Proyecto: Modelo de regresión logística múltiple

Facultad de Ciencias Exactas UJED

Maestría en Estadística Aplicada

Curso: Modelos lineales

Maestrando: Alexis Rangel Calvo

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 regresión logística en el curso 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 de mencionar que por cuestión de tiempo no se ha podido profundizar en el análisis multivariado y PCA, por lo que queda como labor pendiente para mejorar el proyecto.

INTRODUCCIÓN:

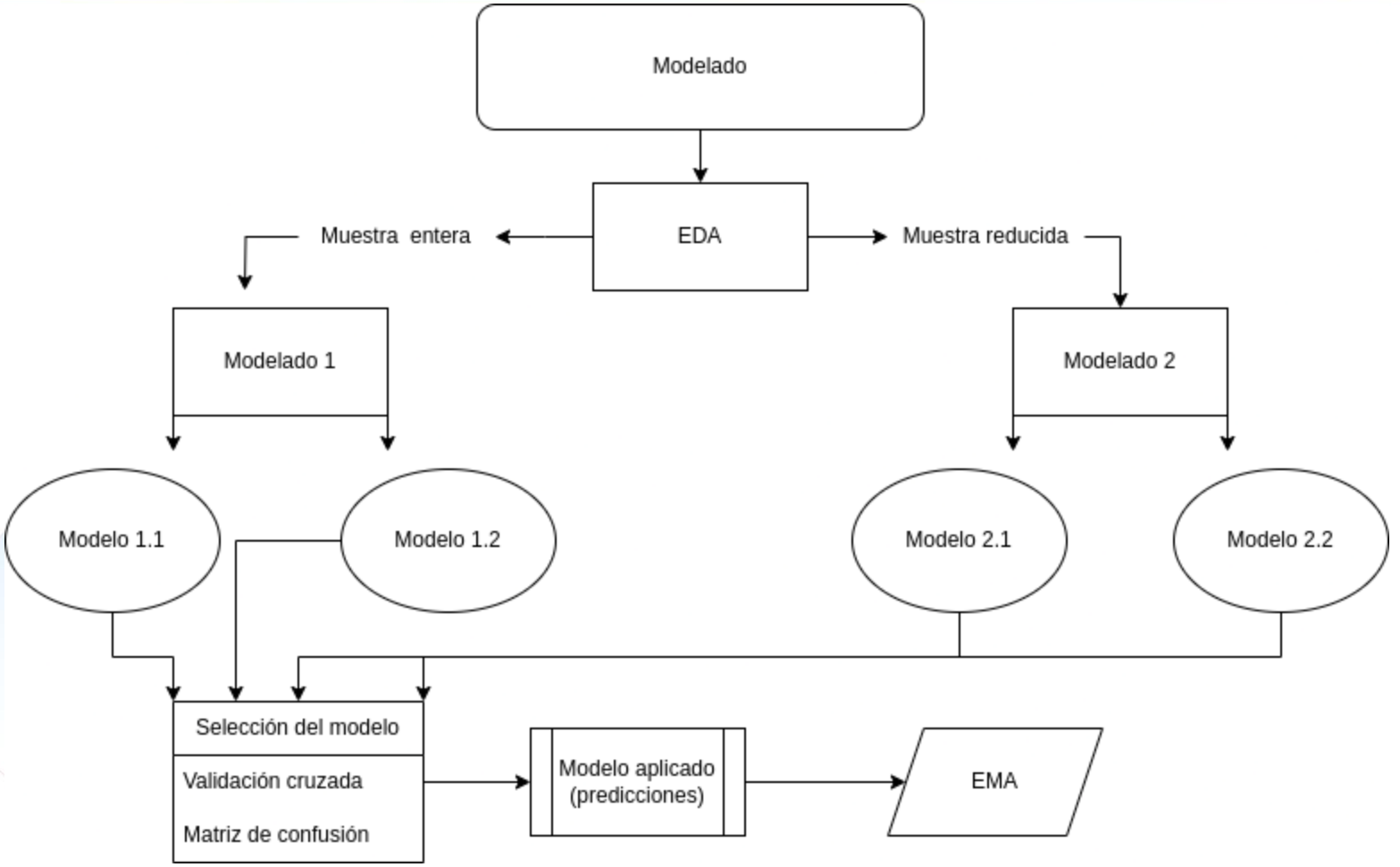
La regresión logística es un método estadístico que trata de modelar la probabilidad de una variable cualitativa binaria (dos posibles valores) en función de una o más variables independientes. La principal aplicación de la regresión logística es la creación de modelos de clasificación binaria.

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'. E-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 limites por variable en donde se hayan detectado patrones del fenómeno y/o valores extremos que puedan perjudicar la generalización de la regresión. Los anterior con el objetivo de poder ajustar más de un modelo que pretenda clasificar de la mejor forma las observaciones nuevas y nunca vistas.

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.

