Pontificia Universidad Católica de Chile
Facultad de Ciencias Económicas y Administrativas

Segundo Semestre 2022

Curso	:	**Machine Learning para Negocios**
Sigla	:	**EAA3707**
Profesores	:	**M Ignacia Vicuña**

Tarea 1

Pregunta 1:

`TidyTuesday` es un proyecto semanal de R for Data Science online learning community cuyo objetivo es proporcionar a los usuarios de R distintos conjuntos de datos reales para que puedan aplicar sus habilidades de análisis de datos en un contexto del mundo real.
En esta oportunidad, trabajaremos con los datos `transit_cost.csv`, donde cada observación es un proyecto de construcción de transporte público. Entre la información interesante que contiene, está el año de inicio de la construcción, el año de finalización del proyecto, la ciudad en la que se desarrolla, la longitud de la línea de tránsito, la longitud del túnel, si es que lo hay, y el costo por kilómetro. La Tabla 1 contiene la descripción de todas las variables que contiene el conjunto de datos.

A partir de un modelo de regresión lineal múltiple, se intentará modelar el costo real del proyecto de tránsito en función de variables predictoras.

(a) Realice un análisis bivariado de las variable `real_cost` con todas las variables independientes. Este análisis debe contener los gráficos de dispersión de cada variable independiente con la variable objetivo y además sus correlaciones de pearson. Interprete los resultados.

(b) A partir del análisis bivariado realizado en (a), seleccione aquellas variables cuya correlación con la variable objetivo sea superior a 0.3.

(c) Con la librería Tidymodels de R ajuste un modelo de regresión lineal múltiple con las variables seleccionadas en el paso anterior. Para ello realice lo siguiente:

- Utilice la semilla set.sed(3707) y realice una partición al 75 % para los datos de entrenamiento.
- Ajuste un modelo de regresión lineal múltiple y reporte los coeficientes estimados del modelo. Con un nivel de significancia al 5 % ¿Qué variables predictoras son significativas?
- Con la muestra de validación, reporte las medidas de bondad de ajuste MAPE, R2.
- Realice un gráfico de dispersión entre el costo real del proyecto y su valor predicho.
- Ajuste nuevamente la regresión utilizando 10-fold cross validación, y reporte las medidas de bondad de ajuste MAPE, R2. Comente y compare con lo obtenido sin reemuestreo.

Pregunta 2:

El boostrap es una técnica de remuestreo que permite estimar el error estándar de un estimador como también un intervalo de confianza, sin asumir que los datos provienen de alguna distribución paramétrica. Este método es muy utilizado en estadística, sobre todo cuando el supuesto de Normalidad de los datos no se cumple o cuando el tamaño de la muestra no es tan grande. Si θ es una cantidad desconocida de la población que se desea estimar,

y $\hat{\theta}$ es un estimador puntual de él, el intervalo de confianza Bootstrap para θ, basado en el método de percentil, se obtiene estimando la función distribución F del estimador $\hat{\theta}$ a partir de un remuestreo de la muestra original.

A partir de los datos `transit_cost.csv` se estimará θ, que representa el costo real medio de los proyectos de tránsito, utilizando boostrap. Realice los siguientes pasos:

(a) Utilice la semilla set.seed(3770) y genere $B = 1000$ muestras bootstrap de a partir de los datos. Las muestras deben ser del mismo tamaño n que la muestra original.

(b) Para cada una de las muestras bootstrap, calcule la media de costo del proyecto, denótelo por $\hat{\theta}_i^*$, $i = 1, ..., B$.

(c) Calcule el promedio y la desviación estándar de las B estimaciones, y denótelo por $\hat{\theta}_{boot}$ y se_{boot}. Compare con la estimación de la media y su error estándar obtenido en la muestra original.

(d) Construya un histograma con las estimaciones bootstrap $\hat{\theta}_i^*$ y superponga la densidad estimada utilizando la función `density( )`. Esta sería la distribución estimada ($\hat{F}$) del estimador $\hat{\theta}$.

(d) Calcule el intervalo de 95 % confianza Bootstrap, el cual se obtiene a partir de

$$[\hat{F}^{-1}(\alpha/2), \hat{F}^{-1}(1-\alpha/2)]$$

donde, $\hat{F}^{-1}(\alpha)$ es el percentil α de la distribución estimada del estimador $\hat{\theta}$.

Nota: Utilice la función `quantile( )` de R para calcular los percentiles de $\hat{F}$.

(e) Calcule el intervalo de 95 % de confianza para θ utilizando la función `t.test( )`. ¿La construcción del intervalo `t.test( )` asume algún supuesto sobre la distribución de $\hat{\theta}$? Compare lo obtenido con el intervalo bootstrap. ¿Cuál intervalo le parece mejor para estimar θ? Justifique..

Pregunta 3:

Para los bancos es de gran importancia contar con un modelo predictivo para otorgar o no un crédito a un nuevo cliente. A partir de datos históricos de clientes, se registra la variable `loan_status`, que toma el valor 1 si el cliente incumplió con el pago, y 0 sino. A partir de las variables predictoras `log_annual_income` que corresponde al logaritmo natural de los ingresos anuales y `home_ownership` que tiene 4 niveles (Mortgage, Other, Own, Rent), se ajustó un modelo de regresión logística obteniéndose el siguiente modelo:

$$\begin{aligned} ln\left(\frac{\pi}{1-\pi}\right) &= 0.565 + 0.084 \cdot \texttt{home\_ownershipOTHER} - 0.008 \cdot \texttt{home\_ownershipOWN} + \\ &\quad 0.007 \cdot \texttt{home\_ownershipRENT} - 0.042 \cdot \texttt{log\_annual\_income} \end{aligned}$$

(a) Calcule la probabilidad de caer en incumplimiento para un nuevo cliente que tiene casa propia y su salario anual en escala logarítmica es de 11 USD.

(b) ¿Cuanto debiera ser el salario anual de un nuevo cliente que arrienda su propiedad, para que su probabilidad de caer en incumplimiento sea de 15 %?

(c) Se construye un clasificador binario utilizando el punto de corte 0.15. Con los datos con que se ajustó el modelo se calcula la matriz de confusión, obteniéndose

	real	
pred	0	1
0	21864	2594
1	915	198

Calcule la precisión del modelo, la sensitividad y la especificidad.

(d) Calcule la razón de chances de caer en incumplimiento para una persona que arrienda una propiedad versus una persona que es dueña de su propiedad. Interprete el resultado.

Tabla 1: Descripción Variables del dataframe transit cost

Variable	Description
e	ID
country	Country Code
city	City where transit tunnel is being created
line	Line name or path
start_year	Year started
end_year	Year ended (predicted or actual)
rr	This is Railroad (0 or 1), where 1 == Railroad.
length	Length of proposed line in km
tunnel_per	Percent of line length completed
tunnel	Tunnel length of line completed in km
stations	Number of stations where passengers can board/leave
source1	Where was data sourced
cost	Cost in millions of local currency
currency	Currency type
year	Midpoint year of construction
ppp_rate	purchasing power parity (PPP), based on the midpoint of construction
real_cost	Real cost in Millions of USD
cost_km_millions	Cost/km in millions of USD
source2	Where was data sourced for cost
reference	reference