Universidad Torcuato Di Tella

Métodos Estadísticos Aplicados a Negocios

Trabajo Práctico 2

Agustín Alba Chicar
ag.albachicar@gmail.com

Joaquín Gonzalez
joagonzalez@gmail.com

6 de mayo de 2019

Índice

1. Instrucciones

El trabajo puede ser grupal (con un máximo de 4 alumnos). Deberán entregar un informe final a través del campus virtual en el cual adjuntarán el Script en `R` para poder replicar sus resultados.

Tienen tiempo para entregar el TP hasta el lunes s 6 de mayo a las 23:55hs. Una entrega tardía implica desaprobar el trabajo práctico.

2. Consignas

Lea detenidamente el caso "Predicting Customer Churn at QWE Inc." de A. Ovchinnikov (2013). Utilice la base de datos adjunta al caso para contestar las siguientes preguntas:

(1) '**(5 puntos)** Describa la base de datos y obtenga los principales estadísticos descriptivos. Comente.

(2) '**(5 puntos)** En forma aleatoria seleccione el 80 % del total de los casos. Dicha submuestra (*training set*) será utilizada para estimar distintos modelos, mientras que el 20 % restante (*testing set*) será utilizado para evaluar los pronósticos. Vuelva a calcular los estadísticos descriptivos para cada sub-muestra, ¿hay diferencias significativas? Comente.

(3) '**(10 puntos)** Utilizando el *training set*, estime un modelo lineal de probabilidad (MLP) que identifique cuáles son los factores que condicionan la probabilididad de que un cliente cancele su contrato (incluya la antigüedad del cliente). Interprete.

(4) '**(20 puntos)** Para la misma muestra y utilizando los mismos regresores, estime un modelo probit y otro logit. Interprete a partir de los odd-ratios en el caso logit.

(5) '**(10 puntos)** A partir de los modelos estimados en el punto anterior, grafique cómo cambia la probabilidad de cancelar el contrato en función de la antigüedad del cliente. Comente.

(6) '**(10 puntos)** Tanto para el modelo probit como para el logit compute la proporción correctamente estimada. Esto es, se considera que la proporción de clientes que cancelan el contrato es correctamente estimada si `churn=1` y la probabilidad estimada del modelo es superior al 50 % o si `churn=0` y la probabilidad estimada del modelo es inferior a 50 %. Comente

(7) '**(10 puntos)** Usando el *testing set*, calcule el error de pronóstico de cada modelo (MLP, Probit y Logit). Obtenga sus principales estadísticos descriptivos. Comente.

(8) '**(10 puntos)** En función de los errores de pronósticos obtenidos en el punto anterior evalúe los pronósticos calculando el RMSE y MAPE en cada caso. Comente.

(9) '**(10 puntos)** Para evaluar la presencia de sesgos sistemáticos en los pronósticos, regrese los errores de pronósticos en función de una constante y evalúe su significatividad. Comente

(10) '**¡Vamos por el 10! (10 puntos)** En forma análoga al cálculo de la proporción correctamente estimada de cada modelo que realizó en el punto 6, esta vez trabaje con el *testing set* y compute la proporción de falsos positivos y falsos negativos. Llamamos falsos positivos a los casos en los que el modelo predecía que el cliente iba a cancelar el contrato cuando en realidad no lo hizo, y llamamos falsos negativos a los casos en los que el modelo predecía que el cliente no iba a cancelar el contrato cuando en realidad sí lo hizo. Comente.

3. Desarrollo

3.1. Consigna 1

3.1.1. Descripción de la base de datos

La base de datos permite al equipo de ventas de QWE analizar la sus clientes para poder trabajar en la fidelización del grupo de riesgo. Riesgo que los mismos cancelen el contrato. Para ello se registran una serie de elementos sobre un grupo de clientes durante Noviembre y Diciembre de 2011. Luego, en Febrero de 2012 se evalúo si los mismos habían continuado con sus contratos o los habían cancelado [**?**].

Listamos y comentamos las columnas de la base:

- *ID*: es el identificador del cliente.
- *Customer_Age*: es la edad en meses del cliente en la plataforma.
- *Churn*: es una variable categórica que permite determinar cuando un cliente finalizó un contrato ("1") o no ("0").
- *CHI_Score_Month*: *CHI* significa *Curtomer Happiness Index* y es una medida del grado de satisfacción del cliente. Esta columna denota el valor del índice en el mes inicial de la muestra.
- *CHI_Score*: es la diferencia en el segundo mes del *CHI* respecto a *CHI_Score_Month*.
- *Support_Cases_Month*: son la cantidad de casos de soporte que recibió QWE Inc. de dicho cliente en el mes inicial de la muestra.
- *Support_Cases*: son la diferencias de número de casos de soporte (*Support_Cases_Month*) en el segundo mes de medición.
- *SP_Month*: es la media de *Support Priority* en los casos de soporte en mes inicial.
- *SP*: es la variación de media prioridad de los soportes medida en el segundo mes respecto al primero.
- *Logins*: es el número de inicios de sesión al sistema.
- *Blog_Articles*: es el número de artículos escritos.
- *Views*: es el número de vistas al sistema.
- *Days_Since_Last_Login*: es el número de días desde el último inicio de sesión.

