Master in Management + Analytics

Descubriendo oportunidades de Cross-Sell en el segundo evento de compras en SAP

Autoría: Peloso, Joaquín
Fecha de defensa de la tesis: 2023

UNIVERSIDAD TORCUATO DI TELLA

MASTER IN MANAGEMENT + ANALYTICS

DESCUBRIENDO OPORTUNIDADES DE CROSS-SELL EN EL SEGUNDO EVENTO DE COMPRAS EN SAP

TESIS

Joaquín Peloso

Mayo 2023

Tutor: Tomás Tetzlaff

Resumen

SAP es una compañía de origen alemán con un gran peso en la industria del software que ha llegado a tener una enorme cantidad de productos a lo largo de su trayectoria. A estos productos se los conoce también como soluciones, y no son otra cosa más que módulos de software dedicados al manejo de distintas áreas de una compañía que pueden venderse por separado, dando lugar a la posibilidad de vender nuevas soluciones a compañías ya clientes. Este concepto también se lo conoce como cross-sell.

Los datos comerciales sugieren que la posibilidad de explotar el potencial de compra de cada cliente no está del todo aprovechado. La mayor cantidad de compañías clientes tienen una sola solución comprada que podría fácilmente permitir una integración hacia otras soluciones del ecosistema. Debido a la inexistencia de una iniciativa similar por parte de la compañía, el presente trabajo se basa en hallar aquellas soluciones que mejor se adecuan a las necesidades de los clientes con mayor potencial de cross-sell, es decir las compañías con una sola solución comprada. Con el fin de lograr tal objetivo, se desarrollan varios modelos predictivos que permiten obtener las probabilidades de venta exitosa de una determinada solución distinta a la ya comprada. Por ejemplo, dada una compañía que ya cuenta con una solución y que tiene ciertas características e historial de compra, se determina la probabilidad que tiene de obtener una solución distinta.

Para alcanzar el objetivo del trabajo, se usan tanto datos comerciales de SAP como también datos propios de las compañías cliente. Es gracias a estos datos que se completa un análisis clasificatorio sobre características y patrones de compra en común entre clientes, independientemente de la cantidad de compras que hayan tenido. Una vez comprendidas cuáles son aquellas variables mas relevantes para el análisis, se procede al armado de varios modelos XGBoost que permiten encontrar las probabilidades de éxito en una venta de tal forma que se formule una recomendación para el sector de planeamiento comercial. La misma está formada por las tres probabilidades más altas junto con la solución correspondiente.

Abstract

SAP is a company of German origin with more than 50 years in the software market that has developed a large number of products throughout its history. These products are also known as solutions and are nothing more than software modules dedicated to the management of several company areas. Due to their module nature, a company can acquire an independent quantity of modules to suit its needs.

Since nothing similar has been accomplished yet, this work goes through how several predictive models were developed in order to obtain the probability of which solution, different from the one already purchased, can be sold to a company that has already purchased an SAP product before. For example, given a company that already has a single solution and has certain characteristics and purchase history, this work will provide the likelihood of purchasing a different solution. As a result, it is expected to have a considerable commercial impact as it would allow cross-sell opportunities to be identified more easily.

The analysis of the data and the development of a classification method resulted in the creation of one predictive model per solution, and from the positive outcome of the same, to choose the best sales probabilities. In other words, by obtaining the probabilities of purchase of the solutions by a company, those in the top 3 of highest probabilities are the recommendations for the commercial sector. Thus, an average accuracy of 72% among the 23 models with average AUC values of 0.81 was achieved. This resulted in several correctly validated recommendations.

Índice

1. Introducción

Las compañías de software siempre tuvieron el fin de crear soluciones para facilitar la administración de una compañía en todas sus áreas, como ser logística, finanzas, facturación, desarrollo de IT y bases de datos, inteligencia de negocios, recursos humanos, y gestión de gastos internos de la compañía, entre otros. Así es como existen varias compañías con un gran catálogo de soluciones que tienen varios puntos a favor y en contra con el objetivo de diferenciarse las unas de otras. Por su parte, gracias al gran volumen de soluciones desarrolladas y disponibles, SAP hizo posible la generación de un ecosistema propio de soluciones creado a partir de la combinación de sus productos. En términos comerciales, esta combinación es la llamada cross-sell, en donde se logra que una compañía ya cliente compre una solución que pertenece a un área distinta de la que ya se tiene. Un claro ejemplo seria que un cliente que posee una solución dedicada a la logística e inventario decida adquirir una solución dedicada al área de recursos humanos de la misma compañía. Es así como se logra que los clientes se mantengan dentro de un mismo ecosistema con varias soluciones compatibles entre ellas.

Por otro lado, es importante entender no sólo el papel que juegan y los beneficios que traen las oportunidades de cross-sell en el ámbito comercial, como ser la generación de una base de clientes fieles, un ecosistema completamente compatible entre sí que permite desarrollar funcionalidades especificas al sistema y que terminan por atraer a mayor cantidad de compañías, sino también cómo generar estas oportunidades. Con anterioridad no se desarrolló un algoritmo que permita detectar cuales son los clientes más propensos a comprar soluciones de otras áreas, lo que causaba que el área comercial se dirija de la misma manera a todas las compañías, independientemente de sus características más importantes como ser la industria a la que pertenecen, la región en la que están ubicadas, el tamaño de la compañía, y principalmente su historial de compras con SAP. En otras palabras, se le da el mismo trato comercial a una compañía que es cliente hace 20 años y tiene unas 10 soluciones, que a otra compañía cliente desde hace 1 o 2 años y solo tiene 1 solución. Esto es lo que hace que muchas de las oportunidades de cross-sell no estén bien aprovechadas.

Esta tesis está estructurada de manera que en el comienzo se explica el contexto y el significado de varios conceptos estrechamente relacionados con los datos utilizados, así como también el posible impacto de este trabajo dadas las circunstancias. Luego, se muestra y explican los datos sobre los cuales esta tesis está basada. En esta sección no sólo se explican los datos en sí, sino que también se ahonda en algunos puntos interesantes de los mismos a la vez que se desarrolla sobre la limpieza y normalización de éstos. En el siguiente paso se exploran y combinan los datos, dando lugar a nuevas variables que ayuden al desarrollo de los modelos.

Una vez entendidos y transformados los datos, se otorga una explicación sobre el enfoque y modelos elegidos, explicando en detalle cuáles fueron las razones. Antes de la conclusión del trabajo, también se muestran los resultados de los modelos y cómo se traducen en términos comerciales.

SAP

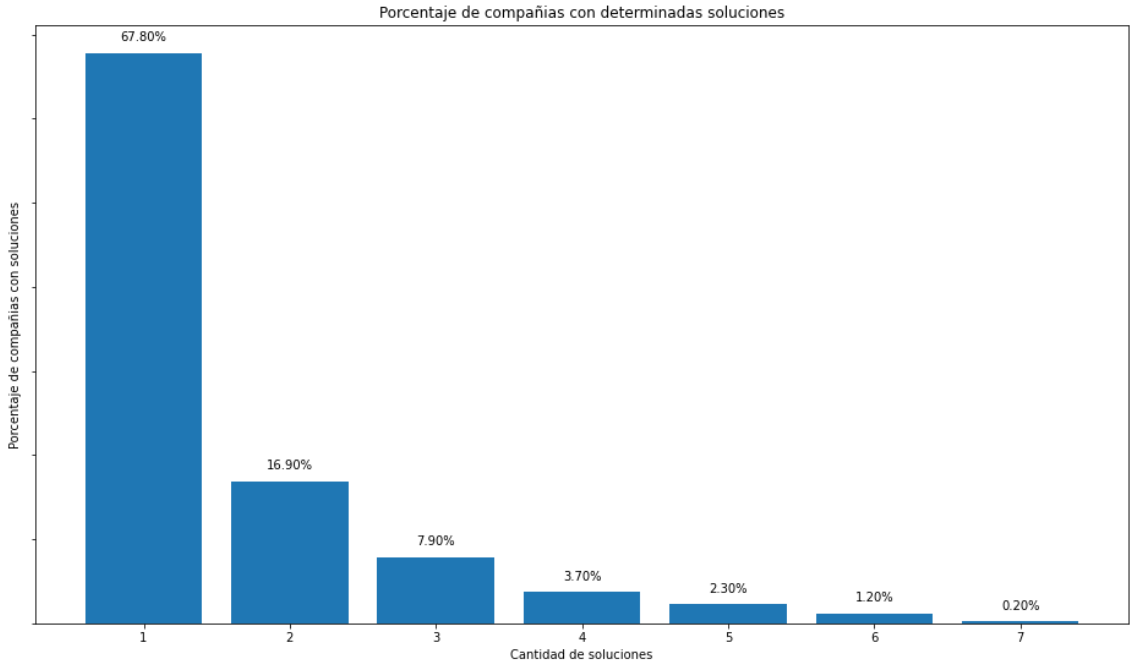
Fundada en 1972, SAP (Systemanalyse und Programmentwicklung, traducido al inglés como System Analysis and Program Development) es una compañía multinacional de origen alemán que hoy cuenta con más de 100.000 empleados a nivel global. Uno de los mayores logros de SAP fue establecer las bases para lo que hoy se conoce como un sistema ERP (Enterprise Resource Planning) que permite facilitar la gestión de múltiples áreas de una empresa, como ser compras, producción, manejo de inventario y materiales, ventas, marketing, finanzas, y recursos humanos, sin importar su tamaño e industria. Hoy, la compañía cuenta con una inmensa cantidad de soluciones y componentes que se ajustan a las necesidades de varias empresas en distintos escenarios.

Explorando el concepto de Cross-Sell

Antes de conocer más en detalle sobre las soluciones que SAP ofrece, es menester comprender en su totalidad el concepto de Cross-sell en este trabajo. Al tratarse de una compañía que abarca una gran cantidad de áreas empresariales con sus soluciones, el objetivo es alcanzar al mayor número de compañías y dentro de éstas, sus áreas. Lograr que una compañía obtenga una solución significa que es un potencial cliente para abrirse a otras áreas y adquirir otras soluciones. Es por esto por lo que un uso de datos descriptivos de cada cliente junto con sus respectivos historiales de compra, podrían probar ser de suma importancia para entender las necesidades y encontrar oportunidades de ventas en ciertos clientes. Es así como pueden hallarse oportunidades de cross-sell, es decir alcanzar con otros productos a clientes ya activos, que pueden hallar interés en alguna solución determinada.

En la figura 1 se puede ve el porcentaje de compañías con determinado número de soluciones, es decir cuantas tienen una o más soluciones.

Figura 1: Porcentaje de compañías con determinado número de soluciones



Concepto de Motor de Recomendaciones

El algoritmo que se desarrollará en este trabajo es una variante de las técnicas de aprendizaje automático conocida como Motor de Recomendaciones. Este algoritmo de recomendaciones, como todos los ya conocidos, utiliza técnicas de aprendizaje automático para proporcionar sugerencias o recomendaciones personalizadas a usuarios, que en este caso se trata del área comercial de SAP. El objetivo principal de un sistema de recomendación es predecir las preferencias de un usuario y utilizar esa información para ofrecerle recomendaciones relevantes.

Este algoritmo tendrá como propósito predecir con cierto grado de precisión cual sería la siguiente compra hecha por el cliente dadas las variables que se tienen en cuenta. En otras palabras, se buscarán patrones en común entre clientes similares para hacer una recomendación comercial a un vendedor o vendedora.

Para evaluar cada una de las compañías y lograr obtener una recomendación basada en cada una de las sub-soluciones, se debe desarrollar un modelo predictivo por sub-solución que nos permita conocer las probabilidades por cada producto. En otras palabras, al obtener por cada una de las compañías cliente el resultado de 22 predicciones basadas cada una en un producto, es posible determinar cuál es la sub-solución con la mayor probabilidad y finalmente llevar a cabo esa recomendación de venta como la recomendación principal. Es posible también otorgar otras recomendaciones que le siguen con el fin de tener más información.

Las variables que el motor de recomendaciones considera para dar un resultado son la industria, dados por el código internacional SIC que permite identificar y clasificar cualquier tipo

de actividad comercial; la región, como ser Latinoamérica, Asia y Pacifico, Europa, etc.; el segmento asignado internamente por SAP, basándose principalmente en el tamaño de la compañía; y el estado de la implementación de la solución determinada por el uso del producto, que puede ser activa, en proyecto, o inactiva.

Sistemas de recomendación orientados a la compra y venta

Los sistemas de recomendación son muy usados en todas las industrias de distintas maneras, pero siempre con el mismo fin, mantener al usuario dentro de la plataforma y ofrecerle productos que puedan de serle de interés. Algunos ejemplos de sistemas de recomendación enfocados en la venta son los de compañías de e-commerce como Amazon, eBay o Alibaba. A modo de encontrar ejemplos similares y comparables con el Motor de Recomendaciones propuesto en este trabajo, se analizarán en mayor profundidad algunos de los sistemas de recomendación de otras plataformas.

Caso Amazon

El sistema de recomendación de Amazon se basa en gran medida en el análisis de las compras y el comportamiento de navegación de los usuarios. Utiliza varias técnicas, entre ellas el filtrado colaborativo y el filtrado basado en contenido. El primero se basa en analizar los patrones de comportamiento de compra de los compradores y encuentra similitudes entre ellos para hacer recomendaciones. En otras palabras, si un cliente ha comprado productos similares a los de otro cliente, entonces se le puede recomendar los productos que el comprador con comportamiento similar ya ha comprado y que este aún no ha adquirido.

El filtrado basado en contenido se utiliza para recomendar productos similares a los que el usuario ha comprado o ha mostrado interés. Si un cliente ha comprado un producto específico, el sistema buscará otros productos que sean similares en términos de categoría, características o propiedades. Además, Amazon también usa técnicas de análisis de texto y minería de opiniones para proporcionar recomendaciones basadas en reseñas y comentarios de otros usuarios. Podría resumirse en que Amazon se basa en ofrecer recomendaciones altamente personalizadas y relevantes para cada usuario.

Caso eBay

eBay se diferencia un tanto de Amazon en su modelo de negocios. Mientras que Amazon tiene un enfoque mucho más basado en tiendas minoristas y la venta de productos nuevos, eBay se concentra principalmente en la compraventa de usuario a usuario, es decir que se encuentran principalmente productos que individuos particulares venden y no de compañías. Aunque no parezca, esto requiere modificar un sistema de recomendación como el de Amazon ya que, al tratarse de una plataforma de compraventa entre individuos, es necesario tener en cuenta la

popularidad de los usuarios y interacciones entre los mismos. Si bien utiliza técnicas de filtrado colaborativo, también incorpora características específicas del vendedor considerando la reputación y la reseña de ellos para brindar recomendaciones más relevantes al comprador. El sistema de recomendación de eBay se centra más en productos populares y en tendencia, basados en la actividad general de la plataforma y las preferencias de los usuarios en términos de categorías y productos específicos.

Caso Alibaba

El sistema de recomendación de Alibaba es muy similar al de Amazon, sin embargo, el enfoque de este es distinto debido al tipo de negocio. Alibaba tiene una gran presencia internacional y su principal negocio esta en el comercio internacional, lo cual hace que el sistema de recomendación tenga que también tomar algunas variables firmográficas de los compradores como ser la región, país, y los distintos tiempos de envío que se ajustan a los servicios disponibles. Esto también dificulta hallar patrones en común entre compradores dispersos por todo el mundo, lo que resulta en que el sistema ponga un mayor peso en los productos relacionados y no tanto en historial de compras.

Influencia en el Motor de Recomendaciones

A partir de los sistemas de recomendación mencionados anteriormente, se llega a la idea de explorar un sistema de recomendaciones aplicado al área comercial. Debido al uso de distintos tipos de filtrados y datos históricos por parte otras compañías, tiene sentido plantear un sistema que use los datos históricos para encontrar patrones de compra en común entre clientes con el fin de hallar su próxima compra.

En primer lugar, un filtrado colaborativo tiene lógica en un escenario en donde se cuenta con datos históricos de compras de compañías, desde la primera compra hasta la última. De esta manera es también posible encontrar patrones respecto a las características de los clientes. En segundo lugar, un filtrado por contenido tiene lugar en este trabajo gracias a conocer exactamente cuál fue la solución comprada por el cliente, de forma que se pueden encontrar las correlaciones entre ciertas características de los clientes con su primera compra.

Finalmente, al tratarse de compañías que pueden estar en cualquier lugar del mundo, es importante también incluir variables relacionadas a la ubicación y a la industria de los clientes, dado que podrían influir en sus siguientes compras.

Concepto de Orden y Evento de Compra

Como parte de las variables independientes que se usarán en este trabajo con el fin de lograr un algoritmo que determine las probabilidades que existen que un cliente con Enterprise Management compre otra solución distinta, se halla el concepto de Orden y Evento de Compra.

Es vital para el análisis diferenciar a aquellos clientes que comenzaron su camino con SAP por medio de la solución Enterprise Management y a aquellos que decidieron elegir otra solución. Esta diferenciación es a la que se le llamara Orden de Compra, teniendo en cuenta el orden del evento de compra. Es decir que no impactará en la probabilidad de la misma manera una compañía cuyo historial de compra podría ser EM → BTP → HXM, que una compañía cuyo historial de compra podría ser HXM → CX → EM.

Además del concepto de Orden de Compra también se encuentra el concepto de Evento de Compra. El mismo hace referencia a cuántas compras hizo el cliente independientemente de la cantidad de productos. Es decir, que un cliente podría tener en su base instalada EM, BTP y HXM, pero siendo sus Eventos de Compra solo dos, por ejemplo, EM → BTP y HXM. Las compras de dos o más soluciones se suelen llamar paquetes ya que en un mismo evento de compra hubo varias soluciones adquiridas.

Variables Independientes

Las variables independientes con las cuales se procura predecir los eventos de cross-sell son cuatro: (1) Industria de la compañía cliente (variable categórica), (2) Compras pasadas de la compañía (variable categórica), (3) Tamaño de la compañía (variable numérica, que da lugar a segmentos en forma de categorías) y (4) Región (variable categórica).

El impacto de las industrias en el análisis

Como parte de las variables independientes a tomar en cuenta, se encuentra la industria de la compañía a analizar. Estas industrias están determinadas por el sistema de nomenclatura Standard Industrial Classification (SIC) que permite clasificar a las compañías en distintos rubros según su principal actividad. A su vez, SAP clasifica cada SIC en Master Codes (MC) para una mayor agrupación de estas industrias en un nivel más alto. Debajo esta el detalle de cada uno de los MC disponibles.

- Aerospace and Defense
- Automotive
- Banking
- Chemicals
- Consumer Products
- Industrial Machinery and Components
- Healthcare
- High Tech
- Insurance

- Oil and Gas
- Life Sciences
- Public Sector
- Retail
- Telecommunications
- Utilities
- Sports & Entertainment
- Media
- Passenger Travel & Leisure
- Higher Education and Research
- Professional Services
- Cargo Transportation & Logistics Service
- Wholesale Distribution
- Defense and Security
- Engineering, Construction and Operation
- Postal
- Mill Prod. & Mining
- SAP Consolidated companies
- Nonclassifiable Est.

Es lógico que una compañía que pertenece a determinada industria pueda no tener las mismas necesidades que una compañía que pertenece a otra, por ejemplo, una compañía con foco en Retail o Wholesale Distribution probablemente busquen con mayor necesidad soluciones relacionadas a la cadena de suministros y operaciones.

El impacto de las compras pasadas en el análisis

Otra de las variables independientes que se usaran para el análisis es el historial de compra de un cliente. Esta es la variable más importante ya que permite hallar un patrón en común en el comportamiento de los clientes sobre SAP. Al conocer las compras pasadas de las compañías, junto con otras características, es posible entender si alguna solución está relacionada a otra en mayor o menor medida, o si una solución esta mayormente relacionada con alguna industria o tamaño de compañía. En otras palabras, si un conjunto de compañías luego de comprar la solución A decide adquirir la solución B, es probable que esto se vea influenciado por alguna de característica que tengan en común. Quizás las compañías más grandes son las que más necesitan una solución para su departamento de RR. HH., por ejemplo.

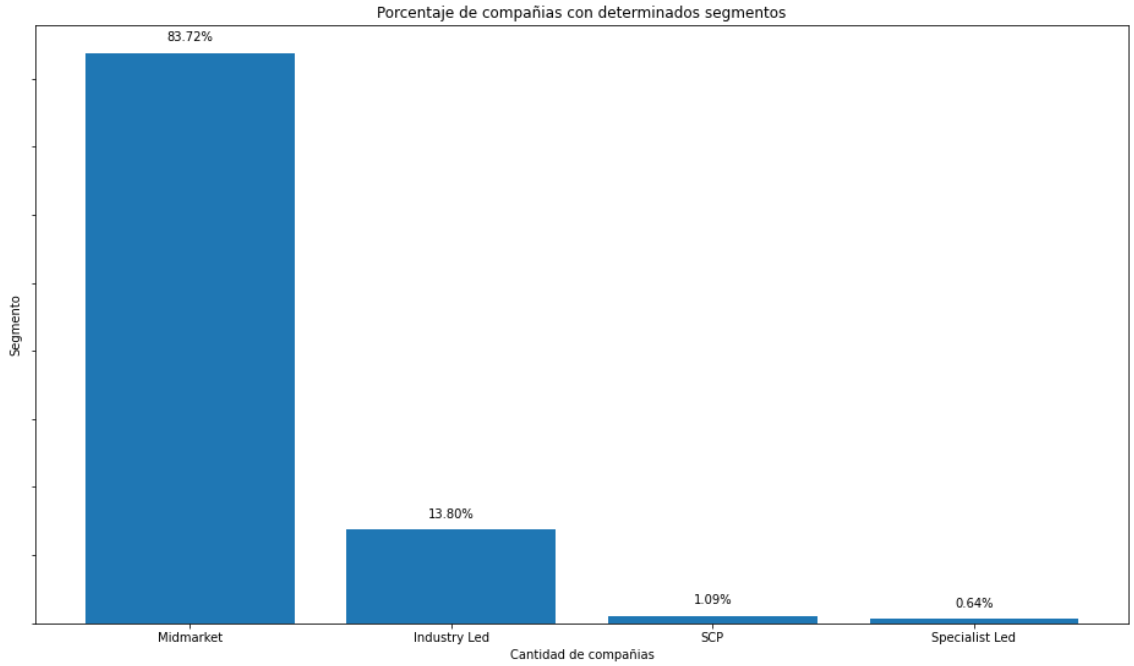
A su vez también se puede determinar la influencia del orden y evento de compra, ya que quizás es más común que una o dos soluciones sean las primeras en ser compradas para luego pasar a una tercera y/o cuarta solución en ese camino. Como se corroborará en los datos, puede suceder que una gran parte de la base instalada de clientes tenga una sola solución, lo cual abre las puertas hacia un análisis no sólo descriptivo, sino que también predictivo que permita encontrar otras posibles soluciones para vender.

El impacto del tamaño de la compañía en el análisis

En este trabajo, el tamaño de una compañía estará estrechamente relacionado con la cantidad de empleados que tiene. Una mayor cantidad de empleados significa que se trata de una compañía grande, mientras que una menor cantidad se traduce en una compañía más chica. Por ejemplo, una compañía de 10 a 49 empleados se considera una compañía pequeña, desde 200 a 500 es una compañía mediana, y una compañía con hasta 5000 empleados se le tratara como una compañía grande.

Si bien hay otras variables que también podrían indicar el tamaño de una compañía, como ser su facturación anual informada, se trata de datos correlacionados dados los lineamientos a seguir y no agregan información significativa. Una compañía grande con una gran cantidad de empleados seguramente facture mucho más y se esté en otro segmento que una compañía más chica con menos empleados. Por este motivo, se decide usar el segmento asignado internamente por SAP que se trata de una combinación de ambos datos, tanto la cantidad de empleados como la facturación anual informada. La figura 2 muestra los tamaños de los segmentos en relación con el total de compañías cliente.

Figura 2: Porcentaje de compañías con determinados segmentos

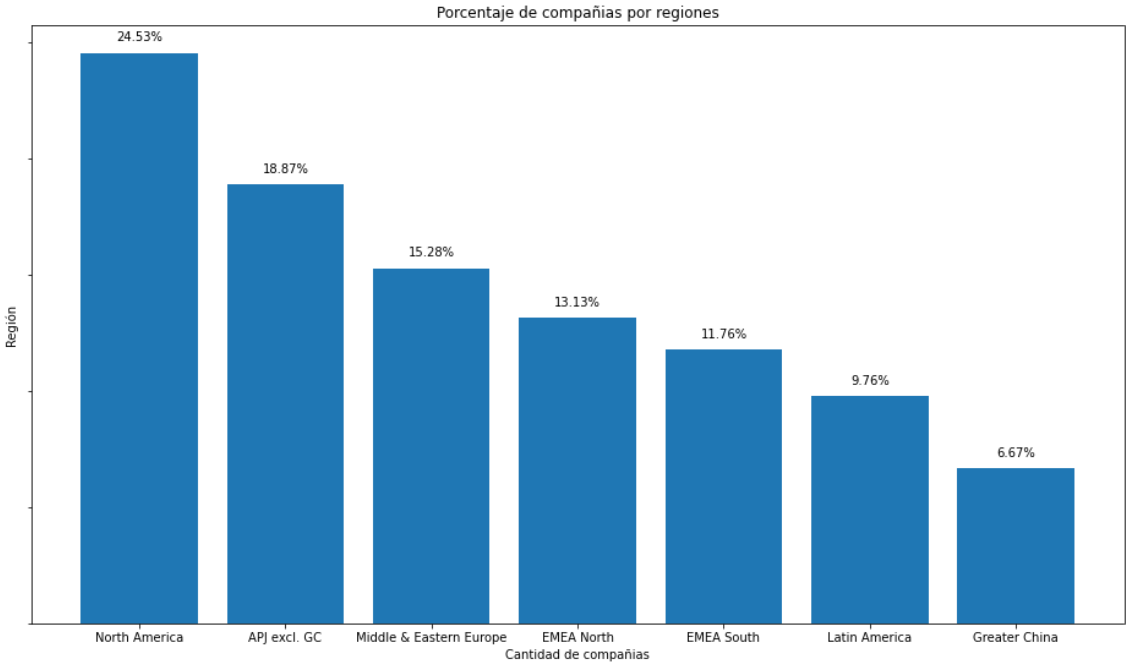


La importancia de la región en las compañías

Finalmente se evaluará también el impacto que puede tener la región a la que pertenece una compañía a la hora de obtener las probabilidades de compra para cierta solución. Quizás aquellas compañías ubicadas en determinada región con una cultura distinta a otra enfoquen sus compras en determinadas soluciones. Lo mismo podría decirse sobre la actividad productiva de

una región que quizá tiene mayoritariamente compañías dedicadas a ciertas industrias. La figura 3 muestra el porcentaje de compañías por regiones, en la cual es claro que la mayor cantidad de compañías se encuentran en Norte América y Asia.

Figura 3: Porcentaje de compañías por regiones.



La situación comercial

En términos comerciales, al tratarse de generar oportunidades de cross-sell, hay varios puntos a mejorar. Uno de ellos, y quizás el más crítico, es la falta de consideración de la relación entre SAP y sus clientes, como ser cuáles fueron las soluciones compradas y bajo que circunstancias se hicieron esas compras. Es decir que no se toman en cuenta datos como el historial de compra y la relación existente entre soluciones, las industrias de los clientes, y el tamaño de las compañías. A su vez, la falta de uso de estos datos con información sobre la actividad en la plataforma de SAP, también dejan un margen para mejorar ya que dependiendo de las búsquedas del cliente en un momento dado también pueden ayudar a determinar el mejor momento para concluir una venta. Si bien se cuenta con la información de la actividad de búsqueda sobre SAP.com, no existe una clara relación entre esta y una probabilidad de venta.

Por otro lado, la fidelidad de los clientes tampoco es tomada en cuenta a la hora de buscar nuevas ventas. Como toda compañía, es necesario no solo expandir su base instalada hacia nuevos clientes, sino que también promover el uso de otras soluciones en clientes ya activos que permitan la integración a un ecosistema compatible con otras soluciones. La forma de acercarse a un cliente para conseguir una venta no puede ser la misma para una compañía que tiene solo una o dos soluciones, que para una que tenga más de 4 o 5. En otras palabras, no se hace foco en la cantidad

de soluciones con las que ya cuenta un cliente y por lo tanto no tiene un trato preferencial ante una compañía que no tiene ninguna solución.

Finalmente, no se debe olvidar el valor que tiene el tiempo en un proceso de venta. A veces el interés de un cliente por una solución se ve afectada por el tiempo que pasó desde su última compra o por algún otro factor externo. Esto hace que tener información integral actualizada lo más rápido posible, tenga un rol crítico en las ventas.

2. Datos

2.1. Revisión de los datos

Los datos provienen directamente de SAP, sin embargo, están enmascarados con el fin de evitar incumplir cualquier regla de confidencialidad. Los datos usados para el desarrollo del motor de recomendaciones son una unificación de la base de datos de compras hechas por clientes y de la base de datos con información adicional de cada una de las cuentas cliente de SAP. Estos datos son actualizados trimestralmente de manera que se obtienen las últimas ventas hechas. Luego de enmascarar los datos, en total se trata de unas 261.908 compañías cliente con 1.200.000 registros de compras aproximadamente.

Las tablas usadas para este trabajo son,

- **Adoption Monitor File**

  El llamado Adoption Monitor es en donde se encuentran datos de más de 1.200.000 compras de las compañías en forma de registros y más de 260.000 compañías clientes en su jerarquía más alta, sin tener en cuenta sucursales de la propia compañía. A su vez, en el mismo dataset están disponibles distintos datos sobre cada compra en forma de variables como un ID, la fecha de compra, nombre de la solución y nombre del componente (cada solución puede tener varios componentes) en detalle. Si bien dentro del dataset no se encuentran datos de las compañías, como ser la industria o región de la compañía, se pueden hallar datos sobre el estado de la implementación de la solución y el estado contractual de mantenimiento, presente entre varias columnas con códigos técnicos sin relevancia para este trabajo.

  Vale aclarar que en ninguna parte se especifica el monto de la compra, ya que no es considerado relevante para el análisis debido a la confidencialidad de los datos ya que esta depende del tipo de contrato acordado entre SAP y las compañías cliente. Este trabajo tiene como objetivo principal hallar las probabilidades de compra a partir de otras variables que pueden influir en el cliente llegado el momento de la compra.

  En total la tabla Adoption Monitor tiene unas 29 columnas, entre las cuales se destacan las siguientes,

  - ID de la compañía registrada en CRM
  - Estado de la implementación de la compra

- Fecha de inicio del contrato
- Tipo de implementación
- Estado contractual del mantenimiento de la solución
- Nombre de la solución
- Nombre de la sub-solución

En las figuras 4 y 5 se grafican las cantidades de compañías clientes por región y por industria, observándose que Norte America y Europa central y oriental son las principales regiones, junto a las industrias de Professional Services y Wholesale Distribution por parte de las industrias.

Figura 4: Distribución del historial de compras entre regiones

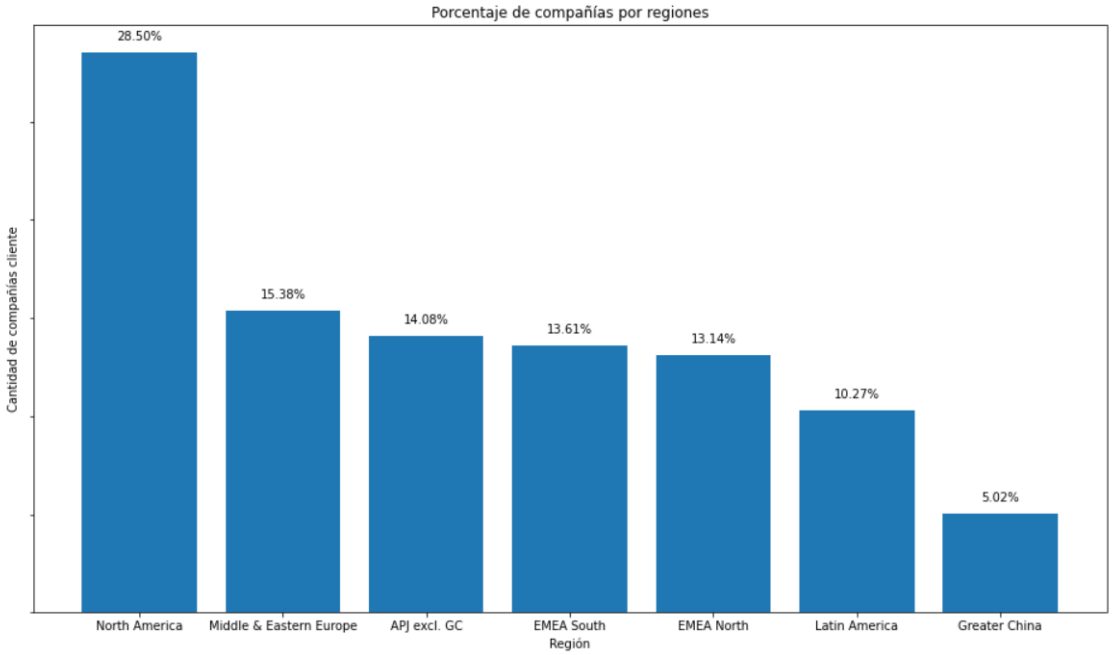
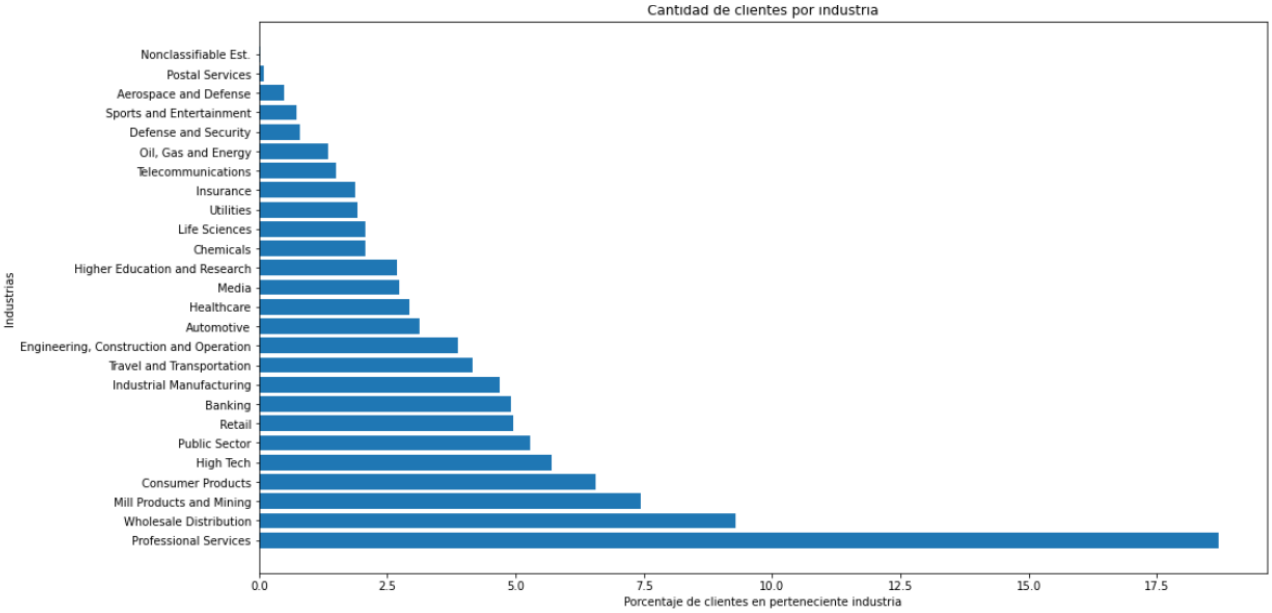


Figura 5: Cantidad de clientes por industria



Parte del trabajo será explorar las distintas combinaciones de solución y componente adquiridos por las compañías ya mencionadas. Para lograrlo es importante entender cómo se conforman las soluciones y los componentes. Existe un total de 8 soluciones, dentro de las cuales existen distintos componentes, por ejemplo, una solución puede tener dentro entre 2 y 5 componentes que pueden o no ser complementarios. A su vez dentro de cada componente es posible seguir explorando en profundidad para encontrarse con subcomponentes, que pueden ser más de 30 por componente. Para los fines de este trabajo, sólo se trabajará con soluciones y componentes, a menos que sea necesario hallar alguna diferenciación entre componentes durante su desarrollo.

- **CRM data**

Se trata de una tabla que guarda todos los principales datos de compañías cliente o potenciales clientes de SAP. Se actualiza de forma diaria con datos como ser el ID único de la compañía, número de empleados e ingreso de la compañía. Otros datos de suma importancia hallados en la tabla son la segmentación interna de acuerdo con los estándares de SAP; la región a la que pertenece la compañía cliente, abarcando todas las regiones del mundo divididas en APJ (Asia Pacific and Japan, excl. Greater China), GC (Greater China), NA (North America), LAC (Latin America and Caribbean), y EMEA (Europe, the Middle East and Africa) /MEE (Middle- and Eastern Europe). Además del país y la región de la compañía, también se cuenta con la industria a la que pertenece la compañía, como por ejemplo industria agropecuaria, automotriz, educación, entre otras.

