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

Clickstream y Sistema Financiero. Innovaciones en las Entidades Bancarias

AUTOR: FEDERICO GARCIA BLANCO

Estructura

Resumen

Introducción

Utilización de Datos Privados a Nivel Corporativo

- Conceptos de Ciencia de Datos
- Aspectos Legales y Privacidad
- Gobernanza en Base de Datos

Metodología

- Fuente de Datos y sus Atributos
- Integración y Preparación
- Algoritmo de Predicción.

Modelo y Procesamiento

- Desempeño del Modelo
- Desarrollo de Código de Programación
- Potencialidad

Conclusión

Resumen

El concepto de Big Data está en auge en la actualidad dado el alto grado de impacto que genera en la población a través de sus aplicaciones prácticas que pueden generar valor agregado en los productos y servicios que se consumen masivamente. En este trabajo, se va a analizar el comportamiento web de los usuarios de un banco en particular para poder predecir el comportamiento de los clientes a futuro. A través de mecanismos de aprendizaje automatizado se pueden aplicar algoritmos matemáticos que permiten evaluar el comportamiento de las personas mediante el análisis de datos.

Palabras Clave

Base de datos; Sistema financiero; Aprendizaje automatizado (ML); Clickstream; Big Data

Introducción

En el trabajo presentado se analizan múltiples secciones de la ciencia de datos aplicando el caso particular del Clickstream bancario para la predicción de conversiones de clientes de una institución. El trabajo sostiene un modelo de predicción basado en algoritmos matemáticos capaces de generar dicha predicción a través del análisis de grandes volúmenes de datos. Para esto, se decide trabajar con lenguajes de programación acordes a la cantidad de datos y a la posibilidad de desarrollo de algoritmos individuales.

Mas allá de la parte técnica que se profundiza con intensidad en los apartados, se propone problematizar sobre aspectos controversiales del rubro como lo son la privacidad de los datos y las estructuras acordes para la presentación de datos pertinente. De esta forma no solo se establecen tratamientos técnicos de programación y estadística, sino también se focaliza sobre ciertos aspectos que son de interés en la actualidad dado el alcance del rubro en una amplia gama de sectores de la vida cotidiana de las personas. El alto grado de avance tecnológico en las poblaciones en su conjunto con la penetración de nuevas herramientas financieras, dan lugar a problematizar sobre estas cuestiones que aún hoy en día no son del todo discutidas.

Con respecto a la estructura de la base de datos, se van a identificar las buenas prácticas a tener en cuenta para la entrega de información de calidad para su posterior análisis. No es lo mismo hacer pública una base de datos bajo ciertos estándares de calidad hacia la comunidad científica de datos que brindar la misma base sin tener en cuenta ciertas cuestiones que aseguran calidad e integridad a una base de datos.

Por otro lado, se problematiza sobre la privacidad de los datos. Este concepto es muy discutido en la actualidad dado que con datos de calidad se han logrado desarrollar herramientas que impactan en lo más profundo de las sociedades como puede ser la elección de sus gobernantes o la discriminación de ciertos sectores de la población entre otras cosas. Dado estos casos, es necesario profundizar sobre esta problemática que tanto está afectando en la actualidad y de la cual no se sabe mucho al respecto por el hecho de que recién se discutiendo en estos días. Si bien puede que se trate de un blindaje mediático sobre la privacidad y uso de los datos por parte de las grandes corporaciones, lo cierto es que como analista es importante tratar el tema para que tanto las generaciones actuales como las que vienen tengan herramientas para iniciar estos debates a edades tempranas del análisis de datos.

Por último, se va a profundizar sobre el desarrollo técnico del modelo de predicción aplicado indicando cuestiones de interés como pueden ser los problemas resueltos en el camino cursado, definición de conceptos alternativos como métricas de error y la explicación de los algoritmos elegidos por el analista. Si bien muchas de estas decisiones son subjetivas, se va a indicar pertinentemente el motivo por el cual se decidió optar por cada una de las variantes para dar lugar a la interpretación por parte del lector.

Apartado 1: Utilización de Datos Privados a Nivel Corporativo

En la actualidad, la posibilidad de tener acceso a grandes volúmenes de datos de los clientes de una organización permite que las autoridades de las compañías puedan hacer uso de dichos datos para identificar los gustos y preferencias de sus clientes. Este proceso en años anteriores no era posible por dos motivos particulares: por un lado, las empresas no captaban la misma cantidad de clientes que pueden captar en la actualidad, y por otro lado eran incapaces de almacenar y procesar grandes volúmenes de datos.

En el particular caso de los bancos, la inclusión financiera de los últimos años permitió que el sistema bancario sea cada vez más amplio y tenga un alcance global tan importante que brindó la posibilidad de que muchas personas comiencen a estar bancarizadas (Tuesta, D., Sorensen, G., Haring, A., & Cámara, N. (2015)). Esta tendencia se ve impactada aún más por las recientes Fintech que se incorporan cada día en el mercado y generan más posibilidades de inclusión financiera aún.

Por otro lado, para resolver la problemática del almacenamiento y análisis de datos, surgieron grandes avances de equipos tecnológicos y desarrollos intelectuales que permitieron a las organizaciones comenzar a gestionar cada vez más grandes volúmenes de datos. El avance de los últimos años que tuvo el hardware de las computadoras y el desarrollo de carreras de análisis de datos, hicieron que la manipulación de la gran variedad de datos con la que cuenta una empresa deje de ser un problema por solucionar.

Consecuentemente, para hacer una medición del comportamiento web de los usuarios de un banco, por ejemplo, se deben almacenar los clicks que los usuarios de la plataforma realicen. Esto es un mecanismo que se está utilizando mucho en la actualidad ya que permite a los bancos personalizar los comportamientos de cada cliente y poder así ofrecer productos y servicios de forma diferencial para maximizar sus retornos y optimizar el uso de capital.

Una de las características fundamentales que implica el concepto de Big Data (Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013)) se refiere a la veracidad de los datos con la que se desea trabajar. Este proceso se ve garantizado si existe un mecanismo de gobernanza de datos tal que permita evitar errores en la carga y manipulación de la base de datos manteniendo ciertos procesos que ayuden a la gestión de esta y regulen quienes toman las decisiones. Es de tal importancia este proceso que las compañías dedican grandes recursos a este sector dado que es preferible tener precauciones a la hora de manipular datos en vez de generar métricas erróneas por un dato mal ingresado o manipulado.

Apartado 1.1 Conceptos de Ciencia de Datos

Actualmente, en las organizaciones se utilizan una gran variedad de términos que refieren a conceptos muy específicos dentro de lo que se llama Ciencia de Datos. Si bien mucha de esta terminología es nueva y hasta puede parecer interesante desde el punto de vista comercial, se ha notado que en el último tiempo los nombres con los que se denomina cada

acción o proceso dentro de la temática son cada vez más siendo que se trata del mismo concepto. Dado esto, se tiende a confundir ciertos conceptos que deberán estar esclarecidos al momento en encarar un trabajo de esta magnitud.

Particularmente, este apartado, se propone definir los conceptos que van a ser necesarios tener en cuenta para poder comprender la totalidad del trabajo en cuestión. Inicialmente se define a la ciencia de datos (Provost, F., & Fawcett, T. (2013)) como un conjunto de principios fundamentales que apoyan y guían la extracción de información y conocimiento a partir del análisis de datos. A pesar de que la definición pueda ser muy amplia, justamente de eso se trata el término dado que el campo de trabajo de la ciencia de datos puede ser muy abarcativo.

Con respeto a la base de datos, se trata de un concepto que refiere al conjunto de información que surge en una misma organización y que puede estar destinada a variedad de sectores dentro de la compañía. A su vez, existe el concepto de Datawarehouse (Hüsemann, B., Lechtenbörger, J., & Vossen, G. (2000)) que se entiende como un recolector de información integrado y dinámico que utiliza procesamientos analíticos en línea para la toma de decisiones. Dentro del Datawarehouse, los usuarios de la base de datos podrían ser capaces de consultar cualquier tipo de información que respecte a la empresa si los permisos internos que tiene se lo permiten.

Para definir un proceso de Aprendizaje Automatizado o Machine Learning según sus siglas en inglés, se debe tener en cuenta el concepto de algoritmo matemático (Bravo, J. A. F. (2005)) que refiere a una “secuencia de pasos operativos para la realización de una tarea o la resolución de un problema”. Dicho esto, un proceso de aprendizaje automatizado (Alpaydin, E. (2009)) es aquel que, dado una cantidad de datos, puede generar una función que aprenda de estos para luego aplicar el conocimiento obtenido con el fin de realizar una acción específica. Durante el proceso de aprendizaje de los datos, el algoritmo matemático cumple un rol esencial dado que va a ser el encargado de moldear la etapa de aplicación sobre datos nuevos. Justamente en esta etapa es primordial el uso de grandes volúmenes de datos para que el algoritmo aprenda de tal forma que permita generalizar sus decisiones dado que entrenó con una gran cantidad de observaciones.

El concepto de entrenamiento se inicia cuando se destina un porcentaje de la base de datos para aplicar la fase de aprendizaje utilizando el algoritmo deseado. Este proceso de partición de la base de datos es de suma importancia porque evita un error muy común de

sobreajustamiento (Lever, J., Krzywinski, M., & Altman, N. (2016)) del modelo que implica el ajuste excesivo del mismo a los ruidos de la muestra tomada que puede ser fruto de razones biológicas o técnicas.