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 (MSPR).

b. Validación cruzada:
    - Validación cruzada por k-pliegues.
    - Bootstraping

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 rmarkdown 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.
- 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 4 diferentes modelos (dos 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 de confusión y la media estimada por remuestreo 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 bootstraping 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 1.1

```
Call:
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
```

- Validación por k-fold para el modelo 1.1

```
```{r}
MSE_kf_1_1 = cv.glm(data = Test_intern_1,
                    glmfit = modelo_1_1, K = 50)
MSE_kf_1_1$delta
```
```

```
[1] 0.2016383 0.1183749
```

Modelo 1.2

```
Call:
glm(formula = Loan.Status ~ Term + Years.in.current.job + Purpose +
    Current.Loan.Amount + Credit.Score + Annual.Income + Monthly.Debt,
    family = "binomial", data = Train_intern_1)
                                                        AIC: 19843
```

- Validación por k-fold para el modelo 1.2

```
```{r, warning=FALSE}
MSE_kf_1_2 = cv.glm(data = Test_intern_1,
                    glmfit = modelo_1_2, K = 50)
MSE_kf_1_2$delta
```
```

```
[1] 0.2019282 0.1184296
```

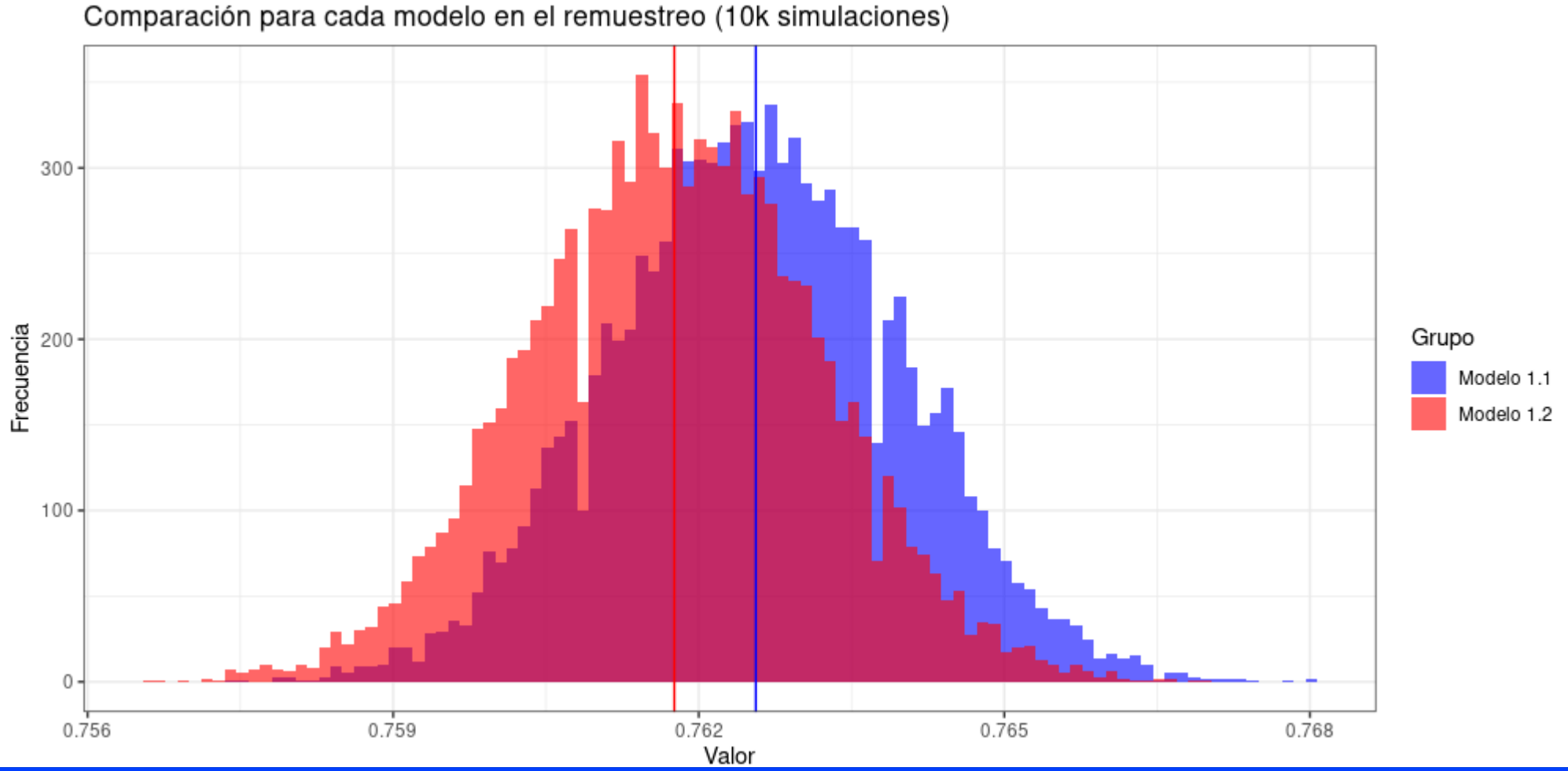
SOBRE LA SELECCIÓN DE VARIABLES:

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.

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 ambos casos el modelo 1.1 se ve favorecido en la comparación.

VALIDACIÓN CRUZADA POR BOOTSTRAPPING


Comparación para cada modelo en el remuestreo (10k simulaciones)
Frecuencia
300
200
100
0
0.756
0.759
0.762
0.765
0.768
Valor
Grupo
Modelo 1.1
Modelo 1.2

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 y 1.2, estimando que en 10,000 conjuntos de prueba simulados a partir del conjunto de entrenamiento, en promedio se lograran clasificar apropiadamente el 76.26% y el 76.18% de las observaciones.

