Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Propuesta integral para generar valor en los datos de audiencia de una empresa de medios. Un caso de uso con Machine Learning.

AUTOR: CLAUDIA PATRICIA MOLINARI

DICIEMBRE 2019

Resumen

Este trabajo toma como unidad de estudio una empresa de medios para la cual se desarrollará una recomendación basada en la gestión y análisis de datos para generar procesos que generen valor al negocio.

Para ello se realizará en primer lugar un relevamiento de sus capacidades e implementaciones actuales para identificar el nivel de digitalización de la empresa. Luego se analizarán las fuentes de datos existentes que pudieran proveer de información útil para desarrollar una solución integral de arquitectura de bases de datos que permita la explotación de los mismos. Como caso de uso se buscará un modelo analítico particular que puede servir como ejemplo o disparador de nuevas ideas. Este modelo se ensayará utilizando distintas técnicas de data mining y machine learning.

La empresa de medios cuenta con más de un sitio web donde los usuarios navegan para hallar información o entretenimiento. El caso de uso a desarrollar será un problema de clasificación donde se buscará predecir si un usuario de un sitio será o no recurrente en un período de tiempo. Para aquellos usuarios que se predicen como no recurrentes, como desarrollo posterior, podría buscarse un modelo de recomendación que pueda generar la recurrencia esperada.

Palabras clave: nivel de digitalización; gestión de datos; machine learning; recurrencia de usuarios.

Introducción

Explotar las nuevas fuentes de información con que cuenta una compañía puede mejorar radicalmente su performance. Y sin duda Big data tiene el potencial de lograr la transformación de los negocios tradicionales en negocios que basen sus decisiones en datos. Sin embargo para iniciar el proceso de transformación es necesario un cambio cultural en las compañías. (McAfee, 2012)

Evaluar y reconocer el nivel de capacidad analítica en que la empresa se encuentra es el primer paso. De este modo la organización estará mejor preparada en términos de desafíos y oportunidades (Lavalle, 2011). Descubrir la potencialidad de los datos propios y reforzarla con otros de fuentes externas pueden generar nuevas oportunidades de negocios y mejorar las existentes.

La gran mayoría de las empresas expresan entender la ventaja de contar con un gran volumen de datos, sin embargo no son muchas las que han desarrollado la capacidad de su análisis e implementación. Contar con datos variados que se van generando día a día y a gran con velocidad implica tener a disposición lo que se conoce como Big Data. Existe una amplia variedad de técnicas para su análisis, que aplican métodos estadísticos tradicionales y otros desarrollados más recientemente, que han podido sofisticarse gracias a la capacidad de procesamiento de las computadoras hoy. Estas técnicas que se basan en algoritmos, se encuadran dentro de lo que se conoce como Data Mining y Machine Learning. Aplicar estos métodos al Big Data de una empresa genera valor en el negocio y aporta un diferencial frente a su competencia.

La empresa de medios a la que se hará referencia en este trabajo trabaja con un modelo de negocio publicitario tradicional, es decir que la monetización se genera a partir de la venta de espacios para publicidad, a diferencia de empresas como Netflix donde el modelo de negocio se basa en la venta de suscripciones (Fernandez-Manzano, 2016). Es por ello que el modelo se centra en el volumen de audiencia y en los targets afines, generalmente demográficos.

En general, las empresas que cuentan con un sitio web tienen implementados mecanismos de etiquetado que les permite detectar cuando una página es desplegada ante el requerimiento de un usuario. Al mismo tiempo, los sistemas trabajan con un

sistema de cookies que permiten identificar entradas recurrentes a un sitio. De este modo los sitios pueden saber la cantidad de usuarios que ingresan, la fecha y hora de la visita, página por la que accede, páginas que visita, tiempo de permanencia, dispositivo y navegador por el que accede, entre otras. Luego, se utilizan herramientas de analítica web para obtener métricas resumen y evaluar la usabilidad del sitio. Sin embargo estas herramientas no proveen un análisis que permita detectar preferencias individuales por usuario.

Surge luego, el planteo de las siguientes preguntas ***¿qué comportamiento puede predecirse a partir de los datos a nivel individualizado?, ¿qué contenidos pueden recomendarse de manera personalizada a partir de dicha predicción?***

La respuesta a estar preguntas abriría una oportunidad de negocio al poder ofrecer al anunciante segmentos específicos de audiencia.

También nos preguntamos ***¿qué diferenciador de audiencia respecto a los competidores nos están mostrando los datos?*** Este diferenciador introduciría ventajas en la comercialización del sitio.

Otras preguntas, y su potencial incidencia en la mejora del negocio pueden surgir de un "brain storming" integrado entre las áreas comercial y de planificación de contenidos.

El objetivo principal de este trabajo es identificar el nivel de digitalización de una empresa particular, que por cuestiones de confidencialidad será referida como "Empresa Medios", y en base a ello y a partir de las fuentes de datos existentes, desarrollar una solución integral para generar valor a partir de los datos, que incluirá el desarrollo de un modelo analítico para un objetivo particular, de los varios que pueden plantearse para esta empresa.

Para cumplimentar el objetivo general se establecen objetivos secundarios. En primer lugar, relevar los ítems necesarios para clasificar el estado de avance en el proceso de digitalización que se plantean en la bibliografía existente (Lavalle, 2011)(McAfee, 2012). Luego, identificar fuentes de datos con potencial valor para el negocio y desarrollar un plan de pasos para generar una base de datos integrada. Como paso siguiente, y entrando en la parte cuantitativa, el objetivo es extraer una base de datos correspondiente a un período determinado y aplicar métodos de data mining y machine

learning para hallar un modelo que permita predecir si un usuario del sitio será o no recurrente en un período de tiempo.

A fin de ubicar a la empresa en alguno de los niveles de digitalización sugeridos en la clasificación de MIT-SLOAN (Lavalle, 2011), se realizará un relevamiento de las características digitales ya implementadas en la empresa, las potenciales y las que aún no posee.

Posteriormente se explorarán las bases de datos de audiencia existentes en las diferentes áreas y del estado de organización y acceso de las mismas.

Según la información relevada se realizará una propuesta para la integración de las bases que puedan ser útil para el desarrollo de un sistema que provea conocimiento de la audiencia de los sitios de la empresa.

Para mostrar la potencialidad de esas bases se extraerán datos que pudieran generar información que resultara útil para algún aspecto del negocio y se realizará la explotación de los mismos a través de algún procedimiento de data mining y/o Machine Learning.

En base al relevamiento realizado y a los resultados obtenidos del proceso de los datos elegidos, se hará una recomendación final de una solución integral.

1 El Camino hacia la transformación Digital

Transitar en el camino de la transformación digital involucra informarse, reflexionar e implementar acciones. Informarse para entender cuál es el foco de la transformación y cómo les fue a empresas que ya están transitando ese camino, y reflexionar sobre cuáles serán los objetivos adecuados para la empresa.

En este primer capítulo se referencian algunos trabajos que muestran la relevancia e impacto del desarrollo de una estrategia basada en datos dentro de una empresa. Al mismo tiempo se reflexiona sobre la responsabilidad que trae aparejada la recolección de datos, y se pretende concientizar sobre los riesgos de un uso inapropiado de los mismos. Luego se hace introduce un marco específico sobre los procesos más utilizados para la organización de una base de datos integrada y los principales métodos de aprendizaje automático.

1.1 Del negocio tradicional a un negocio basado en datos

Innovar en el negocio es ofrecer propuestas de valor a los clientes que permita a la empresa diferenciarse de sus competidores (Lehrer, 2018). Y en vías de lograr innovación, en los últimos años, el dato se ha puesto en el centro de le escena. Capitalizar las oportunidades que ofrecen los datos explotados con técnicas de Big Data es un camino que conduce a generar nuevas propuestas con valor en el negocio.

Sin embargo, todas las empresas, de algún u otro modo, han basado históricamente sus decisiones en datos. ¿Qué es lo que diferencian a estos nuevos tipos de datos que se han popularizado en los últimos años, conocidos como Big Data? Esencialmente se referencia de este modo a aquellos datos que cumplen con las tres características conocidas como las 3V:

Volumen: el avance de la tecnología trajo aparejada la generación de gran cantidad de datos. Mientras que varios años atrás la medida usual de consumo de datos se daba en cantidad de Bytes, actualmente el volumen de datos ha hecho necesario evolucionar en estas medidas. Actualmente se utilizan los zettabytes, habiendo pasado previamente por el kilobyte megabyte, gigabyte, terabyte y petabyte. Según datos de CISCO (CISCO,

2019), en 2017 el total de tráfico de datos de internet fue de 77 Exabytes por mes, y para 2022 pronostica 293 EB. Para 2022 el total de datos móviles será de 77 EB por mes y representará el 20% del total del tráfico por IP. (CISCO, 2019)

Velocidad: los datos que se generan a través de aplicaciones y por el uso de las redes sociales varían en tiempo real, permitiendo establecer status instantáneos y la toma de decisiones adecuadas en el momento oportuno.

Variedad: el avance de la tecnología ha diversificado el tipo de datos que se transmiten: fecha y hora de conexiones, solicitudes a los servidores, mensajes de texto, imágenes, videos, señales de GPS, comentarios en redes.

Además de las 3 V descriptas, actualmente se habla de una cuarta: *Veracidad* haciendo referencia a que la potencialidad de los datos es dependiente de una adecuada gestión de la base de datos y de los datos que dichas bases contienen.

Para aquellas empresas, tales como las empresas de medios, que tienen una interacción directa con los usuarios, las huellas digitales que los mismos van dejando a través de dicha interacción permite a las organizaciones acceder a información de las preferencias de las personas que navegan por su sitio. Las propuestas que surgen a partir de estos análisis crean una ventaja competitiva para las empresas al poder ofrecer servicios más personalizados. Las técnicas de Big Data son el soporte necesario para la creación de valor de estos datos. (George, 2014)

Sin embargo, ingresar en una economía digital tiene sus desafíos: la conectividad es el principal, y es el factor determinante para la digitalización. Por otra parte, la privacidad de los datos y los datos sensibles deben ser temas prioritarios en el procedimiento de tratamiento de datos.

