variables. Con este resultado se elaboró una clasificación manual a aplicar posterior a la predicción del modelo.

Para más información en tanto al Análisis exploratorio realizado, por favor consulta el rmarkdown publicado en rpubs:
https://rpubs.com/alexisrangel/EDA_pf_ml

SOBRE EL PROCESO DE MODELADO:

Con ambos estudios, se ajustaron 6 diferentes modelos (tres por cada modelado), con la finalidad de comparar en cada modelado cada modelo mediante el estadístico de prueba MSPR, la exactitud de la clasificación por matriz y la media estimada por bootstrapping en 10 mil simulaciones.

Para más información en tanto al Modelado 1 y Modelado 2 realizado, por favor consulta los rmarkdown publicados en rpubs: https://rpubs.com/alexisrangel/Modelo_1_pf_lm y https://rpubs.com/alexisrangel/Modelo_2_pf_lm

SELECCIÓN DEL MODELO:

Dado que los conjuntos de entrenamiento empleados para cada modelado tienen diferente tamaño, el único estadístico de prueba que se podría comparar entre estos es la exactitud que nos arroja la matriz de confusión dado que la finalidad del modelo es clasificar correctamente cada observación. En este caso, los métodos de validación cruzada por k pliegues y el estadístico MSPR, como la estimación media de resultados vía bootstrapping resulta solamente tener utilidad para análisis del modelo dentro de cada modelado

Al evaluar cada uno de los modelados, mediante la exactitud clasificadora por matriz de confusión, concluíamos que el mejor par de modelos de los disponibles es el modelado 1, donde los modelos entrenaron con el conjunto de entrenamiento completo y poseen 6 variables explicativas sin interacciones.



SELECCION DE VARIABLES Y VALIDACIÓN CRUZADA POR K PLIEGUES

Modelo 1

```
glm(formula = Loan.Status ~ Term + Years.in.current.job + Purpose +  
  Current.Loan.Amount + Credit.Score + Annual.Income + Monthly.Debt +  
  Years.of.Credit.History, family = "binomial", data = Train_intern_1)  
AIC: 19828
```

```
MSE_kf_1 = cv.glm(data = Train_intern_1,  
  ..... glmfit = modelo_1,  
  ..... K = 50)  
MSE_kf_1$delta  
.....  
[1] 0.1222059 0.1222021
```

Modelo 1.1

```
glm(formula = Loan.Status ~ Term + Years.in.current.job + Purpose +  
  Current.Loan.Amount + Credit.Score + Annual.Income + Monthly.Debt +  
  Years.of.Credit.History, family = "binomial", data = Train_1)  
AIC: 26637
```

```
MSE_kf_1_1 = cv.glm(data = Train_intern_1,  
  ..... glmfit = modelo_1_1, K = 50)  
MSE_kf_1_1$delta  
.....  
[1] 0.2062993 0.1229108
```

Modelo 1.2

```
glm(formula = Loan.Status ~ Term + Purpose + Current.Loan.Amount +  
  Credit.Score + Annual.Income + Monthly.Debt + Years.of.Credit.History,  
  family = "binomial", data = Train_intern_1)  
AIC: 19832
```

```
MSE_kf_1_2 = cv.glm(data = Train_intern_1,  
  ..... glmfit = modelo_1_2, K = 50)  
MSE_kf_1_2$delta  
.....  
[1] 0.1222063 0.1222041
```

SOBRE LA SELECCIÓN DE VARIABLES:

La selección de variables busca encontrar la mejor combinación de variables en lugar de incluir el total, lidiando así con los posibles problemas de multicolinealidad. Para este ejercicio empleamos la función `glmulti()` que ejecuta la selección de modelos automatizada e inferencia multimodelos con GLM, con el método por defecto de selección exhaustiva de todos modelos candidatos.