Como se mencionó anteriormente, la base de datos es aquella que acumula datos externos o internos de la organización. Uno de los tipos de datos que puede acumular puede ser el comportamiento en un sitio web sobre los usuarios. En este caso, el concepto determinado Clickstream (Hind, J. R., Nguyen, B. Q., & Peters, M. L. (2006)) sobre el cual se basa el presente trabajo, refiere justamente a esa temática dado que recopila todos los datos que puede sobre el accionar de los usuarios dentro de una plataforma web. Esto suele componerse de la información que genera el hecho de hacer clicks dentro de una plataforma comercial pudiendo obtener así datos como la fecha, la hora, el tipo de usuario comercial, entre otros atributos. Por lo tanto, el concepto de Clickstream está puro y exclusivamente dedicado a herramientas informáticas donde existe de un lado un usuario de la plataforma y por el otro lado una base de datos capaz de almacenar grandes volúmenes de datos.

Otro concepto novedoso que se menciona durante el trabajo es el Big Data. Este término que gradualmente fue tomado por las organizaciones puede tener un tinte comercial en mayor medida que el resto de los conceptos previamente presentados. Si bien se puede encontrar numerosas referencias bibliográficas que tratan del tema, el Big Data (George, G., Haas, M. R., & Pentland, A. (2014)) refiere a la capacidad de generar datos a partir de una pluralidad creciente de fuentes, incluidos Clickstream, transacciones móviles, contenido generado por usuario, redes sociales, así como también contenido generado a propósito a través de redes de sensores o transacciones comerciales. Algunos de los casos de uso más comunes son aquellos donde compañías multinacionales utilizan datos para generar políticas internas con el fin de optimizar recursos y maximizar retornos. Entre esos datos, suelen recopilar datos geolocalizados, imágenes, sonidos, textos, y variedad de datos provenientes del internet de las cosas que en su mayoría son dispositivos de sensibilidad aplicados a ciertos elementos de necesidad. Es por esto, que se decidió utilizar el término Big Data refiriéndose a aquel proceso de manipulación de una gran cantidad y variedad de datos. A pesar de esto, existen muchas opiniones encontradas de personas idóneas en el tema ya que la definición del término no es absoluta, sino que tiene matices según la visión del profesional que aborde la temática.

Por último, se van a definir ciertos conceptos que están puramente relacionados con el sistema financiero y sus derivados. El sistema financiero en sí es aquel que reúne al conjunto de entidades financieras que operan en el mercado. En este grupo de compañías se suelen destacar a los bancos y aseguradas principalmente, pero también se deben incluir a las casas de cambio, caja de valores y las nuevas Fintech que están penetrando en el mercado tan rápidamente.

Con respecto a las Fintech, (Vives, L. (2015)) son aquellas nuevas compañías con objetivos disruptivos que aplican la tecnología al sector financiero brindando algún tipo de novedad a sus clientes. Si bien a priori no se podría demostrar si el surgimiento de estas nuevas empresas fue consecuencia de la falta de innovación en el sistema financiero formal, se puede notar actualmente en el mercado que estas nuevas empresas captan cada vez más público y obligan a aquellas organizaciones que pertenecen al sistema financiero tradicional desde hace muchos años a generar productos novedosos que impidan la perdida de la participación de mercado. Este concepto de Fintech es sumamente importante dado que se refiere a compañías que son nativas digitales y por lo tanto aplicaron el concepto de ciencia de datos a edades tempranas de su creación. En gran medida gracias a este tipo de emprendimientos la asignatura mencionada fue creciendo rápidamente y no parece dilucidarse un abandono de forma repentina de la actividad, sino todo lo contrario dado que tanto la capacidad de almacenar mayor cantidad de datos como la formación de nuevos profesionales hace que la materia evolucione permanentemente.

Apartado 1.2 Aspectos Legales y Privacidad

El uso y acceso permitido a una base de datos es un eje fundamental en cualquier análisis de datos donde su contenido provenga desde un tercero involucrado. Este concepto no se puede dejar de lado ya que actualmente es un tema muy controversial (Isaak, J., & Hanna, M. J. (2018)) por el gran aumento de casos donde se sospecha un uso indebido de las bases de datos de clientes.

Si bien se trata de un abordaje totalmente nuevo, dado los grandes avances en almacenamientos de datos por parte de las organizaciones, actualmente las compañías ya no pueden actuar pasivamente sobre el tema, sino que deben tomar cartas en el asunto y comenzar a perfilar cual va a ser el uso que se le va a dar a la base de datos con las que se

trabaja. Esto debe incluir también el permiso de las personas que integran dicha base de datos para que la compañía tenga las facultades de utilizar esos datos y algunos casos particulares poder comercializarlos.

Se menciona que es un aspecto totalmente novedoso en los análisis de datos, porque en el pasado talvez era impensado la posibilidad de que una empresa logre réditos económicos por el solo hecho de compartir datos. Los modelos de negocios a lo largo del tiempo fueron cambiando y en la actualidad las bases de datos componen una gran parte del patrimonio de las empresas.

Cuando los Estados y los organismos de control a través de las instituciones estatales no hacen uso de sus facultades para controlar y regular el uso de las bases de datos, permiten que algunas compañías tengan grandes privilegios para manipular y tener libre acción ante los datos que generan o recopilan. La gravedad del asunto es tal porque detrás de toda la información que puede almacenar una organización, pueden existir datos sensibles de los ciudadanos que no deberían ser divulgados ni compartidos, y menos aún vendidos. Por ejemplo, en el caso de Estados Unidos y Europa (Esteve, A. (2017)) ya se viene trabajando sobre el tema y de a poco van generando regulaciones para poder mejorar diariamente sobre el uso que las empresas hacen de sus bases de datos. Este proceso de regulación tiene muy poca historia ya que la tecnología de las grandes bases de datos no hace mucho que se impregnó en el mercado, y por lo tanto seguramente estos procesos vayan mejorando a medida que pasa el tiempo y se tiene más conocimiento sobre el asunto.

En la actualidad, los usuarios finales de las plataformas transaccionales que suelen almacenar grandes volúmenes de datos no suelen tomar dimensión del nivel exposición al que se someten al participar en dichas plataformas. La cultura sobre la privacidad informática y el uso de derechos que las personas poseen sobre el asunto debería mejorar sustancialmente en conjunto con la regulación para poder crear un entorno de desarrollos informáticos más sustentables a lo largo del tiempo haciendo un correcto uso de los datos privados. Este proceso tomaría muchos años para llevarse a cabo dado que es necesario generar un cambio abrupto en la cultura informática de la sociedad explicando y haciendo tomar noción de la exposición a la cual cada uno de los habitantes se expone diariamente.

Existe el caso de algunos países o regiones que tuvieron una iniciativa muy fuerte de compartir las bases de datos que almacenan para que los ciudadanos puedan consultarlas y hasta sean capaces de desarrollar nuevas propuestas con su contenido. Suelen ser Estados

que están a la vanguardia al respecto, dado que problematizan y caracterizan sobre los asuntos recientemente mencionados. Esto brinda una gran ayuda para comenzar un proceso de concientización sobre los datos que las personas son capaces de generar periódicamente. Es de tal magnitud el alcance de estos nuevos avances tecnológicos que se estima que para el año 2020 cada persona en el mundo va a generar 1.7 megabytes de información (Muley, R. (2018)) por segundo.

En este aspecto, se puede problematizar sobre la situación legal en la que se encuentra la base de datos que es usada en este caso para realizar el trabajo, dado que se obtuvo desde un sitio web reconocido y habría sido otorgada por la compañía responsable de la información. Hay tres posiciones que se pueden analizar al respecto del uso de la base de datos. Por un lado, está la compañía que divulga los datos anonimizados de sus clientes con el fin de mejorar un proceso interno de la empresa a través de una competencia de modelos predictivos, por otro lado, están los analistas independientes que hacen uso de dicha base de datos para lograr modelos superadores que mejoren los procesos internos de la empresa en cuestión, y por último está el sitio web donde se divulga y se accede a la base de datos. Estos tres actores son los que están involucrados en el uso y tratamiento de una base de datos que contiene información privada del comportamiento de personas.

Como ya se mencionó anteriormente, existen leyes que regulan este tipo de situaciones para poder controlar que las personas no sufran ningún tipo de daño ni perjuicio mediante la divulgación de datos privados. Inicialmente, se debería saber si el banco tiene las facultades de almacenar datos privados de sus clientes con el finde lograr mejoras en los procesos internos de la institución. En segundo caso, se debería saber qué tipo de acuerdo tienen los clientes del banco para con la institución con respecto al uso de los datos que recopilan de cada uno. Tal vez existan clientes que no están de acuerdo con el mecanismo que utiliza el banco de exponer sus comportamientos de tal forma que sea accesible desde cualquier parte del mundo. Y, por último, se puede problematizar sobre qué ley aplica en cada caso dado que se trata de una situación donde el sitio web que distribuye la base de datos tiene su propia legislación, el banco que está radicado en cierta región también tiene su legislación, y por otro lado los analistas independientes que están distribuidos en todo el mundo deberían tener su propia legislación también. Al tratarse de una cuestión que implica la divulgación a nivel mundial de datos del comportamiento personal de ciertos individuos, no es fácil establecer una línea clara desde la cual se debe abordar legalmente la situación. Se trata de un caso

donde cada legislación puede tener las facultades de actuar, pero a la vez pueden solaparse las leyes de las distintas regiones.

En particular, tratándose de un caso situado en la República Argentina, a priori le correspondería a la entidad ser regulada por la Ley de Protección de Datos Personales N° 25326 sancionada el 04 de octubre de 2000 y promulgada parcialmente 26 días más tarde por el Senado y la Cámara de Diputados de la Nación Argentina. Además, en la Ciudad Autónoma de Buenos Aires existe la Ley de Protección de Datos Personales N° 1845 sancionada el día 03 de agosto de 2005 siendo uno de los pioneros en la región. Ambas leyes intentan prevenir externalidades negativas sufridas por los consumidores a partir del uso y la divulgación de datos personales de parte de las compañías. La sanción ante la solicitud de datos personales no pertinentes que puedan ser utilizados para fomentar la discriminación racial, religiosa o étnica, es parte de los artículos que están comprendidos en ambas leyes como así también el consentimiento por parte de los consumidores para el uso de los datos personales. Aun así, existen numerosos casos en la actualidad donde esos derechos se ven vulnerados y es necesario recurrir a acciones legales para tener soluciones en el corto plazo.