MATRIZ DE CONFUSIÓN O MATRIZ DE COSTOS Y CURVA ROC

Modelo 1.1

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  453 1379
         1    0 6720

               Accuracy : 0.8388
                 95% CI : (0.8308, 0.8465)
    No Information Rate : 0.947
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3405

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.82973
```

Modelo 1.2

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  454 1378
         1    0 6720

               Accuracy : 0.8389
                 95% CI : (0.8309, 0.8466)
    No Information Rate : 0.9469
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3411

 Mcnemar's Test P-Value : <2e-16

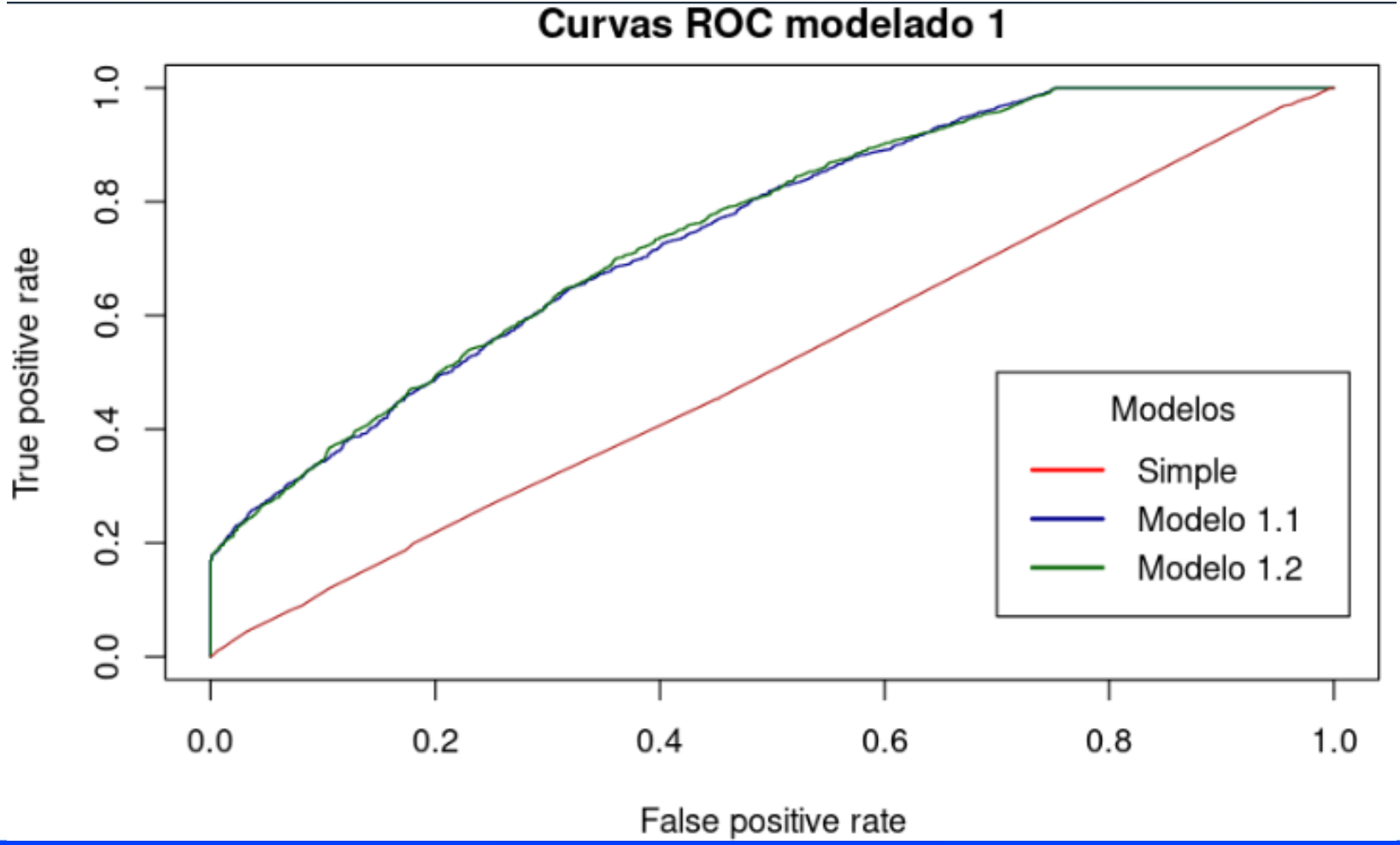
            Sensitivity : 1.00000
            Specificity : 0.82983
```

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 1.1 (Con corte en 0.451) y 1.2 (Con corte en 0.562) clasifica el 100% de los clientes con 'Loan.Status' igual a 'Fully Paid' (verdadero positivo). Por otro lado, dichos modelos clasifican solo el 82.9% 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 satisfactorio (sensibilidad), pero no sea completamente efectivo para clasificar a los clientes que se caracterizan por tener adeudos importantes relativos a la deuda (especificidad).

