FACULTAD DE CIENCIAS SOCIALES
LABORATORIO DE PROGRAMACIÓN: R Y PYTHON
2024-0

TRABAJO FINAL

Docente: Mauricio Vallejos
Correo: mauricio.vallejos@pucp.edu.pe

Entrenamiento de un modelo de Machine Learning usando la Encuesta Nacional de Hogares

Para este trabajo se le entregará una BBDD procesada de la Encuesta Nacional de Hogares (ENAHO).

Paso 1: Selección de variable de interés a elegir
Del set de variables disponibles tendrá que elegir una variable objetivo. Esta puede ser una continua como ingresos o gastos, o una dicotómica. No se elegirán categóricas para este ejercicio.

Paso 2: EDA (11 puntos)
Una vez elegida la variable de interés tendrá que realizar un análisis exploratorio de los datos tanto para la variable objetivo como para los predictores a usar en el entrenamiento de su modelo. En este paso debe realizar lo siguiente:

1. Presentación de estadísticas descriptivas de las variables como la media, mediana, desviación estándar, el mínimo, máximo y los percentiles
2. Evaluación de los valores *missings* existentes y recomendación sobre qué hacer con ellos
3. Evaluación de *outliers* en variables continuas y recomendación sobre qué hacer con ellos
4. Evaluación de la distribución (conteo de valores por categoría) de las variables binarias y categóricas, y análisis de los resultados observados
5. Visualización de datos bivariados entre variables continuas (gráfico de dispersión) y análisis sobre los resultados observados
6. Visualización y análisis de correlación con comentarios sobre los resultados observados

Paso 3: Entrenamiento del modelo (4 puntos)
A continuación, según la elección de su target, tendrá que realizar el ejercicio de entrenamiento de modelos. Sea en el contexto de regresión o clasificación, como mínimo tiene que entrenar dos modelos por caso (por ejemplo, si la variable es continua podría ser un modelo de regresión lineal y otro ridge). Guarde los modelos con nombres descriptivos y comente brevemente cómo funciona cada uno. Recuerde además que el uso de las variables depende del EDA realizado en el paso anterior. No tienen que entrar todas las variables, también puede crear nuevas a su gusto incorporando data externa o modificando las variables existentes.

Paso 4: Evaluación de métricas y elección del mejor modelo (5 puntos)
Con el modelo entrenado, en su data de testing, evalúe las métricas de performance para elegir el mejor modelo que se ajusta a sus datos. Presente una tabla resumen con las métricas de cada modelo y comente. Luego, justique la elección del modelo de su preferencia y comente dos utilidades que podría tener en el campo de al economía, finanzas o negocios.

IMPORTANTE: El archivo a ser entregado tiene que contener la estructura de carpetas con la inclusión de los archivos en estado raw, interm y final dentro de la carpeta *data*; el Jupyter notebook en el que se hace el procesamiento de la información dentro de la carpeta *programs*; y la presentación en la carpeta *Results*. La fecha de entrega será el miércoles 28 de febrero. La data será entregaba el martes 20 al finalizar el día.

Sobre la exposición

Las fechas de las exposiciones serán el lunes 26 y martes 27 de febrero. Se elaborará una presentación con la siguiente estructura.

1. Presentación del caso
2. Análisis de la data
3. Estimación de modelos
4. Evaluación de métricas y elección del modelo final

La duración es de máximo 10 minutos por grupo. Se recomienda ser breves No es obligatoria la inclusión de revisión de literatura.

Lima, 19 de febrero de 2024