**PRUEBA DE CONOCIMIENTOS TÉCNICOS**

Bienvenidos a la prueba de conocimientos para la posición de análisis de datos en el área de Inteligencia de Negocios de la Vicepresidencia de Fortalecimiento Empresarial.

La evaluación se compone de dos partes: en una primera parte se evaluará su capacidad propositiva en cuánto a cuáles son las preguntas que se debe hacer (y como responderlas) para lograr un objetivo general de la empresa. En la segunda se evaluará su capacidad para lograr los resultados técnicos requeridos en el cargo, en esta parte deberá entregar un código que logre una meta clara propuesta por el evaluador.

Esta es una simulación de cómo serán sus interacciones con el equipo de trabajo. Se atenderán dudas colectivas relacionadas con el enunciado de la prueba durante los primeros 10 minutos. Posteriormente, se atenderán dudas de manera personal relacionadas con el desarrollo de la prueba. La parte más importante que usted le va a aportar al equipo de trabajo es la capacidad de procesar y analizar las bases de datos, por lo tanto las dudas que se plantee deberán servir para aclarar el proceso más no sobre el detalle de las respuestas.

Durante toda la duración del examen podrá hacer uso de internet, ingresar a foros y buscar información que lo ayude a responder el examen.

Parte 1

Explicación del modelo de negocio sobre el que usted estaría trabajando:

La Vicepresidencia de Fortalecimiento Empresarial (VFE) de la CCB tiene como objetivo apoyar a las empresas y emprendedores en todas las áreas de una organización. Por ejemplo en temas legales y tributarios, estrategia empresarial, finanzas, etc. Para lograr este objetivo la vicepresidencia tiene el siguiente modelo para atender a los clientes:

1. El proceso comienza con la vinculación de los clientes: consta de la realización de un autodiagnóstico por parte de los emprendedores/empresarios. Este autodiagnóstico (se abrevia con Dx) se transforma automáticamente en un conjunto de servicios (llamado Ruta) que le aportan el mayor valor al cliente. La Ruta está determinada por las respuestas del cliente en el cuestionario. Por ejemplo: si en la pregunta "¿tiene sus estados financieros?" el cliente responde que no los tiene pero le interesaría implementarlos, entonces la herramienta incorpora uno o varios cursos que lo ayuden a lograr este objetivo.
   1. Además de esto, cuando el usuario hace el Dx debe escoger un código CIIU que a su vez lo asigna a alguna de las 4 áreas de la vicepresidencia (Industrias Creativas y Culturales, Confecciones, Agroindustrial y Multisectorial)
   2. Luego de que el cliente sabe su ruta, debe inscribir estos cursos a través de internet o acercarse el día de la fecha del servicio y esperar que haya cupos disponibles.
   3. Los servicios que ofrece la cámara se publican en internet a principios de cada mes. Estas publicaciones contienen la agenda de servicios para el mes correspondiente. En general se ofrecen cupos de 30 personas para atender en distintos horarios: mañana, tarde, y extra laboral (tanto entre semana como fines de semana).
   4. Estos servicios son prestados por 7 líneas de acción distintas: ICC, Confecciones, Agroindustrial, Creación, Fortalecimiento, Innovación e Internacionalización.
2. Cuando el cliente asiste a más del 80% de los servicios de la ruta se considera que "ha finalizado su ruta".
3. Una de las metas de la VFE es lograr que la mayor cantidad de clientes finalicen su ruta. En lo preferible se requiere que esto se haga de manera eficiente, es decir, logrando el mayor cubrimiento con la menor cantidad de servicios.
   1. Algunas de las acciones que toma la VFE para lograr este objetivo son:
      1. La divulgación de los servicios a través de telemercadeo
      2. La publicación y prestación de los distintos servicios en distintas fechas y a distintos horarios. Esta prestación de servicios la hacen cada una de las áreas por aparte (ICC, Confecciónes, Agroindustrial y Multisectorial)

Usted como analista de datos se le entregan las siguientes bases de datos con sus respectivos diccionarios:

* Rutas y Servicios: contiene la información de la ruta recomendada para cada diagnóstico y si el cliente ya ha asistido a cada servicio o no.
* Diagnósticos: tiene información relevante del cliente, su puntaje en el diagnóstico, el sector al que pertenece, la ciudad donde fue creado el diagnóstico y otros.
* Asistencias. En estas bases de datos están las asistencias a todos los servicios que se han dado durante el 2015, en ésta no solamente están las asistencias por parte de los clientes vinculados (que hicieron Dx) sino también de los clientes que vienen por cuenta propia y por lo tanto, no tienen un diagnóstico relacionado (estos clientes se conocen como autogestionados).
* Preinscripciones: contiene preinscripciones

(30%) Basándose en esos datos, la meta y las dos posibles acciones que puede tomar la VFE, ¿qué preguntas se haría usted que le permitieran tomar decisiones que maximicen el objetivo de la VFE?, ¿cuáles hipótesis creería usted que valdría la pena responder y por qué?, ¿cómo respondería a estas hipótesis con los datos que tiene disponibles?