Como se puede apreciar, la información está compuesta por tres grandes grupos: información de cliente, información de soporte e información de uso. Con esta información se pretende generar un modelo que permita determinar la probabilidad que un cliente cancele un contrato. La base presenta 6347 entradas.

3.1.2. Descripción de los estadísticos principales

Edad de los clientes Se presenta la edad media de los clientes medida en meses a continuación:

$$\bar{Edad} = 13,9 meses\,\%$$

$$s_{Edad} = 11,16 meses$$

$$IC_{Edad,95\,\%} = [13,62; 14,17] meses$$

Retención La medición de retención se realiza directamente con el estimador de proporción sobre la columna *Churn* y se obtiene:

$$\hat{p}_{retenidos} = 94,91\,\%$$

$$s_{retenidos} = 0,28\,\%$$

$$IC_{p_{retenidos},95\,\%} = [94,34; 95,43]\,\%$$

CHI La información de *CHI* se tiene para aproximadamente un 80 % de los clientes. Para el resto, simplemente se tiene un cero. Esto genera un sesgo en la media, no menor, al momento de analizar la información. Se muestra a continuación (ver tabla 1) como son las medias antes y después (medición inicial y final), con y sin el cero.

-	Todas las muestras	Discriminando el cero
Inicial	87,32	107,53
Final	92,37	111,15

Cuadro 1: Valores de media de CHI.

A su vez, en las imágenes 1 y 2 se muestran como se distribuyen los índices de CHI. Notar que las líneas verticales denotan los valores de media en la tabla 1. Con rojo se muestran las medias contando al cero y con verde sin contar las muestras que son cero.

Casos de soporte En la figura 3 se muestra la distribución de casos de soporte inicial y final. Las líneas punteadas muestran donde se ubica la media de cada una de las series que se resume en la tabla 2.

Momento	Media	Mínimo IC	Máximo IC
Inicial	0,70	0,66	0,75
Final	0,70	0,62	0,78

Cuadro 2: Casos de soporte.

Una particularidad es que existen valores negativos de caso de soporte finales, esto es porque las mediciones son relativas al momento inicial, lo que indica una caída simplemente en los casos de soporte atendidos.

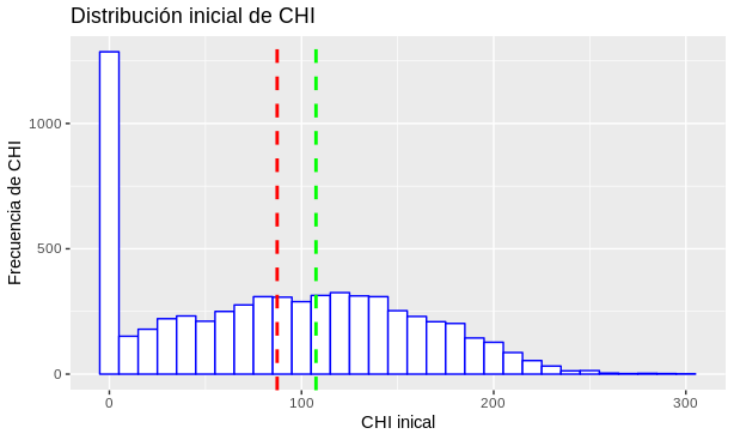

Figura 1: Histograma del índice CHI en la medición inicial.

Prioridad de los soportes Los datos de uso por parte de los usuarios los resumimos en la tabla 3.

Características	Media	Mínimo IC	Máximo IC
Logins	15,73	14,69	16,76
Blog Articles	0,157	0,042	0.271
Views	96,31	18,74	173,87
Days since last login	1,76	1,32	2,21

Cuadro 3: Información de uso.

Nota: todos los intervalos de confianza han sido calculados al 95 %.