Apartado 1.3 Gobernanza en Base de Datos

En estos últimos años se está notando un cambio de paradigma con respecto a la toma de decisiones a nivel empresarial por parte de las autoridades de las organizaciones. Si bien hace muchos años las decisiones de negocio eran tomadas por aquellas personas que conocían ampliamente el contexto y tenían vasta experiencia en el rubro, en la actualidad eso parece estar cambiando porque se agrega un factor fundamental de la mano de la tecnología que son las bases de datos.

Como se ha mencionado anteriormente, los avances tecnológicos de almacenamiento y desarrollo intelectual para el análisis de datos está colaborando fuertemente para asentar este cambio de paradigma en los niveles de alta jerarquía de las corporaciones (McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012)). Esto permite que actualmente, un director ejecutivo de una empresa pueda basar sus decisiones de negocio en la experiencia que posee en el rubro, pero además puede tomar los datos que genera la compañía que preside para ser más preciso en sus decisiones. El hecho de poder apoyarse en

los datos permite que las personas que son encargadas de tomar decisiones puedan justificar de manera más simple el motivo por el cual decidieron tomar ese camino, reforzando la justificación además del mero hecho de tener experiencia en el rubro.

La problemática de este nuevo paradigma está en la confianza que las autoridades de una organización tienen sobre los datos que manejan. En este aspecto, es crucial tener una base de datos confiable que permita ser utilizada por la corporación para evitar tomar decisiones a partir de datos erróneamente manipulados. De este proceso se encarga la gobernanza de datos.

El concepto de gobernanza de datos refiere a un proceso que define un conjunto de reglas y responsabilidades para aquellas personas que manipulan los datos con el fin de lograr un buen uso de estos (Wende, K. (2007)). Para dimensionar esta definición, hay que tener en cuenta que la manipulación de datos en una organización debe ser sumamente restringida dado que solamente aquellas personas que tienen las facultades para hacerlo deberían tener permisos para acceder a la base de datos dejando afuera a todo personal ajeno a la carga de datos. Errores en este proceso que hace parte a la gobernanza de datos puede ser muy nocivo para la calidad del contenido de la base de datos.

Usualmente, en las grandes compañías, existen múltiples procesos en lo que respecta a la manipulación de datos. El acceso a la carga en la base de datos transaccional se suele dar en un sistema que se encuentra apartado de la base de datos desde donde se realizan consultas operativas. Este tipo de mecanismos mitiga claramente los riesgos que corre una compañía a la hora de afrontar posibles pérdidas de información o errores en los sistemas. Dado esto, resulta necesario mencionar la diferencia entre las diferentes etapas ya que no todos los procesos son afectados de la misma forma por la política de gobernanza de datos que posea la compañía. De igual manera, aquellos datos externos que son obtenidos de fuentes que no son propias, tienen otro tipo de reglas para verificar su calidad dado que sería erróneo tratar de la misma forma a datos con distintas fuentes de origen.

En los últimos años, en conjunto con el auge del almacenamiento de gran volumen de datos, las empresas comenzaron a destinar una gran cantidad de recursos para poder controlar la confianza y seguridad de sus bases de datos. Es necesario para mantener un orden en el sistema, tener en claro las personas que van a utilizar la base de datos y que rol va a tener cada uno en la manipulación, la carga, o la extracción.

Este mecanismo sería ideal en una organización siendo que cada uno de los integrantes funciona como un engranaje donde cada actor tiene su rol a cumplir y no debe involucrarse en el trabajo del resto de los actores. Esto no solo permite separar las funciones de cada uno de los trabajadores, sino también permite analizar qué equipo tuvo la falla en caso de encontrar alguna inconsistencia en la base de datos. Los permisos en este tipo de sistemas son fundamentales y deben ser aplicados lo antes posible para evitar errores en el futuro.

Si bien la mayoría de las bases de datos no comienzan a operar en busca de una buena gobernanza de datos, a medida que transcurre el tiempo las autoridades suelen notar ciertas inconsistencias en los datos que hace que comiencen a diagramar un sistema de perfeccionamiento de la base. Esto genera cambios estructurales dentro de la empresa, pero sobre todo requiere una gran dedicación por parte de las autoridades de alta jerarquía de hacer foco en el proceso que no solo tomará recursos, sino también absorberá tiempos y seguramente genere momentos tensos entre los diferentes sectores. Dentro de las organizaciones, algunos sectores que tal vez solo tengan contacto con los datos para realizar consultas o mecanizar un proceso en particular, pueden verse afectados si la compañía decide tomar el objetivo de mejorar su gobernanza de datos dado las molestias que esto genera.

Por otro lado, la gobernanza de datos tiende a generar un estándar de carga de datos, es decir, proponer y hacer cumplir un mecanismo de carga objetivo que no pueda ser vulnerado por aquellos sectores que solo desean cargar el dato sin tener noción del problema que pueden generar si el dato se carga de forma errónea. Establecer campos como obligatorios, articular límites en los campos, imponer formatos de ciertos datos, entre otros, son los mecanismos más utilizados a la hora de gestionar de buena forma la base de datos. Además de esto, el hecho de generar un estándar de exposición de la información ayuda al momento de querer anidar otro tipo de datos externos que puede ser relevante para la toma de decisiones. Con el auge del crecimiento de las bases de datos, algunas organizaciones, sobre todo estatales, han generado un estándar de presentación de información de forma tal que pueda ser fácilmente adicionada si se quiere trabajar con datos de otras regiones.

Además, una buena práctica para mantener una correcta gobernanza de datos es la prioridad de carga computacional evitando que las personas manualmente sean las encargadas de cargar el dato y de esta forma poder reducir los errores en la base de datos. A través de interfaz de programación de aplicaciones (API) se pueden cargar ciertos datos evitando errores humanos que suelen ocurrir en las organizaciones.

Particularmente, en el caso de estudio, se trata de una base de datos muy bien organizada con respecto a su diseño dado que evita la redundancia a través de un proceso de normalización donde se pueden notar variedad de tablas secundarias que aportan contenido a una tabla central que contiene sobre todo códigos poco intuitivos.

Se identifica fácilmente que se trata de una base de datos que no contiene valores nulos, no contiene valores extremos, tampoco contiene valores que carecen de significado. Esto haría destacar que la base de datos tiene recursos de gobernanza por detrás donde se especifica muy bien en la fuente que tipo de dato se tiene que cargar, y dado el contexto de la base de datos, supone un buen manejo de APIs para la carga.

Una vez que se cuenta con una base de datos completa y sin inconsistencias, las autoridades de las organizaciones pueden comenzar a usar los datos que recopilan sin tener temores sobre la confianza que les inspira su propia base. El desafío en todo caso sería encontrar un equipo adecuado que procese y analice los datos de forma tal que permita obtener conclusiones y métricas correctas para la toma de decisiones.

Apartado 2 Metodología

En este apartado se va a explicar en detalle el mecanismo que se utiliza para lograr predecir el comportamiento de los clientes a través de la recopilación de datos de los clicks dentro de una plataforma web. No solo se quiere dar a conocer el lenguaje que se utiliza y los algoritmos que fueron seleccionados para este proceso, sino también se mencionará aquellos inconvenientes que atravesó el analista para llevar a cabo su objetivo.

Luego de haber problematizado en la sección anterior sobre la privacidad de los datos y el uso que se le puede dar a información sensible, en este caso se va a explicar el origen de los datos y la composición de las tablas que se ven involucradas. Este proceso es importante ya que para todo proyecto donde se manipulen datos críticos, se debe realizar un cuestionamiento moral y ético con respecto al uso de estos.

El nivel de profundidad de análisis durante la explicación metodológica que se lleva a cabo puede resultar algo confuso dado que se requiere un nivel técnico básico del lenguaje de programación en cuestión para poder comprender la totalidad del proceso. A pesar de que muchos aspectos técnicos fueron evitados para no entorpecer la lectura del trabajo, puede

suceder que el lector requiera de más profundidad sobre algún tema en particular que no está siendo detallado suficientemente. En esos casos se recomienda tomar el apéndice del trabajo donde se puede evaluar con más detalle la totalidad del código de programación y sus comentarios. Aun así, desde el punto de vista del analista este proceso solo debe ser llevado a cabo por aquellas personas idóneas en el tema dado que puede resultar de más confusión aún en lectores que no se desenvuelvan correctamente en los lenguajes de programación utilizados.

Apartado 2.1 Fuente de Datos y sus Atributos

La base de datos con la que se va a desarrollar el trabajo fue obtenida desde una competencia de ciencia de datos que comenzó a mediados de Julio de 2019 por el transcurso de dos meses, siendo facilitada a través del Banco Galicia que volcó sus datos de Clickstream de sus clientes para analizar su comportamiento. El mecanismo por el cual se compartió dicha base de datos fue a través del sitio web de Kaggle donde se almacenan la mayor cantidad de competencias de este estilo a nivel mundial.

Este proceso de creación de competencias a nivel internacional a través de plataformas como la recientemente mencionada, está ganando mucho terreno en las soluciones que las compañías requieren para hacer uso de sus datos y poder así sacar conclusiones a partir de estos. Algunas empresas u organismos de renombre internacional ya han hecho el mismo proceso obteniendo modelos generados por científicos de datos independientes a lo largo de todo el mundo. Usualmente utilizan premios para incentivar a los integrantes de la competición a pesar de que la página además vela por el avance del conocimiento exponiendo los códigos utilizados de los usuarios que así lo deseen.

