INTRODUCCION PROYECTO DATA SCIENCE PARA SOC

Este proyecto de datascience surge por la necesidad del área de encontrar una forma innovadora de detectar a los clientes que realmente están siendo afectados por las incidencias en la red. La metodología actual resulta anticuada y no proporciona certeza en cuanto a la información que se entrega a otras áreas, lo que llevó a realizar un análisis de la metodología y determinar si era una buena base de inicio o si se debía cambiar completamente el enfoque.

Durante el desarrollo de este modelo, se aprendió algo importante: la adopción del uso de métodos científicos para obtener "verdades" fundamentadas en datos objetivos puede revelar dos posibles cosas. En primer lugar, que los procedimientos actuales para obtener información valiosa están bien estructurados y la metodología de datascience solo ayuda a escalar la lógica o demostrar que la lógica actual está mal diseñada y que se deben advertir a quienes toman decisiones de que sus decisiones se basan en datos erróneos. En este último caso, el problema para el desarrollador de soluciones basadas en ciencia de datos podría estar en conflicto con los que toman decisiones si estos se resisten a aceptar los resultados.

Es importante destacar que aquellos que nos apasionamos por los temas del conocimiento científico no solo debemos ser buenos en la abstracción de sistemas, entidades y comportamientos matemáticos, sino que a veces debemos ser valientes y honestos. Esto se debe a que, en última instancia, nuestro objetivo es encontrar soluciones efectivas y fiables para los problemas que enfrentamos en nuestra área de trabajo.

SOBRE EL OBJETIVO DE ESTE DESARROLLO

El objetivo inicial que se me planteo como ingeniero fue "encontrar una manera innovadora de detectar los usuarios realmente afectados por una incidencia en un red móvil 3G, 4G, 5G.

SOBRE EL ESTADO DEL METODO DE INVESTIGACION ACTUAL

De las Incidencias

Los nodos declarados en incidencias es por criterio binario, o está disponible o esta indisponible

Estos nodos son registrados manualmente en un portal por el NOC donde se convierte en información oficial para la detección de los clientes en las incidencias

De todos los nodos declarados en el portal se seleccionan todos los usuarios vistos en los últimos 7 días y se consideran usuarios potencialmente afectados

Luego, pasada la incidencia se consulta a esos usuarios potenciales por su tráfico en la ventana de la incidencia, si fue menor a 1MB es un usuario afectado, si es mayor no se considera que esta afectado

SOBRE LA EXPLORACION DE LOS RESULTADOS DE LA METODOLOGIA ACTUAL

Los resultados de la metodología actual demostraron no son fiables, ya que se detectaron errores en el procedimiento de cómo obtener los datos de red para análisis y por tener un enfoque lógico rudimentario que no atiende con rigor el verdadero comportamiento de una red móvil ante las incidencias que estas enfrentan y una estructura de administración de las incidencias poco comprensible que inducen a errores y provoca desconfianza en los resultados

Al explorar los datos se observaron casos en la asignación de tickets era confusa y la valorización de gravedad estaba no hacia justicia a la realidad de los datos

Se asigna como P1 (Máxima Gravedad) una zona con 10% de tráfico, mientras otras con 20% de tráfico perdido estaban siendo clasificadas como P3 (Gravedad laxa).

Sobre el tamaño de la incidencia, el método actual separa el tamaño real de la incidencia en diferentes tickets con diferentes valorizaciones de gravedad en base a las comunas, lo que lleva a una sobrevaloración del número de tickets y una pérdida de perspectiva del tamaño real de la incidencia.

En términos del planteamiento del concepto se observa que el diseñador anterior solo considero que los nodos que están indisponibles pueden estar en incidencias , ignorando el hecho de que , bajo cierto tipo de incidencias los usuarios migran a nodos de su entorno, arrastrando a estos nodos de manera indirecta a la incidencia

Este último punto es importante para atender el objetivo inicial de este desarrollo, ya que el cambio de consideraciones de que nodo esta en incidencia impactara en los resultados de la identificación de los clientes que están en las misma

