CD3002C Inteligencia Artificial con Impacto Empresarial
Módulo 3 – Modelos de IA para Datos Estructurados
Febrero – Junio 2024

Actividad 1 – Modelos de Regresión

Instrucciones: Seleccionar una de las dos opciones de bases de datos i) automobile_insurance_claims o ii) health_insurance. A partir de dicha selección realizar las instrucciones 1 – 5. En el desarrollo del archivo de R-Markdown, por favor incluir *data storytelling* de los resultados del análisis exploratorio de los datos (EDA) así como la interpretación de los resultados estimados.

Lectura Sugeridas:

Supervised Machine Learning: Classification and Regression
https://medium.com/@nimrashahzadisa064/supervised-machine-learning-classification-and-regression-c145129225f8

What is Supervised Learning?
https://www.ibm.com/topics/supervised-learning

A Beginner's Guide to Supervised Machine Learning Algorithms
https://towardsdatascience.com/a-beginners-guide-to-supervised-machine-learning-algorithms-6e7cd9f177d5

1) Brevemente responder con tus propias palabras 2 de las siguientes 3 preguntas:

i) ¿Qué es Supervised Machine Learning y cuáles son algunas de sus aplicaciones en Inteligencia de Negocios?

ii) ¿Cuáles son los principales algoritmos de Supervised Machine Learning? Brevemente describir con tus propias palarbas 5 – 7 de los principales algoritmos de Supervised Machine Learning.

iii) ¿Qué es la R^2 Ajustada? ¿Qué es la métrica RMSE? ¿Cuál es la diferencia entre la R^2 Ajustada y la métrica RMSE?

2) Desarrollar Análisis Exploratorio de los Datos (EDA) que incluye los siguientes elementos:

a. Identificación de NA's
b. Reemplazo de NA's
c. Medidas descriptivas
b. Medidas de dispersión
c. Identificación de patrones y/o tendencias en los datos mediante el uso de gráficos incluyendo bar plots, line plots, pie plots, histogramas, matriz de correlación, box plot, scatter plot, qq-plot, etc Mostrar al menos 4 – 6 gráficos.

3) A partir de los resultados de EDA describir la especificación del modelo de regresión lineal a estimar. Brevemente, describir cómo es el posible impacto de cada una de las variables explicativas sobre la principal variable de estudio.

4) Estimación de cada uno de los siguientes modelos de Supervised Machine Learning (SML):

a. OLS Regresión
b. SAR
c. SEM
d. XGBoost Regresión
e. Decision Trees
f. Random Forest
g. Neural Networks Regresión

5) Pruebas de Diagnóstico de los Resultados Obtenidos de la Estmación de Modelos de Regresión

a. Multicolinealidad
b. Heterocedasticidad
c. Autocorrelación Serial
d. Autocorrelación Espacial
e. Normalidad de los Residuales

Nota: En caso de que las pruebas de diagnóstico identifiquen cualquiera de los anteriores a) – e) plantear una solución para mejorar la estimación de la especificiación del modelo.

6) Evaluación y Selección de Modelo de Regresión

a. Mediante el cálculo de la métrica RMSE para cada uno de los modelos estimados en 4) seleccionar el modelo que muestra los mejores resultados estimados.

b. Presentar los valores de la métrica RMSE de cada uno de los modelos estimados en 4) en un gráfico de barras.

7) Desarrollar una breve descripción de los 6 – 10 principales hallazgos de:

a. EDA

b. Modelo seleccionado:

i. ¿Cuáles son las variables que contribuyen a explicar los cambios de la principal variable de estudio?

ii. ¿Cómo es el impacto de dichas variables explicativas sobre la variable dependiente?

iii. ¿Los resultados estimados del modelo seleccionado son similares a los otros modelos estimados? ¿Cuáles son las diferencias?

Fecha de Entrega: Lunes 4 de Marzo 2024 a las 11:59 PM (Vía Canvas)
Formato de Entrega: R – Markdown (html o pdf)
Formato de Entrega: Individual | Incluir Nombre Completo al inicio del archivo