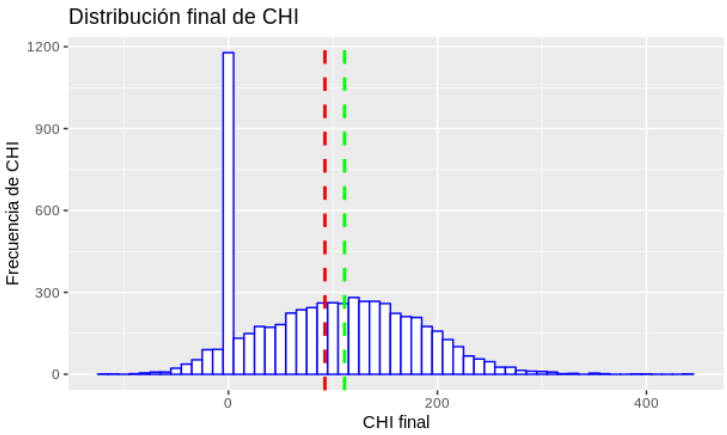

Figura 2: Histograma del índice CHI en la medición final.

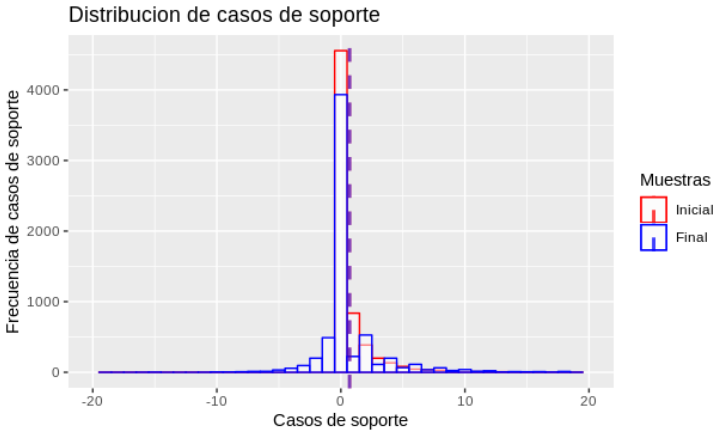

Figura 3: Histograma de casos de soporte.

3.2. Consigna 2

Se solicita realizar un `training set` y un `testing set` compuestos por el 80 % y el 20 % de las muestras respectivamente. Para ello, se samplea el vector de índices logrando así el vector de entrenamiento y luego se toman los elementos restantes para el de evaluación. Acto seguido, se computan los mismos estimadores característicos en prueba de hipótesis para probar que son iguales a los de la muestra entera. El ensayo es escencialemente el mismo código salvo que las muestras pertencen al `training set`.

La tabla 4 muestra los valores de las medias obtenidas con el `training set`, las medias de la muestra completa y *p-valor* de la prueba de hipótesis con un intervalo de confianza al 95 %. En casi ningún caso hay evidencia suficiente para decir que no son iguales.

Característica	Media muestral	Media del `training set`	*p-valor*
Edad de los clientes[meses]	13,9	13,86	0,81
Retención [%]	94,91	94,66	0,42
CHI inicial [u]	87,31	84,42	0,94
CHI final [u]	92,37	92,53	0,89
TODO SP [u]			
Logins [u]	15,72	15,60	0,83
Blog Articles [u]	0,157	0,195	0,578
Views [u]	96,31	55,02	0,022
Days since last login [días]	1,76	1,86	0,68

Cuadro 4: Comparación entre `training set` y muestra.

Como se puede ver la tabla 4 el campo *Views* posee una gran dispersión, y cae por aproximadamente dos unidades fuera del intervalo de confianza ($p-valor < 0,025$).

3.3. Consigna 3

Se solicita realizar un modelo lineal de probabilidad (MLP) que estime la probabilidad de que un contrato sea cancelado teniendo en cuenta al menos la edad del cliente.
El primer acercamiento al problema tomado es hacer hacer un modelo que incluya todas las variables disponibles en la base, aun sabiendo que muchas de ellas no poseen datos y tal vez los mismos no tienen relación alguna con la variable. El modelo será ajustado con el *training set*. Tras la ejecución en `R` de los modelos, obtenemos los coeficientes que se muestran en la tabla 16. Llamamos modelo A al modelo que utiliza todas las variables de regresión y modelo B a aquel que sólo presenta coeficientes con una significancia al 95 % ó más.

Se estima usando SE robustos a heterocedasticidad, corrección necesaria para modelos lm.

Coeficientes	MLP A	MLP B
Intercepto	0.0582424150 (***) 0.0057501548	0.05764147250 (***) 0.00502058583
Curstomer_Age	0.0008684575 (**) 0.0003007570	0.00093860826 (**) 0.00027496819
CHI_Score_Month	-0.0002313576 (***) 0.0000586953	-0.00024882702 (***) 0.00004843773
CHI_Score	-0.0003639595 (**) 0.0001142282	- -
Support_Cases_Month	-0.0025915872 () 0.0028502697	- -
Support_Cases	0.0034365895 () 0.0023668773	- -
SP_Month	-0.0002902072 () 0.0040970350	- -
SP	-0.0010982449 () 0.0031889886	- -
Logins	0.0000777291 () 0.0000851680	- -
Blog_Articles	0.0000475717 () 0.0006270186	- -
Views	-0.0000003086 () 0.0.0000008792	- -
Days_Since_Last_Login	0.0006292039 (***) 0.0001667353	0.0009964122 (***) 0.0001982972

Cuadro 5: MLP coeficientes de los modelos A y B.