MATRIZ DE CONFUSIÓN O MATRIZ DE COSTOS, CURVA ROC Y VALIDACIÓN CRUZADA INTERNA

Modelo 1

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	197	0
1	617	3010

Accuracy : 0.8387

Sensitivity : 0.24201

Specificity : 1.00000

Pos Pred Value : 1.00000

Neg Pred Value : 0.82989

Modelo 1.2

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	198	0
1	616	3010

Accuracy : 0.8389

Sensitivity : 0.24324

Specificity : 1.00000

Pos Pred Value : 1.00000

Neg Pred Value : 0.83012

Modelo 1.1

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	201	3
1	613	3007

Accuracy : 0.8389

Sensitivity : 0.24693

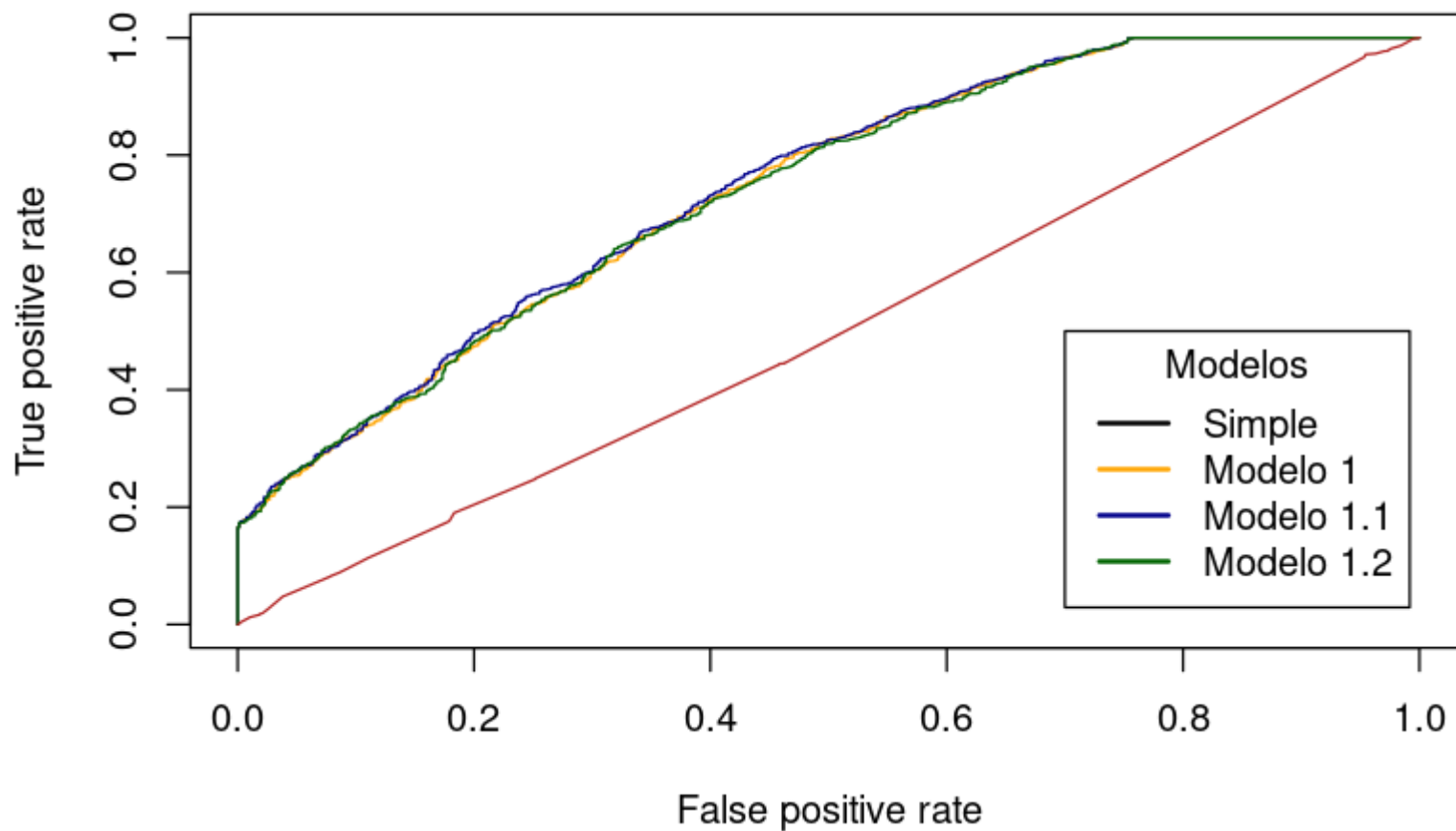
Specificity : 0.99900

Pos Pred Value : 0.98529

Neg Pred Value : 0.83066

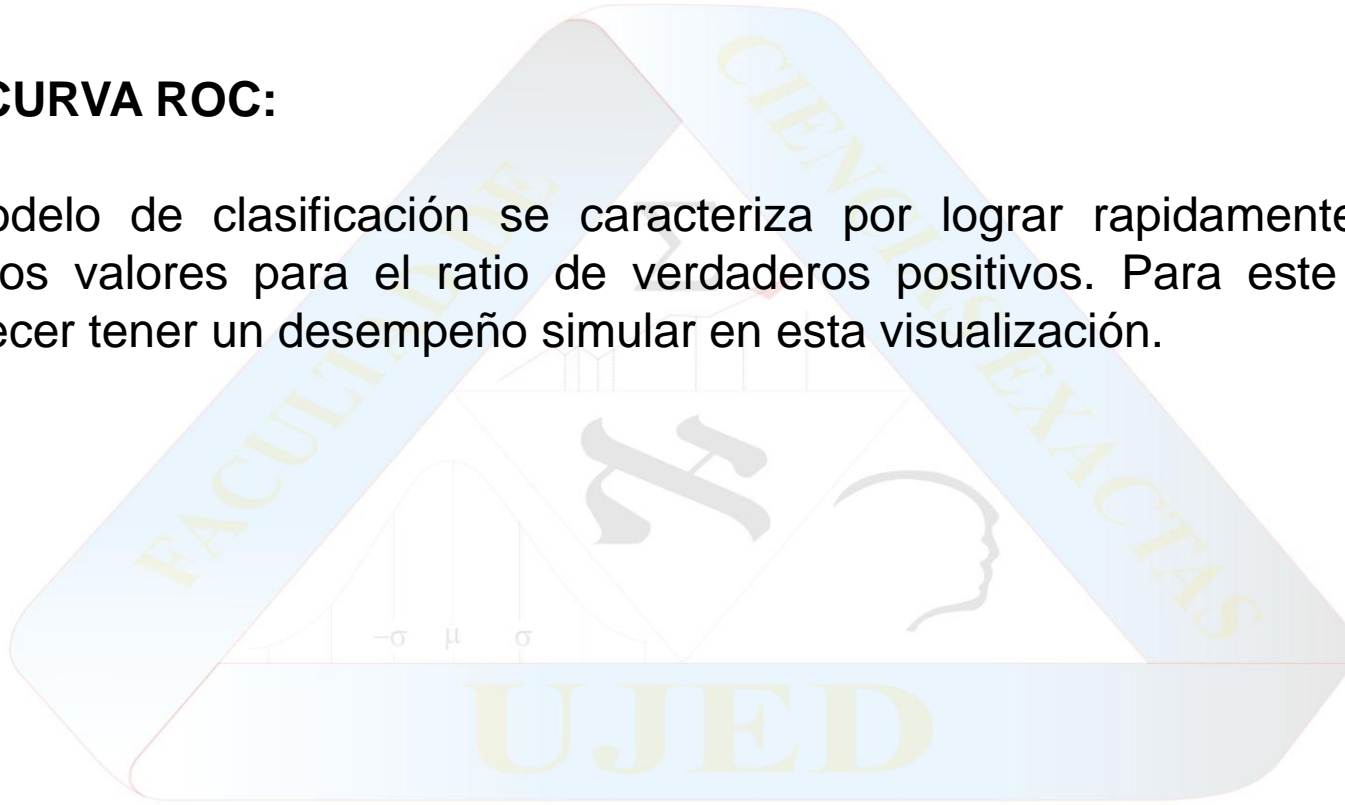
Prevalence : 0.24207

Curvas ROC modelado 1



SOBRE LA CURVA ROC:

Un buen modelo de clasificación se caracteriza por lograr rápidamente y de forma sostenida altos valores para el ratio de verdaderos positivos. Para este caso, ambos modelos parecer tener un desempeño similar en esta visualización.