Los datos que se encuentran en la tabla son obtenidos por medio de procesos internos de la compañía que no son relevantes para este trabajo. Sin embargo, la principal fuente de datos son compañías con grandes trayectorias que recolectan y validan estos datos de compañías actualizándose según sea conveniente. A su vez, también hay que remarcar que, si bien la mayoría de los datos de compañías cliente están disponibles, puede que datos como ser la cantidad de empleados o los ingresos de la compañía no estén disponibles en algunos casos. Esto se debe a que la cobertura de algunas compañías que recolectan datos se ven afectadas por limitaciones gubernamentales o legales en algunos países, dado que su conocimiento público puede no ser obligatorio.

Como se analizará más adelante en el trabajo, se evaluará si existe una correlación entre algunos de estos campos, evaluándose si se pueden desestimar algunos datos debido a la falta de estos. Es así como el número de empleados y el tamaño de la compañía podría, en principio, estar correlacionada con el segmento interno a la que pertenece según SAP.

Es así como esta tabla complementa los datos de compras históricas hechos por compañías clientes hallados en el Adoption Monitor.

Las columnas con las que se cuentan en la tabla son:

- ID de la compañía
- Número de empleados
- Ingreso declarado de la compañía
- Industria a la que pertenece
- Región
- País
- Unidad de negocios
- Segmentación interna

2.2. Exploración de los datos

Análisis preliminar de los datos

Antes de comenzar a trabajar con los datos, se debe entender del contexto de estos. Un análisis preliminar de los datos permite determinar que técnicas usar, de qué manera pensar en nuevas variables, encontrar potenciales errores en los datos, como valores nulos, y hacer los ajustes necesarios.

Los datos de compras históricas por cuentas cliente abarcan desde su primera compra hasta la última, siendo la compra más antigua aproximadamente en el año 1970 y la más actual el año 2023. De misma manera, la tabla con los datos de CRM contiene los datos actuales que se hallan en la base de datos sin tener en cuenta datos históricos, por ejemplo, si una compañía cambió su industria o segmento.

En la figura 6, que abarca el número de ventas durante los años 1970 hasta la actualidad, se pueden ver las variaciones en las ventas en términos porcentuales.

Figura 6: Evolución del número de ventas hasta la actualidad

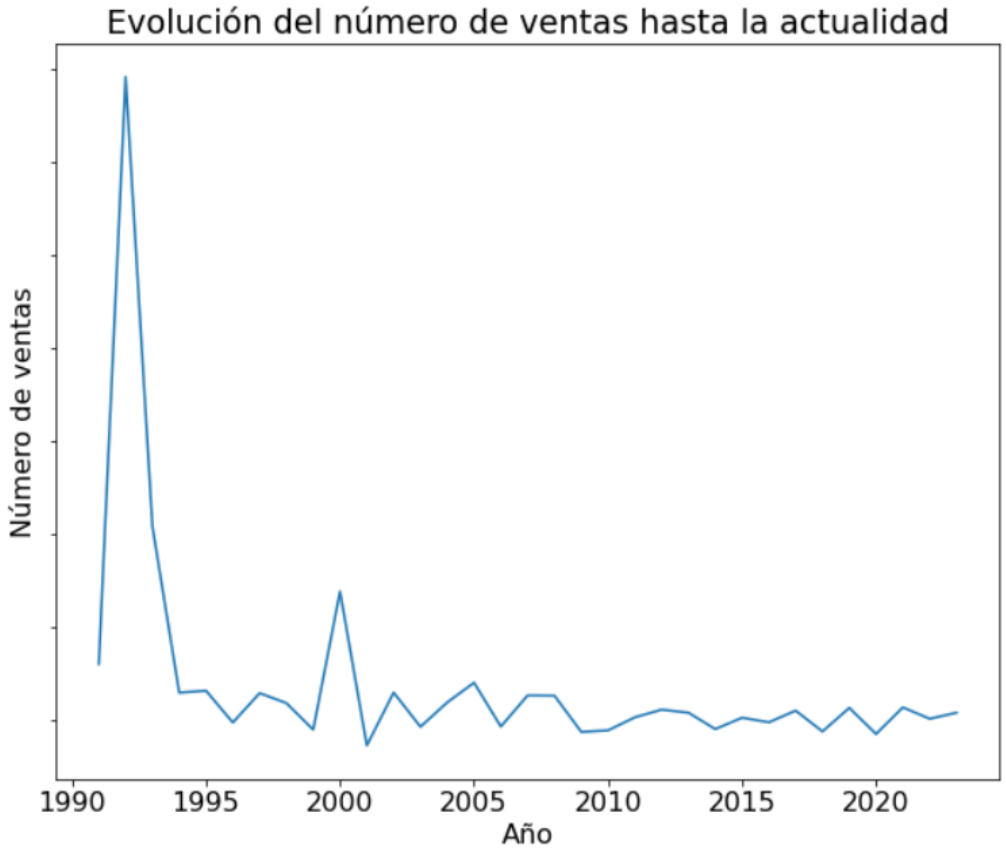
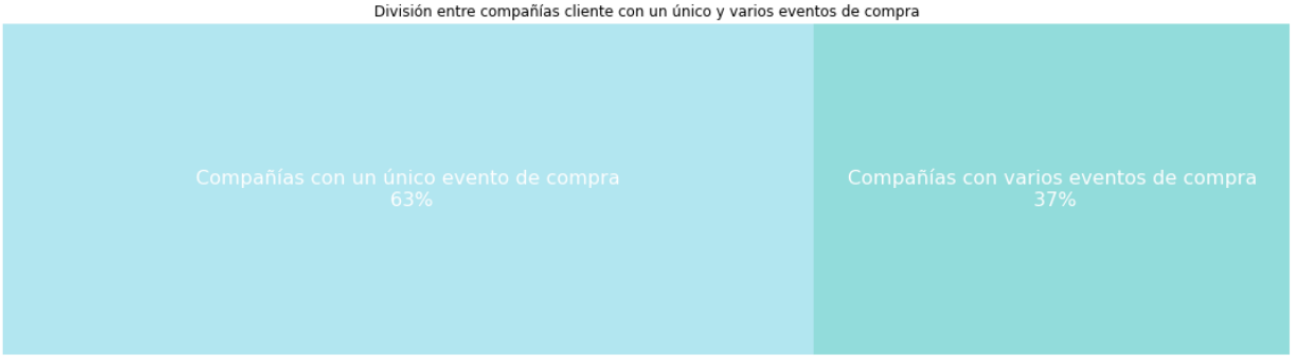


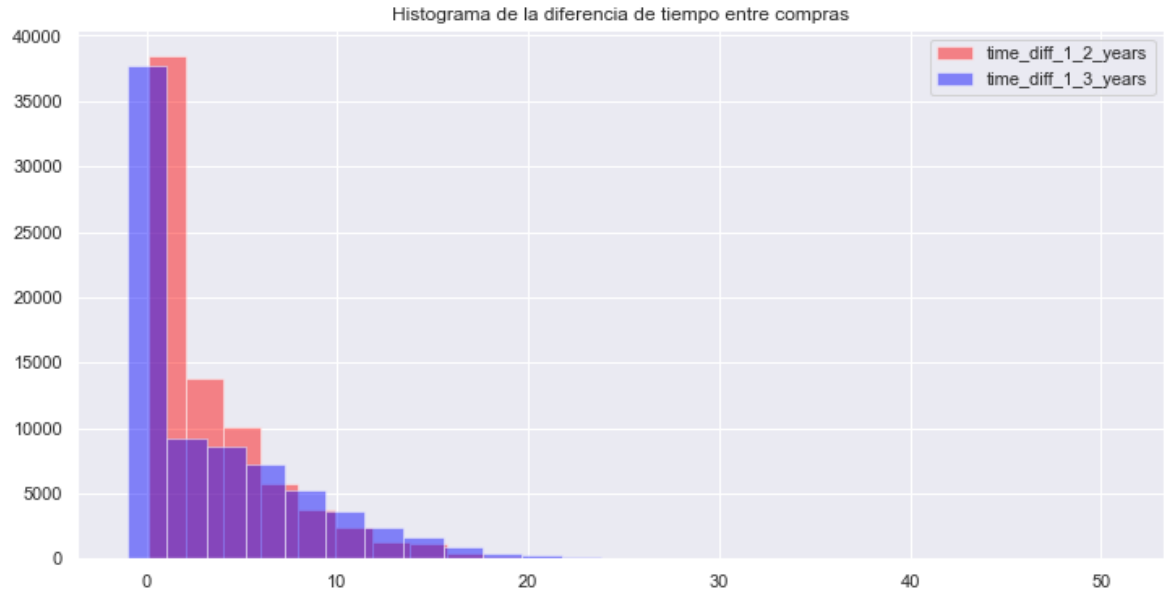
Figura 7: División entre compañías cliente con un único y varios eventos de compra.



Los datos muestran que más de 165.000 clientes (aproximadamente un 70% de la base de clientes de SAP) sólo tienen un evento de compra, sin éxito en llegar más lejos. A su vez, es posible observar que el número de compañías se reduce a medida que aumentan la cantidad de compras, con un salto brusco entre aquellas que tuvieron dos eventos de compras y las que decidieron comprar una tercera vez. Una posible interpretación de estos datos es la falta de campañas o acciones por parte de SAP hacia sus clientes más nuevos o más chicos, asumiendo que la cantidad de compras esta correlacionada positivamente con el tamaño de la compañía. En lo que amerita a este trabajo, esto se traduce en que una campaña que se aboque a las compañías con determinado tamaño y que compraron una sola vez tiene más sentido que buscar clientes más grandes y con mayor cantidad de compras.

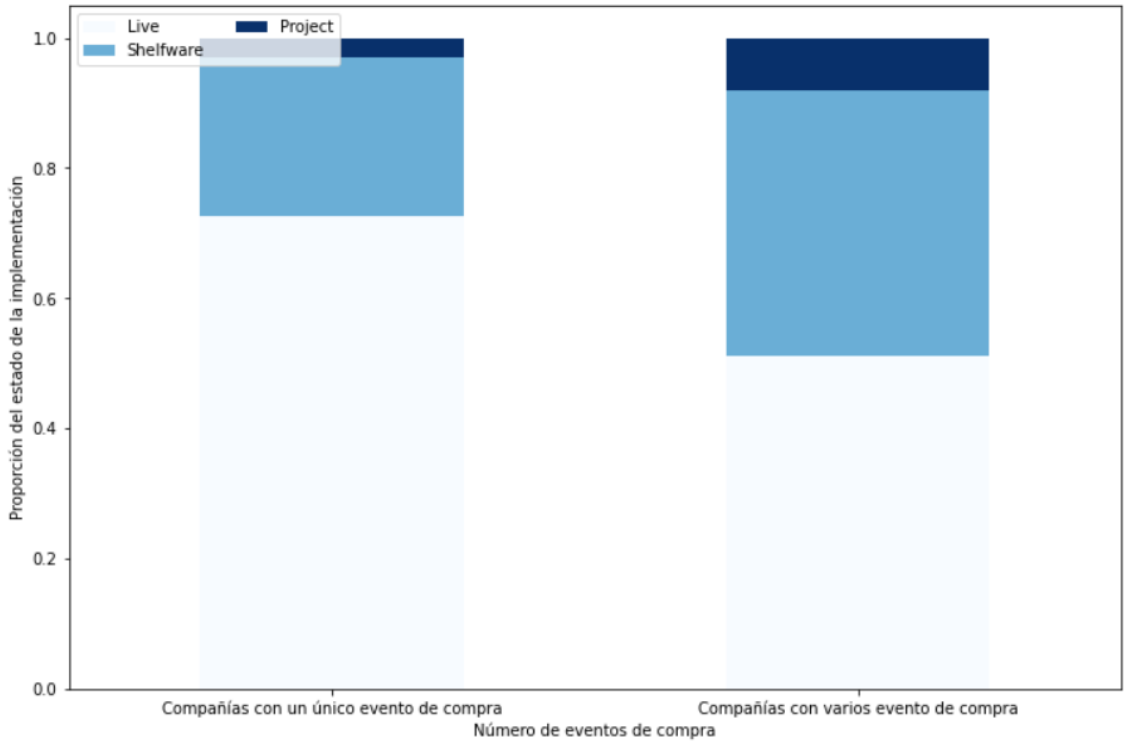
Así mismo, también podría probar ser útil explorar luego variables como la diferencia de tiempo que existe entre una compra y la otra, más específicamente entre la primera compra y la segunda, así como también entre la primera compra y una tercera. Esto permite entender en mayor profundidad el comportamiento de la mayoría de los clientes a la hora de adquirir nuevamente un producto de SAP. Como se ve en la figura 8, la diferencia de tiempo entre compras es importante a la hora de lograr obtener la probabilidad de que una compañía haga una determinada compra o no a futuro.

Figura 8: Histograma de la diferencia de tiempo entre compras.



En línea con lo anterior, cuando se trata de evaluar el impacto que podría tener el estado de la implementación de software en una compañía cliente, se hicieron dos análisis a modo comparativo. Por un lado, una exploración de la proporción de instalaciones en vivo o en proyecto y aquellas que no están en uso y son consideradas "shelfware" para aquellas compañías cliente que sólo llevan una compra hecha. A esta exploración se la compara con el mismo análisis para aquellas compañías cliente que tienen varias compras ya hechas.

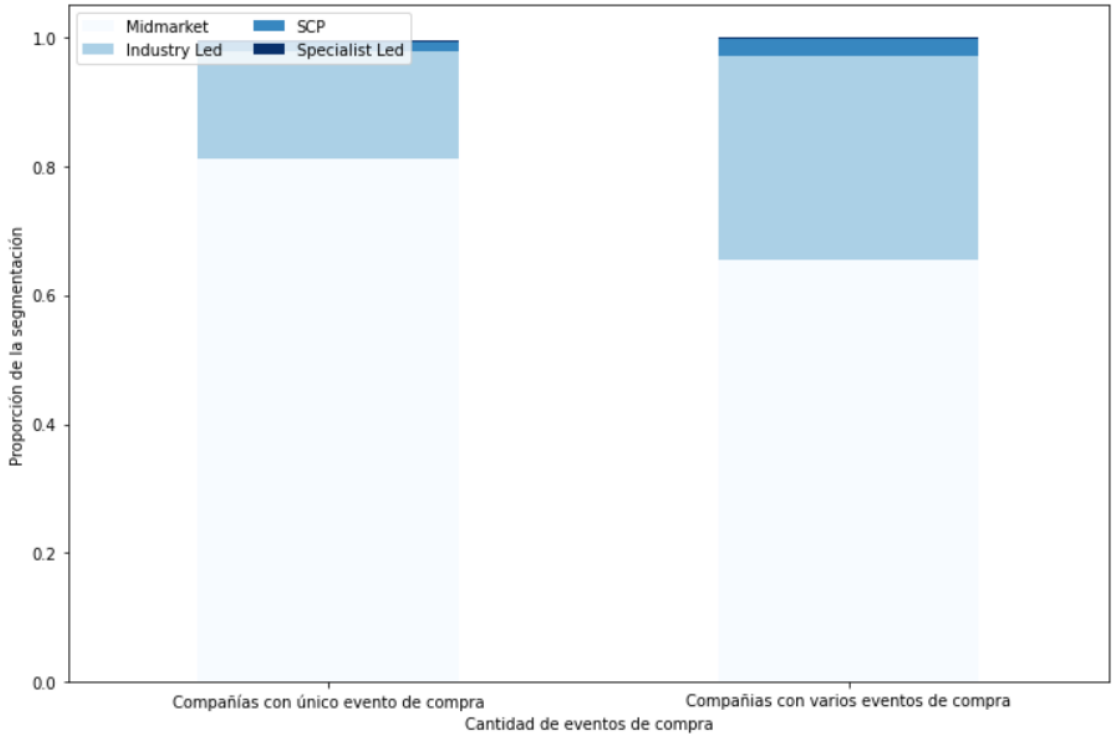
Figura 9: Estado de implementaciones de soluciones entre compañías con un único y varios eventos de compra.



La diferencia en proporciones es notable, ya que, en aquellas compañías con una sola compra, se mantiene activo el uso de su software en más de un 20% comparándolo con aquellas que tienen varias compras. A su vez, la diferencia se trata de todavía más de un 20% cuando se profundiza en el software inactivo o como "shelfware" siendo la proporción superadora de aquellas compañías con varias compras.

Dados estos datos, podría concluirse que en lo que respecta a la implementación y actividad del software, aquellas compañías con más de una compra en SAP suelen tener un índice de 20% mayor de "abandono" en sus soluciones.

Figura 10: Segmento de compañías con un único y varios eventos de compra.



Ahora bien, el siguiente punto importante a considerar es la relación entre el tamaño de una compañía y su capacidad de compra. Al revisar las proporciones, haciendo la comparación entre aquellas compañías con una sola compra y el resto con más de una, el resultado acompaña el supuesto que cuanto más grande es una compañía, mayor cantidad de compra tendrá. En la figura 10 se muestra claramente que las cuentas con una sola compra son aquellas que principalmente se encuentra en el segmento "midmarket", es decir el segmento relacionado al negocio en volumen y a compañías de menor tamaño.

Por el otro lado, la columna que muestra el segmento en aquellas compañías con más de una compra sostiene que la proporción de cuentas pertenecientes al segmento "Industry led" es casi el doble que en su comparativa. Es decir que las compañías con mayor tamaño tienden a hacer más compras.

De todas formas, un punto crucial también son aquellas compañías segmentadas como "SCP", que se trata de una porción mínima del mercado, pero con una alta importancia a nivel estratégico. En estos casos, la diferencia entre ambas proporciones es mínima o casi nula y el tamaño de la compañía parece no hacer gran diferencia.

Valores nulos en las tablas

Es posible encontrar valores nulos tanto en las tablas de Adoption Monitor file como en la de CRM data. Debajo se explicarán los pasos a seguir en cada caso:

- **Adoption Monitor File**

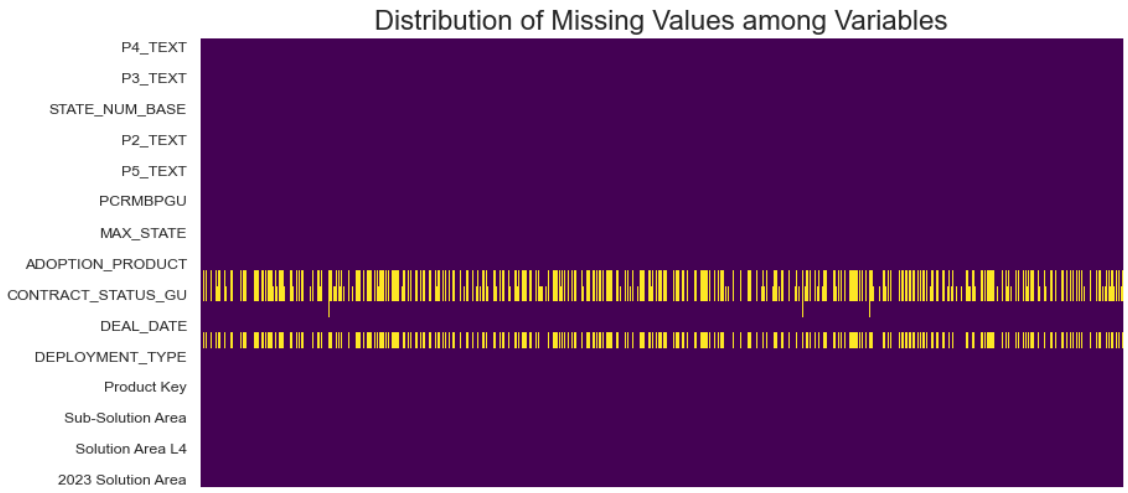
Como puede apreciarse en la figura 11, en lo que respecta al Adoption Monitor File, la cantidad de valores nulos se ve reducida gracias a que se trata de información interna de las compras de las compañías cliente. Sin embargo, la base de datos no queda exceptuada de tener valores nulos. Las razones detrás de estos valores faltantes son varias. La principal razón es el tiempo que pasó de la compra ya que contiene compras hechas desde los años 70, con décadas de diferencias con compras actuales. Otra razón es la falta de información contractual en la base de datos ya que no siempre se encuentra actualizada.

- **Fechas de inicio y fin del contrato:** Estos campos contienen la gran mayoría de valores nulos y, como indican los nombres, contienen información contractual. En lo que respecta a este trabajo, la información contractual no tiene gran utilidad debido a que no siempre está conectada en tiempo real con las adquisiciones de las cuentas cliente. Para eso, es importante tener en cuenta el campo "DEAL_DATE" o, traducido por su significado, fecha de venta. Este último no tiene ningún valor nulo.

- **Estado contractual:** El estado contractual contiene información financiera de la relación entre la compañía cliente y SAP. En otras palabras, informa si el cliente renovó o no una suscripción o si hay algún detalle contractual afectándolo. Si bien es información que podría ser valiosa para el trabajo, tampoco es un campo actualizado en tiempo real y, por lo tanto, su volumen de valores nulos lo hace inútil para la inclusión al análisis.

Tanto en las figuras 11 y 12 se muestran traspuestas las variables del CRM data de manera horizontal, es decir que cada registro del CRM data está ubicado en forma de columna y aquellos valores nulos son coloreados de amarillos para que resalten a simple vista. El color púrpura en una observación significa que su valor no es nulo, mientras que el color amarillo denota la falta de datos en ese registro.

Figura 11: Distribución de valores nulos entre las variables de Adoption Monitor.



- **CRM data**

Como se puede ver en la figura 12, las variables con mayor cantidad de valores nulos son aquellas que dependen de datos externos, es decir, cantidad de empleados e ingreso anual declarado por la compañía. Sin embargo, las variables relacionadas a la unidad de negocio, al tamaño y al rango del ingreso también tienen una considerable cantidad de valores nulos.

El tratamiento de estos valores nulos dependerá principalmente del volumen de datos y de la importancia de la variable.

- **Número de empleados e ingreso anual de la compañía:** Debido al alto volumen de valores nulos y a la correlación de estos valores con la segmentación interna de la compañía, el mejor tratamiento es quitar estas variables del análisis ya que no aportan una cantidad suficiente de datos para ser considerados relevantes en el trabajo.

- **Unidad de negocios:** La variable de unidad de negocios tampoco aporta datos nuevos al análisis que no estén estrechamente correlacionados con la información de la región o país. Los datos que contiene son puramente divisiones estratégicas de regiones que no impactan en la futura tendencia de una compañía en compañía en comprar una solución o no. Esta variable se quitará del análisis.

- **Tamaño y rango de ingreso de la compañía:** Similar a los campos de número de empleados e ingresos anuales de las compañías, son datos obtenidos de manera externa, pero con la diferencia que son campos compuestos por un cálculo interno. Por ejemplo, el campo “tamaño” se trata de una serie de categorías entre las cuales se dividen las compañías según su tamaño, es decir, que existe la categoría “chica”, “mediana”,

"grande", etc. Su cantidad de valores nulos está ligada a la del número de empleados, aunque en muchos casos aquellos valores nulos que se traducen como "0" se convierten en parte de la categoría "chica".

De la misma manera se calcula el campo de rangos sobre el ingreso anual de las compañías. El rango está compuesto por categorías que son dependientes del campo ingreso anual de la compañía descripto anteriormente.

Figura 12: Distribución de valores nulos en las variables del CRM data.

Selección de variables para el análisis

Luego de comprender los datos y revisar su validez, para este trabajo se seleccionarán las variables más importantes para la creación del algoritmo final. Como es sabido, la selección de variables es sumamente importante para construir un modelo predictivo. El objetivo de la selección de características es reducir la dimensionalidad del conjunto de datos, conservando tanta información como sea posible. Esto es importante porque ayuda a evitar el sobreajuste y mejora la precisión, la interpretabilidad y el rendimiento del modelo. Al seleccionar sólo las variables más importantes, el modelo puede centrarse en los factores clave que impulsan la predicción e ignorar la información irrelevante o redundante. Esto también reduce la complejidad computacional y hace que el modelo sea más eficiente. La selección de variables puede realizarse mediante varias técnicas, como los métodos de filtro, envolvente e incrustado. En general, la selección de características es un paso esencial en la construcción de un modelo predictivo que sea preciso, robusto y generalizable a nuevos datos.

Debido a que usamos dos datasets con distintas variables, la selección se hará sobre cada uno para luego crear el dataset final.

Adoption Monitor File

En cuanto al Adoption Monitor File, la selección de variables es un tanto más complejo. En primer lugar y luego de quitar las variables con una alta cantidad de valores nulos, es importante determinar cuáles son las variables que podrían ser útiles para la creación de las nuevas variables que dependen del orden y cantidad de compra. En segundo lugar, si bien hay varios niveles de soluciones en términos de productos, en este trabajo solo se usará el segundo nivel.

Las variables seleccionadas de este dataset son el ID de la compañía (al tratarse de una clave foránea que relaciona esta tabla con la de CRM data), la fecha de compra de la solución, el nombre de la solución comprada, y el estado y tipo de implementación que tuvo.

De esta manera, se obtiene un bosquejo inicial del dataset a usar para el trabajo, uniendo la información de la tabla de CRM data junto con la tabla Adoption Monitor File.

CRM data

Como fue aclarado con anterioridad, debido a la cantidad de valores nulos hallados en las variables relacionadas a la cantidad de empleados y al ingreso anual de las compañías, las variables a usar en el trabajo serán ID de la compañía (con el fin de identificar los resultados), industria, región, país, y segmento interno de la compañía cliente.

Feature Engineering

Como siguiente paso, se procederá a la creación de nuevas variables que ayuden a encontrar patrones. Feature engineering es un proceso de selección y transformación de variables o características para mejorar el rendimiento de un modelo predictivo. El objetivo del feature engineering es crear nuevos rasgos que sean más informativos o predictivos que los originales. Esto suele hacerse seleccionando o creando nuevas variables que sean más relevantes para la variable objetivo, o transformando las variables existentes de forma que capten sus relaciones con el objetivo. Feature engineering es un paso crucial en el proceso de modelado predictivo, ya que puede tener un impacto significativo en el rendimiento del modelo.

Si bien en el presente trabajo existe la variable de fecha de compra, la verdadera utilidad de tal campo radica en entender no solo el día de la compra, sino el significado que hay detrás del momento en el que la compañía cliente decide hacer la compra. Es decir, usar el campo de fecha de compra para determinar si existe un patrón en términos de diferencia de tiempos entre compras y el orden de estas. Aquella compra que primero tuvo como solución Enterprise Management Private, como segunda Core HR and Payroll, y como tercera Procurement, no es lo mismo que aquella compra que tuvo como primera solución Core HR and Payroll, luego Procurement, y finalmente Enterprise Management Private. En tal caso, el orden afecta el resultado, asi como

también es afectado por el momento. Un patrón que podría indicar que las compras de Procurement son en general seguidas en el corto plazo por Core HR and Payroll, también indica un posible patrón de compra.

Orden de compra o secuencia

En primer lugar, la variable de orden de compra se encuentra dada intuitivamente por el ordenamiento de los datos de fechas de manera ascendente, es decir aquella compra que sucedió en primer lugar se ubica por encima de las consecuentes. Por otro lado, debido a la enorme cantidad de fechas que amenaza con no hallar un resultado consistente, se tomó la decisión de transformar las fechas en el primer día del mes en el que tuvo lugar, por ejemplo, la compra sucedida con fecha 21-04-2013 es convertida en 01-04-2013, de tal forma que se logra agrupar las fechas por mes y año, dejando de lado el día.

En segundo lugar, una vez alineadas las fechas, el siguiente paso es crear una nueva columna con el nombre "sequence" que contenga el orden de las compras. De esta forma se puede visualizar fácilmente cual es el orden de las compras por parte de las compañías. Vale la pena destacar que hay casos en donde varios productos fueron comprados en conjunto, o sea en la misma fecha. Por esto mismo es que el número asignado en la columna "sequence" se repite como se ve en la tabla 3. Esta nueva variable no sólo hace visible el orden de las compras, sino que también el evento de la compra.

Tabla 3: Extracto del Adoption Monitor File.

	PBUP_AC	Sub-Solution Area	DEAL_DATE	sequence
347065	10717878	Planning and Analytics	2011-01-01	1.00
347061	10717878	Database and Data Management	2011-06-01	2.00
347063	10717878	Finance and Q2C	2012-11-01	3.00
347059	10717878	AppDev/Automation and Integration	2014-09-01	4.00
347060	10717878	Core HR and Payroll	2014-09-01	4.00
347062	10717878	Enterprise Management Private	2014-09-01	4.00
347064	10717878	Industry-specific Applications	2020-12-01	5.00
347066	10717878	SuccessFactors Cross	2022-01-01	6.00

Diferencia de tiempo entre compras

La diferencia de tiempo entre compras es otra variable se creó a partir de la fecha de compra de la solución. El objetivo de este campo es entender en mayor profundidad si existe una similitud entre algunos clientes que compran determinada solución y luego esperan un tiempo para adquirir la siguiente. A su vez también es útil para conocer la distribución que existe entre la primera segunda o tercera compra en términos de tiempo. Comprender qué porcentaje de las compañías cliente decide adquirir su segundo o tercer producto en una determinada ventana de tiempo puede impactar en la probabilidad que tenga una determinada compañía en comprar una solución.

Para determinar la diferencia de tiempo entre compra basta con realizar una operación de resta entre las fechas, dando como resultado el número de días que pasan entre una y la otra. De esta forma también es posible dividir tal número por 365 para encontrar la cantidad en años y finalmente hacer un agrupamiento en grupos para una mejor visualización como se ve en las figuras 13 y 14.

Figura 13: Intervalos de tiempo entre la primera y segunda compra.

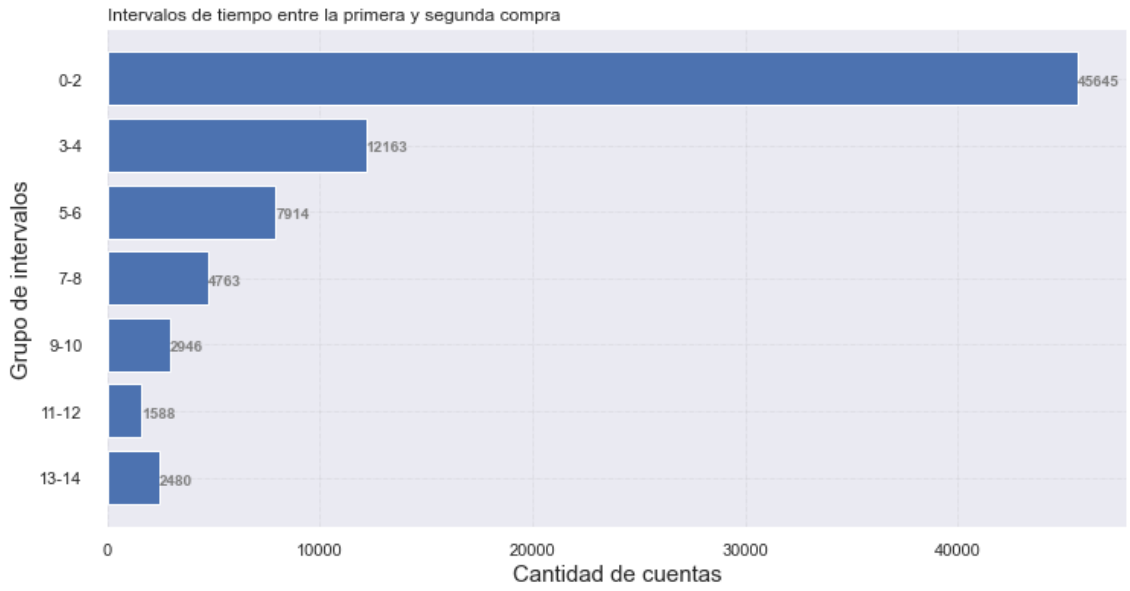
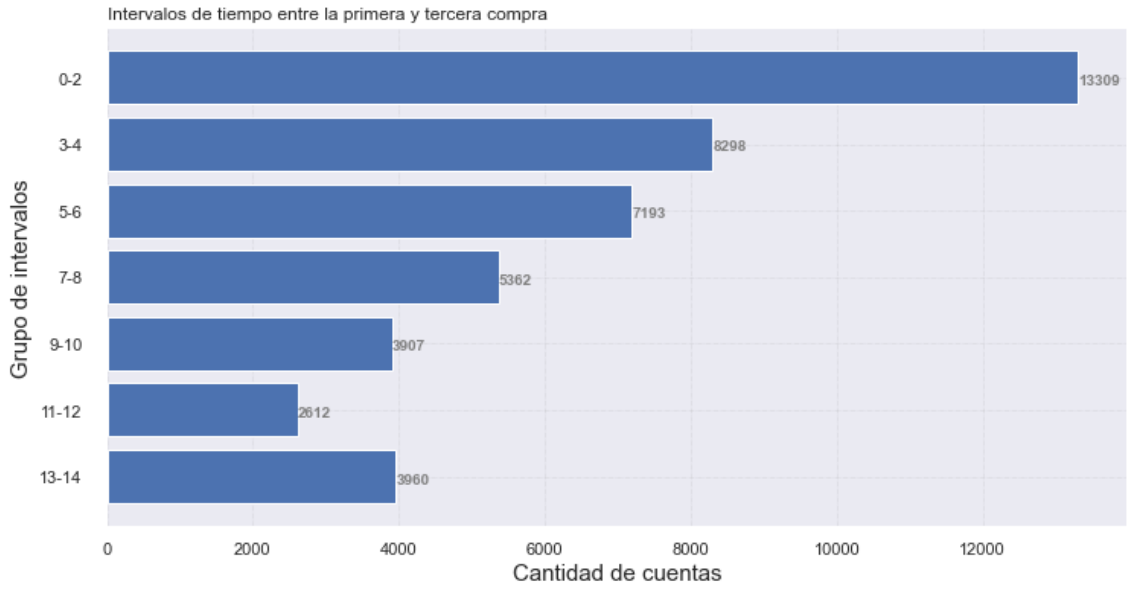


Figura 14: Intervalos de tiempo entre la primera y tercera compra.



Como se aprecia en las figuras, la mayor cantidad de compras que suceden tanto en segundo como en tercer lugar tienen lugar durante los primeros 2 años que la compañía adquirió su primera solución, volviendo estos clientes más propensos a tener una siguiente compra. Un dato interesante es que también se muestra la distribución sobre los intervalos de compra, manteniéndose la misma tendencia entre ambos intervalos. Sin embargo, es interesante el comportamiento luego del intervalo de los 6 años, en donde el volumen de ventas decrece rápidamente. En términos de negocios, podría decirse que es mucho más difícil vender un producto a una compañía una vez que han pasado más de 6 años.

One-hot encoding

Una vez que cuentan con ambos datasets, tanto CRM data como Adoption Monitor File, libres de valores nulos y únicamente con variables relevantes para el análisis, el siguiente objetivo es unificarlos y utilizar la técnica de one-hot encoding para clasificar cada compañía.

Para unir ambos datasets y conseguir que tanto la información firmográfica de las compañías estén unidas a sus historiales de compras, se usará el ID único de cada compañía presente en cada uno de los datasets como “PBUP_AC”. Este campo actúa como una clave primaria en el caso de CRM data y como clave foránea en el caso del Adoption Monitor File.

Tabla 4: Extracto del resultado de la unión de ambos datasets

	PBUP_AC	Sub-Solution Area	DEAL_DATE	sequence	SAP_Mastercode	Region_Label	Internal_Account_Classification	MAX_STATE
0	30771	Core HR and Payroll	2000-02-01	1.00	Insurance	EMEA North	SCP	Shelfware
1	30872	Enterprise Management Private	1999-05-01	1.00	Consumer Products	Middle & Eastern Europe	Industry Led	Live
2	30873	SuccessFactors Cross	2013-06-01	10.00	Consumer Products	EMEA North	SCP	Shelfware
3	30873	Planning and Analytics	2008-03-01	9.00	Consumer Products	EMEA North	SCP	Shelfware
4	30873	Learning and Talent	2006-12-01	8.00	Consumer Products	EMEA North	SCP	Shelfware
5	30873	Finance and Q2C	2006-12-01	8.00	Consumer Products	EMEA North	SCP	Shelfware
6	30873	Digital Supply Chain	2004-06-01	6.00	Consumer Products	EMEA North	SCP	Shelfware
7	30873	Marketing	2005-01-01	7.00	Consumer Products	EMEA North	SCP	Shelfware
8	30873	Procurement	2002-05-01	4.00	Consumer Products	EMEA North	SCP	Shelfware
9	30873	AppDev/Automation and Integration	2002-05-01	4.00	Consumer Products	EMEA North	SCP	Shelfware
10	30873	Database and Data Management	1998-12-01	3.00	Consumer Products	EMEA North	SCP	Shelfware
11	30873	Enterprise Management Private	1997-01-01	2.00	Consumer Products	EMEA North	SCP	Shelfware
12	30873	Core HR and Payroll	1994-10-01	1.00	Consumer Products	EMEA North	SCP	Shelfware

Luego de llegar a un único dataset es el momento de avanzar con la técnica de one-hot encoding. One-hot encoding es una técnica utilizada para convertir datos categóricos en datos numéricos binarios que puedan utilizarse en modelos de aprendizaje automático. Funciona representando cada valor posible de una variable categórica como un vector binario, en el que cada componente del vector corresponde a un valor posible y se establece en 1 si el valor está presente y en 0 en caso contrario. La técnica es especialmente útil para los problemas de clasificación porque permite al modelo tratar cada categoría como una característica separada e independiente, sin imponer ningún orden o clasificación particular a las categorías.