1.2 La transformación responsable

Evaluar y reconocer el nivel de capacidad analítica en que la empresa se encuentra es el primer paso si se quiere pasar de empresa tradicional a empresa data-driven. De este modo la organización estará mejor preparada en términos de desafíos y oportunidades (Lavalle, 2011). Descubrir la potencialidad de los datos propios y reforzarla con otros

de fuentes externas pueden generar nuevas oportunidades de negocios y mejorar las existentes.

Podemos preguntarnos, ¿qué aspectos evidencian el grado de evolución de una empresa en el proceso de transformares en una empresa "data driven"?,¿Cómo se podría desarrollar una clasificación en base a dichos aspectos?

Varios autores se han referido al tema, con distintas aproximaciones. Al mismo tiempo que varios también han expuesto la necesidad de incluir en el concepto de manejo y uso responsable de los datos.

Un grupo de estudio del MIT Center for Digital Business trabajó sobre la hipótesis que una empresa que basa sus decisiones en el manejo de datos tiene mejor performance que aquellas que continúan con un sistema tradicional (McAfee, 2012). Este grupo llevó a cabo una encuesta a 330 ejecutivos de compañías de Norte América sobre sus prácticas en la administración de datos y tecnología, y al mismo tiempo recolectaron información sobre sus reportes anuales. Obtuvieron que las compañías en el tercio más alto de la industria en lo que a manejo de datos se refiere, tenían en promedio **un 5% mayor productividad y un 6% más de rentabilidad** que sus competidores.

Por su parte, el MIT Solan Management Review en asociación con el IBM Institute for Business Value llevaron a cabo una encuesta a 3000 ejecutivos, gerentes y analistas, cubriendo 30 industrias y 100 países (Lavalle, 2011). Como principal resultado obtuvieron que las organizaciones que mejor performance tuvieron, utilizan analítica de datos **cinco veces más** que aquellas que performance más baja. Además ponen los datos en un rol central de manera que son dos veces más probables de utilizar analítica para definir estrategias futuras como también como sustento imprescindible para el día a día de la operación. Sus decisiones se basan en un análisis riguroso de los datos en un porcentaje que supera en el doble a aquellas empresas con bajo rendimiento.

Los autores desarrollaron una clasificación de las empresas en tres estratos de acuerdo al grado de desarrollo en su capacidad analítica y en el uso de datos para la toma de decisiones: Aspiracional, Experimentadas, Transformadas, las cuales se detallarán en el capítulo siguiente (Metodología a Utilizar)

En otro trabajo, MacAfee y Brynjolfsson (McAfee, 2012) puntualizan que para la transformación de una compañía hay cinco aspectos fundamentales para generar el cambio:

- ✓ Contar con líderes que sepan establecer claramente los objetivos y plantear las preguntas correctas a responder a través del tratamiento de Big Data.
- ✓ Contar con profesionales con la habilidad y el conocimiento de trabajar con grandes volúmenes de datos, como por ejemplo, los científicos de datos.
- ✓ Contar con la tecnología adecuada para manejar el volumen, la velocidad y la variedad de los datos con estas propiedades.
- ✓ Colocar en un mismo lugar a los que pueden extraer la información y aquél que comprende el problema.
- ✓ Cambiar la cultura de la empresa desde el "¿Qué pensamos? hacia el "¿Qué sabemos?". Evitar que los datos justifiquen una decisión ya tomada, sino por el contrario tomar la decisión en virtud de la información que los datos proveen.

Además sugieren cuatro pasos prácticos para la transformación de una empresa:

1. Comenzar la transformación a partir de una unidad de negocios particular de la empresa como caso testigo. Elegir la que cuente con un líder con capacidad cuantitativa que cuente con un equipo de científicos de datos.
2. Identificar cinco puntos claves relacionados con cinco oportunidades de negocio basadas en Big Data y proyectar una semana a cada uno, cinco en total y que puedan ser abordados de equipos de no más de 5 personas.
3. Implementar un proceso para innovación de 4 pasos: experimentar, medir, compartir, replicar.
4. Compartir en internet los desafíos analíticos que puedan surgir a grupos interesados de todo el mundo.

Forrester (Boris Evelson-Forrester, 2018), define como empresas data-driven a aquellas que recolectan datos a través de sus operaciones o transacciones, los integran en data

lakes, data-warehouses o data-marts (Acosta, 2019), y realizan analítica para extraer información de esos datos. Es interesante el planteo que surge a partir de preguntarse si efectivamente los procesos de BI y analítica diseñados por una compañía extraen conocimiento de los datos que conducen a decisiones tangibles para el negocio. Introducen así el concepto de *insights-driven business*, que marcaría un nivel más, en la madurez de una empresa.

Bowen y Smith, por su parte, dan una serie de recomendaciones para una adecuada gobernanza de datos (Bowen, 2014), que si bien lo contextualizan en el ámbito de una empresa del área de la salud, son claramente extensibles a cualquier área. Las autoras puntualizan que una adecuada gobernanza de datos es clave para la gestión del riesgo ya que vehiculiza la posibilidad de contar con datos precisos, actualizados y consistentes, al tiempo que permite optimizar tiempo y recursos. Sugieren una organización de datos que cuente con distintos elementos:

- un diccionario de los mismos que contenga definiciones, descripciones, fuentes y criterios de consistencia.
- una base de datos maestra con datos ya "limpios" sin duplicados ni errores.
- un listado de políticas para la integridad de datos y educación de los usuarios de los datos
- proveedores de IT que garantice la adecuada implementación de los puntos anteriores.

La generación de la base de datos maestra es un punto clave. Actualmente la existencia de datos de diferente tipo se ve potenciada por las transacciones generadas por los individuos a través de distintos sistemas, por la interacción a través de dispositivos móviles, por la utilización de las redes sociales, por la creciente transmisión de archivos multimedia tales como videos, fotos y audios. Todos ellos son factores que contribuyen a las 3V que definen al Big Data, y que generaron en las organizaciones un impacto que llevó a cambios disruptivos. En efecto, las empresas debieron establecer nuevas estrategias para el almacenamiento y administración de estos datos de gran volumen, constantemente en cambio, y de gran variedad.

"El potencial para la toma de decisiones es enorme, pero la evolución tecnológica también posibilita usarlos en tiempo real, introduciendo algoritmos en los procesos o productos, que hagan uso de estos y los dote de mayor inteligencia, eficiencia e innovación."(Fernandez Blanco, 215)

En el mismo trabajo de Bowen y Smith, las autoras puntualizan que desde la dirección de la empresa debe establecerse la integridad de los datos como una prioridad, generando una atmósfera general de la importancia de preservar dicha integridad. Sugieren la creación de un Comité Ejecutivo de Seguimiento formado por representantes de diferentes áreas que valide el diccionario de datos y revise su actualización y que sea el encargado de establecer las políticas mencionadas anteriormente y verificar su cumplimiento en el tiempo.

El inicio de una empresa en el camino de transformación involucra no sólo el desarrollo de una estrategia para alcanzar un uso eficiente de los datos sino también un análisis profundo sobre un uso responsable de los mismos que devenga en acciones concretas para asegurarlo. La importancia de este aspecto se pone de manifiesto a través de las diferentes iniciativas de regulación que van surgiendo en el mundo. Como referente más importante debe nombrarse la GDPR (General Data Protection Regulation), que es la más actual y posiblemente más específica regulación surgida a raíz de Big Data. Rige en Europa pero aplica a todos los datos que se comercialicen en este continente, aunque sean generados en otro país. Fue sancionada en mayo de 2018. Regula la extracción, almacenamiento y uso de datos personales.

Al encaminarse hacia una empresa data-driven, la organización debe establecer un comité o entidad que tenga a su cargo la gobernanza de los datos. Es en este ámbito donde debe darse la discusión sobre la responsabilidad en el uso de los mismos. Caben reflexionar sobre la forma en que los datos son extraídos: ¿el usuario es consciente de la forma en que serán utilizados sus datos? ¿el usuario dejó sus datos con la intención de que fueran utilizados? ¿Se le dio aviso al usuario en algún momento que sus datos podrían ser utilizados?

La discusión ética sobre el uso de los datos plantea otros cuestionamientos y no sólo los ligados a la concientización de los usuarios sobre su uso. Se deben plantear también

temas asociados a las consecuencias del tipo de uso que se hará de los mismos, por ejemplo, ¿las acciones implementadas a partir de los datos pueden colateralmente generar discriminación de algunas personas? ¿pueden generar desigualdad en el acceso al conocimiento? Tufekci (TUFEKCI, 2015) aborda el punto de cómo los algoritmos manejan lo que resulta visible a los usuarios y los que no, y sus posibles consecuencias, y cómo las redes pueden influir en movimientos sociales y en cuestiones políticas como las elecciones.

En relación al primer tipo, como ejemplo concreto de asimetría de la información, hace un comparativo entre un periódico, donde un editor es quien toma la decisión sobre la publicación o no de la nota, y las noticias en las redes sociales, donde un algoritmo decide qué se muestra a cada persona. En el primer caso la decisión resulta igualmente visible para todos los lectores. No hay quienes vean una nota diferente del resto, todos ven lo mismo. Lo mismo ocurre con las noticias en un canal de televisión. Un suceso puede informarse o no pero todos reciben lo mismo. En el caso de Facebook, por ejemplo, los algoritmos deciden en forma individualizada el contenido que resulta visible para cada usuario, quienes a su vez, en la gran mayoría, desconocen esta manipulación del News Feed.

Como ejemplo de violación de privacidad pone en cuestionamiento lo que se conoce como publicidad "customizada" ya que en la intención de usar algoritmos para predecir preferencias de la gente puede llevar a revelar orientación sexual, opinión en temas religiosos y/u orientación política. De esta forma podrían estar revelando cuestiones privadas de las personas que las mismas no quieren mostrar e incluso aspectos o situaciones que la persona desconocía.

En relación a la discriminación que pueden generar los algoritmos el autor cita el caso de la protesta en Ferguson, Missouri, en 2014 originada luego que un policía mató a un adolescente afro-americano que estaba desarmado. Esta protesta generó posteriormente una serie de demostraciones a nivel nacional sobre la desigualdad racial en los pueblos pequeños, en lo que a la justicia y al comportamiento de la policía respecta. El autor asegura haber documentado que el algoritmo para el News Feed de Facebook había suprimido las noticias sobre las protestas, considerando que no cumplía las condiciones para considerarlo "noticia relevante". Por su parte Twitter , que no utilizó filtrado a

través de algoritmos, reflejó el impacto de la protesta a través de los millones de tweets de los ciudadanos indignados. Si Twitter hubiera también filtrado las conversaciones sobre el tema, el problema hubiera quedado circunscripto a Ferguson y el problema de desigualdad racial no hubiera surgido a nivel país.