Sobre las queries utilizadas para la detección de los clientes reamente afectados, en el análisis exploratorio demostró que sus resultados no responden a lo que se busca. Los cliente detectados con menos de 1MB durante la incidencia eran clientes con >1MB de manera persistente, días antes y días después. El rango de detección de estos clientes con problemas estructurales en estas queries estuvo entre el 30% al 70%, los que se informan como clientes afectados para campañas de compensación

Por último y para cerrar este capítulo agregar que en la comparación de los resultados de este desarrollo se detectó que los nodos declarados en el portal son inconsistentes (en Arauco habiendo 22 nodos NOC declaro 54 Nodos en la comuna con problemas) lo que en sí mismo ya deberá generar desconfianza ante sus resultados

Por todo lo anterior se estimó desechar completamente el método anterior y replantear los problemas para obtener información confiable y comprensible por cualquier interesado en consumir estos nuevos resultados

MODELADO DATASCIENCE

Los problemas identificados son los siguientes:

- La necesidad de medir el tamaño de la incidencia y cuantificar su nivel de gravedad adecuadamente: esto significa que se necesita tener una forma precisa de medir cuándo ocurre una incidencia en la red y qué tan grave es.
- La necesidad de detectar nodos afectados directa e indirectamente por las incidencias: en algunos casos, los nodos que no están directamente involucrados en una incidencia pueden ser afectados indirectamente. Es importante detectarlos para tener una visión completa de la situación.
- La necesidad de la detección y evaluación de nodos en incidencias con base en análisis de series temporales y de manera automática: en lugar de simplemente clasificar los nodos como afectados o no afectados, se debe analizar su comportamiento a lo largo del tiempo para determinar si están en una incidencia o no.
- La necesidad de verificación que los recursos en incidencia volvieron al 100% de sus capacidades pasadas: es importante verificar que los recursos que estaban afectados por una incidencia han vuelto a su capacidad normal después de que se hayan tomado medidas correctivas.

Una vez que se resuelvan estos problemas, se podrá detectar a los clientes reales afectados por una incidencia en la red y clasificar su afectación en las categorías de "No Navega", "Navega Lento", "Navega igual" y "Navega Mejor".

Para lograr esto, es necesario entender completamente el comportamiento de la red y sus entidades basadas en información del sistema de OSS, sin intervención humana en el proceso.

SOBRE LOS PROBLEMAS DETECTADOS

DE LA ADMINSTRACION DE LAS INCIDENCIAS

Fundamentalmente los problemas detectados respecto de la detección de los nodos en incidencias son producidos por falta de una administración automática de las mismas

Al enfrentar la automatización de la administración de las incidencias, lo primero que se requiere son definiciones que sean aceptadas por todos los que toman decisiones

Estas definiciones o enunciados guiaran el desarrollo y es donde se deberán reflejar las modificaciones del diseño

No hay que olvidar que lo más importante del diseño de una solución basada en ciencia de datos, es el modelamiento del problema en sí basado en definiciones y que las matemáticas o las tecnologías usadas para resolver estos problemas son solo las herramientas que permiten resolver estas cosas

Por lo tanto, es importante tener una comprensión clara del problema y sus definiciones para poder diseñar una solución efectiva y precisa

De aquí en adelante este documento aborda los temas que serán atendidos para la automatización de la administración de la incidencia

SOBRE LA RED, SU COMPORTAMINETO Y LA DETECCION DE EVENTOS

DE LOS DATOS DE LA RED

En la compañía tenemos básicamente 2 tipos de informaciones sobre aspectos técnicos de las conexiones, los datos de red, que son generados por las todas las entidades de la red y son almacenadas por el OSS. Este tipo de datos, conocidos como contadores PM son utilizados por todos los operadores y se utilizan para generar KPI que son discutidos por quienes toman decisiones

Estos contadores básicamente miden en el tiempo diferentes dimensiones de las conexiones de usuarios tales como el trafico cursado, el volumen de usuarios atendidos, candidad de caidad de conexiones, tipo de servicio que utiliza el usuarios

Entre las entidades de red que generan datos temporales PM están los Nodos que son las entidades de interés en este desarrollo

La idea es utilizar esta información y consultarla periódicamente para que vaya actualizando los resultados de la esta investigación

Como son "datos de red" no existe información que identifique a los usuarios, individualmente pero existen contadores que muestran el volumen de usuarios únicos atendidos por las entidades, pero no identifica cuales

DE LOS DATOS DE LOS USUARIOS

En el caso nuestro, estos datos son obtenidos desde la solución XXXXXXX que está basada en sondas que extraen datos de los usuarios temporalmente

Para respetar las políticas de privacidad de nuestros clientes esta información solo será utilizada para la detección de ellos en alguna incidencia y los resultados solo apuntaran a saber cómo le afecto la misma.

