PROYECTO PILOTO DE BUSINESS INTELLIGENCE Y DATA SCIENCE

Grupo de Empresas COMGRAP

Abstract

Propuesta inicial para la creación de procesos de extracción, procesamiento y análisis de datos de negocio haciendo uso de inteligencia de negocio y ciencia de datos junto con la aplicación de algoritmos de aprendizaje automático para la predicción de ventas.

Eng. M.Data.Sc. Felix Enzo Garofalo

felixenzogarofalo@gmail.com

Contenido

RESUMEN

En la empresa COMGRAP es vital contar con un adecuado estimado de ventas futuras a fin de establecer metas sostenibles para el equipo de vendedores(as), conocido como V1 dentro de la empresa. El estimado de ventas permite a la gerencia (en este caso conformada por el CEO y la Gerente de Operaciones) hacer un mejor seguimiento del desempeño del equipo.

COMGRAP es así mismo partner de diversas empresas tecnológicas, tales como Autodesk, Adobe, Microsoft y HP. Dichos partners solicitan que sus asociados cumplan con cuotas de ventas mensuales en servicios, software y consumibles. Lo que es un punto extra que resalta la necesidad de contar con información acertada para la gestión del equipo de ventas.

En el presente trabajo se propone la creación de un algoritmo predictivo utilizando Machine Learning que facilite y mejore la estimación de ventas futuras de la empresa tomando en cuenta factores macroeconómicos del mercado en cuestión y datos históricos de ventas. También se propone la creación de un tablero de inteligencia de negocios que facilite la tarea general de gestión del equipo de trabajo, más allá de la creación del algoritmo, que presente datos útiles y oportunos para la toma de decisiones relacionadas con el equipo V1.

Los resultados obtenidos con el entrenamiento del algoritmo predictivo resultan aceptables en exactitud a fines de establecer metas razonables y oportunas. Al mismo tiempo, gracias al desarrollo de tablero para BI se ha facilitado el análisis de datos por parte de la alta gerencia, quien ahora tiene la intención de invertir en este tipo de desarrollo para otros departamentos.

INTRODUCCIÓN

Vivimos en lo que comúnmente se conoce como la era digital. Un ecosistema superpuesto de tecnologías digitales, cada una de las cuales se basa en las anteriores y cataliza las que vendrán, está transformando no solo nuestras vidas personales y comunitarias, sino también la dinámica de los negocios para organizaciones de todos los tamaños en todas las industrias (Rogers, 2016).

Las tecnologías digitales están transformando no solo algún aspecto de la gestión empresarial, sino prácticamente todo aspecto. Están reescribiendo las reglas de los clientes, la competencia, los datos, la innovación y el valor. Responder a estos cambios requiere más que un enfoque fragmentario; exige un esfuerzo integrado total: un proceso de transformación digital integral dentro de la empresa.

En reconocimiento de la realidad expuesta en las líneas anteriores COMGRAP ha decidido emprender su propio viaje hacia la transformación digital. Esta es sin duda una propuesta ambiciosa y por ello se ha establecido que el proceso sea paulatino, orientado a objetivos estratégicos y, sobre todo, irreversible. La necesidad de volver los procesos de la empresa a ser más orientados a los datos, o Data-Driven, se ha vuelto imperativa en los últimos años, donde ha servido de catalizador el efecto de pandemia sobre el mercado local. De ahí la importancia de que la toma de decisiones se base en hechos medibles, sin dejar de lado la experiencia humana que es de inmenso valor. Sino más bien haciendo una fusión entre ambos (datos y experiencia) que resulte en el beneficio de la organización y sus miembros por individual.

Como ya se dijo, el proceso de transformación digital de una organización es de por sí una empresa ambiciosa. Y como tal, debe ser abordada de forma adecuada para que no se pierda el

impulso inicial y para que la frustración no opaque los logros que puedan obtenerse en etapas tempranas. Es por ello que se ha decidido hacer un estudio de algunas tareas clave de gestión dentro de la empresa que sean especialmente susceptibles a ser mejoradas con el análisis de datos y cuya complejidad resulte adecuada para una primera etapa de implementación. Al poner el foco en optimizar dichas tareas clave, se promueve la obtención de resultados tangibles a la vez que se gana experiencia que resultará útil en las siguientes etapas de implementación.