Como un ejemplo de la influencia del algoritmo de Facebook en lo político el autor del artículo menciona que en 2010 la red social realizó un experimento en el cual demostró que Facebook podía cambiar la intención de voto de las personas, manipulando el tipo de mensaje que se les mostraba.

1.3 La organización de las bases de datos

Con la generación de datos que cumplen con las 3V descriptas en la sección anterior, las empresas debieron establecer nuevas estrategias para el almacenamiento y administración de estos datos. *"El potencial para la toma de decisiones es enorme, pero la evolución tecnológica también posibilita usarlos en tiempo real, introduciendo algoritmos en los procesos o productos, que hagan uso de estos y los dote de mayor inteligencia, eficiencia e innovación."*(Fernandez Blanco, 215)

Para poder explotar el potencial de los datos es necesario considerar arquitecturas que permitan almacenar grandes volúmenes de datos y al mismo tiempo provean de un procedimiento eficiente para su actualización o cambio. También deben contemplar la concurrencia de usuarios, o sea la posibilidad de que varios usuarios soliciten información del sistema al mismo tiempo. Los sistemas tradicionales de bases relacionales no resultan ser la opción más adecuada para solucionar los desafíos del Big Data. Son necesarias otras opciones de almacenamiento y de gestión. Las premisa de tener bases de datos normalizadas, con reglas que priorizan la y buscan evitar la redundancia, tuvo que cambiar hacia el uso de bases no relacionadas (también llamadas, aunque incorrectamente, noSQL).

Estos nuevos diseños resignan algunos de los principios de las bases relacionales y están más enfocados a asegurar persistencia y disponibilidad, aún cuando exista redundancia, la cual muchas veces es necesaria para el manejo de los datos para permitir por ejemplo, acceder a los registros completos de un individuo.

1.4 Business Analytics-Data Mining

Se entiende por Business Analytics al proceso de generar conocimiento de valor a partir de la información que aportan los datos. En la actualidad este proceso incluye gran variedad de métodos sofisticados que involucran modelos estadísticos y, algoritmos de data mining y de machine learning que permiten el análisis exploratorio de los datos, análisis de correlaciones y predicción. Estos métodos pueden dividirse en cuatro grupos:

- Clasificación: se examinan los datos donde las unidades están clasificadas según las categorías de un atributo y se desarrollan reglas que permiten clasificar nuevas unidades cuya clasificación es desconocida.
- Predicción: es similar a la clasificación pero se quiere predecir un valor numérico:
- Recomendación: son sistemas de recomendación on-line que utilizan las preferencias de los usuarios a nivel individual para ofrecer recomendaciones personalizadas.
- Análisis de sentimiento: sistemas orientados a captar la intención de las personas que se expresan a través de conversaciones sociales.

(Shmueli, 2018)

Son varios los métodos que pueden ensayarse en la búsqueda de un modelo adecuado para predicción (Gareth James, 2013):

- Naive Bayes
- para el caso de una variable respuesta binaria, Regresión Logística.
- K-vecinos más cercano
- Árboles de clasificación
- Random Forest
- Boosting
- Support Vector Machines
- Redes Neuronales

Cualquiera sea el método aplicado, la decisión de adoptarlo se basa en la performance del modelo, que se refiere a la capacidad predictiva que del mismo. Hay varios criterios

que pueden utilizarse en un problema de clasificación, entre ellos la accuracy, recall, el área para la curva ROC.

Accuracy: porcentaje de casos bien predichos (positivos y negativos) sobre todos los casos.

Precisión: porcentaje de casos predichos como positivos correctamente, sobre el total de casos predichos como positivos.

Recall: porcentaje de casos predichos como positivos correctamente, sobre el total de positivos verdaderos.

f1-score: pondera la precisión y el recall. Su valor se calcula como:

$$2 \text{ x precisión x recall}/(\text{precisión} + \text{recall})$$

Curva Roc: es una medida representación del porcentaje de falsos positivos vs verdaderos positivos. Si la clasificación se hiciera al azar, la predicción sería 50% para cada caso y por lo tanto los puntos se alinearían sobre la recta y=x y el área bajo dicha recta sería 0.5. Luego, el área bajo la curva es un indicador de la capacidad de predicción del clasificador, cuanto mayor sea de 0,5 mejor predictor. (Del Rosso), (Gareth James, 2013)

Para la estimación del modelo no deben utilizarse los mismos datos que para su validación, por ello suele dividirse la base de datos en dos conjuntos: conjunto de entrenamiento y conjunto de validación. Suelen además, utilizarse para ello un 80% vs 20% de los datos.

Sin embargo, para asegurarse que el modelo no es sensible a la división de la base, es recomendable utilizar Validación Cruzada que valida el error promediando los errores que se obtienen utilizando diferentes selecciones de conjuntos de entrenamiento y validación.

2 Metodología y técnicas utilizadas

En este capítulo se describirán los procesos específicos utilizados en cada etapa del trabajo realizado para cumplimentar los objetivos planteados. Se referencia el criterio de clasificación utilizado en la evaluación general de la empresa en el camino a la digitalización, el tipo de arquitectura de base de datos en que se basará la sugerencia de organización de la base de datos, y se especifican los criterios y métodos aplicados a la base de datos específica utilizada.

La unidad de análisis para el primer objetivo fue la empresa, sobre la cual se sugerirá un proceso de avance de uso de la información digitalizada y la generación de un sistema de base de datos integrada.

Por otro lado y como eje principal del análisis cuantitativo las unidades de análisis fueron los usuarios de los sitios web de la empresa. Sobre ellos se buscó obtener información a partir de las fuentes de datos integradas que expliquen el perfil de la audiencia de los sitios según los atributos para los cuales se pueda obtener información como ser lugares de acceso, dispositivo, horas de mayor acceso, secciones más visitadas, demográficos de los usuarios, cantidad de notas vistas. Como un caso de uso se consideró una base de datos particular sobre la cual se aplicaron métodos de análisis y algoritmos de Machine Learning para obtener un modelo que predijera la recurrencia de un usuario, característica muy importante de un usuario para un sitio, generalmente denominada como "engagement".

2.1 Clasificación de la empresa según su nivel de digitalización

La Empresa Medios cuenta con un alto nivel tecnológico en la producción de contenido audiovisual.

Los usuarios que ingresan a sus sitios web diariamente dejan registro de su navegación. Generan así un gran volumen de datos, con gran velocidad ya que los ingresos son continuos durante las 24horas del día, y de gran variedad ya que quedan registrados datos de tipo cuantitativo, como puede ser la cantidad de minutos que permanecen en el sitio sino también de tipo categórico como el tipo de dispositivo de dónde se conectan (móvil, desktop), y de tipo no estructurado como los comentarios que dejan en las notas. Estas características encuadran el problema dentro del concepto de Big Data.

Para analizar el estado de digitalización de la empresa se utilizó la clasificación del MIT-SLOAN:

- *Aspiracional*: son las que están más alejadas de basar sus procedimientos y acciones en el análisis de los datos. Están más preocupadas por reducir costos y cuentan con menos elementos para lograr incorporar el uso de analítica. En general su uso de analítica no es para tomar decisiones, sino para justificar las ya definidas. Tienen una habilidad limitada en la organización y administración de los datos, y el principal obstáculo que presentan en vías de un avance en materia analítica es la falta de entendimiento de que elevar el nivel de análisis de los datos genera valor para el negocio.

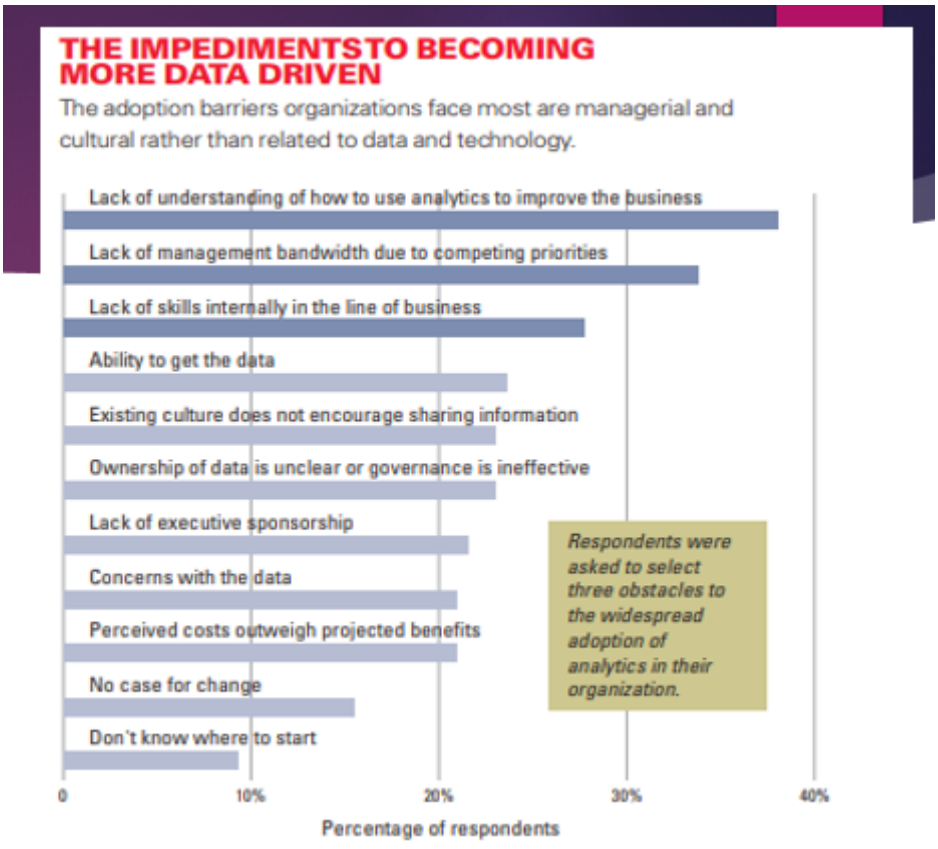
- *Experimentadas*: cuentan con cierta experiencia en analítica y están en desarrollo de mejoras en la colecta, administración y análisis de los datos. Sin embargo la propiedad de los datos no es clara o la gobernanza no es efectiva. Tienen una habilidad moderada en la captura, integración y análisis de los datos, pero tienen una capacidad limitada para compartir la información y los descubrimientos que surgen a partir de los datos. Cuentan con algunos intentos rigurosos del uso de datos para la toma de decisiones pero aún deben crecer en el uso de los mismos para establecer estrategias a futuro.