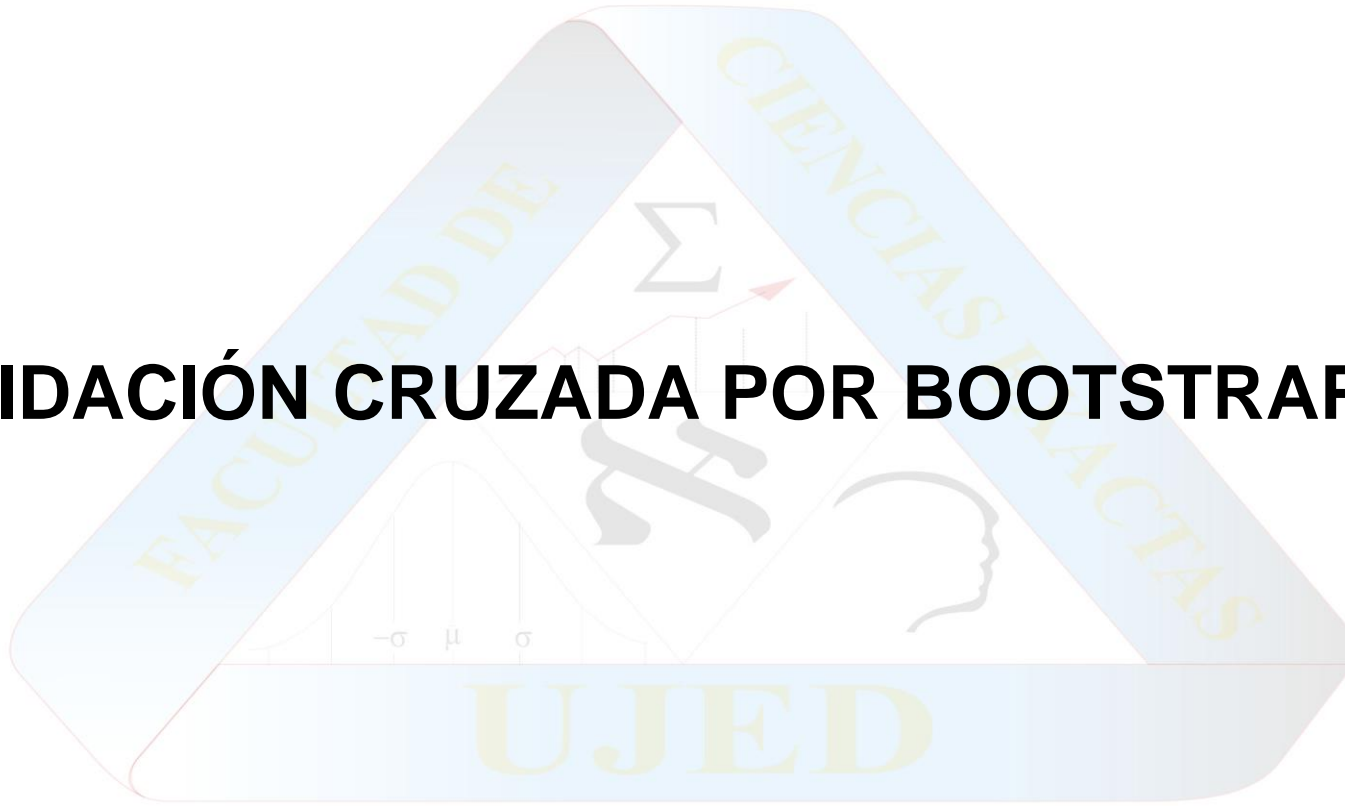
SOBRE LA VALIDACIÓN CRUZADA INTERNA

La validación cruzada interna consistió en generar dos subconjuntos de forma aleatoria a partir del conjunto de entrenamiento: el subconjunto de entrenamiento interno (representando el 75% del conjunto de entrenamiento) y el subconjunto de prueba interno (25% del conjunto de entrenamiento). El subconjunto de entrenamiento interno sirvió como conjunto para ajuste del modelo por selección de variables, como para la validación cruzada con k-pliegues y MSPR. Con el subconjunto de prueba se elaboraron predicciones a partir de registros “nunca antes vistos por el modelo”, pero con la variable respuesta correspondiente, y se observó el resultado de la clasificación en la matriz de confusión para evaluar el desempeño y seleccionar el mejor modelo que se ajuste a la lógica de negocio.

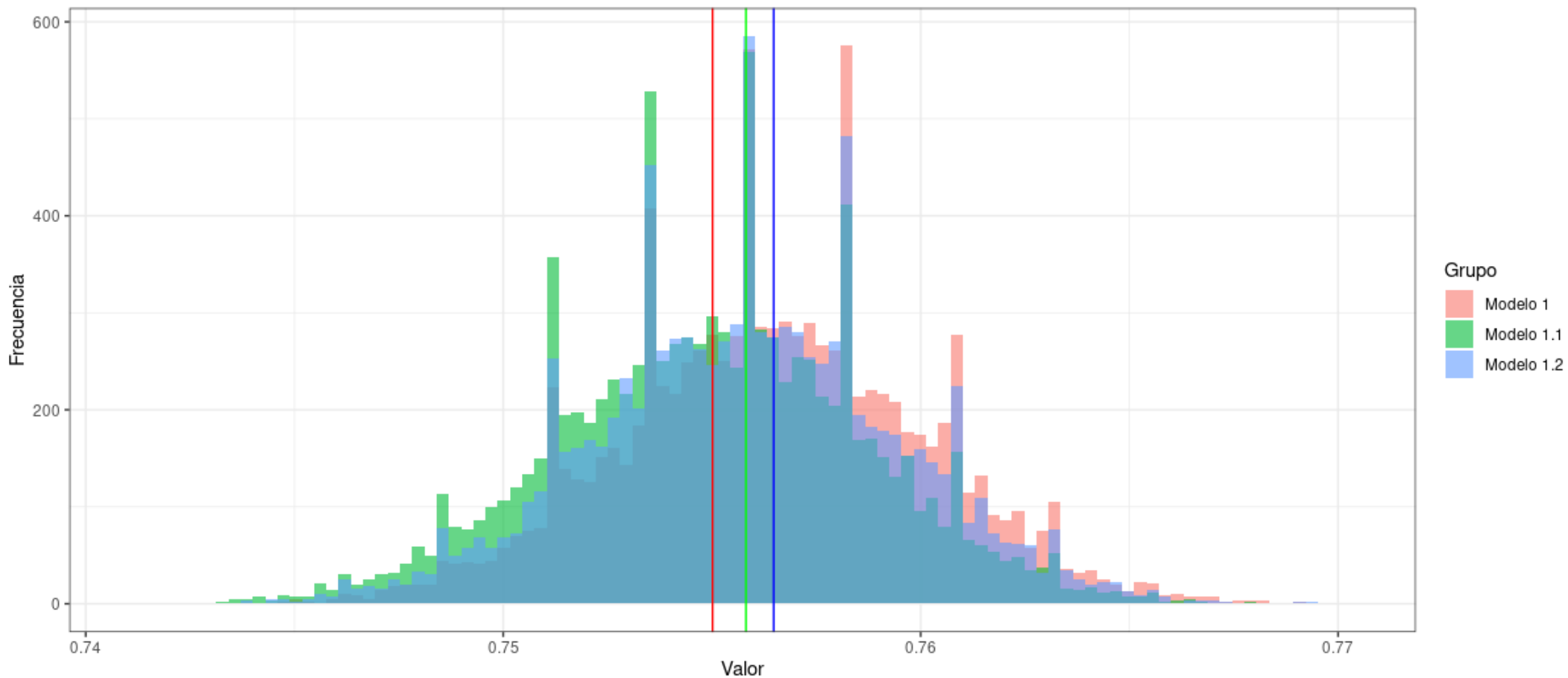
SOBRE LA VALIDACIÓN CRUZADA CON K PLIEGUES:

Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, $k-1$ grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Se busca que la estimación de error del modelo lineal sea el menor. Dicho error calculado por la función `cv.glm()` devuelve dos resultados, uno con corrección de continuidad y otro sin ella. Para los 3 casos, en donde se empleo el subconjunto de entrenamiento interno como muestra, el modelo 1.1 se ve ligeramente favorecido en la comparación.

