Trabajo Práctico 1: Business Case

Fridman Axel	Hsueh Noé	Salas Héctor
527/20	546/19	Postgrado

Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Machine Learning Práctico

Motivación Un accidente cerebrovascular isquémico (ACV) ocurre cuando se interrumpe o se reduce el suministro de sangre a una parte del cerebro, lo que impide que el tejido cerebral reciba oxígeno y nutrientes.[1] Cada año, casi 800.000 personas tienen un accidente cerebrovascular, más de 140.000 mueren y muchos sobrevivientes quedan con discapacidades.[2] Un ACV es una emergencia médica cuyo tratamiento inmediato es crucial para reducir el daño cerebral y futuras complicaciones. Si bien los síntomas son altamente reconocibles, su identificación en el momento es difícil. En el caso de Argentina, la Cruz Roja cuenta con 50.000 personas capacitadas para la identificación del ACV,[3] el 0.11 % de la población.

Descripción del caso de negocio Una clínica importante quiere reducir las muertes de sus pacientes por ACV por medio de una campaña de concientización de sus pacientes de alto riesgo. Para lograr una temprana detección, debemos capacitar a las personas con mayor riesgo y a su entorno a identificar los síntomas más comunes. Eso nos dejaría con una incógnita, ¿quiénes son las personas con más riesgo de tener un ACV?

En este caso, nuestro objetivo será identificar los usuarios con riesgo de ACV y proveerles una notificación temprana. Nuestro KPI será la reducción de las muertes por accidentes cardiovasculares anuales en un 20 % en un lapso de 5 años en nuestro hospital. La métrica offline que trataremos de mejorar y refinar será la F1 de nuestro modelo predictor de probabilidad de accidente. Claramente, poder predecir de manera correcta nuestros pacientes de alto riesgo, nos permitirá tomar mayores recaudos y reducir las muertes. Consecuentemente, esto impactará en nuestro KPI. Cabe destacar que para nuestro caso, consideramos que es más importante la detección de aquellos pacientes con riesgo de ACV frente a los que no tendrán ACV pese a que el modelo los considere como tal.

Presentación del dataset El dataset elegido se encuentra en el siguiente link, es tomado de *Kaggle* y cuenta con 5.000 observaciones y diversas variables, algunas muy relacionadas a la salud y fisiología, como por ejemplo: si es hipertenso, la edad, el sexo, si tuvo alguna enfermedad del corazón, el nivel de glucosa y el índice de masa corporal. Mientras que otras variables son más relacionadas a cuestiones sociales o de hábitos: estado civil, tipo de trabajo, tipo de residencia y si es fumadora o no lo es/ lo dejó.

No escapa nuestra atención, que si bien es posible encontrar que la residencia o el tipo de trabajo pueden correlacionarse bien con si tuvo o no un ACV, podría deberse a otro factor en común como pobreza o falta de acceso al sistema de salud. Es por eso que si bien puede que exploremos su relación con las otras variables, tenemos cautela en su valor explicativo.

Este dataset fue descargado 87.535 veces con lo cual ya fue sumamente estudiado y se lo considera de alta calidad para el desarrollo de un modelo simple.

Análisis de viabilidad

Nro.	Hipótesis	Experimento para validar o refutar	Costo de hacer el experimento	Impacto esperado de saber la respuesta
1	La edad es el mayor predictor de riesgo.	Implementar árboles de decisión y validar si es de las preguntas que 'más arriba' está (i.e. reduce más la entropía).	Realizar este tipo de experimento tiene solo costo temporal. Estimamos 3 hs totales entre capacitarnos en las herramientas y tomar conclusiones.	Poder interpretar mejor nuestro modelo y darnos capacidad explicativa sobre por qué pondera el riesgo de una persona de tener un ACV de una manera u otra.
2	El `work_type` no es relevante.	Agregar la variable a un modelo ya implementado y observar que la métrica usada no mejora sustancialmente.	Ídem experimento (1).	Ídem experimento (1).
3	El impacto del modelo asumiendo un modelo perfecto.	Elaboración de un modelo perfecto asumiendo que se puede obtener todos los datos necesarios de los pacientes para que el modelo funcione.	Acceso a los datos por parte del hospital y la realización del modelo.	Capacidad de evitar en el país los 126 mil casos de ACV por año, de los cuales 18 mil terminan en muerte.[4]
4	Fumar causa ACV	Establecer un modelo y estudiar su causalidad.	Capacitación en estadística para determinar causalidad entre variables.	Establecer la causalidad del ACV permite prevenirlo de mejor forma.

5	El género del paciente varía en importancia según la edad.	Se divide el dataset en 4 grupos etarios, se realiza para cada uno de ellos un modelo. Luego, se explora cómo varía la precisión de los modelos al agregar y sacar el feature `genero`.	Realizar este tipo de experimento tiene solo costo temporal. Estimamos 3 hs totales entre capacitarnos en las herramientas y tomar conclusiones.	Determinar si el género es un buen predictor de riesgo.
6	Los pacientes están dispuestos recibir una notificación de posible ACV siendo este un posible falso positivo.	Llamar a una cantidad entre 50 y 100 pacientes de un hospital preguntándoles si estarían dispuestos a recibir un llamado de advertencia de ACV sabiendo que nuestro modelo puede equivocarse.	Aproximadamente 5-6 hs de trabajo entre llamadas y documentación.	Nos permitirá saber cuán preciso deberíamos hacer nuestro modelo para no "sobreestimar" la cantidad de potenciales ACV y tener una viabilidad del proyecto.

Experimentos Se seleccionaron los experimentos 1 y 5 para realizarse, accesible por medio de este link.

Referencias

[1] Mayo Clinic. *Accidente Cardiovascular*. [Artículo]

[2] Centers for Disease Control and Prevention. *Prevención de muertes por accidentes cerebrovasculares.* [Artículo]

[3] Cruz Roja. [Página web]

[4] Chavez, Valeria. *En la Argentina se produce un ACV cada 9 minutos: tres señales de alerta para detectarlo a tiempo.* En *Infobae* [Nota periodística]