Predicción de Abandono

Isidoro Garcia

2021

```
library(tidyverse)
library(data.table)
library(broom)
library(knitr)
library(lubridate)
library(RCT)
library(gamlr)
library(ranger)
library(tree)
library(parallel)
library(tidymodels)
```

Contexto

Cell2Cell es una compañía de teléfonos celulares que intenta mitigar el abandono de sus usuarios. Te contratan para 1) Encontrar un modelo que prediga el abandono con acierto y para usar los insights de este modelo para proponer una estrategia de manejo de abandono.

Las preguntas que contestaremos son:

1. Se puede predecir el abandono con los datos que nos compartieron?
2. Cuáles son las variables que explican en mayor medida el abandono?
3. Qué incentivos da Cell2Cell a sus usarios para prevenir el abandono?
4. Cuál es el valor de una estrategia de prevención de abandono focalizada y cómo difiere entre los segmentos de los usuarios? Qué usuarios deberían de recibir incentivos de prevención? Qué montos de incentivos

Nota: Voy a evaluar las tareas con base en la respuesta a cada pregunta. Como hay algunas preguntas que no tienen una respuesta clara, al final ponderaré de acuerdo al poder predictivo de su modelo vs las respuestas sugeridas.

Datos

Los dotos los pueden encontrar en `Cell2Cell.Rdata`. En el archivo `Cell2Cell-Database-Documentation.xlsx` pueden encontrar documentación de la base de datos.

Cargemos los datos

```
load('Bases input/Cell2Cell.Rdata')
```

1. Qué variables tienen missing values? Toma alguna decisión con los missing values. Justifica tu respuesta

2. Tabula la distribución de la variable `churn`. Muestra la frecuencia absoluta y relativa. Crees que se debe hacer oversampling/undersamping?

3. (2 pts) Divide tu base en entrenamiento y validación (80/20). Además, considera hacer oversampling (SMOTE) o undersampling. (Tip: Recuerda que el objetivo final es tener muestra ~balanceada en el traning set. En el validation la distribución debe ser la original)

Model estimation

Pondremos a competir 3 modelos:

1. Cross-Validated LASSO-logit
2. Prune Trees
3. Random Forest

4 (2 pts). Estima un cross validated LASSO. Muestra el la gráfica de CV Binomial Deviance vs Complejidad

5. Grafica el Lasso de los coeficientes vs la complejidad del modelo.

6 (2 pts). Cuál es la λ resultante? Genera una tabla con los coeficientes que selecciona el CV LASSO. Cuántas variables deja iguales a cero? Cuales son las 3 variables más importantes para predecir el abandono? Da una explicación intuitiva a la última pregunta

7. Genera un data frame (usando el validation set) que tenga: `customer`, `churn` y las predicciones del LASSO.

8. Estima ahora tree. Usa `mindev = 0.05, mincut = 1000` Cuántos nodos terminales salen? Muestra el summary del árbol

9. Grafica el árbol resultante

10. Poda el árbol usando CV. Muestra el resultado. Grafica Tree Size vs Binomial Deviance. Cuál es el mejor tamaño del árbol? Mejora el Error?

11. Gráfica el árbol final. (Tip: Checa `prune.tree`)

12. Genera las predicciones del árbol pruned. Guardalas en la base de predicciones. Guarda el score y la prediccion categorica en la misma data frame donde guardaste las predicciones del LASSO

13 (4pts). Corre un Random Forest ahora. Cuál es la B para la que ya no ganamos mucho más en poder predictivo?

- Corre para `num.trees=100,200,300, 500, 700, 800`
- En cada caso, guarda únicamente el `prediction.error`

14. Escoge un random forest para hacer las predicciones. Grafica la importancia de las variables. Interpreta

15. Genera las predicciones OOS para el random forest. Guardalas en la misma data.frame que los otros modelos

16 (2pts). Corre el mismo forest pero ahora con `probability = T`. Esto generará predicciones númericas en lugar de categóricas. Genera las predicciones continuas y guardalas en el mismo data frame

17 (4 pts). Genera graficas de las curvas ROC para los tres modelos. Cual parece ser mejor?

18. Genera una tabla con el AUC ROC. Cuál es el mejor modelo ?

19 (2pts). Escoge un punto de corte para generar predicciones categoricas para el LASSO basado en la Curva ROC. Genera las matrices de confusión para cada modelo. Compáralas. Qué tipo de error es mas pernicioso?

20 (2pts). Finalmente, construye una lift table. Esto es, para 20 grupos del score predecido, genera 1) El promedio de las predicciones, 2) el promedio del churn observado. Existe monotonía? El mejor algoritmo es monotónico? (Tip: usa `ntile` para generar los grupos a partir de las predicciones)

21. Concluye. Que estrategia harías con este modelo? Cómo generarías valor a partir de el?