Comparamos en la tabla 6 los estadísticos más representativos de los modelos. Vemos que la performance es bastante similar uno a otro, tanto en SER como en R_a^2 para una cantidad de grados de libertad bastante similar. Sin em-

Estadísticos	MLP A	MLP B
SER	0.2232	0.2232
Grados de libertad	5065	5071
R^2	0.01613	0.01482
R_a^2	0.014	0.01385
F	7.55	15.26

Cuadro 6: Estadísticos de los modelos A y B.

bargo, cualquiera de los dos modelos son deficientes. Ninguno permite estimar de forma correcta casos con alta probabilidad de rechazo, puesto que dentro del *training set*, aproximadamente el 5 % de los contratos son cancelados, lo cual hace que el MLP no performe correctamente para modelar este escenario. Las figuras 7 y 12 muestran como ambos modelos no tienen capacidad de discriminar un escenario del otro al haber, escencialemente, superposición en los intervalos de confianza de las probabilidades estimadas.

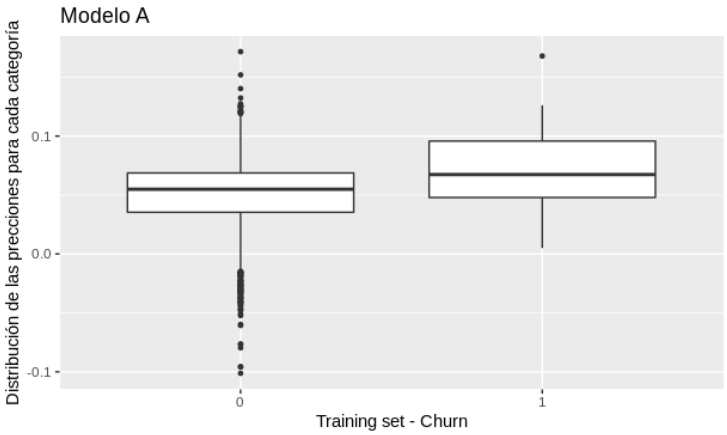

Figura 4: Boxplots de la distribución de probabilidades estimadas para cada caso con el modelo A.

Estos problemas se pueden entender por las siguientes limitaciones de los modelos lineales para este tipo de problemas:

- No se garantiza que la probabilidad este entre 0 y 1
- El efecto de X en p(y=1—beta) es SIEMPRE LINEAL
- El error del modelo es heterocedástico, no es constante para distintos valores de X, entonces debe calcularse con errores estándares robustos

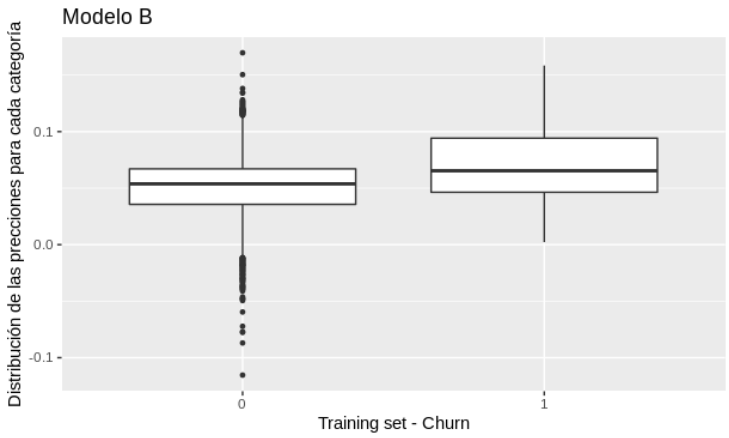

Figura 5: Boxplots de la distribución de probabilidades estimadas para cada caso con el modelo B.

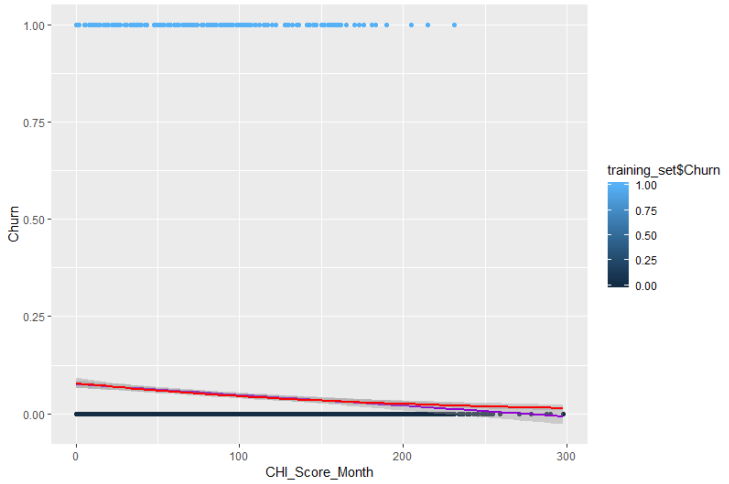

Figura 6: Modelo lineal versus modelo no lineal para CHI Score Month.

