Modelamiento Estadístico y Sistemas Recomendadores: Foro 4

Patricio Águila Márquez

Instrucciones

Considere los datos de una empresa dedicada a la venta de paquetes de viajes a destinos exclusivos en Europa. La empresa tiene información histórica de 704 clientes, a lo largo de tres países de américa latina, y le interesa realizar recomendaciones para un grupo selecto de clientes, a los cuales se desea fidelizar.

La empresa cuenta con destinos en tres ciudades: Madrid, Londres y París. De estas tres ubicaciones, Londres y París corresponden a destinos recientemente incorporados, mientras que Madrid corresponde al primer destino en la cual la empresa comenzó su operación. La empresa cuenta con cierta información personal sobre los clientes. Adicionalmente, ha registrado el motivo más usual del viaje, y si han viajado a su locación original (Madrid) en el pasado.

Recientemente, la empresa ha decidido lanzar un programa de fidelización para un grupo selecto de clientes poco activos en su plataforma web, en el cual desea plantear atractivas ofertas a los dos nuevos destinos exclusivos ubicados en las ciudades de Londres y París. Para ello, la empresa ha decidido desarrollar un sistema recomendador que le permita maximizar el interés de estos clientes en su oferta. La información de estos clientes se encuentra en el archivo "fidelización.csv".

La empresa no sabe a priori si estos clientes tienen interés o no en los destinos. Sin embargo, tiene acceso a los registros de comportamiento web de los clientes presentes en el conjunto de datos "travel.csv", en particular, si han cotizado o no los destinos a Londres y París.

El objetivo de este Foro consiste en construir dos sistemas recomendadores, basados en filtrado colaborativo (modelando el problema de recomendación como uno de clasificación), que le permitan a la empresa realizar recomendaciones efectivas al grupo selecto de clientes.

Las variables que se encuentran en el conjunto de datos se describen en la siguiente tabla:

Variable	**Descripción**
Gender	Género ("male" o "female").
Children	Indica 1 si viaja con hijos, 0 si no.
Country	País de residencia (Argentina, Perú o Chile).
Motive	Indica "travel" si el motivo del viaje es turismo, y "business" si está asociado a negocios.
Age	Edad.
Madrid	Indica 1 si el año anterior viajó a madrid, 0 si no.
London	Indica 1 si manifestó interés en viajar a Londres, 0 si no.
París	Indica 1 si manifestó interés en viajar a París, 0 si no.

Para lograr lo anterior, desarrolle las siguientes actividades:

1. Responda y argumente en el Foro. ¿Por qué es posible interpretar el sistema recomendador deseado como un problema de clasificación binaria?.

- *Resp: se puede interpretar como un problema de clasificación binario, debido a que las variables de interés del ejercicio presentan solo dos estados (viajó/no viajó, tiene interés/no tiene interés).*

2. Cargue el conjunto de datos en la sesión de trabajo de `R` usando la función `read.table`.

```
# Carga de datos de archivo "travel.csv"
datos <- read.csv("../04 Foro 4/travel.csv",header=TRUE, sep=",")
```

3. Separe los datos en *datos.paris* y *datos.londres*. El objetivo de esta separación es desarrollar un sistema recomendador para cada destino, basado en la data de cada usuario, si aprovecharon o no la oferta a madrid el año anterior, y si manifiestan interés en viajar a la ciudad en cuestión.

- *Resumen de los datos: París*

```
##     gender    children      country          motive          marital_status
##  female:226   0:188    argentina:285   business :101   married      :484
##  male  :478   1:516    chile    :126   bussiness:110   single       :190
##                        peru     :293   travel   :493   widow/widower: 30
##
##
##
##       age        madrid  paris
##  Min.   :20.00   0:229   0:243
##  1st Qu.:34.00   1:475   1:461
##  Median :41.00
##  Mean   :41.24
##  3rd Qu.:48.00
##  Max.   :65.00
```

- *Resumen de los datos: Londres*

```
##     gender    children      country          motive          marital_status
##  female:226   0:188    argentina:285   business :101   married      :484
##  male  :478   1:516    chile    :126   bussiness:110   single       :190
##                        peru     :293   travel   :493   widow/widower: 30
##
##
##
##       age        madrid  london
##  Min.   :20.00   0:229   0:361
##  1st Qu.:34.00   1:475   1:343
##  Median :41.00
##  Mean   :41.24
##  3rd Qu.:48.00
##  Max.   :65.00
```

4. Desarrolle un sistema recomendador que le permita determinar si debe o no ofrecer una oferta que tenga como destino la ciudad de París. Para ello:

a. Seleccione de manera aleatoria 70% de las observaciones para crear sus datos de entrenamiento y guarde el 30% restante para objeto de validación, tal como lo hizo en el Foro 2.