- *Transformadas*: son las más avanzadas en el uso de analítica. Tienen experiencia en el uso de datos como motor del funcionamiento de la empresa. Dicha característica les aporta un diferenciador competitivo. Están menos focalizadas en bajar costos que las Aspiracionales y las Experimentadas, en parte porque ya automatizaron varios de sus procesos para la obtención de información. El uso de los datos es determinante en sus acciones y han logrado hacer crecer sus ingresos y aumentar la adquisición y retención de los clientes. Son efectivos en compartir la información y los descubrimientos que surgen de los datos. Sus mayores obstáculos son comprender aún como aumentar el nivel de analítica para generar más valor, gestionar la banda ancha para lograr ser más competitivos y mejorar la accesibilidad a los datos. Tienen un nivel avanzado en la captura, integración y análisis de los datos.

En vías de realizar una recomendación para elevar el nivel de digitalización se partió de las sugerencias del mismo trabajo del MIT, donde los autores enumeran una serie de recomendaciones:

- Pensar de entrada en el mayor desafío, ya que tener en vista una gran idea minimizará la dificultad de superar desafíos menores.
- Comenzar por establecer preguntas y no datos. Los datos tienen que ser el vehículo para responder una pregunta.
- Generar conocimiento a partir de los datos que guie acciones concretas para el negocio.
- No quedarse con las funcionalidades desarrolladas sino estar en la búsqueda continua de una gama más amplia de capacidades y de un uso más avanzado de las ya implementadas.

También se tuvieron en cuenta las barreras que resultaron más frecuentemente mencionadas en la encuesta llevada a cabo por el grupo del MIT:



Infografía según Encuesta sobre 3000 Ejecutivos, gerentes y analistas de 30 industrias y en 100 países. (Lavalle, 2011)

Considerando que una parte importante del proceso de digitalización involucra la implementación de una gobernanza de datos, se tuvieron en cuenta para el diagnóstico de esta empresa en particular y para la posterior recomendación, las sugerencias de Naomi Clarke (Clarke, 2019), quien establece un marco de trabajo basado en el control interno, para mejorar la precisión y calidad de la información basada en datos para la toma decisiones, poniendo en el centro de la escena a la gobernanza de datos.

Clarke refiere a la gobernanza de datos como de vital importancia para responder a al reclamo de los clientes de contar con datos precisos. Más aún, la define como el mecanismo para hacer cumplir reglas y estándares internos para el control de calidad y administración de los datos dentro de una empresa. A fin de definir el marco para una adecuada gobernanza de datos enumera los componentes a tener en cuenta:

1. Ambiente de control: se refiere a cómo la parte directiva y gerencial se involucra en el proceso.

2. Gestión del Riesgo: es el núcleo del marco. Describe cómo se identifica, evalúa y mitiga el riesgo a partir de los controles adecuados.

3. Actividades de Control: se refiere al monitoreo de las actividades establecidas que aseguren que las mismas continúan siendo efectivas para minimizar el riesgo por el cual que fueron concebidas. Pueden definirse Key Control Indicators (KCIs) a tal fin, que podrían incluir medidas sobre la calidad de los datos que pueden a su vez indicar el estado del negocio en la arquitectura de datos definida.

4. Información y Comunicación: la información debe ser confiable, a tiempo, y accesible para la toma de decisiones.

5. Monitoreo de Actividades: la idea es monitorear la información para identificar nuevas líneas de trabajo y para implementar cambios que pudieran resultar necesarios.

Para evaluar en particular a la Empresa Medios se realizó un relevamiento de los datos disponibles, de la existencia de reportes derivados de los mismos, y de su uso en la toma de decisiones.

2.2 Sobre la configuración de una base de datos integrada

Para elaborar una propuesta sobre una posible estructura de base de datos se tuvo en cuenta el tipo de datos presentes en cada fuente de datos y los nuevos tipos de almacenamiento.

La gran cantidad de datos que las empresas recolectan ha exigido un cambio en la forma de almacenamiento. Surgieron así los data warehouses como potentes almacenes de datos para el caso de bases estructuradas.

Sin embargo, la diversidad en el tipo de datos a los que actualmente puede accederse necesita de otras soluciones. Surge entonces el concepto de *data lake*.

Un data lake es un almacén donde se guardan datos de todo tipo de datos provenientes de las distintas áreas del negocio. Estos datos pueden ser estructurados, o no estructurados tales como logs de chats, emails, imágenes, audios y videos.

Un data lake es relevante en dos aspectos principales que resultan característicos de la empresa Medios en estudio:

- para el negocio de la empresa no sólo son importantes datos de tipo cuantitativo sino que puede encontrarse valor en la analítica de otro tipo de datos como ser las opiniones de los usuarios en los mismos sitios o en las redes. Por lo tanto se hace necesario encontrar una forma de almacenar datos de gran cantidad y de diferente tipo.
- puede no haber un plan actual de la utilidad de explotar datos de cierta fuente, pero existe una visión de utilizarlos en un futuro.

2.3 Análisis de un conjunto de datos

Con el objetivo de presentar un caso de aplicación al negocio de la empresa, utilizando métodos de Machine Learning, se consideró una base de datos extraída por el área de sistemas de la empresa. Esta base contenía un registro de cada usuario que ingresó en los sitios de la empresa durante un período. En cada caso proveía los datos del tipo de acceso (dispositivo, navegador, ubicación geográfica) y las urls que visitó. Dicha url constituye una evidencia de los intereses del usuario.

Como un usuario puede ingresar más de una vez en el período, se generó una nueva base conteniendo los datos de manera agregada y que resultaba más adecuada para el análisis. Se consideró la creación de nuevas variables como también la codificación de algunas de las existentes. Se analizó la existencia de valores perdidos y se evaluó su naturaleza y el impacto que podrían tener en el análisis para así tomar la decisión de eliminarlos o imputarlos.

Para evaluar la incidencia de cada variable en el conjunto de datos analizados se aplicaron técnicas de resumen, visualización y limpieza de datos .

El objetivo del análisis fue generar un modelo que permitiera predecir si un usuario que ingresa del sitio podía resultar ser un usuario recurrente o no recurrente. Esta variable es de tipo cualitativa y binaria y fue generada a partir de la cantidad de entradas de cada usuario. Luego, se ensayaron los métodos mencionados en la sección 1.4 adecuados para un problema de Clasificación.

Para evaluar la capacidad de predicción del método ensayado se utilizó accuracy y el Área bajo la curva ROC (AUC) .

La lectura de los datos, tratamiento y posterior análisis e implementación del proceso de Machine Learning se realizó con R.

3 Resultados

A continuación se exponen los resultados obtenidos para el relevamiento del estado de digitalización de la empresa, sobre las bases de datos disponibles , su origen y elementos, y finalmente se detallan los procedimientos realizados sobre la base de

datos, los procesos de Machine Learning aplicados , los resultados de cada uno y el modelo seleccionado.

3.1 Estado de digitalización de la empresa

Para evaluar el estado de digitalización de la Empresa Medios en primer lugar se analizó el contexto en que se ha desarrollado el modelo de negocios y la evolución del mismo en los últimos años.

CPM es una empresa de medios, por lo cual su modelo de negocios principal ha sido tradicionalmente, la venta de publicidad. Desde su inicio la oferta publicitaria fue en la TV lineal, y hoy sigue siendo el corazón del negocio, pero la apertura de sitios de internet asociados a los canales de TV de la empresa incorporó al negocio la oferta de espacios publicitarios digitales.

En este contexto, la empresa siempre utilizó datos para sus decisiones ya que las transacciones de publicidad se basan en ellos: los espacios en TV lineal basan su precio en el rating del programa, y la venta de espacios digitales se basa en la cantidad de usuarios y/o páginas vistas del sitio. Más aún, los datos actualmente disponibles en la empresa permiten analizar la composición de la audiencia por segmentos demográficos, lo cual aporta una visión cualitativa de la audiencia para la toma de decisiones en la generación de contenido.

¿En qué consistiría entonces para esta empresa ingresar en un proceso de transformación data-driven, si siempre trabajó basada en datos?

La diferencia la aportan los datos que los usuarios van dejando con sus huellas digitales: navegación en los sitios, sintonización de programas en TVs con cajas digitales, y en dispositivos móviles. La digitalización de los sistemas permite contar con datos propios, y no sólo con datos provistos por terceros. Se habla actualmente de "first party data" y de "third party data", diferenciando los datos que una empresa obtiene por sus propios registros, de los datos que un tercero le aporta.

Hasta ahora la Empresa Medios, tanto en TV como en Digital ha trabajado con datos provistos por empresas de mediciones que aportan resultados consolidados totales o para segmentaciones. Hoy el desafío es explotar los datos propios, que permiten

explorar con tanta granularidad como los mismos datos permiten según la información que traen, y utilizarlos de un modo inteligente, integrando las distintas fuentes internas y las fuentes externas disponibles, abriendo posibilidades de nuevas vías de monetización y/ o de mejorar los actuales.

Teniendo en cuenta la clasificación propuesta por el MIT-SLOAN, y a partir del relevamiento realizado, la Empresa Medios podría clasificarse como una ***empresa Experimentada***. Sin embargo, se detectan características comunes con empresas Aspiracionales en algunos aspectos analizados. A continuación se detalla el estado de situación observado:

- *Sobre el uso de analítica de datos*:

La empresa cuenta con varios sistemas de reporte de sus audiencias tanto para TV lineal, como para sus sitios de internet, sus videos on-line, y sobre sus seguidores en las redes sociales. Estos sistemas consisten en herramientas desarrolladas por terceros, que son adquiridas a través del pago de una licencia, y que permiten conocer el volumen de la audiencia como también características demográficas. En el caso particular de la audiencia on-line permite conocer además el origen de los usuarios tanto en lo referido a su ubicación geográfica como al dispositivo desde el cual se conectan. Puede decirse entonces que la empresa cuenta con experiencia en analítica de datos cuando éstos ya están consolidados en un sistema de reporte.

