UNIVERSIDAD CENFOTEC

DATA ANALYTICS / BIG DATA

CURSO 5 ENTREGA #1

***Get Started with Data Science and Python***

ALFONSO JIMENEZ MORA

304050577

19 FEBRERO 2020

**1** Declaración escrita de la (s) meta (s):

Nuestra meta para este proyecto es ayudar a Credit One a desarrollar una solución que le permita predecir cuando un cliente va a caer en retrasos al realizar los pagos o caer en mora, sobre el crédito que se le otorgo.

**2** Un marco de proceso de ciencia de datos bien definido y las razones por las que lo propone.

Como marco de proceso de ciencia de datos hemos decidido utilizar BADIR, este marco a partir de 5 pasos nos permite tomar decisiones efectivas utilizando datos, a continuación, detallo cada una de ellas:

1. Preguntas del negocio.

Atravez de preguntas tradicionales, “de qué, quién, dónde, cuándo, por qué y cómo” , nos ayuda a identificar el problema en el contexto. Comprender el contexto, el segmento afectado y las posibles razones según lo entendido por el negocio pueden proporcionar un camino más rápido para resolver el problema

1. Un plan de analisis.

Los problemas y las preguntas de negocio pueden ser de gran alcance, por lo que es importante reducir los objetivos o dividirlos en partes, donde cada parte va a tener su propio análisis de objetivos.

1. Definir las fuentes de datos.

Es importante recopilar datos de fuentes relevantes, que a la vez nos van a servir para validar nuestra hipótesis y garantizar la metodología que escogimos.

1. Generar ideas.

En esta etapa vamos a combinar la metodología con las fuentes de datos, por lo cual debemos:

Revisar los patrones: nos va a ayudar a determinar si existe un problema real con los datos o si hay patrones inusuales en los mismos.

Aquí podremos probar o desechar nuestras hipótesis, dejando las mas valiosas para nuestro análisis.

1. Recomendaciones.

Involucramos a los interesados del proyecto presentándoles un conjunto breve, conciso y eficaz de recomendaciones.

Ser percibido como un socio comercial efectivo al presentar recomendaciones creíbles.

Nos dirigimos a los interesados presentando acciones que generen impacto al resolver el problema comercial

Ademas Badir es un marco de proceso que puede usarse para tratar cualquier tipo de problemas analíticos.

**3** Descripciones y ubicación de las fuentes de datos relacionadas.

Como fuente de datos tenemos un archivo csv con 30000 observaciones, las cuales estan distribuidas en 25 variables.

La informacion se distribuye de la siguiente manera:

Cantidad de Hombres: 11888, Representa un 39,6% de la población.

Cantidad de Mujeres: 18112, Representa un 60,3 % de la población.

De los cuales la cantidad por estado de relación es:

|  |  |  |  |
| --- | --- | --- | --- |
| Casados(1) | Solteros(2) | Divorciados(3) | Otros(4) |
| 13659 | 15964 | 323 | 54 |
| 45,5 % | 53,21% | 0,01% | 0% |

Su nivel de estudio es:

|  |  |  |  |
| --- | --- | --- | --- |
| Graduad Escuela(1) | Universidad(2) | Segundaria(3) | Otros(0,4,5,6) |
| 10585 | 14030 | 4917 | 468 |

|  |  |  |  |
| --- | --- | --- | --- |
| Edad | Mínima | Máxima | Promedio |
|  | 21 | 79 | 35.5 |

Las fuente de datos, dispone de las siguientes variables:

X1: Importe de la dada de crédito (en dólares NT): incluye tanto el crédito al consumo individual y su / su familia de crédito (complementario).

X2: Sexo (1 = masculino; 2 = mujer).

X3: Educación (1 = escuela graduado; 2 = universidad; 3 = la escuela secundaria; 0, 4, 5, 6 = otros).

X4: Estado civil (1 = casado; 2 = único; 3 = divorcio; 0 = otros).

X5: Edad (años).

X6 - X11: Historia de pago pasado. Tenemos un seguimiento de los últimos registros de pagos mensuales (de abril a septiembre de 2005) de la siguiente manera: X6 = el estado de devolución en septiembre de 2005; X7 = el estado de devolución en agosto de 2005; . . .;

X11 = el estado de devolución en abril de 2005. La escala de medición para el estado de amortización es: - 2: No Consumo; -1: A cargo en su totalidad; 0: El uso de crédito renovable; 1 = retraso en el pago durante un mes; 2 = retraso en el pago por dos meses; . . .; 8 = retraso en el pago durante ocho meses; 9 = retraso en el pago durante nueve meses o más.

