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.

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.

Transformación de Datos

La variable que vamos a predecir es 'Match'que tiene posibles valores $0, 1$. El valor 1 de *match* las dos partes de la cita deben de estar de acuerdo en salir en una segunda cita. Partiendo del supuesto, alejado de la realidad, de que una persona acepta una segunda cita el $50 porciento$ de las veces. Entonces la predicción mas ingenua indica que el porcentaje de precisión promedio es de 25 por ciento. El objetivo de nuestra predicción es superar esta precisión.

Realizamos una reducción de dimensionalidad al set de datos, conservamos solamente 33 de las 150 variables iniciales. Cada uno de los renglones del set de datos representa una interacción entre dos agentes que tienen asociado un *iid*. Eliminamos la mitad de los registros

[3]ECM=Error cuadrático medio

porque para cada interacción existían dos renglones de información.

Lo primero que hicimos fue construir para cada uno de los 512 participantes un set de datos en el cual mapeamos la información de preferencias de cada uno de los participantes. Para cada una de las interacciones teníamos dos agentes, construimos un nuevo renglón de registro a partir de la diferencia que había en cada una de las preferencias. Por ejemplo, la persona A tiene una preferencia de $[0, 10]$ sobre Conciertos, en particular tiene una preferencia de 8. La persona B tiene una preferencia de 2. El resultado en la columna artificial que creamos esta dada por la formula: $\text{ABS}(pc_A - pc_B)$. En este caso tendría un valor de 6. Esta misma lógica la seguimos para las preferencias en los temas: importancia de raza, importancia de religión, deportes, ver deportes en la televisión, ejercicio, salir a cenar, visitar museos/galerías, arte, caminata (*hiking*), videojuegos , salir de antro, leer, ver televisión, teatro, películas, conciertos, música, ir de compras, yoga, felicidad esperada de las citas y numero esperado de personas que te interesan. La magnitud de diferencia en estas diferencias representan 21 de las 31 variables que vamos a utilizar para realizar la predicción. Las otras 10 variables son las siguientes:

1. **Diferencia de edad:** Indica la diferencia de edad entre los participantes.
2. **Área de estudio**: Valores $[0, 1]$. Existen 18 áreas de estudio distintas, si los participantes eligieron el mismo, entonces la variable toma el valor 1, en caso contrario 0.
3. **Raza:**Valores $[0, 1]$. Existen 6 razas distintas, si los participantes son de la misma raza entonces la variable vale 1, en caso contrario 0.
4. **Lugar de origen:** Valores $[0, 1]$. En esta sección los participantes describen el su país, estado o ciudad de origen. En caso de que sea el mismo la variable vale 1, en caso contrario 0.
5. **Código Postal:**Valores $[0, 1]$. En caso de que tengan el mismo código postal la variable vale 1, en caso contrario 0.
6. **Ingreso:** Se realizo un promedio del ingreso de los participantes para rellenar los registros vacíos. Esta variable indica la diferencia en valor absoluto de los ingresos de los participantes.
7. **Objetivo de asistir al evento:** Valores $[0, 1]$. Existen 6 diferentes opciones las cuales representan el objetivo que tienen las personas del evento. Si los participantes tienen el mismo objetivo la variable vale 1, en caso contrario 0.
8. **Frecuencia con la que van en citas:** Esta variable representa la diferencia en valor absoluto entre la frecuencia de citas que tiene cada uno de los participantes en su vida cotidiana.
9. **Frecuencia con la que sales en las noches:** Esta variable representa la diferencia en valor absoluto entre la frecuencia de salidas nocturnas que tiene cada uno de los participantes en su vida cotidiana.

10. **Carrera profesional:** Valores $[0, 1]$. Existen 17 clasificaciones de careras profesionales distintas, si los participantes eligieron el mismo, entonces la variable toma el valor 1, en caso contrario 0.

Modelos

Lo primero que se tiene que hacer antes de empezar a usar modelos y ajustarlos para optimizar el desempeño es dividir los datos en tres sets: Entrenamiento (68 por ciento), Validación (12 por ciento) y Pruebas (20 por ciento).