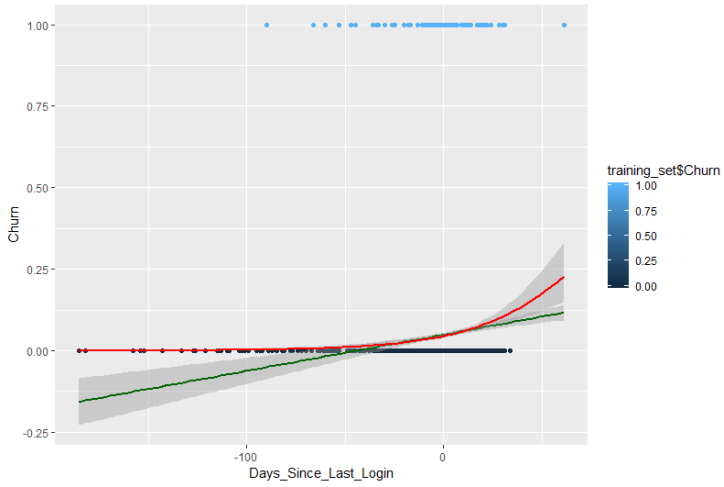

Figura 7: Modelo lineal versus modelo no lineal para Days Since Last Login.

3.4. Consigna 4

Se solicita realizar, para la misma muestra y mismos regresores, un modelo no lineal de probabilidad probit y otro logit. Además, realizar un análisis, a partir de los Odds-ratios (OR), en el caso del modelo logit.

Es importante mencionar que para analizar el efecto marginal de los regresores en estos modelos deberá compararse caso por caso ya que su variación no es lineal como se puede observar en las figuras 6 y 7 de la consigna 3. Además, los regresores no serán un indicador directo de probabilidad, sino indicadores de como cambia el logaritmo de la razon de chances $\frac{p}{1-p}$. Esto se puede observar a través de las expresiones de la función logit.

$$P(churn = 0|X_1, ..., X_n) = \beta_0 + ... + \beta_n X_n \tag{1}$$

Si analizamos para una sola variable independiente, tenemos

$$P(X) = \beta_0 + \beta_1 X \tag{2}$$

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X \tag{3}$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X} \tag{4}$$

$$p = e^{\beta_0+\beta_1 X}(1-p) \tag{5}$$

$$p = e^{\beta_0+\beta_1 X} - pe^{\beta_0+\beta_1 X} \tag{6}$$

$$p + pe^{\beta_0+\beta_1 X} = e^{\beta_0+\beta_1 X} \tag{7}$$

$$p(1 + e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X} \tag{8}$$

$$p = \frac{e^{\beta_0+\beta_1 X}}{(1 + e^{\beta_0+\beta_1 X})} \tag{9}$$

Podemos analizar los signos de los regresores de la tabla 7. Observamos que *Customer_Age* y *Days_Since_Last_Login* son los únicos con signo positivo. Esto quiere decir que ante un aumento en una de estas variables independientes el logaritmo de la razón de chances aumentará, siendo mas probable que churn sea 1, es decir, que el cliente cancele la subscripción. Intuitivamente, estos valores tienen sentido ya que para el caso de *Days_Since_Last_Login* mientras más tiempo pasa un usuario sin loguearse a la plataforma, quiere decir que la está utilizando poco ya sea por que ha dejado de satisfacer sus necesidades o por que esta teniendo problemas. Con *Customer_Age*, podría explicarse que la probabilidad de dejar la plataforma aumenta mientras mas tiempo pasa en ella por el hecho de que aumenta el tiempo en el que no la ha dejado. El resto de los regresores tienen signo negativo, por lo que generan el efecto contrario en la razón de chances.