COMGRAP, empresa en cuestión, es partner de diversas empresas tecnológicas, tales como Autodesk, Adobe, Microsoft y HP. Dichos partners solicitan que sus asociados cumplan con cuotas de ventas mensuales en servicios, software y consumibles. Esto vuelve una tarea vital de gestión la estimación de ventas en meses o trimestres futuros. Y permite identificar la creación de algoritmos de predicción como un objetivo estratégico primario.

En el mismo orden de ideas, y a través de diversas conversaciones con la gerencia de ventas, se ha identificado que muchas de las pérdidas en la gestión del equipo están relacionadas con la falta de información oportuna y actualizada para dar soporte a la toma de decisiones. En el presente contexto, por pérdidas nos referimos a acciones (o falta de acciones) que no resultan en aportes para la cadena de creación de valor. Algunos ejemplos incluyen, pero no se limitan a: la reacción oportuna a eventos que pueden afectar las finanzas (tales como el cambio frente a moneda extranjera o la creación de políticas por parte de estado), el entendimiento de costos de oportunidad frente a fechas claves de marketing, el desempeño individual de vendedores y su gestión de carteras con clientes importantes.

En el pasado, ambos objetivos estratégicos (estimación de ventas e inteligencia de negocio) han sido abordados haciendo uso de herramientas de ofimática como Excel y PowerPoint. En el caso de estimación, se usan los valores de ventas del año anterior para predecir la venta del año actual y se le agrega el valor de varianza del mes correspondiente del año anterior como prospección de crecimiento. En el caso de la inteligencia de negocio está consiste principalmente en la creación de gráficos de Excel y su divulgación está basada en presentaciones de PowerPoint.

La propuesta del presente proyecto busca mejorar el alcance y efectividad de los objetivos estratégicos planteados haciendo uso de la ciencia de datos y la automatización de procesos relacionados con la captura, transformación y carga de dichos datos. Se plantea, por lo tanto:

- La creación de un algoritmo predictivo que saque provecho de datos externos que definen el mercado mientras se saca provecho del conocimiento interno de los clientes.
- La creación de un primer tablero de inteligencia de negocio enfocado en presentar de manera práctica y útil la información pertinente a la gestión del equipo de ventas.

Si bien es cierto que los conceptos y tecnologías planteadas no son nuevas para muchas empresas alrededor del mundo, sí que resulta una propuesta innovadora para la organización en cuestión. La respuesta que se formule a la cuestión de la transformación digital, será vital para la supervivencia y competencia de la empresa en años futuros. Por lo tanto, resulta más que adecuado y oportuno el lanzamiento de este proyecto piloto.

El procedimiento seguido para lograr la solución planteada fue el siguiente:

1. Obtener datos de fuentes principales: como base del proceso digital se encuentra la recolección oportuna de datos. Como fuente de datos interna se utilizó el CRM de la empresa. Y como fuente de datos externa se utilizó la página oficial del Banco Central de Chile de donde se tomaron datos de variables macroeconómicas.

2. Proceso ETL: a fin de limpiar los datos de valores espurios se automatizó el proceso de lectura, tratamiento y guardado en un formato oportuno de las tablas extraídas directamente de las fuentes de información. El proceso ETL fue vertido en un cuaderno de Jupyter para facilitar la documentación y entendimiento a actores futuros.

3. Análisis Exploratorio de Datos (EDA): en base a los datos previamente tratados se llevó a cabo una exploración más profunda de los datos desde un punto de vista estadístico. Y luego se procedió a buscar patrones útiles con la creación de gráficos e incluso algunos modelos estadísticos. Este paso igualmente quedó plasmado en un cuaderno de Jupyter.

4. Ingeniería de Características y Modelos: utilizando conocimiento del dominio se procedió a seleccionar y validar las variables a partir de los datos originales, buscando mejorar de esta manera el rendimiento de los algoritmos a ser creados. Finalmente se crearon diversos modelos predictivos, comparando su rendimiento, para seleccionar el de mejor desempeño en el cuaderno de Jupyter. Este modelo será llevado a producción en etapas futuras de desarrollo.

5. Preparación de Datos para Inteligencia de Negocio: debido a que en los tableros de visualización en necesario mostrar métricas específicas y personalizadas, se procedió a crear un cuaderno de Jupyter específicamente dirigida al pre-procesado de dichas métricas y KPIs de interés para el equipo de ventas.