- *Sobre el flujo de la información*:

La empresa cuenta con un flujo importante de la información resultante del análisis de los datos. En efecto, existe un área que genera reportes diarios, semanales y mensuales sobre TV y mensuales sobre los datos de los medios digitales hacia las distintas gerencias y áreas de programación. El área de contenidos digitales cuenta con un sistema de información continuo, con frecuencia diaria e incluso en tiempo real, que se utiliza para la evaluación y toma de decisiones para el contenido de los sitios.

- *Sobre el uso de la información*:

Los descubrimientos que pueden surgir del análisis de la información son utilizados para tomar decisiones pero también se observa que, en muchos casos, se utilizan para

justificar decisiones que están basadas en la experiencia y la intuición, lo cual es característica de las empresas Aspiracionales.

- *Sobre el valor de los datos para el negocio:*

No se observa un convencimiento en todas las jerarquías de la organización de que el avance en materia analítica generará valor para el negocio, y éste es el punto en común de la empresa con las llamadas *Aspiracionales*. En el contexto económico actual del país y del mercado, hay una intención de disminuir costos sin tomar consciencia que la inversión en gestionar y utilizar datos es un camino a la optimización de tareas y resultados. Hay una aspiración a ser una empresa data-driven más bien impulsada por las tendencia y el contexto actual del mundo de los negocios a nivel internacional, que por un conocimiento pleno del beneficio que ello implicaría. La falta de dicho convencimiento genera impulsos esporádicos que no llegan a transformarse en proyectos concretos al obstaculizarse por cuestiones de costos y/o de recursos. Hay impulsos individuales pero que se ven frustrados por la falta de generación de un proceso concreto apoyado por la plana más alta de la organización.

- *Sobre la colecta de los datos*:

Existe en la empresa colecta de datos digitales sobre los accesos a los sitios y el uso de los mismos, como también sobre las redes sociales y plataformas de consumo de video. Sin embargo estos datos no son almacenados en un repositorio accesible para realizar un análisis integrado de los mismos. Los datos quedan en repositorios externos y no son utilizados de manera directa, sólo a través de los sistemas de proveedores externos que ofrecen datos consolidados. Sin embargo hay algunos desarrollos particulares, realizados para un objetivo muy específico.

- *Sobre la integración de los datos:*

Si bien hay iniciativas para logar una visión 360 de los datos y del negocio, no existe una política de integración de datos. Se han implementado procedimientos para ofrecer pauta más individualizada a partir de la segmentación de audiencia, pero son limitados y está basados en una única dirección de los datos, sin integración con distintas fuentes.

- *Sobre la propiedad y gobernanza de los datos*:

Este punto es fundamental para el proceso de transformación de una empresa, y es el punto más débil en la Empresa Medios. Al no existir un repositorio consolidado de datos, no hay una consciencia de la necesidad de centralizar la gobernanza de los mismos, que a su vez establezca una política de acceso y uso de los datos. Sólo se observa una estructura de este tipo para el data-warehouse de la empresa que contiene datos sobre clientes e ingresos de publicidad.

3.2 Bases de datos disponibles

Existe acceso a fuentes de datos de la audiencia individualizada de los sitios, pero a través de interfaces proporcionadas por plataformas que se encargan de la recolección y almacenamiento de los mismos.

En la mayoría de los casos, la individualización no permite una identificación del usuario ya que se provee un Id o una cookie anónima. Es el caso de los datos de navegación que se almacenan a través del servicio de Google Cloud y del servicio de Cxense. Éste último tiene la particularidad de que la cookie solo está vigente por 30 días.

A través de la plataforma de YouTube vía consulta a la API es posible acceder a data más granular que la obtenida a través de la interfaz analítica, pero del análisis de las métricas y dimensiones disponibles no se observa que pueda llevarse al nivel de individualización de usuario aún cuando fuera en forma anónima.

También a través de las APIs correspondientes es posible acceder a datos del tráfico de las redes sociales vinculadas a los sitios. En estos casos la información también es agregada por grupos de usuarios y no permite identificación.

Se debe tener en cuenta que los datos provenientes de las API de YouTube y de FB no sólo proveen datos cuantitativos sino también de tipo cualitativo como los comentarios en las publicaciones y videos .

En cada base se encuentran disponibles los siguientes datos y atributos :

- En Google Cloud: puede obtenerse datos, a nivel granular por usuario, de todos los datos de navegación por sesión, es decir,

→ datos de localización, por ej, fecha, hora de inicio, hora de finalización, geografía. Son datos de tipo nominal u ordinal.

→ datos del dispostivo de conexión, por ej, tipo de dispositivo, modelo, navegador. Son datos de tipo nominal u ordinal.

→ datos de la conexión, por ej,ancho de banda, velocidad de descarga. Son datos de tipo cuantitativo.

→ datos del recorrido, por ej, origen de la visita (dato nominal), duración, cantidad de páginas vistas (datos cuantitativos).

- En Cxense: pueden obtenerse datos similares, pero no más que para un período de 1 mes, por lo cual, si se quiere contar con datos históricos, se deben ir guardando las consultas. En particular en este sistema el Id de Usuario resulta una clave externa para relacionar la base con otra proveniente de otro sistema (Gigya) que permite incorporar información demográfica y de intereses para algunos usuarios. La empresa no sigue una política de registro de usuarios en sus sitios ya que no los contenidos son considerados públicos y no se restringirá el acceso a los mismos a través de un muro de pago o de registro. Solamente en algunas circunstancias de participación específica se invita al usuario a registrarse. Para aquellos que se lo hacen y se registran a través de una red social es posible obtener los datos del perfil que el usuario hizo público en dicha red. Esta opción es la única que permitiría combinar los datos de navegación en el sitio con una fuente de datos externa y poder salir del ecosistema de los propios sitios.

- En YouTube: los datos que pueden recolectarse son a partir de la conexión a la API y resultan en general datos de localización, conexión y navegación pero se agregan datos específicos del uso de videos: cantidad de videos reproducidos, duración total de las reproducciones, tiempo promedio de visionado. Todos estos datos resultan de tipo cuantitativo. Al ser también una plataforma de red social se cuenta con datos no cuantitativos tales como cantidad de suscriptores, cantidad de interacciones y comentarios.

- En Redes sociales (diferentes de YouTube):

Pueden obtenerse datos tales como:

→ cantidad de usuarios (alcance),

→ cantidad de interacciones,

→ cantidad de impresiones,

→ tiempo promedio de reproducción de un video

→ cantidad de impresiones del post

Todas estas métricas son de tipo cuantitativo.

- Sistema interno de Datos de Videos: es un sistema desarrollado por la empresa y no por terceros, por lo cual tiene disponible la data al nivel más crudo pero restringida a los atributos implementados para medición. En particular, cantidad de plays, cantidad de pre y post rolls, duración de la reproducción.

Existe un data warehouse con datos propios del área comercial pero que se excluye del alcance de este trabajo donde se está enfocando obtener datos para obtener información de valor sobre la audiencia.

3.3 Procesamiento de datos. Un caso de uso: predicción de usuario recurrente

Para el análisis se partió de una base de datos descargada especialmente por el área de Sistemas de la empresa, desde la API del sistema Cxense.

La base de datos disponible constaba con el registro de 690.341 entradas de usuarios a algún sitio de la empresa, y para cada una de las cuales se contaba con 45 atributos indicadores del lugar desde donde se accedió, tipo de dispositivo, tipo de navegador, zona geográfica, url que se visualizó, origen de la visita(si provenía de una búsqueda orgánica, o de una red social o el ingreso había sido en forma directa), entre otras. No todas las variables resultaban de interés en el problema, por lo cual se utilizó una subconjunto de ellas, que se detalla a continuación:

Variable	Descripción	Valores
UserId	Identificador único por usuario	Texto único por cada navegador conectado

Device Type	Tipo de dispositivo desde el que ingresó el usuario	Desktop, móvil,tablet
Host	Dominio donde se aloja el sitio	los 3 dominios de los sitios analizados
Referrer Host Class	Tipo de dominio desde donde llega el tráfico.	Directo, Buscador, Red Social, Interno
ReferrerSocialNetwork	Indica la red social de que viene referido el usuario, si vino referido de alguna	Facebook, Instagram, Twitter, Pinterest, Google.
OS	Sistema Operativo	IOS, Android,Linux,Macntosh, Windows, Otros
URL	Dirección URL de la página visitada	Es un texto que contiene el nombre del sitio y las palabras claves del contenido de la nota que visita el usuario.
Pais	Ubicación geográfica de la IP	Abreviatura indicando el país.

El objetivo del análisis fue buscar un modelo basado en técnicas de Machine Learning que clasifique a los usuarios como recurrentes y no recurrentes en base a los atributos que indican la forma de acceso y la geografía, pero también en base a los intereses.

En primer lugar se realizó un análisis descriptivo de las variables. Todas ellas eran de tipo nominal.

La variable pais se recodificó utilizando los países más frecuentes en la base: Argentina, Uruguay, Paraguay, España, Estados Unidos y Resto. De manera similar se trabajó con el sistema operativo (os).

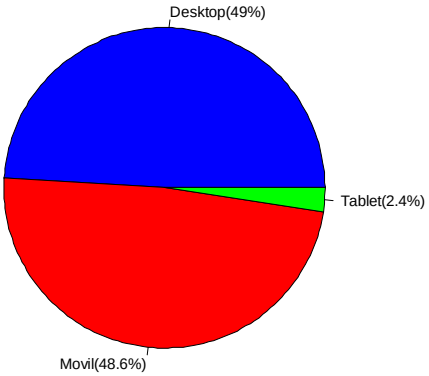
A continuación se presentan gráficos de torta para cada atributo:

- **UserId**:

Permite reconocer las sesiones de un mismo usuario. La base sin consolidar contiene un total de 690.341 identificadores pero que se repiten con cada entrada del mismo usuario. Hay 219.985 usuarios diferentes.