- *Conjunto de entrenamiento (70% de los datos de París)*

```
##     gender    children      country          motive          marital_status
##  female:155   0:130    argentina:190   business : 69   married      :343
##  male  :337   1:362    chile    : 86   bussiness: 77   single       :129
##                        peru     :216   travel   :346   widow/widower: 20
##
##
##
##       age         madrid  paris
##  Min.   :20.00   0:158   0:173
##  1st Qu.:34.00   1:334   1:319
##  Median :41.00
##  Mean   :41.68
##  3rd Qu.:49.00
##  Max.   :65.00
```

- *Conjunto de validación (30% de los datos de París)*

```
##     gender    children      country         motive          marital_status
##  female: 71   0: 58    argentina:95   business : 32   married      :141
##  male  :141   1:154    chile    :40   bussiness: 33   single       : 61
##                        peru     :77   travel   :147   widow/widower: 10
##
##
##
##       age         madrid  paris
##  Min.   :20.00   0: 71   0: 70
##  1st Qu.:33.00   1:141   1:142
##  Median :38.50
##  Mean   :40.21
##  3rd Qu.:47.00
##  Max.   :65.00
```

b. Entrene al menos 3 de los algoritmos de clasificación vistos en clase, que tengan como variable objetivo el interés en viajar a parís.

```
# Entrenamos modelos de clasificación

# Árbol de decisión
recomendador.tree.paris <- rpart(paris ~ ., data=datos.trabajo.paris,
                                 parms = list(split = "gini"))

# Bayes Ingenuo
recomendador.nb.paris <- naiveBayes(paris ~ .,data=datos.trabajo.paris)

# Bagging
recomendador.bagging.paris <- bagging(paris ~., data=datos.trabajo.paris,
                                      mfinal=10)

# Boosting
recomendador.boosting.paris <- boosting(paris ~., data=datos.trabajo.paris,
                                        mfinal=10)

# Random Forest
recomendador.rf.paris <- randomForest(paris ~ ., data=datos.trabajo.paris,
                                      ntree=100, proximity=TRUE)
```

c. Evalúe el desempeño de los clasificadores a través de realizar una predicción en el conjunto de prueba utilizando la métrica de exactitud, y comente en el foro qué algoritmo resulta ganador.

EVALUACIÓN DE DESEMPEÑO USANDO LA MÉTRICA DE 'EXACTITUD'.

Clasificador: Árbol de decisión

```
## [1] 95.75
```

Clasificador: Bayes Ingenuo

```
## [1] 76.89
```

Clasificador: Bagging

```
## [1] 95.28
```

Clasificador: Boosting

```
## [1] 95.75
```

Clasificador: Random Forest

```
## [1] 94.81
```

- *Resp: se evaluó el desempeño de los clasificadores usando 2 semillas distintas (1 y 16). Para efectos de presentación de resultados se utilizó* `set.seed(1)`.
- *Para el caso de la semilla igual a 1, los mejores modelos fueron Árbol de Decisión y Boosting.*
- *Para el caso de la semilla igual a 16, los modelos con un índice más elevado fueron Árbol de Decisión y Bagging*
- *En ambos casos, el modelo con un valor más alto para la métrica de exactitud fue* ***'Árbol de Decisión'***.

d. Responda en el foro: ¿Por qué puede ser conveniente utilizar la métrica de exactitud, y no otra, desde el punto de vista de la recomendación?.

- *Resp: en este ejercicio, en donde tenemos poca cantidad de datos y poco feedback por parte del usuario, sí conviene utilizar la métrica de exactitud, ya que no solo buscamos predecir y acertar a una condición positiva (verdaderos positivos), sino que también predecir y acertar a una condición negativa (verdaderos negativos) con el propósito de no ofrecer recomendaciones que al cliente no le interesen. Es decir, buscamos maximizar: ([Verdaderos Positivos + Verdaderos Negativos]/Población Total).*
- *Por otra parte, la métrica de exactitud puede llevar a interpretaciones erróneas si el set de datos no está balanceado [1]*
- *Ahora, si se tratara de un set de datos con miles/millones de usuarios e ítems, como por ejemplo Spotify y Netflix, ya no sería tan atractivo medir solamente la exactitud de los sistemas recomendadores, sino también incluir métricas como 'diversity' (variabilidad de ítems presentes en la lista de recomendaciones), 'novelty' (habilidad de recomendar al usuario ítems que no haya experimentado anteriormente) y 'serendipity' (qué tan sorprendente es para el usuario una lista de recomendaciones) [2][3].*