No se considera detectar patrones de su comportamiento que estén fuera de este alcance

SOBRE LA DETECCION DE NODOS CON AFECTACION DIRECTA E INDIRECTA

DE LOS NODOS EN INCIDENCIA POR CAUSA DIRECTA

Lo primero que atenderé es separar lo relevante de lo irrelevante para simplificar el diseño de investigación

El objetivo de este "modulo", que llamare "**módulo de detección**" es poder separar el universo de los nodos de la red en dos grandes grupos:

- NODOS CON INCIDENCIA
- NODOS SIN INCIDENCIA
- NODOS FUERA DE SERVICIO

Esta clasificación nos pone de frente al tema sobre qué es lo que entendemos por incidencia, de manera formal

ENINCIADO DE INCIDENCIAS

Una incidencia de red es todo evento adverso que desvíe el rendimiento del tráfico respecto de agrupaciones de nodos agrupados, tanto lógico como políticamente

En términos de una red RAN las causas probables que expliquen las desviaciones de tráfico son 3:

- Por problemas de cobertura
- Por problemas en el Transporte
- Por fallo general de cobertura y/o Transporte
- Por cambios de comportamiento de los usuarios

SOBRE LOS PROBLEMAS DE COBERTURA

La cobertura y sus niveles disparan eventos de movilidad que son soportados por las tecnologías móviles. Estos se disparan normalmente por movilidad de usuarios, o sea los niveles de señal de la cobertura cambia en una conexión porque el cliente se mueve. Si el cliente permaneciera estático lo más probable es que no se disparen eventos de movilidad y que los niveles no sufran grandes variaciones

En una incidencia de cobertura este escenario cambia radicalmente. Los clientes que están estáticos de pronto ven que el entorno electromagnético ha cambiado y debe proceder con sus procesos de handover o reseleccion porque "la red se movio" en términos coloquiales

Pero ¿Qué puede generar que se produzcan estos cambios?

Los nodos generan la cobertura de la red a través de sus celdas. O sea un nodo puede tener 1 o más celdas en 1 o varias frecuencias emanado la cobertura

Se entiende entonces que la cobertura está estrechamente ligada a la cantidad de celdas que están disponibles en un área de agrupación de nodos

Los problemas de cobertura entonces puedes ser detectado por la cantidad de nodos disponibles. Si cae una celda veremos que otras celdas intentaran absorber el tráfico de la celda a menos, por los procedimientos de movilidad de la red, ya sea en el propio nodo o en algún otro servidor cercano

Por lo tanto si vemos detectamos que en número de celdas disponibles disminuye de su capacidad habitual podremos clasificar aquel nodo con pérdida de celdas generando problemas de cobertura

Otra forma de que varíe la cobertura es por optimización del sistema radiante. Aquílos ingenieros de optimización alteran, en su afán de optimizar la red, las inclinaciones y azimut de las antenas y niveles de potencia de radiación. En este caso, veremos que no se producen perdida de celdas. De todos modos si la optimización genera un evento demasiado adverso se disparara como una incidencia en este desarrollo y podrá ser diagnosticado que es un problema por configuración observando el conjunto de variables que se han seleccionado para este desarrollo

SOBRE LOS PROBLEMAS DE TRANSPORTE

Así como los problemas de cobertura tienen sus particularidades que la hacen identificables en término de incidencia de red, los problemas de transporte también tienen un comportamiento característico

El transporte de un nodo es la conexión que une al nodo con el core. Una degradación de esta conexión afecta a todos los usuarios del nodo en cuanto a su capacidad de transmitir datos, pero no afecta al nodo en cuanto a su cobertura