X12-X17: Cantidad de estado de cuenta (en dólares NT). X12 = cantidad de estado de cuenta, en septiembre de 2005; X13 = cantidad de estado de cuenta en agosto de 2005; . . .; X17 = cantidad de estado de cuenta en abril de 2005.

X18-X23: Cantidad de previo pago (en dólares NT). X18 = cantidad pagada en septiembre de 2005; X19 = cantidad pagada en agosto de 2005; . . .; X23 = importe pagado en abril de 2005. Y: el comportamiento del cliente; Y = 0, entonces no por defecto, Y = 1 entonces por defecto".

**4** Una explicación de cómo administrará los datos del proyecto.

Para administrar los datos de proyecto vamos a proceder de la siguiente manera:

1. Verificamos si existen datos incompletos, si existen los eliminamos de la fuente de datos.
2. Verificamos cuales columnas debemos renombrar para realizar un análisis mas eficiente de la información.
3. Verificamos que tipos de datos podemos convertir a otros tipos para manipular mas eficientemente los resultados.
4. Debemos verificar si con los datos que poseemos, podemos dar respuesta a las hipótesis que hemos planteado, por medio de nuestro marco de procesos BADIR.
5. Debemos evaluar si es necesario segmentar la información, para poder realizar un análisis mas eficiente y poder brindar mejores respuestas.

**5** Cualquier problema conocido con los datos y cómo planea abordarlos.

Durante el análisis inicial del proyecto no hemos detectado o conocido problemas con las fuentes de datos que tenemos a disposición.

**6** Un diagrama de flujo que visualiza el proceso detallado que seguirá, anotado con los posibles problemas que haya identificado y sus soluciones propuestas para dichos problemas.

Análisis del problema con el interesado

Analizar los datos

Los datos están correctos

Revisar y optimizar

Seleccionar modelos a aplicar

Entrenar los modelos

Escoger el mejor modelo

Probar el modelo con la datos reales

Analizar e interpretar los resultados

Presentar los resultados al cliente.

Marco de proceso Badir.

1. Preguntas del negocio.

**¿Cuál es la pregunta comercial planteada?**

Se pueda ayudar a disminuir el riesgo de crédito de los clientes de Credit One.

¿**Cuál es la intención subyacente a la pregunta (por ejemplo, cuál es el contexto, cuál es el segmento afectado y cuáles son los pensamientos actuales de los interesados ​​sobre las razones subyacentes?**

CreditOne, en el ultimo año de operaciones se ha experimentado un aumento en el numero de clientes que han caído en incumplimiento de pago de crédito otorgado y a la

vez teme que si la situación no se controla se puede poner en riesgo varios negocios.

**¿Qué consideraciones comerciales (p. Ej., Partes interesadas, cronograma y costo) pueden afectar el análisis?**

**Matriz de Interesados**

|  |  |
| --- | --- |
| Interesados | Nivel |
| Dueños CreditOne | Alto |
| Clientes | Medio |
| Empleados de servicio al cliente | Alto |
| Equipo de analistas | Medio |

1. **Un plan de analisis.**

* ¿Cuál es el objetivo del análisis?

El objetivo del analisis para CreditOne es poder determinar que clientes van a caer en estado de morosidad.

* ¿Qué hipótesis hay que probar?

CreditOne logra determinar que cliente va caer en estado de morosidad.

* ¿Qué datos se requieren / están disponibles para probar las hipótesis?

Los datos que requerimos son :

* 1. Monto del credito otorgado.
  2. Ultimo pago realizado.
  3. Historial de pagos de cliente.
  4. Monto de la cuota mensual.
  5. Caracteristicas del cliente.
  6. Estatus actual del cliente.
* ¿Qué metodología (s) empleará?

Emplearemos la siguiente metodología:

* 1. Análisis del cliente.
  2. Obtención de datos del cliente.
  3. Limpieza y normalización de datos.
  4. Escogencia de modelos de análisis.
  5. Entrenamiento de los modelos.
  6. Evaluación de las métricas.
  7. Escoger el mejor modelo.
  8. Predecir la información con el modelo.
  9. Presentar los resultados al cliente.

1. **Definir las fuentes de datos.**

Tenemos a disposicion un archivo csv, con la informacion recopilada desde el mes

De Abril a Setiembre del 2005.

1. **Generar ideas.**

* ¿Qué patrones ves en los datos?
* ¿Cada una de las hipótesis está probada o refutada?
* ¿Cuánta confianza deben depositar los interesados ​​en los resultados?
* ¿Cómo clasifica sus hallazgos en términos de impacto cuantificado en el negocio?

1. **Recomendaciones.**

**¿Cómo puede presentar de manera más efectiva los resultados de su análisis a sus partes interesadas (en términos que puedan entender y en alineación con la información que valorarán)?**