- Tenemos en total: 4184 registros.
- Set de entrenamiento: 2844 registros.
- Set de Validación: 503 registros.
- Set de Prueba: 837 registros

Primero solo vamos a trabajar con el set de datos de entrenamiento. Vamos a hacer un primer intento y meter los datos a modelos de clasificación que acepten variables continuas. Vamos a usar: Naive Bayes, Árbol ID3 y Random Forest.

Naive Bayes

Naive Bayes es un algoritmo que asume independencia entre variables y en base a probabilidades condicionales calcula una probabilidad de clasificar como 1 y 0. En la practica es muy útil porque funciona como punto de referencia.

Figura 1: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.85	0.94	0.89	424
1	0.29	0.14	0.19	79
avg / total	0.77	0.81	0.78	503

		Valor Real	
		0	1
Prediccion	0	397	68
	1	27	11

Árbol

Técnica que utiliza busca encontrar el mejor atributo para dividir a los datos de acuerdo al valor de predicción. Hace esto en múltiples niveles hasta que divide por completo las observaciones.

Figura 2: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.84	1	0.91	424
1	0.33	0.01	0.02	79
avg / total	0.76	0.84	0.77	503

		Valor Real	
		0	1
Prediccion	0	422	78
	1	2	1

Random Forest

Técnica que crea un numero determinado de arboles de clasificación tipo ID3, cada uno de los arboles selecciona de manera aleatoria variables para hacer la clasificación, es esta aleatoriedad por lo que se conoce a la técnica como 'Random Forest'.

Figura 3: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.84	1	0.91	424
1	0	0	0	79
avg / total	0.71	0.84	0.77	503

Prediccion	Valor Real 0	Valor Real 1
0	424	79
1	0	0

2nda Reducción de dimensionalidad

Realizamos una 2nda reducción de dimensionalidad utilizando arboles, debido a que tuvieron un mejor desempeño que Random Forest. Cada uno de los arboles realiza un análisis de significancia de las variables, es decir, en que medida explican el resultado de Match o no Match. Tomando un nivel de significancia mínimo para las variables de 0.03.

Las variables que resultaron mas significativas fueron:

Realizamos una segunda etapa de modelos. En esta etapa, solamente vamos a usar las variables que salieron significativas en el paso anterior. La hipótesis es que al reducir el numero de variables y conservar solamente las importantes los modelos vana tener un mejor desempeño.

Naive Bayes

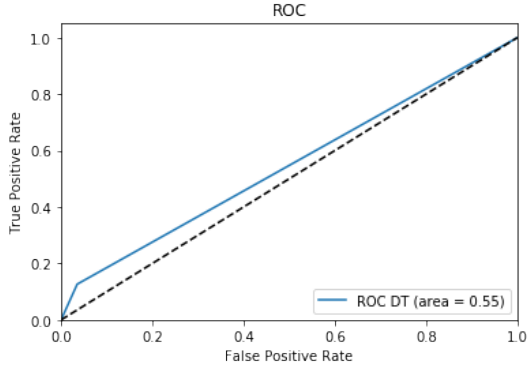
Los resultados de *Naive Bayes* mejoraron en precisión , recall y f1-score. Lo que significa que incremento en las tres métricas que están bajo consideración. Tiene un ROC que esta ligeramente por la linea con pendiente 1, lo que resulta en *Área bajo la curva* (AUC) de .55.