Una de las principales ventajas del one-hot encoding en problemas de clasificación es que elimina cualquier ambigüedad o sesgo que puedan introducir otros métodos de codificación, como la codificación de etiquetas, que asigna valores numéricos arbitrarios a variables categóricas. Esta técnica garantiza que cada categoría esté representada por un conjunto único de características, lo que es crucial para los problemas de clasificación en los que el objetivo es distinguir con precisión entre diferentes categorías. Además, puede ayudar a reducir la dimensionalidad de los datos eliminando características redundantes o irrelevantes, lo que puede mejorar el rendimiento del modelo y reducir el sobreajuste.

En este trabajo, la codificación es implementada para pasar todas las variables categóricas a valores numéricos. Las variables de industria, región, segmento y estado de la implementación fueron transformadas de forma que se les asigna un 0 o un 1 en el caso que pertenezcan a determinada categoría. En el caso de la solución adquirida, la codificación permite ubicar cada solución en forma de columna lo cual también es útil para la agrupación de compañías.

Como fue mencionado, el procedimiento se orienta a hallar satisfactoriamente cuales son aquellas soluciones con mayores probabilidades de formar parte del segundo evento de compras. Es por esto por lo que sólo se trabajará con los primeros y segundos eventos de compra, es decir con aquellos registros en donde la variable "sequence" es igual a 1 y 2, por el momento dejando de lado aquellos que vienen a futuro. De esta forma, al trabajar únicamente con esos dos eventos de compras, es posible separarlos y codificarlos individualmente. Dicho en otras palabras, por un lado, se crea una tabla con solo los primeros eventos de compra, aplicando la técnica de one-hot encoding y agrupándolos por número identificatorio de la compañía. Luego, se hace lo mismo, pero con aquellos segundos eventos de compra, pero sin los datos propios de compañías ya que serían repetitivos.

Una vez obtenidos ambos datasets con todas sus variables categóricas codificadas, es el momento de unirlos para lograr un único dataset completamente codificado y con ambos eventos de compra separados. Es así como nacen dos tipos de variables para el trabajo predictivo, las soluciones pertenecientes al primer evento de compra, y las soluciones pertenecientes al segundo evento de compra. Debajo se muestra el listado de columnas del dataset unificado, explicando a qué evento de compra pertenece cada columna.

Columnas con datos de compañías y el primer evento de compra	Columnas con el segundo evento de compra
PBUP_AC	AppDev/Automation and Integration_y
DEAL_DATE	Business Network_y
Datos de industria	Commerce_y
Datos de región	Core HR and Payroll_y
Datos de segmento	Customer Data Solutions_y
Datos de estado de implementación	Database and Data Management_y
AppDev/Automation and Integration_x	Digital Supply Chain_y
Business Network_x	ERP for SME_y
Commerce_x	Enterprise Management Private_y
Core HR and Payroll_x	Enterprise Management Public_y
Customer Data Solutions_x	External Workforce_y
Database and Data Management_x	Finance and Q2C_y
Digital Supply Chain_x	Industry-specific Applications_y
ERP for SME_x	Learning and Talent_y
Enterprise Management Private_x	Marketing_y
Enterprise Management Public_x	Planning and Analytics_y
External Workforce_x	Procurement_y
Finance and Q2C_x	SAP Signavio_y
Industry-specific Applications_x	Sales Performance Management_y
Learning and Talent_x	Sales and Service_y
Marketing_x	SuccessFactors Cross_y
Planning and Analytics_x	Training and Adoption_y
Procurement_x	Travel and Expense_y

SAP Signavio_x
Sales Performance Management_x
Sales and Service_x
SuccessFactors Cross_x
Training and Adoption_x

Como es apreciable en la tabla 5, aquellas columnas con el nombre de la solución seguido por la letra X, son las soluciones adquiridas en el primer evento de compra, mientras que aquellas columnas con el nombre de la solución y la letra Y, refieren a las soluciones adquiridas en el segundo evento de compra.

Tabla 5: Extractos del dataframe una vez aplicada la técnica de one-hot encoding

	PBUP_AC	DEAL_DATE	Aerospace and Defense	Automotive	Banking	Chemicals	Consumer Products	Defense and Security	Engineering, Construction and Operation	Healthcare	High Tech	Higher Education and Research	Industrial Manufacturing
0	30873	1994-10-01	0	0	0	0	1	0	0	0	0	0	0
1	30875	1994-01-01	0	0	0	0	0	0	0	0	1	0	0
2	30876	1995-12-01	0	0	0	0	0	0	0	0	0	0	0
3	30881	1994-12-01	0	0	0	0	0	0	0	0	0	0	0
4	30884	1997-05-01	0	0	0	0	0	0	0	0	0	0	0
5	30886	1996-12-01	0	0	0	1	0	0	0	0	0	0	0
6	30888	1997-03-01	0	0	0	0	0	0	0	0	0	0	0
7	30889	1993-04-01	0	0	0	0	1	0	0	0	0	0	0
8	30890	1997-03-01	0	0	0	0	1	0	0	0	0	0	0
9	30891	1994-04-01	0	0	0	0	1	0	0	0	0	0	0
10	30892	2000-05-01	0	0	0	0	0	0	0	0	0	0	0
11	30894	1997-04-01	0	0	0	0	0	0	0	0	0	0	0
12	30898	1998-02-01	0	0	0	0	1	0	0	0	0	0	0
13	30900	2000-01-01	0	0	0	0	0	0	0	0	0	0	0
14	30901	1996-04-01	0	0	0	0	0	0	0	0	0	0	0
15	30902	1994-01-01	0	0	0	0	0	0	0	0	0	0	0
16	30904	2004-04-01	0	0	0	0	0	0	0	0	0	0	1
17	30905	1993-11-01	0	0	0	0	0	0	0	0	0	0	0
18	30906	1994-12-01	0	0	0	0	1	0	0	0	0	0	0
19	30907	2006-01-01	0	0	0	0	0	0	0	0	0	0	0
20	30908	2007-01-01	0	0	0	0	0	0	0	0	0	0	0

AppDev/Automation and Integration_x	Business Network_x	Commerce_x	Core HR and Payroll_x	Customer Data Solutions_x	Database and Data Management_x	Digital Supply Chain_x	ERP for SME_x	Enterprise Management Private_x	Enterprise Management Public_x	External Workforce_x	Finance and Q2C_x
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0

AppDev/Automation and Integration_y	Business Network_y	Commerce_y	Core HR and Payroll_y	Customer Data Solutions_y	Database and Data Management_y	Digital Supply Chain_y	ERP for SME_y	Enterprise Management Private_y	Enterprise Management Public_y	External Workforce_y	Finance and Q2C_y
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	1	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0

Con esta transformación de las variables categóricas es posible seguir con un análisis de correlaciones entre variables, específicamente para encontrar las correlaciones entre las soluciones pertenecientes a los primeros eventos de compra y la solución del segundo evento. A su vez, también permite etiquetar las compañías según sus compras, haciendo más fácil el trabajo para el desarrollo de un algoritmo de clasificación.

Con el fin de obtener un dataset de entrenamiento y otro de validación, el dataset final se dividió en dos. Por un lado, el dataset de entrenamiento con el 70% de los datos mezclados al azar, y el dataset de validación con el 30% restante de los datos mezclados al azar también. De esta manera, en los siguientes pasos se analizarán las correlaciones y se entrenará el modelo con el dataset de entrenamiento.

Análisis de correlaciones

Como siguiente paso, se verificarán las correlaciones que existen entre las variables, incluyendo las soluciones. Vale aclarar que la correlación se produce cuando dos o más variables están relacionadas entre sí de forma sistemática, y puede manifestarse como una asociación positiva o negativa entre las variables. Un valor de correlación alto indica que hay una fuerte asociación entre dos variables y el valor, a su vez, describe la fuerza y la dirección de la relación entre dos variables. Un valor de correlación Pearson, que mide un grado de asociación lineal, se mide en una escala de -1 a 1, donde -1 indica una correlación negativa perfecta, 0 indica que no hay correlación y 1 indica una correlación positiva perfecta.

Cuando dos variables tienen un valor de correlación alto, significa que a medida que cambia el valor de una variable, la otra variable también cambia de manera predecible. Por ejemplo, si se observa una correlación alta entre la compra de una solución que tuvo lugar en el primer evento de compras, y la compra de una solución que tuvo lugar en el segundo evento de compras, esto significa que cuando se daba la primera compra, se daba la segunda de manera predecible por el primer evento.

Es importante tener en cuenta que una correlación alta no necesariamente implica una relación causal entre las variables, y es posible que existan otros factores que contribuyan a la asociación observada, ya que no dependerá únicamente de esa variable.

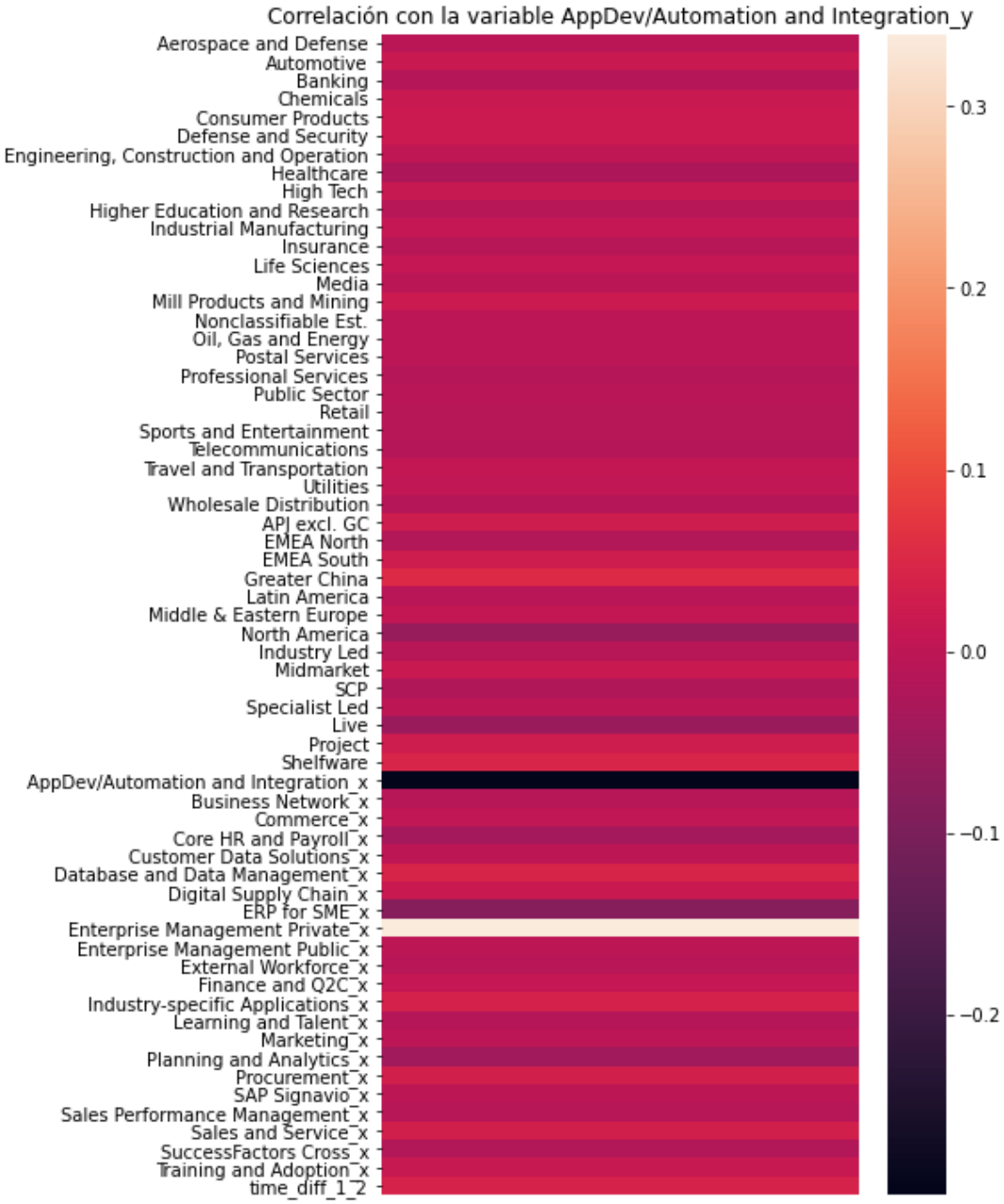
Dada la naturaleza de este trabajo en donde se trata de analizar la correlación de acuerdo con cada solución presente en el segundo evento de compra, no sólo los datos de compañías fueron considerados en la matriz de correlaciones, sino también el peso de la primera solución. En otras palabras, dados los casos positivos y negativos para determinada solución en el segundo evento de compra, cuál es la correlación con su primera solución. Debajo se encuentra el listado de soluciones que forman parte del primer evento de compra con el mayor valor de correlación para determinada solución en el segundo evento.

Tabla 6: Soluciones del primer evento de compra con un mayor valor de correlación con soluciones de un segundo evento de compra.

Soluciones del segundo evento de compra	Soluciones del primer evento de compra con un mayor valor de correlación
AppDev/Automation and Integration_y	Enterprise Management Private_x
Business Network_y	Database and Data Management_x
Commerce_y	Marketing_x
Core HR and Payroll_y	AppDev/Automation and Integration_x
Customer Data Solutions_y	Marketing_x
Database and Data Management_y	Planning and Analytics_x
Digital Supply Chain_y	AppDev/Automation and Integration_x
ERP for SME_y	Database and Data Management_x
Enterprise Management Private_y	Core HR and Payroll_x
Enterprise Management Public_y	AppDev/Automation and Integration_x
External Workforce_y	SuccessFactors Cross_x
Finance and Q2C_y	AppDev/Automation and Integration_x
Industry-specific Applications_y	AppDev/Automation and Integration_x
Learning and Talent_y	AppDev/Automation and Integration_x
Marketing_y	Commerce_x
Planning and Analytics_y	Database and Data Management_x
Procurement_y	AppDev/Automation and Integration_x
SAP Signavio_y	Database and Data Management_x
Sales Performance Management_y	Industry-specific Applications_x
Sales and Service_y	AppDev/Automation and Integration_x
SuccessFactors Cross_y	Learning and Talent_x
Training and Adoption_y	AppDev/Automation and Integration_x
Travel and Expense_y	Planning and Analytics_x

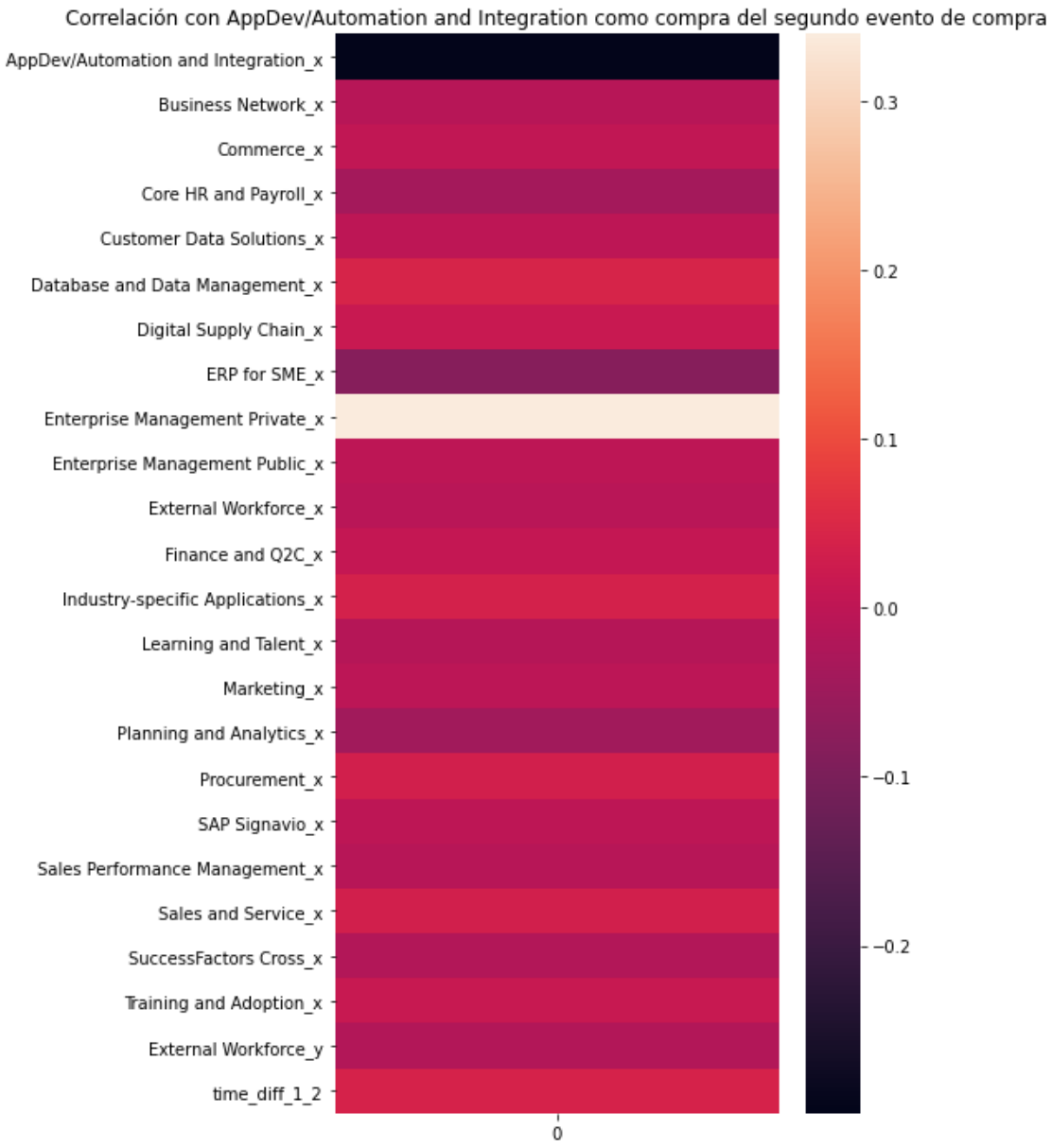
Debajo también se muestran algunas de las correlaciones de algunas soluciones a modo de ejemplo. En el primer caso, AppDev/Automation and Integration muestra que, si bien hay una correlación moderada de las variables, la solución de Enterprise Management Private es la que más resalta.

Figura 15: Correlación entre las variables independientes y la solución AppDev/Automation and Integration como parte del segundo evento de compra.



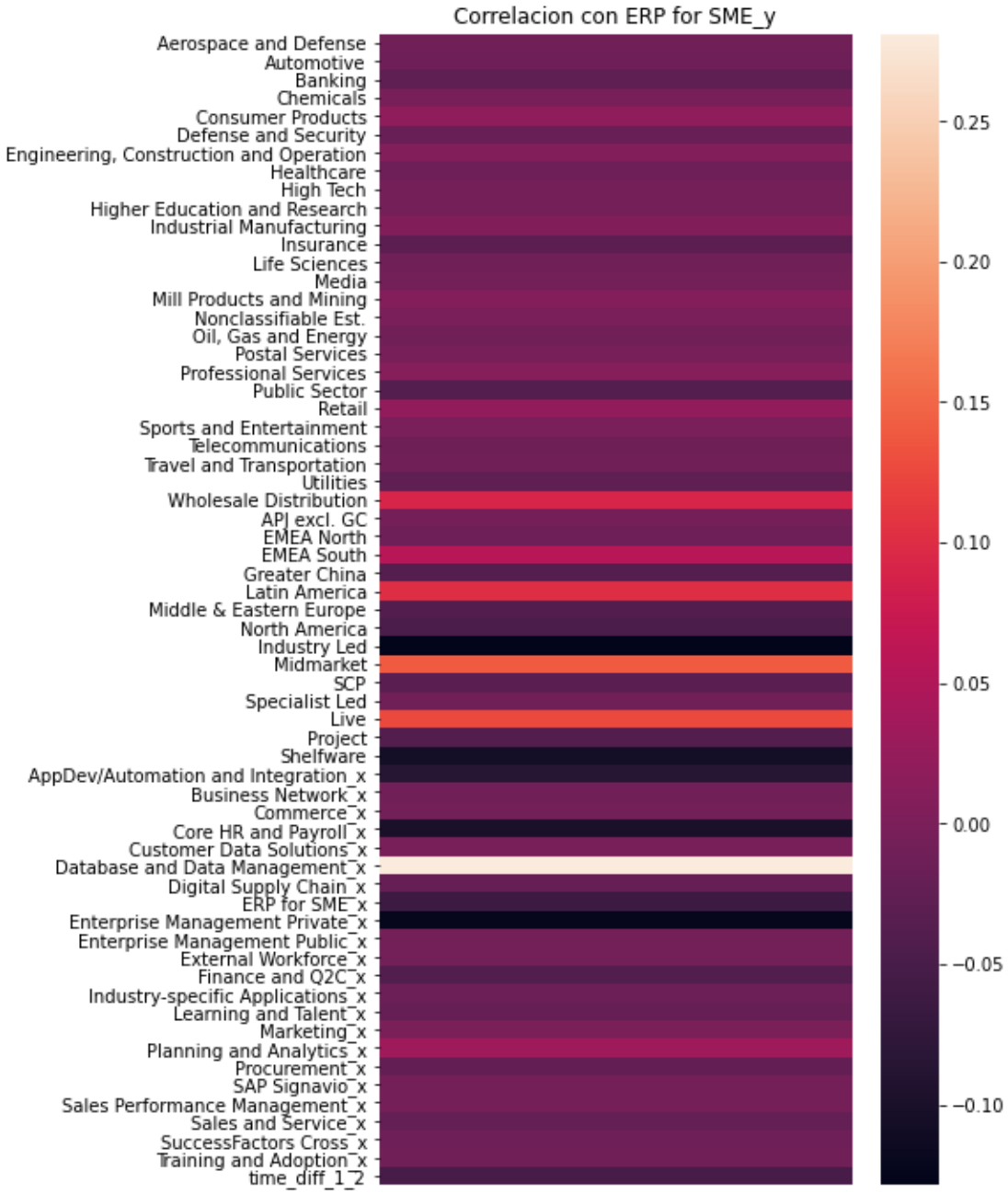
A modo de un vistazo con mayor detalle, es claro cómo, por un lado, la misma solución en el primer evento de compra tiene una correlación completamente negativa, lo cual es razonable por la poca probabilidad de que se adquiera en ambos eventos la misma solución, mientras que Enterprise Management es la que más destaca.

Figura 16: Correlación entre las soluciones y AppDev/Automation and Integration, dado que las soluciones forman parte del primer evento de compra.



Por otro lado, en el caso de ERP for SME como parte del segundo evento de compra, se puede apreciar que no sólo Database and Data Management es la solución del primer evento de compra que más correlacionada está, sino que también variables como la industria de Wholesale Distribution, la región de Latinoamérica, y el segmento de Midmarket, son los que también están más correlacionadas con ERP for SME como segunda solución.

Figura 17: Correlación de las variables independientes con ERP for SME como parte del segundo evento de compra.



Como conclusión de este análisis, resulta muy interesante cómo por un lado variables firmográficas afectan al resultado de la segunda compra, dependiendo de la industria, región, segmento, o incluso estado de implementación en algunos casos. Por otro lado, también es útil conocer cuáles son las soluciones que más correlacionadas están de manera positiva entre sí, otorgando una idea sobre el peso de determinada solución en términos de negocios.

2.3. Desarrollo del método de clasificación

Elección de modelos

Para lograr el mejor resultado de un modelo, es necesario definir qué modelos se usarán, con que métricas se evaluarán, y los posibles parámetros a considerar en cada uno. Además, el desempeño de cada modelo estará impactado no sólo por los parámetros, sino que también por el tamaño del dataset y la cantidad de casos positivos y negativos que existen tanto para el entrenamiento como para la validación. Es así como también se debe observar la performance de los modelos para definir cuál sería el mejor balanceo de los datos.

Para comenzar a elegir cuál sería el mejor modelo, se debe comprender el tipo de problema que se trata de resolver y planificar con claridad cuál es el mejor acercamiento al mismo. Al existir dos grandes ramas de problemas, siendo de regresión o clasificación, lo primero es entender cuál es la variable por predecir. En un problema de predicción de clasificación, la variable a predecir es categórica y discreta, es decir, toma valores de un conjunto finito de clases. Por ejemplo, predecir si un resultado es positivo o negativo, o si determinado registro pertenece a una categoría u otra. En un problema de regresión, la variable a predecir es numérica y continua. Por ejemplo, predecir el valor de una variable dadas ciertas variables.

En este trabajo, debido a que se intenta calcular la probabilidad que se realice una venta, es decir una variable categórica, se trata de un problema de clasificación. En otras palabras, cuáles son las probabilidades de éxito para una venta, sin punto intermedio. La venta puede o no realizarse. Por lo tanto, una manera eficiente de resolver este problema de clasificación seria realizar no sólo un modelo, sino un modelo por cada solución a predecir. Es decir, entrenar unos 23 modelos de machine learning para que cada uno devuelva la probabilidad que determinada compañía sea positiva para esa solución. De esta manera, se obtienen unas 23 probabilidades o resultados, que, al compararlos entre sí, se pueden seleccionar solo las mejores 3 probabilidades de encontrar casos positivos. La recomendación se basa en otorgar esas 3 probabilidades junto al nombre de la solución.

Ahora bien, habiendo definido el problema como un problema de clasificación y formado un plan para atacar el problema, todavía existen otros puntos para tener en cuenta antes de elegir un modelo a entrenar. Al contar con datos etiquetados, es decir que están explícitamente resaltados si son casos positivos o no, el siguiente paso está relacionado con el tipo de información disponible. En este trabajo, el dataset y la variable a predecir no precisan del análisis de texto, por lo cual ayuda a descartar en primer lugar a Naive Bayes como modelo de machine learning.

Por otro lado, modelos como K-Nearest Neighbors (KNN), Support Vector Classification y Gradient Boosting Machines podrían resultar útiles. Sin embargo, en el caso de KNN, si bien se trata de un modelo muy fácil de implementar, es un modelo computacionalmente muy costoso y que frente a una gran cantidad de datos podría probar ser muy lento, especialmente si se trata de 23 modelos. Sin mencionar que depende altamente del valor de K, teniendo que buscar en cada caso el valor óptimo, aumentando aún más su nivel de complejidad. En el caso de Support Vector Classification sufre de un problema similar a KNN, ya que no es escalable a conjuntos de datos muy grandes, incrementando rápidamente su tiempo de procesamiento, y a su vez no es el mejor modelo para casos con datos desbalanceados como podrían ser estos.

Por último, el caso de los métodos de ensamble, en particular Extreme Gradient Boosting (XGBoost), podría demostrar ser el modelo que más se ajusta a las necesidades del problema. Si bien algunas de sus desventajas son el ajuste de hiperparámetros y la necesidad de un poder computacional más alto, lo primero puede ser mermado por el uso de un algoritmo que permita obtener los mejores valores de hiperparámetros dado el resultado de una métrica como ser GridSearchCV. Además, su capacidad de evitar el sobreajuste de los datos y facilidad para ser escalable a grandes cantidades de datos, lo hace el principal candidato para este trabajo.

Presentando los modelos a usar

El método boosting se basa en la premisa de mejorar las predicciones resultantes de un árbol de decisión. Para mejorar los resultados de lo que sería un árbol de decisión, el método boosting ejecuta varios árboles de decisión sobre un mismo dataset de entrenamiento aprendiendo de los errores de cada uno a medida que los ejecuta, de tal manera que obtiene distintas probabilidades por cada uno de los árboles. El resultado final del modelo es aquella probabilidad más alta que mejor performance tuvo de acuerdo con la métrica seleccionada. Es decir, que el método usa el mismo dataset original, pero aprende basándose en los errores de cada uno de los árboles de decisión definidos por la profundidad del modelo. En este trabajo se utiliza el método de XGBoost debido a su robustez frente a datos atípicos.

Los principales parámetros que pueden ajustarse manualmente para una mejor performance del modelo según el problema son,

- Número de árboles (nrounds).
- Profundidad máxima de los árboles (max_depth).
- Tasa de aprendizaje (eta).
- Mínima reducción del error para generar un corte (gamma).
- Variables por muestrear y considerar en cada árbol (colsample_bytree).
- Mínima cantidad de observaciones en los hijos para considerar un corte. (min_child_weight).

- Muestreo de observaciones a considerar en cada árbol (subsample).

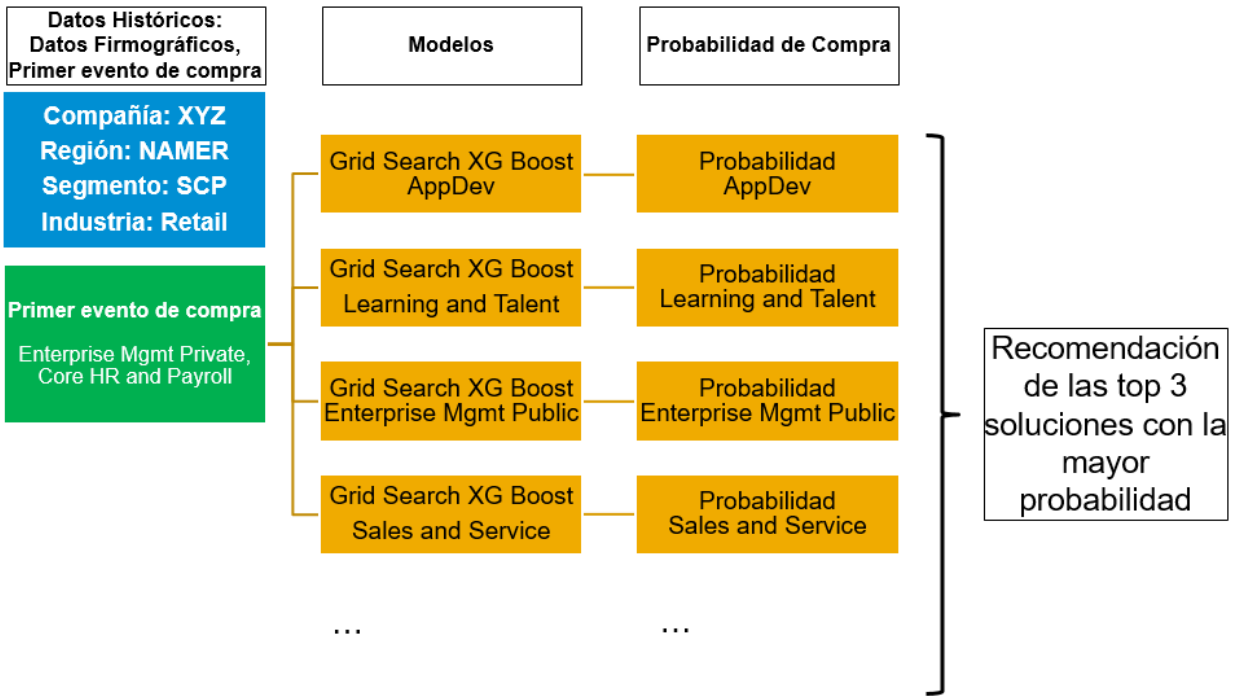
Algunas de las ventajas del método boosting son la capacidad de evitar el sobreajuste de los datos con buenos hiperparámetros, la mejora del rendimiento del método de árboles de decisión, la buena performance ante datos desequilibrados y manejo de outliers, así como también una reducción del sesgo ante la combinación de varios árboles de decisión. Por otro lado, como puntos de desventaja ante otros métodos, el método boosting sufre la sensibilidad ante la elección de hiperparámetros en donde una modificación en un hiperparámetros puede tener un gran impacto en el resultado del modelo. Por otra parte, el alto requerimiento computacional, el impacto que la predicción inicial y calidad de datos pueden tener sobre el modelo también se transforman en desventajas.

Metodología aplicada

En el caso de este trabajo en particular, debido a que las probabilidades a obtener por cliente son de 23, es muy difícil hallar un conjunto de hiperparámetros óptimos para cada uno de los modelos cuyos casos positivos y negativos varían. Por este motivo es que se decide aplicar el método de GridSearchCV junto con cada uno de los modelos, de tal forma que cada uno se ejecuta con un set de hiperparámetros óptimos según los datos de entrenamiento.

En la figura 18, se muestra un ejemplo de la lógica detrás del enfoque para conseguir las recomendaciones.

Figura 18: Diagrama del enfoque de los modelos de recomendación.



Optimizando los modelos

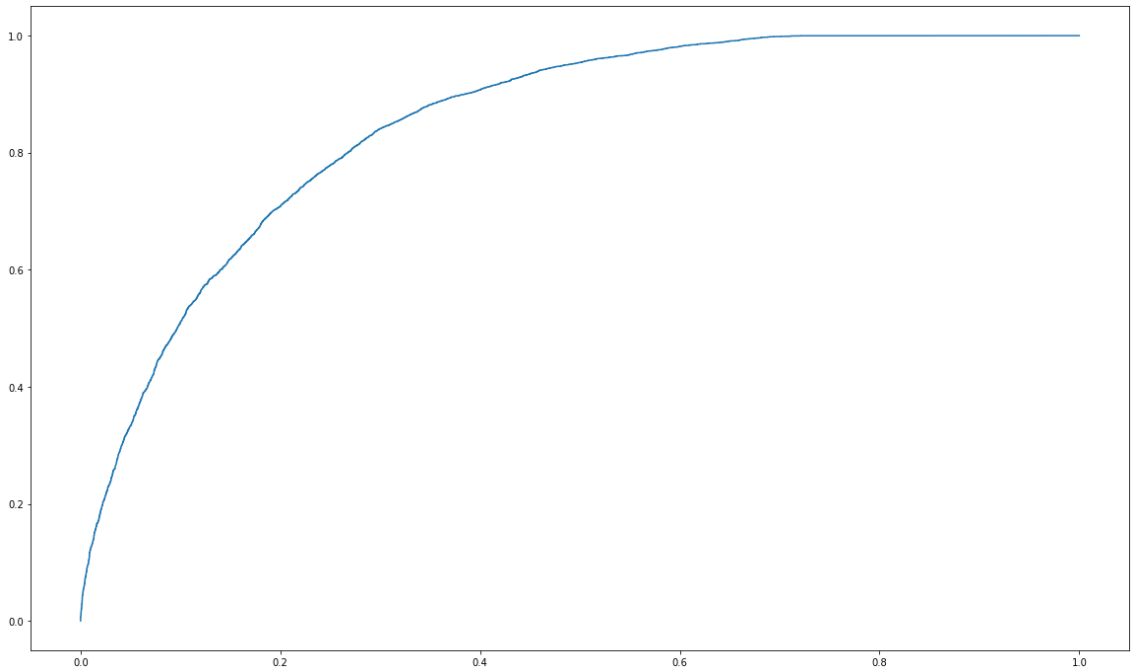
Una vez definido tanto el enfoque del problema y el modelo a usar, el siguiente paso es confirmar qué tan bien balanceados se encuentran los datos, especialmente cuando se trata de un problema de clasificación.

Debido a la gran cantidad de variables objetivo y de casos positivos y negativos necesarios (teniendo en cuenta que son 23 soluciones), considerar el oversampling o undersampling no parecería ser la mejor opción. Al quitar casos negativos de una solución, se le quitan casos positivos a otra, mientras que sumar casos tampoco tendría mucha eficacia ya que cada caso es positivo para una determinada solución. Sin embargo, el peso de clase o classweight, podría ser útil ya que no trata de quitar o sumar casos, sino de asignarles un determinado peso a cada caso minoritario según la solución. En otras palabras, no requiere de un cambio en el conjunto de datos, y mejora el rendimiento del modelo.

Otro paso relevante para optimizar el modelo es la optimización de hiperparámetros. En este caso, al tratarse de 23 modelos, la mejor forma de encontrar los mejores hiperparámetros es ajustándolo a cada uno. Sin embargo, debido a que los datos podrían cambiar de trimestre a trimestre, tampoco es del todo inteligente trabajar con hiperparámetros que sean estáticos y requieran de ajustes cada trimestre. Por esta razón es que se decidió usar un enfoque relacionado al GridSearchCV que permita la búsqueda de los mejores hiperparámetros para cada uno de los modelos, quedándose en este caso con los hiperparámetros que resulten en el mejor AUC (Area Under the Curve), una métrica que calcula el área debajo de la curva ROC (Receiver Operating Characteristic). El AUC proporciona una medida de la capacidad del modelo para distinguir entre muestras positivas y negativas. El AUC tiene un valor en el rango de 0 a 1, donde un valor de 0 indica que el modelo es incapaz de distinguir entre las clases y un valor de 1 indica una capacidad perfecta para distinguir entre las clases. Al lograr optimizar cada modelo de manera independiente y mediante una técnica que se ajusta, tampoco es crítico el reajuste de los parámetros del GridSearchCV para encontrar los mejores resultados de un trimestre a otro.

En la figura 19, se muestra la curva ROC correspondiente al resultado del modelo de “AppDev/Automation and Integration”, con un valor de AUC de 0,84.

Figura 19: Curva ROC correspondiente al modelo de AppDev/Automation and Integration.



Métricas para evaluar los resultados

El paso restante es elegir la métrica correcta para evaluar la performance de los modelos. Al tratarse de 23 modelos independientes, cada uno puede tener una performance distinta. Junto con los modelos se obtiene el resultado de AUC obtenido por cada uno, de manera que permite observar fácilmente que tan bueno fue el desempeño de un modelo, y ver cambios en su performance si se hicieron cambios para mejorarlo. Vale la pena aclarar que estos resultados son a partir del uso de los modelos de machine learning con datos de validación.