Como no hay afectación de la cobertura, los clientes no sufren variaciones en sus niveles de señal y como consecuencia ningún procedimiento de movilidad es disparado y los clientes quedan conectados a la red pero con menor capacidad

La otra característica de la falla de transporte es la simultaneidad de caída de tráfico como consecuencia de su arquitectura de dependencias de nodos en una cascada y el comportamiento de sus curvas de tráfico en conjunto con su comportamiento de usuarios es bien característico, por lo que lo hace fácilmente diagnosticable

SOBRE LOS ESTADOS DE FUERA DE SERVICIO

Si se presenta un fallo general tanto por cobertura o transporte el tráfico se irá a 0. En estos casos esos nodos clasificaran como fuera de servicio

ACERCA DEL CAMBIO DE COMPORTAMINETO DE USUARIOS

Fines de semanas largos, vacaciones, eventos masivos son eventos temporales que rompen la rutina de los usuarios con respecto a la red.

Estas si no son bien atendidas podrían confundirse con un fallo de red cuando en realidad es una situación generada por los usuarios y que exigen a la red en periodos determinados

Aquí el comportamiento fundamental es la disminución o el aumento de usuarios atendidos en una zona. En este caso observaremos que no existen problemas que generen alteraciones en la cobertura y tampoco se detectan problemas de transporte

También será útil observar el comportamiento de las llamadas al 103 para asegurarnos que estamos ante una migración o absorción de usuarios y no ante un fallo de red (Falso negativo)

SOBRE LAS INCIDENCIAS INDIRECTAS

Los nodos en incidencias por causas indirectas también tienen un patrón fácil de identificar. Esto es su absorción de usuarios nuevos

Si la incidencia es de cobertura, los clientes de los recursos de radio que desaparezcan irán a otros nodos a mantener el servicio y estos nodos son identificables por el aumento en simultáneo en los usuarios que atiende respecto de los nodos en incidencias directas

Esta detección será útil para la identificación de los usuarios realmente involucrados en una incidencia y podremos conseguir clasificaciones más exactas del impacto de la incidencia (No Navega, Navega Lento, Navega igual, Navega Mejor)

Debemos dejar en claro que, por sus características, los eventos de transporte no generan incidencias indirectas.

En este punto hare un resumen de lo que he planteado:

Se habla de donde provendrán los datos para los análisis

Se establecen categoría de clasificación de los nodos para un orden en la investigación

- NODOS CON INCIDENCIA
- NODOS SIN INCIDENCIA
- NODOS FUERA DE SERVICIO
- NODOS CON INCIDENCIA INDIRECTA

Además se determinan las categorías para la clasificación de diagnóstico de falla para los nodos en incidencia directa

- Problemas de cobertura
- Problemas de transporte
- Comportamiento de usuarios

También se establece que una incidencia es un desvío en el tráfico y que las clasificaciones aquí planteadas determinan las características de las incidencias

SOBRE LA ADMINSITRACION DE INCIDENCIAS

SOBRE LA ADMINISTRACION DE LOS NODOS EN INCIDENCIAS, SU AGRUPACION Y SU VALORIZACION DE GRAVEDAD

Las incidencias la componen todos los nodos que comparten aspecto de temporalidad y espacialidad

Así podríamos decir que una incidencia de red en su dimensión temporal la componen todos los nodos que simultáneamente presentan desviaciones en su tráfico,

Así al inicio de una incidencia se tendrá un volumen de tráfico que no se pudo cursar por alguna causa que ya hemos identificado. El volumen de la perdida nos dará un indicador de gravedad (P1, P2, P3...y así)

Respecto del tamaño de la incidencia, se debe considerar lo siguiente:

El tamaño de la incidencia es el área que involucran todos los nodos detectados con incidencias de tráfico. Estas agrupaciones en un mapa se verán como "hoyos" de degradación al lado de los nodos sin eventos adversos

Así el verdadero tamaño de la incidencia el clúster lógico que crean las agrupaciones de nodos. Así por ejemplos dos agrupaciones de nodos una en la I Región y otra en la 8 Región a la misma hora, estaríamos frente a 2 incidencias

