Quiz: Evaluación de modelos de clasificación con datos numéricos y categóricos

Puntos totales 90/100

Se ha registrado el correo del encuestado (**a21310402@ceti.mx**) al enviar este formulario.

0 de 10 puntos

Nombre(s) *

Luis Felipe

Apellidos * .../10

Hernandez Briseño

Sigue las instrucciones y contesta las preguntas. 90 de 90 puntos

Crea un notebook y abre el set de datos ames_housing_no_missing.csv adjunto en esta actividad utilizando el siguiente código:

```
import pandas as pd
ames_housing = pd.read_csv("ames_housing_no_missing.csv")

target_name = "SalePrice"
data, target = ames_housing.drop(columns=target_name), ames_housing[target_name]
target = (target > 200_000).astype(int)
```

La columna SalePrice contiene la variable objetivo que es el precio de venta de cada casa. El set de datos por naturaleza está diseñado para crear un modelo de regresión para la predicción del precio de venta de una casa. Sin embargo, lo convertiremos a un poblema de clasificación para predecir si una casa es costosa o no considerando que es costosa si el precio SalePrice supera los $200,000 dólares.

Utiliza data.info() y data.head() para examinar las columnas del dataframe y responde a la siguiente pregunta: 10/10

El set de datos contiene:

- ○ Solo características numéricas
- ○ Solo características categóricas
- ◉ Ambas, tanto características numéricas como categóricas

¿Cuántas características se tienen disponibles para predecir si una casa es costosa o no? 10/10

- ◉ 79
- ○ 80
- ○ 81

¿Cuántas características están representadas con números? Nota: puedes usar el método df.select_dtypes o la función sklearn.compose.make_column_selector para saberlo. 10/10

- ○ 0
- ◉ 36
- ○ 42
- ○ 79

De las siguientes columnas, ¿cuáles contienen valores numéricos cuantitativos, excluyendo categorías ordinales?. Selecciona todas las que apliquen. Para responder considera la descripción del set de datos disponible en este link https://www.openml.org/search?type=data&sort=runs&id=42165&status=active: 15/15

- [x] LotFrontage
- [x] LotArea
- [ ] OverallQual
- [x] YearBuilt

Crea un modelo predictivo utilizando las siguientes columnas numéricas: 15/15

numerical_features = ["LotFrontage", "LotArea", "MasVnrArea", "BsmtFinSF1", "BsmtFinSF2", "BsmtUnfSF", "TotalBsmtSF", "1stFlrSF", "2ndFlrSF", "LowQualFinSF", "GrLivArea", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces", "GarageCars", "GarageArea", "WoodDeckSF", "OpenPorchSF", "EnclosedPorch", "3SsnPorch", "ScreenPorch", "PoolArea", "MiscVal",]

El modelo predictivo deberá ser un pipeline que incluya sklearn.preprocessing.StandardScaler para escalar los valores numéricos, sklearn.linear_model.LogisticRegression para el modelo de clasificación y cross_validate para la validación del modelo.

¿Cuál es el accuracy obtenido con 10-fold en la validación cruzada?

- ( ) ~0.5
- ( ) ~0.7
- (•) ~0.9

Ahora crea un modelo predictivo pero utilizando las variables numéricas anteriores y las categóricas restantes. Crea un pipeline que procese las variables numéricas con StandardScalery las categóricas con OneHotEncoder. Para evitar cualquier problema con las categorías extrañas que pudieran presentarse solo durante la predicción, configura el parámetro handle_unknown="ignore" en el One-hot. Configura el parámetro max_iter=500 en la regresión logística.

