DESAFÍO 1

DESCRIPCIÓN

El analista de datos, cuando desarrolla un proyecto, pasa la mayor parte del tiempo preparando los datos para que estén listos para crear visualizaciones y análisis poderosos. Más específicamente, realizando la limpieza y transformación.
Por eso es muy importante que el analista de datos desarrolle habilidades de tratar los datos y dejarlos listos para la parte más interesante: crear modelos y análisis.
Este desafío se trata de desarrollar esta habilidad, para eso debe aplicar procesos de tratamiento y limpieza de datos.

TEMA: Analizar los gastos de los parlamentarios paraguayos
Fuente de datos: datos que los legisladores declararon por año
https://datos.hacienda.gov.py/data/pgn-gasto
Entidad: 001-CONGRESO NACIONAL
http://datos.congreso.gov.py/opendata/index
https://www.controlciudadanopy.org/

Los datos pueden contener una serie de problemas que pueden dificultar la creación de análisis más profundos. Una de las primeras cosas que debe realizar es identificar tales inconsistencias, como campos que poseen valores nulos o duplicados, convertir campos de fecha que son cargados como texto, formatar campos como RUC etc.

¿Qué tal juntar datos de varios años en un único dataset y aplicar técnicas de limpieza y procesamiento de los datos?
Podría obtener datos de los últimos cuatro años y aplicar lo que uso en este ejercicio.
Recomiendo que documente el proceso realizado. Así, cualquier persona que consulta su trabajo sabrá cuál fue su intuición y las técnicas utilizadas, además de facilitar la reproducibilidad.
No existe una receta para las técnicas que debe utilizar en la limpieza de datos, eso varía de proyecto para proyecto. Sin embargo, a continuación algunas cosas que puede hacer inicialmente:

- Lidiar con datos nulos (ej. borrar o imputar un valor nuevo)
- Remover columnas que no traen ninguna información
- Procesar fechas que están en formato incorrecto
- Alterar el tipo de columna (ej. una columna que es número pero está como texto)
- Remover duplicados.

DESAFÍO 2

DESCRIPCIÓN

La comunicación es una de las habilidades más importantes de un científico de datos.
Saber comunicar efectivamente sus resultados de una forma que las personas del negocio y no técnicas consigan entender es una de las soft skills más apreciadas del mercado, y también una de las más difíciles de dominar.
Una de las formas más efectivas de comunicarse está en una técnica llamada ***Storytelling*** (Contar historias, traducción libre). Esto es, crear una narrativa sobre su trabajo para que sea más fácil entender su raciocinio y entender sus resultados. La visualización de los datos no puede faltar en este tipo de comunicación. La combinación de buenos gráficos y visualizaciones con una historia bien contada, es capaz de hacer a cualquier audiencia entender los resultados de su trabajo - sea la explicación del resultado de un test de hipótesis o un análisis estadístico, o la comunicación de resultados de su modelo de Machine Learning. Podemos comenzar a investigar los datos y generar visualizaciones para identificar patrones extraños o estadísticas interesantes.

Piense en las preguntas que los datos pueden responder. Estos son los indicadores que muestran información interesante. Por ejemplo, si analizamos los datos de los parlamentarios podríamos querer ver: ¿cuál rubro declaró más despensas? Cuál fue el porcentaje que los parlamentarios gastaron? ¿hubo algún rubro con más del 90%? En año de elecciones los parlamentarios gastan más? Entre otras preguntas.
Puede comenzar a generar algunas estadísticas bien simples, como contar valores de alguna columna, sumarlas, hacer agrupaciones, ordenarlas, etc.

A partir de estos análisis macro, conseguirá filtrar y hallar una buena historia.

Puede utilizar herramientas como Power BI, Tableau, Excel o Knime para visualizar sus datos y crear interacciones con el usuario.
Un consejo importante al contar historias es disminuir al máximo la carga cognitiva de su audiencia. En otras palabras, no queremos que nuestros espectadores piensen mucho.
Cuando un gráfico está bien estructurado es capaz de explicar varios minutos sobre lo que significa el eje X e Y.

DESAFÍO 3

DESCRIPCIÓN

El trabajo del científico de datos generalmente gira en torno a dos frentes principales: descripción (entender qué sucedió y porqué sucedió) y predicción (prever lo que puede suceder y lo que debe ser realizado).

Es probable que en el desafío anterior haya aplicado algunas de esas técnicas descriptivas, utilizando análisis de datos y visualizaciones para entender lo que sucedió. Por ejemplo, en el desafío anterior usted analizó los gastos de los legisladores.

Para las tareas de predicción, son utilizadas técnicas de Machine Learning y herramientas estadísticas para prever el futuro.

Forecasting es una de las técnicas más conocidas y utilizadas por los científicos de datos para prever indicadores de negocios. Por ejemplo, en el área de finanzas puede querer que usted realice una previsión sobre el fracturamiento de la empresa; o entonces su equipo de marketing puede querer prever cuántos usuarios irán registrarse en la plataforma en los próximos meses.

En este desafío deberá crear su propio modelo de forecasting.

TEMA: Crear un modelo que consiga prever cuanto los senadores van gastar en los próximos tres meses.

Puede crear su propio dataset, pero será muy importante que tenga datos de más de un año. Por ejemplo de 4 años.

Puede utilizar el siguiente dataset: dataset_ceaps_forecasting. Los datos que debe utilizar son de DS (fecha de reembolso), Y (suma de reembolsos de los legisladores de aquel día).

DS	Y
2018-05-30	116443.69
2018-05-31	106078.24
2018-06-01	159585.05
2018-06-02	17243.94

CONSEJO: Comience sus previsiones de forma simple. Un modelo inicial (también llamado de baseline) es utilizar promedios. Por ejemplo, para los próximos tres meses, puede decir que el valor de “y” será el promedio de los últimos 3 meses.

Luego use métodos autorregresivos, ARIMA, Regression Linear y Machine Learning.

Para evaluar su modelo, debe utilizar una métrica. Para problemas de Forecasting, puede utilizar métricas de problemas de regresión, como RMSE y MAPE.