Lamentablemente, solo podemos realizar un análisis cualitativo en este caso. Por esta razón es que trabajaremos con los *Odd_Ratio* de los modelos probit y

Coeficientes	Logit	Probit
Intercepto	-2.81087146 (***) 0.11927494	-1.57884113 (***) 0.05539772
Curstomer_Age	0.01259426 (***) 0.00591806	0.00732904 (**) 0.00279750
CHI_Score_Month	-0.00498508 (***) 0.00119898	-0.00242609 (***) 0.00053988
CHI_Score	-0.00891517 (***) 0.00251761	-0.00395046 (***) 0.00116718
Views	-0.00008081 () 0.00005401	-0.00004273 (***) 0.00002659
Days_Since_Last_Login	0.02174173 (***) 0.00498183	0.00886789 (***) 0.00227636

Cuadro 7: Modelos Logit y Probit.

logit, enfoque que nos permitirá realizar un análisis cuantitativo respecto de la variación del módulo de los regresores.

En la tabla 8 observamos los OR para el modelo logit calculado.

Coeficientes	OR	Intervalo de confianza (97.5 %)
Intercepto	0.06015255	[0.04732941, 0.0755672]
Curstomer_Age	1.01267390	[1.00070787, 1.0242222]
CHI_Score_Month	0.99502732	[0.99266877, 0.9973487]
CHI_Score	0.99112445	[0.98623565, 0.9960089]
Views	0.99991919	[0.99982319, 1.0000134]
Days_Since_Last_Login	1.02197980	[1.01235425, 1.0322408]

Cuadro 8: Odd-Ratios modelo logit.

Valores de $OR > 1$ indican un aumento en la razon de ocurrencia de la variable dependiente cuando aumenta en una unidad la variable independiente. Un $OR < 1$ indica lo contrario, mientras que si es igual a 1 implica que el factor explicativo no produce efectos. En este caso observamos que tenemos $OR \approx 1$ por lo que, si bien las variables son estadísticamente significativas (intervalos de confianza no incluyen al 1), no generan un gran efecto sobre la probabilidad de $Y = churn$. Se observa consistencia respecto al summary(logit) realizado anteriormente ya que los $OR > 1$ son los que indican una probabilidad mas alta de $churn = 1$.

3.5. Consigna 5

Se solicita, a partir de los modelos estimados en el punto anterior, graficar como cambia la probabilidad de cancelar el contrato en función de la antigüedad del cliente.

$$P(churn = 0|X_1, ..., X_n) = \beta_0 + ... + \beta_n X_n \tag{10}$$

Para un rango de $[1 : 1500]$ se estima la probabilidad de cancelar el contrato en función de la edad del cliente. Si bien el máximo de $Customer\_Age$ es mucho menor, se usa un rango mayor para poder visualizar la tendencia.Se observa la tendencia esperada según los valores de los regresores calculados anteriormente.

Figura 8: Probabilidad de cancelar contrato en funcion de Customer Age.

Figura 9: Probabilidad de cancelar contrato en funcion de Customer Age.

3.6. Consigna 6

Se solicita, a partir de los modelos estimados en el punto anterior, computar la proporción correctamente estimada.

Las predicciones del modelo son realizadas sobre el data frame de testing y comparado con las observaciones.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	Verdadero Positivos (VP)	Falsos Negativos (FN)
Negativos (observación)	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Cuadro 9: Matriz de Confusión.

De la matriz, podemos definir la proporción correctamente (PCE) estimada como:

$$PCE = \frac{VP + VN}{N} \tag{11}$$

Este indicador nos dará información sobre que porcentaje de los datos ha sido calificado correctamente por los modelos.

La matriz de confusión del modelo logit se observa en la tabla 10, la del modelo probit en la tabla 11.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	0	256
Negativos (observación)	0	4821

Cuadro 10: Matriz de Confusión modelo logit.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	0	116
Negativos (observación)	0	2804

Cuadro 11: Matriz de Confusión modelo probit.

Las matrices calculadas no varían entre modelos.

Se observa un problema en las detección de cancelaciones por parte de ambos modelos, no se han podido predecir correctamente las mismas. Podemos afirmar que los modelos están sesgados para la clasificación de no cancelación.

Estos problemas pueden ser causa de un error en en análisis o en el cálculo del modelo. También, podrían estar relacionados a la naturaleza de la muestra elegida para trabajar. A pesar de estos errores en la estimación, la proporción correctamente estimada es muy alta, esto se debe a que hay una proporción de cancelaciones muy baja respecto del volumen de la muestra.

$$PCE_{logit} = \frac{VP + VN}{N} = 0{,}9495765 \tag{12}$$

$$PCE_{probit} = \frac{VP + VN}{N} = 0{,}9495765 \tag{13}$$

3.7. Consigna 7

Se pide, utilizando el testing set, calcular el error de pronótisco para los modelo MLP, probit y logit respectivamente.

