Un análisis de un experimento de citas rápidas usando aprendizaje de maquina

Ilan Jinich
Juan Pablo Rodriguez

What influences love at first sight?

Descripción del problema

Kaggle es una plataforma en la que estadísticos y científicos de datos compiten en producir el mejor modelo para predecir y describir un conjunto de datos. En particular uno de los conjuntos de datos que se encuentran en la pagina es el de *Speed Dating Experiment* [1], los datos fueron juntados entre el 2002 y el 2004 y traen información sobre los eventos que se realizaron de citas rápidas en esa época y sus participantes, cada participante tuvo un promedio de 20 citas con una duración de cuatro minutos cada una.
Durante los eventos se les realizo a los participantes una serie de preguntas y con base a estos se obtuvieron datos de edad, estudios, calificación de SAT, raza, importancia de raza (¿qué tan importante es para usted que en una relación la otra persona sea de la misma raza a usted?), importancia de religión (¿qué tan importante es para usted que en una relación la otra persona sea de la misma religión a usted?), meta en el evento (conocer gente, divertirse, etc.), frecuencia de citas (¿cuantas veces a la semana tiene una cita?), frecuencia de salidas (¿cuantas veces a la semana sales?(no necesariamente en citas),interés en música, interés en yoga, interés en hacer deporte, interés en ver deportes, interés en hacer ejercicio, interés en comida, interés en museos, interés en arte, interés en excursionismo, interés en vídeo juegos, interés en antros, interés en leer, interés en televisión, interés en el teatro, interés en películas, interés en conciertos, interés en ir de compras [1] y felicidad esperada (¿qué tan feliz espera ser con las personas que conozca en el evento?) para cada uno de los participantes. Además para cada una de las citas se recopilaron datos sobre la calificación de atractividad que le daba cada uno de los participantes a su cita y si hubo o no un *"match"* en la cita [2].

Análisis de regresión

La primeras preguntas que nos hicimos basándonos en [2] fue ¿qué tanto influye la raza de la otra persona en la calificación que le das de atractividad? ¿sera igual para los hombres que para las mujeres esta relación?¿la gente prefiere salir con gente de su propia raza?
La variable de raza tenia cinco posibles valores: raza blanca, raza negra, raza hispana, raza asiática y otro. Dado que la gente que contesto otro, no sabemos de que raza son decidimos considerarlos como si no pertenecieran a ninguna raza en lugar de ignorar los datos.
El análisis lo hicimos de la siguiente manera:

[1] Las variables de interés son una calificación del 1 al 10.
[2] Como vera son muchas variables ¿quién dijo Big Data?

1. Dividimos a los participantes en 8 grupos de acuerdo a su genero y su raza.

2. para cada grupo hicimos el modelo de regresión $y = \beta_0 + \beta_1 negro + \beta_2 blanco + \beta_3 hispano + \beta_4 asiatico$. Donde y es una variable de respuesta que marca que tan atractiva es la persona y las variables de raza (*negro*, *blanco*, *hispano* y *asiatico*) son variables explicativas categóricas que toman el valor de uno si la persona pertenece a la raza y cero en otro caso.

Los resultados que obtuvimos fueron los siguientes:[3]

grupo	β_0	β_1	β_2	β_3	β_4	ECM
mujeres negras	7.000000	-0.1111111	-0.77142857	-0.37500000	-1.2982456	3.920570
mujeres blancas	5.877698	0.4840043	0.42478523	0.09267253	-0.5929152	4.040844
mujeres hispanas	5.129032	0.4499151	0.54596774	0.67096774	-0.7711375	4.830665
mujeres asiáticas	5.682927	0.2726287	0.26806985	0.33402232	-0.3854585	2.952910
hombres negros	6.545455	0.4545455	0.54150198	1.1767677	-0.2431290	2.515770
hombres blancos	6.369697	-0.4353904	0.15464563	0.4257576	-0.2815106	3.560819
hombres hispanos	6.652174	-0.7771739	0.05893720	0.1478261	-0.5674282	3.004950
hombres asiáticos	6.276923	-0.6389920	0.36022595	0.3546559	-0.2054945	3.207121

El error cuadrático medio para todos los grupos fue muy grande considerando que la y toma valores del cero al diez y por lo tanto las predicciones del modelo no son las mejores. Como el modelo describe el comportamiento de las personas, a pesar de los errores grandes, decidimos dejarlo como una alternativa descriptiva valida.

Conclusiones

1. Análisis predictivo

En la dinámica de *Speed Dating* del experimento se recopilo información en dos distintas etapas: antes de las citas y después de todas las citas. El análisis que presentamos a continuación fue realizado con la información que se tenia de cada participante antes de empezar la ronda de citas.

Datos

La variable que vamos a predecir es 'Match'que tiene posibles valores $0, 1$.

En particular nos enfocamos en la serie de preferencias que tenia cada participante sobre diferentes temas y actividades. Incluimos también

[3]ECM=Error cuadrático medio

Random forest

Neural net

SVM

Referencias

[1] Kaggle-Speed Dating Experiment
`https://www.kaggle.com/annavictoria/speed-dating-experiment` (última consulta en diciembre de 2017).

[2] R. Fisman et.al., "Racial Preferences in Dating "
`http://faculty.chicagobooth.edu/emir.kamenica/documents/racialpreferences.pdf` (última consulta en diciembre de 2017).