* Responda estas preguntas de una forma clara y concisa. Se recomienda una estructura basada en "bullet points" e indentaciones para hacer más fácil de entender para el lector.
* Incluya detalles de cómo procesaría los datos que tiene para obtener los resultados, y los indicadores que lo ayudarían a tomar las deicisiones.

RESPUESTA ALDO

¿cómo maximizo el objetivo de la VFE a partir de: la prestación de los distintos servicios en distintos horarios y días, y el telemercadeo de los demás?

El objetivo de la VFE es maximizar el número de clientes que terminan su ruta. Por lo tanto para poder lograrlo es necesario lograr que la mayor cantidad de clientes asista a los servicios en su ruta. Una forma de atacarlo sería a través de los siguientes pasos:

1. Identificar cuáles son los servicios necesarios para completar la ruta de los clientes.
   1. Aún mejor, ordenar los servicios de mayor a menor por el número de veces que aparecen en las rutas faltantes de los clientes. Los que queden en los primeros puestos serán más importantes que los que queden en los últimos.
2. Una vez identificados analizar cuáles servicios que son necesarios para que se completen las rutas tienen gran demanda y cuáles no. Los servicios muy demandados no necesitarán de mucha publicidad a través de tele-mercadeo, mientras que los servicios con poca demanda implicarán un mayor esfuerzo en esta área.
   1. La medición de la demanda se hace a través de la base de preinscripciones. La demanda es proporcional al número de preinscripciones que tiene el servicio.
3. La prestación de los servicios deberá hacerse en los horarios que más satisfagan a la comunidad. Para saber esto se deberá generar un indicador que nos muestre la afinidad del horario con la comunidad. Un posible indicador es el promedio de asistencias en el horario/día durante el 2015.
   1. Para calcular este indicador sería necesario sumar el número de asistencias en los distintos horarios y luego dividir por el número de veces que se ofreció un servicio en ese horario.
4. Otra posible mejora sería publicar la agenda de los servicios con un mes de anticipación. Si el primero del mes se publican los servicios que se van a dar el día siguiente, las personas no alcanzan a enterarse o a planear su tiempo, mientras que si el primero del mes se publican los servicios que se van a dar en el mes siguiente las personas tendrían más tiempo para organizarse.
   1. Si se quiere analizar si esto es un efecto real sobre la asistencia habría que mirar si el promedio de las asistencias justo después de las fechas de publicación (a principios de mes) es menor al promedio de asistencia a mediados o finales del mes.

Parte 2:

En esta parte se evalúan las habilidades técnicas para lograr resultados concretos y se requiere lo siguiente:

Consideraciones:

* Cree distintos scripts (do-files) para organizar sus respuestas y en el nombre del archivo especifique los literales que responde en cada uno.
* Intente que la ejecución de los scripts desde el computador del evaluador sea sencilla: utilice directorios por defecto, basados en la ubicación del script o cree un proyecto desde el programa Rstudio o Stata; mantenga una estructura organizada de las carpetas de donde lee los datos, donde están los scripts y donde pone las visualizaciones y los reportes.
* En este documento escriba muy bien los pasos que va a seguir para responder a cada una de las preguntas, de tal forma que al leer el código se pueda entender cuál fue la metodología.
* Comente muy bien sus scripts, ya que en estos estará basada su calificación.
* Los resultados debe entregarlos en Excel y debe explicar cómo están organizados para facilitar la lectura por parte del evaluador (puede ser en un README.txt).
* Se tendrá en cuenta la facilidad en la lectura de las tablas de excel, el orden de las variables y el formato.
* Recuerde que siempre se atienden preguntas de forma individual
* También puede consultar toda la información que necesite en internet
* En esta parte deberá utilizar solo 3 de las bases de datos para responder las preguntas
* Tenga en cuenta que a veces las bases tienen espacios, ceros a la izquierda, tildes, letras o mayúsculas que podrían afectar el cruce de las bases y la agregación/agrupación de los datos.
* No espere encontrar todas las entradas de una base en otra cuando esté cruzandolas, pero asegúrese de estar capturando la mayor cantidad de relaciones.
* Durante todo el examen podrá utilizar cualquier software entre Stata o R pero sus reportes de resultado deberán estar en PDF o Excel.
* Las áreas de los Dx dependen directamente de la variable ***sector.DX.*** de la siguiente manera:
  + Agroindustrial: "AGROINDUSTRIAL" y "AGRÍCOLA".
  + Confecciones: "TEXTIL Y CONFECCIÓN"
  + ICC: "INDUSTRIAS CREATIVAS Y CULTURALES"
  + Multisectorial: el resto

