UNIVERSIDAD NACIONAL DE TUCUMÁN

FACULTAD DE CIENCIAS ECONOMICAS
UNIVERSIDAD NACIONAL TUCUMAN

ANALÍTICA DE NEGOCIOS EN UNA FINANCIERA DE SAN MIGUEL DE TUCUMÁN. APLICACIÓN DEL MODELO LOGIT PARA EVALUAR LA MOROSIDAD EN LOS CLIENTES

Autor: De La Rosa, Milagro Del Valle

Director: Garcia, Javier Antonio

2019

Trabajo de Seminario: Licenciatura en Administración

INTRODUCCIÓN

Durante el ejercicio de las actividades en una organización, esta genera información con la que nutre cada uno de sus procesos internos de decisión. En la actualidad, las organizaciones generan una masiva cantidad de datos a partir de las operaciones que se realizan dentro de ella, estos son resguardados en los distintos sistemas de información y gestión con los que la compañía. Ante la abundancia de los datos se cree erróneamente que los problemas relacionados a la falta de información serán resueltos, sin embargo, la información de calidad solo será una realidad para la organización si esta cuenta con métodos de gestión, manipulación y análisis correctos de los datos.

El presente trabajo se desarrolla a partir de la necesidad de dar respuesta acerca de los métodos de explotación y análisis de los datos conocidos. Entre los métodos desarrollados se encuentran el business Intelligence y data mining, aplicados mediante técnicas estadísticas a una compañía regional de créditos con el fin de explorar como estos impactan en la toma de decisiones y la determinación de las estrategias. La metodología utilizada es de análisis cuantitativo de datos con diseño no experimental de corte transversal.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

Sumario: 1. Determinación del problema 2. Justificación 3. Objetivos de la Investigación 4. Hipótesis de la investigación 5. Metodología de la investigación

1. Determinación del problema:

La información que corre a lo largo y a lo ancho de la estructura organizativa de una empresa y el estudio de esta, logra que la toma de decisiones se adapte a la compañía y haciéndola única durante el curso de sus operaciones, entonces se debe destacar que el punto de partida de todo proceso decisorio de una empresa nace en el análisis de los datos y la información que en si misma se genera, siendo las decisiones las que marcan el futuro y la prosperidad de las mismas.

Las grandes compañías de hoy en día generan un gran cumulo de datos y son almacenados en grandes repositorios de datos diseñados especialmente para que estos se encuentren al alcance de todos los usuarios de la información, dichas estructura que alojan la masividad de datos fueron fruto de los avances de las tecnologías y la inversión de las compañías en una estructura tecnológica sólida que permite el acceso a los datos de manera eficaz, principalmente con fines de mantener la disponibilidad de la información por periodos prolongados, ya que esta es, en muchos casos el respaldo de las transacciones diarias de una compañía, sin embargo, dicha era de la masividad de datos almacenados trajo acarreado costos que debían ser

justificados más allá de su tarea principal de reguardo y explotar los frutos de la información. Con esto se refiere al hallazgo de dar un uso a la estructura de información de una compañía y hacer que esta permita contar con información más detallada a la hora de tomar la decisión convirtiéndola en una herramienta de competitividad para la organización.

En el ámbito regional una de las problemáticas que prevalece es que la era de la masividad de los datos ha desembarcado en las empresas de la región sin que estas cuenten aún con una estructura tecnológica sofisticada y más allá de esto, la realidad de estas empresas es que continúan realizando su proceso de toma de decisiones a partir de un análisis precario de los datos que generan diariamente sin aprovechar estos de manera completa y mejorar los diferentes tipos de estrategias para volver a la compañía más competitiva, estos hace que el crecimiento de estas empresas sea de menor impacto y también exponiéndolas a quedar obsoletas y sin poder competir a la altura de las empresas de la industria. A partir de lo narrado, surge un problema de ausencia de explotación analítica de los datos y administración de la información que les provee las actividades de una empresa.

Dentro de una industria específica, como ser la de banca y de servicios financieros, la tecnología ha logrado transformarlos para que los mismos puedan ofrecer sus productos a tan solo un clicks de distancias, logrando que los servicios que ofrecen sean más fáciles de adquirir por el consumidor. A partir de estos cambios tecnológicos pueden obtener la mayoría de los datos necesarios de los clientes, en cuanto a identificación y monitoreo de las actividades, por medio digital, por lo que la era de los servicios digitales también han logrado sumergir a la industria de servicios financieros al mercado de la masividad de los datos. Es por esto que la estructura de la información que recorren los circuitos de negocios de estas empresas pueden crecer y con esto, tienen una mayor facilidad para implementar el análisis de los datos de manera mucho más sofisticadas con el fin de determinar la clasificación de sus clientes para analizar el comportamiento o bien el nivel de riesgo de cada uno

de los perfiles. En las grandes empresas del país, en muchas áreas ya se pueden ver los resultados de la automatización de algunos procesos y como la tecnología soporta el gran volumen de datos generados diariamente, pese a esto, las pequeñas empresas de la región a un no lograr instaurar en sus metodologías de recopilación y análisis, la riqueza de la gran cantidad de datos que les proporciona el giro habitual de la compañía y sus operaciones.

El presente trabajo surge a partir de percibir la problemática de masividad de datos con falta de explotación por los referentes en una Compañía Financiera dedicada a la venta de micro-créditos a cientos de clientes de las principales provincias del noroeste argentino. Grandes volúmenes de datos valiosos y la falta de explotación para dirigir la gestión de los clientes conllevan al siguiente trabajo a búsqueda de alternativas, por medio de la investigación académica, que sean provechosos para el mejor análisis de la información y aportar valor al proceso decisorio realizado por la alta gerencia de la empresa.

A partir de observar la existencia de la problemática mencionada en el párrafo anterior en el desarrollo de las actividades de algunas empresas, el presente trabajo propone el estudio de una alternativa de análisis de datos usando como ejemplo el caso de una pequeña empresa tucumana dedicada a la venta de microcréditos para el consumo en las provincias de Santiago del Estero, Tucumán y Salta, la misma será mencionada a lo largo del trabajo como “La Financiera”.

A lo largo del trabajo se desarrollará el uso de los datos disponibles para el análisis generado por la Financiera con métodos estadísticos multivariantes que se caracterizan por relacionar las múltiples variables con el fin de aprovechar la información que aportan en conjunto.

2. Justificación

Es importante plantear que el mundo se encuentra en una era de transformación digital constante y esta solo puede aportar su máximo potencial si se explotan el poder de los datos. En efecto, las organizaciones se encuentran atravesando una época de revolución de los datos, esto es impulsado no solo por la abundancia de datos actual, sino por las tecnologías fundamentales que cambian la forma en que se reúne, almacena, analiza y transforma la información. Todo lo mencionado abre desafíos que se deben hacer frente con tal de no poner a la compañía fuera de competencia. Las organizaciones deben trabajar activamente para hacer arrancar el motor de la explotación de los datos y promover con ello el nacimiento de nuevas estrategias que la mantengan en crecimiento.

Los datos se deben reunir, almacenar, analizar y transformar para brindar beneficios que pueden ser prácticos. Estos procesos se encuentran en el centro de la etapa de innovación de los datos –la derivación de un valor inmenso a partir de cantidades enormes de información que es, de otro modo, improductiva.

Se pretende que el presente trabajo exponga una problemática real que pone en una situación de desventaja a empresas de la región y más puntualmente trabajar en cómo puede realizar la explotación de los datos usando como caso de ejemplo el de la Financiera de la región en la cual se basa la investigación.

3. Objetivos de la Investigación

El objetivo general del presente trabajo es proponer una metodología que permita obtener información relevante para la toma de decisiones mediante la explotación de los datos, por lo que a lo largo del desarrollo del presente trabajo se analiza un ejemplo práctico del caso de una empresa de San Miguel de Tucumán. A su vez, como parte de la investigación también se

definieron los siguientes objetivos específicos, recolectar y depurar la base de datos de préstamos de la empresa e identificar las variables que revisten mayor importancia en el análisis, modelar los datos por medio de métodos estadísticos (modelo logit) y simular un espacio para la toma de decisiones.

Los instrumentos de recolección de información para la presente investigación se basaron en entrevistas y extracción de datos de origen desde base de datos por medio de consultas lógicas.

4. Hipótesis de la Investigación

En función al objetivo planteado anteriormente para la investigación, la hipótesis sobre la cual se sustenta el presente trabajo trata del siguiente enunciado:

"La metodología de exploración, explotación y análisis de los datos de una empresa es de gran importancia para el desarrollo de nuevas estrategias y proceder a la toma de decisiones a partir del estudio exhaustivo de la información generada en la propia compañía"

5. Metodología de la Investigación

La metodología aplicada para el desarrollo del presente trabajo es de análisis cuantitativo de datos con diseño no experimental de corte transversal, ya que el mismo se basa en el estudio de modelos estadísticos como alternativa de solución al problema planteado.

Las investigaciones cuantitativas tienden a fragmentar la realidad y se trabaja con variables específicas que se cuantifican y se expresan en valores numéricos. En estos casos, es importante la fiabilidad, la validez y la realización de la limpieza de los datos, con lo que se pretende que los resultados sean objetivos y generalizarlos, utilizando técnicas estadísticas para el análisis de los datos.[1]

[1] SAMPIERI, Metodología de la Investigación. 5ta Edición. Editorial McGraw-Hill, (México 2010), pág. 33

Para que exista un método cuantitativo se requiere que entre los elementos del problema de investigación haya una relación cuya naturaleza sea representable por algún modelo numérico, ya sea lineal, exponencial o similar.

Durante el desarrollo de una investigación de corte cuantitativo y el análisis de los datos se pueden distinguir tres etapas fundamentales:

- Análisis exploratorio inicial de los datos que consiste en la depuración de la matriz de datos y en análisis descriptivos de los mismos.
- Una segunda fase de análisis bivariable a través de estudios inferenciales.
- Una tercera fase basada en análisis multivariantes.

En el análisis exploratorio inicial, también llamado EDA (exploratory data análisis), tiene especial importancia la depuración de la matriz de datos y primera aproximación al análisis descriptivo univariable a través de distribuciones de frecuencia, representaciones gráficas univariables, medidas de tendencia central, de variabilidad o dispersión, de posición, de asimetría, de curtosis, de comprobación de supuestos paramétricos, etc. (1)

Las pruebas de decisión estadística constituyen un aspecto importante del análisis de datos cuantitativo. Ejemplos de estas pruebas son la correlación de Pearson, prueba t de Student, análisis de la varianza (ANOVA), pruebas no paramétricas como X^2 (jicuadrado), T de Wilcoxon, U de Mann-Whitney prueba de McNemar, etc.

El análisis multivariante es propio de los estudios más sofisticados. Se justifica en el principio de causación múltiple, es decir, que los fenómenos complejos obedecen a múltiples causas y no a una sola. Esto tiene especial importancia, dado que las investigaciones en ciencias sociales y especialmente en el área educativa se ven sometidas a un gran número de variables que intervienen.

Algunos ejemplos de análisis multivariante son el análisis factorial de la varianza (ANOVA), análisis de la covarianza (ANCOVA), análisis multivariante de la varianza (MANOVA), el análisis factorial (AF) de componentes principales, el cluster analysis, análisis discriminante, etc

CAPITULO II

MARCO TEÓRICO: ANALÍTICA Y EXPLOTACIÓN DE DATOS

Sumario: 1. Datos: Explotación, análisis y toma de decisiones 2. Bussiness Intelligence: Mineria de datos y Análisis de datos 3. Mineria de datos 4. Técnicas utilizables del data mining

1. Datos: Explotación, análisis y toma de decisiones.

Durante estos últimos años han surgido múltiples herramientas que proponen a las compañías emprender, en base al conocimiento, análisis y explotación de datos, la gestión y el proceso de toma de decisiones enfocado en la analítica y a la inteligencia de negocio (Data Mining, Data Analitics, Business Intelligence,). En este capítulo se hará foco en los conceptos claves de estos métodos de análisis que fundamentan el objetivo del presente trabajo.

2. Business Intelligence: Minería de datos y Análisis de datos

El aumento exponencial en el volumen de datos ha llevado a una revolución de la información y el conocimiento, siendo hoy un aspecto clave de la investigación y la construcción de estrategias para recopilar información y conocimientos significativos de los datos existentes. Toda esta información es almacenada en un gran almacén de datos, que luego se utiliza con fines de

clarificar toda aquella información confusa y desordenada, en información útil para la una organización o bien el descubrimiento de nuevos conocimientos.

Cuando en una organización ocurren esta revolución de la información se deben unir los esfuerzos para sacar provecho a la gran masa de datos que invaden todo el circuito organizacional. Es por esto que **las grandes empresas** buscaron (y encontraron) métodos eficaces para poner fin a la problemática.

Uno de los métodos que se han utilizan para dar pelea a la revolución de la información es el Business Intelligent, siendo este método capaz de transformar los datos en información procesable. El business Intelligent ayuda a optimizar las decisiones comerciales estratégicas y tácticas de las organizaciones utilizando las aplicaciones, la infraestructura y las herramientas, y las mejores prácticas que facilitan el acceso a los hechos y cifras operativos de una organización.

Es importante entender que las herramientas y modelos para la implementación de Business Intelligence en una organización depende de la calidad de los datos originados por la misma, la infraestructura tecnológica que soportan dichos datos y el compromiso de la gerencia en emprender la gestión de la organización a partir de la explotación de los datos.