Para este caso en particular, se utilizaron los datos del Clickstream de los usuarios del Banco Galicia del año 2018 para predecir comportamientos del año 2019. Este nuevo concepto, consta de captar digitalmente mediante un registro los movimientos que los usuarios realizan dentro de la página web de la institución o a través de su dispositivo móvil. Mediante este mecanismo, la institución puede recolectar mucha información que podría ser relevante para su análisis y la toma de decisiones a nivel jerárquico.

La base de datos en este caso contiene múltiples tablas donde cada una brinda información sobre ciertos atributos que se pueden dar en cada página donde el usuario presiona el click,

es decir, va a mencionar que tipo de cliente es según la sección donde ingresó, que tipo de acción está queriendo hacer según la página donde ingresó, entre otras cosas.
A continuación, se van a detallar las tablas mencionadas y los conceptos de sus atributos tal cual son expresados en el sitio web de la competencia:

Pagesviews.csv:
USER_ID - Id del usuario
FEC_EVENT - Fecha
PAGE - código de la página
CONTENT_CATEGORY - Categoría de la página
CONTENT_CATEGORY_TOP - Categoría de más alto rango para la página
CONTENT_CATEGORY_BOTTOM - Categoría de más bajo rango para la página
SITE_ID - Id del sitio visitado
ON_SITE_SEARCH_TERM - Palabra clave buscada

Device_data.csv:
USER_ID - Id del usuario
FEC_EVENT - Fecha
CONNECTION_SPEED - Categoría de conexión
IS_MOBILE_DEVICE - Categoría de si es teléfono móvil o no

sampleSubmission.csv: es un ejemplo de entrega correcta

conversiones.csv:
USER_ID - Id del usuario
ANIO - Año
MES - Mes
CONVERSIONES – Cantidad de conversiones

Archivos con descripciones textuales de variables categóricas:

CONTENT_CATEGORY.csv

CONTENT_CATEGORY_TOP.csv

CONTENT_CATEGORY_BOTTOM.csv

SITE_ID.csv

PAGE.csv

En este contexto, las conversiones realizadas por el usuario se refieren a la cantidad de préstamos que tomó a través de la página web del banco o desde su dispositivo móvil mediante la aplicación. Justamente, este concepto es la clase que se desea predecir y el motivo fundamental por el cual se creó la competencia que arrojará múltiples modelos.

En los archivos con descripciones textuales de variables categóricas se pueden encontrar múltiples ejemplos para tomar noción a qué término se refiere cada variable. En ese sentido funciona como los metadatos de la base de datos donde se explica a que corresponde cada uno de los atributos.

En el caso del atributo PAGE, se refiere al código de la página que se está visitando. Dado que se trata de una plataforma web, existen múltiples paginas resultantes de la navegación dentro de esta. En total son 1835 códigos que representan a cada opción. Como ejemplo se puede mencionar el código 39 que es la página de inicio cuando se desea ingresar a la opción de inversiones dentro de la plataforma del banco.

Por otro lado, se encuentra la variable binaria ON_SITE_SEARCH_TERM que refiere a si la persona realizó una búsqueda rápida en el buscador que ofrece la plataforma.

También los SITE_ID vinculados con el código del sitio que se está visitando. El Banco Galicia de forma interna identifica las páginas visitadas en 4 clasificaciones nominales que son BANCOGALICIAPROD, MINORISTA, PRUEBANET, o HACETEGALICIA.

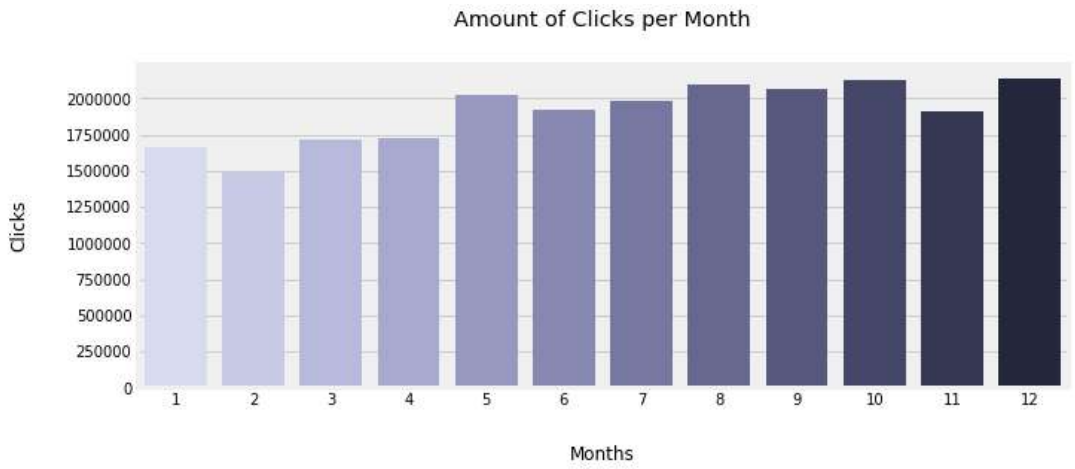
Por último, vale la pena destacar las variables CONTENT_CATEGORY y sus respectivos extremos. El archivo CONTENT_CATEGORY refiere a la categoría de la página que se está visitando. Para ejemplificar se puede nombrar el código 42 que se identifica con el metadato BANCA:ONLINE:WEB:EMPRESAS. Mientras que la categoría superior de la página está contemplada en la variable CONTENT_CATEGORY_TOP. Se puede

mencionar como ejemplo el código 9 que corresponde a la descripción BANCA:ONLINE:WEB:MOV que indica que se trata de un usuario que está navegando en el servicio MOVE de Banco Galicia. Y finalmente CONTENT_CATEGORY_BOTTOM indica una clasificación de la página un poco más precisa donde se puede poner como ejemplo la descripción :BANCA:ONLINE:WEB:RRHH codificada bajo el número 59.

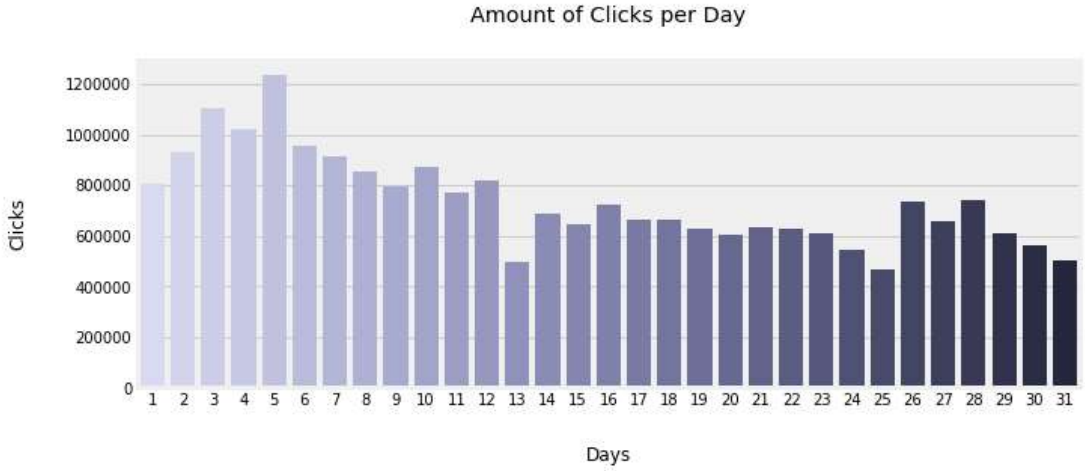
Si bien estas clasificaciones no son muy intuitivas, es la forma que tiene la entidad para identificar los diferentes comportamientos de los usuarios dentro de la plataforma web. Una vez que se investiga un poco la base de datos, el proceso puede más engorroso aún porque se va a notar mucha correlatividad entre las variables lo cual podría dar lugar a un proceso de componentes principales para reducir la dimensionalidad de la base de datos con la que se trabaja.

Luego de haber explicado los atributos que contiene la base de datos, se propone identificar ciertas características que puedan ser útiles para el abordaje del trabajo y sobre todo para dimensionar la cantidad de datos con los que se está trabajando. Inicialmente, puede sorprender la cantidad de datos que se obtienen de un proceso de Clickstream siendo que, si bien hay 11500 usuarios aproximadamente, se trata de más de 22 millones de clicks realizados en todo el año 2018. Esto es importante resaltarlo para identificar a un proceso de Clickstream con mecanismos de manipulación de grandes volúmenes de datos que son a su vez muy homogéneos entre sí, es decir, si bien son muchos datos la variedad de atributos son relativamente pocos ya que no superan las 15 variables.

En segundo lugar, se puede ver a continuación la distribución de los clicks a lo largo de todo el año estableciendo los 12 meses como formas de agrupación. Se puede notar que en los meses de enero y febrero disminuye significativamente la cantidad de clicks realizados por los usuarios. El motivo de dicha disminución talvez podría darse según los días no laborales de una gran parte de la población ya que coincide con la temporada de verano en Argentina en la cual se suelen brindar vacaciones a aquellas personas que trabajar en relación de dependencia.