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


Curvas ROC modelado 1
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Modelos
Simple
Modelo 1.1
Modelo 1.2

SOBRE LA CURVA ROC:

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

SELECCIÓN FINAL DEL MODELO:

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

- Ambos modelos presentan el mismo nivel de exactitud.
- El MSPR resulta ser inferior en el modelo 1.1.
- La media estimada de porcentaje sobre 100 en la clásifiación es mayor en el modelo 1.1.

Para más información en tanto al las predicciones para el conjunto de prueba con el modelo seleccionado y ajustado, por favor consulta el rmarkdown publicado en rpubs:
https://rpubs.com/alexisrangel/predicciones_fp_lm

INTERPRETACIÓN PRACTICA DEL MODELO

Hasta este punto, hemos ajustado el mejor modelo de clasifiación 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 probabilidadades de que la dependiente adopte un 1 (para la respuesta dicótoma)

ODDS RATIOS VARIABLES NOMINALES MÁS IMPORTANTES:

- Si la variable `Purpose` adopta el `buy a car`, este evento estara asociado con un aumento del 265% en las probabilidades del pago al corriente (`Fully Paid`)

- Si la variable `Purpose` adopta el valor `take a trip`, este evento estara asociado con un aumento del 198% en las probabilidades del pago al corriente (`Fully Paid`)

- Si la variable `Purpose` adopta el valor `house improvement`, este evento estara asociado con un aumento del 116% en las probabilidades del pago al corriente (`Fully Paid`)

- Si la variable `Term` adopta el valor `Short`, este evento estara asociado con un aumento del 69% en las probabilidades del pago al corriente (`Fully Paid`) comparado con el evento cuando `Term` adopa el valor de `Long`.

ODDS RATIOS VARIABLES CONTINUAS DEL MODELO:

- Si la variable `Credit.Score` aumenta en 100 unidades, este evento estara asociado con una disminución del 13% en las probabilidades del pago al corriente (`Fully Paid`).

- Si la variable `Annual.Income` aumenta en 100,000 unidades, este evento estara asociado con un aumento del 32% en las probabilidades del pago al corriente (`Fully Paid`).

- Si la variable `Monthly.Debt` aumenta en 10,000 unidades, este evento estara asociado con una disminución del 11.8% en las probabilidades del pago al corriente (`Fully Paid`).

- Si la variable `Current.Loan.Amount` aumenta en 1,000,000 unidades, este evento estara asociado con un aumento del 0.77% en las probabilidades del pago al corriente (`Fully Paid`).

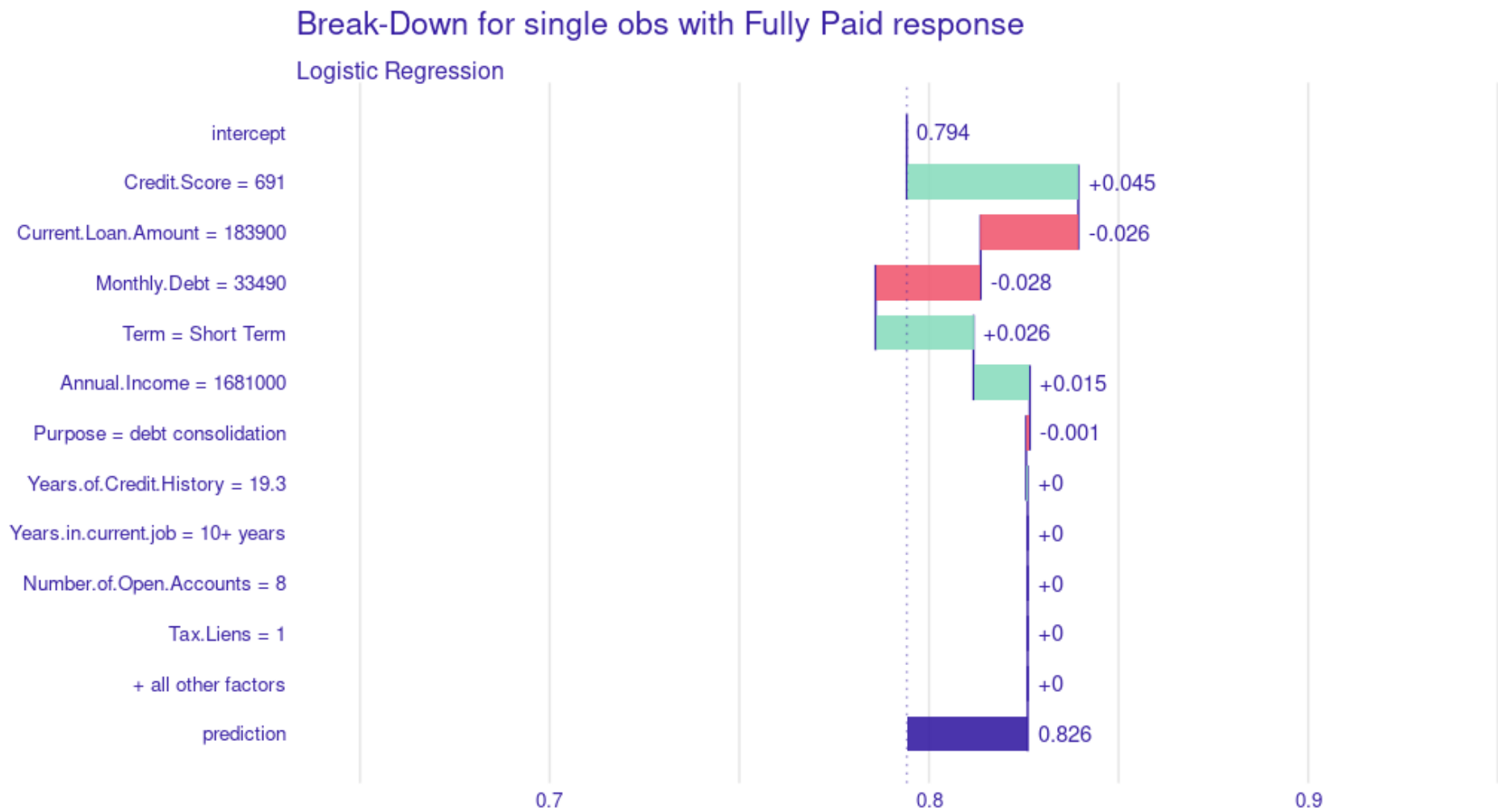
EXPLANATORY MODEL ANALYSIS:

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

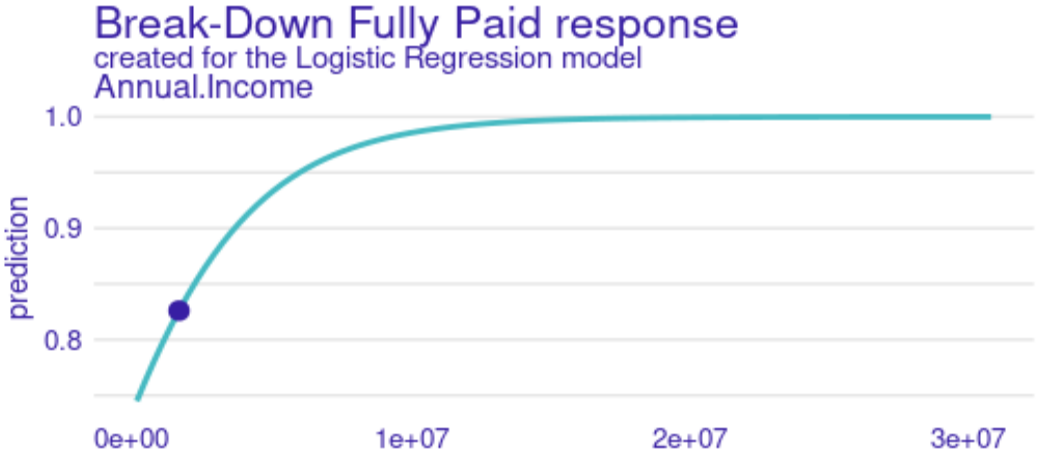
De este amplio campo del aprendizaje automatico, se presenta un breve acercamiento con enfoque agnóstico para estudiar la relación entre los inputs y outputs del modelo de regresión logístico al predecir una observación aislada, proceso llamado también explicaciones de nivel.

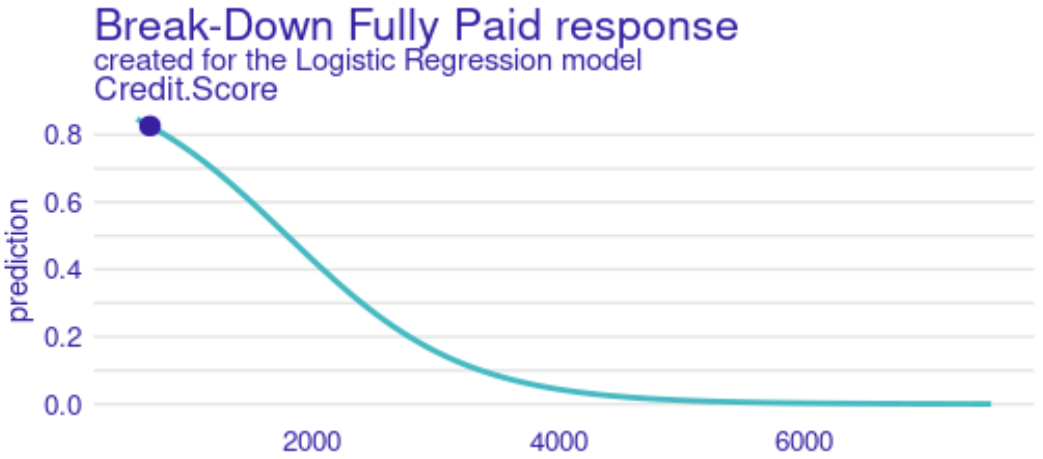
Para más información en tanto al explanatory model analysis, por favor consulta el rmarkdown publicado en rpubs:
https://rpubs.com/alexisrangel/EMA_pf_ml