En lenguaje sencillo, Business Intelligence analizará los datos complejos sin procesar de una organización y los transformará en información útil según lo requiera la empresa. Al utilizar esta información útil, la empresa sabrá qué está funcionando, qué no, cuál es el futuro y cómo puede mejorar su empresa, por lo que consiste en explorar grandes volúmenes de datos para asistir a los ejecutivos en la toma de decisiones.

En términos más técnicos Business Intelligence es la práctica de exploración iterativa y metodológica de los datos generados dentro de una organización, utilizando modelos de análisis estadísticos, enfocándose en recabar y combinar una gran cantidad de datos con el objetivo de derivar en ideas que no llegan a ser percibidas en una escala menor.

Para que se pueda obtener un mejor entendimiento de cómo se debe implementar business Intelligence se desarrollará a continuación los procesos involucrados:

1. Recolectar los datos complejos sin procesar de una organización
2. Estudio y análisis de los datos
3. Visualización de los resultados
4. Sobre la base de estos, las empresas tomarán decisiones inteligentes para el bienestar de la organización.

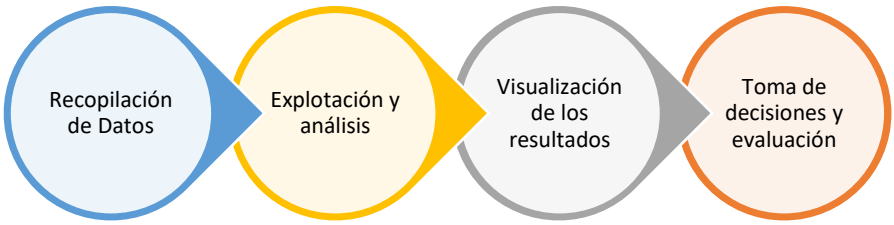


Figura 1: Proceso de implementación del Business Intelligent.

Durante la recopilación de los datos, ya sea de uno de los procesos administrativo dentro de la organización o bien, de todo el circuito empresarial de la una entidad, los datos son solo piezas que se guardan con el fin de que en algún momento sean utilizados para sustentar hechos dentro de la organización. Por ejemplo, durante el proceso de venta de un producto a un nuevo cliente, una empresa se interesa por responder las siguientes preguntas ¿Quién es? ¿Cuál es el producto que requiere? ¿Cuántas unidades? ¿Cuáles serán los términos de pago?, para responder dichas preguntas el cliente debe ser identificado, ya que en caso que el mismo realice la compra pueda reclamarse su pago de manera correcta, luego debe haber requerido al vendedor el producto que necesita y cuantas unidades y haber expuesto dentro de su identificación el método de pago requerido. Durante todo este

proceso, como habrá podido notar, se ha generado una gran cantidad de datos dentro de solo ese circuito, sin embargo, el nombre del cliente no será tan indicativo dentro de un análisis de perfil de cliente como si lo será su calificación crediticia y el tipo de producto que compra, es por esto que se necesita de un estudio más a fondo de los datos que se resguardan.

El Análisis de datos (Data Analitycs) y la minería de datos (Data mining) son dos subconjuntos de Business Intelligence que trabajan para darle mayor riqueza a los datos recolectados, es por esto que es importante definirlos detalladamente y entender cada uno de estos subprocesos dentro del BI.

Minería de datos, es un proceso sistemático y secuencial de identificar y descubrir patrones ocultos e información en un gran conjunto de datos, en cambio cuando se habla de análisis de datos se hace referencia en un superconjunto que consiste en la extracción, limpieza, transformación, modelado y visualización de datos con la intención de descubrir información significativa y útil que puede ayudar en la obtención de conclusiones y tomar decisiones, es decir es un conjunto completo de actividades que se encarga de la recopilación, preparación y modelado de datos para extraer conocimientos o conocimientos significativos, es por eso que se entiende la minería de datos como una actividad dentro del análisis de datos.

La minería de datos identifica y descubre un patrón oculto en grandes conjuntos de datos, mientras que el análisis de datos proporciona ideas o pruebas de hipótesis o modelos de un conjunto de datos. El objetivo de Data Mining es hacer que los datos sean más utilizables, no necesita ninguna hipótesis preconcebida para identificar el patrón o la tendencia en los datos, sin embargo, el análisis de datos prueba una hipótesis dada, utilizando para esto todo el conocimiento sobre el negocio y su estructura organizativa, acompañando sus resultados con la visualización de estos. La minería de datos es una tecnología usada para descubrir información oculta y desconocida, pero potencialmente útil, a partir de las fuentes de información

de una organización, esta ha sido el resultado de la evolución del sistema de información de las compañías, permitiendo tomar el proceso más allá del acceso y navegación retrospectiva de los datos, sino hacia la entrega de información prospectiva y proactiva.

Bases para la comparación	Minería de datos	Análisis de los datos
Definición	Es el proceso de extraer un patrón específico de grandes conjuntos de datos.	Es el proceso de ordenar y organizar datos sin procesar para determinar ideas y decisiones útiles.
Área de experiencia	Implica la intersección del aprendizaje automático , las estadísticas y las bases de datos.	Requiere el conocimiento de ciencias de la computación , estadística, matemáticas, conocimiento de las materias, IA / Machine Learning
Sinónimos	También se conoce como descubrimiento de conocimiento en bases de datos.	El análisis de datos es de varios tipos: exploratorio, descriptivo, análisis de texto, análisis predictivo , minería de datos, etc.
Perfil de trabajo	El especialista en minería de datos generalmente desarrolla algoritmos para identificar estructuras significativas en los datos. Un especialista en minería de datos sigue siendo un analista de datos con un amplio conocimiento del aprendizaje inductivo y la codificación práctica.	Un analista de datos generalmente no puede ser una sola persona. El perfil del trabajo implica la preparación de datos sin procesar, su limpieza, transformación y modelado y, finalmente, su presentación en forma de visualizaciones gráficas / no basadas en gráficas.
Responsabilidades	Es responsable de extraer y descubrir patrones y estructuras significativos en los datos.	Es responsable de desarrollar modelos, explicaciones, pruebas y proponer hipótesis utilizando métodos analíticos.
Salida	El resultado de una tarea de minería de datos es un patrón de datos	El resultado del análisis de datos es una hipótesis verificada o una visión de los datos.
Ejemplos	Una de las principales aplicaciones de la minería de datos se encuentra en el sector de comercio electrónico, donde los sitios web muestran la opción de "los que compraron esto también vieron"	Un ejemplo de análisis de datos podría ser el "estudio de series temporales del desempleo durante los últimos 10 años"

Fuente: Internet www.educba.com

3. Minería de datos

La minería de datos es el proceso de evaluar los patrones no reconocidos en los conjuntos de datos sin procesar de gran tamaño, según las diferentes perspectivas para clasificar los datos en información útil. (2)2 La minería de datos ayuda a encontrar información o conocimiento útil de un océano de datos. No hay valor para los datos hasta que los convierta en información valiosa. Es necesario analizar estos datos y convertirlos en información valiosa, por lo tanto, la minería de datos ayudará a extraer esta valiosa información de enormes conjuntos de datos disponibles.

El data mining es posible gracias a la evolución de los datos dentro de una organización, y su aplicación puede realizarse gracias a que existen tres pilares que sostienen su existencia:

•Recolección masiva de datos: hoy en día las organizaciones cuentan con una gran cantidad de datos gracias a sus circuitos organizacionales y la intercomunicación que les permite el entorno.

•Potentes computadoras con multiprocesadores: los avances tecnológicos en infraestructura y software permite la explotación de los grandes volúmenes de datos existentes.

•Algoritmos de Data Mining: Los algoritmos de Data Mining utilizan técnicas que han existido por lo menos desde hace10 años, pero que sólo han sido implementadas recientemente como herramientas maduras, confiables, entendibles que son más consistentes en sus resultados que métodos estadísticos clásicos.

Los otros procesos involucrados en Data Mining son:

(2) HAND D. J. Construction and Assessment of Classification Rules. J. Wiley Editorial (United States 1997)

- Limpieza de los datos: manejará datos irrelevantes, inexactos e incompletos.

- Integrando los datos: combina múltiples fuentes de datos en información significativa

- Selección de datos: los datos, que son significativos para el análisis, se recuperarán de la base de datos.

- Transformación de dato: convierte datos en formas específicas que son relevantes para la minería

- Minería de datos: extraerá los patrones de datos necesarios

- Evaluar los patrones en los datos: extraerá patrones que representan información o conocimiento dependiendo de medidas interesantes.

- Presentación de información o conocimiento: presentará el conocimiento extraído a la empresa utilizando diferentes visualizaciones con ayuda del data analitycs (análisis de datos).

De acuerdo a lo que menciona una definición general la minería de datos o exploración de datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, supone aspectos de gestión de datos y de bases de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

Simplificando dicho concepto se puede decir que la minería de datos es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos con el objetivo de encontrar patrones que puedan aportar información valiosa en la toma de futuras decisiones.

En la actualidad dicho concepto es mayormente mal utilizado, debido a que el mismo ha llegado en conjunto con la revolución tecnológica y se trata encajar en él cualquier forma de datos en gran escala o procesamiento de la información.

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación).

El proceso de minería de datos tiene normalmente cuatro etapas principales:

- Determinación de los objetivos

- Procesamiento de los datos

- Determinación del modelo

- Análisis de los resultados

En el primero de los pasos se trata el tipo de información que el cliente desea extraer de la base de datos. La segunda etapa es la que requiere más trabajo ya que se tiene de seleccionar, limpiar, enriquecer, reducir y transformar la base de datos que haya facilitado el cliente. Una vez que se procesa se encuentra lista para implementar el modelo que se haya escogido. Por último, el siguiente paso se centra en su interpretación.

Pasos para la explotación de Datos

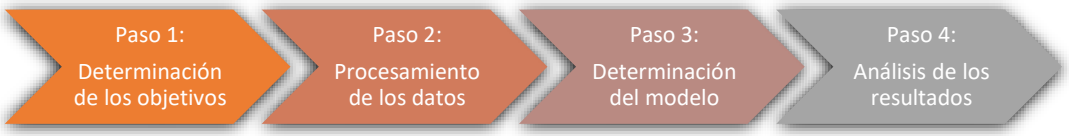


Figura N°2: Pasos para la explotación de datos. Fuente: propia

Esto se utiliza para muchos fines, según cada empresa y sus necesidades. Entre algunos de sus usos posibles se incluyen los siguientes:

- Pronósticos y riesgos: analizar datos para determinar el origen de desaciertos pasados (por ejemplo, la cantidad de clientes que con impagos) podría ayudar a un vendedor a tomar mejores decisiones sobre la estrategia de venta de préstamos y el perfil de los clientes morosos

- Agrupación: los datos proporcionados por los clientes les permiten a las empresas agrupar usuarios de muchas maneras; por ejemplo, demográficamente en función del sexo, la edad, los ingresos, el lugar en el que viven y sus hábitos de consumo. Esto les permite dirigirse eficientemente a los usuarios adecuados.

- Análisis de comportamiento: examinar los datos les permite a las empresas comprender el tipo de estímulos a los que los clientes responden. El análisis ayuda a determinar qué se puede hacer para evitar los comportamientos negativos de los consumidores que perjudican a su empresa.

La minería de datos o el KDD (Knowledge Discovery in Databases) se ha definido también como “el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles”. Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones. Para conseguirlo harán falta técnicas de aprendizaje automático, estadística, bases de datos, técnicas de representación del conocimiento, razonamiento basado en casos, razonamiento aproximado, adquisición de conocimiento, redes de neuronas y visualización de datos. Tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc.

Los datos recogen un conjunto de hechos y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). La explotación de datos involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros.

El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados.

4. Tecnologías de Apoyo para la Minería de datos:

Las tecnologías de apoyo para poder implementar el análisis de datos con Minería de datos, son importantes, debido que estas son el punto donde los datos se transforman.

Las más importantes entre estas tecnologías son los métodos estadísticos y el aprendizaje automático. Los métodos estadísticos han producido varios paquetes estadísticos para computar sumas, promedios, y distribuciones, que han ido integrándose con las bases de datos a explorar. El aprendizaje automático consiste en la obtención de reglas de aprendizaje y modelos de los datos, para lo cual a menudo se necesita la ayuda de la estadística.

Durante la búsqueda de herramientas que habrán de dar soporte a las nuevas formas de utilizar la información que nutren las organizaciones se deberá tener en cuenta el objetivo principal del trabajo a realizar y que es lo que se busca estudiar con dicha información, por lo que es interesante determinar cuáles serán los datos relevantes para el estudio y que permitan tomar las decisiones correctas. Es necesario entonces poder adaptar nuestros conocimientos para utilizar de manera razonable y provechosa tanto los datos como las nuevas tecnologías que darán apoyo.

Para la correcta implementación de la explotación de los datos se deberá tener en cuenta que la estadística es el nexo que vinculará directamente con los resultados y las conclusiones del caso de estudio, es por eso que es importante entender más en profundidad donde esta ciencia donde se une con su sabiduría y como llega a nutrirse de los datos para poder obtener información confiable. Adicionalmente se deberá garantizar que la recolección de los datos y la búsqueda de respuesta no esté dirigida a partir de las preguntas incorrectas. El uso de una técnica estadística sobre los datos inapropiados será tan en vano como utilizar los métodos incorrectos para el estudio del modelo.

A partir de lo antes mencionados es importante entender las siguientes pautas para poder realizar un trabajo de explotación de datos de calidad:

- Se debe delimitar el **poder y las restricciones** de las herramientas estadísticas de apoyo, estas son importantes ya que responderán al caso bajo estudio, si no se conoce respecto a qué tipo de problemática responden sus resultados, lo más probable que es que haya sido un trabajo de largas horas con la herramienta inadecuada.
- Tomar poder sobre los datos con la herramienta de apoyo como ser el software que procese estos que permita tomar el control de los análisis de datos, acceder a las últimas técnicas estadísticas, ahorrar tiempo, producir gráficos efectivos, generar trabajos reproducibles.