5. Repita el paso anterior para el sistema recomendador asociado a la ciudad de Londres.

- *Conjunto de entrenamiento (70% de los datos)*

```
##     gender    children      country          motive          marital_status
##  female:155   0:130    argentina:190   business : 69   married      :343
##  male  :337   1:362    chile    : 86   bussiness: 77   single       :129
##                        peru     :216   travel   :346   widow/widower: 20
##
##
##
##       age        madrid  london
##  Min.   :20.00   0:158   0:250
##  1st Qu.:34.00   1:334   1:242
##  Median :41.00
##  Mean   :41.68
##  3rd Qu.:49.00
##  Max.   :65.00
```

- *Conjunto de validación (30% de los datos)*

```
##     gender    children      country         motive          marital_status
##  female: 71   0: 58    argentina:95   business : 32   married      :141
##  male  :141   1:154    chile    :40   bussiness: 33   single       : 61
##                        peru     :77   travel   :147   widow/widower: 10
##
##
##
##       age        madrid  london
##  Min.   :20.00   0: 71   0:111
##  1st Qu.:33.00   1:141   1:101
##  Median :38.50
##  Mean   :40.21
##  3rd Qu.:47.00
##  Max.   :65.00
```

EVALUACIÓN DE DESEMPEÑO USANDO LA MÉTRICA DE 'EXACTITUD'.

Clasificador: Árbol de Decisión

```
## [1] 98.58
```

Clasificador: Bayes Ingenuo

```
## [1] 73.58
```

Clasificador: Bagging

```
## [1] 98.58
```

Clasificador: Boosting

```
## [1] 98.58
```

Clasificador: Random Forest

```
## [1] 99.06
```

- *Resp: se evaluó el desempeño de los clasificadores usando 2 semillas distintas (1 y 16). Para efectos de presentación de resultados se utilizó* `set.seed(1)`.
- *Para el caso de la semilla igual a 1, el mejor modelo fue Random Forest.*
- *Para el caso de la semilla igual a 16, los modelos con un índice más elevado fueron Boosting y Random Forest.*
- *En ambos casos, el modelo con un valor más alto para la métrica de exactitud fue* ***'Random Forest'***.

6. Una vez construidos ambos sistemas recomendadores, cargue el conjunto de datos *fidelizacion.csv*, en la cual se presentan clientes pertenecientes al programa de fidelización. Para estos clientes, solamente se sabe si viajan o no a madrid el año anterior, por lo cual las columnas "london" y "paris" han sido dejadas en un valor nulo.

- *Resumen clientes pertenecientes al programa de fidelización.*

```
##     gender   children      country         motive         marital_status
##  female:14   0: 9     argentina:17   business : 6   married      :28
##  male  :24   1:29     chile    : 9   bussiness: 2   single       : 9
##                       peru     :12   travel   :30   widow/widower: 1
##
##
##
##       age        madrid
##  Min.   :21.00   0: 9
##  1st Qu.:30.75   1:29
##  Median :42.50
##  Mean   :42.53
##  3rd Qu.:53.75
##  Max.   :65.00
```

7. Utilizando los recomendadores, realice una recomendación hacia las ciudades de Londres y París para cada cliente.

```
# Recomendación a París, usando modelo mejor evaluado (Árbol de Decisión)
pred.fid.paris <- predict(recomendador.tree.paris,
                          newdata = datos.fid,
                          type='class')

# Recomendación a Londres, usando modelo mejor evaluado (Random Forest)

pred.fid.londres <- predict(recomendador.rf.londres,
                            newdata = datos.fid,
                            type='class')

datos.fid.pred <- cbind(datos.fid, pred.fid.londres, pred.fid.paris)
```

8. En base a las recomendaciones realizadas, calcule:

a. La cantidad de clientes a cuales se les recomiendan ambas ciudades:

```
sum((pred.fid.londres == 1) & (pred.fid.paris == 1))
```

```
## [1] 13
```

b. La cantidad de clientes a los cuales se les recomienda solamente Londres:

```
sum((pred.fid.londres == 1) & (pred.fid.paris == 0))
```

```
## [1] 1
```

c. La cantidad de clientes a los cuales se les recomienda solamente París:

```
sum((pred.fid.londres == 0) & (pred.fid.paris == 1))
```

```
## [1] 12
```

d. La cantidad de clientes a los cuales no se les recomienda ningún viaje:

```
sum((pred.fid.londres == 0) & (pred.fid.paris == 0))
```

```
## [1] 12
```

Recordar que se usó una semilla igual a 1, por lo cual si se elige otra, las recomendaciones por destino de interés pueden variar.