Para cada caso, se calculara el error de pronóstico para cada observación, así como el error medio y la varianza del error medio.

$$Error\_de\_pronóstico = y_i - \hat{y}_i \tag{14}$$

$$ME = \frac{1}{n} \cdot \sum e_i \tag{15}$$

$$EV = \frac{1}{n} \cdot \sum (e_i - ME)^2 \tag{16}$$

Se observa que la magnitud del error es altamente significativa para los falsos negativos, esto es especialmente claro en las figuras 10, 11 y 12. El sesgo se mantiene independientemente del modelo utilizado. Errores medios bajos y varianzas grandes respecto de esa media (50 %)

	Error Medio	**Varianza del error**
MLP	0.09516783	0.05630573
probit	0.08509942	0.04781444
logit	0.08510169	0.04777062

Cuadro 12: Matriz de errores medios y varianzas de modelos MLP, probit, logit.

Figura 10: Error de estimación respecto de valor observado en modelo MLP.

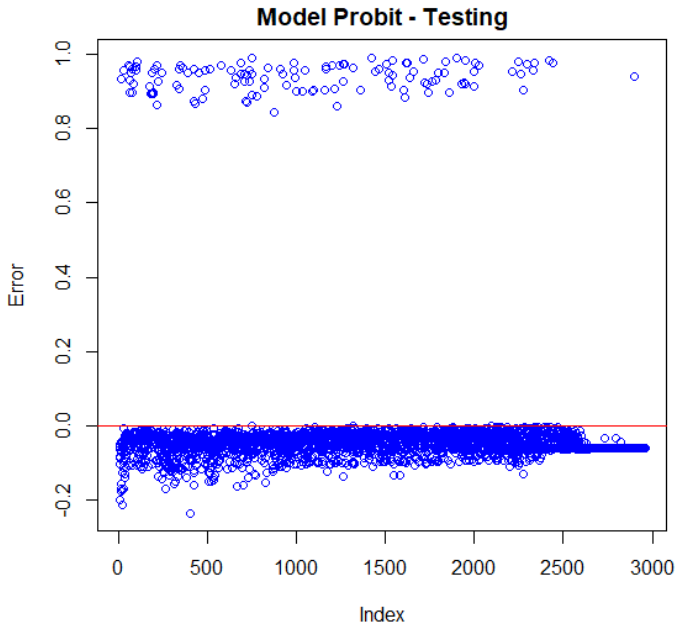

Figura 11: Error de estimación respecto de valor observado en modelo probit.

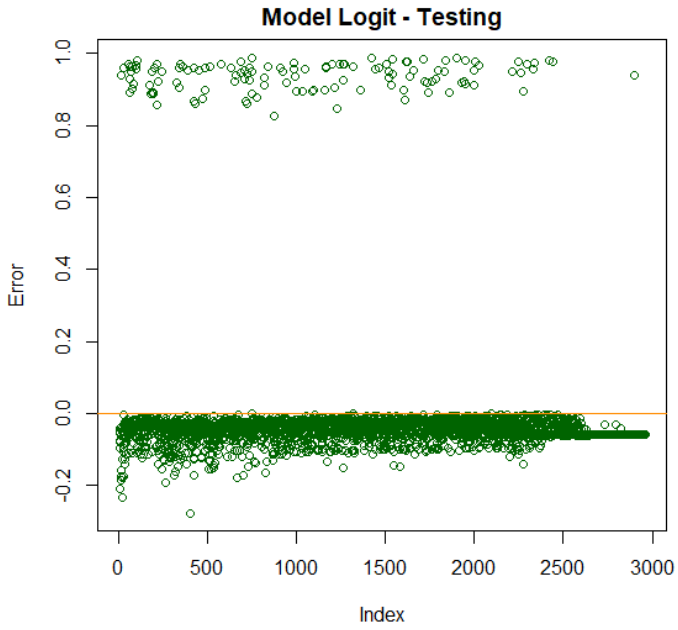

Figura 12: Error de estimación respecto de valor observado en modelo logit.

3.8. Consigna 8

Se pide, en función de los errores de pronósticos obtenidos en la consigna 7, evaluar los pronósticos calculando RMSE y MAPE en cada modelo.

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum e_i 2} \tag{17}$$

$$MAPE = \sqrt{\frac{1}{n} \cdot \sum \frac{|e_i|}{y_i}} \cdot 100 \tag{18}$$

En el caso particular del MAPE obtenemos resultados inconsistentes ya que la variable dependiente y que se intenta explicar toma valores 0 y 1. Esto hace que el error medio absoluto porcentual tienda a infinito debido a la división por cero.

	RMSE	**MAPE**
MLP	0.2173679	∞
probit	0.1987087	∞
logit	0.198537	∞

Cuadro 13: Matriz de errores medios y varianzas de modelos MLP, probit, logit.

RMSE es la raíz cuadrada de la varianza de los residuos. Es una medida de presición de los pronósticos del modelo que depende de la escala del error. Mientras mas bajo sea, mejores serán los pronósticos realizados.