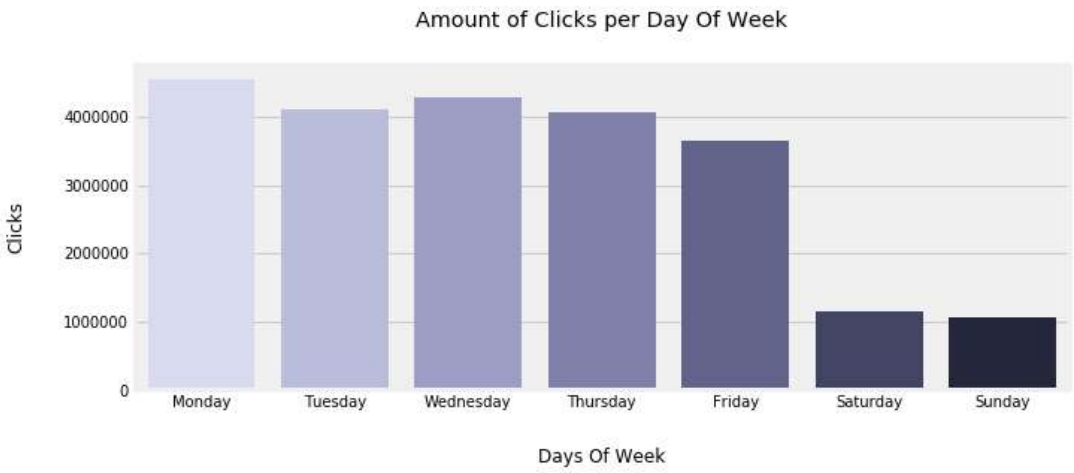


Por otro lado, se expone un análisis grafico de la cantidad de clicks por día que realizan los usuarios del banco detallando algunas situaciones curiosas como el aumento de clicks en los primeros días del mes y también la disminución muy marcada de clicks en el día 13 de cada uno de los meses. La razón por la cual se puede ver incrementada la cantidad de clicks en los primeros días de cada mes puede darse por el hecho del cobro de un salario que suele suceder en los primeros días del mes, no obstante, el motivo por el cual el día 13 de cada mes obtiene pocos clicks es una incógnita para el analista.



Por último, se desea mostrar cómo es la distribución de clicks según los días de la semana. En un análisis exante, puede suponerse que los fines de semana no son días muy concurridos en la plataforma web del banco o sus aplicaciones móviles dado que las personas suelen tener sus días de descanso y por lo tanto no usarían tanto los servicios del banco. El gráfico

presentado a continuación reafirma esa idea ya que los días que menos actividad tienen son los sábados y domingos que se diferencian fácilmente del resto de los días de la semana.



Apartado 2.2 Integración y Preparación

A lo largo del proceso, existieron múltiples complicaciones que afectaron los tiempos de entrega y las formas de resolver los diferentes problemas que se enfrentaron. Aun así, todos fueron sorteados con suma responsabilidad evitando generar errores en la base de datos que podían recaer en métricas y modelos erróneos por falta de capacidad a la hora de manipular la base de datos.

Uno de los mayores problemas con lo que se tuvo que lidiar fue unificar las acciones en una sola instancia. Como se mencionó anteriormente, la tabla principal indicaba por cada instancia un click que realizó el usuario en un momento del tiempo dado. Por ejemplo, el usuario con id 54 el día 4 de marzo de 2018 a las 15 horas 25 minutos y 36 segundos hizo click en la página de inicio de sesión de su cuenta de home banking. Pero a su vez, es mismo usuario hizo click en esa misma acción en repetidas ocasiones por lo cual la tabla generaba múltiples instancias para una misma acción dado que fue si bien fue realizada por el mismo usuario, el tiempo en que se realizó el click fue diferente. Al querer tener una acción individual por instancia, se decidió concatenar las páginas para generar las acciones en una columna nueva, y luego tomar los valores únicos de esa columna. De esta forma se generó

una lista de todas las acciones posibles que pueden realizar los usuarios quedando a disposición una tabla con clicks en cada instancia y otra tabla con acciones.

Este procedimiento fue elegido por el analista, pero no por eso es la mejor opción, sino que puede existir otro tipo de mecanismo que permita evaluar el comportamiento de los usuarios tomando otro proceso de seguimiento.

De alguna forma, se debía llegar a una tabla final donde cada registro sea un id de usuario único para poder caracterizarlo. Para esto se utilizó en Python la librería pandas que contiene una función de suma utilidad en estos casos llamada crosstab. Esta función permite individualizar cierto atributo generando una función agregada en el resto de los atributos cumpliendo el rol similar al de una tabla pivote tan frecuentemente utilizadas en la actualidad. Una vez que se obtuvo este proceso, la función agregada fue la de contar instancias con esa misma acción. Mediante este proceso se obtuvieron dos tablas finales resultantes, una con cada id de usuario en las filas y con las acciones en las columnas, y por otro lado una tabla con cada id de usuario en las filas, y con las páginas en las columnas. Es decir, una tabla tenía el comportamiento de cada usuario según la cantidad de veces que presionaba el click en cierta acción, y otra tabla tenía el comportamiento de cada usuario según la cantidad de veces que presionaba el click en cierta página. Vale la pena destacar nuevamente que cada acción es una concatenación de la página con sus características. Al existir más de 1800 páginas, por ejemplo, se llegaba a tener tablas con 2500 columnas aproximadamente que dimensiona la magnitud del alcance de la cantidad de datos que se manejan en el trabajo.

Si bien ese fue el problema con más relevancia en lo que respecta a la preparación de los datos, también se tuvo que decidir la forma en que se incorporaba el contenido sobre la velocidad de internet y la conectividad del usuario en el momento de iniciar sesión dentro de la página web del banco o su dispositivo móvil. Para esto se decidió tomar la mediana mensual de los usuarios en lo que respecta a la velocidad de conexión y la moda muestral del tipo de dispositivo móvil con el que el usuario ingresa a los servicios del banco. Dado que la distribución de este último atributo es binario, era indiferente usar la mediana o la moda ya que el resultado iba a ser el mismo porque la persona se conecta desde una computadora o bien desde su celular.

Finalmente, las tablas quedaron expresadas de la siguiente forma. La primera con los id de usuario como filas (se trata de unos 11500 usuarios aproximadamente) y las acciones

posibles como columnas, y la segunda con los id de usuario como filas y las páginas como columnas. Los valores que comprenden las tablas son la cantidad de veces que el usuario realizo la acción o ingresó a la página correspondiente, donde luego esos valores fueron normalizados para el mejor análisis de los atributos.

El motivo por el cual se decidió elaborar dos tablas diferentes fue con el propósito de probar diferente ingeniería de datos para ver cual tenía una evaluación mejor. Es decir, se decidió probar el modelo en ambas bases de datos resultantes del proceso mencionado recientemente para ver si alguna forma de tratar la base de datos era más conveniente que otra. El resultado que se expone a lo largo del trabajo fue el resultante de trabajar con la tabla que comprende las acciones de los usuarios, es decir, a la concatenación de los atributos correspondientes a cada click realizado.

Apartado 2.3 Algoritmo de Predicción.

Habiendo explicado anteriormente el origen de los datos, la problemática con respecto al tratamiento de la privacidad de estos, el proceso de integración de información y la posterior preparación de la tabla final a utilizar, se puede comenzar a explicar el procedimiento de aplicación de algoritmo seleccionado. Este proceso no solo incluye a la función matemática en particular sino también a una serie de mecanismos relacionados al Machine Learning que optimizan el uso de los datos y el desempeño del modelo.

Inicialmente se va a retomar el concepto de base de entrenamiento incorporando un nuevo termino llamado validación cruzada que es de mucha utilidad en este tipo de proyectos de datos. Como se mencionó anteriormente, un error muy común cuando se aborda este tipo de temáticas es el sobreajustamiento del modelo a los datos utilizados. Esto se refiere a la poca precisión que puede tener un modelo cuando se lo desea aplicar a datos totalmente nuevos de los cuales no tiene experiencia. Este proceso llamado “overfitting”, es un campo de estudio al que se le presta mucha atención dado que se puede confundir un modelo muy bueno que en la realidad no aplica bien para el tema abordado. El riesgo al que se expone el analista en caso de que pase por alto este proceso es muy elevado dado que puede proponer un modelo con un muy buen desempeño, pero la realidad le indica otro tipo de resultado.

El proceso de validación cruzada consta en separar la tabla con la que se desea trabajar en dos partes mutuamente excluyentes que confirman el 100% de la base de datos, pero no comparten ningún registro. Una vez establecida esta separación, se denomina a una parte la base de entrenamiento y a la otra parte la base de validación. Usualmente, la base de entrenamiento capta entre el 60% y el 90% del total haciendo que la base de validación sea el resto. Esto se divide en ese sentido porque el modelo va a desempeñarse de mejor forma cuanto mejor pueda aprender de una variedad grande de observaciones, y justamente por eso se decide entrenar con la base de mayor magnitud. En cambio, la otra base de validación solo se va a utilizar para evaluar que el modelo pueda ser generalizado evitando así problemas de sobreajustamiento.

Hasta ese momento solo se hizo una división de la base de datos para que el modelo pueda aprender de la base de entrenamiento y pueda generalizar aplicando el modelo en la base de validación. Aun así, el proceso no concluye dado que existe la posibilidad de que particularmente la separación de la base de datos se haya hecho de forma tal que la base de entrenamiento tenga una varianza menor que la base de validación o viceversa y que por lo tanto a la hora de aplicar el modelo tenga un desempeño por debajo de lo esperado. Es por este motivo que se requiere de un procedimiento de validación cruzada donde se pretende mitigar el riesgo mencionado anteriormente estableciendo una variedad de muestreo sobre la base de datos total.

Este mecanismo consta en tomar una cantidad n de separaciones entre base de entrenamiento y base de validación. Es decir, al final de la separación van a existir n casos distintos donde cada uno va a tener su propia separación entre entrenamiento y validación siendo cada caso independiente del otro. De esta forma, se tienen n bases de entrenamiento y n bases de validación donde cada tramo de entrenamiento corresponde a su tramo complementario de validación.