6. Creación de Tablero en Power BI: finalmente se creó un tablero siguiendo el esquema DAR (Dashboard - Analysis - Report) para facilitar la inspección y seguimiento de métricas para el equipo de ventas.

Las predicciones obtenidas con el algoritmo creado demostraron capacidad de tomar en cuenta variaciones significativas de ventas, incluso a frecuencia diaria. Esta información facilitará el establecimiento de metas objetivas en función de las condiciones del mercado. Igualmente, el tablero de Power BI para inteligencia de negocios referido al equipo de ventas ha facilitado el seguimiento por parte de la gerencia y se planea repetir el procedimiento para otros departamentos.

A continuación, se describe el estado del arte en el desarrollo de algoritmos predictivos, haciendo énfasis en el tratamiento de series temporales con el uso de Machine Learning. Con esto mente se establecerán los objetivos específicos de la presente investigación, la solución planteada, los criterios de evaluación establecidos como criterio de éxito y los resultados obtenidos. Finalmente, se ofrecen conclusiones y recomendaciones para desarrollos futuros.

OBJETIVOS

OBJETIVO GENERAL

Crear un algoritmo predictivo de montos de ventas a ser alcanzados por el equipo de ventas a fin de definir adecuadamente metas para el mismo equipo.

OBJETIVOS ESPECÍFICOS

- Extraer datos desde fuentes internas (CRM) y fuentes externas (Banco Central de Chile) y automatizar el proceso de preparación y limpieza para ser utilizados en pasos siguientes.
- Llevar a cabo un Análisis Exploratorio de Datos para comprender la naturaleza de los datos de entrada e identificar patrones explotables por el algoritmo de predicción.
- Comparar el rendimiento de diversos modelos predictivos con distintas configuraciones e identificar el que mejor se ajuste al alcance de las metas propuestas.
- Crear un tablero de visualización utilizando el esquema DAR para mejorar la toma de decisiones relativa al equipo de ventas.

SOLUCIÓN PLANTEADA

A fin de lograr los objetivos propuestos se siguió una metodología de seis (6) pasos que englobaron el ciclo completo de los datos desde su obtención hasta su uso en el entrenamiento de redes neuronales. Tal y como se mencionó brevemente en la introducción los pasos fueron los siguientes:

1. Obtener datos de las fuentes principales.
2. Automatizar el proceso de Extracción, Transformación y Carga de datos.
3. Realizar un Análisis Exploratorio de Datos.
4. Ejecutar la ingeniería de características y de modelos predictivos.
5. Preparar datos para inteligencia de negocios.
6. Crear un tablero de inteligencia de negocios para el equipo de ventas.

El detalle de cada uno de esos pasos se desglosa a continuación.

ETAPA 1: OBTENCIÓN DE DATOS

En la base de cualquier proceso digital se encuentra la recolección oportuna de datos. En aras de crear algoritmos predictivos que tomaran en cuenta el comportamiento del mercado fue necesario hacer uso de datos internos a la empresa y datos externos a la misma.

Datos internos: en este caso se utilizaron los datos internos provenientes del CRM de la empresa. Este es un CRM programado a la medida para COMGRAP que un desarrollador particular creó según las especificaciones de la organización. Aunque actualmente se está migrando a un CRM con más potencialidades (HubSpot) por el momento la información contenida en el CRM fue más que suficiente. Otra ventaja fue que los datos de ventas del ERP se consolidan en el CRM mensualmente para facilitar el trabajo del equipo de ventas. Por lo tanto, con el CRM se obtuvo acceso a toda la información de

ventas necesaria. El CRM actual no cuenta con un API, solo permite la exportación de datos en formato Excel. Así que fue necesario exportar manualmente datos de ventas mensuales.

Datos externos: el objetivo de contar con datos externos es poder tomar en cuenta cómo las variables económicas del mercado pudieran afectar el desempeño interno en la empresa. La fuente más confiable para la obtención de variables macroeconómicas es la página oficial del Banco Central de Chile (https://www.bcentral.cl/). Esta página ofrece estadísticas de indicadores diarios de tipos de cambio, tasas de interés, producto interno bruto, entre otros. La página oficial del Banco Central de Chile tiene una API para la ingesta de datos por parte de terceros. Sin embargo, el acceso a esta API no es público, se debe hacer una solicitud al Banco para recibir la credenciales de acceso. Esta solicitud fue enviada, pero a la fecha no hubo respuesta. Así que se procedió a descargar los datos de forma manual desde el front end de acceso público. Los datos fueron descargados en formato CSV.