Por otro lado, otras métricas que permiten evaluar el desempeño de los modelos son:

- Accuracy: Es la proporción de muestras clasificadas correctamente por el modelo en relación con el total de muestras.

- Precision: Es la proporción de muestras positivas clasificadas correctamente en relación con el total de muestras clasificadas como positivas. Es decir, trata de evaluar el desempeño del modelo a partir de cuantas veces acertó a los casos positivos.

- Recall: Es la proporción de muestras positivas clasificadas correctamente en relación con el total de muestras positivas en el conjunto de datos. En otras palabras, evalúa el porcentaje de compras correctamente predichas.

En la tabla 7, se encuentran los resultados de los 23 modelos llevados a cabo con el método GridSearchCV XGBoost, cada uno ajustado para una solución a predecir. A modo de aclaración, en el caso de Customer Data Solutions, los valores son extremadamente bajos ya que hay un volumen muy bajo de compras en total, lo cual afecta la predictibilidad de éste. A medida que pase el tiempo y aumenten las ventas, los resultados podrían mejorar.

Tabla 7: Valores de Accuracy, Recall y Precision por cada uno de los modelos.

Predicted SubSol	Accuracy	Recall	Precision
AppDev/Automation and Integration	77%	80%	51%
Business Network	90%	84%	7%
Commerce	82%	67%	2%
Core HR and Payroll	76%	80%	24%
Customer Data Solutions	98%	0%	0%
Database and Data Management	76%	77%	44%
Digital Supply Chain	74%	64%	32%
Enterprise Management Private	86%	83%	63%
Enterprise Management Public	90%	87%	42%
ERP for SME	92%	84%	33%
External Workforce	90%	76%	31%
Finance and Q2C	72%	66%	18%
Industry-specific Applications	77%	71%	11%
Learning and Talent	79%	58%	27%
Marketing	89%	48%	2%
Planning and Analytics	74%	76%	31%
Procurement	61%	75%	5%
Sales and Service	73%	81%	57%
Sales Performance Management	93%	65%	21%
SAP Signavio	88%	60%	34%
SuccessFactors Cross	87%	69%	11%
Training and Adoption	76%	64%	17%
Travel and Expense	98%	84%	60%

3. Resultados

Análisis numérico de resultados

Una vez que se definieron tanto el enfoque como los modelos a usar junto con sus métricas, el siguiente paso es analizar los resultados de éstos.

Para comenzar, se pueden verificar los valores correspondientes a la AUC de cada modelo. Como se puede ver, los resultados varían entre un puntaje de 0,69 hasta llegar a un 0,93, sin tener en cuenta el resultado del modelo para la solución Travel and Expense que resultó en un AUC de 0,98. La tabla 8 presenta cada uno de los valores de AUC correspondientes a los modelos.

Vale aclarar que se eligió la métrica AUC ya que es la que mejor funciona en problemas de clasificación de este tipo.

Tabla 8: Valores de AUC correspondientes a cada modelo.

Variable target según el modelo	AUC correspondiente al modelo
AppDev/Automation and Integration_y	0,848
Business Network_y	0,906
Commerce_y	0,707
Core HR and Payroll_y	0,844
Customer Data Solutions_y	0,765
Database and Data Management_y	0,834
Digital Supply Chain_y	0,728
ERP for SME_y	0,933
Enterprise Management Private_y	0,932
Enterprise Management Public_y	0,889
External Workforce_y	0,881
Finance and Q2C_y	0,736
Industry-specific Applications_y	0,782
Learning and Talent_y	0,718
Marketing_y	0,797
Planning and Analytics_y	0,801
Procurement_y	0,693
SAP Signavio_y	0,887
Sales Performance Management_y	0,830
Sales and Service_y	0,698
SuccessFactors Cross_y	0,824
Training and Adoption_y	0,695
Travel and Expense_y	0,983

En el apéndice se encuentran disponibles las ROC de cada modelo, incluyendo el puntaje de AUC correspondiente para cada caso. La razon por la cual el valor de AUC pueden variar entre modelos es debido a la variabilidad de los datos, por lo cual es importante balancear los datos y reducir el ruido en los mismos. Casos como el de Travel and Expense, demuestran que el valor de AUC es elevado ya que en la mayoria de los casos se halla un mismo patrón de compra. Algo

opuesto sucede con los casos de Sales and Service, Procurement o Training and Adoption, en donde los valores son menores debido a la gran variablidad de los datos.

De todas maneras, si bien los valores de AUC son menores para algunas soluciones, es aquí donde se define la importancia de las mismas. Existen soluciones con un mayor grado de impacto comercial en términos de volumen, tales como aquellas ligadas a las bases de datos y el análisis de ventas. Por esta razon es que comercialmente las probabilidades ligadas a soluciones como “Training and Adoption” o “Learning and Talent” no conllevan el mismo peso que soluciones como "AppDev/Automation and Integration", “Business Network”, o “ERP for SME”. Por lo tanto un valor menor en la performance de deteminados modelos no influye de la misma manera sobre la toma de decisiones.

Cuando se trata de la interpretación de los resultados, se puede verificar las probabilidades y resultados por cada uno de los registros. La tabla 9 hallada debajo es un extracto del resultado final de los modelos, en donde se halla la información firmográfica de cada compañía (como ser industria, región, segmento, y el estado de la implementación), seguido por las probabilidades más altas por registro, resultando en la recomendación de tres soluciones como parte del siguiente evento de compra. Junto a estas recomendaciones, las siguientes columnas a la derecha llamadas “real_purchase” contienen la información sobre la compra hecha por la compañía cliente, a modo de validación. Como puede observarse, en la mayoría de los casos los modelos predijeron de manera exitosa cual sería la próxima solución por venderse, aunque en algunos no fue de esa manera. En la figura 20 se puede ver un ejemplo de los resultados de los modelos.

Figura 20: Ejemplo de los resultados de los modelos.

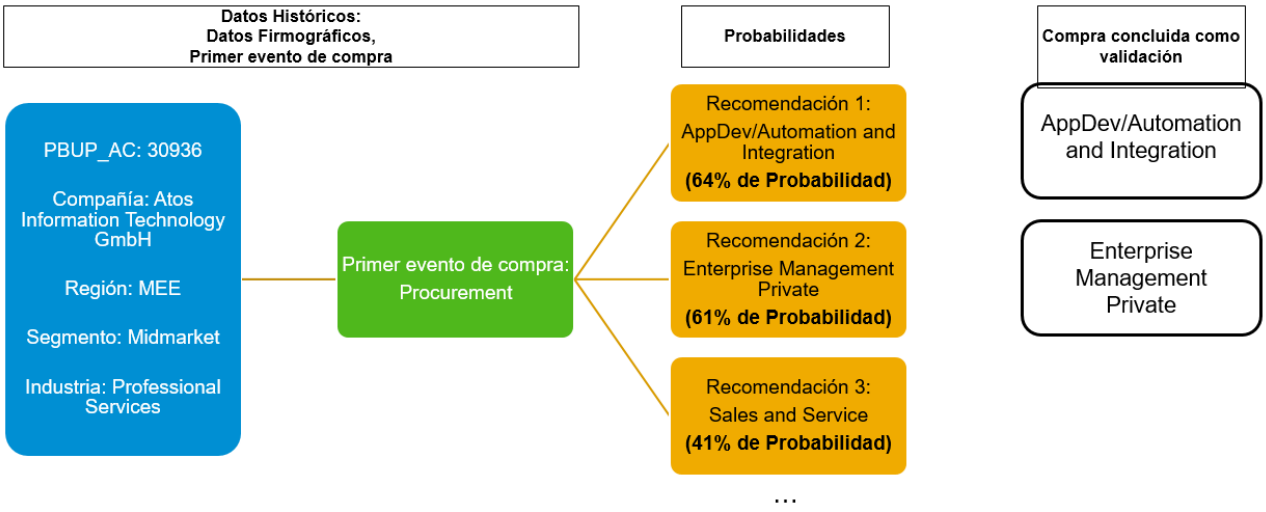


Tabla 9: Extracto de los resultados de los modelos

PBUP_AC	Industria	Región	Segmento	Estado de la implementación
30908	Oil, Gas and Energy	Middle & Eastern Europe	Industry Led	Project
30909	Mill Products and Mining	Middle & Eastern Europe	Industry Led	Live
30912	Oil, Gas and Energy	Middle & Eastern Europe	Midmarket	Shelfware
30936	Professional Services	Middle & Eastern Europe	Midmarket	Shelfware
31075	High Tech	Middle & Eastern Europe	Industry Led	Live
31076	Wholesale Distribution	EMEA North	Midmarket	Live
31102	Banking	Middle & Eastern Europe	Industry Led	Project
31114	Professional Services	EMEA North	Midmarket	Shelfware
31125	Consumer Products	EMEA North	Midmarket	Shelfware
31133	Professional Services	EMEA North	Midmarket	Shelfware
31148	High Tech	EMEA North	Midmarket	Shelfware
31152	Chemicals	EMEA North	Industry Led	Live
31168	Wholesale Distribution	EMEA North	SCP	Shelfware
31170	Insurance	EMEA North	Industry Led	Live
31284	Mill Products and Mining	EMEA North	Midmarket	Project
31293	Utilities	EMEA North	Midmarket	Shelfware
31297	Retail	EMEA North	Industry Led	Live
31302	Banking	EMEA North	Midmarket	Shelfware
31303	Mill Products and Mining	Middle & Eastern Europe	Industry Led	Shelfware
31304	Public Sector	EMEA North	Industry Led	Shelfware
31306	Wholesale Distribution	EMEA North	Industry Led	Live
31321	Consumer Products	Middle & Eastern Europe	Industry Led	Live
31322	Professional Services	EMEA North	Midmarket	Shelfware
31331	Oil, Gas and Energy	EMEA North	Midmarket	Live
31340	Professional Services	EMEA South	Midmarket	Shelfware
31348	Life Sciences	EMEA North	Midmarket	Live
31356	Industrial Manufacturing	EMEA North	Midmarket	Live
31360	Consumer Products	EMEA North	Midmarket	Live
31475	Chemicals	EMEA North	Industry Led	Live
31477	High Tech	EMEA North	Midmarket	Shelfware
31509	Consumer Products	EMEA North	Industry Led	Live
31513	Industrial Manufacturing	EMEA North	Midmarket	Shelfware
31529	Automotive	EMEA North	Midmarket	Live
31534	Defense and Security	EMEA North	Industry Led	Shelfware
31672	Professional Services	EMEA North	Midmarket	Shelfware
31684	Consumer Products	EMEA North	SCP	Live

Rec_score_1	Rec_sol_1	Rec_score_2	Rec_sol_2
56%	AppDev/Automation and Integration	26%	Database and Data Management
93%	Enterprise Management Private	63%	Finance and Q2C
64%	Digital Supply Chain	55%	Industry-specific Applications
83%	Enterprise Management Private	45%	AppDev/Automation and Integration
90%	Enterprise Management Private	23%	AppDev/Automation and Integration
89%	AppDev/Automation and Integration	38%	Planning and Analytics
79%	Enterprise Management Private	26%	Digital Supply Chain
81%	AppDev/Automation and Integration	58%	Finance and Q2C
75%	Finance and Q2C	55%	Enterprise Management Private
77%	Sales and Service	60%	Database and Data Management
90%	Sales and Service	61%	Digital Supply Chain
90%	Enterprise Management Private	24%	AppDev/Automation and Integration
94%	Enterprise Management Private	20%	Finance and Q2C
94%	Enterprise Management Private	16%	Finance and Q2C
71%	AppDev/Automation and Integration	5%	Finance and Q2C
94%	Enterprise Management Private	41%	Industry-specific Applications
72%	Enterprise Management Private	32%	Planning and Analytics
78%	Planning and Analytics	55%	ERP for SME
89%	Enterprise Management Private	30%	AppDev/Automation and Integration
84%	Enterprise Management Private	33%	AppDev/Automation and Integration
66%	Digital Supply Chain	53%	Core HR and Payroll
94%	SuccessFactors Cross	81%	AppDev/Automation and Integration
66%	AppDev/Automation and Integration	63%	Planning and Analytics
96%	Industry-specific Applications	47%	Learning and Talent
75%	AppDev/Automation and Integration	64%	Sales and Service
94%	Enterprise Management Private	67%	Procurement
15%	AppDev/Automation and Integration	12%	Planning and Analytics
61%	Database and Data Management	33%	Finance and Q2C
63%	Enterprise Management Private	22%	Core HR and Payroll
58%	AppDev/Automation and Integration	21%	Database and Data Management
90%	Enterprise Management Private	38%	AppDev/Automation and Integration
87%	AppDev/Automation and Integration	37%	Database and Data Management
44%	Enterprise Management Private	38%	AppDev/Automation and Integration
97%	Enterprise Management Private	18%	Learning and Talent
74%	Digital Supply Chain	70%	AppDev/Automation and Integration
73%	AppDev/Automation and Integration	70%	Finance and Q2C

Rec_score_3	Rec_sol_3	real_purchase_1	real_purchase_2
16%	Planning and Analytics	AppDev/Automation and Integration_y	
22%	Digital Supply Chain	Enterprise Management Private_y	
53%	AppDev/Automation and Integration	Industry-specific Applications_y	
40%	Finance and Q2C	AppDev/Automation and Integration_y	Enterprise Management Private_y
13%	Finance and Q2C	Enterprise Management Private_y	
20%	Finance and Q2C	AppDev/Automation and Integration_y	
23%	Finance and Q2C	Enterprise Management Private_y	
28%	Digital Supply Chain	AppDev/Automation and Integration_y	
12%	AppDev/Automation and Integration	Enterprise Management Private_y	
59%	AppDev/Automation and Integration	Database and Data Management_y	
57%	AppDev/Automation and Integration	Training and Adoption_y	
11%	Finance and Q2C	Finance and Q2C_y	
3%	Database and Data Management	Enterprise Management Private_y	
13%	AppDev/Automation and Integration	Finance and Q2C_y	
1%	Database and Data Management	AppDev/Automation and Integration_y	
7%	AppDev/Automation and Integration	Sales and Service_y	
30%	AppDev/Automation and Integration	Enterprise Management Private_y	
47%	AppDev/Automation and Integration	AppDev/Automation and Integration_y	
9%	Finance and Q2C	Enterprise Management Private_y	
10%	Training and Adoption	Enterprise Management Private_y	
53%	AppDev/Automation and Integration	Planning and Analytics_y	
68%	Core HR and Payroll	Core HR and Payroll_y	
41%	Core HR and Payroll	Procurement_y	Sales and Service_y
44%	Finance and Q2C	Procurement_y	
40%	Database and Data Management	Learning and Talent_y	
46%	Commerce	Enterprise Management Private_y	
4%	Finance and Q2C	Planning and Analytics_y	

21%	ERP for SME	Database and Data Management_y	
16%	Planning and Analytics	Enterprise Management Private_y	
11%	Finance and Q2C	Procurement_y	Sales and Service_y
28%	Procurement	Enterprise Management Private_y	
13%	Finance and Q2C	AppDev/Automation and Integration_y	
26%	Planning and Analytics	Enterprise Management Private_y	
8%	AppDev/Automation and Integration	Enterprise Management Private_y	

Dados los datos presentados, podría decirse que los modelos predicen con un 72% de accuracy que por lo menos sea una de las soluciones recomendadas en el top 3 fue comprada como parte del segundo evento de compra.

Importancia de las variables

Respecto a la importancia de las variables, los resultados fueron distintos de solución a solución. Si bien en todos los casos lo más importante está definido por la compra anterior, el resto de las variables tales como industria o segmento también fueron relevantes en algunos escenarios. Por ejemplo, en el caso de la solución "AppDev/Automation and Integration" se nota que la solución "Enterprise Management Private" es la variable mas importante, seguida por la de "ERP for SME" y otras variables que incluyen una región en mucho menor medida. Sin embargo, la solución "ERP for SME" muestra un ejemplo completamente distinto, donde el top 3 de variables relevantes están conformadas por "Core HR and Payroll", "Database and Data Management" y el segmento "Midmarket", seguidas por la industria "Defense and Security". Esto demuestra que en mayor o menor medida todas las variables son parte de la probabilidad resultante.

Impacto comercial de los resultados

Mas allá de los resultados numéricos y la evaluación de la performance de cada uno de los modelos, también debe traerse a mención el impacto comercial que tienen estos resultados. En otras palabras, es gracias a estas recomendaciones que se hallaron patrones útiles en la manera de abordar un cliente que pueden llevar a un cumplimiento de la estrategia comercial de SAP.

Para comenzar a entender este impacto es importante conocer cuál sería la compra óptima o ciclo de compra óptimo que una compañía podría tener, independientemente si ya es cliente o no. Este ciclo de compra optimo este dado por las soluciones que mayor valor tienen para SAP, es decir aquellos que no sólo tienen un alto rendimiento financiero, sino que también dan lugar a la compra de otras soluciones. Tal es, por ejemplo, el caso de las soluciones de ERP for SME,

soluciones relacionadas a Enterprise Management, y AppDev/Automation and Integration. Estas soluciones son las principales por las cuales las compañías clientes suelen, en su gran mayoría, volcarse luego a otras soluciones de ecosistema. Es así como los resultados de este trabajo muestran que hay compañías que en primer lugar compraron una solución que es parte de este núcleo de alto valor y que tienen una mayor probabilidad que su siguiente compra sea otra solución de este núcleo.

Por otro lado, también existe una minoría de clientes que su primera compra no fue un producto que formara parte de este núcleo de alto valor. Por ejemplo, su primera compra pudo tratarse directamente de una solución basada en recursos humanos, financiera o de capacitación. Es aquí donde la siguiente compra tiene un papel fundamental en el desarrollo de la relación comercial con el cliente. Por ejemplo, si la primera compra de una compañía fue Finance and Q2C, la mejor opción seria llevar a esa compañía a que su siguiente compra sea una solución de alto valor para SAP. Es por esto por lo que las probabilidades resultantes de este trabajo ayudan a identificar las oportunidades de venta de SAP que podrían llevar a una mayor integración al ecosistema.

Finalmente, otro beneficio de este trabajo es que, debido a que una gran parte de los clientes tienen una sola compra, el uso de estas probabilidades de cross-sell junto a la búsqueda de integración de los clientes al núcleo de soluciones de alto valor de SAP, podría llevar a replantear el abordaje comercial en determinados clientes.

4. Conclusión

El proceso de ventas en cualquier compañía requiere de mucho trabajo, en especial la parte de detectar posibles compradores de acuerdo con sus perfiles, encontrar cuales serían los productos que mejor se ajustan a sus necesidades, y a su vez llegar a una oferta que realmente pueda ser más que interesante para el comprador.

Varias son las compañías que, muchas veces sin notarlo, dan el mismo trato a clientes o compradores que buscan distintos productos o tienen distintas necesidades, como por ejemplo compradores que pertenecen a diferentes industrias o que manejan un ingreso completamente distinto. Es por esto por lo que es importante no solo hallar el perfil adecuado de comprador, sino también llegarle con una oferta que sea interesante desde un comienzo.

El trabajo de análisis realizado, junto con los modelos de predicción desarrollados en base a las características y comportamientos de las compañías compradores, podría significar un paso positivo tanto para el departamento comercial de la compañía como para la formación de una estrategia corporativa a futuro. Al tratarse del primer trabajo de características similares hecho con datos propios de SAP, los resultados muestran un claro potencial hacia el uso más eficiente de los recursos de la compañía para la venta, y más importante aún, un modelo de negocios que da lugar a muchas mejoras en los modelos.

Cabe destacar que habiendo trabajado con datos de compañías y de compras anteriores de compañías clientes, y por las limitaciones de tiempo, el trabajo se basó en descubrir que tan bien era posible predecir con éxito las soluciones que serían parte del segundo evento de compra por parte de una compañía, sin importar eventos de compra futuros. Esto permitió no sólo limitar el trabajo y facilitarlo en términos de complejidad, sino que también significa que, con cierto esfuerzo, es posible trabajar sobre el 37% de la base de clientes que tienen más de un evento de compra con SAP. Visto de otra forma, el resultado del trabajo demuestra que sería posible predecir el segundo evento de compra para el 63% de la base total de clientes de la compañía, ayudando así a lograr ventas en una base de clientes que prácticamente tienen mucho que descubrir con SAP.

Por medio de los resultados, se muestran dos conclusiones al problema enfocado de cross-sell. La primera conclusión es que es posible encontrar un patrón en común entre las compañías cliente y, lo que es mejor, usarlo para predecir cuál sería el mejor enfoque de ventas a futuro sobre cada una de las compañías, teniendo en cuenta sus características firmográficas. Si bien todavía queda lugar para mejorar los modelos, un aproximado de 72% como accuracy indica que en varios casos la predicción fue correcta. Sin embargo, vale aclarar que este trabajo se basó en encontrar la solución que sería parte del segundo evento de compra, sin tomar en cuenta eventos de compra

que suceden en tercer, cuarto, o quinto lugar por decir algunos. Los modelos toman únicamente los valores existentes hasta el segundo evento de compra, tanto para entrenamiento como para validación.

La segunda conclusión es que este trabajo muestra que se puede mejorar el uso de los recursos de la compañía, tanto recursos humanos como también recursos tecnológicos, para alcanzar una mejora en los niveles de venta. Por un lado, se verifica la hipótesis sobre la necesidad de cambiar el trato hacia compañías que ya son clientes de SAP, siendo las probabilidades de éxito en las ventas son mayores si se abordan con determinada solución por medio. Por ejemplo, segmentando vendedores de soluciones de acuerdo con las probabilidades de venta para un grupo de clientes. Por otro lado, también se demuestra que enfocar estos mismos recursos en los clientes correctos o el desarrollo interno de tecnología en una mejora constante de un sistema de recomendación, mejorarían en gran parte las ventas. Todo esto se ve respaldado por los resultados de los modelos, siendo que se halló un conjunto de soluciones que forman un núcleo de gran valor para futuras ventas. Explicado de otra manera, se halló que soluciones como "AppDev/Automation and Integration", "ERP for SME", o "Enterprise Management" son fundamentales para llevar al cliente a un alto nivel de integración al ecosistema SAP.

En términos comerciales, estos resultados muestran que modelos de machine learning que ayuden a hallar la probabilidad de una venta pueden tener un alto impacto tanto en niveles comerciales como estratégicos. Por un lado, comercialmente el algoritmo podría ayudar al uso más eficiente de los recursos de la compañía y enfoque de los vendedores, logrando así aumentar las ventas de ciertas soluciones donde el vendedor entienda que hay una mayor probabilidad. Si bien el desarrollo de estos modelos no tiene bajo ningún punto de vista el reemplazo de la experiencia que puede tener un vendedor, tiene como objetivo complementarlo con el fin de detectar oportunidades con mayor rapidez y profundizar aún más el entendimiento de sus posibles compradores. Otra ventaja en términos comerciales, dado que se trabaja con compañías que ya habían comprado con SAP, es el incremento del customer engagement, logrando que los clientes exploren y sean aún más parte del ecosistema propuesto por SAP.

El trabajo no impacta únicamente en el área comercial, sino que también en la parte estratégica de la compañía. Una visión más amplia sobre la elección de las compañías cliente a la hora de decidir por una segunda solución podría ayudar a tener un mejor entendimiento del mercado, hallar tendencias de compra en determinados sectores, como por ejemplo el automotriz o retail, mejorar la imagen de la compañía como resultado de una mejora en las ventas, y finalmente pensar en nuevas ideas ya sean productos o métodos de alcanzar un sector.

Si bien se considera que se obtuvieron resultados interesantes, esto no quita que los modelos tengan lugar para mejoras. Algunas de las mejoras sobre las cuales podría tratarse en el

corto y mediano plazo pueden incluir el uso de datos obtenidos a partir de terceros, como por ejemplo las soluciones que una compañía cliente puede tener de otra compañía que compite con SAP, o el uso de más datos relacionados al tiempo de uso que una compañía cliente le puede dar a una solución de SAP. Otro punto que podría incluirse a futuro es la implementación de una mayor cantidad de reglas de negocios que sirvan como límites para el modelo, es decir que si determinada solución tiene determinada características que podrían significar una mejor compatibilidad con otra, podría aumentar sus probabilidades de venta. A su vez, podría darse el escenario en el que SAP decide concentrar su estrategia en determinadas soluciones, con lo cual también podría ser de utilidad para guiar a los modelos.

Otros puntos por mejorar que dependen solamente del desarrollo de los modelos es el nivel de complejidad que éstos tienen. Debido a que se trata de 23 modelos ejecutados de manera iterativa, el requerimiento computacional hace que tome cerca de 24 horas en ejecutarse por completo, lo cual puede ser un problema si necesita ser ejecutado varias veces en el transcurso de una semana, por ejemplo. Sin embargo, debido a que la información de historiales de compra se actualiza trimestralmente, en un comienzo el tiempo de ejecución no resulta el principal obstáculo a trabajar. Por esta razón es que se decidió seguir con los modelos más allá de su tiempo de ejecución.

Finalmente, siempre es útil mencionar que la performance de los modelos mejorará a medida que se recolecten más datos de compañías y compras. En el futuro el enfoque de los modelos podría ampliarse para tomar en cuenta también eventos de compra que suceden luego del segundo y así agregar compañías que tienen muchos más eventos de compra con SAP, dado que aquí hemos incluido solamente las compañías que tienen exactamente dos eventos de compra.

Este trabajo demostró que es posible hallar un patrón de compra entre las muchas compañías cliente de SAP y el valor comercial que podría tener un análisis de este tipo.

5. Apéndice

Un recorrido por todas las soluciones de SAP

Debido a que en este trabajo se mencionarán la distintas soluciones y las correlaciones que pueden existir entre ellas, es importante mencionarlas y describirlas. Es así como dentro de las 10 soluciones que SAP ofrece se encuentran ERP y Finance, CRM and Customer Experience, Network and Spend Management, Digital Supply Chain, HR and People Engagement, Experience Management, Business Technology Platform, Digital Transformation, Small and Midsize Enterprises, e Industry Solutions. Para entender un poco más sobre cada una, debajo se hace una breve descripción de sus capacidades.

- ERP y finanzas: Estas soluciones tienen como fin ayudar a la planificación y análisis financieros, cierres financieros y contables, gestión de tesorería, gestión de cuentas por cobrar, facturación e ingresos, gobernanza, riesgo y cumplimiento y ciberseguridad. Cada una de estas herramientas permite el uso de datos en tiempo real y la automatización de varios procesos para ganar tiempo.

- CRM y experiencia del cliente: CRM (Customer Relationship Management) o "gestión de relaciones con el cliente" en español, se trata de un software que analiza y gestiona las interacciones y los datos del cliente a lo largo de todo el ciclo de vida del cliente, creando así mejores experiencias de cliente, un mejor servicio al cliente y mejores relaciones de negocio a la vez que aumenta los ingresos. Por otro lado, la experiencia del cliente (CX) es la manera en que las interacciones con su marca o producto hacen sentir a los clientes. Un cliente puede ser un negocio (B2B) o una persona (B2C) y el espíritu de honrar su recorrido es el mismo. Es así como estas herramientas pueden lograr convertir satisfactoriamente cada oportunidad de mercado, generar fidelidad con clientes, y escalar una empresa.

- Gestión de red y gastos: Estas soluciones ayudan a cualquier empresa a desarrollar una comprensión más profunda de sus gastos con el objetivo de reducirlos y entender como impacta cada de decisión en los gastos empresariales. Cada gasto a su vez es parte de una integración de punta a punta con los procesos de negocios como ser procesos de pagos. Otro punto a tener en cuenta es que esta herramienta también permite capturar y analizar datos sobre gastos para lograr un ahorro eficiente.

- Cadena de suministro digital: La cadena de suministro digital o DSC (por sus siglas en inglés, Digital Supply Chain), es una solución que apunta hacia el área operativa de una

compañía. Tiene el fin de crear una cadena de suministros totalmente interdependientes para mantenerse resiliente, ágil, productivo y sostenible frente a situaciones que pueden surgir. Una cadena de suministros sustentable implica la integración de prácticas viables desde el punto de vista ambiental y financiero, dentro de todo el ciclo de vida de la cadena de suministro, desde diseño y desarrollo del producto hasta selección de materiales (incluso la extracción de materias primas y la producción agrícola), fabricación, embalaje y transporte. Una cadena de suministros resiliente es la capacidad de una cadena de suministro para persistir, adaptarse o transformarse frente al cambio. Esto se vio principalmente en el impacto del COVID-19 durante todo el año 2020 y varios modelos de negocios que no lograron adaptarse a cambios bruscos debido a una disrupción continua, los cambios en la demanda y la incertidumbre. A su vez, también se intenta que cada cadena de suministros sea lo más ecológica y transparente posible, teniendo un bajo impacto ambiental y transparencia ante socios y organismos comerciales.

- RR. HH. e interacción con el personal: HXM (Human Experience Management por sus siglas en inglés) se trata de un software con enfoque en el área de Recursos Humanos de una compañía. Dentro de sus herramientas, se encuentran la gestión de experiencia del empleado, entendiendo mejor sus objetivos, necesidades y deseos; RR. HH. centrales y nómina, con un seguimiento de horarios y gestión de beneficios en la nube; gestión de talento, gracias al cual se puede trabajar en las habilidades de los empleados por medio de capacitaciones, evaluación de desempeño, remuneraciones, y desarrollo; y analíticas de RR. HH., que permite analizar y entender en mayor profundidad cuales son los puntos fuertes y a mejorar del sector.

- Gestión de experiencias: La gestión de experiencias sirve no solo para mejorar el compromiso del empleado por medio de encuestas e informes, sino que también mejora la experiencia del empleado, gestionar el ciclo del empleado, evaluar su rendimiento de manera integral, optimizar los beneficios otorgados, y mejorar la experiencia de un candidato.

- Business Technology Platform: BTP agrupa la gestión de datos, analíticas, inteligencia artificial, el desarrollo de aplicaciones, la automatización e integración en un único entorno unificado. BTP crea experiencias personalizadas en todos los procesos de negocio, construye aplicaciones, analíticas e integraciones más rápido, y opera la innovación crítica con confianza en la infraestructura de los principales proveedores en la nube. Los desarrolladores tienen lo que necesitan para conectar, extender y enriquecer los procesos de negocio críticos rápidamente. Los usuarios de negocio pueden

automatizar tareas, crear flujos de trabajo rápidos y flexibles o personalizar interfaces sin codificación.

- Transformación digital: La transformación digital de una compañía es vital para una estrategia de transformación de negocios general. Las tecnologías adecuadas, junto con personas, procesos y operaciones, les dan a las organizaciones la capacidad de adaptarse rápido a la disrupción y/o las oportunidades, cumplir con las nuevas y cambiantes necesidades del cliente, e impulsar el futuro crecimiento e innovación, a menudo de maneras inesperadas. SAP logra impulsar esta transformación digital por medio de su ERP en la nube que va de la mano con tecnologías digitales inteligentes como la inteligencia artificial, machine learning, redes de internet de las cosas, analíticas avanzadas y automatización robótica de procesos (RPA por sus siglas en inglés) con el fin de transformar una compañía y sus procesos de negocio, modelos de negocios, y su organización y cultura.

- Pequeñas y medianas empresas: SAP también es capaz de dar soporte a pequeñas y medianas empresas por medio de soluciones adaptadas a sus necesidades. Algunas de estas soluciones son SAP Business One como ERP, HCM como módulo de RR. HH., CRM y CX, DSC, gestión de red y gastos con SAP Business Network, análisis de datos y resultados con BTP, y paquetes de partners especializados en distintas áreas e industrias. Es así como por medio de tecnologías modernas, SAP puede llevar el negocio a la nube.

- Soluciones para la industria: En el caso de algunas industrias, aplicaciones más específicas son necesarias. SAP ofrece RISE with SAP para acelerar la transformación digital de las compañías. RISE es una solución integral que se compone de productos clave, como SAP S/4HANA Cloud, así como prácticas futuras de la industria, business process intelligence y servicios de SAP y Partners.

A su vez, además de las soluciones también se hallan, en un nivel más granular, la sub-soluciones. Estas son aquellas que pertenecen a una determinada solución y que cumplen con una función específica. Hoy en día, SAP cuenta con 29 sub-soluciones que pertenecen a distintas soluciones. Debido a que en este trabajo también se trataran la sub-soluciones, se hará un breve repaso de cuáles son los sub-soluciones.

Business Technology Platform (BTP)

- Database and Data Management

- Planning and Analytics
- AppDev/Automation and Integration

Intelligent Spend & Business Network

- Business Network
- Procurement
- External Workforce
- Travel and Expense

SuccessFactors (HXM)

- Core HR and Payroll
- Learning and Talent
- Sales Performance Management

Customer Experience (CX)

- Commerce
- Customer Data Solutions
- Marketing
- Sales and Service

SAP Signavio

- SAP Signavio

Cloud Enterprise Resource Planning (Cloud ERP)

- Digital Supply Chain
- ERP for SME
- Finance and Q2C
- Industry-specific Applications
- Enterprise Management Private
- Enterprise Management Public

Supporting Solution Areas

- SuccessFactors Cross
- Training and Adoption

Es importante saber para este trabajo, que se analizaran aquellas compañías clientes que cuentan con el Cloud ERP, más específicamente Enterprise Management Private (EM Private) y Enterprise Management Public (EM Public), como parte de su base instalada.

Gráficos y Tablas

Tabla 1: Extracto de Adoption Monitor File

	PBUP_AC	DEAL_DATE	MAX_STATE	2023 Solution Area	Sub-Solution Area
0	5026459	20220630.00	Shelfware	Business Technology Platform	Database and Data Management
1	19847728	20220601.00	Shelfware	Business Technology Platform	Planning and Analytics
2	11134855	20230101.00	Live	Intelligent Spend & Business Network	Travel and Expense
3	1188223	20090201.00	Shelfware	Business Technology Platform	Database and Data Management
4	2716173	20140620.00	Live	Cloud ERP	ERP for SME
5	35500243	20211228.00	Live	Cloud ERP	ERP for SME
6	9167040	20140623.00	Live	Cloud ERP	ERP for SME
7	13521979	20220526.00	Shelfware	Cloud ERP	Digital Supply Chain
8	7196343	20071129.00	Shelfware	Business Technology Platform	Database and Data Management
9	50012079	20211004.00	Live	Business Technology Platform	Database and Data Management
10	5735316	20220101.00	Live	Customer Experience	Sales and Service
11	1160265	20180921.00	Live	Business Technology Platform	AppDev/Automation and Integration
12	1084812	20111221.00	Shelfware	Cloud ERP	Industry-specific Applications
13	25026689	20230101.00	Project	Business Technology Platform	AppDev/Automation and Integration
14	163221	20110517.00	Live	Customer Experience	Sales and Service
15	2918824	20100504.00	Live	Business Technology Platform	AppDev/Automation and Integration
16	30168949	20181214.00	Live	Cloud ERP	ERP for SME
17	33091809	20230118.00	Project	Intelligent Spend & Business Network	Business Network
18	39405	20221001.00	Shelfware	Business Technology Platform	Planning and Analytics
19	17122185	20220530.00	Shelfware	Cloud ERP	ERP for SME
20	13745699	20150519.00	Shelfware	Business Technology Platform	Database and Data Management
21	7814300	20190627.00	Live	Cloud ERP	Enterprise Management Private
22	3868609	20220330.00	Project	Supporting Solution Areas	Training and Adoption
24	7433002	20070216.00	Shelfware	Business Technology Platform	Database and Data Management
25	8379289	20230101.00	Live	Intelligent Spend & Business Network	Travel and Expense
26	32394180	20230101.00	Live	Business Technology Platform	AppDev/Automation and Integration
27	1235790	20010906.00	Shelfware	Business Technology Platform	AppDev/Automation and Integration
28	14346861	20130812.00	Shelfware	Business Technology Platform	AppDev/Automation and Integration
29	50010913	20171222.00	Live	Business Technology Platform	Planning and Analytics
30	14360455	20130613.00	Live	Business Technology Platform	Database and Data Management

Tabla 2: Extracto de CRM data

	PBUP_AC	EmployeeSize	CompanyRevenue	SAP_Mastercode	Region_Label	Country	Market_Unit_Label	Internal_Account_Classification
0	7064318	nan	nan	Aerospace and Defense	North America	USA	Northeast	Industry Led
1	8780083	nan	nan	Professional Services	North America	USA	West	Midmarket
2	9396153	nan	nan	Banking	North America	USA	West	Midmarket
3	12355346	nan	nan	Mill Products and Mining	North America	USA	West	Midmarket
4	35775090	nan	nan	Professional Services	North America	USA	Midwest	Midmarket
5	33841442	nan	nan	Engineering, Construction and Operation	North America	USA	Midwest	Midmarket
6	8723512	nan	nan	Wholesale Distribution	North America	USA	Northeast	Midmarket
7	10560201	nan	nan	Banking	North America	USA	Northeast	Midmarket
8	34405721	nan	nan	Defense and Security	North America	USA	Regulated Ind	Midmarket
9	21288178	nan	nan	Professional Services	North America	USA	South	Midmarket
10	36464870	38.00	nan	Professional Services	North America	USA	West	Midmarket
11	12367658	nan	nan	Professional Services	North America	USA	West	Midmarket
12	12320644	nan	nan	Engineering, Construction and Operation	North America	USA	South	Midmarket
13	23156897	nan	nan	Professional Services	North America	USA	South	Midmarket
14	26360185	nan	nan	Professional Services	North America	USA	South	Midmarket
15	50251192	nan	nan	Professional Services	North America	USA	South	Midmarket
16	11571496	nan	nan	Professional Services	North America	USA	Midwest	Midmarket
17	50244514	nan	nan	Professional Services	North America	USA	South	Midmarket
18	33855012	nan	nan	Professional Services	North America	USA	Midwest	Midmarket
19	34632624	nan	nan	Professional Services	North America	USA	Northeast	Midmarket
20	50102022	nan	nan	High Tech	North America	USA	Northeast	Midmarket
21	17394171	nan	nan	Utilities	North America	USA	Regulated Ind	Midmarket
22	27974962	18.00	nan	Higher Education and Research	North America	USA	Regulated Ind	Midmarket
23	33920395	10.00	nan	Professional Services	North America	USA	Midwest	Midmarket
24	27226744	190.00	nan	Wholesale Distribution	North America	USA	Midwest	Midmarket
25	27271268	208.00	nan	Consumer Products	North America	USA	Northeast	Midmarket
26	168845	200.00	nan	Higher Education and Research	North America	USA	Regulated Ind	Industry Led
27	10362138	nan	nan	Wholesale Distribution	North America	USA	South	Midmarket
28	29276177	nan	nan	Professional Services	North America	USA	Midwest	Midmarket
29	28916763	nan	nan	Professional Services	North America	USA	Northeast	Midmarket

Gráficos sobre las correlaciones entre variables independientes y dependientes

Figura 21: Correlación entre las variables independientes y la solución AppDev/Automation and Integration como parte del segundo evento de compra.