VALIDACIÓN CRUZADA POR BOOTSTRAPPING



Comparación para cada modelo en el remuestreo (10k simulaciones)



SOBRE LA VALIDACIÓN CRUZADA MEDIANTE BOOTSTRAPPING:

Una muestra bootstrap es una muestra obtenida a partir de la muestra original por muestreo aleatorio con reposición, y del mismo tamaño que la muestra original. Muestreo aleatorio con reposición (resampling with replacement) significa que, después de que una observación sea extraída, se vuelve a poner a disposición para las siguientes extracciones.

Con esta técnica de validación cruzada y con los modelos 1, 1.1 y 1.2, estimando que en 10,000 conjuntos de prueba simulados a partir del subconjunto de prueba, en promedio se lograran clasificar apropiadamente el 76.06%, 76.01% y el 76.03% de las observaciones.

SOBRE LA MATRIZ DE CONFUSIÓN:

El mejor modelo de clasificación será aquel que clasifique a los positivos como verdaderos positivos, y a los negativos como verdaderos negativos. Por un lado, para este caso el modelo 1 (Con corte en 0.418), el modelo 1.1 (Con corte en 0.616) y 1.2 (Con corte en 0.561) clasifican en promedio el 83% de los clientes con 'Loan.Status' igual a 'Fully Paid' (verdadero positivo). Por otro lado, dichos modelos clasifican en promedio el 99.5% de los clientes con 'Loan.Status' igual a 'Charged Off'.

En otras palabras, se espera que el modelo clasifique apropiadamente a los clientes con pago insatisfactorio (especificidad), pero no sea completamente efectivo para clasificar a los clientes con pago satisfactorio (sensibilidad).

Es aquí donde se valoran los costos de cada escenario, y se ajusta el corte de la clasificación para maximizar la sensibilidad, la especificidad o la exactitud.

SELECCIÓN FINAL DEL MODELO:

Hasta este punto, se concluye que el mejor modelo del primer modelado resulta ser el modelo 1.1 dado que:

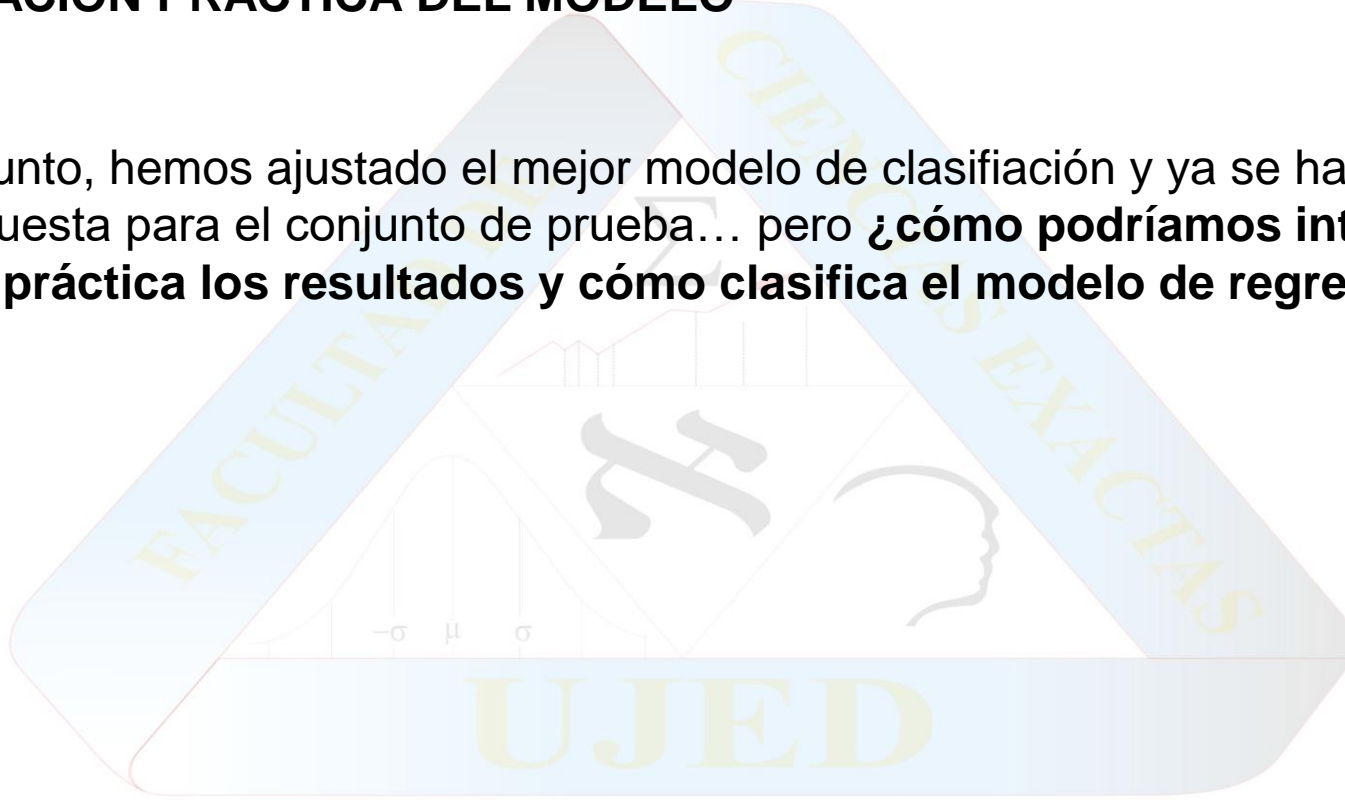
- Los 3 modelos presentan prácticamente el mismo nivel de exactitud, sensibilidad y especificidad.
- El modelo 1.1 se ajusta más a la necesidad que la lógica de negocio pretende atender con este modelo de clasificación.

Para más información en tanto al las predicciones para el conjunto de prueba con el modelo seleccionado y ajustado, por favor consulta el markdown publicado en rpubs:

https://rpubs.com/alexisrangel/predicciones_fp_lm

INTERPRETACIÓN PRACTICA DEL MODELO

Hasta este punto, hemos ajustado el mejor modelo de clasificación y ya se ha estimado la variable respuesta para el conjunto de prueba... pero **¿cómo podríamos interpretar de una manera práctica los resultados y cómo clasifica el modelo de regresión logística ajustado?**



ODDS RATIOS (OR)

Matemáticamente un OR corresponde a un cociente entre dos odds, siendo un odds una forma alternativa de expresar la posibilidad de ocurrencia de un evento de interés. Es decir que el odd ratio es la probabilidad de que se de el evento A contra probabilidad de que no se de el evento A, de acuerdo con el caso de estudio.