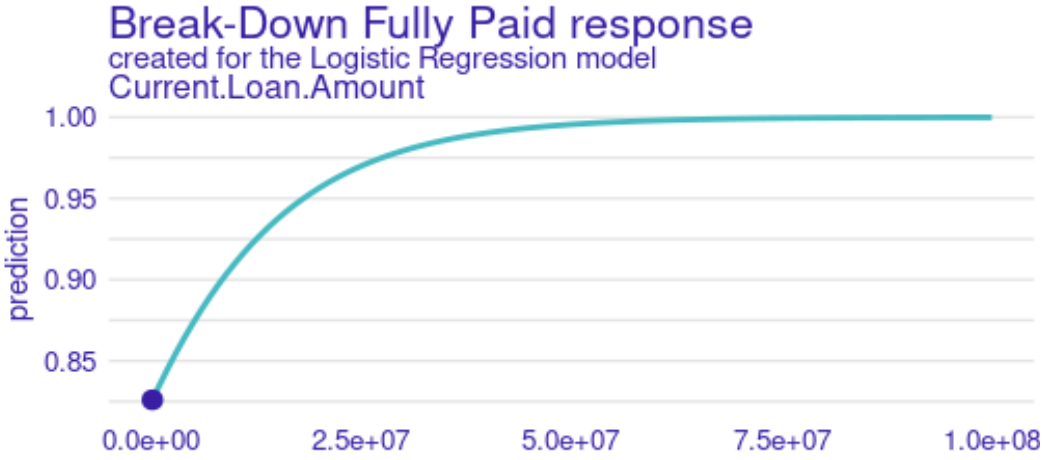
Predicción individual no. 1 (id = 8972)	
Credit.Score	691
Current.Loan.Amount	$ 183,900
Monthly.Debt	$ 33,490
Term	Short Term
Anual.Income	$ 1,681,000
Purpose	Debt consolidation

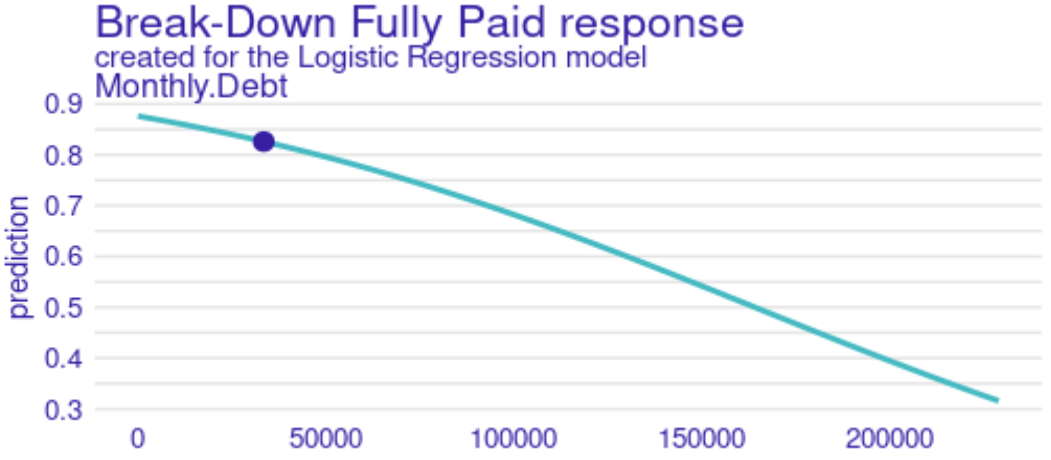

Break-Down for single obs with Fully Paid response
Logistic Regression
intercept
Credit.Score = 691
Current.Loan.Amount = 183900
Monthly.Debt = 33490
Term = Short Term
Annual.Income = 1681000
Purpose = debt consolidation
Years.of.Credit.History = 19.3
Years.in.current.job = 10+ years
Number.of.Open.Accounts = 8
Tax.Liens = 1
+ all other factors
prediction
0.794
+0.045
-0.026
-0.028
+0.026
+0.015
-0.001
+0
+0
+0
+0
+0
0.826
0.7
0.8
0.9