20/20

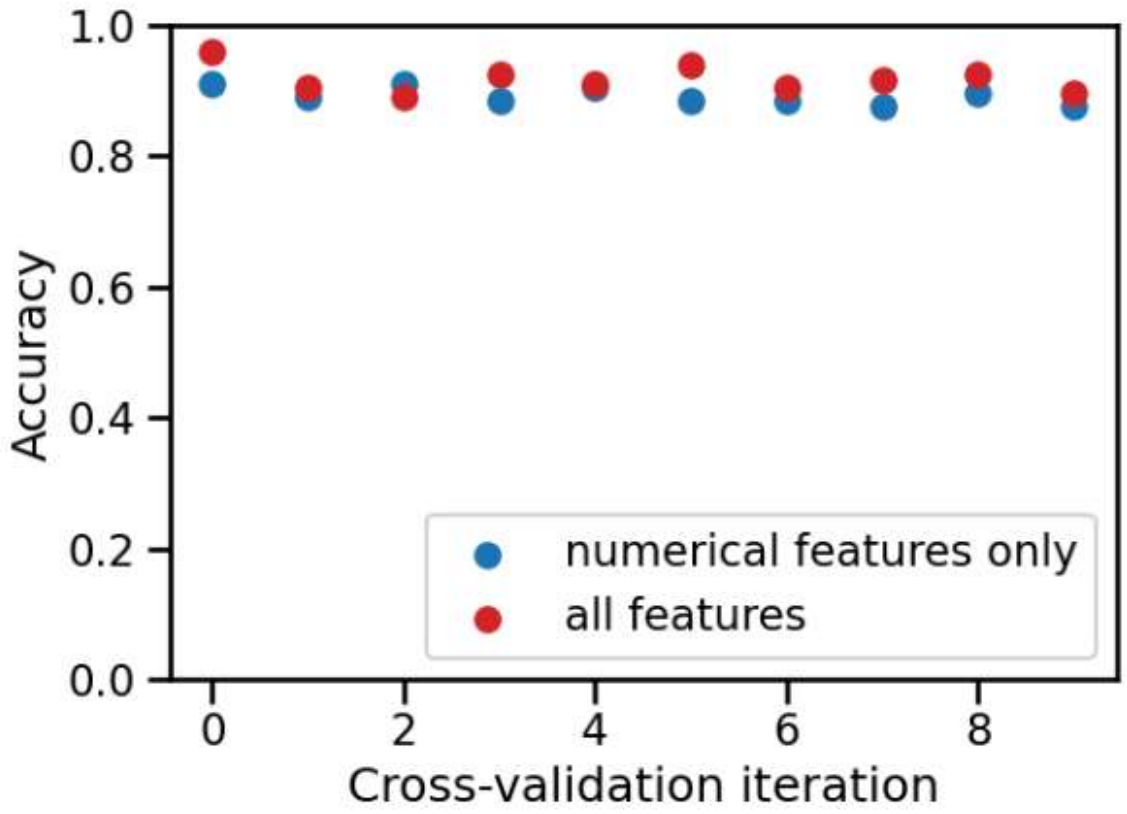
¿Cuál es el accuracy obtenido con 10-fold en la validación cruzada?

- ~0.7
- ~0.9
- ~1.0

Una forma de comparar dos modelos es con sus promedios de accuracy, como se hizo en el ejercicio anterior. Pero pequeñas diferencias en el accuracy entre los modelos podría haber sido por coincidencia. Otra forma de comparar los modelos es revisando cuando un fold tiene mejor resultado en un modelo que en el otro. Esto provee mayor información en cuanto a si una partición de los datos hace la clasificación más fácil o difícil para cada modelo. 10/10

El accuracy para cada modelo en cada fold se puede apreciar en la siguiente figura, en donde el fold 1 corresponde al valor 0 en el eje x y el fold 10 al 9.

El rango de folds en donde el modelo que utiliza todas las características (numéricas y categóricas) se desempeña mejor (tiene mejor accuracy) que el modelo que solo utiliza las variables numéricas es (considera que los folds van del 1 al 10:



- ○ [1,3]: el modelo que utiliza todas las características se desempeña constantemente peor
- ○ [4,6]: ambos modelos se desempeñan de forma similar
- ◉ [7,10]: el modelo que utiliza todas las carecterísticas es consistentemente mejor