Cuando el OR es igual a 0, se nos indica que no hay relación entre la variable dependiente y la independiente (porque euler elevado a la 0 siempre sera igual a 1). Cuando el OR es mayor a 1 la relación entre eventos es positiva, si es menor se da el caso contrario.

Es decir, si el OR es mayor o menor a 1, por cada unidad que se incremente la variable independiente, dicho evento estará asociado a un aumento o una disminución en las probabilidades de que la dependiente adopte un 1 (para la respuesta dicótoma)

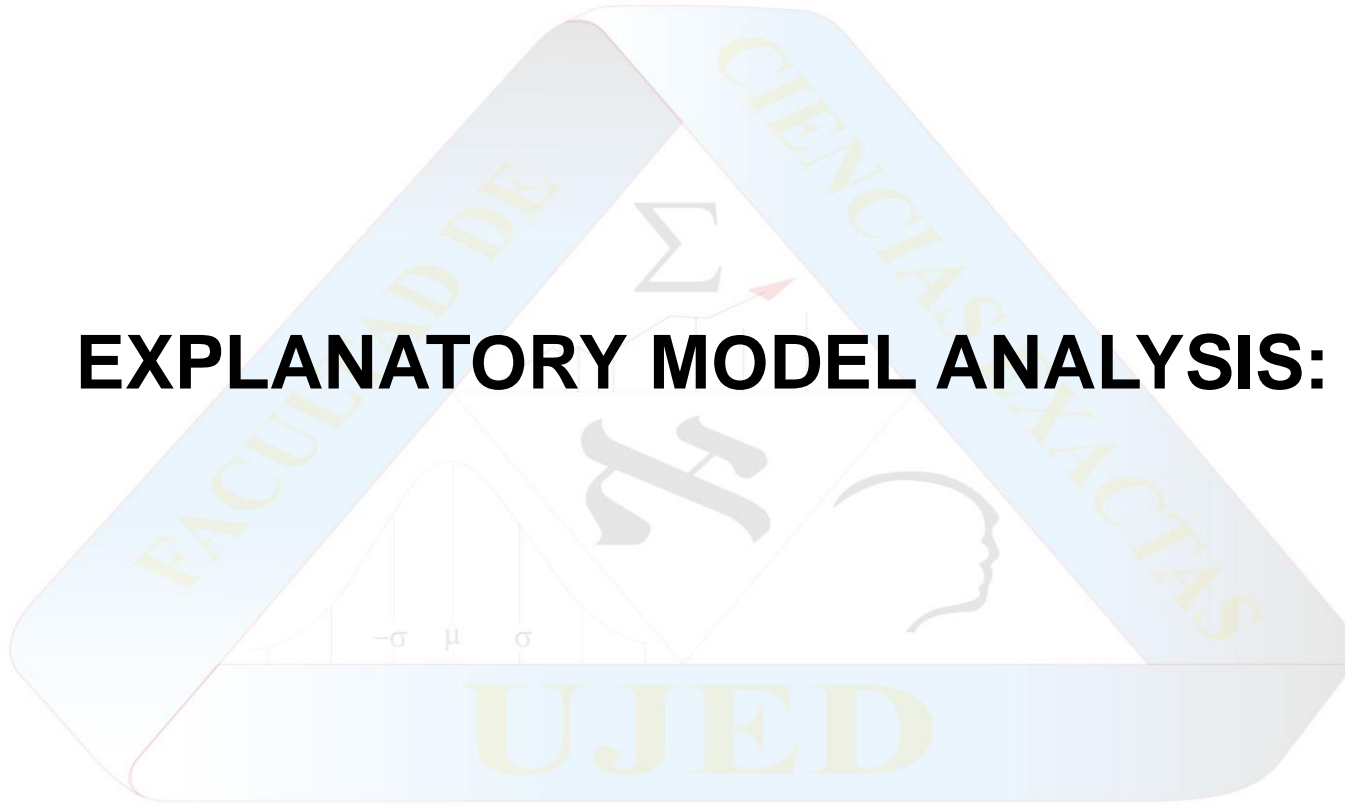
ALGUNOS DE LOS ODDS RATIOS DE VARIABLES NOMINALES:

- Si la variable `Purpose` adopta el `buy a car`, este evento estara asociado con un aumento del 231% en las probabilidades del pago al corriente (`Fully Paid`) comparado con el evento cuando `Purpose` adopta el valor de `Business Loan`
- Si la variable `Purpose` adopta el `buy a house`, este evento estara asociado con un aumento del 30% en las probabilidades del pago al corriente (`Fully Paid`) comparado con el evento cuando `Purpose` adopta el valor de `Business Loan`
- Si la variable `Years.in.current.job` adopta el `9 year`, este evento estara asociado con un aumento del 26.5% en las probabilidades del pago al corriente (`Fully Paid`) comparado con el evento cuando `Years.in.current.job` adopta el valor de `< 1 year`
- Si la variable `Years.in.current.job` adopta el `3 years`, este evento estara asociado con un aumento del 26.7% en las probabilidades del pago al corriente (`Fully Paid`) comparado con el evento cuando `Years.in.current.job` adopta el valor de `< 1 year`

ALGUNOS DE LOS ODDS RATIOS VARIABLES CONTINUAS DEL MODELO:

- Si la variable `Current.Loan.Amount` aumenta en 10,000,000 unidades, este evento estara asociado con un aumento del 76.5% en las probabilidades del pago al corriente (`Fully Paid`).
- Si la variable `Credit.Score` aumenta en una 100 unidades, este evento estara asociado con una disminución del 13.74% en las probabilidades del pago al corriente (`Fully Paid`).
- Si la variable `Annual.Income` aumenta en 1,000,000 unidades, este evento estara asociado con un aumento del 34.4% en las probabilidades del pago al corriente (`Fully Paid`).
- Si la variable `Monthly.Debt` aumenta en 10,00 unidades, este evento estara asociado con una disminución del 13.4% en las probabilidades del pago al corriente (`Fully Paid`).

EXPLANATORY MODEL ANALYSIS:



Explanatory Model Analysis for models is whats Exploratory Data Analysis is for Data

De este amplio campo del aprendizaje automatico, se presenta un breve acercamiento con enfoque agnóstico para estudiar la relación entre los inputs y outputs del modelo de regresión logístico al predecir una observación aislada, proceso llamado también explicaciones de nivel.