Una vez que se tiene el proceso de validación cruzada diagramado, solo resta aplicar el mecanismo de aprendizaje automatizado para evaluar el modelo. Este paso refiere a la aplicación de un modelo previamente seleccionado para que pueda aprender de cada una de las bases de entrenamiento y luego pueda evaluar el modelo en cada una de las bases de validación. Dado que trabaja con n cantidad de casos, se van a obtener n cantidad de evaluaciones del modelo expresadas como una variable aleatoria en sí misma. Es decir, va a

tener su media aritmética y su varianza si por ejemplo se seleccionara la métrica de precisión para evaluar el modelo.

Una vez que se obtiene la distribución de la nueva variable aleatoria que surge de evaluar métricamente el modelo aplicado, se debe interpretar los resultados para decidir si el modelo seleccionado tuvo un buen desempeño o bien se encontró por debajo de lo esperado. Esta decisión corresponde pura y exclusivamente al analista dado que se requiere una experiencia en la asignatura para ser capaz de tomar la decisión adecuada. Aun así, no existe una única visión ni una verdad absoluta de cómo proceder en estos casos ya que la interpretación subjetiva de cada profesional en conjunto con la visión de negocio va a determinar el modelo a utilizar.

En algunos casos, los analistas prefieren un modelo que luego de haber sido expuesto a un proceso de validación cruzada, tenga una varianza acotada para brindar más estabilidad en sus predicciones, pero en otros casos la visión de negocio talvez exija que la media aritmética sea la más indicada para seleccionar el modelo. Entre un modelo con mayor estabilidad dado que tiene una varianza menor, y otro modelo con mayor media de precisión, pero a su vez tiene mayor varianza, el analista va a tener que transmitir la interpretación de dichos resultados al encargado del negocio para poder tomar una decisión con la información técnica visible.

Si bien se nombró el mecanismo de selección de modelos y de métrica de evaluación, durante la explicación del proceso mencionado anteriormente no se hizo énfasis en estos conceptos que son muy importantes tenerlos en claro. Tanto la selección del algoritmo a aplicar como la métrica de evaluación suelen tomar una gran parte de las horas consumidas en un proyecto que aplique ciencia de datos para una organización, es por eso que tomarse el tiempo adecuado para problematizar al respecto debe ser un proceso obligado dentro de la toma de decisiones.

Para este caso en particular, el algoritmo que se utilizó para desarrollar el trabajo fue el método del descenso del gradiente el cual tiene una gran base teórica dentro de la estadística aplicada al aprendizaje automático. Aun así, se podría haber optado por otro tipo de algoritmo que tenga un buen desempeño para casos de clasificación binaria como lo pueden ser los árboles de decisión o las regresiones logísticas tan frecuentemente utilizadas en estos últimos años. Luego de haber hecho una prueba muy simple en el código de Python, se decidió optar por el inicialmente mencionado porque justamente posee una varianza menor

con una media de precisión mayor trabajando con el mismo set de datos. A pesar de esto, no se descarta que una medida de optimización de parámetros para el resto de los modelos pueda resultar superadora al modelo propuesto y por lo tanto se pueda mejorar la predicción.

El accionar del algoritmo del descenso del gradiente consta en minimizar la función de error del proceso para encontrar mínimos a través del cálculo de derivadas parciales con respecto a cada atributo que se desea utilizar como regresor. El conjunto de derivadas parciales conforma un vector que indica la dirección en la que la pendiente asciende, denominando dicho vector como el gradiente. Lo que se desea lograr es descender a través de la utilización del opuesto de dicho gradiente para luego repetir el proceso hasta encontrar el mínimo global en donde los casos donde pequeñas variaciones de los atributos no impliquen una mejora significativa en la función de error. Si bien este es el proceso ideal que implica aplicar este tipo de algoritmo, no siempre se puede llegar a un mínimo global, sino que talvez, el método queda atrapado en mínimos locales lo cual es problemático.

Un aspecto a tener en cuenta es el ratio de aprendizaje que introducimos en el proceso dado que va a indicar cuanto afecta el gradiente a la actualización de los parámetros seleccionados en cada iteración para realizar la predicción. Este último concepto va a ser el encargado de evitar una convergencia en mínimos locales para llevar la solución a mínimos globales. Aun así, la incorrecta definición de este ratio puede generar no solo que se encuentren mínimos locales, sino también que el proceso no converja nunca y se genere un bucle infinito que no tiene solución. Este último caso suele suceder cuando el ratio es considerablemente alto y por lo tanto el proceso no es capaz de converger a una solución. En cambio, cuando el ratio es muy bajo, el desempeño del algoritmo seguramente sea bastante bueno pero a costa de calcular múltiples derivadas parciales lo cual puede generar demoras y hacer ineficiente el proceso.

Por último, se va a explicar la importancia de establecer un concepto de métrica adecuado para el modelo generado dado que va a ser el encargado de evaluar si el desempeño del modelo es el esperado o bien se decide abandonar el camino cursado para seleccionar otro tipo de algoritmo. En este caso dado que la variable de salida del modelo es binaria como se mencionó anteriormente, la métrica debe ser establecida para casos en el cual se pueden evaluar este tipo de variables.

Para casos de variables binarias existen dos tipos de errores, el error de tipo 1 o falsos positivos que mide aquel error donde una observación es verdadera y la estimación indica

que es falsa, y está el error de tipo 2 o falso negativo que mide aquel error donde una observación es falsa y la estimación indica que es verdadera. Es de suma importancia comprender esta diferenciación dado que, según el modelo de negocio que se encuentre vinculado al análisis se va a tener una tendencia en preferir cometer un por sobre el otro.

En este caso la métrica de error seleccionada fue el área bajo la curva de ROC que representa gráficamente la sensibilidad frente a la especificidad de un clasificador binario según sea el umbral de discriminación. Es decir, mide como va variando los falsos positivos y los verdaderos positivos a medida que el umbral de discriminación va modificándose. Esto genera una curva en un plano que a su vez se le puede calcular el área bajo la curva que estará comprendida entre 0 y 1 donde 0.5 es el azar en la estimación y 1 es la estimación perfecta. Mas cercano al siguiente apartado se va a profundizar sobre estos conceptos que son fundamentales para evaluar modelos de aprendizaje automático.

Modelo y Procesamiento

En este apartado se va a detallar en mejor medida la propuesta del trabajo indicando los mecanismos que fueron utilizados a lo largo del proceso y explicando las decisiones que fueron tomadas por el analista en cada caso. Esto permitirá generar una evaluación sobre el procedimiento aplicado y poder así establecer conclusiones de forma crítica sobre el tema abordado a lo largo del trabajo.

Como ya se mencionó anteriormente, no solo se pretende cumplir con el objetivo de predecir las conversiones de los clientes del banco en este trabajo, sino también se intenta brindar algún tipo de explicación que pueda ser de relevancia a la hora de analizar los parámetros del modelo con respecto al comportamiento de los usuarios web. Estas funciones de los algoritmos matemáticos suelen complementarse muy bien porque a pesar de que generalmente se relacionan los conceptos con un hecho meramente predictivo, puede ser de mucha utilidad para el proceso de toma de decisiones comprender el objetivo del trabajo con una visión más holística.

Apartado 3.1 Desempeño del Modelo

En los apartados anteriores del trabajo se mencionó que la métrica de evaluación que se utilizará para medir el desempeño del modelo será aquella que mide el área bajo la curva de ROC (Narkhede, S. (2018)) (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor). Esta métrica es muy utilizada en el ambiente de la ciencia de datos ya que permite analizar de forma rápida e intuitiva el desempeño de un modelo matemático que evalúa el comportamiento de un algoritmo que estima valores binarios como es el caso de este trabajo. De lo contrario, se si se tratase de predecir valores que no son binarios, así sean variables continuas, variables ordinales, o variables nominales, se debería apuntar a otro tipo de métrica que funcione de manera correcta para esos casos particular.

El área bajo la curva de ROC, indica justamente un área comprendida entre el plano de los valores reales positivos y una curva resultante del modelo aplicado. Los ejes del plano refieren al ratio de verdaderos positivos en el eje vertical, y a los falsos positivos en el eje horizontal. Además, algunos autores identifican al ratio de verdaderos positivos como el índice de sensibilidad y al ratio de falsos positivos como el valor resultante de la unidad menos el índice de especificidad del modelo.

Para realizar esta métrica de desempeño de un modelo, es crucial elaborar la matriz de confusión resultante de la aplicación del algoritmo dado que con dicha matriz se va a desarrollar el grafico correspondiente a la curva de ROC. En esta matriz que tanto se está utilizando actualmente para identificar problemas en valores de predicción en diversas áreas de aplicación, va a expresar los valores correspondientes a los verdaderos positivos, falsos positivos, falsos negativos y por último los verdaderos negativos. A continuación, se detalla el sentido de cada uno de estos valores:

- Verdaderos positivos (VP): son aquellos casos observados donde la predicción los clasificó como clase positiva cuando en realidad eran observación de clase positiva. Hubo acierto en la predicción.
- Falsos positivos (FP): son aquellos casos observados donde la predicción los clasificó como clase positiva cuando en realidad eran observación de clase negativa. No hubo acierto en la predicción.
- Falsos negativos (FN): son aquellos casos observados donde la predicción los clasificó como clase negativa cuando en realidad eran observación de clase positiva. No hubo acierto en la predicción.

- Verdaderos negativos (VN): son aquellos casos observados donde la predicción los clasificó como clase negativa cuando en realidad eran observación de clase negativa. Hubo acierto en la predicción.