Figura 4: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.86	0.96	0.91	424
1	0.4	0.13	0.19	79
avg / total	0.78	0.83	0.79	503

Prediccion	Valor Real 0	Valor Real 1
0	397	68
1	27	11

Figura 5: ROC



Árbol

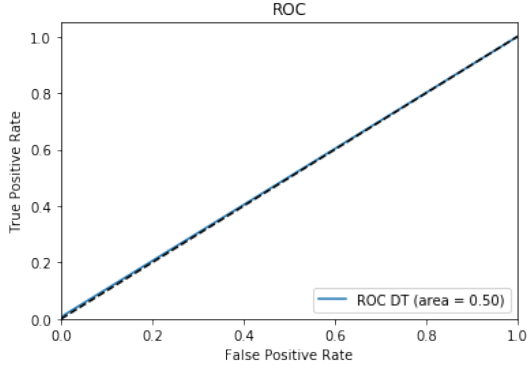
El árbol mantuvo el mismo desempeño que si usamos todas las variables. Esto se debe a que ne la etapa anterior uso solamente las variables que conservamos para hacer una predicción, no sufrió ningún cambio al eliminar el resto de las variables que no tenían significancia,

Figura 6: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.84	1	0.91	424
1	0.33	0.01	0.02	79
avg / total	0.76	0.84	0.77	503

		Valor Real	
		0	1
Prediccion	0	422	78
	1	2	1

Figura 7:



Random Forest

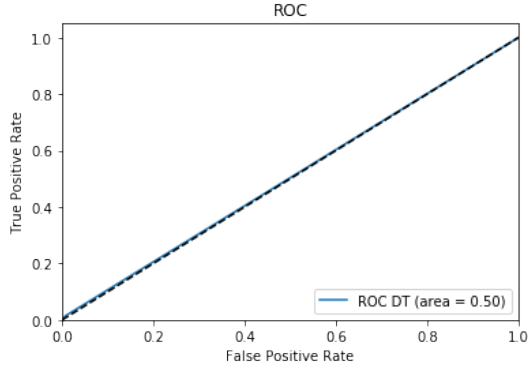
Esta técnica aumento su precisión en 25 porciento, sin embargo, el resto de las variables conservaron su valor. El cambio si resulto benéfico para el modelo pero no supera el desempeño de *Naive Bayes*.

Figura 8: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.85	0.99	0.91	424
1	0.4	0.03	0.05	79
avg / total	0.78	0.84	0.78	503

Prediccion	Valor Real 0	Valor Real 1
0	421	77
1	3	2

Figura 9:



Set de Prueba

El mejor modelo para determinar si una pareja de personas hacen *Match* fue *Naive Bayes* con una precisión de promedio de 78 porciento. En el caso de los valores 1 que son los mas importantes porque son los *Matches*, obtuvo un 40 porciento. Ademas cuenta con el mayor soporte.

Después de seleccionar el modelo con el mejor desempeño, tomamos los datos que separamos en un principio y los enchufamos en nuestro modelo para ver que tan buena es la predicción. Recuerden que estos datos nunca los ha visto el modelo.

Figura 10: Matrices de clasificación y Confusión

	precision	recall	f1-score	support
0	0.84	0.96	0.9	697
1	0.27	0.07	0.11	140
avg / total	0.74	0.81	0.76	837

		Valor Real	
		0	1
Prediccion	0	670	130
	1	27	10

Figura 11:

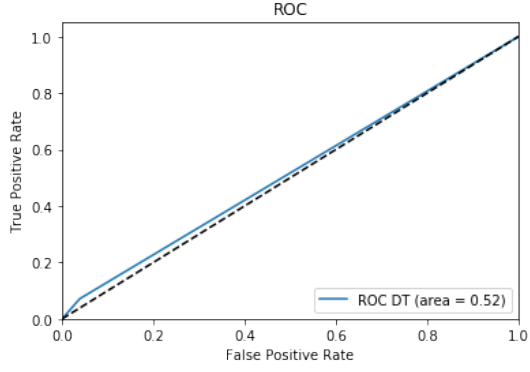


Figura 12: Variables Relevantes con su correlación con Match

	Match
Match	1
Ingreso	0.03344
Salidas en la noches	-0.064659
Numero Esperado de citas exitosas	0.037851
Museos	0.025033
Arte	0.018317
Antros	-0.021505
Television	-0.022507
Teatro	0.008577
Conciertos	0.003754
Musica	0.009057

Conclusión

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).