SPIKE

Spike Challenge - Predicción de atrasos de vuelos en SCL

¡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 responde hasta donde puedas. Si por algún motivo hay alguna parte que no lograste completar, no hay problema.
3. Tendrás hasta el Martes 16 de Octubre a las 23:59 para enviar tus respuestas al desafío.
4. Solo se aceptarán *ipynb*, *rmarkdown* o *rnotebook* como formatos de entrega y solamente *python* o *R*. La idea es 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://s3.amazonaws.com/spike-challenge/SpikeChallengeOct2018.zip

Saludos!
Spike

Instrucciones

Como dice el título, vamos a predecir la probabilidad de atraso de los vuelos que aterrizan o despegan del aeropuerto de Santiago de Chile, SCL. Para eso armamos un dataset usando datos públicos y reales de la Dirección General de Aeronáutica Civil (DGAC), donde cada fila corresponde a un vuelo que aterrizó o despegó de SCL durante todo el 2017. Para cada vuelo se cuenta con la siguiente información:

- Fecha-I: Fecha y hora programada del vuelo.
- Vlo-I: Número de vuelo programado.
- Ori-I: Código de ciudad de origen programado.
- Des-I: Código de ciudad de destino programado.
- Emp-I: Código aerolínea de vuelo programado.
- Fecha-O: Fecha y hora de operación del vuelo.
- Vlo-O: Número de vuelo de operación del vuelo.
- Ori-O: Código de ciudad de origen de operación
- Des-O: Código de ciudad de destino de operación.
- Emp-O: Código aerolínea de vuelo operado.
- DIA: Día del mes de operación del vuelo.
- MES: Número de mes de operación del vuelo.
- AÑO: Año de operación del vuelo.
- DIANOM: Día de la semana de operación del vuelo.
- TIPOVUELO: Tipo de vuelo, I=Internacional, N=Nacional.
- OPERA: Nombre de aerolínea que opera.
- SIGLAORI: Nombre ciudad origen.
- SIGLADES: Nombre ciudad destino..

Desafío

1. Analiza el dataset *baseSCL2017.csv*. ¿Qué puedes decir de los datos, distribuciones, missing, etc? ¿Hay algo que te llame la atención? Entregable: texto/imágenes.
2. Genera las siguientes columnas adicionales:
    a. Periodo_dia: mañana (entre 5:00 y 11:59), tarde (entre 12:00 y 18:59) y noche (entre 19:00 y 4:59), en base a Fecha-I.
    b. Flag_temporada_alta: 1 si Fecha-I está entre 15-Dic y 3-Mar, o 15-Jul y 31-Jul, o 11-Sep y 30-Sep, 0 si no.
    c. Dif_min: diferencia en minutos entre Fecha-O y Fecha-I.
    d. Atraso15: 1 si Dif_min>15, 0 si no.
    Entregable: csv con dataset y código (si es que usaste).
3. ¿Cómo compone la tasa de atraso por destino, aerolínea, mes del año, día de la semana, temporada, tipo de vuelo? ¿Qué variables esperarías que más influyeran en predecir atrasos? Entregable: texto/imágenes.
4. Entrena uno o varios modelos (usando el/los algoritmo(s) que prefieras) para estimar la probabilidad de atraso de un vuelo. Siéntete libre de generar variables adicionales y/o complementar con variables externas. Entregable: código de entrenamiento.
5. Evalúa tu modelo. ¿Qué performance tiene? ¿Qué métricas usaste para evaluar esa performance y por qué? ¿Por qué elegiste ese algoritmo en particular? ¿Qué variables son las que más influyen en la predicción? ¿Cómo podrías mejorar la *performance*? Entregable: texto/imágenes.

Envía un .zip conteniendo todas las respuestas, con el formato ¨[Apellido-Nombre].zip¨ a jobs@spikelab.xyz usando el asunto: “Spike Challenge”.