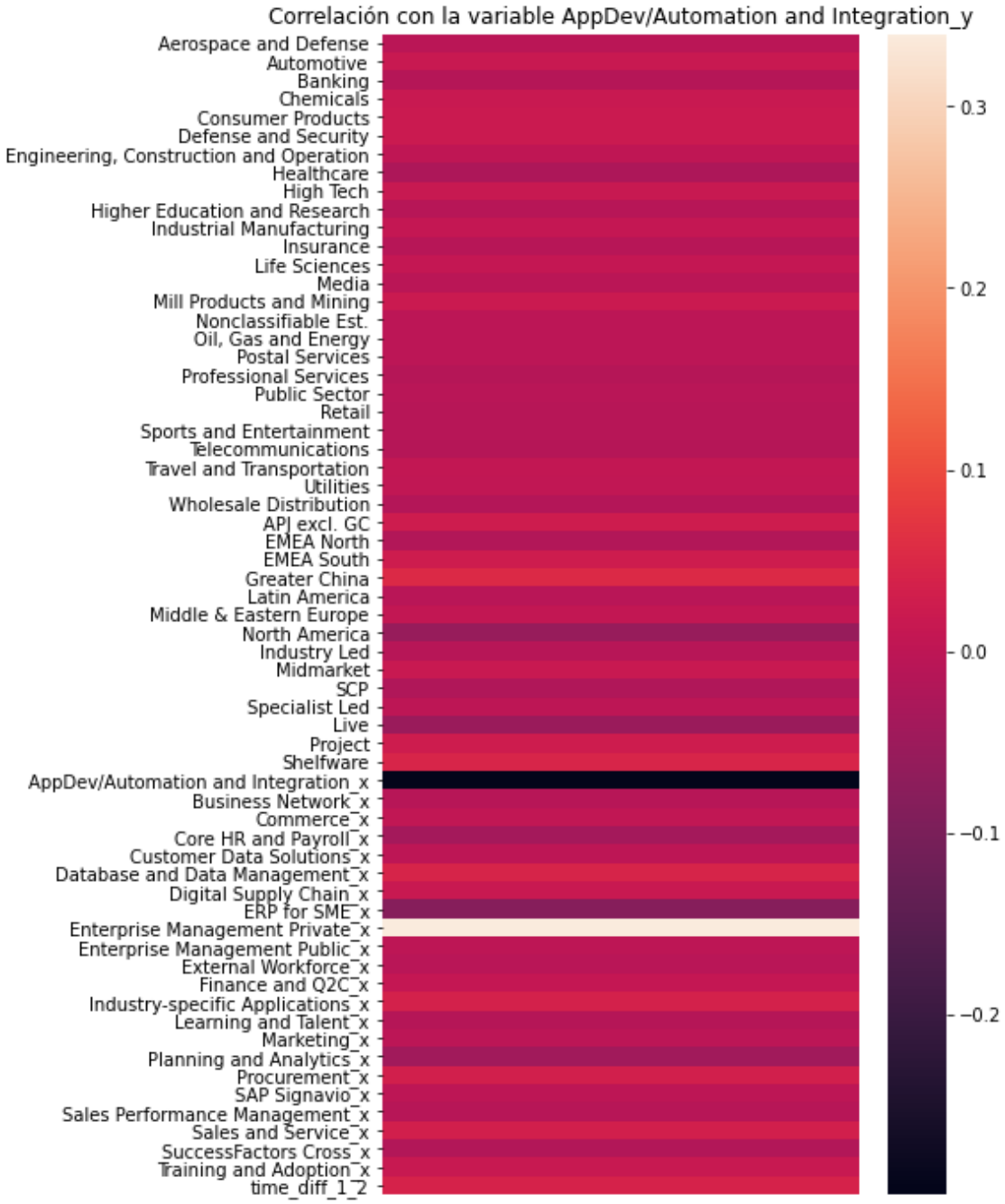


Figura 22: Correlación entre las variables independientes y la solución Business Network como parte del segundo evento de compra

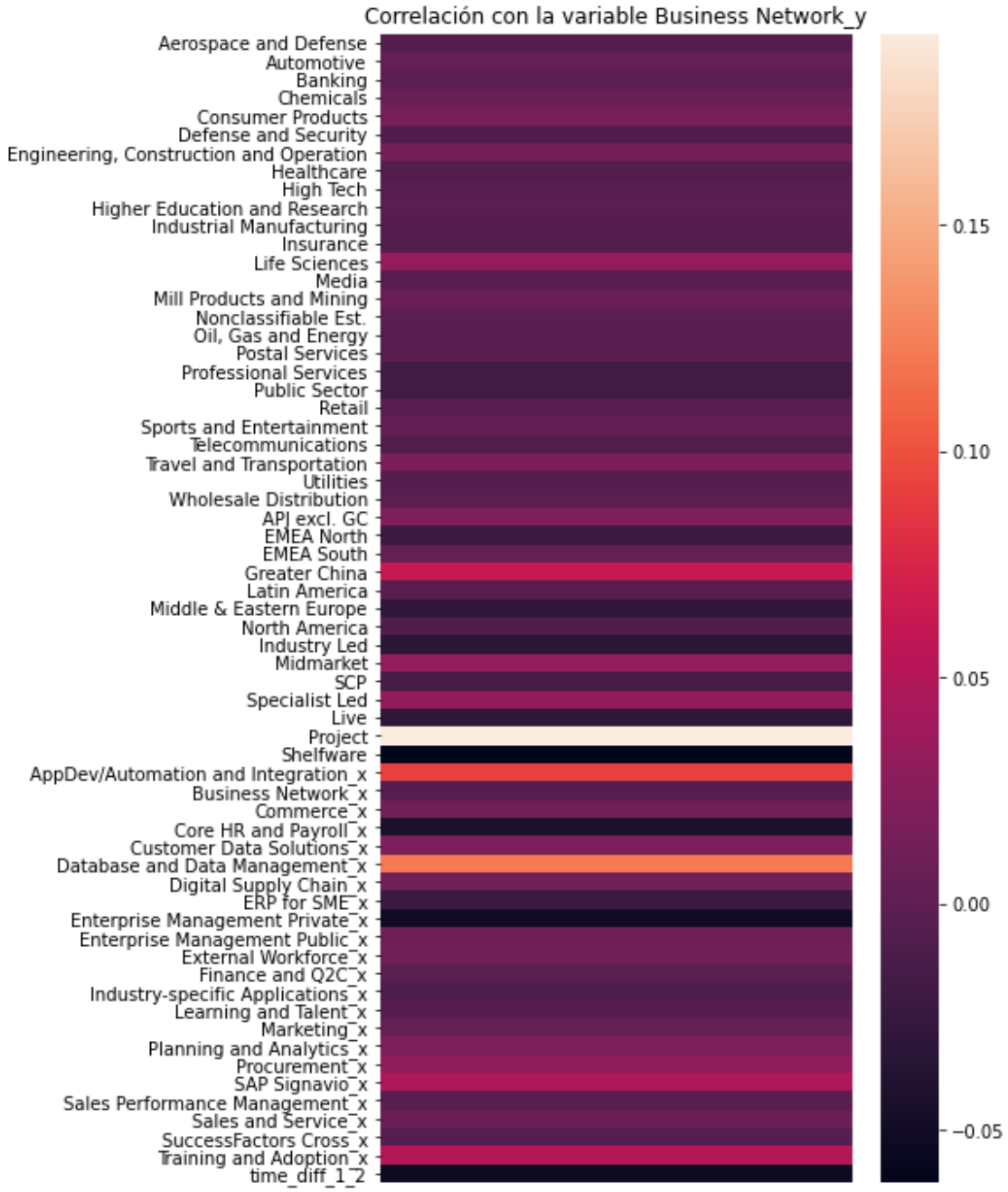


Figura 23: Correlación entre las variables independientes y la solución Commerce como parte del segundo evento de compra.

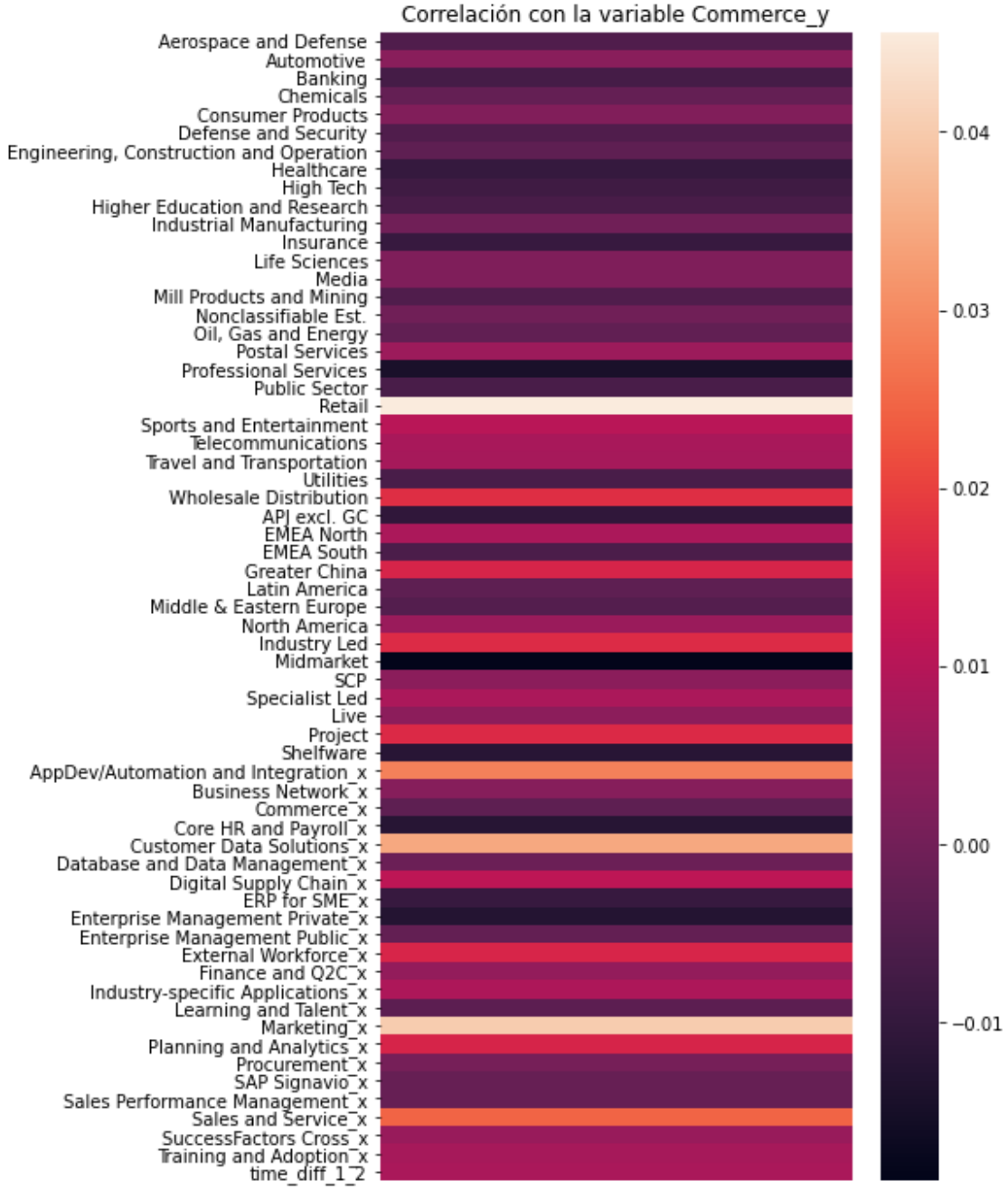


Figura 24: Correlación entre las variables independientes y la solución Core HR and Payroll como parte del segundo evento de compra.

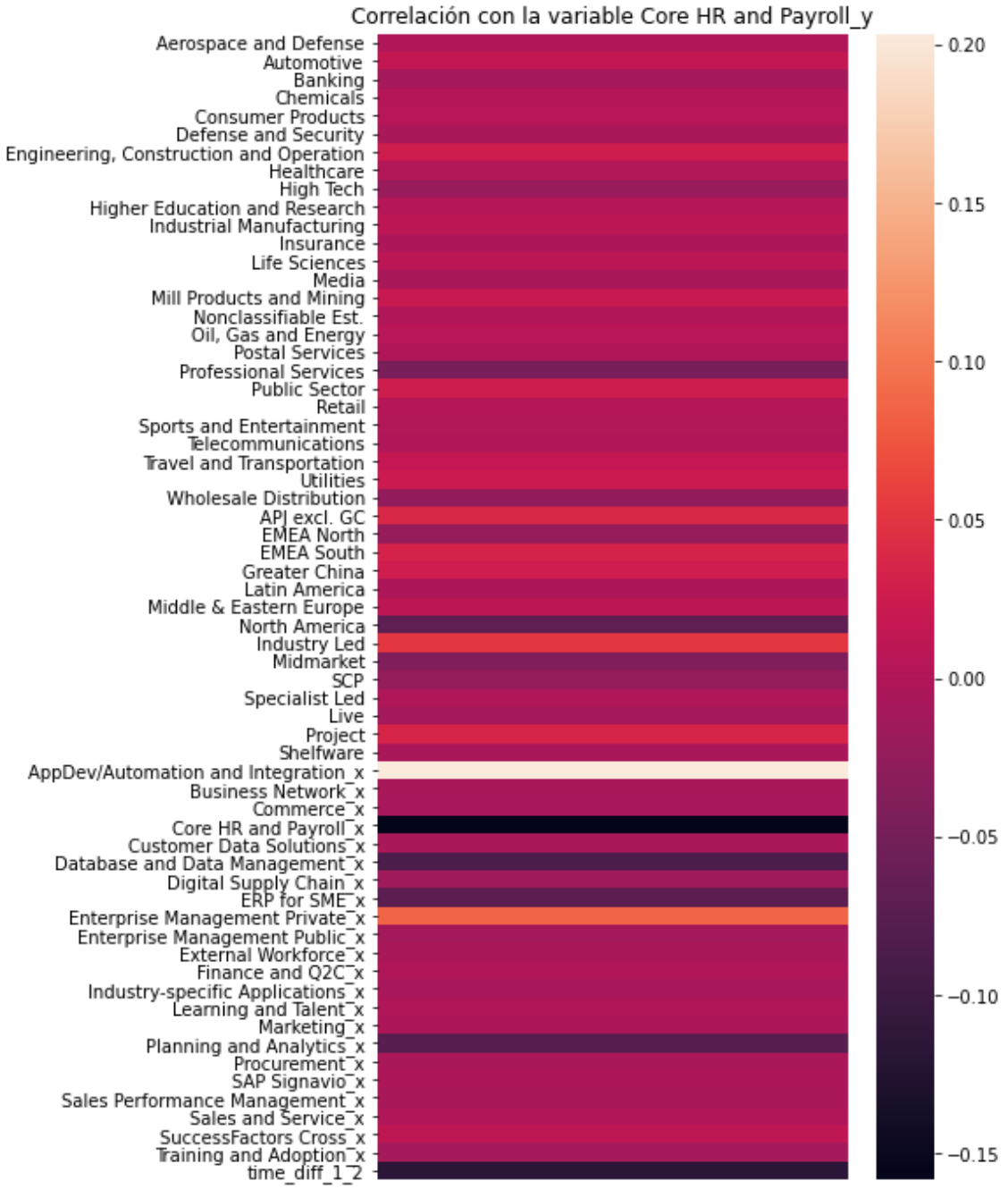


Figura 25: Correlación entre las variables independientes y la solución Customer Data Solutions como parte del segundo evento de compra.

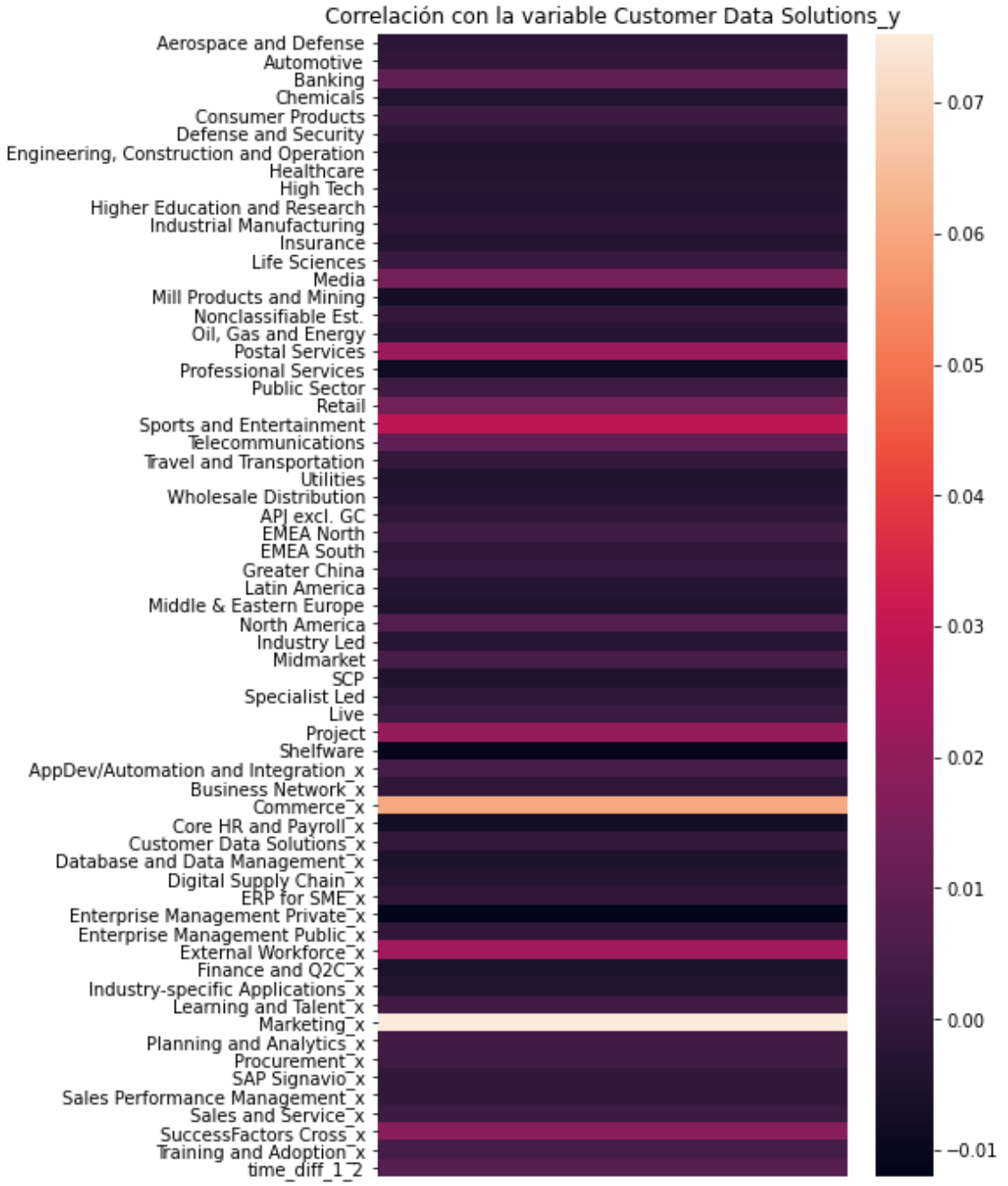


Figura 26: Correlación entre las variables independientes y la solución Database and Data Management como parte del segundo evento de compra.

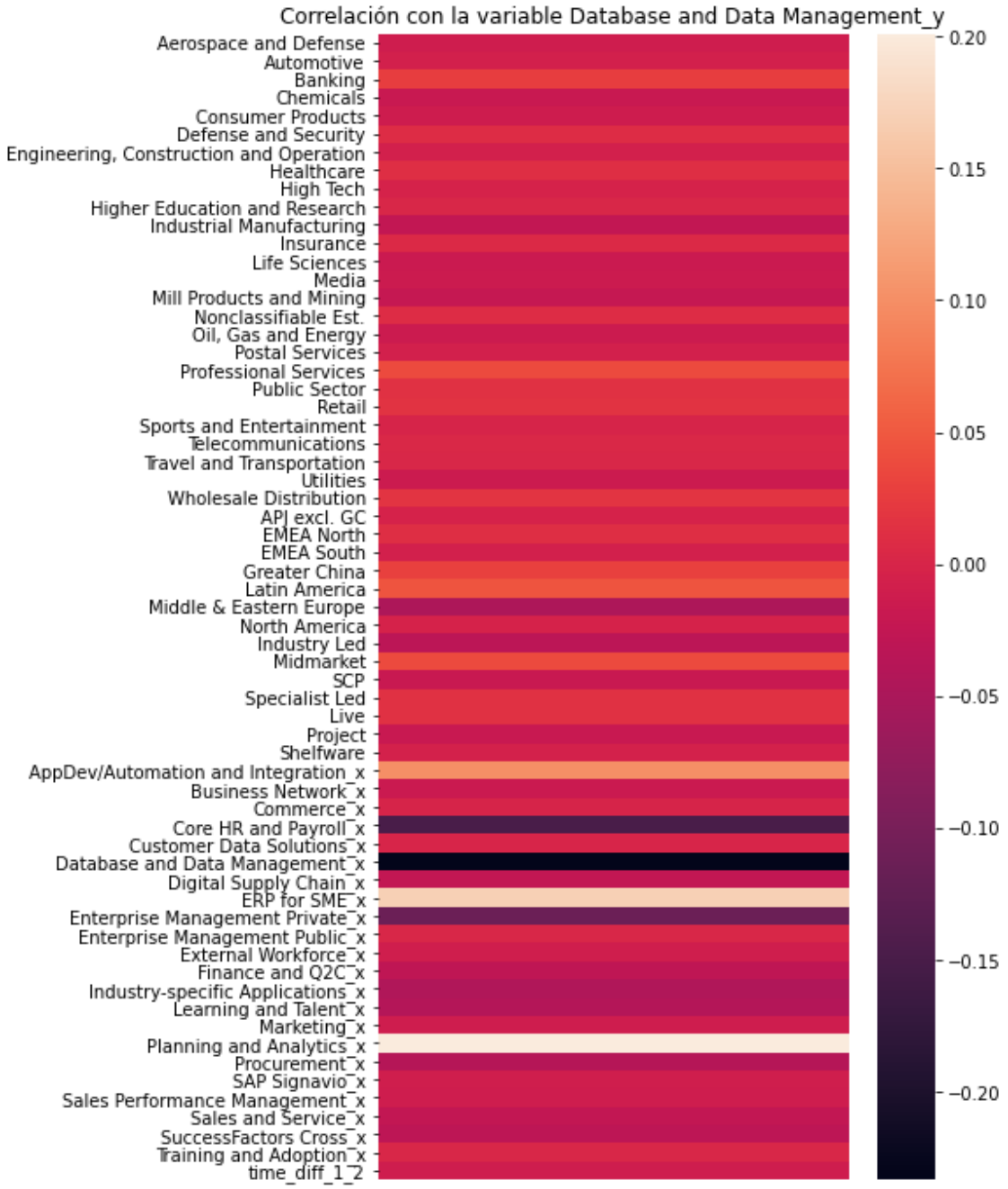


Figura 27: Correlación entre las variables independientes y la solución Digital Supply Chain como parte del segundo evento de compra.

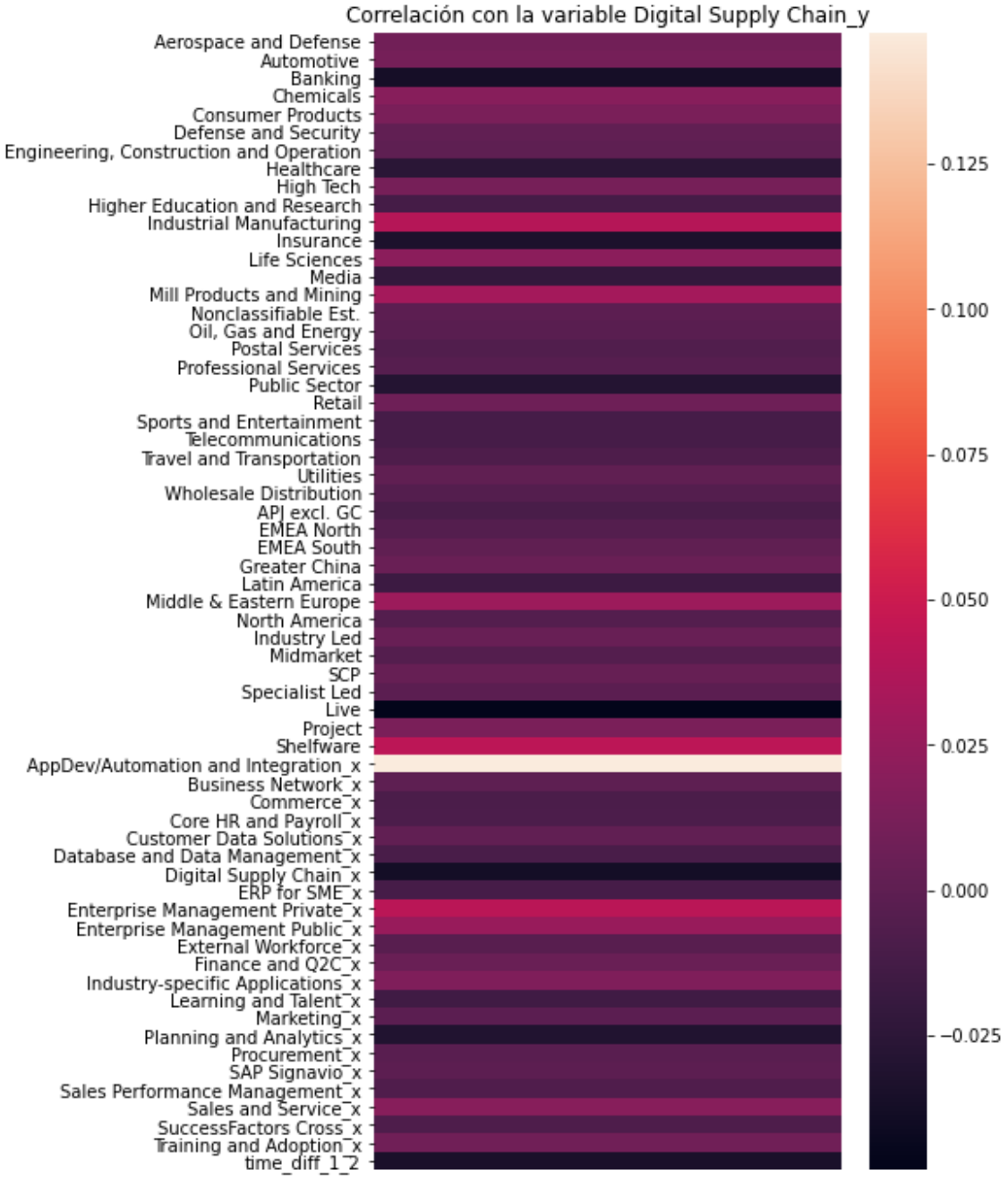


Figura 28: Correlación entre las variables independientes y la solución ERP for SME como parte del segundo evento de compra.

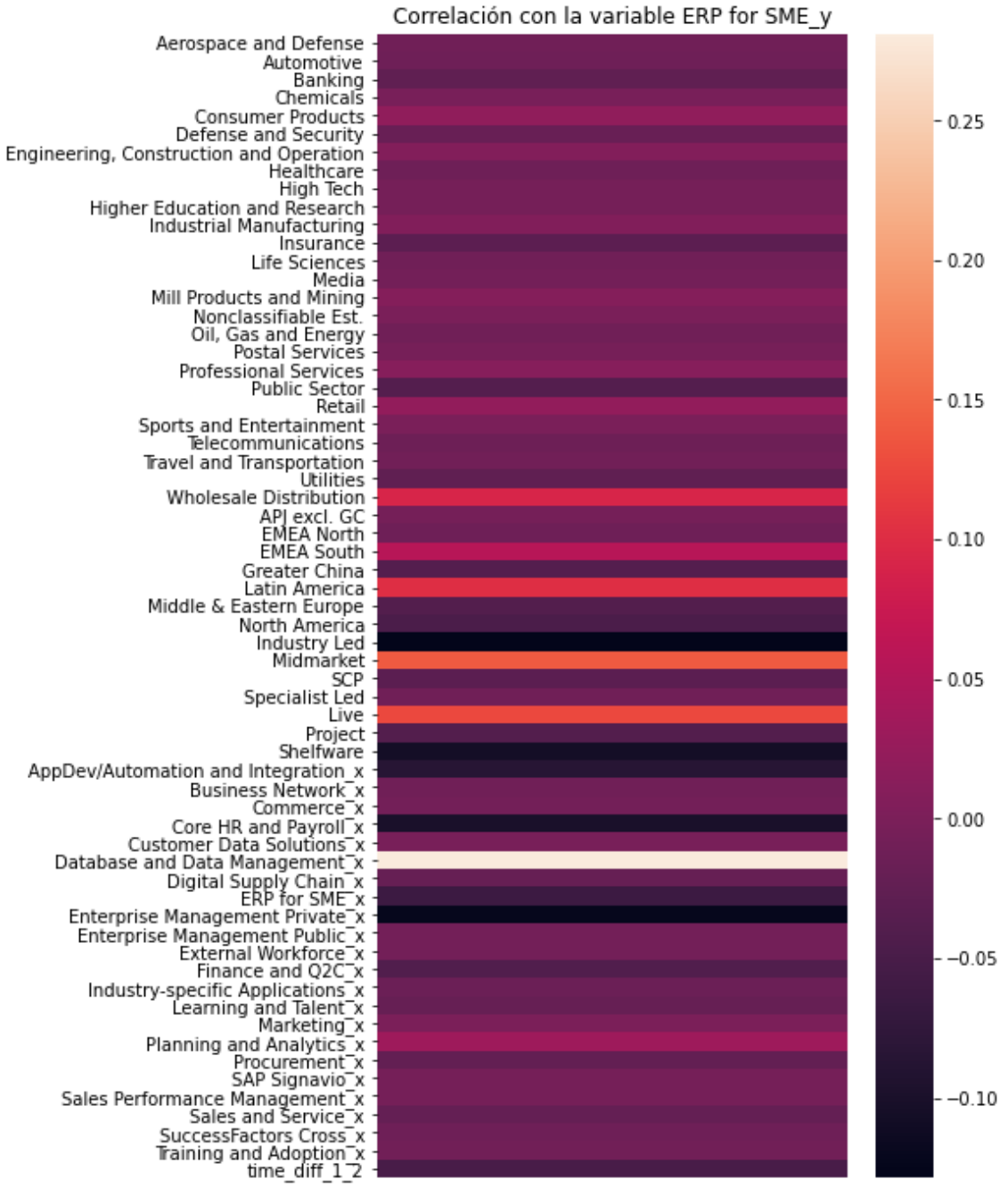


Figura 29: Correlación entre las variables independientes y la solución Enterprise Management Private como parte del segundo evento de compra.

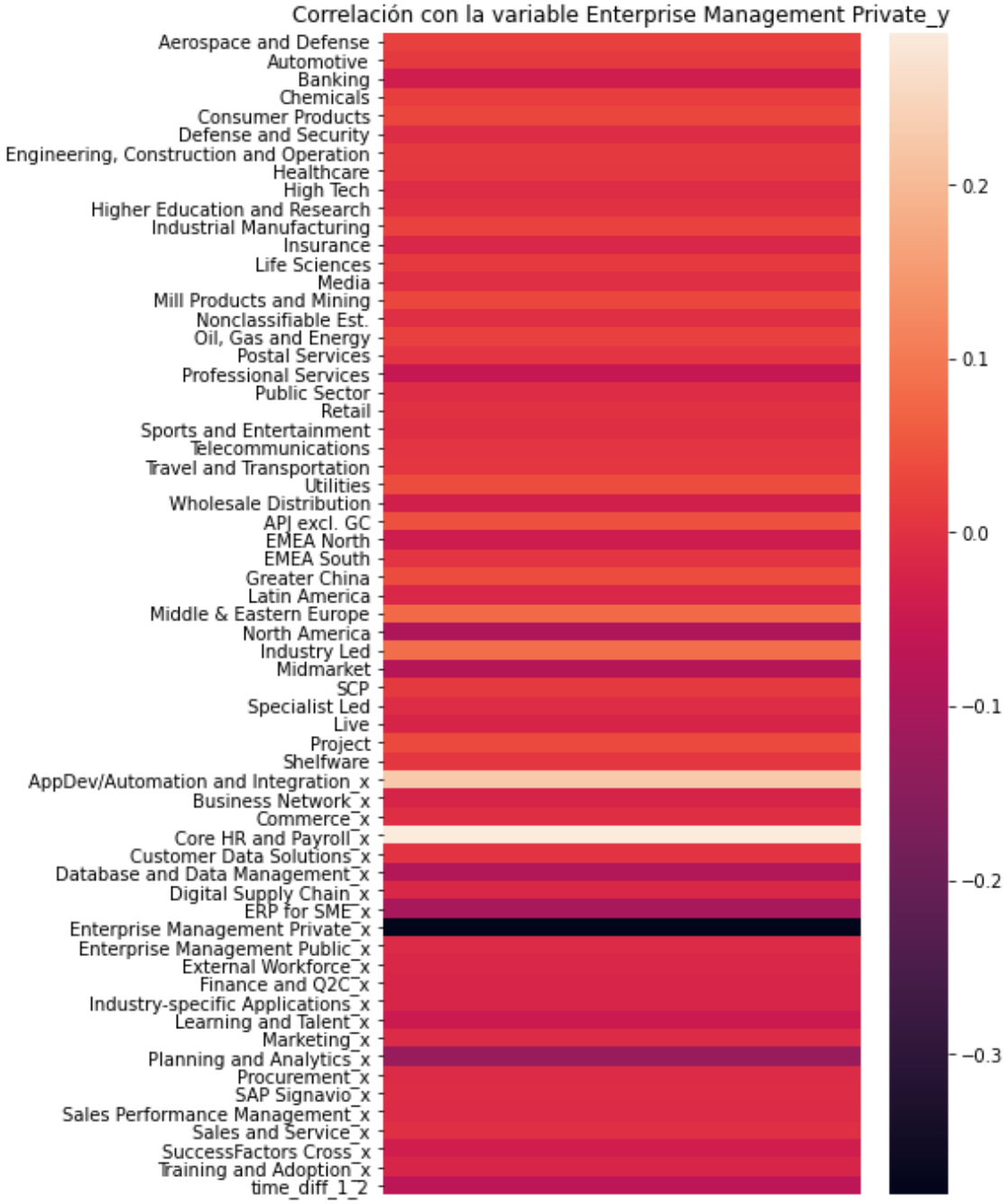


Figura 30: Correlación entre las variables independientes y la solución Enterprise Management Public como parte del segundo evento de compra.