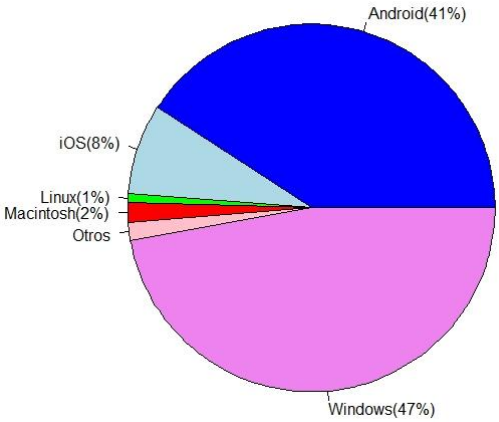
- **Dispositivo**:

No hay datos perdidos, todos los registros tienen un dispositivo asociado. 49% de los ingresos se produjeron a través de una computadora o laptop, el 48.6% por un teléfono móvil y un 2.4% desde una Tablet.



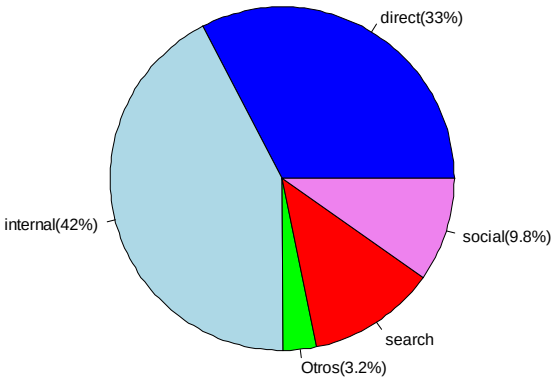
- **Sistema Operativo:**

Las categorías con al menos 1% de frecuencia son Android, Windows, IOS, Linux, Macintosh. El resto se agrupó en la categoría Otros. Había algunos casos sin sistema operativo identificado.



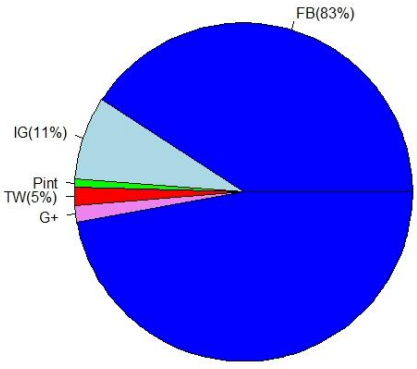
- **Fuente del tráfico (Referrrer host Class**):

Los usuarios visitan la mayoría de las páginas referidos desde otra página del mismo sitio o ingresando directamente la dirección web en barra de búsqueda. El atributo se refiere al origen de la visita, no del usuario, ya que un usuario puede entrar una vez direccionado desde un lugar y otra vez por otro.



- **Red Social Referida**:

De los usuarios que ingresan a través de una red social, el 83% proviene de FB.



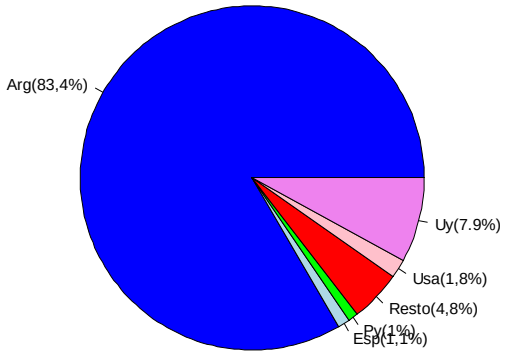
- **Host**: Indica el sitio que aloja la página visitada.

Host	Frec%
Sitio de Noticias del Espectáculo	2.4%

Sitio de Entretenimientos	4%
Página de móviles del sitio de Not de Espect	3.3%
Página para móviles de sitio de Noticias	5%
Sitio de Noticias	2%

- **Pais:**

La mayor parte de los ingresos se producen desde Argentina, seguidos por Uruguay, pero en proporción mucho menor.



Fue necesario consolidar la base de datos y crear nuevas variables para contar con una base adecuada para la aplicación de los métodos de aprendizaje automático.

Dado que las filas correspondían a ingresos a los sitios de individuos que podían repetirse ya que podían acceder más de una vez, se buscó una nueva configuración donde para cada usuario diferente (cookie) en las filas se obtuviera la frecuencia de las categorías de los atributos para el total de ingresos de ese usuario.

Luego, el preparado de la base de datos consistió de los siguientes pasos:

1. Para cada usuario se calculó la cantidad de sesiones registradas.
2. Se calculó la frecuencia de las categorías que presentaba cada usuario en el total de sus entradas.
3. Se generó una nueva base de datos compatibilizando los datos de los pasos 1. y 2. de manera que cada fila del nuevo set de datos corresponda a un

usuario diferente y cada atributo de la base original se desglosó en tantas como categorías presentó.

4. A partir de la cantidad de sesiones de un mismo usuario, se generó la variable binaria Recurrente que vale 1 si el usuario tuvo más de una sesión. Ésta será la variable de clasificación a predecir.

5. El único atributo indicativo de los intereses o atractivos por los que el usuario había ingresado al sitio, fue la url, ya que en la misma se indica el sitio en el que ingresó y las palabras claves de la nota que visualizó. Debido a ésto la url fue tratada como un texto al cual se le aplicaron funciones disponibles en el R para Natural Language Processing.

6. Con los procesos seleccionados de NLP se logró representar cada url en un vector de palabras individuales ("tokens").

7. La cantidad de palabras diferentes para el conjunto total de urls era muy grande, por lo cual se calculó la frecuencia de cada una y se seleccionaron para su uso aquellas que al menos había aparecido en el 1% de los casos.

8. Para cada una de las palabras seleccionadas se calculó la cantidad de veces que estuvieron presentes para cada usuario diferente.

9. Se consolidó una única base que contenía en cada fila registros agregados por categorías de las variables, incluyendo como atributos también las palabras resultantes del proceso de "tokenización" de las urls. De este modo la base final que se conformó como entrada para los procesos de Machine Learning contenía atributos numéricos que eran indicativos de las frecuencias con que dichos atributos se presentaban.

10. La variable a explicar: NoRecurrente, era la única de tipo Factor.

A continuación se detalla la metodología aplicada para los pasos 5 a 9 anteriores:

Para cada fila de las 690.341 entradas se desglosó la url en palabras utilizando el paquete tm de R. Quedó así conformado una lista de R que en cada elemento contenía la "tokenización" de la url. No todas las filas corresponden a usuarios diferentes, sino a ingresos (sesiones) diferentes , por lo cual había que asignar a

cada usuario la frecuencia de cada palabra indicativa de un interés. Se conformó un conjunto de datos nuevo que en la primera columna contenía el Id del usuario, y en la segunda una palabra de interés, por lo cual el tamaño del set de datos pasó a tener 4.069.033 filas. La cantidad de palabras diferentes que se formaron (utilizando un filtro de stopwords en español) fueron 12.943. Como es de esperarse, algunas palabras aparecían en muy pocos casos, por lo cual se aplicó un filtro a fin de no forzar el tiempo ni la memoria de procesamiento utilizando atributos que no son de interés. Se filtraron así las palabras que aparecían en al menos 35.000 casos del total de 690.341 (por lo menos en el 1% del total de filas).

Utilizando una subrutina en R se extrajeron, para cada usuario, los palabras de todas las urls que había visitado en las distintas sesiones. Luego, se calculó la frecuencia de los tokens y se incorporó al data frame que ya tenía agregada la información con las otras variables.

Los datos faltantes se reemplazaron por 0 dado que corresponden a categorías que no estuvieron presentes para el usuario y los atributos en las columnas corresponden a frecuencias observadas.

Los datos resultaron bastante balanceados ya que 61% de los casos son recurrentes y 39% no.

Considerando la base ya consolidada Se ensayaron los siguientes modelos particionando la muestra en un 80% para testing y un 20% para validación, obteniéndose la precisión indicada:

Método	Medida de precisión
Naive Bayes	Accuracy=70%, ROC= 75,2%
Árbol de Clasificación	Accuracy= 99,6%, ROC=99,7%
Random Forest	Accuracy= 99,9%, ROC=100%

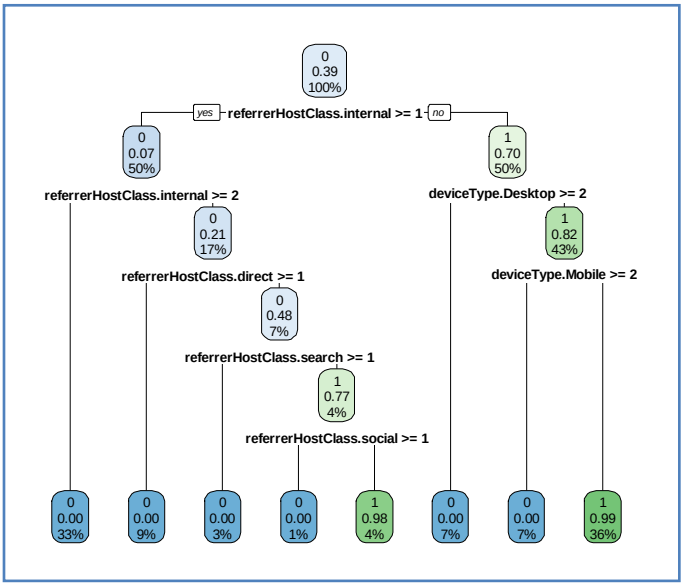
Como era esperable, el método de Naive Bayes fue el de menor precisión.

Un árbol de clasificación mejoró la precisión llegando a un 99.6%. Al analizar la importancia de cada variable en la decisión del árbol se obtiene que la variable que

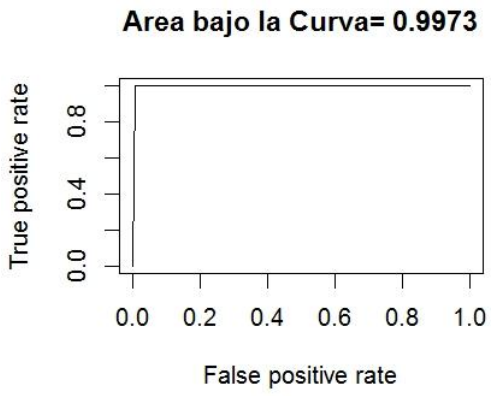
inicia la clasificación es el Host de tipo Interno, o sea, cuando el individuo proviene de uno de los sitios de la empresa.

A partir de allí separa el dispositivo de acceso, según sea Desktop o no. Le siguen en importancia, la procedencia por una búsqueda directa, orgánica y por una red social.

