crehana

Proyecto Final:

Analítica predictiva y modelos de regresión en Python, por David Zarruk

Technology & Data

Parte 1: Análisis exploratorio de los datos

Para este proyecto debes descargar la base de datos abalone_train.csv que hace parte de los adjuntos del curso. Esta base de datos contiene información sobre un conjunto de abalones, que son un tipo de molusco. Las variables de la base de datos son:

1. *sexo: masculino (M), femenino (F) o indefinido (I)*
2. *longitud*
3. *diametro*
4. *altura*
5. *peso_entero*
6. *peso_sin_cascara*
7. *peso_visceras*
8. *peso_cascara*
9. *edad*

El objetivo final del proyecto es predecir la edad de un abalón, utilizando información de las variables disponibles en la base de datos. En este primer avance, el objetivo es realizar el análisis exploratorio de los datos. En particular, debes realizar lo siguiente:

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? **Nota:** 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.
    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 abalone_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 la edad de cada abalón 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 edad y escoger cuál modelo es mejor entre los dos propuestos. En particular, es importante que hagas lo siguiente:

1. Grafica la variable edad contra la edad predicha en cada modelo (haz un gráfico separado para el modelo_2 y para el modelo_3). ¿Qué modelo pareciera ser mejor cuando se hace el análisis gráfico?
2. Calcular las métricas RMSE, MAE y MAPE para los dos modelos y construir una tabla que permita comparar los dos modelos.
3. ¿Qué modelo es mejor entre modelo_2 y modelo_3?
4. ¿Qué tan bueno es el mejor modelo? Es decir, si utilizamos el mejor modelo, ¿qué tan grande es el error que vamos a tener, en promedio, a la hora de predecir la edad de un abalón?

Parte 3: Proyecto final

Ahora sí, llegó la hora de hacer analítica predictiva en Python. El objetivo de esta tercera entrega del proyecto es que construyas tu propio modelo de regresión para estimar la edad de un abalón.

Para este propósito, deberás usar las bases de datos abalone_test.csv y abalone_train.csv. El objetivo es que construyas distintos modelos para predecir la edad de un abalón 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 la edad de un abalón?
2. Toma los datos del conjunto de entrenamiento abalone_train.csv. La idea es entrenar 3 distintos modelos de predicción con estos datos. Recuerda que las variaciones a tus modelos pueden ser:

    a. Utiliza distintos modelos: regresión lineal univariada, regresión lineal multivariada, Random Forests con distintos parámetros
    b. Utiliza distintas variables en cada modelo. Por ejemplo, en un modelo puedes utilizar longitud, diametro, altura y peso_entero. En otro modelo puedes utilizar todas las variables.

c. Puedes construir nuevas variables.

Nota: si intentas entrenar los modelos con la variable sexo, vas a obtener un error, puesto que esta variable es categórica y toma los valores no numéricos F, M e I. Para utilizar esta variable, ejecuta primero estos dos líneas de código:

```
df_train[['sexo_F', 'sexo_I', 'sexo_M']] = pd.get_dummies(df_
train['sexo'])
df_test[['sexo_F', 'sexo_I', 'sexo_M']] = pd.get_dummies(df_
test['sexo'])
```

Con esto, estás agregando al DataFrame tres columnas con valores 0 y 1: sexo_F, sexo_I, sexo_M. Para estimar los modelos, debes incluir estas tres variables en vez de sexo.

3. Toma los datos del conjunto de validación abalone_test.csv. Para cada uno de los modelos del punto 2., haz la gráfica que muestra la edad real vs la edad predicha por cada modelo. Según el análisis gráfico, ¿qué modelo pareciera ser el mejor?
4. Calcula las métricas RMSE, MAE y MAPE y agrégalas en una tabla que permita comparar el desempeño de los tres modelos. ¿Qué modelo es mejor en predecir la edad de un abalón? ¿Qué tan bueno es el mejor modelo?tiene diabetes? ¿Qué tan bueno es el mejor modelo?