5. Técnicas utilizables del data mining

Cualquier problema o situación empresarial que necesite estudio y para el que existan datos históricos es un problema susceptible de ser tratado mediante técnicas de Minería de Datos.

A continuación, se detalla de forma meramente enunciativa algunas de las técnicas en las que se puede dar tratamiento a los datos históricos.

- Búsqueda de lo inesperado por descripción de la realidad multivariante: cuantas más variables se hayan obtenido para describir un fenómeno serán mejor, más ricas, más globales y más coherentes las descripciones y más fácil será detectar lo inesperado, esto es, aquello que no se había previsto y que resulta valioso para entender mejor el comportamiento de algún grupo de individuos, lo cual se ve favorecido por el hecho de trabajar con muestras grandes. Las muestras aleatorias son suficientes para describir la regularidad estadística global, pero no para detectar comportamientos particulares de subgrupos. En estos casos se pueden utilizar
    - *Análisis de Factoriales Descriptivos:* Permiten hacer visualizaciones de realidades multivariantes complejas y,

por ende, manifestar las regularidades estadísticas, así como eventuales discrepancias respecto de aquella y sugerir hipótesis de explicación.

- Búsqueda de asociaciones: un cierto suceso, ¿está asociado a otro suceso?, ¿se puede inferir que determinados sucesos ocurren simultáneamente más de lo que sería esperable si fuesen independientes?, ¿es posible determinar que un cliente potencial tendrá con un comportamiento similar a un cliente histórico de la compañía? Cuando es necesario responder a los interrogantes anteriores se puede optar las siguientes:
    - Técnicas de «clustering». Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

- Definición de tipologías: Los consumidores son, a efectos prácticos, infinitos, pero los tipos de consumidores distintos son un número mucho más pequeño. Detectar estos tipos distintos, su perfil y proyectarlos sobre toda la población, es una operación imprescindible a la hora de programar una política de ventas. Por otro lado, las tipologías no tienen que ser necesariamente de consumo, pueden ser de opiniones, valores, condiciones de vida, etc.;
    - Arboles de decisión. Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

- Detección de ciclos temporales: todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se sitúa cada consumidor ayudará a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

    - Series Temporales. A partir de la serie de comportamiento histórica, permite modelizar las componentes básicas de la serie, tendencia, ciclo y estacionalidad y así poder hacer predicciones para el futuro, tales como cifra de ventas, previsión de consumo de un producto o servicio, etc.

- Predicción: a menudo se deberá efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿llover ‘a mañana? Estas y muchas más son preguntas que se deberán responder, para ello se construirá un modelo a partir de los datos históricos. Si la variable de respuesta es continua (p. e. la rentabilidad de un cliente) se dirá que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) se concluirá que se trata de un problema de clasificación.

    - Redes bayesianas: Consiste en representar todos los posibles sucesos en que se está interesado mediante un grado de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

    - Previsión local: La idea de base es que individuos parecidos tendrán comportamientos similares respecto de una

cierta variable de respuesta. La técnica consiste en situar los individuos en un espacio euclídeo y hacer predicciones de su comportamiento a partir del comportamiento observado en sus vecinos.

Un enriquecimiento de las posibilidades de análisis son los sistemas híbridos, esto es, la combinación de dos o más técnicas para mejorar la eficiencia en la resolución de un problema, como, por ejemplo, utilizar un algoritmo gen ético para inicializar una red neuronal, o bien utilizar un árbol decisión como variable de entrada en una regresión logística.

CAPITULO III

BUSINESS INTELLIGENCE PARA EL PROCESO DECISORIO

Sumario: 1. Introducción al proceso de decisión 2. Conceptos generales del proceso decisorio 3. Calidad en la toma de decisiones 4. Objetivos del data mining y data analitycs en el proceso de decisión 5. Análisis de las consecuencias de la TTDD

1. Introducción al proceso de decisión

En el capítulo anterior se hizo hincapié en describir de lo que se trata el Business Intelligence y se destacó el hecho que este se encuentra conformado por subprocesos que de implementarse dentro de una organización puede lograr que el proceso decisorio se encuentre sustentado con información de calidad y disminuyendo en gran medida la incertidumbre, obteniendo análisis detallados y concisos respecto a la situación de la organización relevada a partir de la recolección de los datos a lo largo de todos sus procesos internos. Para darle secuencia al trabajo, durante este capítulo se abordará los conceptos e ideas de la relación entre el Business Intelligence y el proceso decisorio.

El análisis de los datos dentro de una empresa tiene el potencial de transformar la forma en que sus líderes toman las decisiones. Para lo cual es importante destacar no solo el potencial de la explotación y el análisis de datos para influenciar en el proceso decisorio, sino en cómo puede mejorar la calidad de las decisiones y, sobre todo, como desarrollar las habilidades para

aprovecharlo de una manera eficiente y eficaz. En este sentido, Bob McDonald, presidente de Procter & Gamble, afirmó: “He escuchado mucho sobre Big Data, pero no se trata de datos. Se trata de cómo somos capaces de utilizar los datos”.[3] (McDonald, 2013)

La información que genera una organización durante su interacción con el mercado nutre a la misma a partir de pequeñas partículas que se van integrando a cada uno de los procesos administrativos y productivo junto con las áreas de captación de estas partículas. Esto existe desde el día uno desde que se empieza a transitar por la vida de una empresa, sin embargo, la toma de decisiones sustentadas en la ciencia y el estudio de los datos ha ido evolucionando a lo largo de los años.

Desde hace varios años los directivos y gerentes a cargo de la toma de decisiones dentro de las organizaciones han ido trabajando sobre la información que circula por las arterias de los procesos administrativos core de una organización para poder tener conciencia de la realidad por la que transita la compañía día a día y poder lograr una mejor elección de las estrategias orientadas al crecimiento. En un principio el análisis de la información que se realizaba con el fin de responder preguntas simples, como “¿Cuántos han sido nuestras ventas en un periodo determinado?” “¿Cuáles han sido los productos que más se han vendido?” “¿Quiénes son nuestros clientes más rentables?” “¿Qué impacto ha tenido la inversión en la fuerza de mercadotecnia en las ventas del 1er semestre?”. Para responder estas tan solo se tenía en cuenta un fragmento de la información total que era obtenida en planillas de cálculo, sin ir más allá de un análisis univariado y aislado de los datos, estos tipos de análisis son llamados como forma tradicional de análisis de datos. Luego, las organizaciones se han ido convirtiendo en entidades más complejas por lo que no bastaba responder estas preguntas con análisis por encima de las relaciones de la información ni desestimando el valor de la

[3] MCDONALD, Bob. Three steps to analytics driven business. InformationWeek. (England 2013).

información que nutre todas las áreas, por lo que surge la idea de desarrollar herramientas para tomar decisiones impulsadas por datos y sus relaciones y es partir de este tipo de estudio que se empezó a realizar análisis estadísticos y matemáticos para tomar decisiones para llegar a enfocarse en una enorme, no estructurada y de rápido movimiento cantidad de información.

2. Conceptos generales del proceso decisorio

Cuando se hace referencia al proceso decisorio, se debe tener en cuenta que decidir es un proceso voluntario, sistemático, que, a través de un análisis subjetivo, en ejercicio del razonamiento y con la emoción propia del ser humano, obtiene la elección/acción de una alternativa (o curso de acción) para cumplir con los fines, objetivos, propósitos previamente definidos, clarificados y ponderados por el sujeto que se llamará decisor.

Decidir es seleccionar una alternativa entre un conjunto de alternativas. Esta selección está a cargo de un determinado sujeto, en un determinado momento, y luego de un proceso de reflexión. [4]

No existen decisiones sin emociones. Estas afectan las conductas humanas, provocan la revisión de objetivos, de creencias, del sistema de preferencias, además de desdibujar el escenario general que conforma el espacio decisorio. No hay un manual de procedimientos que pueda establecer todas las variantes posibles que puedan presentarse. Y esto se debe fundamentalmente a la gran incertidumbre que abunda en el mundo real.

El proceso de información es un proceso tendiente a disminuir la incertidumbre. La información es un conjunto de datos, con cierto significado, que provee conocimiento acerca de algún tema. Por lo que obtener información es adquirir conocimiento, es decir, implica aprender, y en principio, reducir la incertidumbre. La información es incertidumbre con signo contrario.

[4] HOOFT, A. Influencia del Big Data en el proceso decisorio, Revista Científica de UCES Vol. 24, N° 1. (Buenos Aires 2019) pág. 61-73.

Como una herramienta para ayudar a superar su limitación racional, la utilización del Data Analitycs o Data Mining definitivamente lleva al decisor a obtener mejores predicciones, y por consiguiente mejores predicciones producen mejores decisiones. En efecto, en razón el análisis disciplinado y científico de los datos, los administradores pueden medir, y por lo tanto conocer, radicalmente más acerca de sus negocios, y directamente trasladar ese conocimiento en mejores procesos de decisión y mejor desempeño.

En la era de la revolución de la información, la incertidumbre al parecer tendría poco espacio, sin embargo, la masividad de la información poco estructurada puede llegar a generar un mayor problema de incertidumbre o creer en que una alternativa es la correcta cuando dentro de su análisis se tomaron información incorrecta. Se puede disponer de gran cantidad de información, pero ella no siempre es confiable y muchas veces es contradictoria entre sí. Ante ella lo único que se puede hacer es seleccionar la información que realmente útil y dentro de ella la más confiable.

Para la implementación de la explotación, análisis de datos masivos y la inclusión de estos análisis dentro del proceso decisorio formal de una compañía, el directivo de una organización debe preguntarse durante la toma de una decisión importante ¿Qué dicen los datos? Y luego, ¿de dónde proviene esta información? ¿Qué tipo de análisis fueron hechos? ¿Cuán seguros estamos de los resultados? Estas preguntas ayudaran al decisor a entender mejor los resultados obtenidos del trabajo de los datos para obtener mejor información para la toma de decisiones, sin embargo, este debe permitirse a ser desautorizado por la evaluación de los datos y dejar que estos desaprueben su intuición.

Sin embargo, la dificultad empírica en la utilización de la explotación y la analítica de los datos hace necesario el desarrollo de capacidades cognitivas y dinámicas en quienes toman decisiones en las organizaciones. Las capacidades cognitivas de gestión en el nivel individual se refieren también

a la habilidad de usar el conocimiento que –como un recurso clave- afecta la forma en que se comportan los consejos de administración de las organizaciones. En ese nivel individual, los directores y administradores necesitan desarrollar las habilidades o capacidades cognitivas de gestión para percibir, analizar y procesar cambios en el ambiente. Estas capacidades cognitivas de gestión se refieren a la capacidad individual de los administradores de desempeñar actividades mentales.

Como la analítica de los datos provee nuevos conocimientos dentro de las tendencias en el ambiente, su uso puede mejorar también las capacidades dinámicas de las organizaciones al apoyar a quienes toman decisiones para adaptarse y responder rápidamente a las dinámicas demandas del ambiente. Como resultado las organizaciones necesitan integrar, construir y reconfigurar competencias para enfrentar los cambios en el ambiente que la explotación y análisis de los datos resalta.

3. Calidad en la toma de decisiones

A continuación, se hará hincapié en los ejes que influyen en la calidad dentro de la toma de decisiones:

a) Realidad compleja de la información que se genera en la organización. Cuando la información se vuelve más grande, más compleja y más inexplicable, la capacidad limitada de los humanos plantea dificultades en descifrar e interpretar la relación de la información disponible con el problema a solucionar. En estos casos el decisor trata de renunciar a la riqueza de los datos con el fin de hacerlo más limitado y acotar la solución del problema a la capacidad máxima que el dispone para el análisis de las alternativas (pocas e inexactas). En un principio se cree que se deben considerar como una limitación del entorno, sin embargo, a veces el decisor debe entender que las

decisiones en un entorno complejo deben tratarse de manera complejas para brindar un set de alternativas estratégicas para el éxito dentro de una organización.

b) La falta de conocimiento de la fuente de los datos influye en la calidad de la toma de decisiones. En párrafos anteriores se hizo hincapié que para la implementación del de data analitycs y data mining en una organización sea efectivo debe tener consistencia en los datos con los que se procesa el análisis, como así también se debe considerar que la elección de los modelos de análisis debe ser coherentes y concisos con los resultados de que propone obtener. El decisor debe conocer quiénes (o que) recolecta los datos de los circuitos de información de una organización y como los estudiosos de los datos le agregan valor a la información recolectada. Esto solo tendrá sentido siempre y cuando el decisor se involucre en todo el proceso y se encuentre atento al mismo durante todas las etapas, es decir, desde la captación, recolección y limpieza, hasta la generación de los modelos que soporten el análisis de la información.

c) Inconsistencias dentro de los modelos de análisis utilizado. Cuando se realiza la implementación del data analitycs y data mining, para el aprovechamiento completo de la información, se deberá tener en cuenta cada una de los métodos de análisis existentes y cuáles son los que mejores se adaptan a la problemáticas que se quiere resolver, es por esto que es necesarios que existan profesionales capacitados en la explotación de la información y el análisis de los datos por métodos estadísticos multivariantes, con el fin que estas herramientas sean utilizadas de manera correcta y sin caer en errores groseros por desconocimientos de los métodos y “oscurecer” la información.