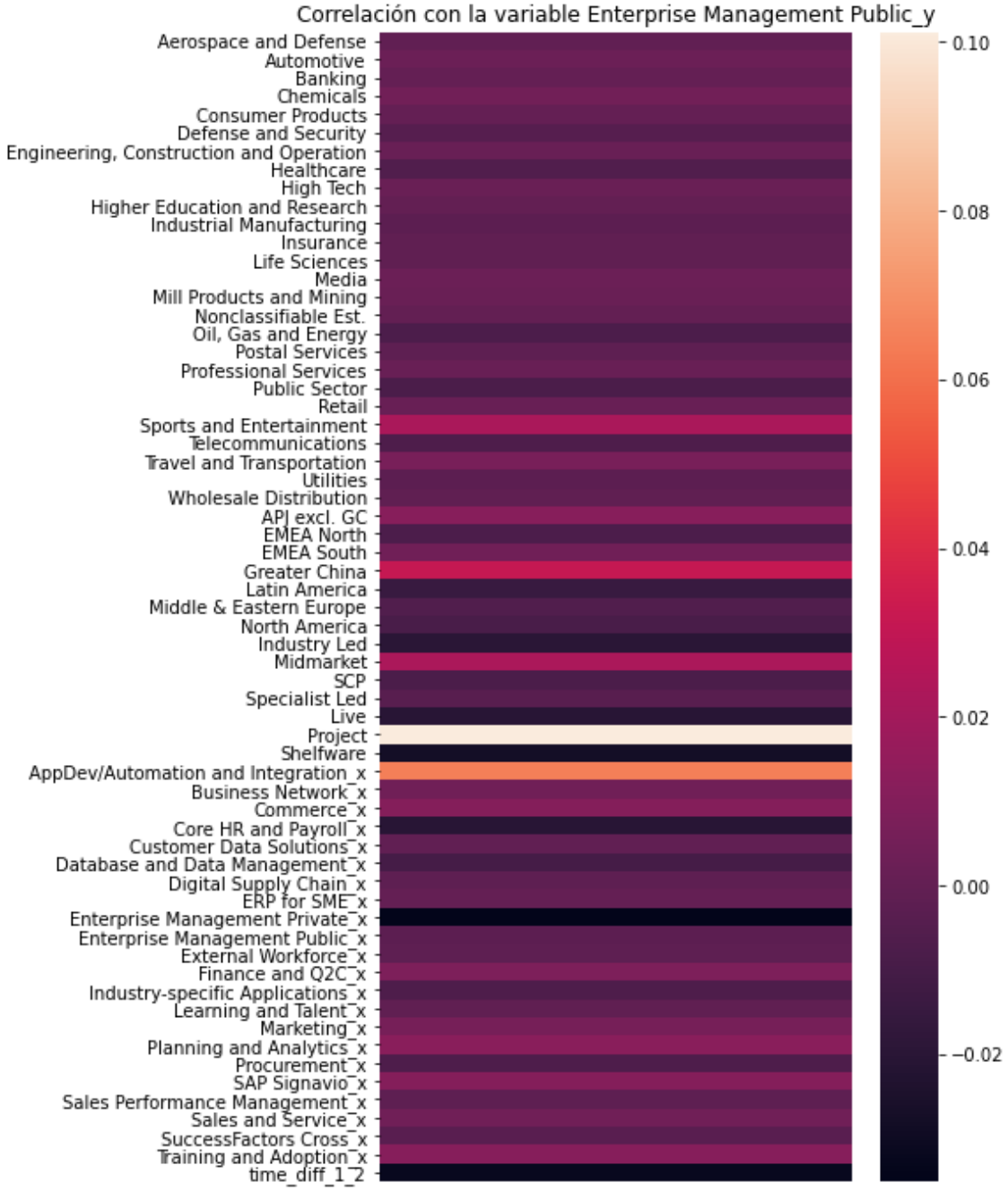


Figura 31: Correlación entre las variables independientes y la solución External Workforce como parte del segundo evento de compra.

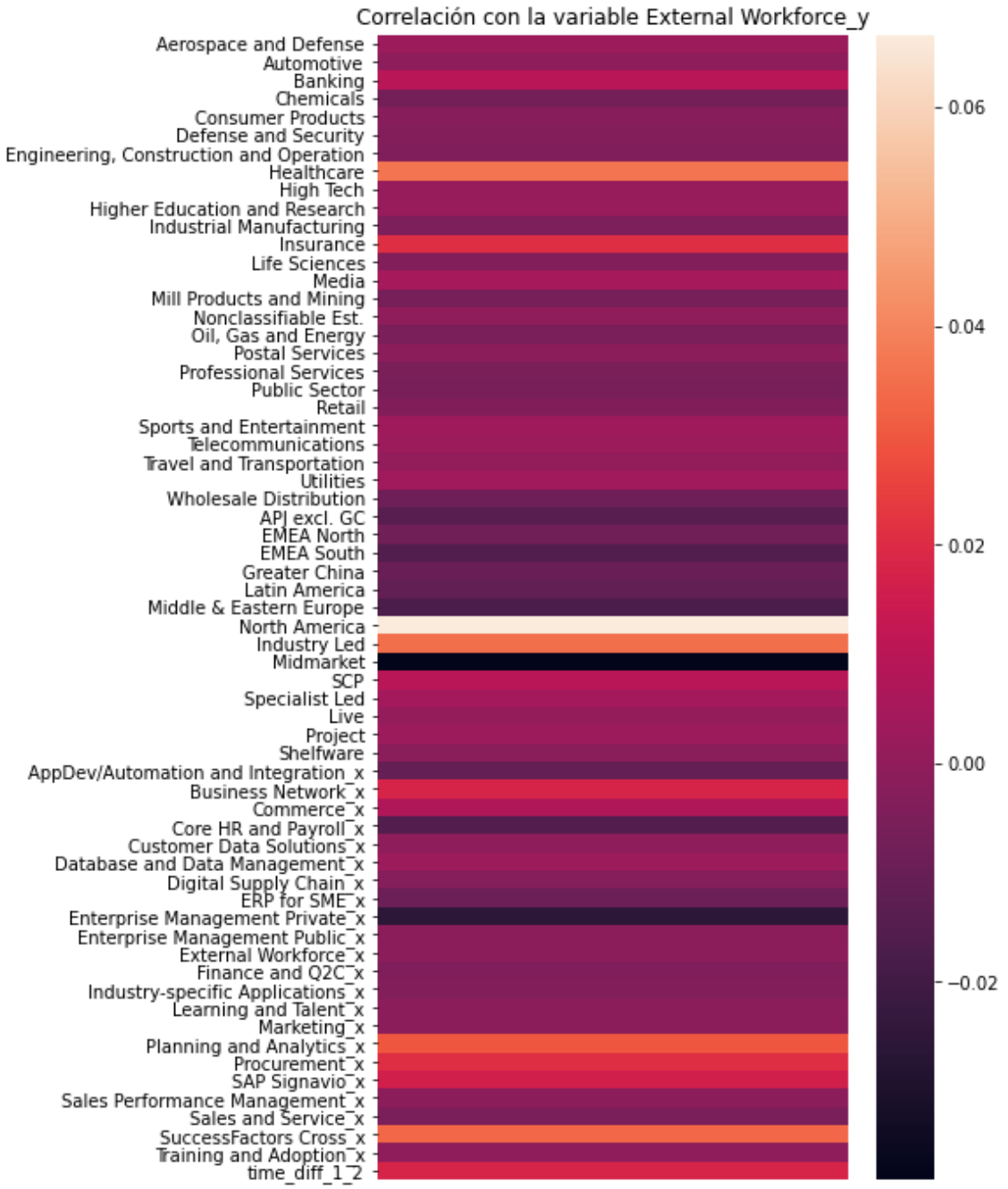


Figura 32: Correlación entre las variables independientes y la solución Finance and Q2C como parte del segundo evento de compra.

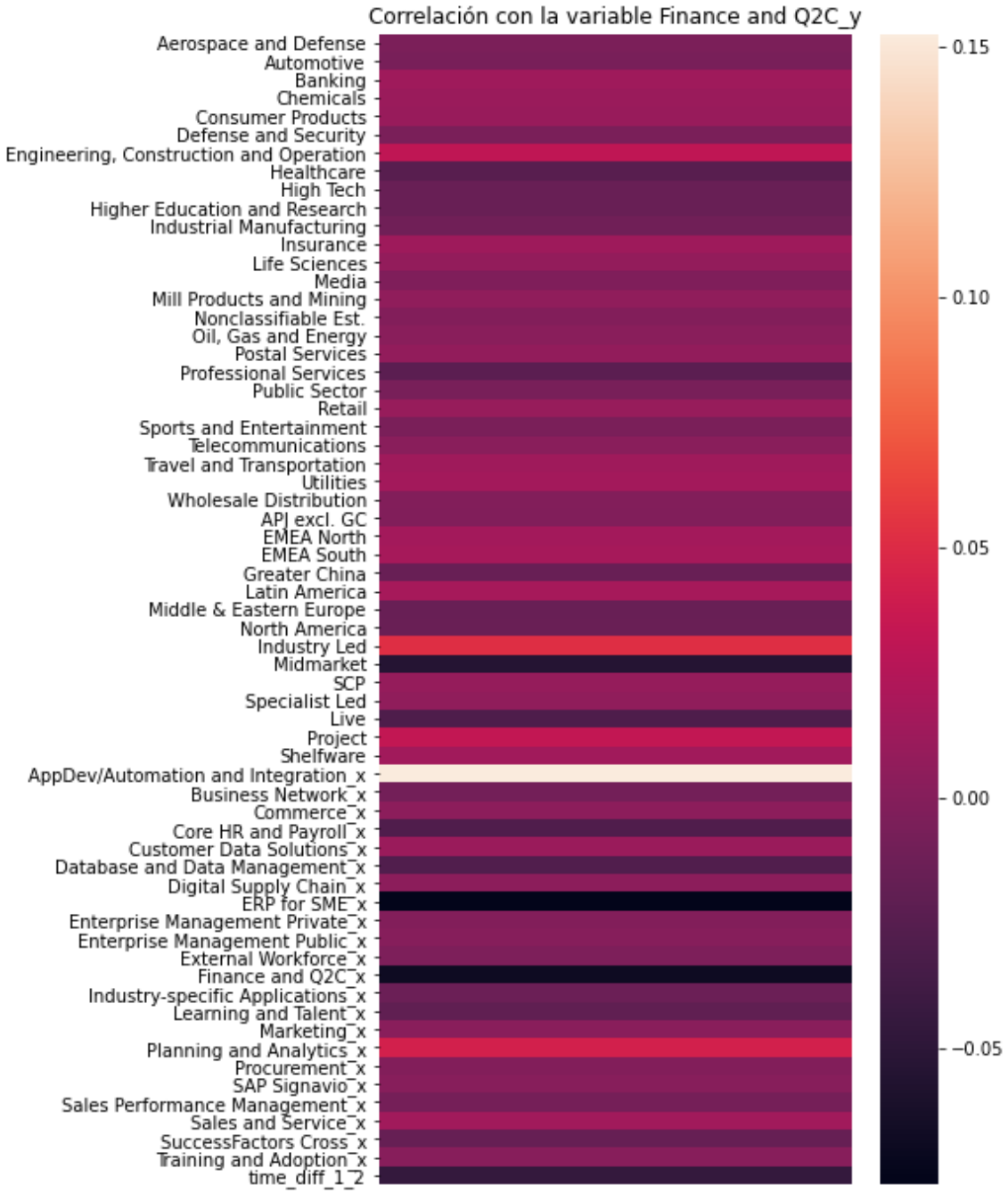


Figura 33: Correlación entre las variables independientes y la solución Industry-specific Applications como parte del segundo evento de compra.

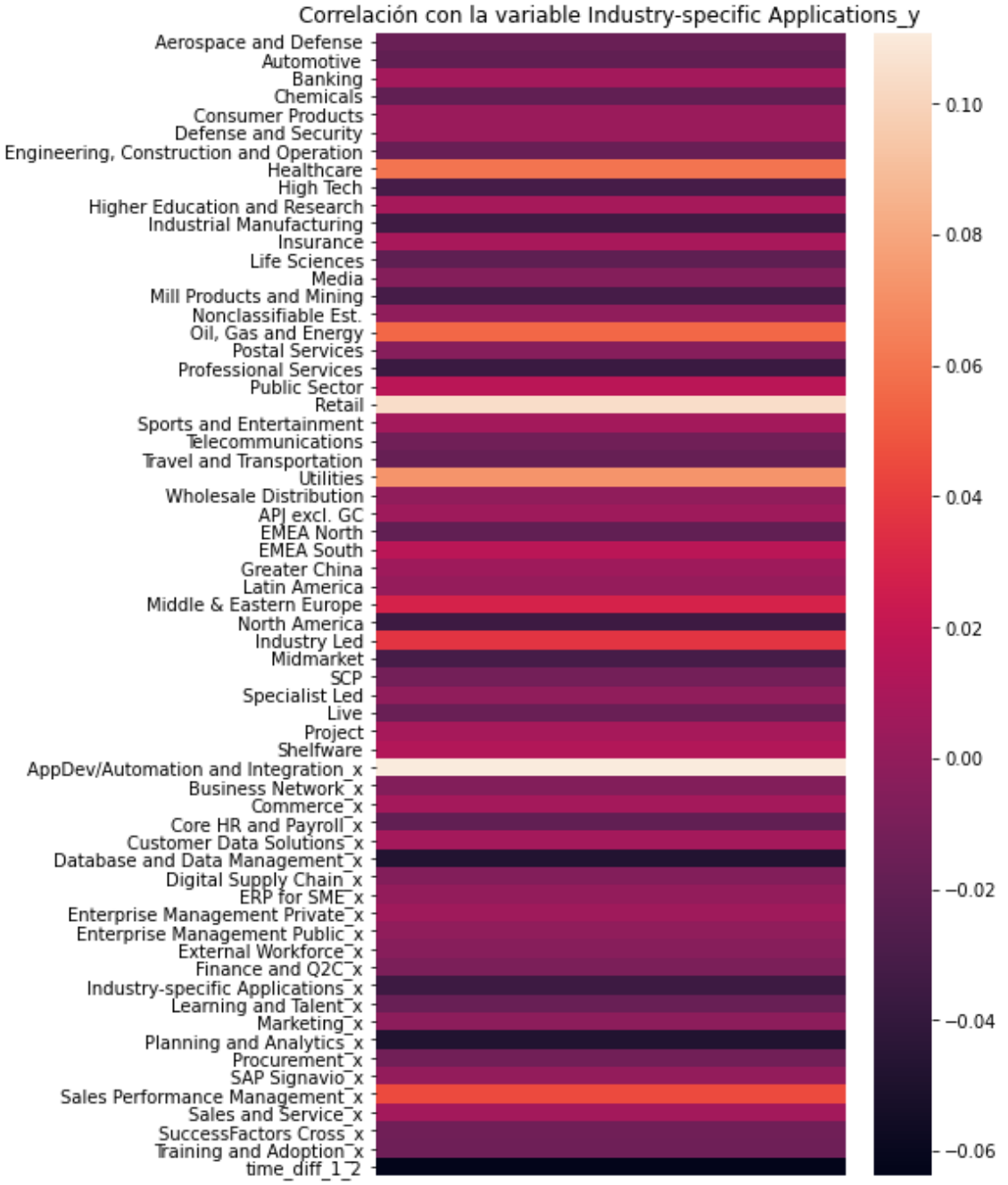


Figura 34: Correlación entre las variables independientes y la solución Learning and Talent como parte del segundo evento de compra.

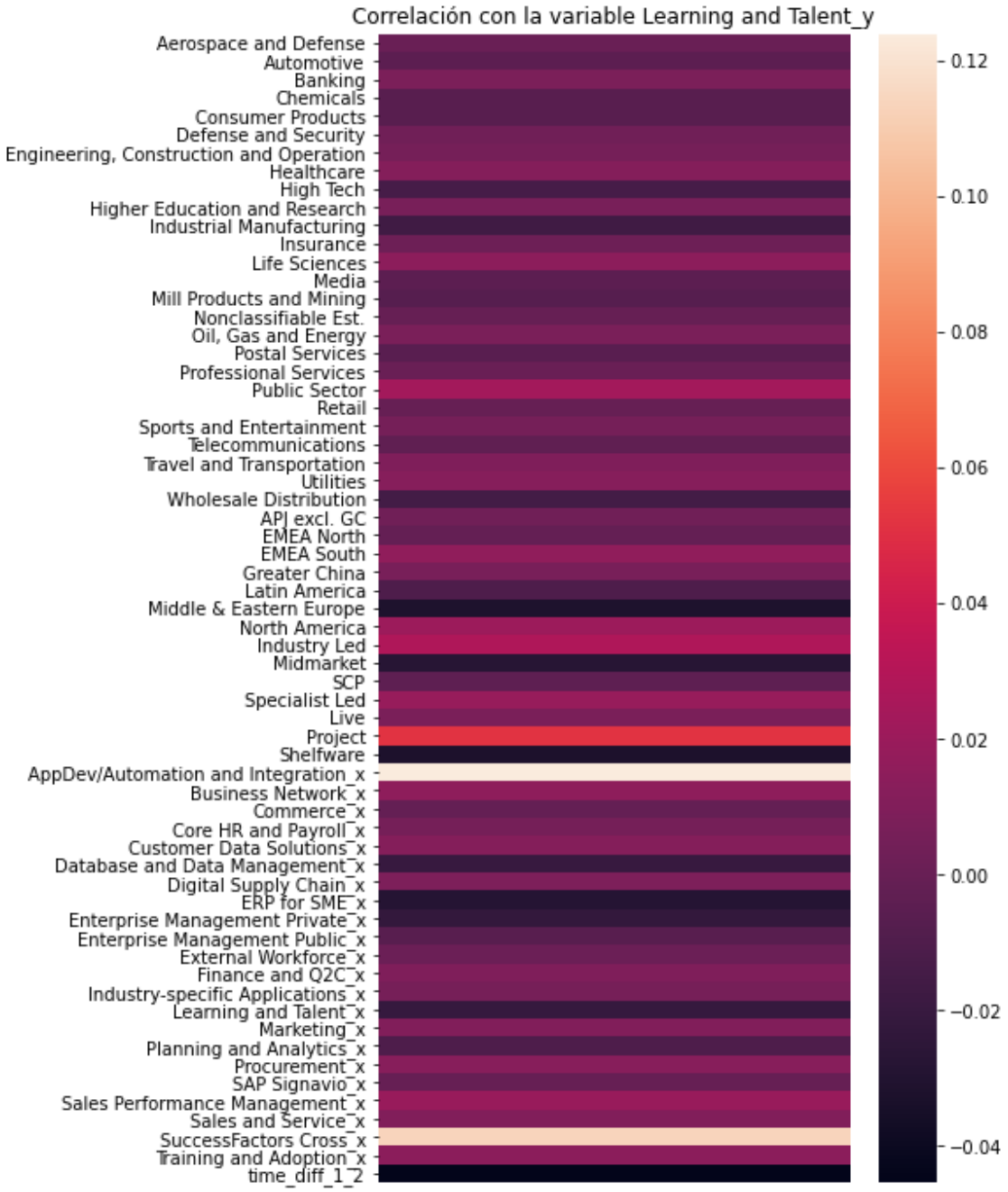


Figura 35: Correlación entre las variables independientes y la solución Marketing como parte del segundo evento de compra.

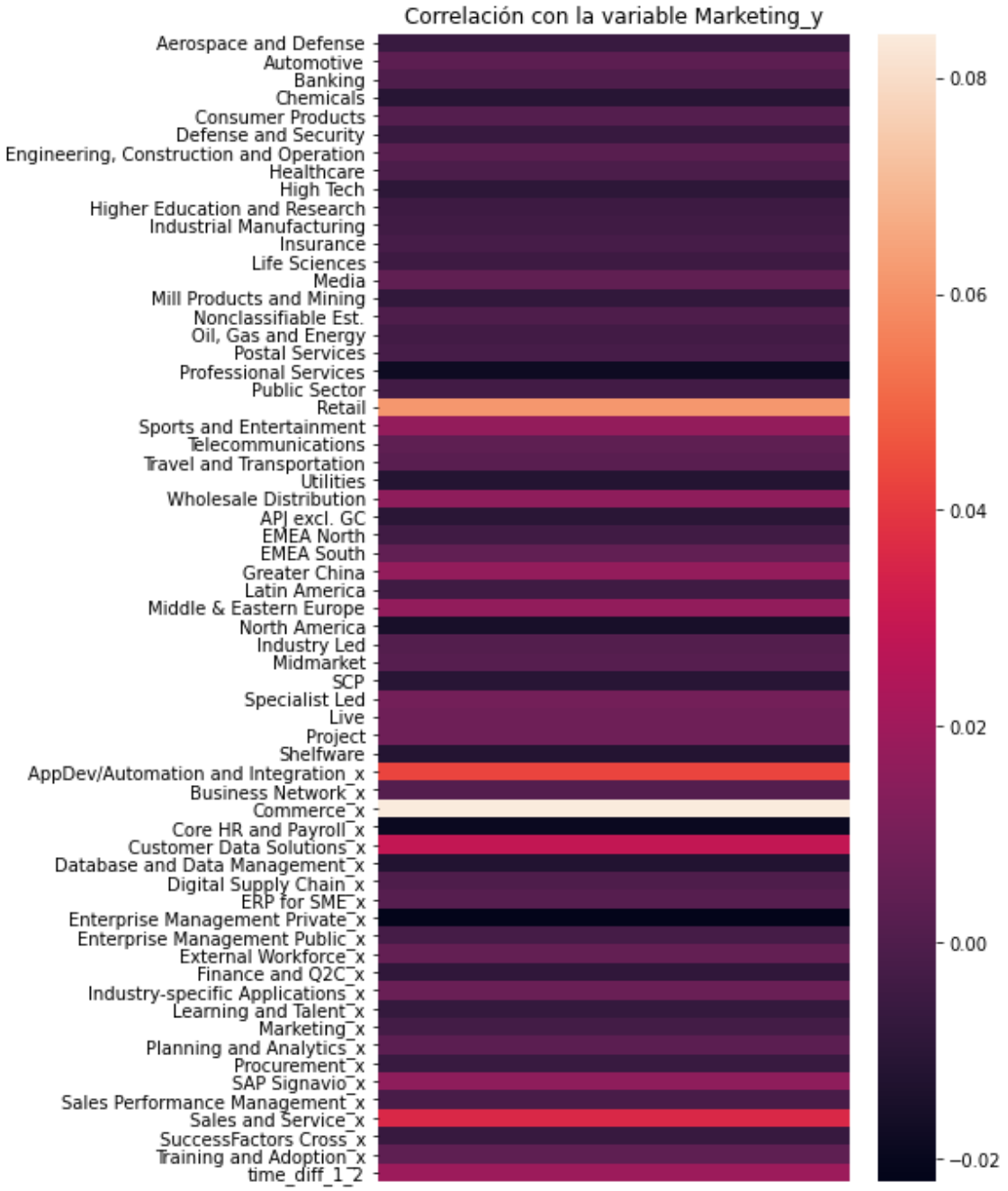


Figura 36: Correlación entre las variables independientes y la solución Planning and Analytics como parte del segundo evento de compra.

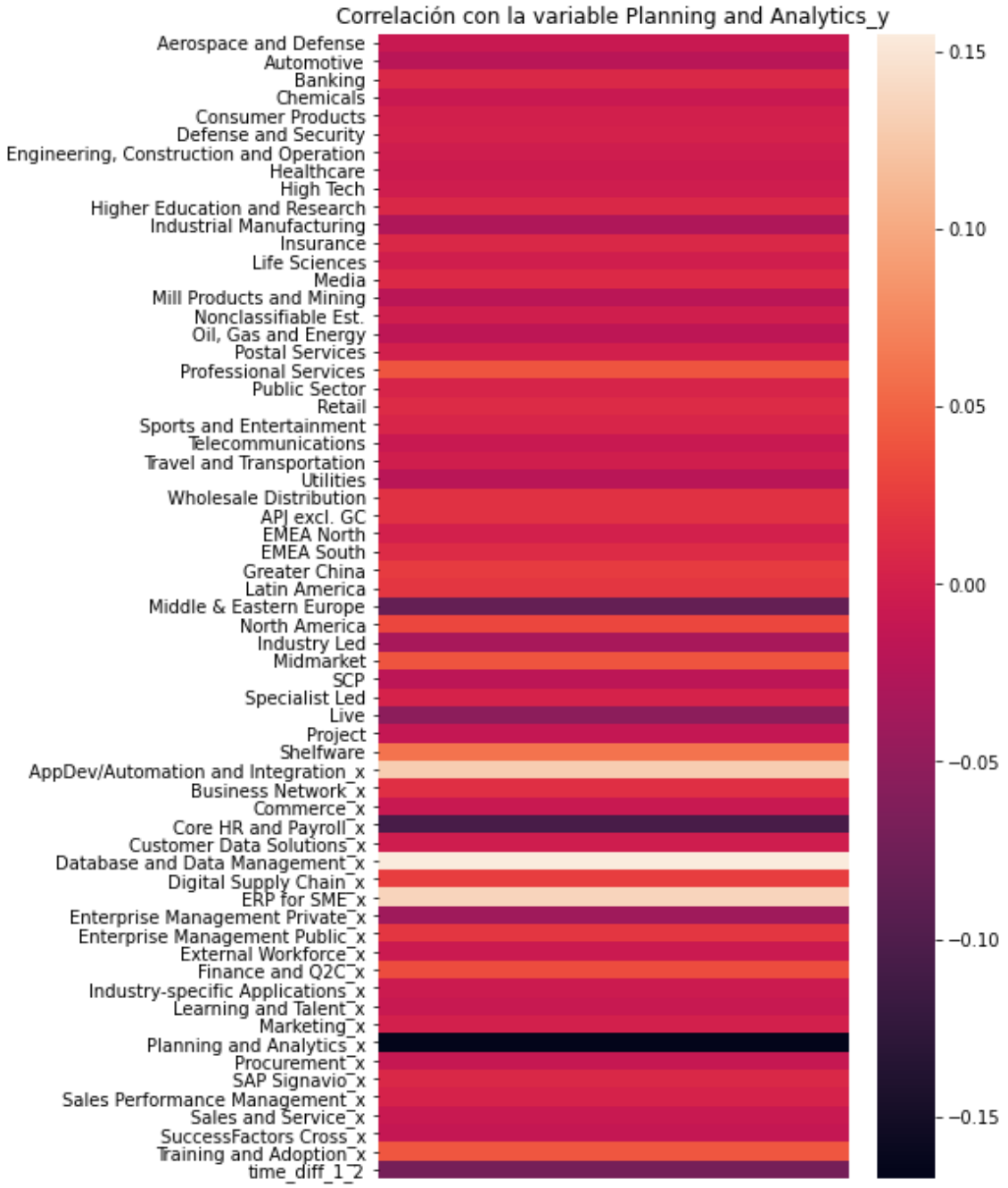


Figura 37: Correlación entre las variables independientes y la solución Procurement como parte del segundo evento de compra.

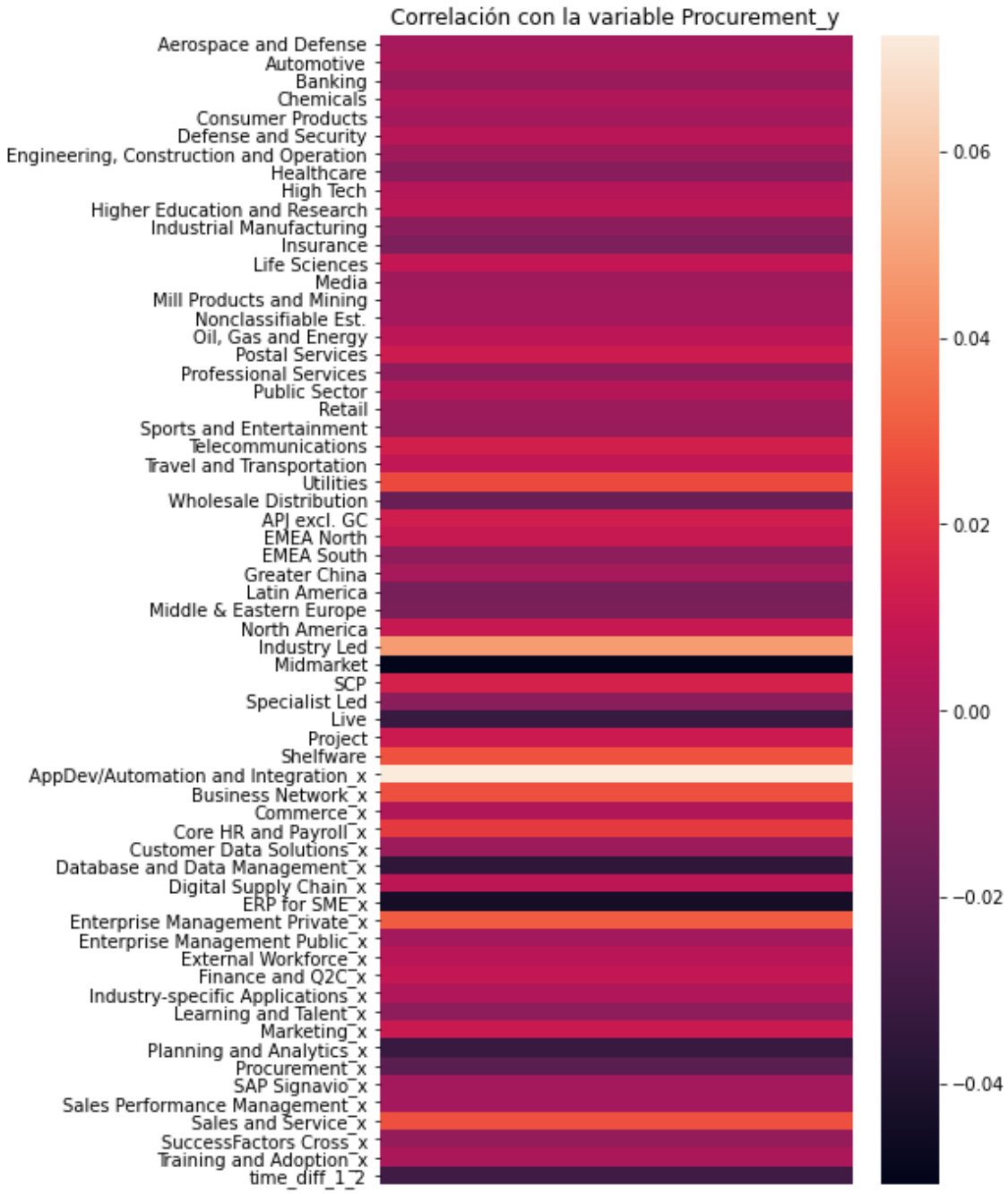


Figura 38: Correlación entre las variables independientes y la solución SAP Signavio como parte del segundo evento de compra.

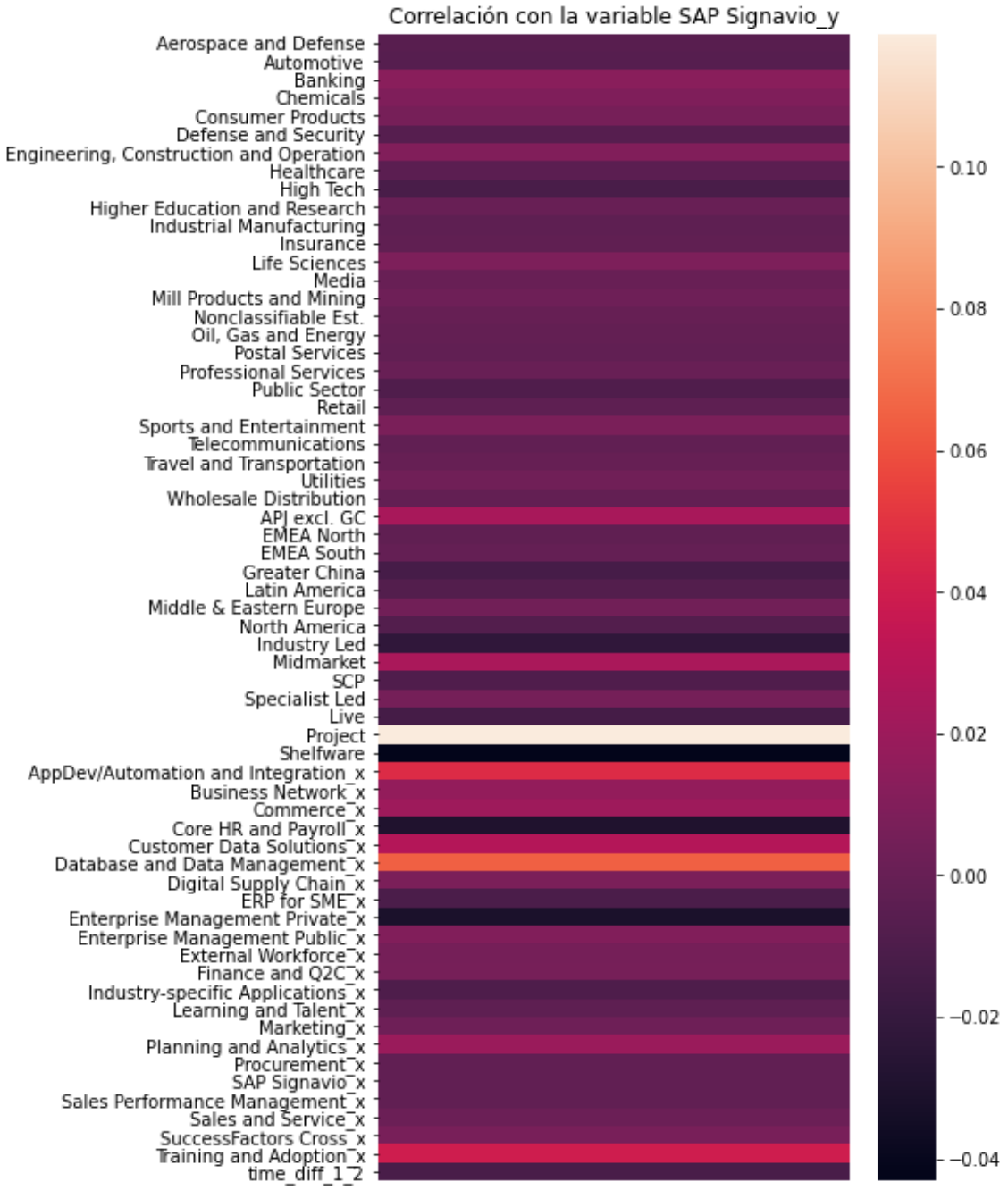


Figura 39: Correlación entre las variables independientes y la solución Sales and Performance Management como parte del segundo evento de compra.

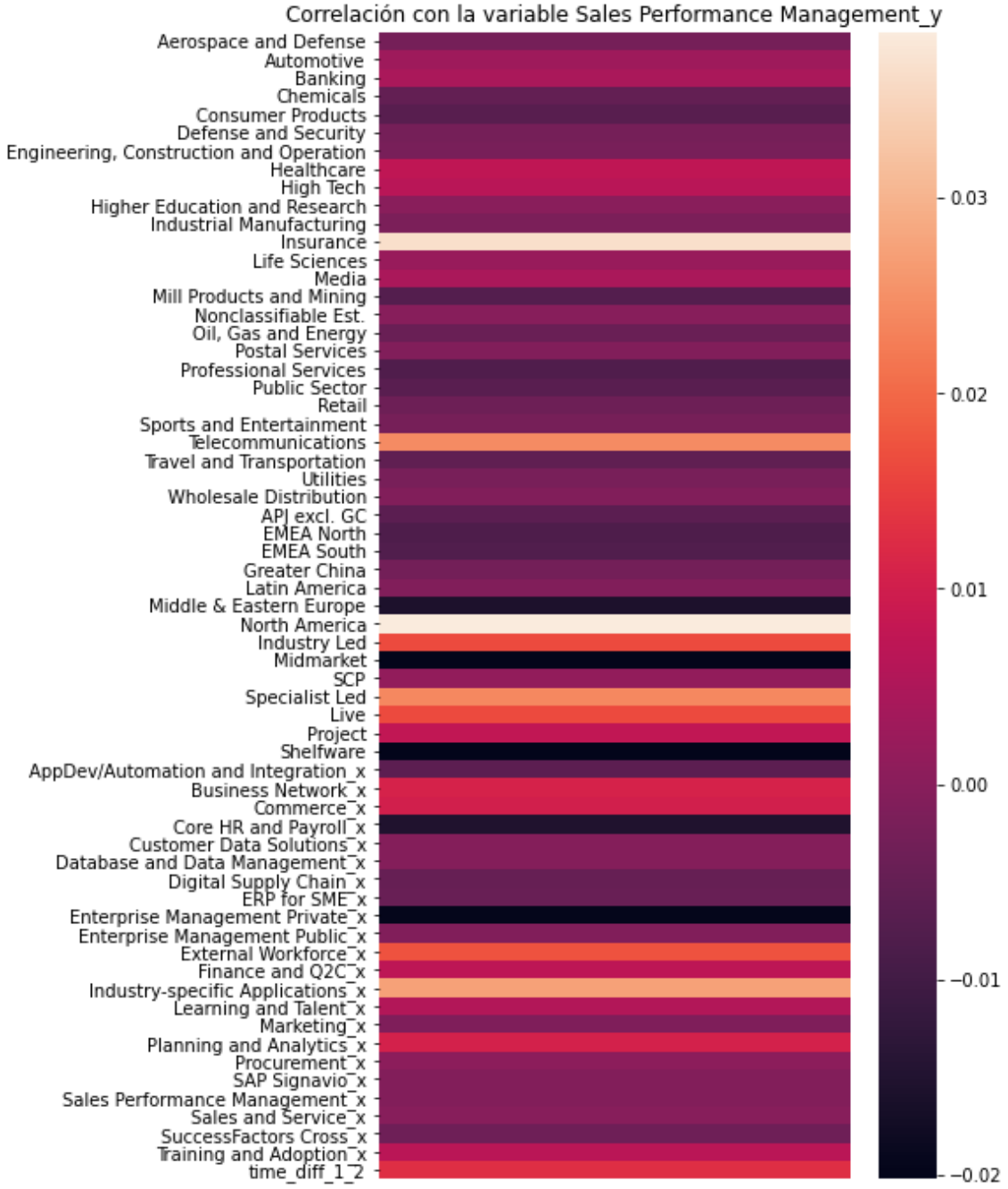


Figura 40: Correlación entre las variables independientes y la solución Sales and Service como parte del segundo evento de compra.