El RMSE más bajo de los modelos analizados es el $RMSE_{logit}$.

3.9. Consigna 10

Se solicita, a partir de los modelos estimados al comienzo del análisis, computar la proporción falsos positivos y falsos negativos sobre el testing set.

Se utiliza el siguiente algoritmo para hallar las proporciones solicitadas:

```
# Corremos predicciones sobre sobre testing_set
p_logit_testing <- predict(
    model_logit,
    newdata = testing_set,
    type="response")

p_probit_testing <- predict(
    model_probit,
    newdata = testing_set,
    type="response")

names(testing_set)
summary(testing_set)

############################
#      MODELO LOGIT        #
############################

# Calculamos y Clasificamos las predicciones
predicted_value_logit <- p_logit_testing
predicted_class_logit <- ifelse(predicted_value_logit>0.5,"Yes", "No")
# Usamos el valor que predice el modelo para verificar
predicted_class_logit_tbshoot <- ifelse(
    predicted_value_logit>0.5,
    sprintf("Yes %f",predicted_value_logit),
    sprintf("No %f",predicted_value_logit))

# Generamos dataframe de observado vs predicciones
# en el conjunto testing
performance_data_logit <- data.frame(
    observed=testing_set$Churn,
    predicted=predicted_class_logit)
performance_data_logit_tbshoot <- data.frame(
    observed=testing_set$Churn,
    predicted=predicted_class_logit_tbshoot)

error_pred_logit <- data.frame(
    observed=testing_set$Churn,
    predicted=predicted_value_logit,
    error=testing_set$Churn-predicted_value_logit)
```

```
total <- length(testing_set$ID)
positive <- sum(performance_data_logit$observed=="1")
negative <-sum(performance_data_logit$observed=="0")
predicted_positive <- sum(performance_data_logit$predicted=="Yes")
predicted_negative <- sum(performance_data_logit$predicted=="No")

# Hallamos los valores de la matriz de confusion
VP <- sum(performance_data_logit$observed=="1" &
     performance_data_logit$predicted=="Yes")
VN <- sum(performance_data_logit$observed=="0" &
     performance_data_logit$predicted=="No")
FP <- sum(performance_data_logit$observed=="0" &
     performance_data_logit$predicted=="Yes")
FN <- sum(performance_data_logit$observed=="1" &
     performance_data_logit$predicted=="No")

PCR_logit <- VP + VN / total
```

El mismo procedimiento realizamos para el modelo probit.

Se detalla la matriz de confusión de los modelos en la tabla 13. Se observa el mismo comportamiento que el detallado en el ejercicio 6.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	Verdadero Positivos (VP)	Falsos Negativos (FN)
Negativos (observación)	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Cuadro 14: Matriz de Confusión.

De la matriz, podemos definir la proporción correctamente (PCE) estimada como:

$$PCE = \frac{VP + VN}{N} \tag{19}$$

Este indicador nos dará información sobre que porcentaje de los datos ha sido calificado correctamente por los modelos.

La matriz de confusión del modelo logit se observa en la tabla 10, la del modelo probit en la tabla 11.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	0	123
Negativos (observación)	0	2794

Cuadro 15: Matriz de Confusión modelo logit.

	Positivos (predicción)	**Negativos (predicción)**
Positivos (observación)	0	123
Negativos (observación)	0	2794

Cuadro 16: Matriz de Confusión modelo probit.

Las matrices calculadas no varían entre modelos. No se manifiestan diferencias significativas a los resultados obtenidos sobre el dataset training del punto 6.

Se observa un problema en las detección de cancelaciones por parte de ambos modelos, no se han podido predecir correctamente las mismas. Podemos afirmar que los modelos están sesgados para la clasificación de no cancelación.

Estos problemas pueden ser causa de un error en en análisis o en el cálculo del modelo. También, podrían estar relacionados a la naturaleza de la muestra elegida para trabajar. A pesar de estos errores en la estimación, la proporción correctamente estimada es muy alta, esto se debe a que hay una proporción de cancelaciones muy baja respecto del volumen de la muestra.

$$PCE_{probit} = PCE_{logit} = \frac{VP + VN}{N} = 0{,}9578334 \tag{20}$$

A. Uso del script en R

Para el correcto uso del script en R adjunto, se deben instalar los archivos que figuran en el encabezado y modificar la variable `path_to_workspace` con la ruta al directorio que contiene el script. Luego, el script asumirá que la base de datos `UV6696-XLS-ENG-Caso-Customer-Churn.xls` se encuentra en el mismo directorio.

El script depende de las siguientes librerías:

- `dplyr`
- `ggplot2`
- `MASS`
- `ISLR`
- `lmtest`
- `AER`
- `DescTools`