4. Objetivos del *data mining* y *data analitycs* en el proceso de decisión.

El objetivo de cualquier iniciativa de la aplicación de nuevas herramientas para el análisis de la información es encontrar una ventaja competitiva mediante el procesamiento de los datos permitiendo maximizar los beneficios o bien reducir los costos. Con este motivo es interesante explicar cómo la unión entre los insights ofrecidos por los datos, estrategia organizacional y la generación de valor comprenden una trípode de apoyo de cualquier empresa que desee mantener el éxito en los años venideros. Para sustentar lo dicho se procede a leer el siguiente ejemplo:

"Una organización desea disminuir la tasa de incobrabilidad de sus clientes limitando los límites de créditos otorgados y disminuyendo los periodos de cobranzas. La toma de la estrategia fue fundamentada en el análisis de los montos totales que son adeudados a la fecha. Sin embargo, a partir del análisis de los clientes que se encuentran en situación de mora puede determinarse qué tipo de clientes cuentan con mayor probabilidad y poder reducir de la cartera los mismos, ya que si se tratara de manera general a los clientes (tanto como los que cumplen con los pagos a términos, como los que no) con las mismas medidas, la organización puede tener mayor impacto en la disminución de las ventas a créditos."

En el ejemplo enunciado se trató de destacar que el análisis de la información de manera correcta mantendría las relaciones con los clientes que cumplen y limitando el acceso al crédito a aquellos clientes quienes incurrirían en mora en los próximos meses.

A pesar que nuestro ejemplo parece ser simple y lógico, en ciertas ocasiones los análisis de manera rápida y las alternativas drásticas acaban con debilitar la vida organizacional y las relaciones con los clientes. Lo que se propone realizar es analizar la información existente respecto a los clientes y

los créditos otorgados, clasificándolos a aquellos que presentan mora y estudiando las características del mismo que pueden haber generado la incobrabilidad (temporalidad del crédito, tasas, condiciones laborales del cliente, etc.), es decir, relacionar las distintas variables con las que se distingue cada uno de los clientes y determinar si existe relaciones entre ellas y si estas terminan impactando en la condición de mora en el que el mismo se encuentra. Si los resultados obtenidos de dicho análisis permiten clasificar a los clientes, implementar con los mismos las medidas que le parezca correcta a la gerencia con el fin de disminuir la incobrabilidad ya sea antes del inicio de la relación o bien durante la misma.

Los tipos específicos de análisis de negocios incluyen:

- Análisis descriptivo, que realiza un seguimiento de los indicadores claves para comprender el estado actual de una empresa;
- Análisis predictivo, que analiza los datos de tendencias para evaluar la probabilidad de resultados futuros; y
- El análisis prescriptivo, que utiliza el rendimiento pasado para generar recomendaciones sobre cómo manejar situaciones similares en el futuro.

5. Análisis de las consecuencias de la TTDD

El análisis de los datos y su explotación permite conocer las consecuencias de similares medidas en el pasado. Esto aporta racionalidad a las decisiones, alimenta al analista con vistas ricas en información existente y con posibles análisis de sensibilidad.

El Business Intelligence(BI) puede precalcular comportamientos cercanos a las consecuencias de las posibles alternativas aportando racionalidad. Así como brindó facilidades para estudiar las circunstancias previas a una decisión, podrá analizar sus consecuencias.

La debida comunicación de decisiones y resultados favorecerá a toda la organización, transformará conocimiento tácito en explícito para el personal involucrado con la decisión y aplicarlo a sus esferas de actuación y decisión trabajando con mayor racionalidad.

Adquirir hábitos provenientes de experiencias en forma sistemática, permite decisiones más seguras y racionales pues el hábito integra conocimientos emergentes de las decisiones. Se sostienen que es una invalorable oportunidad para los profesionales informáticos que deberían ser el puente de contacto y promoción del intercambio entre las diferentes áreas organizativas. El avance e importancia del BI debe inducirlos a mejorar la escucha a sus usuarios y profundizar la calidad del modelado de datos como apoyo a la toma de decisiones, entendiendo los criterios de evaluación de alternativas. Deben aportar experiencia en la preparación de una estructura de datos eficiente para la construcción del Datawarehouse y la infraestructura soft que dará soporte a la gestión de los datos a lo largo de todo el proceso de información y las decisiones estratégicas.

CAPITULO IV

MODELOS ESTADÍSTICOS DE DATOS EN DATA MINING

Sumario: 1. Estadísticas en la revolución de los datos 2. Modelos estadísticos 3. El análisis multivariado aplicado en la analítica de negocios

1. Estadísticas en la revolución de los datos

La estadística en la revolución de datos tiene gran impronta ya que es una de las herramientas favoritas para dar soporte al análisis de la información y poder modelizar la complejidad de los datos, pues permite a su vez el diseño eficiente de la investigación hacia la búsqueda de las respuestas necesarias para mantener competente en el mercado a las organizaciones. Permite reconocer y explorar sobre la similitud entre los problemas planteados y las diferentes alternativas de análisis, mensurándolas y estimar el impacto de cada una de ellas, optimizando así el análisis de los resultados. Además, permite analizar la validez del resultado por medio de métricas que dan certeza de la calidad del análisis y si este es extrapolable a la realidad.

El análisis de los datos y su exploración, junto con la estadística, apuntan a trabajar juntas por encontrar patrones dentro de los datos que les permita entender la realidad por la que transitan las organizaciones.

La estadística siempre se encuentra relacionada con los datos, sin embargo, la relación no siempre es directa. Siempre se debe empezar con una pregunta que intenta responder, luego se debe entender los datos que se

encuentran disponibles y partir de allí determinar el modelo o los métodos estadísticos que se utilizaran para la explotación de los datos.

Para determinar el modelo adecuado a utilizar para el análisis de los datos se deben preguntar si es un problema de:

- Estimación
- Selección
- Discriminación
- Clasificación

La creciente evolución de los sistemas de comunicación de una organización lleva a una creciente complejidad de ésta y el ritmo acelerado de tales cambios evolutivos significa que los lideres deben tomar medidas para poder aprovechar esta gran disposición de información que les brinda el torrente diario de operaciones.

La explotación de los datos no es tratar a los mismos como registros históricos, sino más bien usarlos para poder lograr una mejor gestión de la incertidumbre y anticiparse a futuros acontecimientos. Una de la mayor debilidad competitiva de los grandes líderes es que sienten incapacidad a la hora organizar los datos para conseguir con ellos información de calidad.

Para poder enfrentar dicha problemática de datos dispersos y “alocados” corriendo a lo largo y a lo ancho de la organización, los líderes de la compañía deben utilizar la modelación de los datos.

En términos matemáticos, se habla de “modelar” debido a que se está creando un modelo, una reconstrucción simplificada de cómo funciona un proceso observado en el mundo real.

En un modelo de datos, siempre se tiene, al menos,

- Una variable resultante, siempre una sola, también llamada variable “dependiente”,

- Una o más variables predictores, también llamadas “explicativas”

El modelado de datos puede ser utilizado para dos propósitos:

- Predecir el valor de una variable resultante en base a valores conocidos de las variables predictores. Aquí no interesa tanto entender cómo es que las variables interactúan entre sí, o por qué lo hacen. Mientras las predicciones sean acertadas, o se acerquen lo suficiente, el modelo cumple su cometido.
- Explicar la relación entre una variable dependiente y todas las demás (las explicativas), buscando determinar si la relación es significativa.

Existen varios tipos de modelos estadísticos de datos que intentan simplificar la realidad y poder realizar un análisis más certero de los fenómenos a través de métricas conocidas. Dentro de estos se puede indagar en base a nuestro problema el mejor de los modelos que se ajusten a nuestros datos y que a partir de allí se pueda obtener información rica para tomar una decisión fundamentada en el comportamiento de las variables analizadas.

2. Modelos estadísticos

Los modelos estadísticos utilizan ecuaciones matemáticas para codificar información extraída de los datos. En algunos casos, las técnicas de modelado estadístico pueden proporcionar modelos adecuados de forma rápida.

A continuación, se detalla algunos de ellos

- Modelos de regresión lineal: os modelos estadísticos utilizan ecuaciones matemáticas para codificar información extraída de

los datos. En algunos casos, las técnicas de modelado estadístico pueden proporcionar modelos adecuados de forma rápida.

- Modelos de regresión Logística: es una técnica de estadístico para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal, pero toma un campo objetivo categórico en lugar de uno numérico.
- Análisis de componentes principales/ Factorial: busca combinaciones lineales de los campos de entrada que realizan el mejor trabajo a la hora de capturar la varianza en todo el conjunto de campos, en el que los componentes son ortogonales (perpendiculares) entre ellos. Análisis factorial intenta identificar factores subyacentes que expliquen el patrón de correlaciones dentro de un conjunto de campos observados. Para los dos métodos, el objetivo es encontrar un número pequeño de campos derivados que resuma de forma eficaz la información del conjunto original de campos.
- Análisis discriminante: realiza más supuestos rigurosos que regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos.

3. El análisis multivariado aplicado en la analítica de negocios

Estos datos pueden ser analizados y entendidos con estadística simple, uni- bi-variada, pero en muchos casos se necesitan técnicas estadísticas multivariante, más complejas, para convertir los datos en conocimiento.

La mayoría de los problemas reales son de naturaleza multivariada – lo que significa que existen múltiples variables que contribuyen a ellos. Los

patrones de comportamiento de dichas variables, generalmente, se encuentran determinados por un número de procesos que interactúan y varían en el tiempo y en el entorno que las genera.

El comportamiento de un cliente, por ejemplo, está afectado por varios factores, como, por ejemplo, el económico, sus gustos personales, sus necesidades, el lugar donde desarrolla sus actividades o vive, la sociedad y la cultura del que forma parte, las modas por la que transita, etcétera. Por esta multiplicidad de factores que interactúan en el comportamiento de un cliente ante una acción que pueda realizar que pueda comprometer a la organización, los vuelven difíciles de analizar y complejiza prever su comportamiento y su relación con la organización.

Intentar analizar cada uno de estos factores de manera aislada podría ocasionar que los resultados no sean válidos, ya que se trata de un análisis acotado y enfocado al estado situacional por el que transita el objeto observado.

Las variables relacionadas en mayor o menor grado, si cada variable se analiza aisladamente la estructura completa de los datos puede no ser revelada.

Los datos complejos requieren métodos de análisis que puedan hacer frente a múltiples variables simultáneamente, que no sólo revelen variables influyentes sino también la relación que dichas variables tienen entre sí para comprender completamente la estructura y las características clave de los datos.

El análisis multivariado pretende se reflexione sobre la verdadera naturaleza multidimensional de los datos, detectar y cuantificar las interrelaciones posibles entre las variables, explorando conjunto de datos complejos.

Algunos autores afirman que el propósito del análisis multivariante es medir, explicar y predecir el grado de relación de los valores teóricos (combinaciones ponderadas de variables). Es decir, el carácter multivariante reside en los múltiples valores teóricos (combinaciones múltiples de variables) y no sólo en el número de variables u observaciones. [5]

[5] HAIR, J. ANDERSON, R. TATHAM, R. BLACK, W. Análisis Multivariante, 5ta edición, Prentice Hall (Madrid 1999) pág. 4

CAPÍTULO V

ANÁLISIS LOGIT

Sumario: 1. Regresión logística, aproximamiento teórico 2. Concepto de ODDS o razón de probabilidad, ratio de ODDS y logaritmo de ODDS 3. Ajuste del modelo 4. Evaluación del modelo 5. Condiciones 6. Predicciones 7. Evaluación de ajuste y precisión del modelo

1. Regresión logística, aproximamiento teórico

La regresión logística forma parte del conjunto de métodos estadísticos que caen bajo denominación regresión y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple. Permite modelar un resultado binomial con una o más variables explicativas, pues mide la relación entre la variable dependiente categórica y una o más variables independientes mediante la estimación de probabilidades utilizando una función logística, que es la distribución logística acumulativa.

La regresión logística, también conocida como análisis logit, está restringida en su forma básica a dos grupos, aunque en formulaciones alternativas puede considerar más de dos grupos.[6]

Como su nombre ya indica, la regresión logística es una técnica de análisis de regresión. El análisis de regresión es un conjunto de procesos estadísticos que puede usar para estimar las relaciones entre variables. Más

[6] HAIR, J. ANDERSON, R. TATHAM, R. BLACK, W. Análisis Multivariante, 5ta edición, Prentice Hall (Madrid 1999), pág. 248

específicamente, utiliza este conjunto de técnicas para modelar y analizar la relación entre una variable dependiente y una o más variables independientes. El análisis de regresión le ayuda a comprender cómo cambia el valor típico de la variable dependiente cuando se ajusta una de las variables independientes y otras se mantienen fijas.

Como se indagó en el capítulo IV, hay varias técnicas de regresión. Puede distinguirlos observando tres aspectos: el número de variables independientes, el tipo de variables dependientes y la forma de la línea de regresión.

La predicción de una respuesta cualitativa para una observación puede denominarse clasificación de esa observación, ya que implica asignar la observación a una categoría o clase. Por otro lado, los métodos que a menudo se usan para la clasificación predicen primero la probabilidad de cada una de las categorías de una variable cualitativa, como base para hacer la clasificación.

La regresión lineal no es capaz de predecir la probabilidad. Si usa la regresión lineal para modelar una variable de respuesta binaria, por ejemplo, el modelo resultante puede no restringir los valores Y predichos dentro de 0 y 1. Aquí es donde entra en juego la regresión logística, donde obtiene un puntaje de probabilidad que refleja la probabilidad de ocurrencia en el evento.

