FACULTAD DE CIENCIAS SOCIALES
LABORATORIO DE PROGRAMACIÓN: R Y PYTHON
2024-2

TRABAJO FINAL

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

Se entregará una base de datos de corte transversal correspondiente al año 2017, que incluye características sociodemográficas de un grupo de encuestados. A cada grupo se le asignará un departamento específico, sobre el cual deberán enfocar su proyecto. Una vez recibidas las asignaciones, podrán elegir entre dos tipos de proyectos:

- Proyecto econométrico: Identificar los principales determinantes de los ingresos laborales de los individuos encuestados.
- Proyecto machine learning: Entrenar el mejor modelo posible para predecir los ingresos laborales de los individuos encuestados.

Si optan por el primer proyecto, deberán realizar: i) el procesamiento y análisis de información en Python, y ii) estimaciones econométricas en R. En cambio, si eligen el segundo proyecto, realizarán: i) el procesamiento y análisis de información en R, y ii) el entrenamiento del modelo de machine learning en Python.

IMPORTANTE:

ENTREGA: Enviar su trabajo en un archivo comprimido .zip al correo mauricio.vallejos@pucp.edu.pe con el asunto y nombre de archivo "PYTHON_TF_XX", en la que XX es el número del grupo, hasta el domingo 15 de diciembre a las 11:59 p.m. como máximo. La carpeta debe de incluir el notebook o R script con el desarrollo de los ejercicios, así como cualquier documento adicional solicitado. **Si la tarea no es enviada con los formatos solicitados no será revisada.**

Parte 1: Ordenamiento de espacio y estructura de trabajo (1 puntos)

Se evaluará que las carpetas de trabajo sigan las recomendaciones de clase. Además, que se presente un código limpio con comentarios por secciones y asignaciones de variables bajo el esquema snake_case. El profesor debe de ser capaz de ejecutar sus proyectos usando la opción Run all o a través de un Rproject.

Parte 2: Procesamiento de información (9 puntos)

Realizaremos un análisis exploratorio de la base de datos entregada. A lo largo de cada etapa de esta sección, podrá modificar la codificación de las variables o construir nuevas según lo considere conveniente (por ejemplo, dummies, categóricas, logarítmicas, cruzadas, etc.). Además, enfocando algunos análisis en la variable objetivo (ingresos laborales), debe abordar los siguientes puntos (la creatividad en el análisis es un valor añadido):

- Imprime un resumen estadístico y descripción general de los datos (1 punto)
- Identifique los valores faltantes en el conjunto de datos. En caso de existir, abórdelos usando las recomendaciones de clase (2 puntos)
- Visualice la distribución de las variables continuas mediante histogramas o gráficos de densidad, y los recuentos de frecuencias para las variables discretas (binarias o categóricas) (2 puntos)
- Utilice boxplots para detectar la presencia de outliers en la variable objetivo. Evalúe individualmente y por cruces con categóricas. Ante la presencia de outliers, abordarlos según recomendaciones de clase. (2 puntos)
- Analice la relación entre las variables mediante una matriz de correlaciones (enfocado en la variable objetivo y para detectar multicolinealidad) y gráficos de dispersión (enfocado solo en la variable objetivo contra variables continuas). (2 puntos)
- Sobre los puntos anteriores, comente los patrones de información que considere más relevantes. Añada gráficos o estadísticos de creerlo necesario. (2 puntos)

Parte 3.1 (proyecto econométrico): Estimación de un modelo para encontrar los determinantes (10 puntos)

Ahora es el momento de validar estadísticamente los determinantes más importantes sobre los ingresos laborales. Recuerda que no solo puedes trabajar con los datos de la base de datos entregada, sino también realizar modificaciones o añadir datos nuevos desde otras fuentes. Al hacer las estimaciones, deben probar con al menos tres especificaciones de modelos. Todos deben estar bien identificados y presentar lo siguiente:

- Estimaciones a través de una regresión lineal multivariada. (1 punto)
- Validación de supuestos por cada modelo: heterocedasticidad, normalidad de errores y no multicolinealidad. (3 puntos)
- Evaluación de ajuste del modelo a nivel grupal e individual (1 punto)

- Interpretación precisa de coeficientes sujeto a significancia. Si usa logaritmos adecuar su análisis. (3 puntos)
- Creación de una matriz resumen de las estimaciones, que debe contener los coeficientes, su pvalue, y las pruebas de ajuste grupal. (2 puntos)

Parte 3.2 (proyecto machine learning): Entrenamiento de un modelo para la predicción de salarios laborales (10 puntos)

Ahora es el momento de usar los datos para la creación de un modelo que sea capaz de predecir los salarios laborales. Recuerda que no solo puedes trabajar con los datos de la base de datos entregada, sino también realizar modificaciones o añadir datos nuevos desde otras fuentes. Al hacer el entrenamiento, debes seguir el siguiente procedimiento (no olvides definir una semilla en cada proceso para que tus resultados puedan ser replicados):

- Despliegue de los datos para entrenamiento y testeo. Puedes usar una división 70/30 u 80/20. (1 punto)
- Entrenar de un modelo Lasso para la selección de variables más importantes. Este modelo debe ser entrenado con la selección de mejores parámetros a través de un GridSearch con métrica MAPE. (2 puntos)
- Usando las variables seleccionadas, entrenar de un modelo RandomForest con la selección de mejores parámetros. Utiliza GridSearch con métrica MAPE. (2 puntos)
- Reporte de resultados presentado las métricas MPE, MAPE, R2 y % de observaciones con +-20% de error. (1 puntos)
- Visualización de la distribución de los errores en el grupo de training y testing. (2 punto)
- Cross validation con métrica MAPE para evaluar la robustez del modelo. (2 puntos)

Lima, 28 de noviembre de 2024