Cuestionario

1. *(30%)*. Para este análisis deberá utilizar la base de diagnósticos y la base de rutas. Utilice solo el Dx más reciente por cada cliente. Use solamente diagnósticos en estado diligenciado. Deje únicamente Dx de creación y fortalecimiento. Los diagnósticos deben estar en la jurisdicción de la CCB para que sea válido (hay un archivo con la jurisdicción). Deberá agrupar los sectores de los diagnósticos para crear una variable “***area***” de acuerdo a como se explicó en las *consideraciones*
   1. Un histograma con la distribución de los porcentajes de avance en ruta de todos los clientes: el número de servicios a los que el cliente ha asistido sobre el total de servicios que el mismo tiene en la ruta. Haga un histograma con el total de la muestra
   2. Una tabla con los porcentajes de avance en ruta promedio (entre clientes) por el área a la que pertenece el cliente (ICC, confecciones, agroindustrial y multisectorial) y el tipo de cliente: creador o empresario. Por cada área separe los tipos de cliente.

1. *(40%)*. Para el siguiente ejercicio deberá utilizar las bases de asistencias de 2015. Debe unir las bases de cada período para tener una sola con todas las asistencias del 2015. Haga estos análisis únicamente para las sedes de Chapinero, Kennedy y Salitre, y para las actividades de tipo Taller, Cápsula o Asesoría grupal. Solo tenga en cuenta actividades abiertas y de tipo presencial. En lo posible, válgase de gráficos (y sus interpretaciones) para mostrar sus análisis y conclusiones.
   1. Haciendo análisis de datos responda cuál sería la mejor jornada para ofrecer los servicios. Divida su muestra teniendo en cuenta el horario: antes de las 12m, 12m-4:59pm o después de las 5pm, y el día de la semana (lunes, martes…).
   2. Teniendo en cuenta que las personas que se inscriben por internet no siempre asisten, la cámara quiere aumentar el número de cupos disponibles por web para el público en general en cada uno de sus servicios (actualmente en 30). Sin embargo esta decisión debe tomarse con cautela porque es posible que se desborde la capacidad de los salones (de 30 personas) debido a un aumento irresponsable de estos cupos. Haga un análisis predictivo del número de asistentes totales a un servicio a partir del número de inscritos web y luego calcule un número de cupos web ideal asumiendo una probabilidad de desborde del 15%. (Esto es una aplicación de los modelos de regresión lineal)

RESPUESTA ALDO

Para calificar estos ejercicios debe tenerse en cuenta el procedimiento del candidato. Este cobrará suma importancia ya que la cantidad de factores de error es muy alta y llegar a un resultado concreto será muy improbable. Por este motivo, la mejor forma de evaluar al candidato es mirando el código con el que intentó llegar al resultado y tener en cuenta que haya tenido en cuenta la mayor cantidad de pasos posibles para lograr el resultado.

A continuación, para cada uno de los literales, voy a enunciar los pasos que se deben tener en cuenta en el código para poder llegar a un resultado correcto.

I.

Filtros Base de Dx:

* Deberá agrupar los diagnósticos que tengan el mismo número de cédula y dejar únicamente el que tenga la fecha más reciente.
* Deberá eliminar los diagnósticos con estados distintos a “Diligenciado”
* Deberá filtrar los dx de formalización y nueva línea de negocio
* Deberá leer la base de jurisdicción, y transformarla para que quede directamente en mayúsculas, sin tildes y sin ningún símbolo especial como espacios (a la izquierda o derecha), puntos o comas. Si no hace esto no podrá hacer el cruce para dejar solo las ciudades dentro de la jurisdicción.