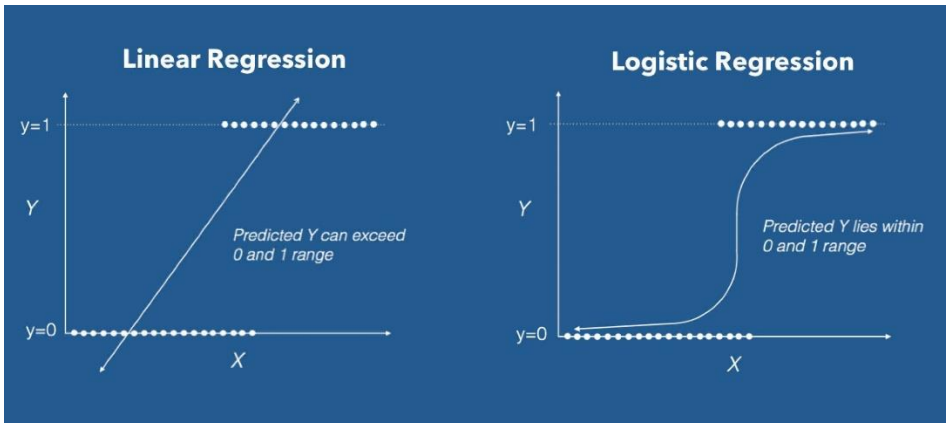


Figura N°3: Diferencia grafica entre regresión lineal y logística. **Fuente** www.hackerearth.com

La Regresión logística va a contestar a preguntas tales como: ¿Se puede predecir con anticipación si un cliente que solicita un préstamo a un banco va a ser un cliente moroso? ¿Se puede predecir si una empresa va a entrar en bancarrota?

En general, la regresión logística es adecuada cuando la variable de respuesta *Y* es politómica (admite varias categorías de respuesta, tales como mejora mucho, empeora, se mantiene, mejora, mejora mucho), pero es especialmente útil en particular cuando solo hay dos posibles respuestas (cuando la variable de respuesta es dicotómica), que es el caso más común.

Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta 0+\beta 1x$. El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

Para evitar estos problemas, la regresión logística transforma el valor devuelto por la regresión lineal (β0+β1X) empleando una función cuyo resultado está siempre comprendido entre 0 y 1. Existen varias funciones que cumplen esta descripción, una de las más utilizadas es la función logística (también conocida como función sigmoide):

función sigmoide=σ(x)=1/1+e^{-x}

Para valores de x muy grandes positivos, el valor de e^{-x} es aproximadamente 0 por lo que el valor de la función sigmoide es 1. Para valores de x muy grandes negativos, el valor e^{-x} tiende a infinito por lo que el valor de la función sigmoide es 0.

Sustituyendo la x de la ecuación 1 por la función lineal (β0+β1X) se obtiene que:

$$P(Y = k|X = x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} =$$

$$\frac{1}{\frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}} + \frac{1}{e^{\beta_0+\beta_1 X}}} =$$

$$\frac{1}{\frac{1+e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}}} =$$

$$\frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

donde *P(Y=k|X=x)* puede interpretarse como la probabilidad de que la variable cualitativa *Y* adquiera el valor *k* (el nivel de referencia, codificado como 1), dado que el predictor *X* tiene el valor *x.*

Esta función, puede ajustarse de forma sencilla con métodos de regresión lineal si se emplea su versión logarítmica, obteniendo lo que se conoce como LOG of ODDs

$$ln(\frac{p(Y=k|X=x)}{1-p(Y=k|X=x)}) = \beta_0 + \beta_1 X$$

2. Concepto de ODDS o razón de probabilidad, ratio de ODDS y logaritmo de ODDS

En la regresión logística se modela la probabilidad de que la variable respuesta *Y* pertenezca al nivel de referencia 1 en función del valor que adquieran los predictores, mediante el uso de LOG of ODDs.

Supóngase que la probabilidad de que un evento sea verdadero es de 0.8, por lo que la probabilidad de evento falso es de 1 - 0.8 = 0.2. Los ODDs o razón de probabilidad de verdadero se definen como la ratio entre la probabilidad de evento verdadero y la probabilidad de evento falso. En este caso los ODDs de verdadero son 0.8 / 0.2 = 4, lo que equivale a decir que se esperan 4 eventos verdaderos por cada evento falso.

La trasformación de probabilidades a ODDs es monotónica, si la probabilidad aumenta también lo hacen los ODDs, y viceversa. El rango de valores que pueden tomar los ODDs es de [0, ∞]. Dado que el valor de una probabilidad está acotado entre [0,1] se recurre a una trasformación logit (existen otras) que consiste en el logaritmo natural de los ODDs. Esto permite convertir el rango de probabilidad previamente limitado a [0,1] a [−∞, +∞].

Los ODDs y el logaritmo de ODDs cumplen que:

Si p(verdadero) = p(falso), entonces odds(verdadero) = 1

Si p(verdadero) < p(falso), entonces odds(verdadero) < 1

Si p(verdadero) > p(falso), entonces odds(verdadero) > 1

A diferencia de la probabilidad que no puede exceder el 1, los ODDs no tienen límite superior.

Si odds(verdadero) = 1, entonces logit(p) = 0

Si odds(verdadero) < 1, entonces logit(p) < 0

Si odds(verdadero) > 1, entonces logit(p) > 0

La transformación logit no existe para p = 0.

3. Ajuste del modelo

Una vez obtenida la relación lineal entre el logaritmo de los ODDs y la variable predictora *X*, se tienen que estimar los parámetros β0 y β1. La combinación óptima de valores será aquella que tenga la máxima verosimilitud, es decir el valor de los parámetros β0 y β1 con los que se maximiza la probabilidad de obtener los datos observados.

La identificación del mejor modelo de regresión logística se realiza mediante la comparación de modelos utilizando el cociente de **verosimilitud**, que indica a partir de los datos de la muestra cuanto más probable es un modelo frente al otro. La diferencia de los cocientes de verosimilitud entre dos modelos se distribuye según la ley de la Chi-cuadrado con los grados de libertad correspondientes a la diferencia en el número de variables entre ambos modelos. Si a partir de este coeficiente no se puede demostrar que un modelo resulta mejor que el otro, se considerará como el más adecuado, el más sencillo.

4. Evaluación del modelo

Existen diferentes técnicas estadísticas para calcular la significancia de un modelo logístico en su conjunto (p-value del modelo). Todos ellos consideran que el modelo es útil si es capaz de mostrar una mejora respecto

a lo que se conoce como modelo nulo, el modelo sin predictores, Interpretación del modelo

A diferencia de la regresión lineal, en la que *β1* se corresponde con el cambio promedio en la variable dependiente *Y* debido al incremento en una unidad del predictor *X*, en regresión logística, *β1* indica el cambio en el logaritmo de ODDs debido al incremento de una unidad de X, o lo que es lo mismo, multiplica los ODDs por $e^{\beta 1}$. Dado que la relación entre *p(Y)* y *X* no es lineal, *β1* no se corresponde con el cambio en la probabilidad de *Y* asociada con el incremento de una unidad de X. Cuánto se incremente la probabilidad de *Y* por unidad de *X* depende del valor de *X*, es decir, de la posición en la curva logística en la que se encuentre.

5. Condiciones

- Independencia: las observaciones tienen que ser independientes unas de otras.
- Relación lineal entre el logaritmo natural de odds y la variable continua: patrones en forma de U son una clara violación de esta condición.
- La regresión logística no precisa de una distribución normal de la variable continua independiente.

6. Predicciones

Una vez estimados los coeficientes del modelo logístico, es posible conocer la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello se emplea la ecuación del modelo:

$$\hat{p}(Y=1|X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

7. Evaluación de ajuste y precisión del modelo

Una de las principales aplicaciones de un modelo de regresión logística es clasificar la variable cualitativa en función de valor que tome el predictor. Para conseguir esta clasificación, es necesario establecer un threshold de probabilidad a partir de la cual se considera que la variable pertenece a uno de los niveles. Por ejemplo, se puede asignar una observación al grupo 1 si p^(Y=1|X)>0.5 y al grupo 0 si de lo contrario.

La regresión logística emplea todos los diferentes conjuntos de métricas. Aquí, se debe tratar con probabilidades y valores categóricos. Las siguientes son las métricas de evaluación utilizadas:

- Criterios de información de Akaike (AIC)

Puede ver AIC como contrapartida del cuadrado r ajustado en regresión múltiple. Es un indicador importante del ajuste del modelo. Sigue la regla: cuanto más pequeño, mejor. AIC penaliza el aumento del número de coeficientes en el modelo. En otras palabras, agregar más variables al modelo no permitiría que AIC aumentara. Ayuda a evitar el sobreajuste. Mirar la métrica AIC de un modelo realmente no ayudaría. Es más útil para comparar modelos (selección de modelos). Entonces, construya 2 o 3 modelos de regresión logística y compare su AIC. El modelo con el AIC más bajo será relativamente mejor.

- Desviación nula y desviación residual

La desviación de una observación se calcula como -2 veces la probabilidad logarítmica de esa observación. La importancia de la desviación se puede comprender mejor utilizando sus tipos:

Desviación nula y residual. La desviación nula se calcula a partir del modelo sin características, es decir, solo intercepción. El modelo nulo predice la clase a través de una probabilidad constante.La desviación residual se calcula a partir del modelo que tiene todas las características. En comparación con la regresión lineal, piense en la desviación residual como la suma residual del cuadrado (RSS) y la desviación nula como la suma total de los cuadrados (TSS). Cuanto mayor sea la diferencia entre la desviación nula y residual, mejor será el modelo. Además, puede usar estas métricas para comparar varios modelos: el modelo que tenga una desviación nula más baja, significa que el modelo explica la desviación bastante bien y es un mejor modelo. Además, disminuya la desviación residual, mejor el modelo. Prácticamente, a AIC siempre se le da preferencia sobre la desviación para evaluar el ajuste del modelo.

- Matriz de confusión

La matriz de confusión es la métrica más crucial comúnmente utilizada para evaluar los modelos de clasificación. El esqueleto de una matriz de confusión se ve así:

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

Como puede ver, la matriz de confusión evita la "confusión" al medir los valores reales y predichos en formato tabular. En la tabla anterior, Clase positiva = 1 y Clase negativa = 0. A continuación se muestran las métricas que se puede derivar de una matriz de confusión:

- Precisión: determina la precisión general prevista del modelo. Se calcula como Accuracy = (True Positives + True Negatives)/(True Positives + True Negatives + False Positives + False Negatives)
- Tasa positiva verdadera (TPR): indica cuántos valores positivos, de todos los valores positivos, se han predicho correctamente. La fórmula para calcular la tasa positiva verdadera es (TP/TP + FN). Además, TPR = 1 - False Negative Rate. También se conoce como sensibilidad o recall
- Tasa de falso positivo (FPR): indica cuántos valores negativos, de todos los valores negativos, se han predicho incorrectamente. La fórmula para calcular la tasa de falsos positivos es (FP/FP + TN). Además, FPR = 1 - True Negative Rate.
- Tasa negativa verdadera (TNR): indica cuántos valores negativos, de todos los valores negativos, se han predicho

correctamente. La fórmula para calcular la tasa negativa verdadera es (TN/TN + FP). También se conoce como especificidad.

- Tasa de falso negativo (FNR): indica cuántos valores positivos, de todos los valores positivos, se han predicho incorrectamente. La fórmula para calcular la tasa de falsos negativos es (FN/FN + TP).

- Precisión: indica cuántos valores, de todos los valores positivos predichos, son realmente positivos. Se formula como: (TP / TP + FP).

- Puntuación F: la puntuación F es la media armónica de precisión y recall. Se encuentra entre 0 y 1. Cuanto mayor sea el valor, mejor será el modelo. Está formulado como 2((precisión * recall) / (precisión + recall)).

- Característica del operador receptor (ROC)

ROC determina la precisión de un modelo de clasificación en un valor umbral definido por el usuario. Determina la precisión del modelo utilizando Área bajo curva (AUC). El área bajo la curva (AUC), también conocida como índice de precisión (A) o índice concordante, representa el rendimiento de la curva ROC. A mayor área, mejor modelo. ROC se traza entre la tasa positiva verdadera (eje Y) y la tasa positiva falsa (eje X). En este gráfico, nuestro objetivo es empujar la curva roja (que se muestra a continuación) hacia 1 (esquina izquierda) y maximizar el área bajo la curva. Más alta la curva, mejor el modelo. La línea amarilla representa la curva ROC en el umbral de 0.5. En este punto, sensibilidad = especificidad

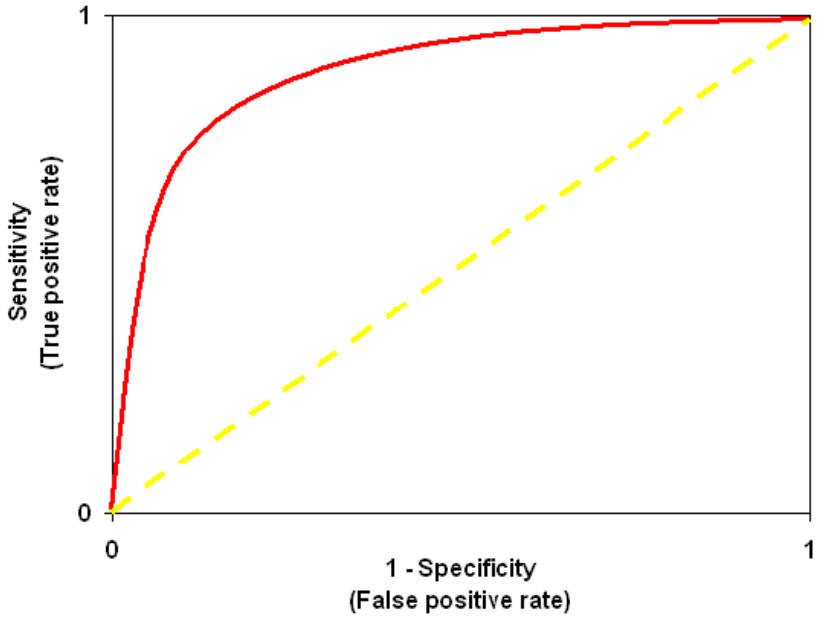


Figura N°4: Curva Roc. Fuente www.hackerearth.com

Las métricas mencionadas en los párrafos ayudarán a comprender las implicancias del modelo, y como este varía en función los umbrales de ocurrencias definidos para la comprensión más acotada del modelo. En el capítulo a continuación se realizará la aplicación del modelo.

