crehana

Proyecto Parte 1: análisis exploratorio de los datos

Analítica predictiva en R, por David Zarruk

Technology & Data

Para el proyecto, debes descargar la base de datos “diabetes.csv” que hace parte de los adjuntos del curso. Esta base de datos tiene información sobre distintas características de un grupo de personas, así como una variable que indica si la persona fue diagnosticada con diabetes cinco años después o no. Las variables de la base de datos son:

1. *num_embarazos:* Número de veces que la persona ha estado embarazada
2. *plasma*: Concentración de la glucosa del plasma
3. *presion_diastolica*: Presión diastólica (mm Hg)
4. *grosor_piel:* Grosor de la piel en el triceps (mm)
5. *insulina:* Insulina sérica de 2 horas (mu U/ml)
6. *bmi:* Índice de masa corporal (peso en kg / altura en metros al cuadrado)
7. *diabetes_pedigree:* Función de pedigree de diabetes
8. *edad:* Edad
9. *diabetes:* 0 si no fue diagnosticado, 1 si fue diagnosticado

El objetivo del proyecto es pronosticar si una persona va a ser diagnosticada con diabetes, dadas las variables 1.-8.. En este primer avance, el objetivo es hacer el análisis exploratorio de los datos. En particular, el alumno debe encontrar:

i. ¿Cuántos datos tenemos para nuestro análisis?
ii. Encontrar estadísticas descriptivas de todas las variables de la base de datos. Entre ellas, deben estar: media, desviación estándar, mínimo, máximo, percentiles 25, 50 y 75.
iii. Hacer un análisis exploratorio para ver qué variables pueden ser las mejores al momento de predecir si una persona tiene diabetes. Para esto, debes generar:

    i. Gráficos de dispersión de las variables 2.-8. contra la variable edad. ¿Cuáles variables parecieran ser buenas para explicar la edad de un abalón?
    ii. Gráfico de calor para ver las correlaciones existentes entre variables. ¿Lo que se observa en el gráfico de calor concuerda con lo encontrado en los gráficos de dispersión? ¿Cuáles variables tienen correlaciones positivas o negativas más fuertes con respecto a la variable edad?
    iii. Gráficos de cajas y bigotes para todas las variables categóricas o discretas. ¿Cuáles variables son categóricas? ¿Qué categorías tienen mayor edad en promedio?

Proyecto parte 2: métricas de desempeño de un modelo

Para esta segunda entrega, debes descargar la base de datos “diabetes_test.csv”. Esta base de datos contiene el conjunto de validación, sobre el que vamos a medir qué tan buen desempeño tiene un modelo particular.

Esta base de datos contiene exactamente las mismas variables que la base de datos de la primera entrega, además de dos nuevas columnas llamadas modelo_2 y modelo_3. Estas variables son la predicción de que un usuario tenga diabetes utilizando dos modelos entrenados por otra persona.

En este segundo avance, el objetivo es que midas el desempeño de cada modelo para predecir la variable diabetes y escoger cuál modelo es mejor entre los dos propuestos. En particular, es importante que hagas lo siguiente:

[1]Si se intenta hacer el gráfico de dispersión de la librería Seaborn para la variable sexo agregándole la línea de regresión como hicimos en clase, va a arrojar error por ser una variable categórica. Por eso debemos excluirla de este análisis si queremos agregar línea de regresión.

1. Computa la matriz de confusión para cada modelo (haz una matriz separada separado para el modelo_2 y para el modelo_3). ¿Qué modelo tiene mayor tasa de falsos negativos? ¿Y de falsos positivos?
2. Calcula las métricas relevantes para un modelo de clasificación:
    a. Accuracy
    b. Recall
    c. Precisión
    d. F1-score

3. ¿Qué métrica crees que es la más relevante para escoger el "mejor" modelo? Justifica tu respuesta.
4. De acuerdo a tu respuesta en el punto 3, ¿qué modelo escogerías?
5. ¿Qué tan bueno es el mejor modelo? Es decir, describe qué significa que el modelo que escogiste tenga ese nivel de precisión y de recall.

Descripción del proyecto final

Ahora sí, llegó la hora de hacer analítica predictiva en R. El objetivo de esta tercera entrega del proyecto es que construyas tu propio modelo de clasificación para predecir si una persona tiene diabetes.

Para este propósito, deberás usar las bases de datos *"diabetes_test.csv"* y *"diabetes_train.csv"*. El objetivo es que construyas distintos modelos para predecir si una persona tiene diabetes y que escojas el mejor entre todos los modelos. Los puntos a entregar son:

1. De acuerdo con el análisis exploratorio de la primera entrega, ¿cuáles variables parecieran ser las que más información tienen para predecir si una persona tiene diabetes?
2. Toma los datos del conjunto de entrenamiento diabetes_train.csv. Entrena 3 distintos modelos de predicción con estos datos. Recuerda que las variaciones a tus modelos pueden ser:

   a. Utiliza distintos modelos: modelo de probabilidad lineal univariado, modelo de probabilidad lineal multivariado, regresión logística, Random Forests con distintos parámetros
   b. Utiliza distintas variables en cada modelo. Por ejemplo, en un modelo puedes utilizar num_embarazos, plasma, presion_diastolica y grosor_piel. En otro modelo puedes utilizar todas las variables.
   c. Puedes construir nuevas variables.

3. Toma los datos del conjunto de validación diabetes_test.csv. Para cada uno de los modelos del punto 2., haz la matriz de confusión.
4. Calcula las métricas de accuracy, precisión, recall y F1-score y agrégalas en una tabla que permita comparar el desempeño de los tres modelos. ¿Qué modelo es mejor para predecir si una persona tiene diabetes? ¿Qué tan bueno es el mejor modelo?

¡Espero que te diviertas con este proyecto!