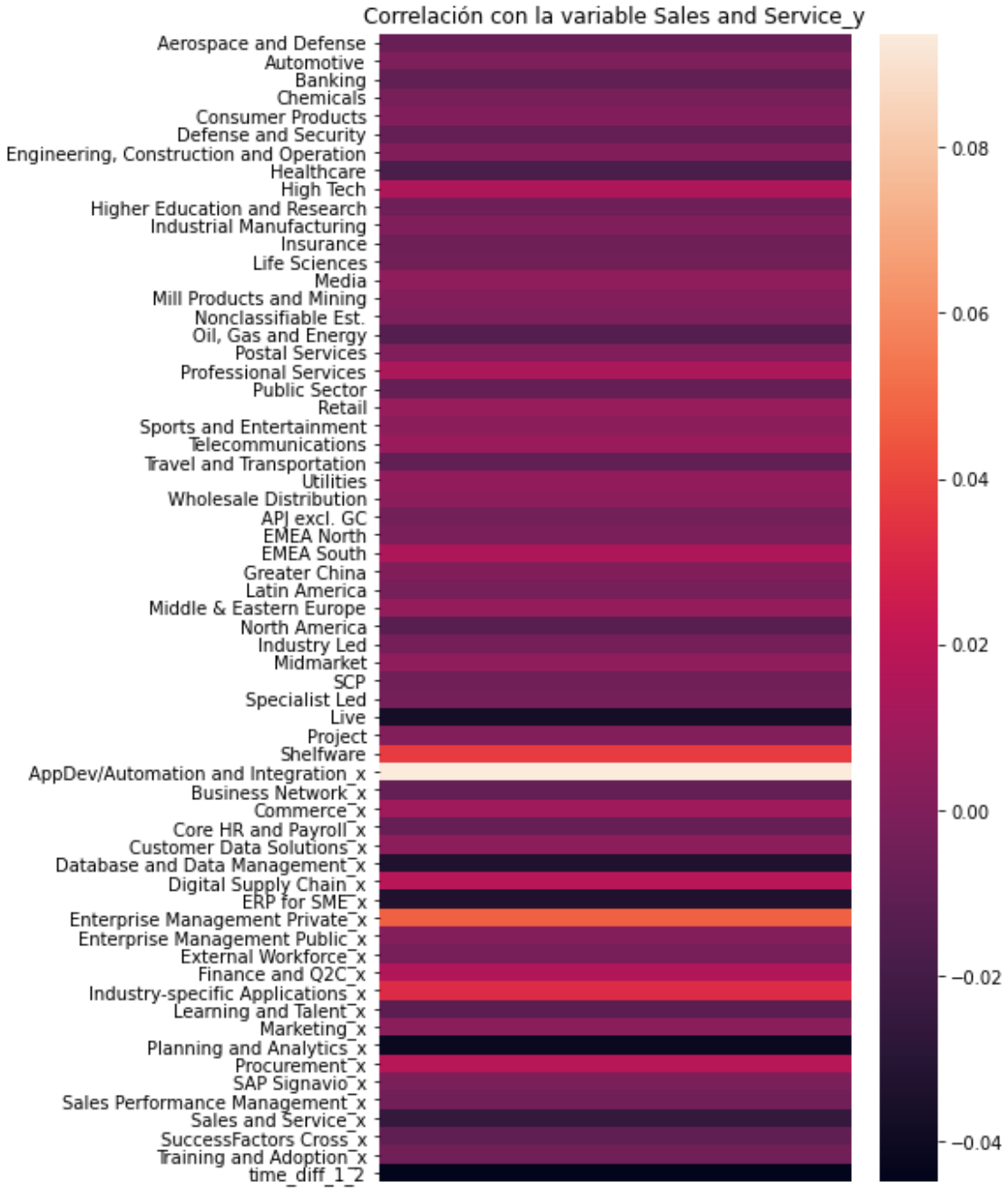


Figura 41: Correlación entre las variables independientes y la solución SuccessFactors Cross como parte del segundo evento de compra.

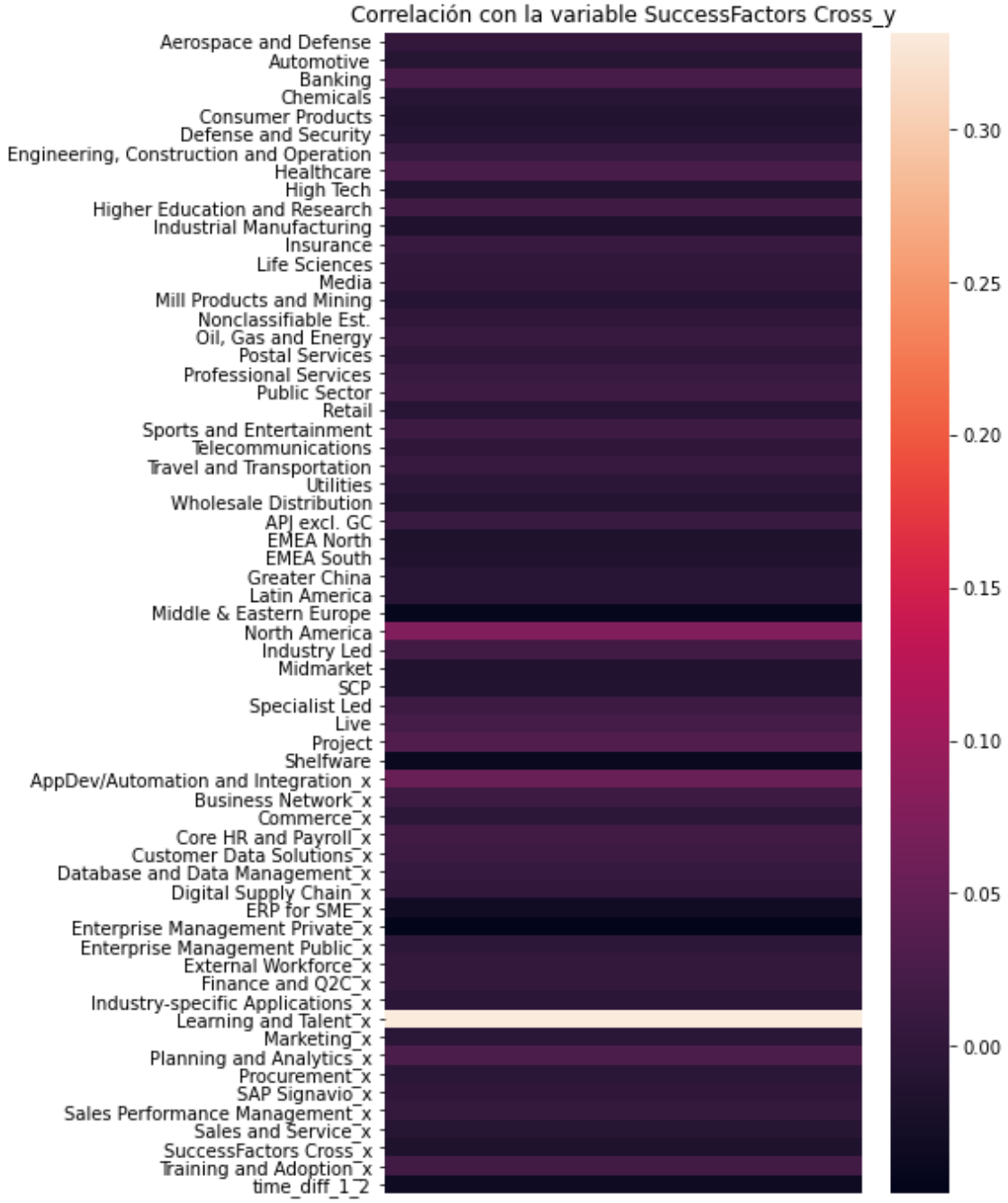


Figura 42: Correlación entre las variables independientes y la solución Training and Adoption como parte del segundo evento de compra.

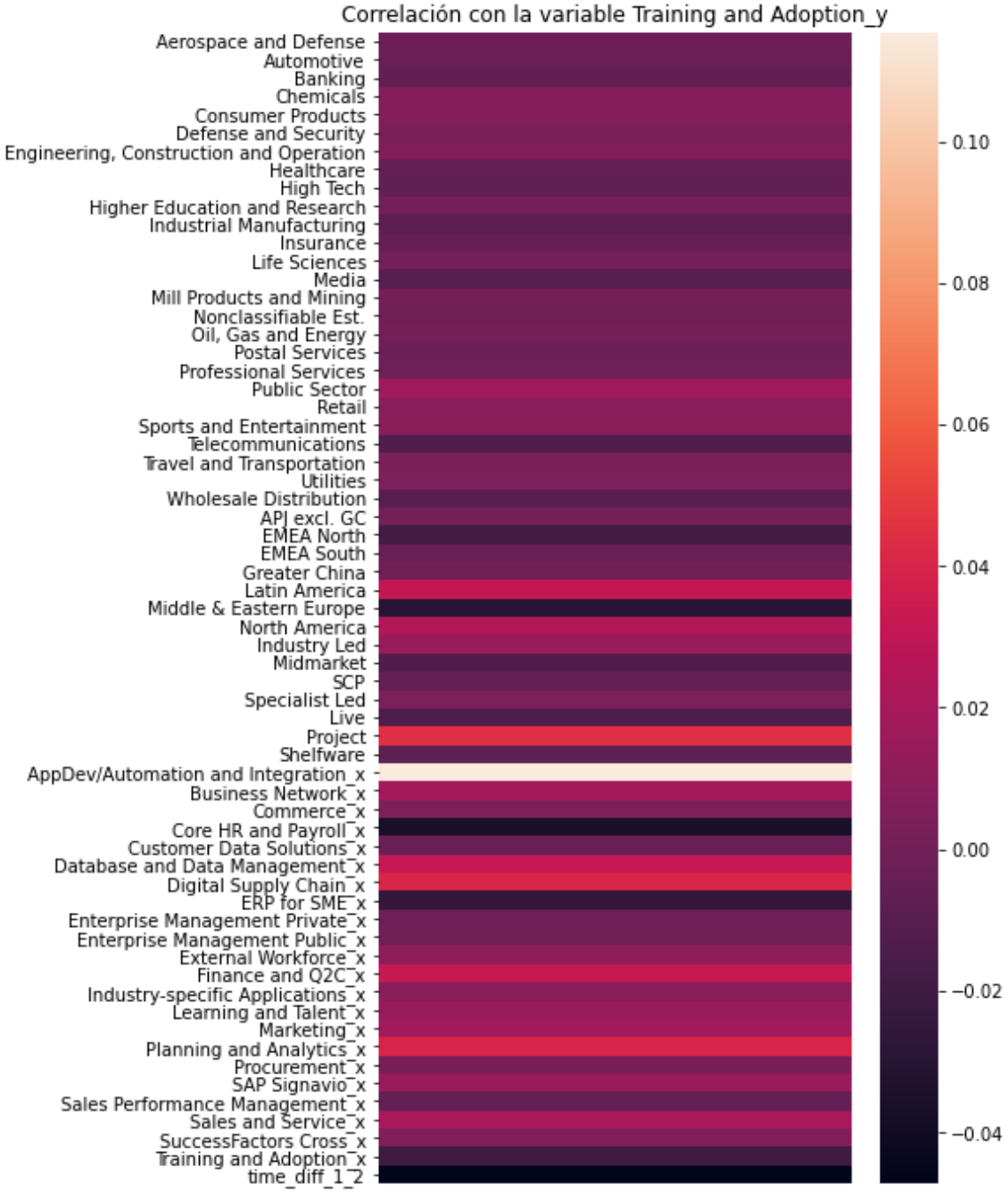
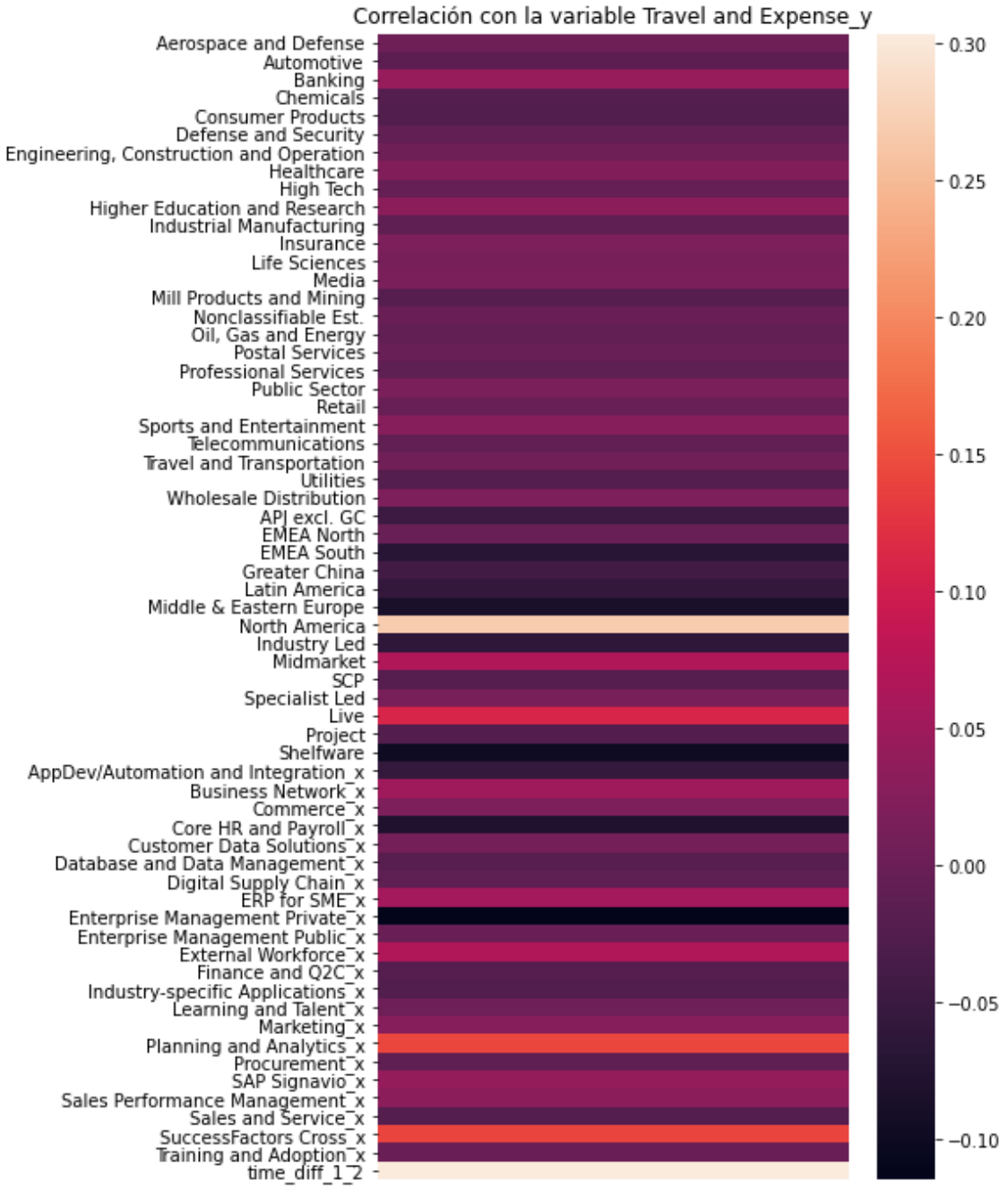


Figura 43: Correlación entre las variables independientes y la solución Travel and Expense como parte del segundo evento de compra.



Gráficos de las curvas ROC de los modelos

Figura 44: Curva ROC correspondiente al modelo de AppDev/Automation and Integration con un AUC de 0,84.

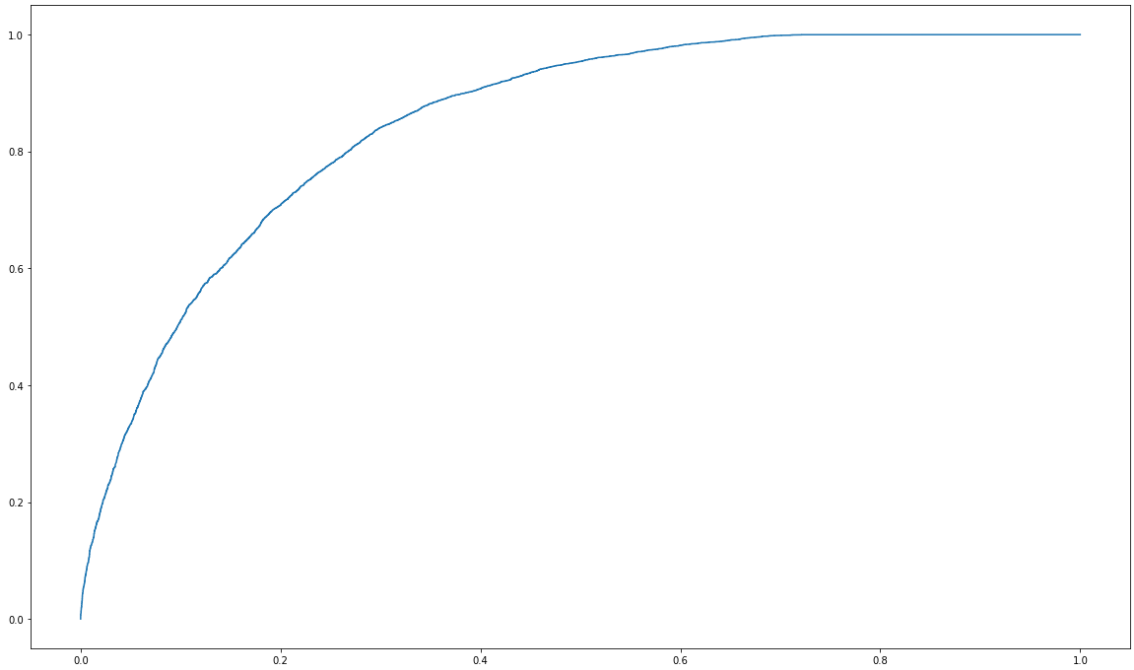


Figura 45: Curva ROC correspondiente al modelo de Business Network con un AUC de 0,9.

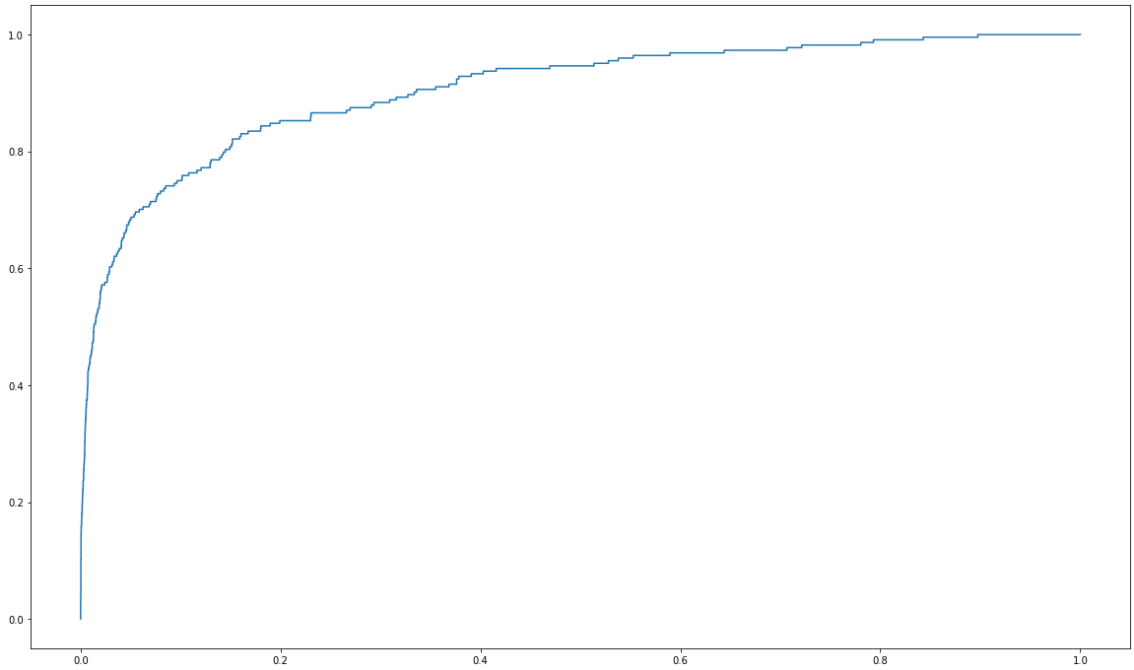


Figura 46: Curva ROC correspondiente al modelo de Commerce con un AUC de 0,7.

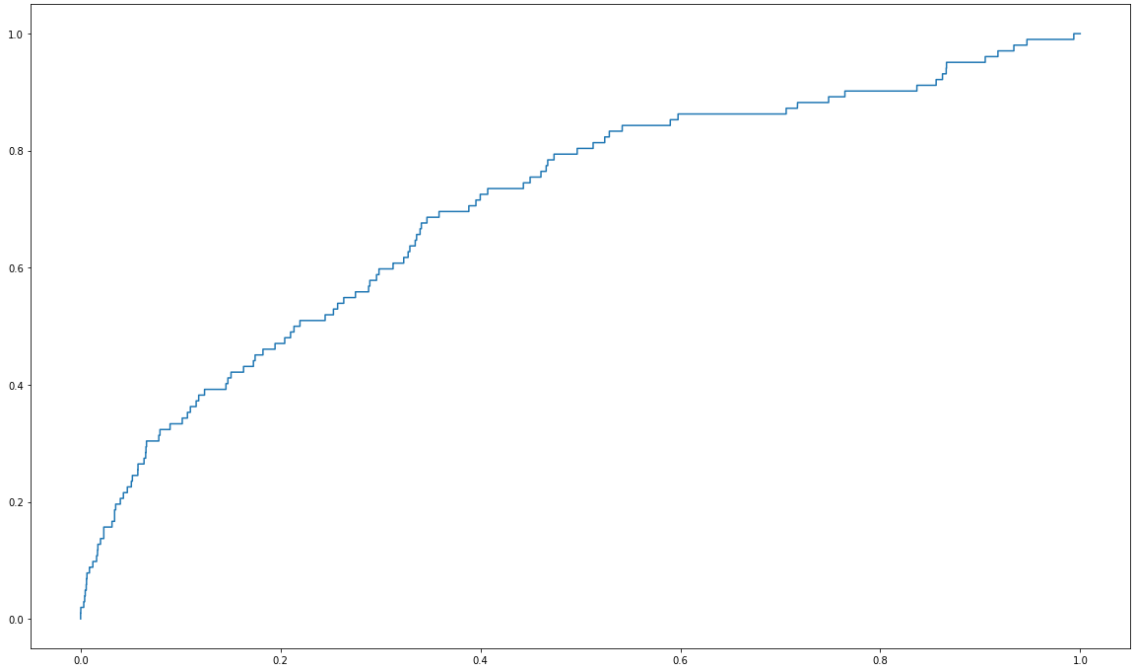


Figura 47: Curva ROC correspondiente al modelo de Core HR and Payroll con un AUC de 0,84.

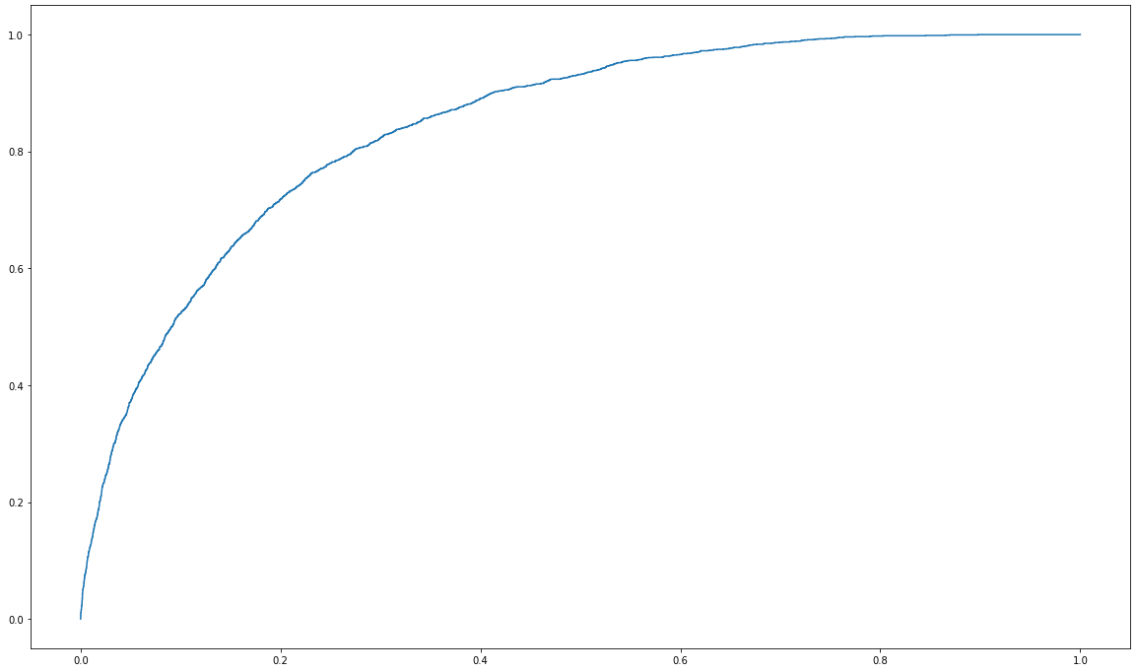


Figura 48: Curva ROC correspondiente al modelo de Customer Data Solutions con un AUC de 0,76.

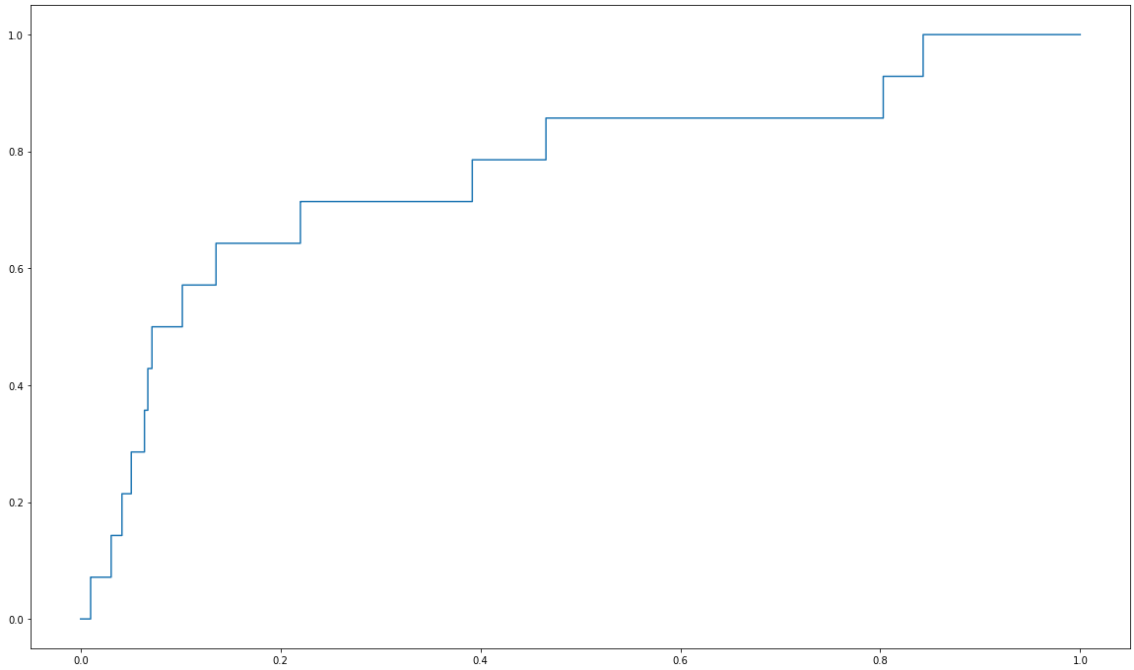


Figura 49: Curva ROC correspondiente al modelo de Database and Data Management con un AUC de 0,83.

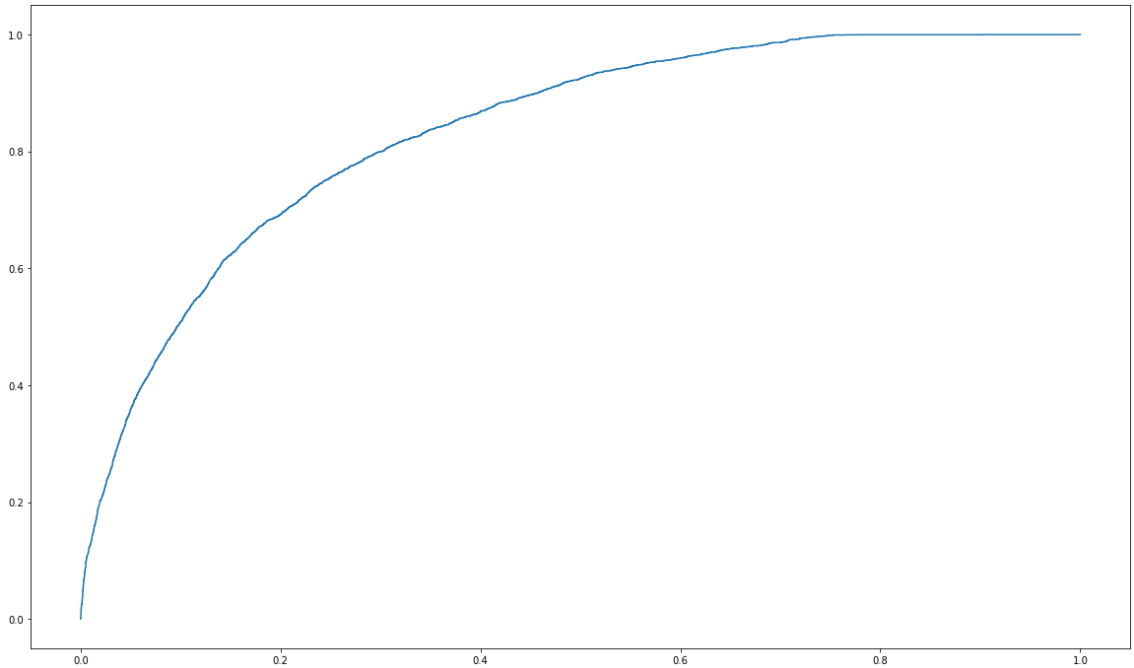


Figura 50: Curva ROC correspondiente al modelo de Digital Supply Chain con un AUC de 0,72.

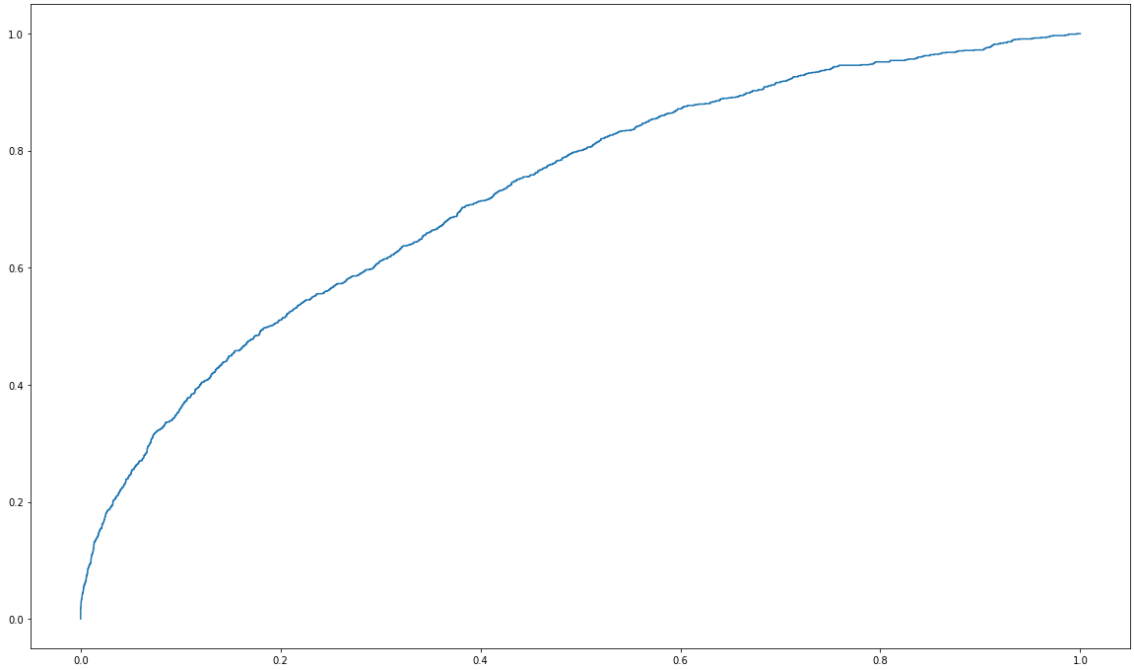


Figura 51: Curva ROC correspondiente al modelo de ERP for SME con un AUC de 0,93.

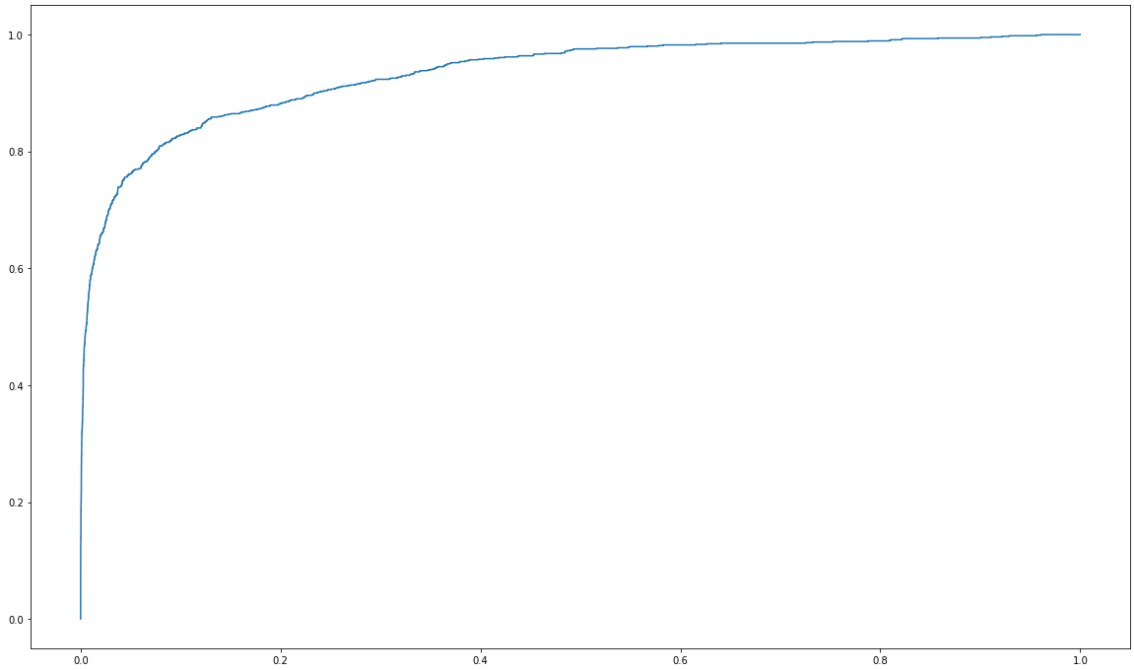


Figura 52: Curva ROC correspondiente al modelo de Enterprise Management Private con un AUC de 0,93.

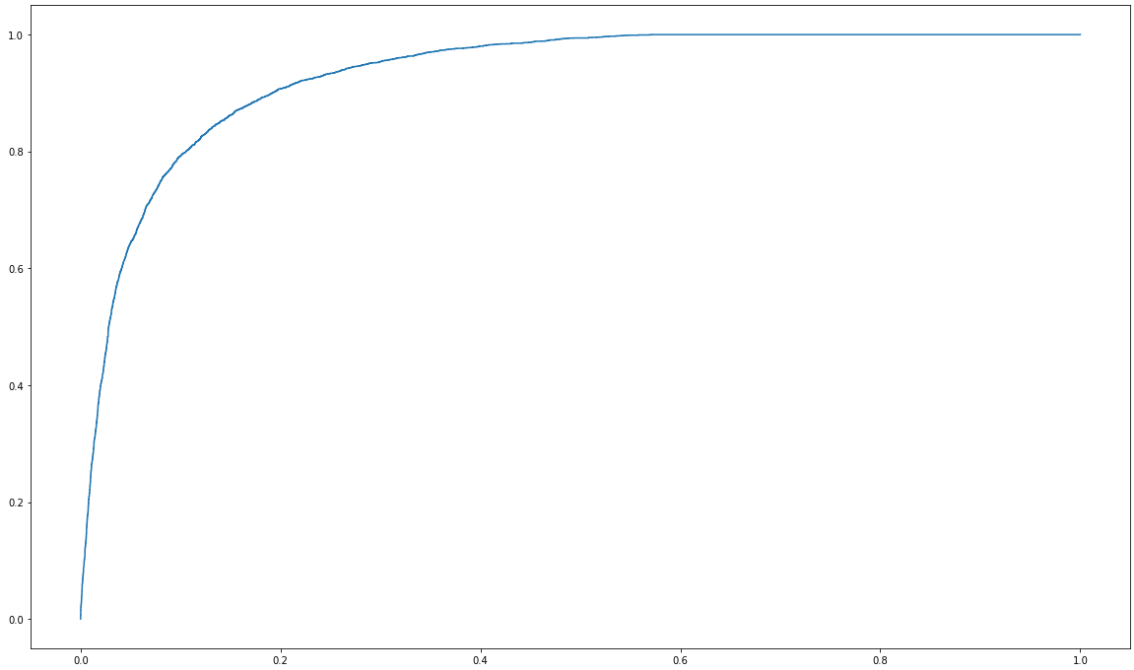


Figura 53: Curva ROC correspondiente al modelo de Enterprise Management Public con un AUC de 0,88.