La distribución de las observaciones a lo largo de la matriz va a estar dada por un umbral de decisión que permitirá decidir en qué cuadrante cabe cada una de las observaciones dado el nivel de predicción que posea cada una. Si se fija un umbral de 50% por ejemplo, todas las observaciones que tengan como probabilidad de predicción un valor mayor a 0.5, serán clasificada como clase positiva. En cambio, todas las observaciones que tengan como probabilidad de predicción un valor menor a 0.5, serán clasificadas como clase negativa. A pesar de que en este caso se plantea un umbral de decisión de 0.5, se pueden aplicar múltiples umbrales según el modelo de negocio que se esté analizando.

A priori es un simple pensar que a medida que más verdaderos positivos y verdaderos negativos pueda obtener el modelo, mejor desempeño tendrá. Aun así, si bien esto es cierto, hay cierto umbral que decidir por parte del analista o por parte de la persona idónea en el negocio que va a influir en la cantidad de observaciones que se aplique para cada caso. Justamente este umbral es tomado por la métrica del área bajo la curva de ROC para medir el desempeño del modelo de forma dinámica y por lo tanto se puede evaluar la mejor decisión a tomar según el modelo de negocio abordado. A pesar de que pueda parecer poco intuitivo un caso donde el umbral tenga importancia, se pueden mencionar casos médicos como ejemplo donde los falsos negativos tienen un peso mucho mayor que los falsos positivos y por lo tanto el umbral a tomar debería modificarse en estos casos. Cuando se tiene el caso de tomar la decisión de la aplicación de una vacuna para una enfermedad importante que salva la vida de aquellas personas que tienen la enfermedad y a su vez genera efectos secundarios menores a aquellas personas que no tienen la enfermedad, la definición del umbral suele ser muy baja en estos casos porque es preferible afectar en menor medida a personas sanas que no salvar a aquellas personas que están sufriendo la enfermedad.

Con este ejemplo, se puede ver que existen dos tipos de errores en estos casos predictivos, pero no necesariamente dichos errores son igual en sus magnitudes. Estos errores se denominan error de tipo I y error de tipo II refiriéndose a los falsos positivos y a los falsos negativos respectivamente. A pesar de que idealmente el mejor de los casos es no cometer

ningún error y predecir perfectamente cada una de las observaciones, en la realidad suelen suceder y la forma de minimizarlos es aplicar el mejor algoritmo posible e identificar claramente los costos que pueden tener cada error según cada caso.

Cuando los autores se refieren a los costos de los errores, no siempre se refirieren a valores monetarios, sino que pueden referirse a otro tipo de valuaciones que pueden estar dadas por fuera de lo nominal lo cual hace muy difícil cuantificar los costos. Aun así, es un proceso que se suele cuantificar con ayuda de las personas idóneas en el negocio porque permite mejorar la toma de decisiones y brindar una explicación intuitiva a aquellas personas que no están involucradas con los conceptos de matriz de confusión y métricas de desempeño.

Habiendo definido estos conceptos, se puede mencionar que la curva de ROC apunta justamente a evaluar este tipo de controversias donde hay que decidir la ubicación del umbral para minimizar esos costos mencionados. Al situarse en un punto sobre la curva de ROC en el plano, se puede establecer el umbral deseado conociendo los errores que tuvo el algoritmo y por lo tanto conociendo los Falsos positivos y los Falsos negativos obtenidos durante el proceso. Cabe destacar que a medida que se disminuyen los falsos positivos, de la forma contraria estarán aumentando los falsos negativos dado que se trata de un juego de suma cero donde para mejorar un aspecto necesariamente el proceso se ve obligado a empeorar otro. En los casos extremos, el proceso dispondrá de un umbral cercano a 0 cuando quiera que todas sus predicciones sean de clase positiva, y por el contrario dispondrá de un umbral cercano a 1 cuando quiera que todas sus predicciones sean de clase negativa.

Los valores que arrojará el área bajo la curva de ROC deberían estar en comprendidos entre 0.5 y 1 siendo 0.5 el azar mismo y 1 el modelo perfecto que no posee errores ni de tipo I ni de tipo II. A medida que la métrica más se acerca al valor 1, mejor será el modelo que identifica con mayor precisión las observaciones. Algunos autores han mencionado a modo de guía una forma de interpretar los resultados de la métrica en cuestión. A continuación, se detalla cada caso:

- AUC ROC igual a 0.5: es el azar mismo.
- AUC ROC entre 0.5 y 0.6: el modelo es malo.
- AUC ROC entre 0.6 y .075: el modelo es regular.
- AUC ROC entre 0.75 y 0.9: el modelo es bueno.
- AUC ROC entre 0.9 y 0.97: el modelo es muy bueno.

- AUC ROC entre 0.97 y 1: el modelo es excelente.

A pesar de que intuitivamente un modelo con una métrica AUC ROC de 1 pueda ser ideal por el alto grado de aprendizaje del algoritmo, posiblemente se esté tratando de un error de sobreajustamiento o algún error dentro del proceso porque es de esperar que un modelo predictivo cometa ciertos errores analizando variables complejas por lo que una métrica tan buena debe ser analizada con profundidad por el analista con mucha cautela.

Dicho esto, la métrica del área bajo la curva de ROC es de mucha utilidad para comparar modelos que analicen clasificaciones binarias como es el caso del trabajo abordado donde la clasificación se da en torno a la conversión o no dentro de la página web del banco por parte de los usuarios.

Particularmente en el caso resuelto, el área bajo la curva de ROC que se obtuvo fue cercana al 82% lo cual indicaría según la clasificación anteriormente presentada que se trata de un modelo con un buen desempeño. Si bien esta métrica podría continuar mejorando, requeriría de modificaciones en el modelo que permitan generar una métrica superadora.

Apartado 3.2 Desarrollo de Código de Programación

Ya habiendo mencionado el debido proceso para llevar a cabo el presente trabajo, solo resta indicar que tipo de lenguaje de programación se optará para desarrollar el código propiamente dicho. No solamente se define el lenguaje, sino también es necesario definir las aplicaciones que van a ser utilizadas para mejorar la escritura y comprensión del código de programación. Es decir, si bien se puede desarrollar el código en su lenguaje determinado, en estos casos es una buena práctica utilizar diferentes herramientas que ordenen y limpien el código de forma tal que sea fácil de comprender por un tercero.

El lenguaje que se utilizará va a ser Python como ya se mencionó anteriormente dado que se desempeña muy bien en este tipo de proyectos por su versatilidad y facilidad al momento de la escritura. Cabe destacar que se trata de uno de los lenguajes más utilizados en la actualidad según StackOverflow por los desarrolladores con una gran captación de parte de los científicos de datos dado su gran variedad de librerías creadas para esta asignatura. La facilidad de manipular datos en Python y su gran comunidad hace que sea el elegido entre diversos lenguajes de programación destinados a la ciencia de datos como lo pueden ser R o SPSS entre otros.

Para escribir el código de programación en Python se pueden utilizar múltiples aplicaciones que facilitan la escritura del código. Se optó por escribir dentro de Jupyter Lab dado su facilidad para realizar pruebas dentro del mismo código evitando generar demoras en la ejecución de códigos completos. Esta decisión no necesariamente sea la mejor, hay múltiples opciones muy similares que poseen características muy buenas y generan un buen desempeño del código.

Adicionalmente, se decidió trabajar con Anaconda que se trata de una aplicación que comprende múltiples lenguajes de programación destinados a la ciencia de datos y permite trabajar a su vez con muchos editores de texto lo cual brinda una amplia gama de posibilidades. Este paso es pura y exclusivamente decisión del analista por su comodidad al trabajar con este tipo de aplicaciones capaces de generar entornos virtuales variados, pero para este caso no brinda muchos beneficios sustanciales a la hora de desarrollar el presente trabajo.

A lo largo del trabajo, se experimentaron múltiples problemas para con la base de datos y su manipulación para generar las condiciones de desarrollo del modelo predictivo. Si bien muchos de los problemas se sortearon fácilmente y no requieren de mucha complejidad en su abordaje, existieron algunos inconvenientes que vale la pena mencionar y profundizar.

Inicialmente, el primer problema que se tuvo que solucionar fue la toma de decisión de trabajar sobre un horizonte temporal en particular, es decir, los meses con lo que se iba a predecir las futuras conversiones. No es lo mismo intentar predecir las conversiones del mes siguiente contando con el mes actual o contando con los últimos 9 meses. De la misma manera, no es lo mismo intentar predecir las conversiones del próximo trimestre o del próximo mes independientemente de cuantos meses se utilicen para hacer el entrenamiento. Esta decisión debe recaer en parte en el analista dado que requiere no solamente conocimiento técnico del tema sino también conocimiento del rubro de negocio. La decisión tomada fue incluir como variable los meses que se quieran tomar para hacer la predicción y los meses que se quieren tomar para hacer el testeo del modelo. Luego de un amplio análisis al respecto, se decidió predecir las conversiones del siguiente trimestre utilizando los 9 meses anteriores dado que se consideró que incorporaba la mayor información posible.

Por otro lado, se tuvo que resolver el problema de selección de columnas y de interpretación de dichas columnas. Se tomó la opción de trabajar con todas las columnas que comprendía la base de datos, desde las columnas con contenido del sitio web del banco hasta las

columnas referidas al usuario y su conexión móvil. Aun así, la interpretación de los datos puede ser algo confuso porque se optó por tomar la concatenación de los atributos del comportamiento en el sitio web del banco, es decir, tomar como acción única una sola columna que comprenda todo el accionar que realizo el cliente en cada click conteniendo las categorías. Justamente para realizar este paso se utilizó una simple concatenación que permitía ver el comportamiento único que tuvo el cliente en una sola columna.

