SPIKE

Spike Challenge - Predicción de precios de insumos básicos en Chile

¡Gracias por participar en el proceso de selección de Spike! Como parte del proceso, este desafío nos ayudará a entender la manera en que te enfrentas a problemas nuevos y, además, podremos evaluar tus conocimientos actuales.

Algunos puntos importantes,

1. Este desafío no te debiera tomar más de 5 horas de tu tiempo. Por lo mismo, no esperamos respuestas muy pulidas ni perfectas.
2. Las preguntas irán aumentando en dificultad, por lo que intenta responder hasta donde puedas. Si por algún motivo hay alguna parte que no lograste completar, no hay problema.
3. Tendrás hasta el lunes 19 de abril a las 23:59 para enviar tus respuestas al desafío.
4. Solo se aceptarán *Jupyter notebook* (recomendado), *R Markdown* o *R Notebook* como formatos de entrega y solamente *python* o *R*. La idea es que sea fácil para nosotros correr lo que ustedes escribieron (que sea reproducible).
5. Lee bien las instrucciones!

Accede a este link para encontrar las instrucciones y el dataset para el desafío:
https://github.com/SpikeLab-CL/desafio_spike_precios

Saludos!
Spike

Nota: Este desafío es optativo si ya has hecho un desafío con Spike. Si quieres que revisemos un desafío anterior tuyo, avísanos!

En este desafío vamos a ver si somos capaces de predecir el precio de un insumo básico, como la leche, a partir de variables climatológicas y macroeconómicas. No siempre estos datos nos entregan toda la información que nos gustaría, como por ejemplo señales claras del avance de la sequía a lo largo del país, sin embargo, nos permite entender otro tipo de efectos, como movimientos en ciertos sectores de la economía. En esta línea, te iremos guiando para construir un análisis y algunos modelos que nos ayuden a concluir.

1. Datos: Precipitaciones, Indicadores Económicos Banco Central

- Cargar archivo *precipitaciones.csv* con las precipitaciones medias mensuales registradas entre enero 1979 y abril 2020. (Unidad: mm).
- Cargar archivo *banco_central.csv* con variables económicas.

2. Análisis de datos. Creación de variables

- Realiza un análisis exploratorio de la base de datos, ¿Qué puedes decir de los datos, sus distribuciones, valores faltantes, otros? ¿Hay algo que te llame la atención?
- Realiza una limpieza de datos para que las series de tiempo no tengan duplicados ni valores incorrectos.

3. Visualización

- Crea una función que permita graficar series históricas de precipitaciones para un rango de fechas determinado. Para esto la función debe recibir como argumentos el nombre de una región, fecha de inicio y fecha de término (asegúrate de verificar en tu función que tanto el nombre de la región como las fechas ingresadas existan en el dataset).
- Usa esta función para graficar las precipitaciones para la Región Libertador General Bernardo O'Higgins y para la Región Metropolitana entre las fechas 2000-01-01 y 2020-01-01.
    - ¿ Qué observas con respecto a estacionalidades y tendencias?

- Crea una función que, para una región, grafique múltiples series de tiempo mensuales de precipitaciones, donde cada serie de tiempo corresponda a un año. La función debe recibir como argumento una lista con los años que queremos graficar (2000, 2005,..) y el nombre de la región. El eje X debe indicar los meses (enero, febrero, etc…).
- Usa esta función para graficar las precipitaciones para la Región del Maule durante los años 1982, 1992, 2002, 2012 y 2019.
    - ¿Qué puedes concluir de estos gráficos?

- Crea una función que permita visualizar dos series históricas de PIB para un rango de fechas determinado. Para esto la función debe recibir como input el nombre de cada serie, fecha de inicio y fecha de término.
- Grafica las series de tiempo del PIB agropecuario y silvícola y la del PIB de Servicios financieros desde el 2013-01-01 hasta la fecha más reciente en que haya datos.
    - ¿Qué puedes decir de cada serie en particular?
    - ¿Hay alguna relación entre estas dos series?

4. Tratamiento y creación de variables

- ¿Cómo podríamos evaluar la correlación entre las distintas series de tiempo y cómo se tienen que correlacionar para entrenar un modelo? ¿Mucha correlación, no correlacionadas, da igual?
- Para el entrenamiento del modelo, queremos predecir el precio de la leche para el productor en Chile. Para eso, descarga el archivo *precio_leche.csv* y haz un merge con las bases de datos de precipitaciones y datos del Banco Central.

  *Este archivo tiene una columna de año, mes y *precio_leche* (que corresponde al precio nominal, sin IVA, en pesos chilenos por litro), por lo que vas a tener que crear la columna de fecha que calce con la de las otras bases.

- Crea las variables:
  - A partir de la variable fecha, crea nuevas variables para el año, mes, trimestre.
  - Lags y estadísticas acumuladas (por ejemplo: promedio, varianza) de las variables que consideres relevantes.

5. Modelo

- Entrena un modelo que permita predecir el precio de la leche el próximo mes, en función de los datos entregados.
  - Si necesitas crear variables adicionales que pueden aportar información al modelo, tienes total libertad.
- Construye una base de test (o de cross validation). ¿Cuál fue tu definición de tiempo/cantidad de datos para este set de datos? Explica por qué la elegiste así.
- ¿Qué datos adicionales te gustaría tener?¿Qué datos son necesarios para que este modelo funcione/mejore las métricas?
- ¿Cómo evalúas el resultado del modelo?¿Qué métricas tiene sentido mirar?
- ¿Para qué aplicaciones puede servir un modelo de este tipo? En particular, ¿Cómo podría ayudar a combatir el cambio climático?