Para más información en tanto al explanatory model analysis, por favor consulta el rmarkdown publicado en rpubs:
https://rpubs.com/alexisrangel/EMA_pf_ml

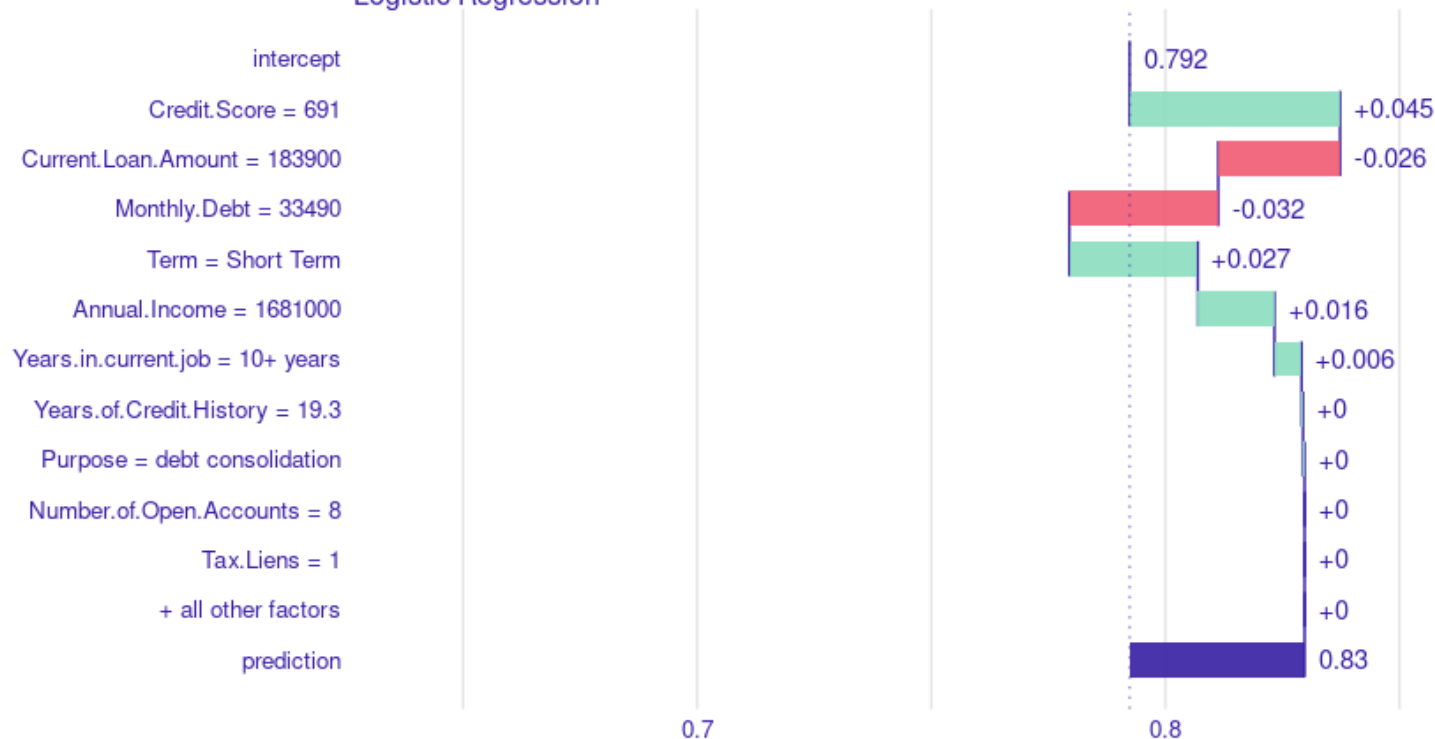
Predicción individual no. 1 (id = 8972)

Credit.Score	691
Current.Loan.Amount	\$ 183,900
Monthly.Debt	\$ 33,490
Term	Short Term
Anual.Income	\$ 1,681,000
Purpose	Debt consolidation

EXPLICACIÓN DE NIVEL PARA OBSERVACIONES AISLADAS

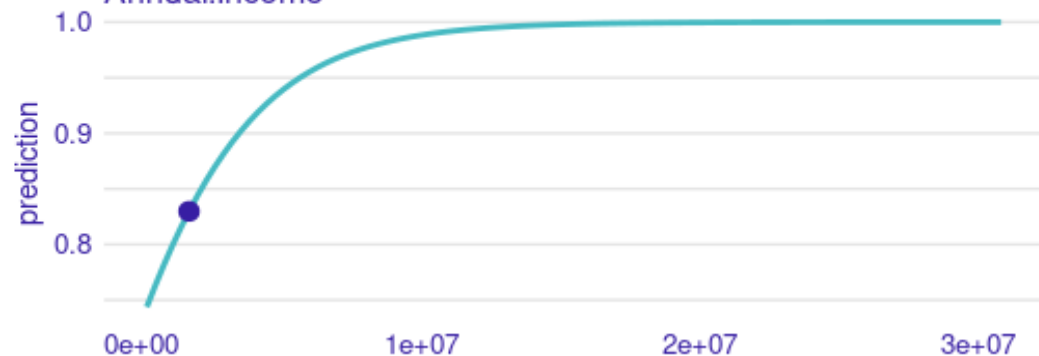
Break-Down for single obs with Fully Paid response