Árbol de Clasificación:

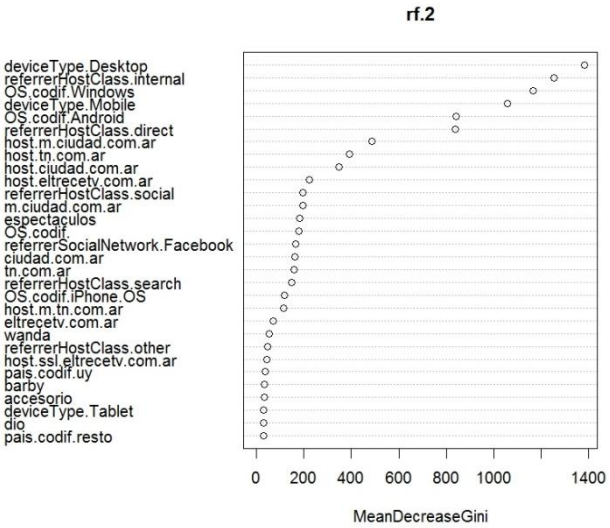


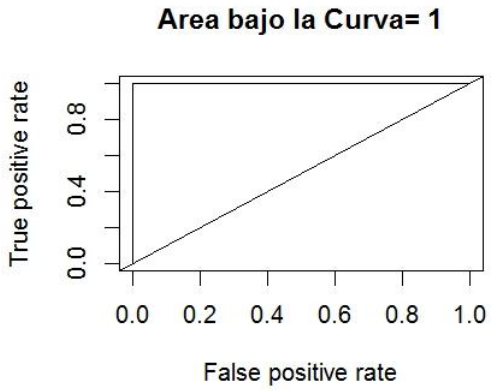
Curva ROC para el árbol de decisión:



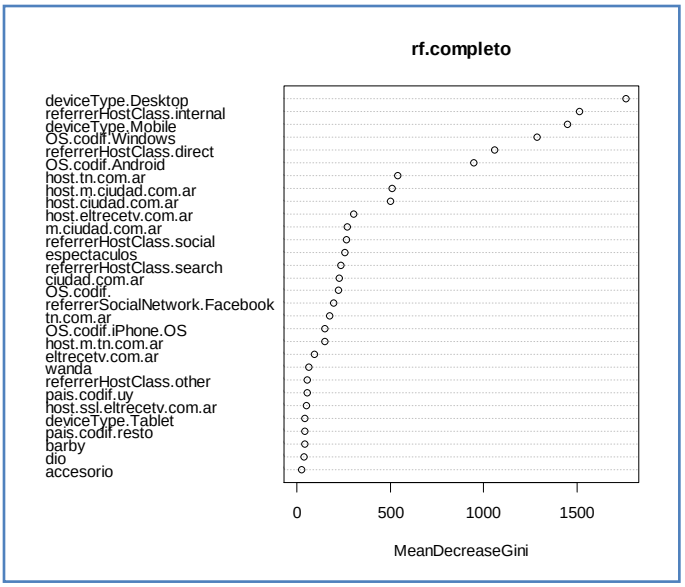
El modelo aplicando Random Forest mejoró aún más la precisión, seleccionando luego éste para el modelo final utilizando todos los datos. En los siguientes gráficos

se observa la importancia de los atributos al utilizar este modelo y la curva ROC para la predicción.





Para el modelo a implementarse utilizando todos los datos, se obtiene el siguiente diagrama de importancia de las variables:



4 Propuesta -Recomendaciones

El objetivo integral de este trabajo fue presentar una evaluación del estado de la empresa en el camino de transformación en una compañía "data driven" y en base a ello, generar una propuesta de avance, al mismo tiempo que se desarrolló uno de los tantos procedimientos basados en Machine Learning que podrán implementarse para mejorar el negocio.

En este apartado se detalla una propuesta concreta en base al relevamiento realizado y el análisis de datos llevado a cabo.

4.1 Sobre el avance en la transformación

En miras de elevar el nivel de la empresa en el proceso de conversión hacia una empresa data-driven, el primer objetivo debe ser eliminar lo que queda de aspiracional. Luego avanzar sobre la potencialidad que tiene la compañía en el manejo de datos para conformar un sistema de integración, administración y uso de los mismos.

Se presentan a continuación las recomendaciones y acciones propuestas para estos objetivos.

➡ Un punto inicial, y a la vez determinante, consiste en lograr un convencimiento al nivel más alto de jerarquía de la dirección, que la analítica de datos propios elevará el nivel del negocio y mejorará su rendimiento. Para ello es necesario presentar casos de

empresas de medios que se encuentren más avanzadas en este proceso y que muestren mejoras sustanciales. Sin un convencimiento al nivel de la alta gerencia, no será posible enfrentar la transformación.

➡ Es necesario que surja de la capa más alta de gerenciamiento la directriz de que las decisiones a futuro deben estar avaladas por un análisis inteligente de los datos y empezar a abandonar los criterios propios basados en la intuición o en la opinión de personas influyentes ("Hippo"-Highest Paid Person Opinion). *La capacidad de quienes toman las decisiones debe reflejarse en el entendimiento de la información provista por los datos.*

➡ Se debe poner foco en la extracción de datos que son útiles para la compañía: no es suficiente entender los datos con que se cuenta, sino que es necesario *saber por qué y para qué serían utilizados y qué beneficio traería a la organización*. Esto necesita de un espacio de discusión inter-áreas y un listado de los objetivos con un nivel de prioridades. No debe dejarse de lado la incorporación de fuentes de datos que, aunque en una primera instancia no vayan a ser explotadas, puedan significar una mejora futura.

➡ Se deben explorar fuentes externas que puedan contribuir al entendimiento de la data propia (por ejemplo, Trending Topics en Twitter o en Google).

➡ Es un punto fundamental centralizar en un área específica las decisiones sobre la arquitectura de una base de datos integrada, y su gobernanza. *Es fundamental la existencia de un programa de Master Data Management (MDM) con un responsable que tenga el rol de Chief Data Officer*. Se propone la organización de las siguientes entidades para su organización:

- IM (information Management): donde se toman las decisiones sobre la información a extraer, y las bases a integrar. Involucra la adquisición o creación de datos, el acceso, la seguridad, el almacenamiento, su resguardo y mantenimiento. También se ocupa de la integridad, consistencia, disponibilidad, y precisión de la información. Este área sería la responsable de generar un repositorio de Metada que almacena información sobre definiciones en las bases de datos, el uso, cambio, referencias. Es un área netamente tecnológica, sin soporte del negocio en sí. Es atribución del CDO proveer ese enfoque para evitar esfuerzos innecesarios y sólo

enfocar en lo que realmente genere valor. Este área debe estar integrada por profesionales del área de Sistemas.

- Arquitectura: con la información provista por IM debe establecerse una arquitectura de datos adecuada para que los datos puedan ser utilizados por el área de BI y de Analítica que son los encargados de extraer información de los datos y generar valor para el negocio. Este área debe estar integrada en la parte operativa por un profesional del área de Arquitectura de Bases de Datos.

- CDO (Chief Data Officer): responsable de la administración de los sistemas anteriores con un punto de vista estratégico y de planificación. Identifica los datos que interesan a la compañía y las necesidades tecnológicas para su incorporación o extracción y uso. Es la persona que gestiona servicios internos y externos. Tiene conocimiento del negocio pero también tiene capacidad de analítica de datos. Dentro de la gobernanza de datos se ocupa de la estrategia relacionada con los datos estableciendo políticas para la explotación efectiva de los datos y su uso dentro de la empresa. Impulsa estrategias de flujo y análisis de la información relacionadas de manera directa con objetivos del negocio. Es el responsable de integrar áreas para transmitir las necesidades a las áreas de IM y Arquitectura.

Para dar inicio a un proceso formal, y tomando como base el relevamiento realizado sobre las bases de datos que pueden integrarse, se sugiere encarar el proyecto de integración y uso de datos utilizando una metodología ágil del tipo SCRUM. Son muchas las aplicaciones que pueden implementarse en la empresa Medios para generar valor a partir de los datos. Cada una de ellas puede trabajarse como un sprint de un proceso SCRUM. Para cada aplicación puede ser necesario generar integraciones de varias fuentes que pueden realizarse en forma simultánea y no secuencial, por lo que una metodología del tipo ágil resulta recomendable. Además cada entregable al final del sprint puede generar una retroalimentación para mejorar el proceso, pero el entregable puede ponerse en práctica.

4.2 Sobre la base de datos

De todas las bases de datos que se relevaron en la empresa, sin duda los datos con mayor granularidad y que contienen un Id de usuario son los almacenados en la nube de

Google Cloud. En particular el Id de usuario permite seguir a los usuarios a través de los sitios, de manera anónima, o sea sin identificación particular del mismo, sino a través de una cookie, permitiendo conocer el "user journey". Se recomienda programar la extracción periódica de estos datos y almacenarla en un repositorio del tipo Data Lake.

En el mismo repositorio debe almacenarse la data extraída del sistema Gigya que contiene los usuarios registrados ya que contiene el mismo Id y permite matchear los datos a través de dicho Id que actúa como una foreign key.

Respecto a los datos provenientes de YouTube y Facebook se sugiere generar también consultas programadas. Estas consultas generan archivos de tipo csv que pueden direccionarse para ser almacenadas en una base de datos del mismo repositorio anterior.

Todos los datos almacenados en el Data Lake son aptos para ser procesados con R o con Phyton aplicando procesos de data mining y de Machine Learning. En particular se sugiere poner a discusión la factibilidad de adquirir herramientas de ML que procesan los datos a través de plataformas específicas como podrían ser las comercializadas por Google, Microsoft o IBM.

Si bien no es posible identificar un usuario de los sitios que también resulta usuario de YouTube o Facebook se propone aplicar procesos de clusterización independientes a partir de cada base para armar estrategias de contenido y de publicidad basada en similitud de perfiles.

Dado que los sitios de la empresa ofrecen información, no sólo de actualidad política, económica y social, sino también de deportes y espectáculos, se sugiere integrar en el data lake el resultado de un scrapping periódico sobre redes que permita obtener los temas de conversación social vigentes.