CETERIS PARIBUS PROFILE

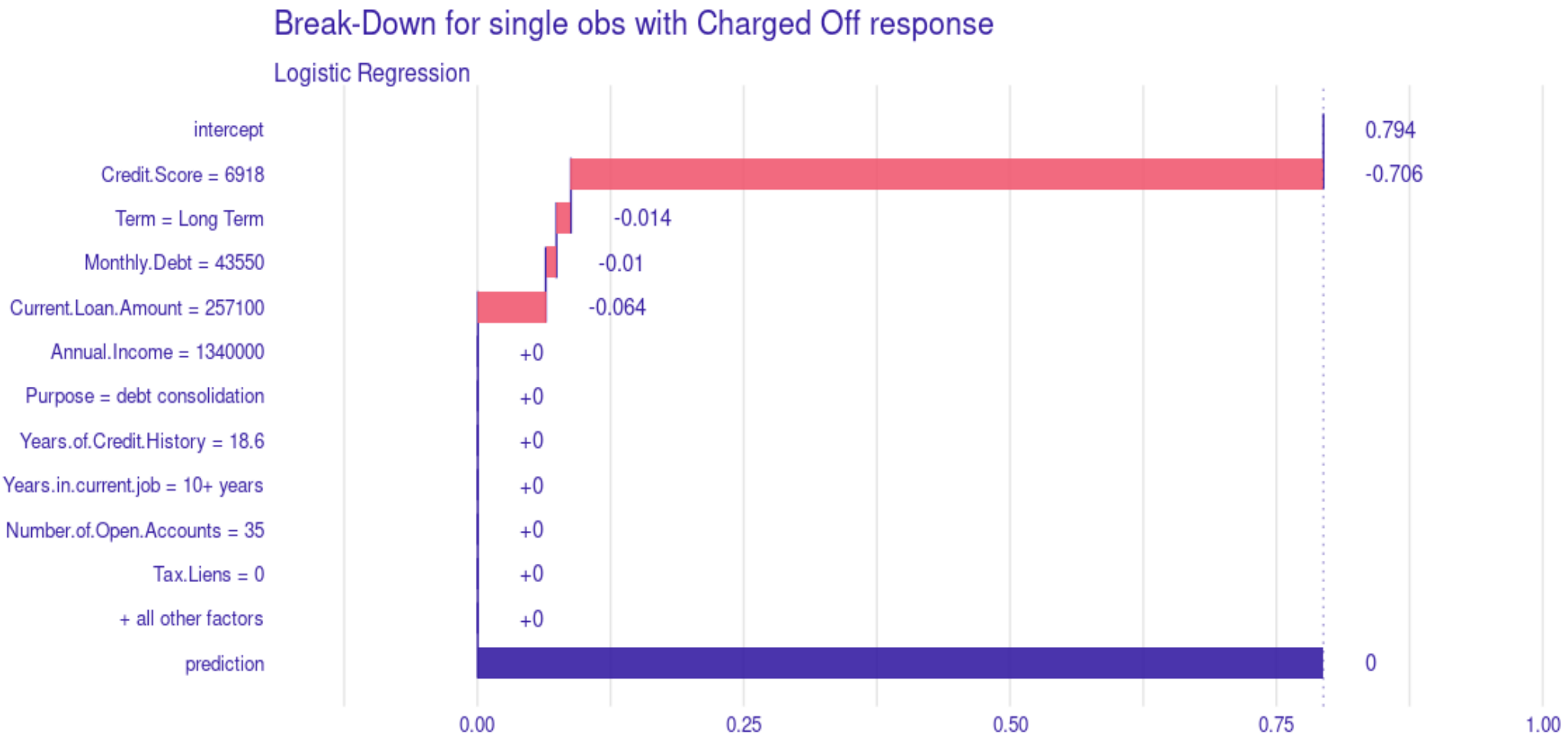




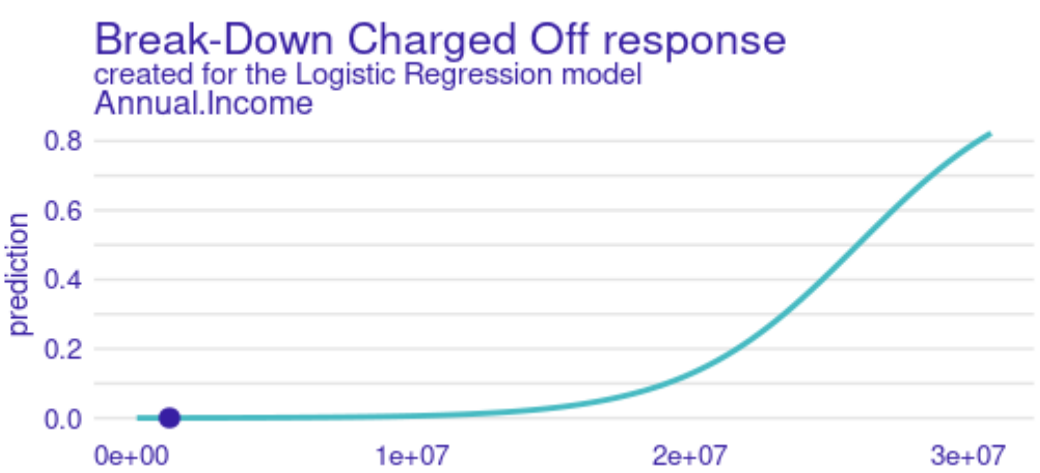


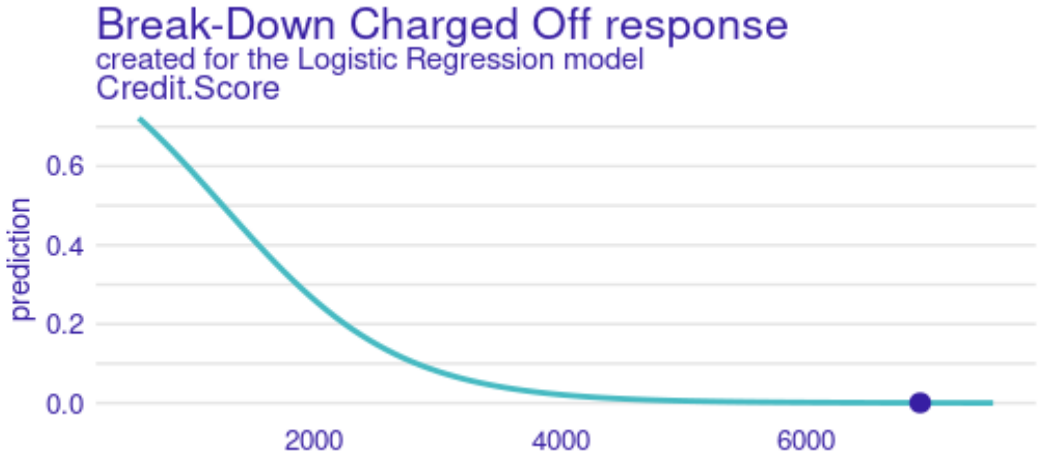


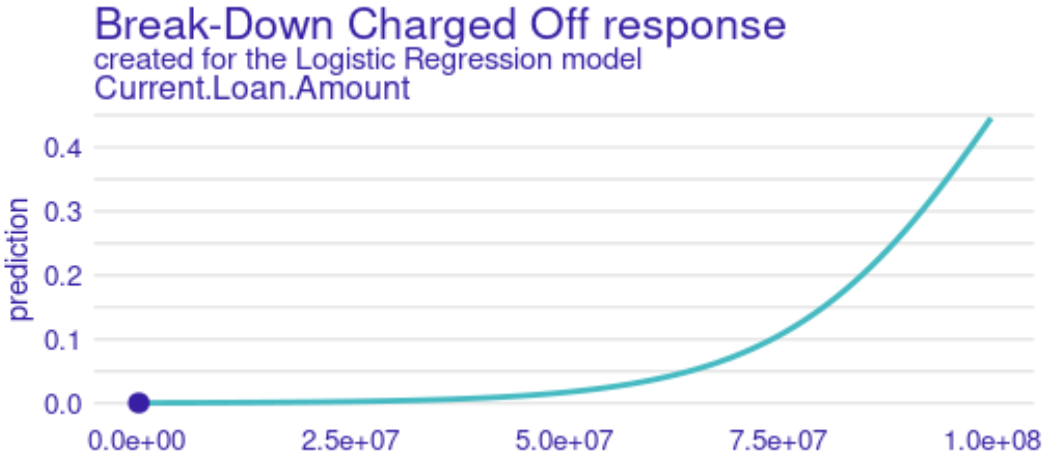
Predicción individual no. 2 (id = 11434)	
Credit.Score	6,918
Current.Loan.Amount	$ 257,100
Monthly.Debt	$ 43,550
Term	Long Term
Anual.Income	$ 1,340,000
Purpose	Debt Consolidations

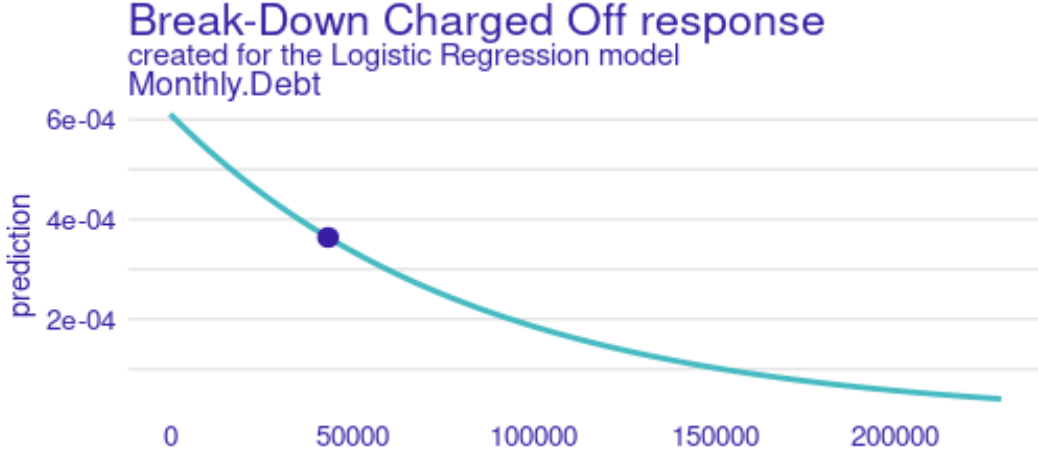

Break-Down for single obs with Charged Off response
Logistic Regression
intercept
Credit.Score = 6918
Term = Long Term
Monthly.Debt = 43550
Current.Loan.Amount = 257100
Annual.Income = 1340000
Purpose = debt consolidation
Years.of.Credit.History = 18.6
Years.in.current.job = 10+ years
Number.of.Open.Accounts = 35
Tax.Liens = 0
+ all other factors
prediction
0.794
-0.706
-0.014
-0.01
-0.064
+0
+0
+0
+0
+0
+0
+0
0
0.00
0.25
0.50
0.75
1.00

CETERIS PARIBUS PROFILE









CONCLUSIONES

Concluimos con un modelo de regresión logística que presenta:

- Una exactitud de 83.8% en la clasificación para el conjunto de entrenamiento, clasificando el 100% de los casos en donde la variable respuesta adopta el valor de “Fully Paid” y un 82.9% de los casos en donde la variable adopta el valor de “Charged Off”.
- Una media estimada del 76.26% de aciertos o registros correctamente clasificados, esto mediante una simulación por bootstraping de 10 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 fueron exitosos para un modelo más generalizable

AREAS DE OPORTUNIDAD:

Se considera que:

- Un EDA más extenso, que incluya un análisis multivariado profundo, así como el empleo de tecnicas 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 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, Bootstraping. https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap

Link al repositorio del proyecto en github:
https://github.com/alexisrangelcalvo/FP-LM-MAEA-UJED-1S