Logistic Regression

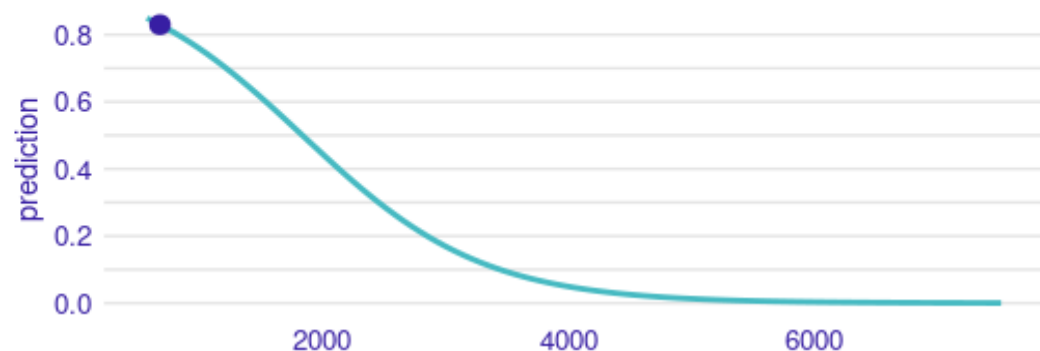


CETERIS PARIBUS PROFILE

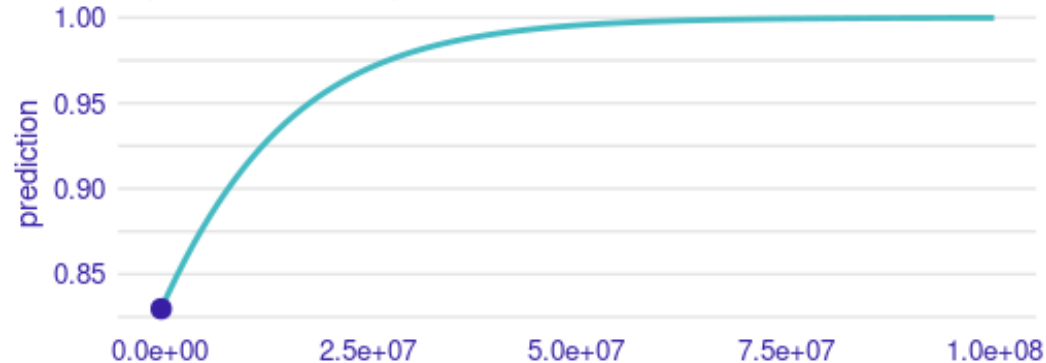
Break-Down Fully Paid response
created for the Logistic Regression model
Annual.Income



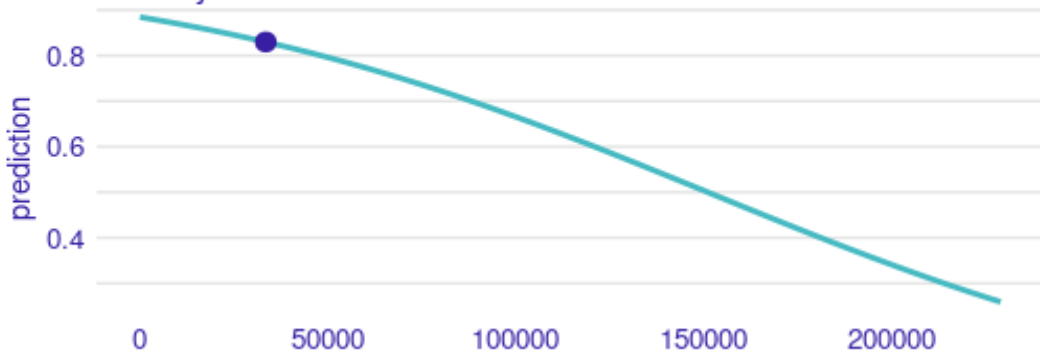
Break-Down Fully Paid response
created for the Logistic Regression model
Credit.Score



Break-Down Fully Paid response
created for the Logistic Regression model
Current.Loan.Amount



Break-Down Fully Paid response
created for the Logistic Regression model
Monthly.Debt



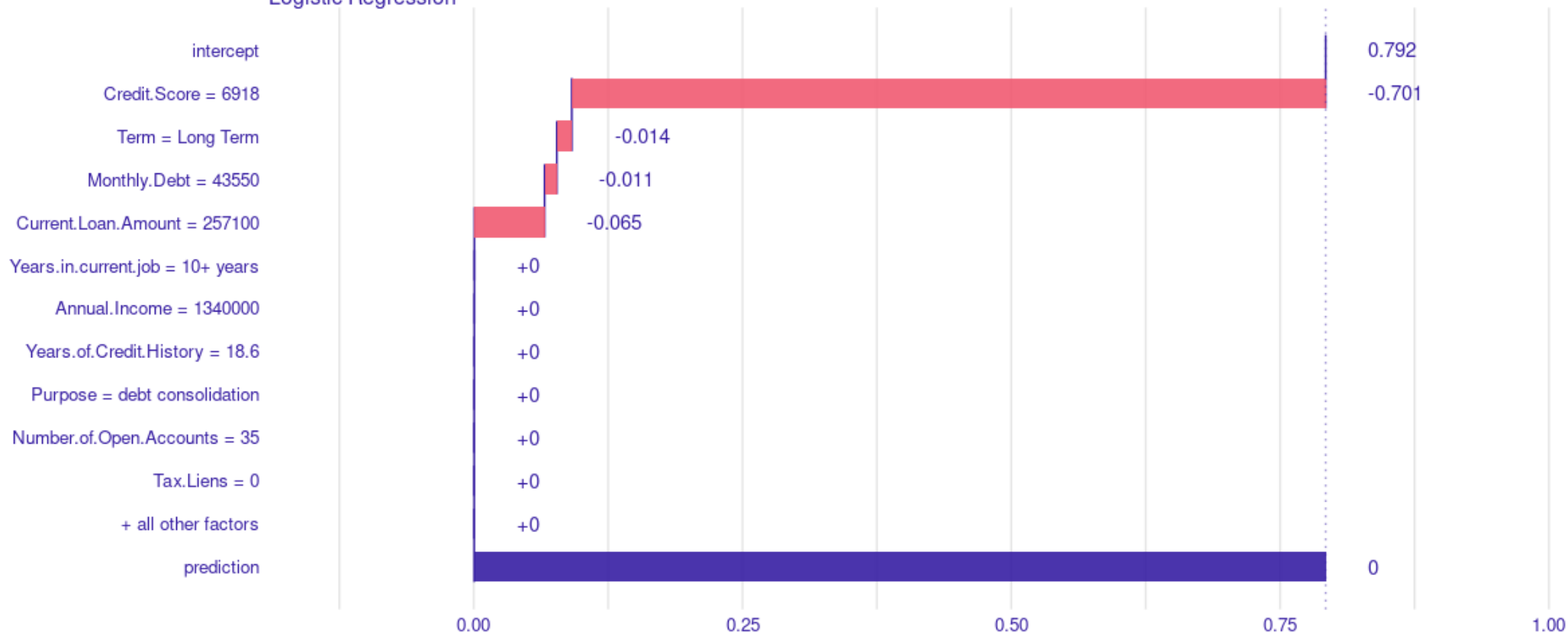
Predicción individual no. 2 (id = 11434)

Credit.Score	6,918
Current.Loan.Amount	\$ 257,100
Monthly.Debt	\$ 43,550
Term	Long Term
Anual.Income	\$ 1,340,000
Purpose	Debt Consolidations