CAPÍTULO VI

PRESENTACIÓN DE CASO DE APLICACIÓN PRACTICA

Sumario: 1. Introducción al caso de estudio 2. Exploración y análisis de la información disponible 3. Elección del modelo estadístico de análisis o 4. Preparación y exploración del dataset 5. Ejecución del modelo de regresión 6. Evaluación de ajuste del modelo 7. Clasificación y evaluación de las predicciones del modelo 8. Conclusiones del caso de aplicación.

1. Introducción al caso de estudio

Para poner en práctica lo estudiado en los capítulos anteriores se desarrollará un caso de estudio sobre una organización real de la Ciudad de San Miguel de Tucumán dedicada a la venta de microcréditos para el consumo destinados a la población en general. La misma se encuentra en el mercado de servicios financieros desde el año 2006, y desde sus inicios hasta la actualidad ha logrado acrecentar sus carteras de clientes y los distintos productos financieros ofrecidos. Cuenta con 5 sucursales en la provincia, emplazadas en dentro del microcentro tucumano y otras en localidades vecinas de la ciudad, como Tafí Viejo, Yerba Buena y el Manantial. Adicionalmente, la entidad cuenta con convenios con comercios que les permite ofrecer a los clientes financiación en el acto.

Durante un largo periodo de sus actividades la empresa mencionada realizó la venta de sus créditos de manera masiva con solo requisitos de identificación de sus clientes sin solicitar información respecto al comportamiento de pago o bien su capacidad de pago. Esta es una de las

razones por la que en la actualidad la empresa se encuentra en una reestructuración de la cartera de crédito a fin de disminuir el riesgo de incobrabilidad.

“La Financiera” cuenta con un sistema de información customizado, en el mismos se registran todas las ventas de créditos y son resguardas en una base de datos de la organización. Los datos disponibles a partir de la registración de las operaciones de la compañía son:

- Capital
- Préstamos otorgados
- Intereses
- Numero de cuotas
- Fechas de inicio del crédito
- Sexo del cliente al que se le otorga
- Situación laboral del cliente
- Edad
- Métodos de cobranzas
- Grado de mora del cliente
- Entre otros…

A los fines de dar provecho a los datos disponibles a la compañía, se precedió a realizar un trabajo de exploración y análisis de los mismos enfocados en la estrategia de reestructuración de la cartera de clientes y obtener alternativas de decisión analíticas con estimaciones más cercanas a la realidad expresada por la información que las operaciones revelan.

2. Exploración y análisis de la información disponible

A partir de los datos recolectados de la base de datos La Financiera, se realizó un estudio para seleccionar las variables que darán respuesta a la problemática planteada por la compañía.

En la primera etapa de la exploración de la información se correspondió con la limpieza de la información y al alineamiento de las variables. Es decir, se analizó el universo de datos contenido en las tablas extraídas desde las bases de datos y se eliminaron aquellas en las que contenían valores nulos, o bien aquellas variables que no sumaban al análisis deseado. La base original obtenida contaba con 46 variables, en las cuales se detallaba datos como: Banco, Tipo de notificación de resumen, liquidación del crédito, usuario que registró el préstamo, información personal del cliente, tipo de plan de pago, entre otras. De estas se extrajeron aquellas que contenía datos completos y que agregaban valor al análisis.

Las variables que se seleccionaron para el análisis son las siguientes:

- Capital: Capital otorgado en préstamo.
- Cuotas: Número de cuotas en la que se particionará el pago del total del crédito.
- Total: El valor total del capital más los intereses que conforman el préstamo.
- Alta: Fecha de alta del crédito.
- Situación laboral: situación laboral que acredita el cliente al momento del crédito.
- Sexo: Femenino o Masculino
- Edad
- Importe de cuota: Cuota parte del préstamo que el cliente paga mensualmente.
- Canal de cobro:
- Grado de mora: clasificación de los clientes en función a su grado de morosidad.

La variable “grado de mora” en la base original se trataba de una variable categórica la cual contenía varios niveles que se definían a partir del

número de meses que el cliente presentaba morosidad, a los fines de la aplicación del modelo se realizó una reclasificación de la variable, ya que es interesante interpretar a esta como una de los indicativos de la probabilidad del cliente de caer en impago y niveles de mora severa.

La herramienta para la exploración, preparación y análisis de los datos utilizada será R, cuyo lenguaje de programación permite utilizar modelos estadísticos multivariante a partir del procesamiento de grandes cantidades de datos, ya que la base de información resultante contiene 60170 observaciones, a partir de 6016 líneas y 10 variables de clasificación.

3. Elección del modelo estadístico de análisis

Para la elección del modelo estadístico a utilizar se hizo foco en responder una de las preguntas que “La Financiera” se encontraba planteándose a fin de mejorar la calidad de sus clientes y mantenerse rentable. La organización le interesaba saber cuáles eran aquellos clientes que tenían mayor probabilidad al caer en incobrabilidad, quienes luego de haber mantenido más de 6 meses de impago total de sus obligaciones con La Financiera se convierten en un costo para la compañía que se encuentra compuesto tanto por la deuda pendiente más el costo de ejecución legal de la deuda.

A partir de la observación de la estructura de los datos y pensando en dar respuesta a la problemática al que La Financiera deseaba responder, se decidió realizar un análisis de Regresión Logística, ya que la misma permitirá clasificar los datos en función las variables de entradas seleccionadas.

A continuación, se procederá a realizar una breve introducción teórica al modelo estadístico utilizado, continuando con la validación del mismo en el software Estudio.

4. Preparación y exploración del dataset

Como primer paso, se procede a cargar la base a analizar en el software estadístico Rstudio. En la figura debajo se muestra el comando que se ejecuta para levantar el dataset del archivo de origen de los datos. Una vez ejecutado el comando, el software de procesamiento muestra lo siguiente:

```
In [24]: base_orig1<-read.delim("clipboard", dec=",")

In [25]: str(base_orig1)

'data.frame':   6016 obs. of  11 variables:
 $ Capital          : num  15000 20000 5000 1500 5000 ...
 $ Cuotas           : int  15 18 12 6 6 9 6 12 12 12 ...
 $ Total            : num  35250 52400 9719 2208 7360 ...
 $ Alta             : Factor w/ 229 levels "1/1/2017","1/11/2017",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Situación_Laboral: Factor w/ 14 levels "Ama_de_Casa",..: 3 3 3 3 5 3 3 3 3 3 ...
 $ Sexo             : Factor w/ 2 levels "Femenino","Masculino": 1 2 2 1 2 2 1 1 2 2 ...
 $ Edad             : int  46 32 63 52 54 47 64 33 60 29 ...
 $ Importe_cuota    : num  2350 2911 810 368 1227 ...
 $ Canal_de_cobro   : Factor w/ 3 levels "CBU","Codigo",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Mora             : Factor w/ 2 levels "Moroso","No_moroso": 2 1 1 1 2 2 1 1 1 1 ...
 $ Grado_de_mora    : Factor w/ 3 levels "Incobrable","Moroso",..: 3 2 1 2 3 3 1 1 1 1 ...
```

La base cargada al sistema contiene un total de 6016 observaciones de 11 variables, ellas son:

- Capital
- Cuotas
- Total
- Alta
- Situación laboral
- Sexo
- Edad
- Importe de cuota
- Canal de cobro
- Mora
- Grado de mora

Como se puede observar en el resumen que muestra el comando, el data set cuenta tanto como variables numéricas como categóricas. Entre las variables categóricas se obtuvieron a Situación Laboral, Sexo, Canal de cobro

y Mora. Una vez cargado el dataset, se procedió a explorarlo para corroborar que el mismo no cuente con valores nulos.

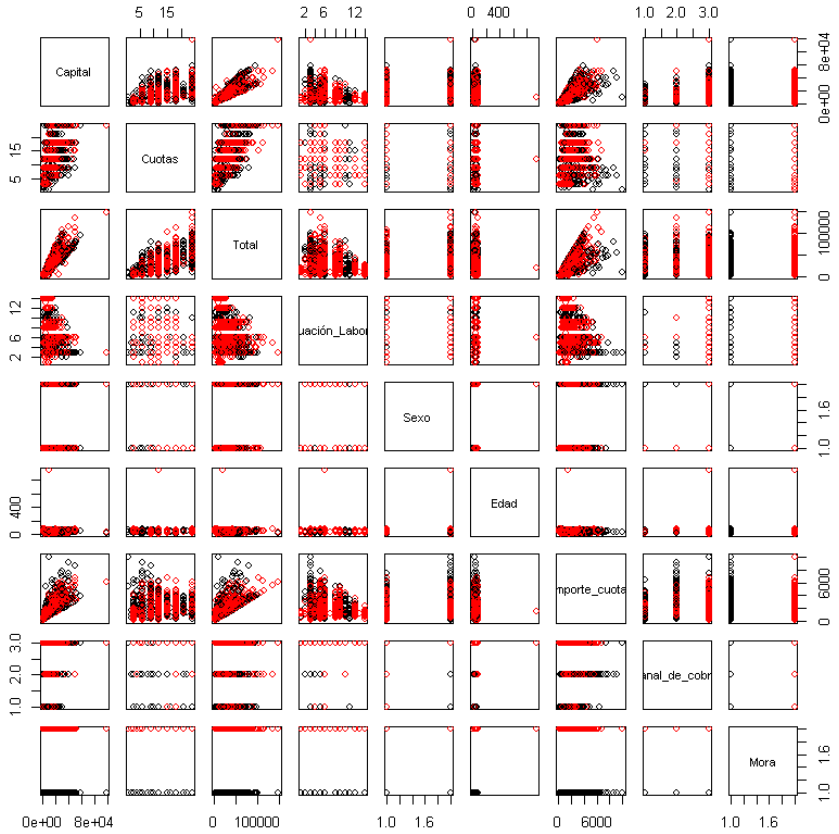
Para obtener un resumen de los datos y de algunas medidas descriptivas de las variables bajo análisis, se ejecutó la función Summary, de la cual se obtuvo el siguiente resultado:

```
In [46]: summary (base_sn)
```

```
    Capital            Cuotas          Total
 Min.   :  468.3   Min.   : 1.00   Min.   :   468.3
 1st Qu.: 8000.0   1st Qu.:12.00   1st Qu.: 12926.9
 Median :10101.0   Median :12.00   Median : 21010.1
 Mean   :13239.3   Mean   :12.27   Mean   : 25347.8
 3rd Qu.:16161.6   3rd Qu.:15.00   3rd Qu.: 33776.7
 Max.   :97722.7   Max.   :24.00   Max.   :147720.5

                    Situación_Laboral        Sexo           Edad
 Con_Recibo                   :2576    Femenino :2200   Min.   :  0.00
 Empleado_Privado             :1389    Masculino:3816   1st Qu.: 35.00
 Empleado_Publico             :1150                     Median : 43.00
 Titular_de_Tarjeta_de_Credito: 298                     Mean   : 44.23
 Empleado_Municipal           : 210                     3rd Qu.: 53.00
 Jubilado_a                   : 180                     Max.   :939.00
 (Other)                      : 213
 Importe_cuota            Canal_de_cobro        Mora
 Min.   :  156.1   CBU            :1187   Moroso   :3200
 1st Qu.: 1300.0   Codigo         : 941   No_moroso:2816
 Median : 1750.8   Pago_voluntario:3888
 Mean   : 2000.0
 3rd Qu.: 2496.8
 Max.   :10000.0
```

Como se observa en la imagen anterior se puede ver que las variables cualitativas pueden tomar varios valores dentro de ellas. Más adelante en el desarrollo del trabajo se transformarán estas variables categóricas para que puedan ser analizadas dentro del modelo elegido.

Adicionalmente, se realiza un análisis de correlación entre las variables.



La imagen anterior muestra que las variables con mayor correlación son Capital, Cuotas, Total, Importe de cuota y Mora, lo que tiene sentido, ya que estas son productos de las otras. Se tendrá en cuenta la correlación de estas variables dentro de la construcción del modelo para evitar que exista redundancia en los datos.