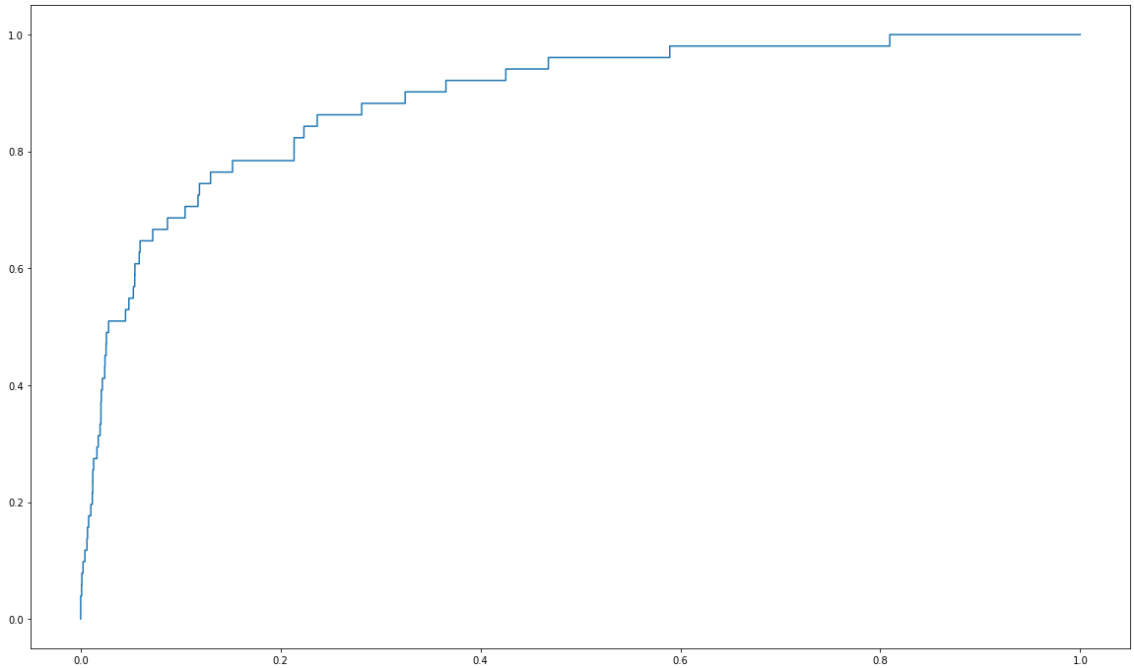


Figura 54: Curva ROC correspondiente al modelo de External Workforce con un AUC de 0,88.

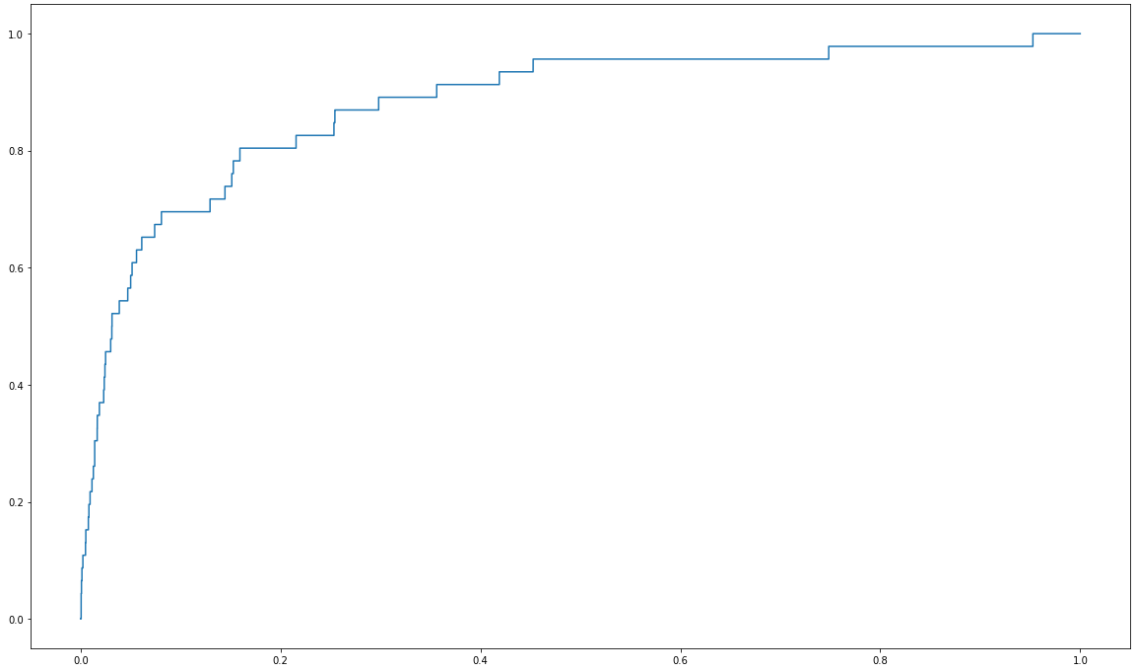


Figura 55: Curva ROC correspondiente al modelo de Finance and Q2C con un AUC de 0,73.

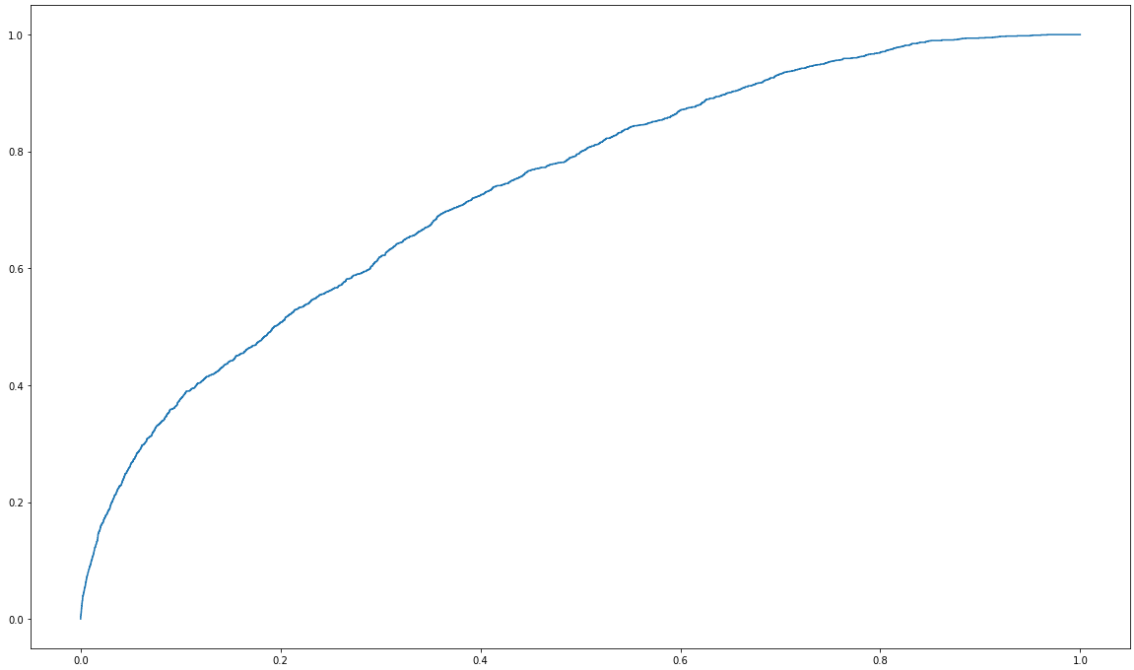


Figura 55: Curva ROC correspondiente al modelo de Industry-specific Applications con un AUC de 0,78.

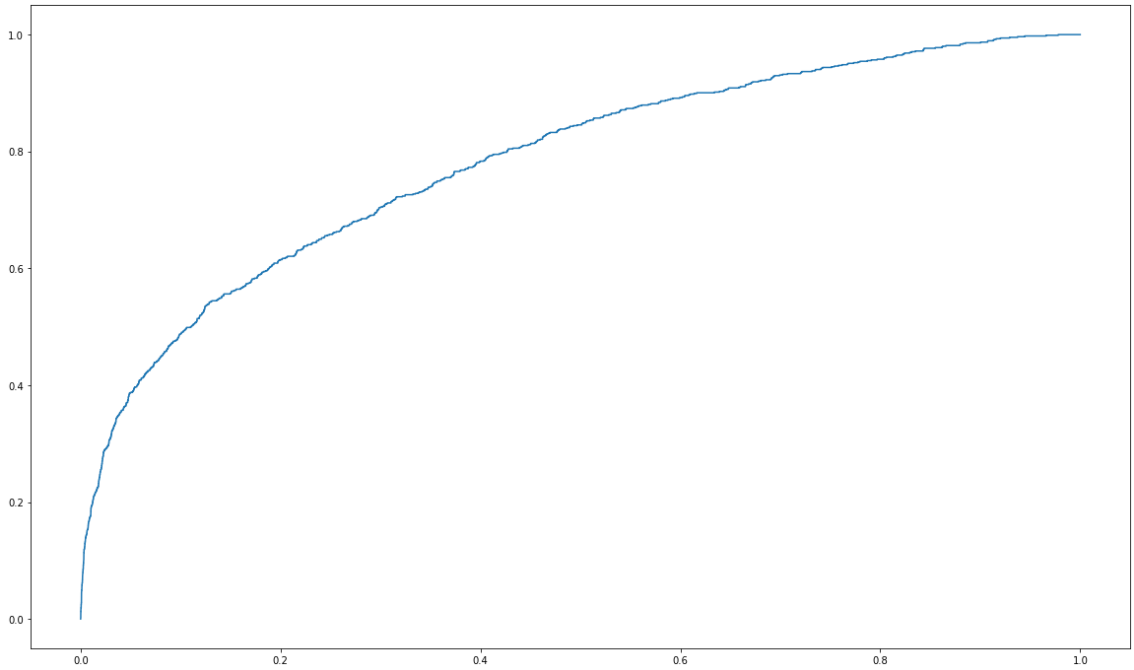


Figura 55: Curva ROC correspondiente al modelo de Learning and Talent con un AUC de 0,71.

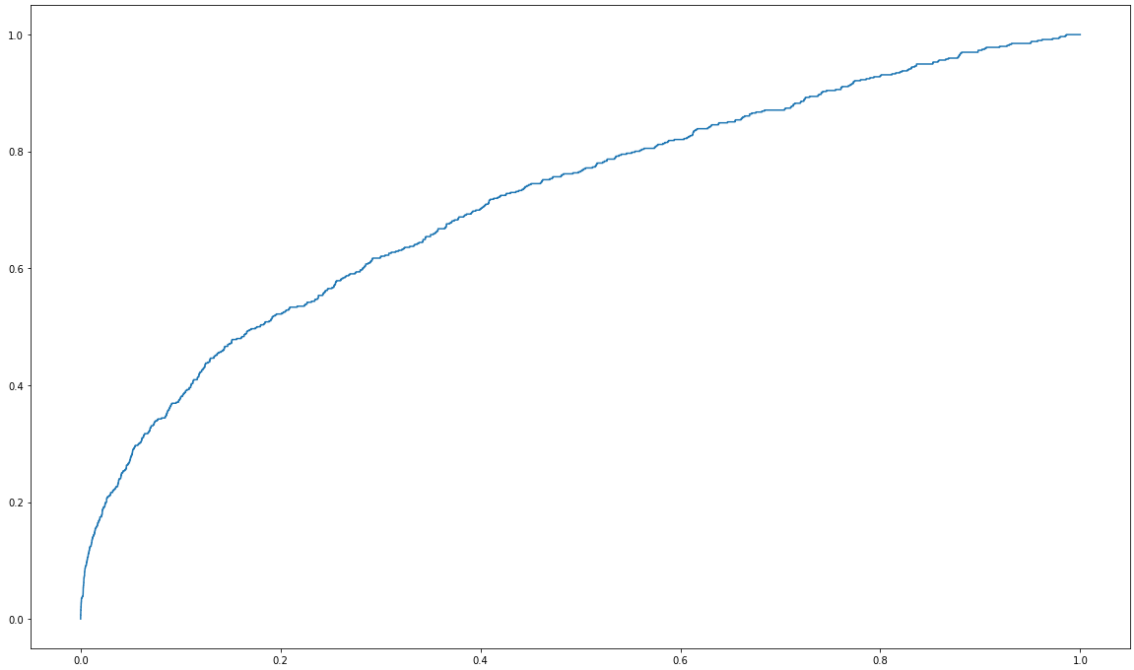


Figura 56: Curva ROC correspondiente al modelo de Marketing con un AUC de 0,79.

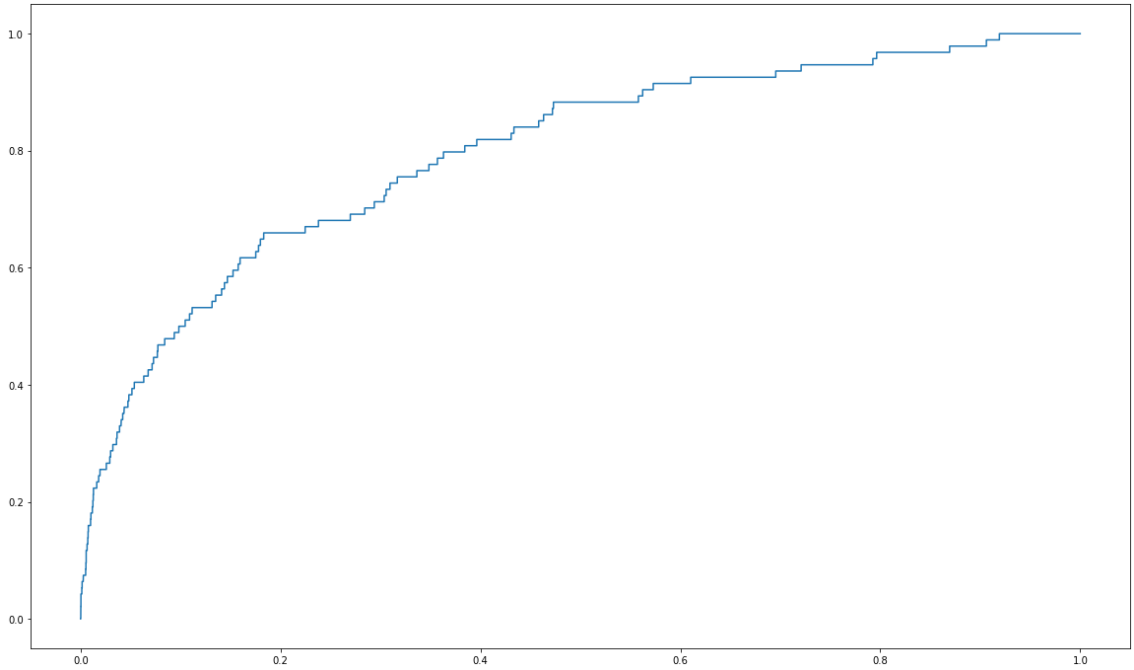


Figura 57: Curva ROC correspondiente al modelo de Planning and Analytics con un AUC de 0,8.

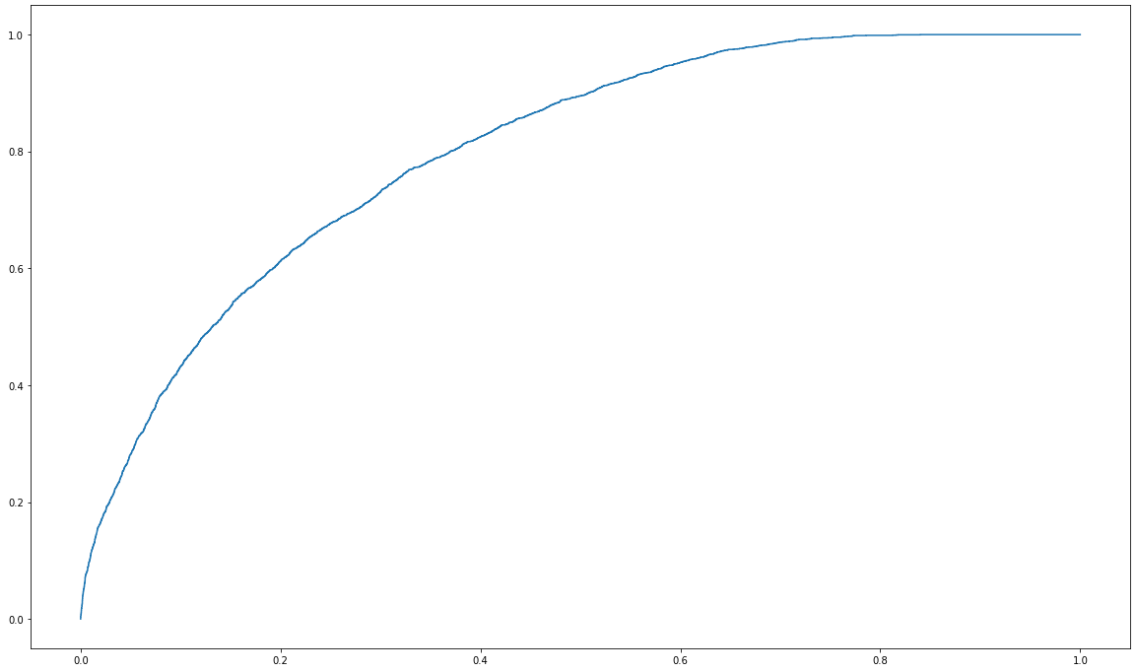


Figura 58: Curva ROC correspondiente al modelo de Procurement con un AUC de 0,69.

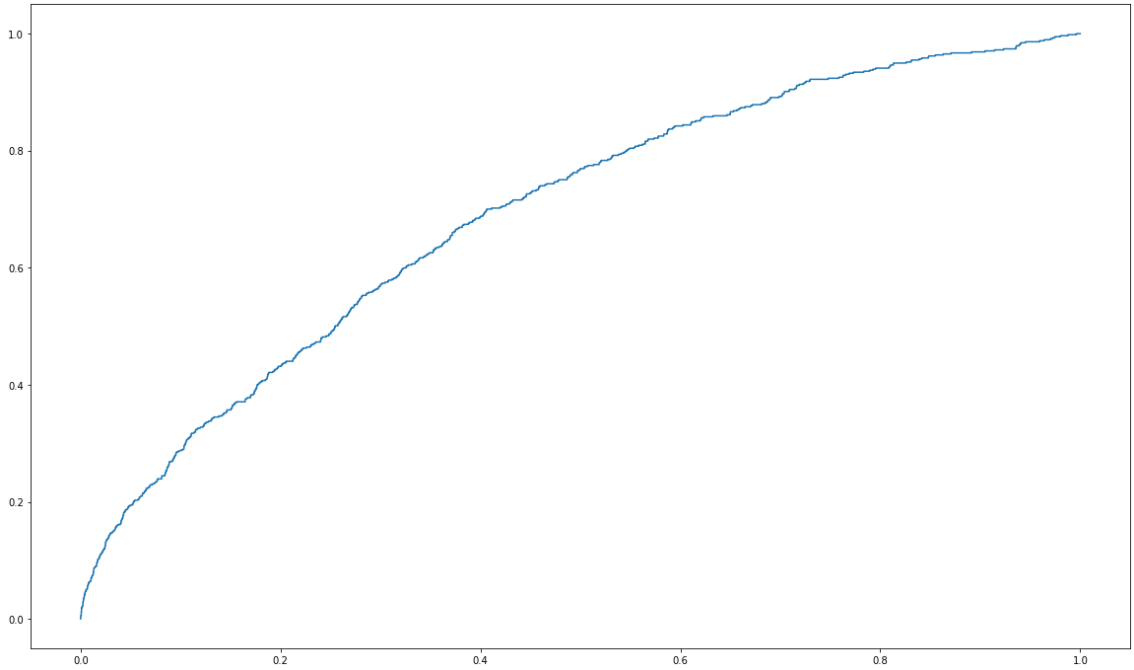


Figura 59: Curva ROC correspondiente al modelo de SAP Signavio con un AUC de 0,88.

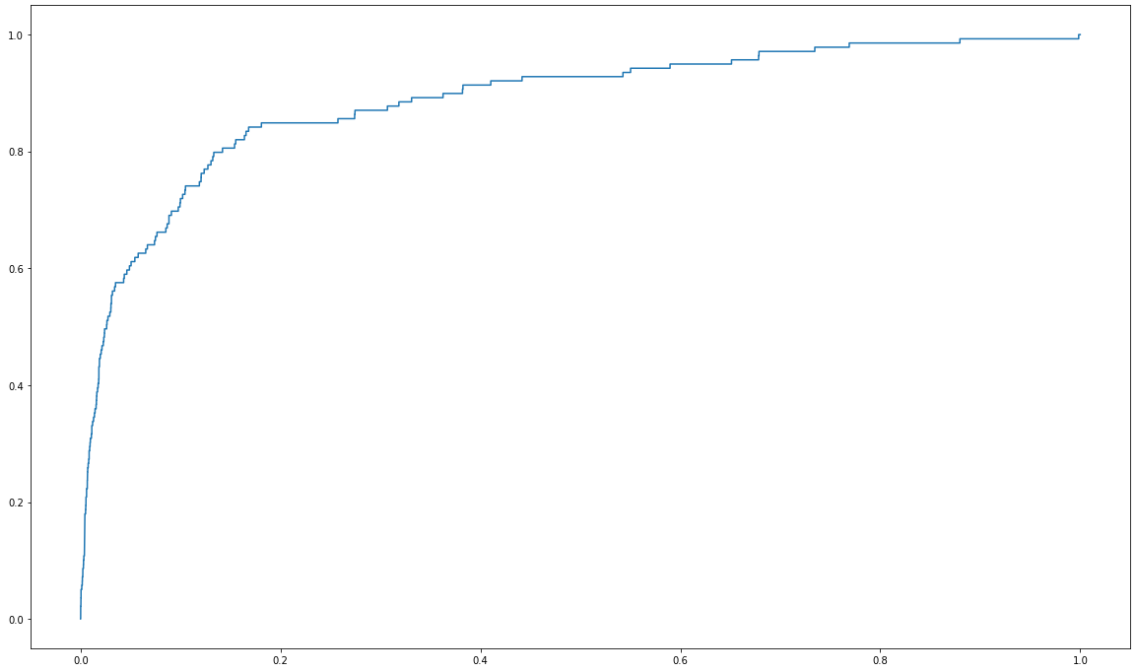


Figura 60: Curva ROC correspondiente al modelo de Sales Performance Management con un AUC de 0,83.

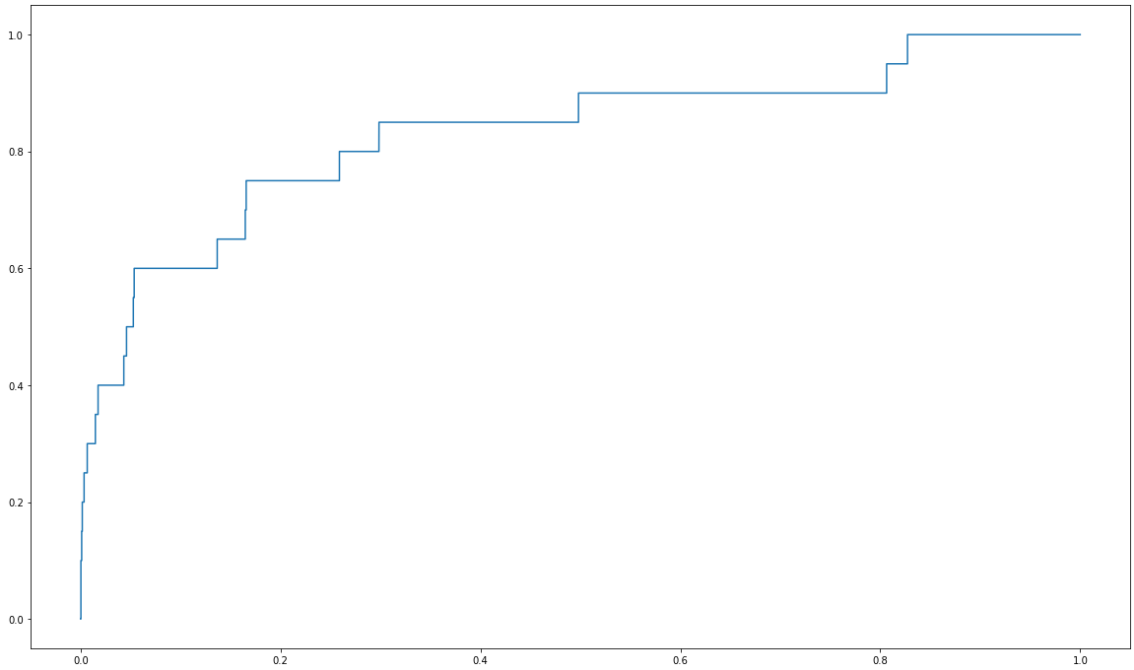


Figura 61: Curva ROC correspondiente al modelo de Sales and Service con un AUC de 0,69.

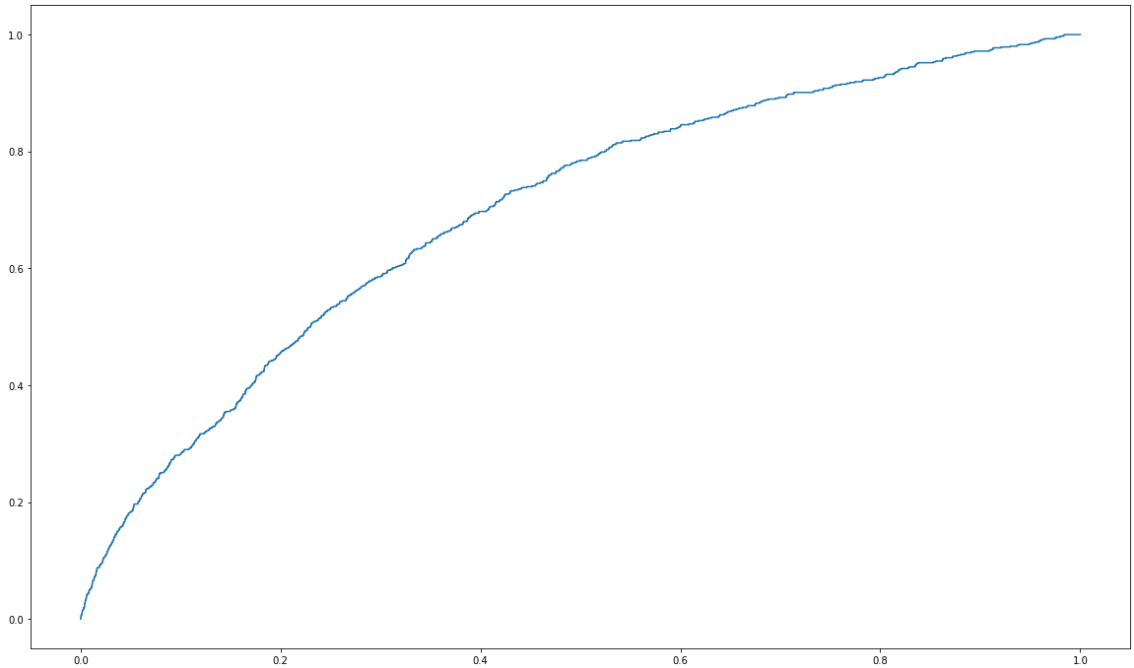


Figura 62: Curva ROC correspondiente al modelo de SuccessFactors Cross con un AUC de 0,82.

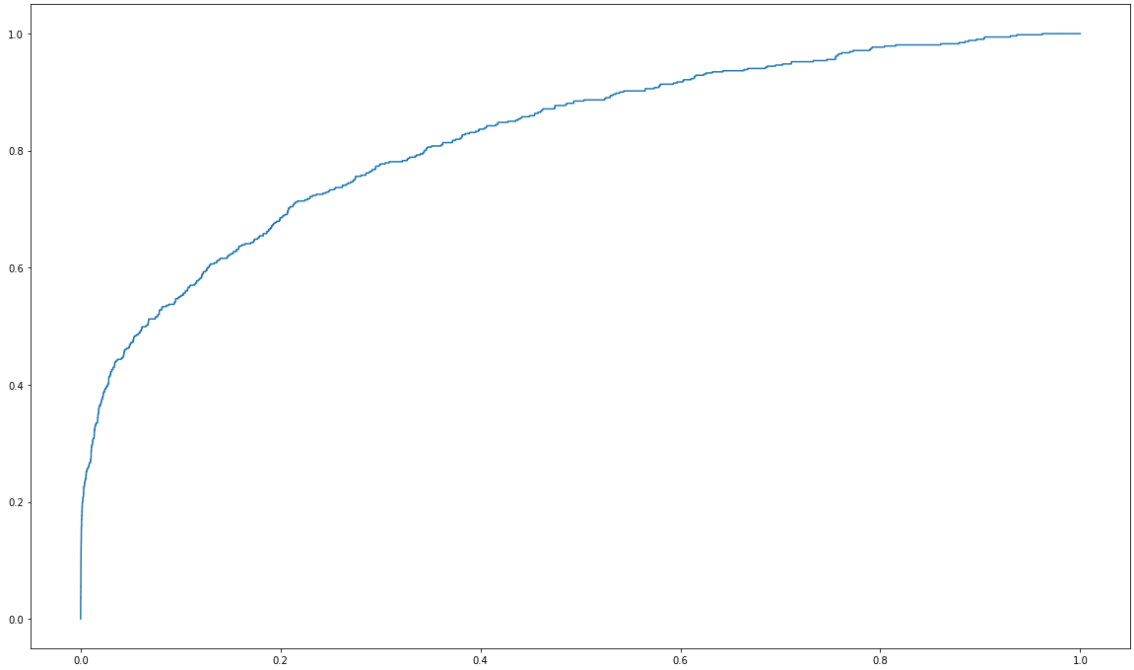


Figura 63: Curva ROC correspondiente al modelo de Training and Adoption con un AUC de 0,69.

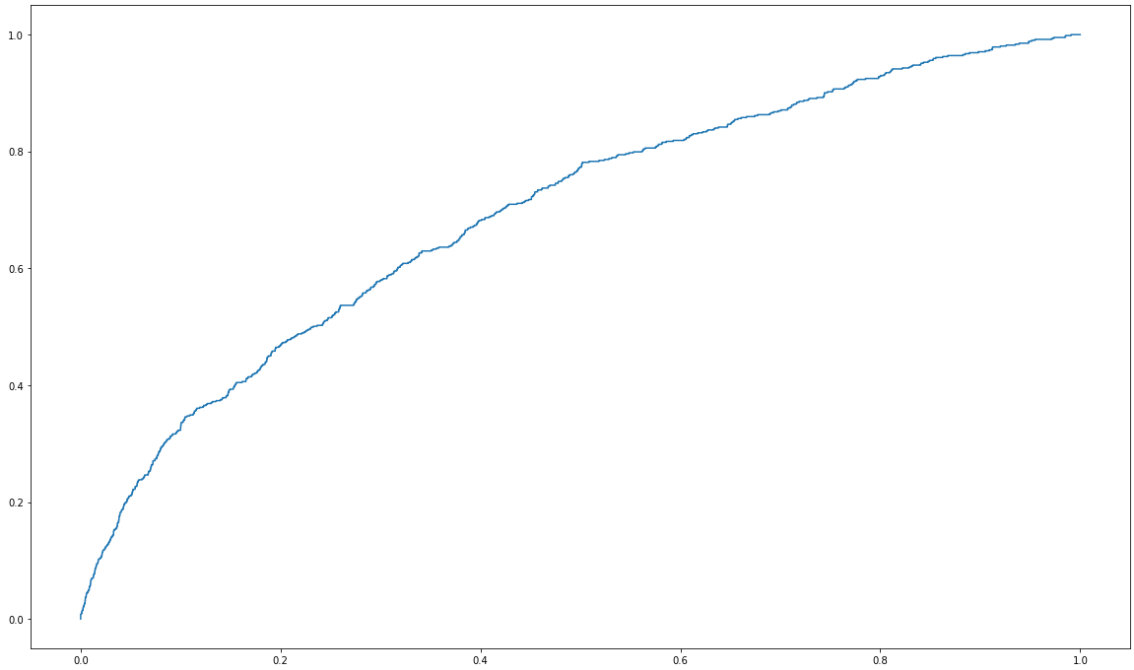
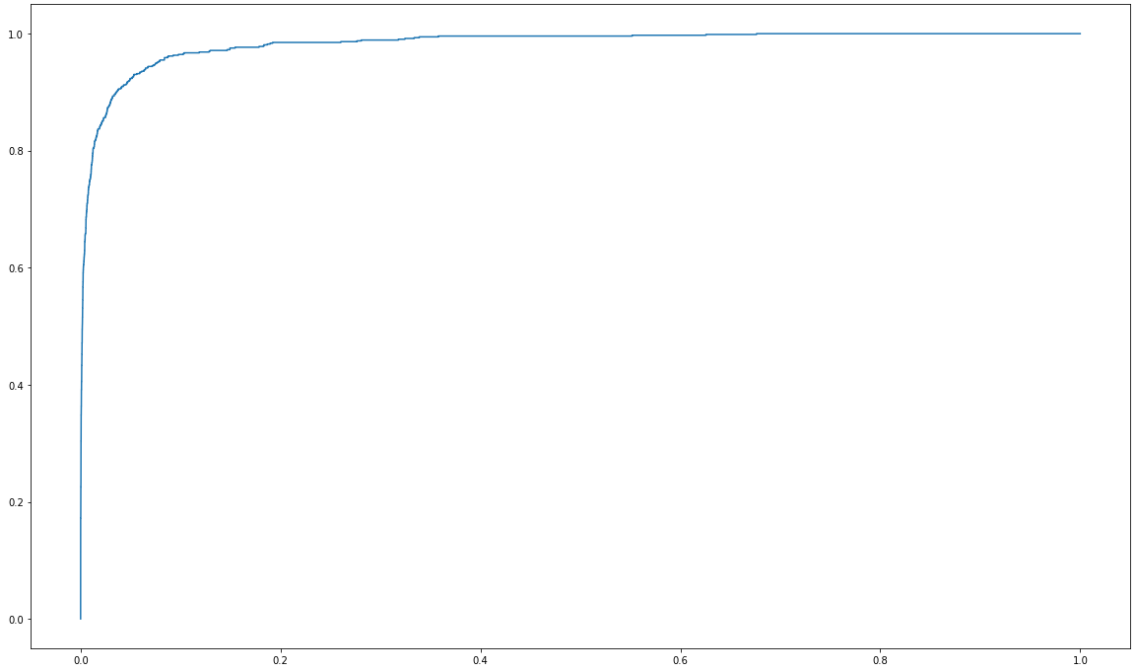


Figura 64: Curva ROC correspondiente al modelo de Travel and Expense con un AUC de 0,98.



6. Bibliografía

- Gareth J., Witten D., Hastie T., y Tibshirani R. (2013). *An introduction to statistical learning*. New York: Springer.
- Max Bramer. (2007). *Principles of data mining*. Vol. 180. London: Springer.
- Tan P.N., Steinbach M., and Kumar V. (2005). *Introduction to Data Mining*. Pearson.
- Alpaydin, Ethem. (2014). *Introduction to machine learning*. MIT press.
- Jerome Friedman, Trevor Hastie, y Robert Tibshirani. (2001). *The elements of statistical learning*. New York: Springer.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.
- Japkowicz, N. (2003). *Class imbalances: are we focusing on the right issue. In Workshop on Learning from Imbalanced Data Sets II*.
- Visa, S., & Ralescu, A. (2005). *Issues in mining imbalanced data sets-a review paper. In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*.
- Chen, T., & He, T. (2015). *Higgs boson discovery with boosted trees. In NIPS 2014 workshop on high-energy physics and machine learning.*
- Corey Wade. (2020). *Hands-on gradient boosting with xgboost and scikit-learn: perform accessible machine learning and extreme gradient boosting with python*. Birmingham: Packt.
- Drew Conway, John Myles White. (2012). *Machine Learning for Hackers: Case Studies and Algorithms to Get You Started.* Sebastopol: O'Reilly.
- Andreas C. Müller & Sarah Guido. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* Sebastopol: O'Reilly.
- William McKinney. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Sebastopol: O'Reilly.
- Allen Downey, Jeffrey Elkner, Chris Meyers. (2002). *Aprenda a Pensar como un Programador (con Python),* Green Tea Press.
- T.H. Cormen, (2009). *Introduction to Algorithms*, tercera edición, MIT Press.
- Nathalie Japkowicz, Mohak Shah. (2014). *Evaluating Learning Algorithms: A Classification Perspective*. New York: Cambridge University Press.