Cruce con la base de Rutas:

* Una vez filtrada la base de Dx será necesario hacer un merge, join o en palabras coloquiales un “cruce” con la base de rutas. En este cruce deberán quedar solo las entradas que tengan entradas en ambas bases de datos. Este cruce deberá hacerse a través del identificador de diagnóstico [ID.DX].
* Solamente es necesaria el estado de la actividad para calcular el porcentaje de avance en ruta

Cálculo de porcentaje de avance en ruta:

* Deberá agrupar por cada valor que tome el identificador del diagnóstico y para cada grupo calcular: el número de servicios con estado asistido y el número de servicios totales.
* La división entre estos dos indicadores entregará el porcentaje de avance de ruta por cada diagnóstico.

a. Para generar el histograma en el eje X deben haber rangos del porcentaje de avance en la ruta y en el eje Y debe estar la frecuencia (o el número de Dx) en el que recaen en cada uno de los rangos de porcentaje que se definan.

b. Deberá agrupar por las distintas combinaciones entre área y tipo de cliente (ICC-Creador, ICC-Fortalecimiento, Agro-Creador, etc.). Luego deberá calcular el promedio de avance en ruta como la suma de los porcentajes de avance en ruta de todos los Dx en cada grupo sobre el número de Dx en este subgrupo. También existen funciones para calcular el porcentaje directamente como average().

II.

Debido a que este punto es más de análisis que de cálculo de cifras, los resultados y la metodología utilizada podrá variar bastante. La metodología acá propuesta consiste en una de las posibilidades para responder a estas preguntas.

Pegado Base Asistencias:

* Deberá leer cada una de las bases de datos teniendo en cuenta que las tablas comienzan desde la fila 10 para cada una de las bases
* Una vez la base quede completa para todo el 2015, es necesario filtrarla para dejar solo las sedes y los tipos de actividad especificados en el enunciado.
* Es necesario arreglar los códigos para quitarle los espacios
* Se eliminan las entradas que no tengan código o que el código esté en blanco

a.

Quizás la parte más complicada de este ejercicio sea la transformación de los horarios de inicio de clase en la jornada a la que pertenece.

Horario - Jornada:

1. Debe identificar la hora de inicio (la primera parte de la variable HORA)
2. Debe identificar si esta hora es am o es pm
3. Luego debe asignar la hora de inicio a una jornada: 8-12 (mañana), 12-17(tarde) y después de 17 (extralaboral)

Día:

* Debe homogeneizar los días debido a que están escritos de distintas formas en la base: por ejemplo Sábado y sabado.

Análisis:

1. Deberá agrupar por jornada y por día de la semana y calcular el promedio de asistencias como: el número de asistencias dividido el número de servicios que se prestaron ese día.
   1. Si no ignora las actividades con asistencias iguales a 0 podría no obtener resultados correctos.
2. Una vez se genera esta tabla es necesario interpretarla. Diciendo que los horarios nocturnos y los sábados son las sesiones con el mejor promedio de asistencias.

Gráfica: Simplemente una gráfica que muestre como hay algunas jornadas y algunos días para los cuales la asistencia es mejor. Es necesario que se pondere por el número de servicios. Puede ser una gráfica de barras con los promedios de asistencia por día-jornada

b.

Este es un problema de predicción. Entonces el algoritmo para usarlo puede ser cualquier algoritmo de regresión estadística. En este caso yo voy a exponer como se haría con una regresión lineal.

Regresión Lineal:

* La variable explicativa deberá ser el número de inscritos por web
* La variable explicada o dependiente deberá ser el número de asistentes a la actividad
* Se debe hacer una regresión lineal sin intercepto, es decir: .
* Una regresión con intercepto podría funcionar pero sería más difícil de interpretar y daría otros resultados:

Una vez se hace la regresión se debe calcular un intervalo de predicción con una probabilidad de error dada. Si uno quiere una probabilidad de acierto de 70% (con error de 15% por abajo y 15% por arriba del intervalo) deberá obtener una ecuación como la siguiente:

Como a la cámara únicamente le afecta un desborde por el límite superior de este intervalo. Entonces la ecuación de los asistentes vs los inscritos web sería:

El número de cupos web que se debe abrir sería tal que el número de asistentes fuera menor a 30. La ecuación acá escrita describe un intervalo para el cual el número de asistentes sería menor al lado derecho de la ecuación con un 85% de probabilidad (porque no nos afecta el desborde inferior). Por lo tanto reemplazando el número de asistentes por 30 y despejando el número de inscritos Web nos entregaría el valor de los cupos que estamos buscando. En este caso sería igual a 54.

Gráfica: hacer una gráfica con puntos donde en el eje X estén los incritos WEB y en el eje Y estén los asistentes. Además agregar una línea de tendencia e inclusive los intervalos para hacerlo más claro.