sumz

Correo de bienvenida

¡Bienvenido al proceso de selección para el cargo de Científico/Científica de Datos en Sumz!
Tu rol principal será desarrollar y/o acompañar el desarrollo de modelos analíticos, estadísticos y probabilísticos para nuestros clientes, asegurándose de que sean efectivos y contribuyan al logro de los objetivos del negocio.

El proceso de selección consta de 4 etapas:

1. **Primera etapa:** entrevista con talento humano.

2. **Segunda etapa:** entrevista técnica con Juan Camilo Azuero, Líder del CoE (Center of Excellence).

3. **Tercera etapa:** Prueba de conocimiento que será enviada por correo electrónico. Se proporcionará la información contextual y los insumos necesarios para la prueba de conocimiento en la comunicación.

4. **Cuarta etapa:** presentación de hallazgos, conclusiones y recomendaciones en una simulación de Comité de Entrega de Resultados.

*En caso de no continuar avanzando en alguna de las etapas del proceso, el equipo de talento humano te notificará.

¡Te deseamos mucha suerte en el proceso de selección y esperamos tener noticias tuyas pronto!

sumz

Prueba técnica

En Sumz, como consultora analítica, nos enfrentamos al constante desafío de desarrollar modelos predictivos para clientes de diferentes sectores económicos. Para esta prueba trabajamos con un cliente del sector retail que necesita un modelo de estimación de demanda para su comercio. La disponibilidad de los productos es fundamental para ofrecer un excelente servicio al cliente, lo que implica considerar dos factores críticos: evitar la falta de stock y no excederse en el abastecimiento de unidades. Así, nuestro objetivo es construir una proyección de demanda precisa que permita optimizar el inventario.

Por lo tanto, te desafiamos a crear el mejor modelo de estimación de unidades vendidas para cada uno de los artículos y para cada uno de los días del año. Este modelo permitirá planear la frecuencia y el volumen de las compras necesarias para garantizar la disponibilidad de los productos requeridos sin incurrir en la rotura de stock o en el exceso de aprovisionamiento.

Nos gustaría ver tu enfoque en el desarrollo de este modelo y cómo abordarías los desafíos específicos de este caso. ¡Buena suerte, esperamos ver tus resultados!

Descripción

El objetivo de esta prueba es evaluar su capacidad para analizar y desarrollar modelos analíticos en un tiempo limitado. Se espera que no le dedique más de 10 horas para completar la prueba, lo que incluye el tiempo para documentar su proceso. Puede utilizar cualquier herramienta de su preferencia (Python, R, Spark, SAS Guide/Miner, SPSS, etc.) y recursos en línea, pero no se permite consultar directamente con otras personas a través de ningún medio.

Tiene libertad para realizar supuestos que considere necesarios, por favor enúncielos explícitamente en la solución. No es obligatorio utilizar todos los datos o variables proporcionados ya que esto depende de la estrategia que se adopte para abordar el problema. Tenga en cuenta que no existe una única solución.

El dataset

El dataset integra la información de las ventas y atributos de producto para las fechas comprendidas entre 01/01/2020 y el 30/11/2022.

Ficheros

- El dataset "catalogo_productos.csv" contiene las características de todos los productos, como su tamaño, categoría, proveedor, entre otros. La clave principal en este dataset es el "id_producto"

- El dataset "demanda.csv" se utilizará para construir un modelo predictivo que estime las ventas diarias por producto. Los datos abarcan desde el 01/01/2020 hasta el 30/11/2022. Este dataset proporciona información sobre la demanda diaria por fecha y por "id_producto"
- El dataset "demanda_test.csv" se empleará para aplicar el modelo construido y evaluar su precisión fuera de muestra. Los datos cubren el período del 01/12/2022 al 30/2/2023. Este dataset contiene dos columnas: fecha y "id_producto"

Variables

- date: momento del tiempo en el que se produce el evento.
- Id_producto: número identificador del artículo.
- Categoría: Corresponde a la categoría del artículo
- Sub_categoria: Toma valores de sub categorías del campo categoría
- Tamaño: toma los siguientes valores (pequeño, mediano, grande)
- Premium: Variable dummy. 0 no es premium y 1 es premium
- Marca_exclusiva: Variable dummy. 0 no es exclusiva y 1 es exclusiva
- Estacional: variable dummy que identifica si el producto tiene estacionalidad. 0 no es estacional y 1 es estacional
- Nit_proveedor: código de identificación del proveedor
- demanda: variable a predecir. Nos indica las unidades vendidas para cada día y cada artículo.

