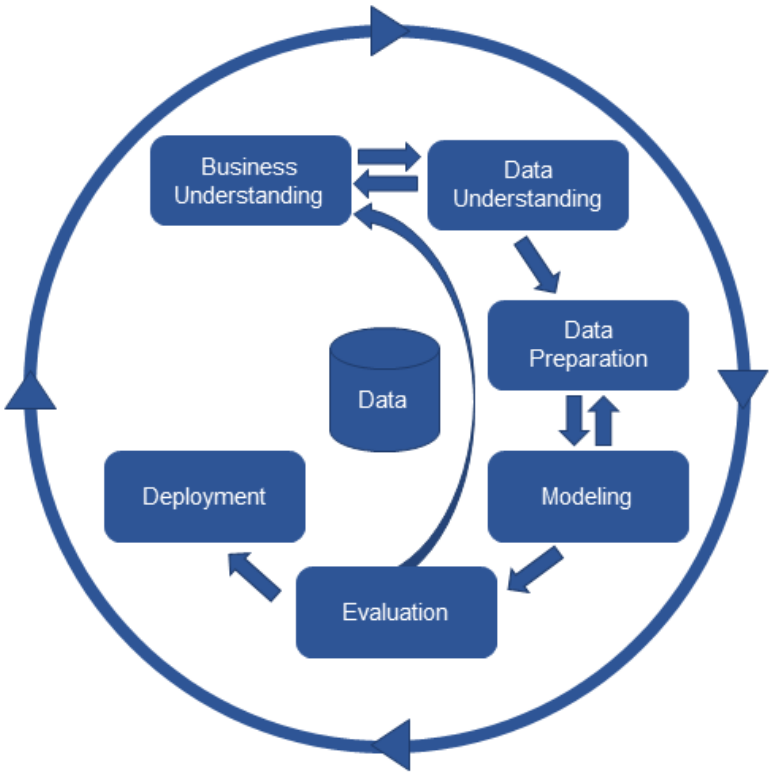
Ejercicio para la posición de Científico de Datos

1. Introducción:

Partiendo del framework CRISP-DM (Cross Industry Standard Process for Data Mining) y usando la herramienta Dataiku (https://www.dataiku.com/), se busca que el candidato participe en una competencia de Kaggle siguiendo los pasos del framework de CRISP-DM construya un modelo predictivo y se evalúe su desempeño.



Framework CRISP-DM

CRISP-DM

Entendimiento del negocio:
Esta fase se enfoca en comprender el objetivo y los requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición de problema y un plan preliminar diseñado para alcanzar los objetivos.

Comprensión de los datos:
Esta fase comienza con una recopilación de datos inicial y continúa con las actividades para familiarizarse con los datos, **identificar** problemas de calidad de los datos, descubrir las primeras perspectivas de los datos o detectar subconjuntos interesantes y formar hipótesis.

Preparación de los datos:
Esta fase cubre todas las actividades para construir el conjunto de datos final (datos que se incorporarán a las herramientas de modelado) a partir de los datos sin procesar iniciales. Las tareas generalmente incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para herramientas de modelado.

Modelado
En esta fase, varias técnicas de modelado y algoritmos se seleccionan y aplican al conjunto de datos, así la calibración de valores óptimos de parámetros e hiperparámetros.

Evaluación:
En esta etapa, es importante evaluar más detalladamente el modelo y revisar los pasos ejecutados para construir el modelo, con el fin de asegurarse de que logre los objetivos comerciales.

Despliegue:
La fase de implementación puede ser tan simple como generar un informe y comunicar resultados o tan compleja como implementar un modelo en un entorno de producción.

2. Objetivo del desafío:

El ejercicio propuesto se evalúa en *Kaggle*.

Kaggle es una plataforma de competencia para la ciencia de datos. Se estará trabajando con la competencia de Kaggle: "Telstra Network Disruption" (https://www.kaggle.com/c/telstra-recruiting-network) y la herramienta para el análisis de datos Dataiku.

El objetivo de esta competencia es predecir la gravedad de la falla de la red de telecomunicaciones de Telstra en un momento determinado en una ubicación determinada en función de los datos y registros disponibles.

Para poder participar es necesario:

1.- Crear una cuenta de Kaggle.
2.- Ir a la página de la competencia dar click en "Join Competition" y aceptar los términos y condiciones.
3.- Raken Data Group te enviará un usuario de DataIku, en donde desarrollaras el ejercicio (mayor detalle en la sección 6).

3. Datos:

En el apartado de “Data” podemos encontrar los datos para la competencia. Cada fila en el conjunto de datos principal (train.csv, test.csv) representa una ubicación y un punto de tiempo. Se identifican por la columna “id”, que es la clave “id” utilizada en otros archivos de datos. La severidad de la falla tiene 3 categorías: 0, 1, 2 (0 significa que no hay falla, 1 significa solo unas pocas y 2 significa muchas). Se extraen diferentes características de los archivos de los registros (logs) y otras fuentes: event_type.csv, log_feature.csv, resource_type.csv, severity_type.csv.