Para poder ejecutar el modelo de Regresión Logística, se procedió a transformar los valores de las variables categóricas en variables dummies, es decir, las categorías que las observaciones podían tomar se convertirán en variables binarias que podrán tomar valores entre cero y uno. Por ejemplo, la variable sexo se dividirá en dos variables femenino y masculino, estas tomaran el valor 1 cuando la variable tome el valor femenino o bien cero cuando tome el valor masculino.

```
In [62]: str(Base_final)

'data.frame':   6016 obs. of  20 variables:
 $ Capital                                          : num  15000 20000 5000 1500 5000 ...
 $ Cuotas                                           : int  15 18 12 6 6 9 6 12 12 12 ...
 $ Edad                                             : int  46 32 63 52 54 47 64 33 60 29 ...
 $ Situación_LaboralAma_de_Casa                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralAutónomo_a                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralCon_Recibo                      : int  1 1 1 1 0 1 1 1 1 1 ...
 $ Situación_LaboralEmpleado_Municipal              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralEmpleado_Privado                : int  0 0 0 0 1 0 0 0 0 0 ...
 $ Situación_LaboralEmpleado_Publico                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralIndependiente                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralJubilado_a                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralMonotributista                  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralSin_Recibo                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralSoldado_voluntarios             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralTaxista                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Situación_LaboralTemporarios_construccion_limpieza: int  0 0 0 0 0 0 0 0 0 0 ...
 $ SexoFemenino                                     : int  1 0 0 1 0 0 1 1 0 0 ...
 $ Canal_de_cobroCBU                                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Canal_de_cobroPago_voluntario                    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MoraMoroso                                       : int  0 1 1 1 0 0 1 1 1 1 ...
```

Como se puede ver en la imagen, el software nos convierte las variables categóricas en variables dummies por medio de una función, por lo que la base ya se encuentra lista para correr el modelo de regresión.

5. Ejecución del Modelo de Regresión

En R, se puede implementar Regresión logística utilizando la función GLM. Ahora bien, se debe comprender e interpretar los aspectos cruciales de los resultados que puede aportar esta función.

a) La estimación representa el valor de los coeficientes de regresión. Aquí, los coeficientes de regresión explican el cambio en el registro (probabilidades) de la variable de respuesta para un cambio de unidad en la variable predictora.

b) Std. error representa el error estándar asociado con los coeficientes de regresión.

c) El valor z es análogo al estadístico t en la salida de regresión múltiple. El valor z> 2 implica que la variable correspondiente es significativa.

d) El valor p determina la probabilidad de significación de las variables predictoras. Con un nivel de confianza del 95%, una variable con p <0.05 se considera un predictor importante. Lo mismo se puede inferir observando estrellas contra el valor p.

Una vez listo la data set en R y teniendo en cuenta los puntos expuestos en el párrafo anterior se procede a correr el modelo de regresión logística.

A continuación, se muestra la llamada a la función que se corrió. La variable respuesta en nuestra función es “MoraMorosos” esta variable, resultado de la transformación de las variables categóricas en dummies. La misma toma valor 1 cuando el cliente es moroso y 0 cero cuando se trata de un no moroso. Continuando con la definición de la función, las variables predictoras son las correspondientes a:

-Capital

- Edad

- Situación Laboral

- Sexo

-Canal de Cobro

```
In [25]: summary(model_rlogit)
```

```
Call:
glm(formula = MoraMoroso ~ Capital + Edad + Situación_LaboralAma_de_Casa +
    Situación_LaboralAutónomo_a + Situación_LaboralCon_Recibo +
    Situación_LaboralEmpleado_Municipal + Situación_LaboralEmpleado_Privado +
    Situación_LaboralEmpleado_Publico + Situación_LaboralIndependiente +
    Situación_LaboralJubilado_a + Situación_LaboralMonotributista +
    Situación_LaboralSin_Recibo + Situación_LaboralSoldado_voluntarios +
    Situación_LaboralTaxista + Situación_LaboralTemporarios_construccion_limpieza +
    SexoFemenino + Canal_de_cobroCBU + Canal_de_cobroPago_voluntario,
    family = binomial, data = Base_final)
```

```
In [25]: summary(model_rlogit)
                                                  z value Pr(>|z|)
(Intercept)                                         4.475 7.65e-06 ***
Capital                                             3.404 0.000664 ***
Edad                                               -0.970 0.332284
Situación_LaboralAma_de_Casa                        1.116 0.264324
Situación_LaboralAutónomo_a                         0.822 0.411218
Situación_LaboralCon_Recibo                        10.417  < 2e-16 ***
Situación_LaboralEmpleado_Municipal                 2.782 0.005402 **
Situación_LaboralEmpleado_Privado                   0.452 0.651563
Situación_LaboralEmpleado_Publico                   1.808 0.070574 .
Situación_LaboralIndependiente                     -0.020 0.983874
Situación_LaboralJubilado_a                         0.901 0.367807
Situación_LaboralMonotributista                     0.148 0.882735
Situación_LaboralSin_Recibo                         2.617 0.008858 **
Situación_LaboralSoldado_voluntarios                0.060 0.952276
Situación_LaboralTaxista                            2.249 0.024541 *
Situación_LaboralTemporarios_construccion_limpieza -0.020 0.983994
SexoFemenino                                       -1.040 0.298343
Canal_de_cobroCBU                                   4.530 5.89e-06 ***
Canal_de_cobroPago_voluntario                     -25.473  < 2e-16 ***
```

El resultado de la función del modelo de regresión logística muestra cómo se obtiene la probabilidad de que un cliente, en función de las variables predictoras.

6. Evaluación de ajuste del modelo

Como se observan en los resultados, el p value para algunas de las variables es mayor a 0,05, por lo que dichas variables no son significativas para el análisis de regresión. Si se ejecuta la función GLM sin esas variables, el valor de AIC será menor, por ende, el ajuste del modelo será mejor. El AIC del modelo es igual a 5763,7

```
In [25]: summary(model_rlogit)
Situación_LaboralJubilado_a                         0.901 0.367807
Situación_LaboralMonotributista                     0.148 0.882735
Situación_LaboralSin_Recibo                         2.617 0.008858 **
Situación_LaboralSoldado_voluntarios                0.060 0.952276
Situación_LaboralTaxista                            2.249 0.024541 *
Situación_LaboralTemporarios_construccion_limpieza -0.020 0.983994
SexoFemenino                                       -1.040 0.298343
Canal_de_cobroCBU                                   4.530 5.89e-06 ***
Canal_de_cobroPago_voluntario                     -25.473  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8315.4  on 6015  degrees of freedom
Residual deviance: 5725.7  on 5997  degrees of freedom
AIC: 5763.7

Number of Fisher Scoring iterations: 13
```

Se probará ejecutar la función GLM, teniendo en cuenta la información que aporta el estadístico, a fin de probar que el ajuste del modelo mejora en función de eliminar el ruido de las variables predictoras no significante para el modelo.

```
In [20]: summary(model_rlogit1)

Call:
glm(formula = MoraMoroso ~ Capital + Situación_LaboralCon_Recibo +
    Situación_LaboralEmpleado_Municipal + Situación_LaboralEmpleado_Publico +
    Situación_LaboralSin_Recibo + Situación_LaboralTaxista +
    Canal_de_cobroCBU + Canal_de_cobroPago_voluntario, family = binomial,
    data = Base_final)
```

Como se aprecia en la figura, la función que se ejecuta en este nuevo modelo de regresión logística solo tiene en cuenta las variables que en el anterior modelo tenían un valor p menor a 0,05. Como, por ejemplo, Capital, Canal de cobro, etc.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7178  -0.6140   0.2175   0.6851   2.0208

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.066e+00  1.176e-01   9.064  < 2e-16 ***
Capital                           1.304e-05  3.704e-06   3.520 0.000431 ***
Situación_LaboralCon_Recibo       1.838e+00  8.244e-02  22.298  < 2e-16 ***
Situación_LaboralEmpleado_Municipal  6.545e-01  2.100e-01   3.117 0.001829 **
Situación_LaboralEmpleado_Publico    2.493e-01  1.066e-01   2.339 0.019317 *
Situación_LaboralSin_Recibo       1.439e+00  5.499e-01   2.617 0.008877 **
Situación_LaboralTaxista          7.342e-01  3.281e-01   2.237 0.025256 *
Canal_de_cobroCBU                 6.979e-01  1.462e-01   4.774 1.81e-06 ***
Canal_de_cobroPago_voluntario    -3.034e+00  1.148e-01 -26.426  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8315.4  on 6015  degrees of freedom
Residual deviance: 5735.8  on 6007  degrees of freedom
AIC: 5753.8

Number of Fisher Scoring iterations: 5
```

Como se puede ver el AIC paso desde 5763,7 a 5753,8, con lo que es fácil decir que se logró mejorar el ajuste del modelo extrayendo las variables poco significativas del modelo.

A partir de la lectura del resultado del modelo se puede decir entonces, por ejemplo, que el algoritmo de odds de que un cliente sea moroso está asociado al monto de capital otorgado en el préstamo siendo esta relación significativa de acuerdo por el p-value expresado (0,00432<0,05).

7. Clasificación y evaluación de las predicciones del modelo

En esta instancia será importante evaluar la manera en la que el modelo de regresión que obtuvo como resultado se comporta comparado con lo realmente observado. A continuación, se realizará una clasificación tabular donde se comparará los valores predichos frente a los observados.

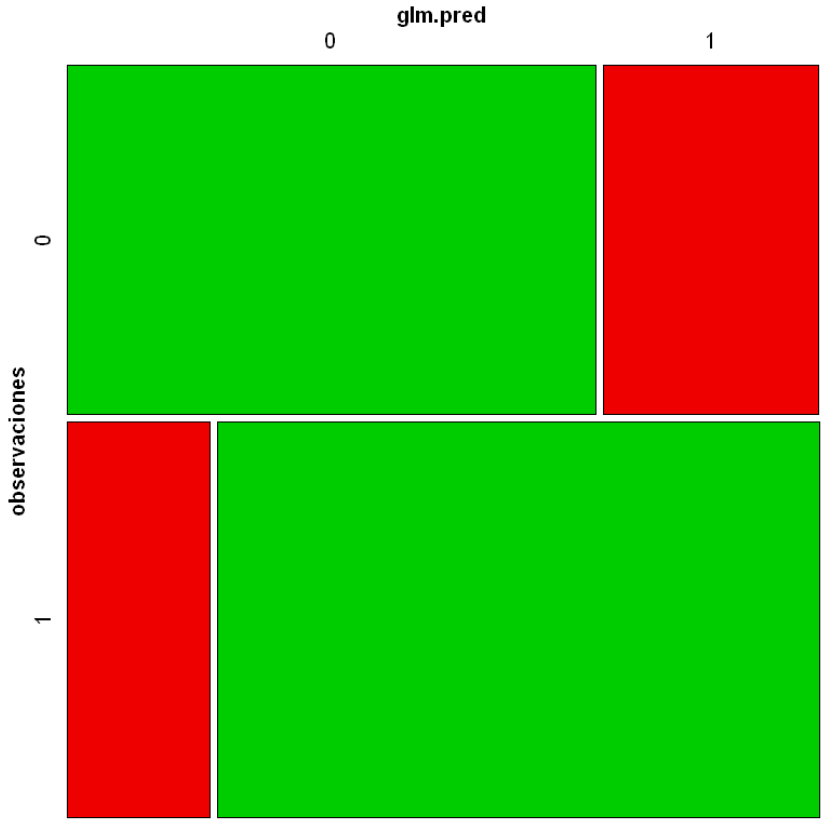


Figura N°5: Matriz de confusión del modelo de regresión Logística

Para obtener la matriz de confusión se deben calcular las probabilidades de ocurrencia de que un cliente sea moroso o no y calcular los valores predichos por este, definiendo un umbral de decisión que se determina una probabilidad mínima a partir de la cual se realiza una clasificación positiva que la observación.

```
In [21]: Modelogit.probs <- predict(model_rlogit1,type = "response")

In [22]: glm.pred <- ifelse(test = model_rlogit1$fitted.values > 0.5, yes = 1, no = 0)

In [23]: matriz_confusion <- table(model_rlogit1$model$MoraMoroso, glm.pred,
                                   dnn = c("observaciones", "glm.pred"))
```

En el caso de estudio se determinó como umbral 0,5 en la primera prueba, para la cual se obtiene la siguiente matriz.