Por último, para la aplicación del modelo se decidió disminuir la cantidad de columnas con las que se desea predecir las conversiones. Este proceso tomó su tiempo porque el límite es muy subjetivo y no está claro que posición tomar al respecto. A pesar de esto, se tomaron las 30 acciones más relevantes para el modelo que ya había sido entrenado para volver a ser entrenado con dichas acciones y por lo tanto reducir la dimensionalidad de la tabla de aprendizaje. Un gran aspecto positivo de haber realizado esta maniobra es la facilidad a la hora de explicar el modelo dado que se dejó de contar con más de 1000 atributos para utilizar finalmente solo 30 de ellos. Una alternativa a este tipo de procesos puede ser la aplicación del proceso de componentes principales que también cumplen la función de reducir la dimensionalidad de una tabla perdiendo la menor variación posible.

Si bien como se mencionó recientemente se decidió trabajar con las 30 variables que más información le brindaban al modelo, cabe destacar que la variable inicial PAGE era la que más explicaba con respecto al resto de las variables iniciales con las que se comenzó trabajando, teniendo en cuenta que había gran correlación entre el resto de las variables. Al momento de realizar la concatenación de las variables para identificar acciones únicas dentro del modelo, se perdió un poco de vista cual era el rol de la variable PAGE dentro del proceso. Aun así, dado la forma en que se concatenaron las variables iniciales se puede identificar al primer código como el correspondiente a la variable PAGE que se está mencionando. Por lo tanto, al tener una lista con las 30 variables más descriptivas para el modelo se puede rescatar algunas cuestiones que pueden ser de importancia. Los códigos 41, 109, 1259, 153 y 65 de la variable inicial PAGE, por ejemplo, son aquellos que mejor explican las predicciones. Estos códigos refieren a las siguientes descripciones:

/PRESTAMOS/INICIO

/PRESTAMOS/SOLICITAR-SELECCIONAR-SUBTIPO-PRESTAMO-PER

/PRESTAMOS/SOLICITAR_SELECCIONAR_SUBTIPO_PRESTAMO_PER

/VISAHOME/VISA-HOME

/TARJETAS/RESUMEN/0

De este análisis se pueden hacer al menos dos salvedades. Existen dos descripciones que son muy parecidas casi iguales lo cual indicaría presencia de valores repetidos en la base de datos que no fueron identificados por sus mínimas diferencias. Y, por otro lado, tendría sentido identificar a las paginas relacionadas con los prestamos como lo son las primeras 3 descripciones con la variable explicada que contiene las conversiones bancarias. Se puede afirmar que la cantidad de veces que el usuario ingresa en las paginas mencionadas recientemente es un gran indicador para que el modelo pueda predecir de la forma mas precisa posible.

Apartado 3.3 Potencialidad

El set de datos que se armó fue con el claro propósito de generar un modelo de predicción que sea capaz de identificar a través del comportamiento de los clicks de los usuarios del Banco Galicia en el año 2018 las conversiones que fueran a realizar los mismos usuarios en el primer trimestre del año 2019. Como se puede notar, se trata de un modelo que no está ejecutándose en tiempo real dado que no tiene como atributo el comportamiento altamente reciente del usuario, sino que solo analizará datos que pueden tener hasta 15 meses de antigüedad. Aun así, el mismo modelo va a ser capaz de obtener la importancia de cada variable y que atributos tomar en cada caso.

Teniendo en cuenta que se tienen muchos datos sobre el comportamiento de los clientes del banco, en estos casos donde se analizan los datos mediante un modelo de aprendizaje automático, se pueden generar nuevas variables que el analista estime a priori que son de importancia para el modelo como lo puede ser por ejemplo la cercanía en fechas de fiestas de fin de año, el cobro de aguinaldo, entre otras. Si bien estas variables no son establecidas por el origen de la base de datos, nada impide que puedan ser creadas por el analista para intentar mejorar el modelo y por lo tanto tener más precisión en su predicción. Seguramente, a medida que se creen dichas variables y se establezcan como atributos en el modelo, algunas podrán ser útiles y otras no tanto, pero como se mencionó anteriormente el modelo va a ser

capaz de diferenciarlas. Con esto, se quiere dar la idea que, si bien no es absoluto, en cuanto más variables para analizar haya mejor va a poder predecir el modelo.

Una vez que se desarrolla el modelo, el banco lo podría poner en producción para sus métricas periódicas donde va a poder estimar el capital que le va a destinar a sus productos crediticios y de esta forma poder optimizar el uso de sus recursos. Además, podría poner en práctica otro tipo de análisis como puede ser la identificación de posibles clientes que en la actualidad no consumen sus productos, o bien la identificación de clientes actuales que por alguna razón deciden contratar servicios con otro banco.

Para finalizar, se puede mencionar que a medida que el tiempo fue evolucionando, los negocios en conjunto con la tecnología también fueron creciendo y adaptándose al mercado actual donde la globalización y el internet nos mantiene permanentemente conectados al resto del mundo. Si se comprende las dimensiones del alcance que los datos tienen en la actualidad, los negocios podrían mejorar sus métricas en muchos aspectos ya que pueden individualizar a cada cliente dado los grandes avances en almacenamiento de información y poder de computo. La decisión de ahora en más recae sobre todo en los altos directivos de las organizaciones y el entusiasmo a innovar en mercados que están creciendo de manera significativa.

Conclusión

Luego de haber tratado sobre una variedad de conceptos vinculados a la ciencia de datos y su aplicación práctica en el ambiento del Clickstream bancario, se notó que los algoritmos generados pueden ser útiles para predecir y explicar las conversiones bancarias en un futuro no muy lejano. Además, se vinculó el análisis con muchas problemáticas que se ven en la actualidad pasando por la privacidad de los datos, la gobernanza de estos, y también el impacto social que generan estas nuevas herramientas informáticas. A pesar de que en la actualidad recién se está profundizando sobre estos conceptos, talvez las generaciones que vienen tendrán un trabajo más crítico acerca de la ciencia de datos y todas las aplicaciones que esta conlleva.

Por otro lado, se problematizó sobre el estado de la base de datos con el que se creó la predicción. Los profesionales que se dedican particularmente al análisis en esta área de

trabajo suelen hacer mucho énfasis en la importancia de contar con una base de datos que se encuentre en óptimas condiciones para que el analista pueda trabajar de forma segura y responsable.

En cuanto al aporte que genera el trabajo presentado, se puede afirmar que un algoritmo con el método del descenso del gradiente es capaz de predecir las conversiones bancarias en este caso con los datos dados a través del Clickstream de clientes. Aun así, hay variedad de modelos a aplicar que pueden ser superadores al presentado y por lo tanto tener mejores métricas de desempeño al momento de identificar las predicciones. Los algoritmos matemáticos que comprenden redes neuronales podrían ser una buena alternativa de aplicación dado su alta capacidad de predicción en grandes volúmenes de datos como es el caso analizado. Identificando los nuevos avances tecnológicos en esta área que suelen ser variados y muy novedosos, se puede intentar replicar modelos exitosos en otras áreas diferentes a la recientemente analizada. Este tipo de desarrollos donde se transfiere el conocimiento de un caso externo al caso particular estudiado, son muy comunes hoy en día dado el alto grado de aprendizaje que son capaces de obtener los nuevos algoritmos.

Adicionalmente, la incorporación de datos externos sería una buena medida a implementar para mejorar más el modelo aún. El origen de los datos externos va a ser una decisión tomada en conjunto por el analista y la persona idónea en el negocio que pueda brindar propuestas desafiantes para la aplicación. A pesar de esto, no siempre los datos externos a incorporar contienen mucha lógica o sentido común, sino que a veces el desconocimiento sobre el negocio puede ser una medida atractiva para la incorporación de estos datos ya que el campo de abstracción de una persona que no conoce en profundidad la situación es mucho más reducido y no contiene barreras conceptuales que se pueden encontrar en individuos idóneos en el tema.

Para finalizar, se propone aplicar lo mencionado recientemente para lograr un modelo superador que implique una mejor predicción de conversiones y lograr así resultados superadores. La generación de modelos de estas características no solo será de utilidad para el sector financiero, sino que también brinda avances al estado del conocimiento del rubro tratado en este trabajo.

Referencias Bibliográficas

Alpaydin, E. (2009). Introduction to machine learning. MIT press.

Bravo, J. A. F. (2005). Avatares y estereotipos sobre la enseñanza de los algoritmos en matemáticas. Junta de Gobierno de la FISEM, 31.

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big data.

Esteve, A. (2017). The business of personal data: Google, Facebook, and privacy issues in the EU and the USA. International Data Privacy Law, 7(1), 36-47.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management.

Hind, J. R., Nguyen, B. Q., & Peters, M. L. (2006). U.S. Patent No. 7,003,565. Washington, DC: U.S. Patent and Trademark Office.

Hüsemann, B., Lechtenbörger, J., & Vossen, G. (2000). Conceptual data warehouse design (pp. 6-1). Universität Münster. Angewandte Mathematik und Informatik.

Isaak, J., & Hanna, M. J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. Computer, 51(8), 56-59.

Lever, J., Krzywinski, M., & Altman, N. (2016). Points of significance: model selection and overfitting.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Muley, R. (2018). Data Analytics for the Insurance Industry: A Gold Mine, Page 7

Narkhede, S. (2018). Understanding AUC-ROC Curve. Towards Data Science, 26.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Tuesta, D., Sorensen, G., Haring, A., & Cámara, N. (2015). Inclusión financiera y sus determinantes: el caso argentino. Documento de Trabajo, (15/04).

Vives, L. (2015). La revolución del" fintech": el caso de Kantox. Harvard Deusto business review, (249), 76-82.

Wende, K. (2007). A model for data governance-Organising accountabilities for data quality management. ACIS 2007 Proceedings, 80.

Anexos/apéndices

Código fuente en el siguiente enlace:
https://www.kaggle.com/fgarciablanco/clickstream-banco-galicia-2019