Archivos

- train.csv - el conjunto de entrenamiento para la severidad de la falla
- test.csv - el conjunto de prueba para la severidad de la falla
- sample_submission.csv – una muestra del formato correcto para la entrada
- event_type.csv: tipo de evento relacionado con el conjunto de datos principal
- log_feature.csv - características extraídas de los archivos de registro
- resource_type.csv: tipo de recurso relacionado con el conjunto de datos principal
- severity_type.csv: tipo de severidad de un mensaje de advertencia que proviene del registro

4. Herramienta (Dataiku)

Para poder realizar el análisis de datos, se debe utilizar la herramienta Dataiku para lo cual se proveerá con un usuario para poder hacer uso de los recursos de Raken (http://raken.ddns.net:11200). La herramienta es muy fácil de usar y tiene capacidades de llevar todo el ciclo de ciencia de datos de principio a fin. Tutoriales para el manejo de la herramienta se pueden encontrar en la siguiente liga:
https://www.dataiku.com/learn/portals/tutorials.html

5. Proceso

En tu rol de científico de datos, el objetivo de este ejercicio es realizar el proceso de ciencia de datos tocando todos los puntos dentro del framework de CRISP-DM utilizando el conjunto de datos de la competencia de Kaggle para finalmente obtener un puntaje para el modelo predictivo. Durante el proceso (Entendimiento del negocio, entendimiento de los datos, etc.) se debe ir llevando una bitácora de cada uno de los puntos para poder ser entregado en un documento final. Al final se busca que se realice un modelo predictivo y se genere un archivo el cual se suba a la plataforma de Kaggle para obtener un puntaje de valoración del modelo predictivo. Puedes subir cuantas veces quieras el resultado del modelo a Kaggle hasta que llegues al mejor resultado que puedas obtener.

Si aplicas como Científico de Datos tienes una semana para resolver el ejercicio. Si eres becario tienes hasta 2 semanas.

**Como becario entendemos que estás incorporándote en una nueva aventura. No te estreses si no puedes por completo, pero queremos ver coherencia y que te esforzaste por resolverlo.*

6. Entregables:

Al final debes generar una presentación **ejecutiva** sobre lo que realizaste, el objetivo es que presentes tus hallazgos, conclusiones de manera ordenada, contando una historia de principio a fin para usuarios no técnicos (imagina que estarás presentando a los(as) jefes de los jefes de tu jefe(a)).

El documento(te sugerimos presentación en la tecnología que quieras) deberá de contener:

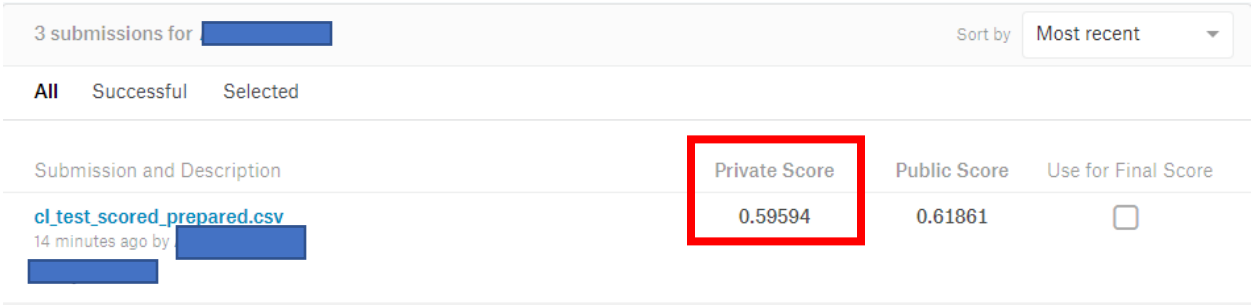
1. Nombre completo acompañado de un apodo o alias
2. Fecha de Inicio
3. Fecha de Entrega
4. Descripción a alto nivel de la solución
5. Historia
6. Screenshot de tu mejor score en kaggle
7. Insights relevantes que hayas encontrado.

Este documento lo expondrás ante varias personas. Sugerimos que te prepares para saltar de tu presentación al código.

7. Criterios de evaluación

Es fundamental que recuerdes que no sólo son matemáticas y algoritmos, si bien es fundamental no es lo más importante en este ejercicio. El resultado, la presentación y la consistencia son los focos de evaluación. Por ello presta especial atención a los siguientes puntos:

1. Contenido de tu presentación
2. Limpieza en el documento
3. Redacción coherente y sencilla
4. Entendimiento del problema
5. Orden en el flujo de DataIku (al menos el Data Prep)
6. Score de Kaggle (Private Score) → Mientras menor sea es mejor. El límite máximo permitido es de 0.7, es decir, tu score debe de ser inferior a 0.7. Si aplicas como becario omite este punto aunque si lo alcanzas es mejor =)



7. Explicación del modelo seleccionado
    a. Qué hace el modelo
    b. Por qué lo seleccionaste
    c. Contra que otros modelos compitió
    d. Modificaste algún parámetro sí/no ¿por qué?
    e. ¿Para qué utilizaste el set de pruebas? ¿Qué hace?

Eres libre de usar los algoritmos y componentes de Dataiku o programar con R/Python. La libertad creativa es fundamental.

Si bien es una plataforma que puede que no conozcas verás lo fácil y sencilla que es. Suena a mucho pero no es tanto, tranquilízate y verás que eres capaz de resolverlo, tienes buen tiempo.

Te deseamos mucha suerte y éxito.

El equipo de DS de Raken Data Group

=)