```
In [27]: matriz_confusion

                glm.pred
observaciones    0    1
            0 1953  863
            1  584 2616
```

A partir de la matriz de confusión se puede obtener un par de métricas para entender mejor lo que el modelo de regresión desea revelar. A partir de esto se puede discernir entre los casos bien clasificados y los que fueron erróneamente clasificados por el modelo

De acuerdo con lo estudiado en el capítulo VI, se puede calcular métricas a partir de la matriz de confusión, de los cual se deberá tener en claro los valores True negative, False Positive, False Negative y True Positive.

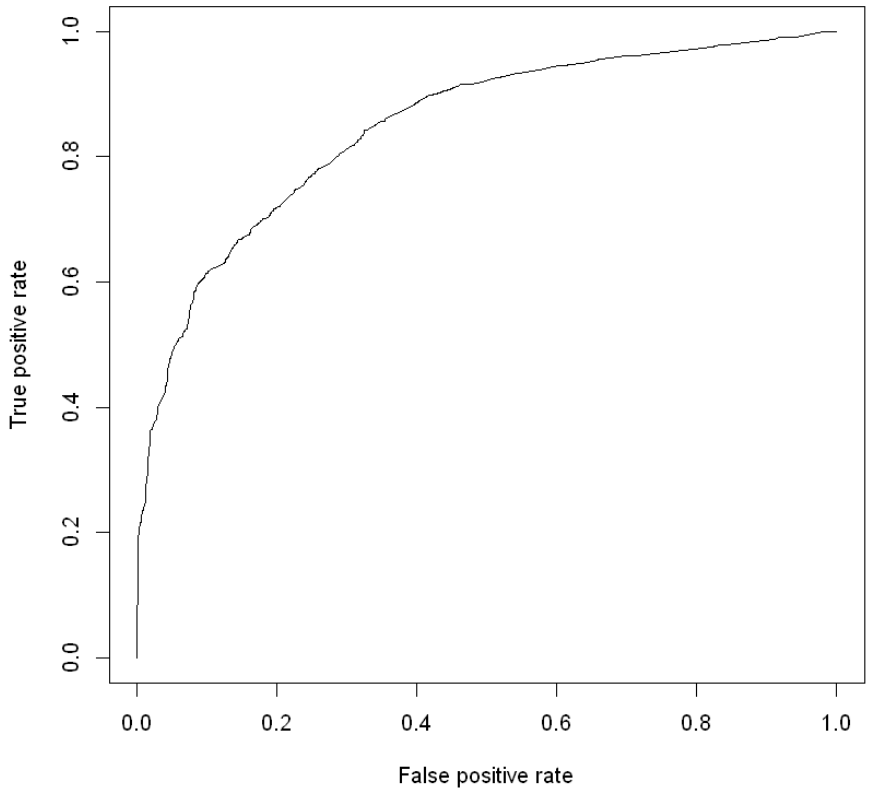
En la matriz de confusión obtenida a partir de un umbral de decisión del 0,5 se obtiene lo siguiente:

- True Negative: 1953
- False Positive: 863
- False Negative: 584
- True Positive: 2616

La primera métrica interesante de entender:

Metrics	Modelo Thr 0,5
Accuracy	0,759474734
Recall (TPR)	0,8175
Precisión	0,751940213
True Positive Ratio (TPR)	0,8175
False Positive Ratio (FPR)	0,306463068
AUC	0,755518466

El Accuracy da una proporción de los datos predichos de manera correcta, mientras que el recall o sensibilidad hace referencia respecto a la proporción de positivos correctamente predichos, esta medida es sensible a los falsos negativos. En cuanto a la medida de presión indica la proporción de verdaderos positivos en función al total de predicciones positivas. El AUC es el área bajo la curva ROC que se presenta a continuación.



Para poder visualizar las distintas medidas de clasificación que puede tomar el modelo en función al cambio de los umbrales de decisión se procedió a probar las métricas para dos niveles de decisión distintos.

Ahora se analizará la matriz para un umbral de 0,6

```
Number of Fisher Scoring iterations: 5

In [21]: Modelogit.probs <- predict(model_rlogit1,type = "response")

In [40]: glm.pred <- ifelse(test = model_rlogit1$fitted.values > 0.6, yes = 1, no = 0)

In [41]: matriz_confusion <- table(model_rlogit1$model$MoraMoroso, glm.pred,
                                   dnn = c("observaciones", "glm.pred"))
```

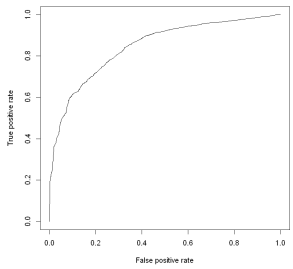
```
In [42]: matriz_confusion
                glm.pred
observaciones    0    1
            0 2562  254
            1 1278 1922
```

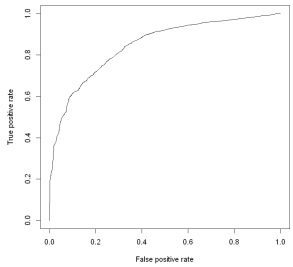
Por lo consiguiente el análisis de la matriz para un umbral de decisión mayor a 0,4 se obtiene lo siguiente:

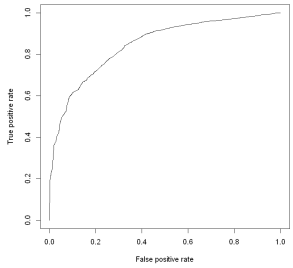
```
In [21]: Modelogit.probs <- predict(model_rlogit1,type = "response")

In [49]: glm.pred <- ifelse(test = model_rlogit1$fitted.values > 0.4, yes = 1, no = 0)

In [50]: matriz_confusion <- table(model_rlogit1$model$MoraMoroso, glm.pred,
                                   dnn = c("observaciones", "glm.pred"))
```

```
In [51]: matriz_confusion
                glm.pred
observaciones    0    1
            0 1633 1183
            1  323 2877
```

Metrics	Modelo Thr 0,5	Modelo Thr 0,6	ModeloThr 0,4
Accuracy	0,759474734	0,745345745	0,749667553
Recall (TPR)	0,8175	0,600625	0,8990625
Precisión	0,751940213	0,883272059	0,70862069
True Positive Ratio (TPR)	0,8175	0,600625	0,8990625
False Positive Ratio (FPR)	0,306463068	0,090198864	0,420099432
AUC	0,755518466	0,755213068	0,739481534







Como se puede apreciar en la tabla de métricas calculadas, las medidas como accuracy y AUC (área bajo la curva ROC) no han tenido variaciones significativas de acuerdo al cambio de los umbrales, esto se debe a que la proporción de predicciones correctas es aceptable, sin embargo, los valores que se castigan más o menos el cambio de umbrales son los siguiente:

- Recall/Sensibilidad: en la tabla se observa que cuando se aumenta la probabilidad umbral la medida disminuye en contraste al valor que se obtiene cuando se baja la probabilidad mínima a 0,4. Es decir que, a mayor umbral, mayor número de predicciones Falsas Negativas, por ende, menor es la proporción de correctos predichos.

- Precisión: Al contrario, a la medida anterior, se observa que la métrica aumenta a mayor umbral, es porque esta es sensible a los Falsos Positivos. Es decir que esta es mayor dado que a mayor umbral de decisión, menor presencia de Falsos positivos y mayor precisión.

- En cuanto al FPR se observa que para los distintos niveles de umbrales de decisión muestra menor o mayor valoración y esto se explica a partir que esta toma la presencia de falsos positivos. Lo ideal sería que el FPR tienda a 0 mientras que el TPR tienda a 1. El modelo que más se aproxima a este ideal es que toma un umbral del 0,5.

- AUC: al observar las gráficas ROC generadas a distintos umbrales de decisión se puede inferir que el nivel de predicción de verdaderos positivos es buena en los tres umbrales, sin embargo, se puede observar en el gráfico, que el área bajo la curva tiene mayor capacidad cuando el umbral de decisión toma el valor a 0,5, es decir que el modelo predecirá un verdadero positivo solo cuando este tenga una probabilidad de ocurrencia mayor a 0,5.

8. Conclusiones del caso de aplicación

Por medio de las métricas, como se puede apreciar, la evaluación del modelo es más clara y la elección de un umbral se corresponderá con la estrategia que se desea abordar para dar respuesta a la problemática. En el caso de la Financiera, sería importante definir entre dos decisiones en cuanto a la ejecución del modelo, ya que podrían tomarse tanto un umbral de decisión a los fines de maximizar el número de predicciones verdaderas positivas y minimizar las falsas negativas.

En este punto se deberá pensar en lo siguiente, La Financiera desea tener un mayor monitoreo sobre los clientes que tienen mayor probabilidad de morosos, según el modelo de regresión, los clientes que caen en mora están explicados significativamente por las variables de situación laboral que indican “con recibo”, “sin recibo”, “empleado público”, “taxista” y “municipal”, como así también las variables canales de cobro y capital. En estos pilares La Financiera debe asentar su estrategia de reducir su tasa de incobrabilidad, ya que esto tiene sentido, el análisis de los clientes parece ser escasa y poco estructurada a la actualidad, el cliente beneficiario del préstamo parece estar a apto a recibirlo con tan solo presentar identificación y el préstamo le es otorgado con intereses acorde al nivel de riesgo, lo que hace que el monto total del mismo se acrecenté y se vuelva contraproducente para el cliente pagar en los plazos que establece.

Más allá de la política de otorgamientos de La Financiera, las métricas del modelo pueden ayudar a decidir frente los objetivos de su problemática. Si la entidad de crédito desea tomar un rol conservador y definir como umbral de decisión 0,4 y disminuir en gran medida el número de predicciones falsas negativas, es decir cuando el modelo predice que el cliente no es moroso, pero si lo es, entonces el modelo castiga con mayor número de predicciones falsas positivas, es decir que clasifica como moroso un cliente pagador. En esta circunstancia, la entidad crediticia correría el riesgo de perder ventas rentables con tal de disminuir el riesgo de incobrabilidad. En estas circunstancias La Financiera debería evaluar el otorgamiento de los préstamos en base a los intercepto de la función de regresión y el umbral de decisión que le permita reducir las predicciones falsas negativas sin perder clientes pagadores que le agregan rentabilidad al negocio.

CONCLUSIONES GENERALES

A lo largo del trabajo se desarrollaron los distintos conceptos que llaman a la minería y análisis de los datos como herramientas fundamentales para la gestión de la información para la toma de decisiones de forma eficaz y eficiente. Se inició el mismo con una breve explicación de cómo estás herramientas pueden dar soporte al proceso de decisión hasta como los modelos estadísticos toman protagonismo para la aplicación de análisis de datos (entre muchas de otras técnicas). La estadística multivariada y la utilización de modelos cuantitativos ofrecen al científico de los datos pruebas y métricas de cómo estos se comportan y como pueden hacerlo en el futuro. Además, la evidencia del caso de aplicación muestra como la ejecución de la analítica por medio de un modelo de regresión sustenta cuantitativamente determinadas clasificaciones de los clientes, obteniendo mayor información respecto a si estos pueden adaptarse o no a los objetivos de la entidad de crédito estudiada.

En breves párrafos se concluye que:

- La era digital ha echado raíces en la época en la que se transita y las organizaciones debe nutrirse de estas para lograr el crecimiento. La potencialidad de las infraestructuras tecnológicas y la rapidez con la cual el flujo de información corre en todas direcciones en

las organizaciones debe proveer a la dirección la base para toda decisión, por lo que es importante entender que se debe trabajar de manera oficial en la mejora constante de las tecnologías de la información y en su infraestructura como punto de partida.

- La infraestructura tecnológica no debe ser lujosa y brillante porque si, el objetivo es dar soporte a la información generada por las distintas áreas y que esta pueda ser resguarda, consultada y estudiada con facilidad.

- Se debe formar a aquellos que darán apoyo en las tareas de explotación y análisis de los datos, ya que estos en su gran mayoría son complejos y ante la masividad, deben ser tratados de manera cuidadosa para que no se pierda calidad en la información de salida.

- Poder de interpretación y análisis. Cuando se trata de la aplicación de herramientas de explotación y análisis en los datos se debe entender que no es tarea fácil. El uso de datos masivos para la mejora en la calidad de la toma de decisiones de una organización requiere de tiempo y dinero, y de estos dos el tiempo es la estrella. El análisis de datos y modelos complejo, lleva tiempo de estudio, de cuestionamientos y pruebas, por lo que es importante preparar un equipo con los conocimientos suficientes, que entienda los objetivos de la información y entienda las problemáticas que se desea abordar con minería y análisis de los datos.

- Conocimientos técnicos en el área. Las personas que integren el equipo que se ocupara de la recolección, análisis e interpretación de los datos debe estar formado con los conocimientos técnicos suficientes con la finalidad de no incurrir en demoras innecesarias o quitar calidad a la información por ignorancia de los modelos utilizados.

- Predisposición a la utilización de los resultados y a dejar que los datos “estudien” las decisiones, más allá de la experiencia

profesional del decisor y sin sesgo del mismo. Esto no quiere decir que el directivo/gerente decisor no deba cuestionar la información brindada por los datos, sino más bien que este debe tener apertura para analizar las nuevas alternativas que antes no las haya considerado dado los límites del análisis humano.

Como se puede observar, las conclusiones se basan en que la organización debe replantear su estructura de información y alinearse a una estrategia donde el centro de atención se posa sobre los datos y la información que nutre la compañía, manteniendo estándares de calidad sobre los mismos y garantizar así mayor grado de certeza durante el proceso decisorio y por ende aumentar las posibilidades de crecimiento de la misma.

ÍNDICE BIBLIOGRAFICO

A) General

HAND, D. J. Construction and Assessment of Classification Rules. J. Wiley Editorial (United States 1997)

HAIR, J. ANDERSON, R. TATHAM, R. BLACK, W. Análisis Multivariante, 5ta edición, Prentice Hall (Madrid 1999)

HERNANDEZ SAMPIERI, Roberto BAPTISTA LUCIO, Pilar y FERNANDEZ- COLLADO, Carlos. Metodología de la Investigación. 5ta Edición. Editorial McGraw-Hill, (México 2010).

HOOFT, A. Influencia del Big Data en el proceso decisorio, Revista Científica de UCES Vol. 24, N° 1. (Buenos Aires 2019)

RAGHUNATHAN, S. Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis. Decision Support Systems, CRC Press Editorial, (United States 2010).

B) Especial

ALUJA, T. La mineria de datos entre la estadística y la inteligencia artificial, Universitat Politecnica de Catalunya. (España 2001)

HAND, D. J. (1998). Data Mining: Statistics and more? The American Statistician, 52, 2, 112-118

JANSSEN, M., VAN DER VOORT, H. y WAHYUDI, A. Factors influencing big data decision-making quality. Journal of Business Research, 70. (2017).

KOSCIELNIAK, H. y PUTO, A. BIG DATA in decision-making processes of enterprises *Procedia Computer Science*, 65. (2015)

MCAFEE, A. y BRYNJOLFSSON, E. Big Data: The Management Revolution. *Harvard Business Review*, (2012).

MCDONALD, Bob. Three steps to analytics driven business. InformationWeek. (England 2013).

C) Otras Publicaciones

Consultas a base de información, en internet: www.educba.com (11/09/2019), www.rpubs.com (27/09/2019), www.bitsandbricks.github.io (29/11/2019), www.hackerearth.com (29/09/2019), www.powerdata.com (8/10/2019), www.isotools.org (18/10/2019) www.digitalhouse.com (24/10/2019)

ÍNDICE