TODOS SOMOS BODESA desde 1887

Prueba de conocimientos

Puesto: Científico de Datos

LA MARINA · El Bodegón · PROGRESSA · Celerity · MAX 4LESS Outlet · NAVIERI

Transformación Tecnológica

Bodesa S. A. P. I. de C. V.

Instrucciones

El propósito de esta prueba de conocimientos es determinar el nivel de dominio en la aplicación de algoritmos de Aprendizaje Automático para proyectos de Ciencia de Datos, así como el dominio en análisis de datos, detección de patrones y procesamiento de información. Esta prueba deberá entregarse el día **22 de julio, a las 9 a. m.**, contestando el mismo correo electrónico en el que se envió esta información, adjuntando el **entregable** de la prueba (más adelante en el documento se describirá qué será este entregable).

1. La fuente de datos que se utilizará para esta prueba se encuentra en el documento CSV adjunto a este correo, llamado «Tabla de pagos». Esta tabla contiene información de 30 mil clientes a quienes se les otorgó un crédito financiero, así como un histórico de 6 meses (desde abril hasta septiembre) de diversas variables relacionadas con su comportamiento de pago. El objetivo principal de la prueba es **predecir el incumplimiento de pago de los clientes.** La tabla contiene 25 columnas, así como un multi-índice de 2 niveles: el primero es una identificación del tipo de variable que se almacena en la columna, y hay dos tipos:
    a. Característica: están identificadas como variables X.
    b. Objetivo: está identificada como variable Y, y es la variable que se pretende predecir.
2. El segundo nivel contiene los nombres de las columnas de la tabla, y representan lo siguiente:
    a. (X0, ID):es el número de identificación del cliente, y cada uno representa a un cliente distinto.
    b. (X1, MONTO_CREDITO): es el monto del crédito que le fue otorgado al cliente. Refleja la deuda total inicial que el cliente contrató.
    c. (X2, SEXO): representa si el cliente es hombre o mujer, siendo 1 = HOMBRE y 2 = MUJER.
    d. (X3, ESCOLARIDAD): representa el grado de estudios del cliente, donde 1 = POSGRADO, 2= LICENCIATURA, 3 = BACHILLERATO, 4 = OTRO.
    e. (X4, ESTADO_CIVIL): 1 = CASADO, 2 = SOLTERO, 3 = OTRO.
    f. (X5, EDAD): años cumplidos del cliente.
    g. (X6:X11, ESTADO_PAGO): estas 6 columnas representan la cantidad de meses antes o después de su fecha límite de pago, hasta la realización del pago. Un valor menor a 0 representa la cantidad de meses que el cliente adelantó su pago, por ejemplo, -2 significa que el cliente realizó su pago 2 meses antes de la fecha límite de pago; el 0 significa que el cliente pagó en tiempo, menos de un mes antes de su fecha de pago, o justo en la fecha límite; finalmente,

los valores mayores que 0 son los meses de atraso del pago. La primera variable en este grupo es del mes de septiembre, y va hacia atrás hasta el mes de abril, al igual que las siguientes columnas.

   h. (X12:X17, ESTADO_CUENTA): refleja el monto que el cliente vio reflejado en su estado de cuenta en cada mes.
   i. (Y, INCUMPLIMIENTO_PAGO): es una variable binaria que funge como etiqueta de incumplimiento del siguiente mes, es decir, de octubre. Un pago incumplido es aquel que no se realiza en la fecha límite de pago. 1 = PAGO INCUMPLIDO, 0 = PAGO PUNTUAL. Esta es la variable que se debe predecir.

3. Todo el ejercicio deberá realizarse en un cuaderno de Jupyter, que será el **entregable** del ejercicio. El nombre del archivo debe ser la palabra **Prueba**, seguido de tu nombre y tu apellido (sólo uno de cada uno), y debe ser guardado en formato `.ipynb`. Ejemplo: **Prueba David Contreras.ipynb**.
   a. Es importante mencionar que este cuaderno de Jupyter es el único entregable que se debe enviar como respuesta de la prueba.
   b. Puedes utilizar el IDE o editor de código de tu preferencia, pero recuerda que la extensión del archivo debe ser .ipynb.
4. Define qué metodología es la que seguirás para realizar el ejercicio, y justifica de manera sucinta por qué la elegiste por encima de otras. Esta metodología debe marcar una serie de pasos a seguir para llevar a cabo un proyecto de Ciencia de Datos, aplicando el análisis de la información y las predicciones correspondientes. Marca cada etapa del proceso en el cuaderno con celdas Markdown, utilizando el símbolo de numeral al inicio del texto. Ejemplo: # Preprocesamiento de los datos
   a. Solamente deben seguirse aquellos pasos que estén relacionados directamente con el análisis de los datos. Cosas como la investigación del negocio, recabación de datos, la definición de objetivos del proyecto u otras partes que no estén directamente relacionadas con el análisis de datos y la aplicación de algoritmos de aprendizaje automático deben ser omitidas.
5. Independientemente de la metodología elegida, es importante incluir un análisis estadístico de las variables de la tabla, como el cálculo de medidas de tendencia central y desviación (media, desviación estándar), la identificación de datos atípicos, y cualquier otro que pudiera resultar útil para detectar patrones importantes en los datos.
6. Menciona, por lo menos, dos tipos de algoritmos de aprendizaje automático que te pueden ayudar a realizar la predicción de incumplimiento de pago, y describe por qué esos algoritmos resultan adecuados para esta tarea. Entrena modelos con estos algoritmos, y compara sus resultados. Define cuál de ellos elegirías como el mejor modelo, y explica las razones por las que lo elegiste.

    a. Genera las métricas de evaluación de los resultados del modelo que se ajusten mejor al tipo de modelo entrenado, esto para cada uno de los modelos que utilices.
7. Estudia los resultados obtenidos, y menciona cuáles consideras que son los factores determinantes en el incumplimiento de pago de los clientes.
8. Todas las respuestas del ejercicio deben anotarse en el cuaderno de Jupyter en celdas Markdown, cada una en su sección correspondiente.

Recomendaciones

- Es mucho más importante contestar todos los cuestionamientos del ejercicio, que realizar un análisis muy profundo de cada punto. Es recomendable que primero sigas las instrucciones paso a paso, de la manera más sencilla y breve posible, y al final regresar a cada punto para tratar de desarrollarlo con más cuidado, si el tiempo te lo permite.
- Procura acompañar tu análisis con visualizaciones adecuadas de los datos. Muestra gráficas y tablas que resulten relevantes para la comprensión de los patrones identificados y los resultados obtenidos.
- Cuida mucho el tiempo de entrega. Procura enviarlo, por muy tarde, una hora antes de la fecha límite, puesto que no se admitirá una respuesta que llegue a partir de las 9:01 del lunes, y dado ese caso, se considerará la prueba como **no entregada**. Si te surge algún problema con la entrega, llama a RR. HH. y comenta la situación **con tiempo**, para que se pueda trabajar en una solución.
- Si al leer estas instrucciones te surge alguna duda, ponte en contacto inmediatamente con RR. HH., para que podamos darte la asesoría necesaria.

Mucho éxito, recuerda que el objetivo principal de la Ciencia de Datos es **revelar lo que está oculto** en los datos.