A medida que pase el tiempo, la gravedad y el tamaño de la incidencia irá disminuyendo en función de los umbrales de gravedad que hayan acordado los administradores de la red y deberá cerrarse una vez se detecte que todos los nodos han vuelto a su comportamiento habitual

En términos prácticos cada vez que se inicie en simultaneo una agrupación de nodos con incidencia y que estos guarden una relación espacial, se le asignara un "Numero de Ticket" y el monitoreo constante de la evolución del trafico ira actualizando el nivel de gravedad para determinar el impacto de las acciones correctivas y el ticket deberá cerrase una vez que todos los nodos han vuelto a sus valores habituales

Con la idea ya planteada de cómo se procederá a clasificar los nodos y sus agrupaciones es hora de hablar sobre los módulos que serán la interfaz entre los datos de la red y los usuarios

Los módulos principales serán:

Módulo de graficas donde solo aparecerán los elementos en incidencias en el tiempo y la cuantificación en términos de tráfico de la misma por agrupación por tickets activos y cerrados

Modulo Geo donde se representaran en un mapa las incidencias

Y las tablas de incidencias donde la incidencia tendrá información más acabada como las comunas involucradas, tipo de incidencia, numero de nodos en la incidencia, numero de nodos fuera de servicio, número de clientes afectados entre otras características que sean necesarias para la correcta Administración de la información

También se considera la creación de módulos que permiten la evaluación del modelo para saber si este requiere algún ajuste o si las predicciones y la detección en el desvío del comportamiento son aceptables. Este módulo es más técnico y está orientado a la Administración por parte de un datascience respecto del rendimiento del modelo

SOBRE LA AUTOMATIZACION DEL PROCESO DE ADMINISTRACION DE INCIDENCIAS

Objetivos:

- Mejorar la eficiencia y efectividad en la administración de incidencias.
- Reducir el tiempo de respuesta y resolución de incidencias.
- Aumentar la satisfacción de los usuarios y clientes.
- Facilitar la gestión y toma de decisiones de los responsables de la red.
- Facilita su implementación

Alcance:

- Automatización del proceso de clasificación, diagnóstico y asignación de ticket y gravedad de incidencias.
- Actualización constante de la información de la red para poder administrarla a nivel nacional.
- Registro histórico de las incidencias y nodos afectados.
- Identificación de nodos que presenten fallas recurrentes.
- Implementación de un plan de acción y un equipo responsable de su implementación y seguimiento.

SOBRE LA AUTOMATIZACION DEL PROCESO DE ADMINISTRACION DE INCIDENCIAS

Objetivos:

- Mejorar la eficiencia y efectividad en la administración de incidencias.
- Reducir el tiempo de respuesta y resolución de incidencias.
- Aumentar la satisfacción de los usuarios y clientes.
- Facilitar la gestión y toma de decisiones de los responsables de la red.
- Facilita su implementación

Alcance:

- Automatización del proceso de clasificación, diagnóstico y asignación de ticket y gravedad de incidencias.
- Actualización constante de la información de la red para poder administrarla a nivel nacional.
- Registro histórico de las incidencias y nodos afectados.
- Identificación de nodos que presenten fallas recurrentes.
- Implementación de un plan de acción y un equipo responsable de su implementación y seguimiento.

Desafíos

Como a la hora de este desarrollo no tengo las autorizaciones necesarias para las tareas de transferencia y carga de archivos en tablas de BD y tampoco cuento con los permisos para utilizar VertexIA para el uso de Python, desarrollaré todo el código en SQL, donde el modelamiento de las predicciones serán realizadas sin ayudas de librerías, sino en base a matemática estadística básica y buen criterio

SOBRE LA DETECCION DE INCIDENCIAS

A esta altura del documento tenemos claro que nuestras variables principales serán:

- Nodos
- Numero de celdas en nodo por FCN
- Número de usuarios cursando tráfico en nodo
- Volumen de tráfico del nodo

DEL ANALISIS INDIVIDUAL DE CADA NODO PARA DETECCION DE INCIDENCIAS

Para la detección de incidencias, se deben considerar varias variables de cada nodo, incluyendo el volumen de tráfico, el número de usuarios traficando en el nodo y el número de celdas. Es importante destacar que el número de celdas es una variable que debería permanecer estable en el tiempo. Si se observa una disminución en el número de celdas durante una incidencia, puede ser una señal de un problema en el nodo.