EXPLICACIÓN DE NIVEL PARA OBSERVACIONES AISLADAS

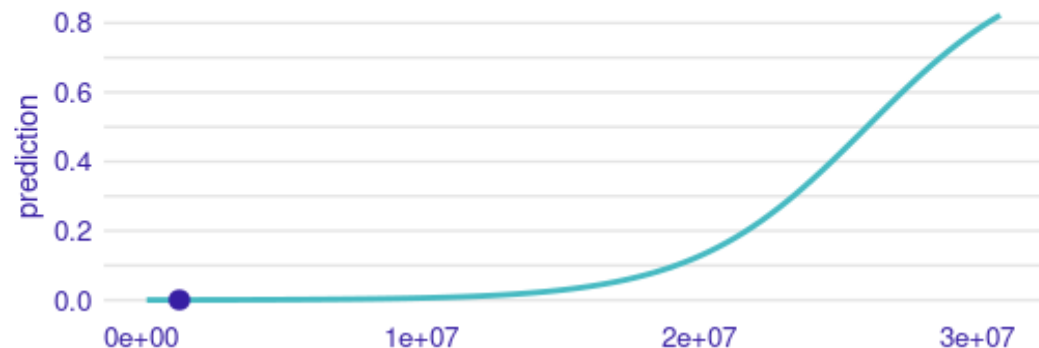
Break-Down for single obs with Charged Off response

Logistic Regression

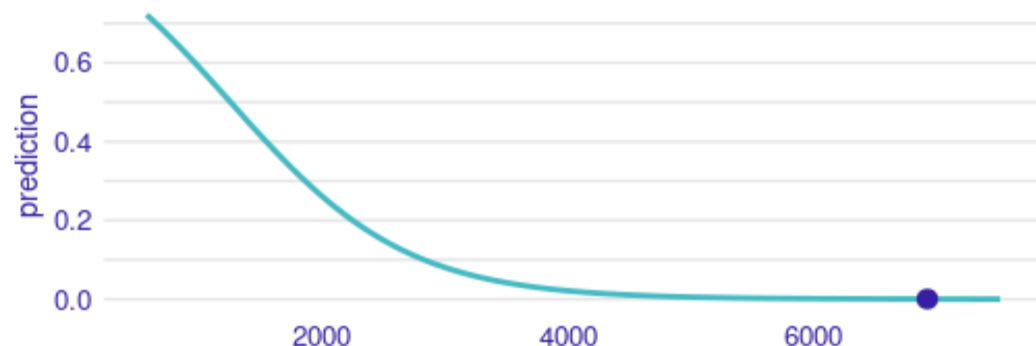


CETERIS PARIBUS PROFILE

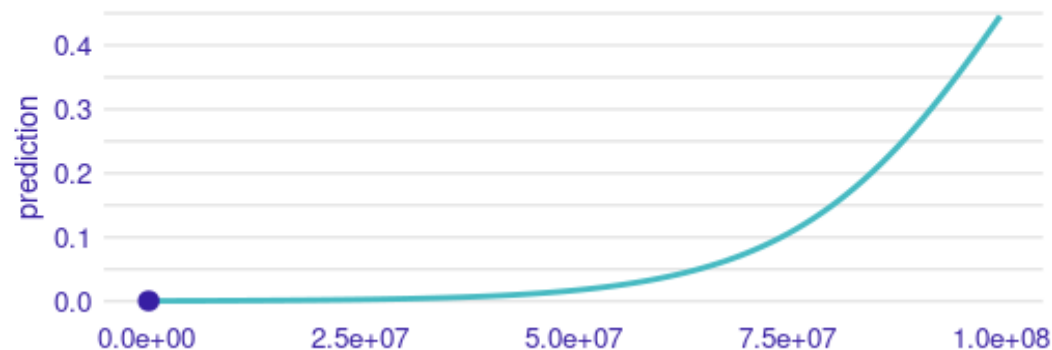
Break-Down Charged Off response
created for the Logistic Regression model
Annual.Income



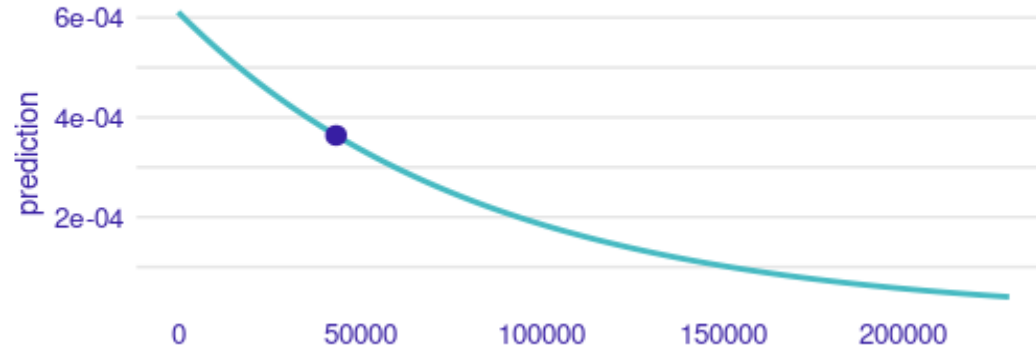
Break-Down Charged Off response
created for the Logistic Regression model
Credit.Score



Break-Down Charged Off response
created for the Logistic Regression model
Current.Loan.Amount



Break-Down Charged Off response
created for the Logistic Regression model
Monthly.Debt



CONCLUSIONES

Concluimos con un modelo de regresión logística que presenta:

- Una exactitud de 83.89% en la clasificación para el conjunto de entrenamiento, clasificando el 83.06% de los casos en donde la variable respuesta adopta el valor de “Fully Paid” y un *98.5% de los casos en donde la variable adopta el valor de “Charged Off”.
- Una media estimada del 75.5% de aciertos o registros correctamente clasificados, esto mediante una simulación por bootstrapping de mil muestras semejantes a la del conjunto de prueba.

Dicho modelo se logró ajustar con el conjunto entero de entrenamiento, debido a que los resultados del EDA no se han concluido y se estima que con un profundo análisis multivariado podamos tener un modelo propuesto en el modelado 2 para poder compararlo con el ya ajustado (modelo 1.1) y volver a definir el mejor modelo.

AREAS DE OPORTUNIDAD:

Se considera que:

- Un EDA más extenso, que incluya un análisis multivariado profundo, así como el empleo de técnicas como *oversampling* para balancear la desbalanceada muestra a clasificar podrían aportar a un mejor ajuste.
- El estudio de interacciones entre variables, como un mejor ajuste de hiperparámetros podrían ofrecer un modelo de clasificación con una mejor exactitud.

Bibliografía:

- glmulti: An R Package for Easy Automated Model Selection with (Generalized) Linear Models. (2010). Journal of Statistical Software, 34(12).

<https://pdfs.semanticscholar.org/24f6/a684dd840e95df2875c23a8ca786dbfccae0.pdf>

- Rodrigo, J. A. (s. f.). Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstrapping.

https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap

- EMA: <https://ema.drwhy.ai/>

Link al repositorio del proyecto en github:

<https://github.com/alexisrangelcalvo/FP-LM-MAEA-UJED-1S>