Formato y escritura

Los datasets con formato csv tienen como estructura:

- Nombres de campo: Incluidos en la cabecera.
- Separador: ","
- Codificación: UTF-8.

Notas

El 2 de julio de 2021 entró en operación una tienda de la competencia a pocos metros.

Preguntas a responder

- ¿Cuál es la estacionalidad de las ventas?, ¿la estacionalidad cambia dependiendo de la categoría del producto?. Por favor estime la estacionalidad semanal, mensual y anual.
- ¿Cúal fué el impacto de la apertura del competidor? Por favor cuantifique el impacto
- ¿Las ventas tienen una tendencia creciente o decreciente? Excluya el impacto de la apertura del competidor del análisis.

sumz

Estos son los entregables que se esperan de la resolución de la prueba:

- Documento explicativo: Se espera un documento detallado que explique el proceso seguido para resolver la prueba, incluyendo todas las etapas que se hayan seguido para la solución de la misma. Si alguna fase del desarrollo analítico no se incluye en este documento, se asumirá que no se hizo.
- Código documentado: Se espera que el candidato entregue el código documentado que respalden el ejercicio analítico. Se puede usar cualquier formato para la entrega, pero se debe explicar cómo reproducir el análisis en el documento explicativo. Dado que el ejercicio es corto, se invita a entregar un único notebook incluyendo el código y el análisis (documento explicativo).
- Archivo CSV: Se debe entregar un archivo CSV con nombre "resultado_prueba.csv" que contenga las predicciones del modelo sobre el dataset “demanda_test.csv”. Debe contener las columnas: date, id_producto y demanda

La calidad del modelo generado y la técnica utilizada para su creación serán evaluadas y valoradas. Se compararán los valores predichos por el modelo con la demanda esperada de acuerdo a la estrategia de generación de los datos. La métrica utilizada será el RMSE.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Siendo:

- “n” el número de casos
- “yi” el valor real
- “ŷ” el valor estimado

Metodología de calificación

Nuestra metodología de calificación evalúa de manera integral las habilidades de los participantes en la prueba a través de cuatro etapas clave: Análisis Exploratorio de Datos, Modelamiento y Presentación de Resultados. Buscamos identificar candidatos con sólidos conocimientos de analitica y capacidad para resolver problemas de negocio para enfrentar los desafíos en el entorno empresarial actual. Valoramos tanto la competencia técnica como las habilidades de comunicación y presentación.

Detalles de los pesos:

- Análisis Exploratorio de Datos (EDA) [Peso: 30%]
    - Análisis exploratorio [Peso: 5%]
    - Análisis de la demanda [Peso: 5%]

    - Respuesta a las 3 preguntas mencionadas arriba [Peso: 20%]
- Modelamiento (usar al menos 2 modelos) [Peso: 20%]
    - Creación y selección de atributos [Peso:4%]
    - Imputación de datos[Peso: 2%]
    - Partición train/test [Peso: 4%]
    - Entrenamiento y ajuste de modelos [Peso: 6%]
        - Selección de hiper parámetros mediante validación cruzada
        - Usar al menos dos modelos
    - Evaluación de modelos y selección del mejor modelo [Peso: 4%]
        - Comparación métricas modelos
        - Visualizaciones de la demanda real y la predicha
        - Conclusiones
- Presentación de Resultados [Peso: 10%]
    - Documentación clara y organizada [Peso: 5%]
    - Comunicación efectiva de los resultados y hallazgos [Peso:5%]
- Error en dataset fuera de muestra (demanda_test.csv) [Peso: 40%]

¡Mucha suerte!