Es posible que las celdas desaparezcan por incidencias o por la acción de alguna persona capacitada para hacer este tipo de alteraciones. En ese caso, la falta de celdas será detectada, pero se deberá marcar como un apagado consciente una vez se haya conseguido la información sobre la justificación del apagado.

Este marcado deberá ser ejecutado por un operador NOC una vez que se haya encontrado la información necesaria. Se espera que estos eventos no sean recurrentes. Si no se encuentra una justificación sobre el apagado del recurso entre los datos de la compañía, el nodo quedará marcado como incidencia hasta conseguir la información que la justifica.

La idea es que, con el tiempo, esta bolsa de nodos disminuya y se pueda tener un control eficiente de lo que ocurre en la red. Además, es importante destacar que una vez que se ingrese la información de justificación, la predicción de este nodo se ajustará automáticamente para su nueva normalidad. Esto permitirá una detección más precisa y un mejor control de incidencias en la red.

SOBRE LA PREDICCION DE VARIABLES CON VARIACION HORARIA DEL TRAFICO Y DEL VOLUMEN DE USUARIOS CON TRAFICO

Antes de abordar los detalles sobre la selección de las muestras y el método matemático a utilizar para realizar las predicciones es importante destacar los desafíos que debemos afrontar para realizar predicciones más adecuadas a las necesidades del problema

DESAFIOS DE LA PEDICCION

La ventana de datos de predicción de este modelo es basado en las muestras de los últimos 60 días, separado por muestras días de semana y muestras fin de semana

Considere la siguientes graficas

En ellas se ha ordenado de mayor las muestras y sirve como base para entender el problema de la predicción

En la gráfica uno se muestra el comportamiento de un nodo para sus respectivos días, para una hora cualquiera. Digamos las 15:00

Las muestras de esta graficas se muestran estables en el tiempo, por lo tanto no habría mucha diferencia en los resultados de la predicción si tomamos el 100% o el 50% de las muestras

En la segunda grafica vemos que que existe una estabilidad alta en el 75% de las muestras mientras se observa una estabilidad disminuida para el 25% restantes en la muestras.

En este caso es evidente que las muestras del grupo de muestras con valores menores que la mayoría indican que hubo alguna incidencia que afecto el comportamiento previsto

En este caso es recomendable excluir estas muestras para la predicción y que los resultados de la misma sean más apegado a la realidad del comportamiento del nodo

En primer lugar, se utilizará la predicción de largo plazo diaria para determinar si un nodo está en incidencia o no. Si la predicción diaria indica que el nodo está en incidencia, entonces se ignorarán todas las muestras horarias correspondientes a ese día.

En el caso en que la predicción diaria indique que el nodo está en incidencia durante más de un día, entonces se deshabilitarán todas las muestras horarias correspondientes a esos días. De esta manera, se garantiza que la predicción horaria no tenga en cuenta los valores anómalos durante el periodo en el que se detectó la incidencia y evitará que nodos que horariamente se ven sin problemas pero lleven más de 60 día en incidencias queden escondidos (Falso buen comportamiento de red) de la vista del analista

Si después de deshabilitar las muestras horarias correspondientes a los días de incidencia, no queda suficiente información para realizar la predicción horaria, se utilizará la predicción horaria de otro nodo que tenga un comportamiento de largo plazo muy similar al nodo en incidencia.

En el momento en que la predicción diaria indique que el nodo ha vuelto a su comportamiento habitual, entonces se utilizarán nuevamente las muestras horarias para la predicción horaria. En este caso, si existe al menos un día completo útil indicado por la predicción diaria, entonces la predicción horaria será igual a los valores de ese día. Si la mejora persiste al transcurrir de los días, se irán agregando más muestras horarias al conjunto de valores aptos para la predicción.

De esta manera, se asegura que la predicción horaria sea lo más precisa posible, al mismo tiempo que se evita que los valores anómalos generados por una incidencia en el nodo afecten la predicción. Además, se garantiza que se utilicen todos los datos disponibles en el momento en que la predicción diaria indique que el nodo ha vuelto a su comportamiento habitual.