9. En base a los cálculos del punto anterior, reflexione: ¿Qué servicio/oferta podría plantear la empresa para cada uno de los grupos anteriores?. Responda fundamentadamente en el Foro.

RESUMEN RESULTADOS POR GRUPO.

a: clientes a los que se les recomienda ambas ciudades.

- *La mayoría de este grupo son personas provenientes de Argentina y Perú, casados, con hijos, cuyo motivo de viaje anterior fue por turismo.*
- *Se sugiere ofrecerles un pack de viaje familiar, con itinerario en ambas ciudades, pase de acceso a entretenciones para niños y visita a sitios patrimoniales.*

```
##     gender   children      country        motive        marital_status
##  female: 2   0: 1     argentina:6   business :3   married      :11
##  male  :11   1:12     chile    :1   bussiness:1   single       : 1
##                       peru     :6   travel   :9   widow/widower: 1
##
##
##
##       age        madrid pred.fid.londres pred.fid.paris
##  Min.   :33.00   0:5    0: 0             0: 0
##  1st Qu.:40.00   1:8    1:13             1:13
##  Median :43.00
##  Mean   :44.46
##  3rd Qu.:51.00
##  Max.   :64.00
```

b: clientes a los cuales se les recomienda solamente Londres.

- *Grupo conformado por un solo cliente, hombre, sin hijos, soltero.*
- *Se sugiere ofrecerle un plan personalizado para visitar sitios patrimoniales y otorgar pase de entrada a lugares de la bohemia local.*

```
##     gender  children      country        motive        marital_status
##  female:0   0:1      argentina:1   business :0   married      :0
##  male  :1   1:0      chile    :0   bussiness:1   single       :1
##                      peru     :0   travel   :0   widow/widower:0
##
##
##
##       age     madrid pred.fid.londres pred.fid.paris
##  Min.   :39   0:1    0:0             0:1
##  1st Qu.:39   1:0    1:1             1:0
##  Median :39
##  Mean   :39
##  3rd Qu.:39
##  Max.   :39
```

c: clientes a los cuales se les recomienda solamente París.

- *La mayoría de este grupo son personas provenientes de Argentina y Perú, casados, con hijos, cuyo motivo de viaje anterior fue por turismo.*
- *Este grupo es muy parecido al 'a', por lo cual se les oferta una propuesta parecida: pack de viaje familiar, con pase de acceso a entretenciones para niños y visita a sitios patrimoniales.*

```
##     gender  children      country        motive        marital_status
##  female:5   0:3      argentina:5   business :3   married      :10
##  male  :7   1:9      chile    :1   bussiness:0   single       : 2
##                      peru     :6   travel   :9   widow/widower: 0
##
##
##
##       age        madrid pred.fid.londres pred.fid.paris
##  Min.   :21.00   0:3    0:12             0: 0
##  1st Qu.:28.75   1:9    1: 0             1:12
##  Median :41.50
##  Mean   :40.67
##  3rd Qu.:49.75
##  Max.   :64.00
```

d: clientes a los cuales no se les recomienda ningún viaje.

- *Grupo mixto de clientes, los cuales el 100% viajaron por turismo a Madrid el año anterior.*
- *Sería necesario recopilar más antecedentes del porqué no manifiestan interés por los destinos sugeridos (por ejemplo, se podría evaluar su interés por otros destinos distintos a ciudades europeas).*

```
##     gender  children      country        motive         marital_status
##  female:7   0:4      argentina:5   business : 0   married      :7
##  male  :5   1:8      chile    :7   bussiness: 0   single       :5
##                      peru     :0   travel   :12   widow/widower:0
##
##
##
##       age        madrid pred.fid.londres pred.fid.paris
##  Min.   :22.00   0: 0   0:12             0:12
##  1st Qu.:23.75   1:12   1: 0             1: 0
##  Median :50.00
##  Mean   :42.58
##  3rd Qu.:54.75
##  Max.   :65.00
```

- En resumen, como acción inmediata, se podría ofrecer una propuesta de viaje similar para los grupos 'a' y 'c', más una oferta personalizada para la persona del grupo 'b'.
- En lo que respecta al grupo 'd', se podría cuestionar si es necesario fidelizarlos, o bien, buscar otras estrategias para captar su interés.

BIBLIOGRAFÍA

[1][Accuracy, https://en.wikipedia.org/wiki/Precision_and_recall, Imbalanced data]

[2][UCL Department of Computer Science, http://www.cs.ucl.ac.uk/fileadmin/UCL-CS/research/Research_Notes/RN_11_21.pdf, Why accuracy is not enough]

[3][James Topor, https://rpubs.com/jt_rpubs/288709, Beyond accuracy: adding greater serendipity to a recommender system]