Para el proceso de armado del data lake es imprescindible un primer paso de analizar las métricas y dimensiones disponibles para construir cada base de datos extrayendo sólo aquellos datos que serán útiles para la empresa y estableciendo objetivos concretos. La consultas a varias de las bases externas tienen un costo dependiente de los parámetros de consulta. Se propone la construcción de un set de preguntas a responder o de acciones a definir, que resulten relevantes para el negocio, ya sea para la mejora del contenido ofrecido o bien para la mejora o ampliación del espacio publicitario a ofrecer.

4.3 Sobre el uso responsable de los datos

Tufekci ha expresado en su trabajo *Algorithmic harms beyond Facebook and Google:emergent challenges of computational agency* (2015) que el uso de algoritmos que toman decisiones generan las mismas preocupaciones éticas que surgen sobre las decisiones humanas: cuestiones de transparencia, de discriminación, de responsabilidad.

En efecto, las decisiones tomadas a partir del uso de algoritmos pueden llegar a generar una potencial discriminación por raza, por género, por incapacidad física, por tamaño de la familia, o cualquier otra razón que podrían incluso estar infringiendo leyes concernientes a ello.

Un ejemplo que referencia el autor es la aplicación de algoritmos en la selección de empleados basado en el criterio de "la distancia hasta el trabajo". Dada la segregación residencial existente en muchas ciudades, contratar gente basado en este criterio podría estar consecuentemente generando discriminación racial.

El uso de los algoritmos se extiende cada vez más sobre todo tipo de áreas, tomándose decisiones que pueden incluso revelar aspectos privados de los individuos que ellos mismos no quieren revelar tales como preferencias sexuales. La existencia y uso de los dispositivos conectados, como los celulares inteligentes, potencian y expanden aún más la capacidad de estas prácticas.

En el caso de la empresa referenciada en este trabajo, los datos sería utilizados para ofrecer contenido y publicidad personalizada. Cabría en este caso varios puntos a reflexionar:

- ¿hasta qué punto se debe personalizar los contenidos ofrecidos y cuál sería un punto de equilibrio para evitar que los individuos queden encerrados en un mismo ecosistema que les resulte confortable y no puedan acceder a nuevas ideas?
- ¿la personalización no estaría potenciando actitudes o pensamientos que pudieran resultar negativos para el individuo y/o la sociedad?
- ¿la publicidad segmentada no está coartando la oportunidad de conocer nuevos productos?
- al no ofrecer ciertos productos a ciertos perfiles, ¿no está generando algún tipo de discriminación?

- ¿cabe alguna posibilidad de que ciertas publicidades ofrecidas generen acciones de los individuos que puedan resultar peligrosas para su salud?

4.4 Sobre la recurrencia de usuarios.

En este trabajo se analizó la construcción de un modelo para predecir la recurrencia de usuarios.

Los métodos aplicados resultaron con una buena performance, y por lo tanto el modelo puede aplicarse a nuevos datos, como también puede mejorarse utilizando datos más recientes.

De manera integrada con el modelo desarrollado en este trabajo para clasificación , debería desarrollarse un modelo de recomendación con el objetivo de generar de generar interés en el potencial usuario No Recurrente para poder transformarlo en Recurrente.

El caso de clasificación de usuarios presentado como un caso de uso de Machine Learning en una empresa de medios es un ejemplo entre muchos otros casos de desarrollo que pueden proponerse para generar valor en los datos ya sea para el área comercial como para el área de contenidos. En particular para el área comercial se propone avanzar en desarrollos enfocados a la optimización de la pauta publicitaria en digital y a la pre evaluación de pautas. Para ello es necesario contar con datos históricos, que en caso que no existieran, deben comenzar a ser generados de inmediato.

No es para descartarse tampoco que un proceso de Machine Learning pudiera contribuir también a mejorar la distribución de la pauta en la televisión lineal, aún cuando ésta permanece direccionada por procedimientos más tradicionales.

Conclusión y trabajo a futuro

Este trabajo se ha basado en el estudio de un caso con el objetivo de realizar un aporte constructivo en el camino de transformación de una empresa en una compañía "data-driven".

Para ello, en el apartado 3 se ha analizado el estado actual de la empresa en el uso de datos, como también las potenciales bases de datos existentes.

En el apartado 4 se han realizado recomendaciones para avanzar en el camino de una mejora en el uso inteligente de datos.

También se ha tomado un caso particular de uso y se han aplicado procedimientos de data mining y machine learning a una base de datos específica como ejemplo de acciones posibles, y con el objetivo de presentar un caso que incentive el proceso de digitalización.

Como resultado de la evaluación del estado actual de la empresa se ha concluido que, si bien la empresa cuenta aún con alguna característica propia de una empresa Aspiracional, puede decirse que por su situación actual cumple con todas las condiciones para ser clasificada como empresa Experimentada de acuerdo a la clasificación propuesta por el MIT-SLOAN.

En lo referente a la organización interna y a la generación y almacenamiento de bases de datos, en este trabajo se han propuesto acciones concretas para avanzar hacia un nivel más alto en miras de convertirse en una empresa data-driven. Más aún, se han listado los grupos de trabajo imprescindibles para una efectiva puesta en práctica.

Si bien se han realizado las recomendaciones para encarar un proceso de mejorar el nivel de la empresa en el camino de la toma de decisiones basada en datos, también se ha sugerido tomar una actitud crítica sobre el uso de los datos, reflexionando sobre cuestiones éticas del uso de los mismos, y realizando un análisis que permita asegurar que no caer en actitudes de segregación o discriminación en el contenido ofrecido, como tampoco en la incitación a actividades que pudieran generar algún daño físico o psicológico con la publicidad ofrecida.

Para el análisis del caso de uso, si bien ha sido posible obtener un modelo con una buena tasa de clasificación, el mismo sería más rico y preciso si pudieran incorporarse más atributos de los usuarios para lo cual serán necesarias acciones concretas orientadas a la recolección de los mismos.

Como conclusión integral puede afirmarse que la empresa Medios se encuentra en un camino ya iniciado hacia la digitalización y cuenta con potencial suficiente para poder avanzar en un tiempo inminente. En este trabajo se han detallado criterios y pasos para encarar ese avance.

El caso de datos que se presentó, tuvo el objetivo de clasificar un usuario como potencial usuario recurrente o no recurrente. Como acción futura, el procedimiento de clasificación puede mejorarse con la incorporación del relevamiento de datos adicionales sobre los usuarios de los sitios y con el desarrollo de un proceso de recomendación, basado en los usuarios recurrentes, que pueda llegar a cambiar el estado de no recurrente a recurrente. Sin embargo el caso aquí presentado relativo a la recurrencia de un usuario es sólo un ejemplo de varios casos de aplicación que pueden implementarse.

Referencias bibliográficas

- Acosta, V. (2019). Descubre la principal diferencia entre Data Mart y Data Warehouse. *Revista Digital INESEM* .
- Boris Evelson-Forrester. (2018). Transform Your Organization From Data-Driven To Insights-Driven.
- Bowen, R. A. (2014). Developing an enterprisewide data strategy. *Healthcare financial management* .
- CISCO. (2019). Cisco Visual Networking Index:.
- CISCO. (2019). Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper.
- Clarke, N. (2019). How to ensure provision of accurate data. *Journal of Securities Operations & Custody* .
- Del Rosso, R. (s.f.). Taller de Integración Final-Material de Clase.
- E.Brewer. (2000). Towards Robust Distributed Systems.
- Ernesto, C. (2019). SQL y Preparación de datos. *Notas de Clases* .
- Fernandez Blanco, C. (215). Big data-El dato en un rol estratégico.
- Fernandez-Manzano, N. C. (2016). Data Management in Audiovisual Business: Netlix as a case study. *elprofesional de la informacion* .
- Gareth James, D. W. (2013). An Introduction to Statistical Learning with applications in R. *Spinger* .
- George, H. (2014). Big Data and Management. *Academy of Management Journal* , 321-326.
- Gilbert, L. (2002). “Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News* .
- Howard, S. . (2016). Automation, Big Data, and Politics: A research Review.
- *https://docs.rapidminer.com/9.1/studio/operators/modeling/feature_weights/weight_by_forest.html*. (s.f.).
- *https://docs.rapidminer.com/9.1/studio/operators/modeling/feature_weights/weight_by_forest.html*. (s.f.).

- *https://planificacionmedios.com/2013/02/17/cobertura-contactos-frecuencia-y-grps/*. (s.f.).
- *https://www.40defiebre.com/que-es/engagement*. (s.f.).
- *https://www.4webs.es/blog/diferencias-usuarios-nuevos-recurrentes-ecommerce*. (s.f.).
- IAB-Glosario. (2016). *http://www.iabargentina.com.ar/metricas.php.*
- Lavalle, L. S. (2011). Big Data, Analytics and the path from Insights to Value.
- Lehrer, W. V. (2018). How Big Data Analytics Enalbes Service Innovation: Materiality, Affordance and the Individualization of Service. *Journal of Management Information Systems* , 424-460.
- McAfee, B. (2012). Big Data: The management revolution. *Harvard Business Review* , 1-9.
- Shmueli, B. Y. (2018). *Data Mining for Business Analytics.* Wiley.
- TUFEKCI. (2015). Algorithmic harms beyond Facebook and Google:emergent challenges of computational agency.
- Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergente challenges of computational agency.

Anexo

Glosario de términos específicos

Los datos utilizados en este trabajo utilizan algunas variables asociadas específicamente a la audiencia web. A continuación se presenta la definición de los términos principales según el InteractiveAdvertising Bureau de Argentina (IAB-Glosario, 2016).

- *Website* (Sitio Web): conjunto de páginas interrelacionadas con un objetivo en común e identificadas con un tipo de contenido, organización o empresa.
- *Domain name* (Nombre del dominio): nombre a través del cual se puede acceder a un sitio web. Por ejemplo: "sitio.com".
- *Page* (Página): es lo que surge en la pantalla del navegador al acceder a una dirección en Internet.
- *UniqueVisitors / UniqueUsers* (Visitantes únicos / Usuarios Únicos): cantidad de individuos que navegaron al menos una vez por un sitio o un conjunto de sitios dado, en un período de tiempo determinado.
- *Visits/Sessions* (Visitas / sesiones): cada una de las visitas que realiza un usuario al sitio web. Generalmente la visita se considera finalizada cuando el usuario abandona el sitio o cuando transcurren 30 minutos de inactividad.
- *Page views*(Páginas vistas): cantidad de veces que fueron visualizadas las páginas de un sitio web en un período dado.