ETAPA 2: EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE DATOS (ETL)

Antes de pasar a analizar los datos y crear algoritmos predictivos es necesario transformarlos adecuadamente para evitar valores inadecuados y finalmente cargarlos en la base de datos a ser utilizada. Este proceso de extracción, transformación y carga (ETL por sus siglas en inglés) se automatizó haciendo uso de python en el archivo *"01_extract_transform_load.ipynb"*.

Fueron tratados tres grupos básicos de datos:

- Datos de clientes (provenientes del CRM)
- Datos de ventas (provenientes del CRM)
- Datos de variables macroeconómicas (Banco Central de Chile)

Limpieza de datos

En el caso de los **datos de clientes**, existían una gran cantidad de valores nulos en buena parte de las columnas. Así que se decidió mantener sólo aquellas columnas cuyo conteo de nulos no superara las tres cifras. Además, para facilitar el procesamiento futuro de esta tabla se ejecutaron las siguientes operaciones: se ordenaron las columnas haciendo que el ID quedara en primer lugar, se eliminaron la filas donde seguían existiendo valores nulos y se cambió el tipo de datos de la columna ID a "int".

En el caso de los **datos de ventas**, debido a que originalmente los datos fueron descargados desde CRM por mes, en primer lugar se procedió a unificar los diversos archivos de ventas por mes en un único DataFrame. Luego, haciendo un conteo de valores nulos por cada columna se estableció el criterio de prescindir de las columnas cuyo conteo de nulos superara los 10.000 valores. En las columnas restantes se siguieron los siguientes criterios para el tratamiento de nulos:

- En la columna industria, se rellenó con la palabra "GENERAL" los valores nan

- En la columna Vendedor se rellenó con "NO APLICA" los valores nan
- En las columnas columnas "Grupo" y "Subgrupo" se rellenó con "SIN GRUPO"
- En las columnas "Cantidad", "Valor", "Costo", "Margen", "Costo Manual", "Recargo IVA" se rellenó con 0.
- En el caso de las fechas, se rellenaron los valores nulos utilizando el método "Forward Fill".
- Se eliminaron las líneas donde el valor de "Factura Venta" resultó nulo.
- El resto de los valores null, se completó con el valor "-1". De tal manera que en procesos futuros que impliquen el uso de los datos, se puedan filtrar estos valores a discreción.

Para el caso de los **datos de variables macroeconómicas**, estos ya venían bastante limpios y preprocesados por parte del ente emisor. Así que no fue necesaria ninguna acción al respecto.

Transformación de datos

En el caso de los **datos de venta** fue necesario formatear adecuadamente las fechas, que originalmente tenían formato de cadena de texto. Así que se convirtieron los valores de fechas a Datetime y los valores enteros a “int”. Finalmente, se ordenaron los registros según su fecha de factura de venta.

Lo mismo fue necesario para el caso de los **datos de variables macroeconómicas**. Se transformaron las fechas a formato Datetime. Finalmente, se unificaron los valores de las distintas variables macroeconómicas en un único DataFrame llamado “vme”.

Persistencia de datos

Cada uno de los DataFrame creados se guardó en una única base de datos en formato HDFS (Hadoop Distributed File System). La razón de escoger este tipo de archivo fue para facilitar la escalabilidad a futuro a medida que el volumen de datos aumente y los alcances de ciencia de datos en la organización se incrementen.

El formato HDFS permite guardar diversas tablas en forma de DataFrames con identificadores clave. El nombre del archivo es “data.h5” y los registros realizados en esta tabla fueron los siguientes:

- Datos de clientes: clave *“clientes”*
- Datos de ventas: clave *“ventas”*
- Datos de variables macroeconómicas: clave *“vme”*.

ETAPA 3: ANÁLISIS EXPLORATORIO DE DATOS

El Análisis Exploratorio de Datos se ejecutó haciendo uso de Python y herramientas de análisis sobre DataFrame de Pandas. El proceso completo está registrado en el archivo *“02_exploratory_data_analysis_.ipynb”*.

Principal objetivo de análisis

Se busca establecer patrones de relación en los datos de ventas de la empresa que permitan describir y predecir ventas futuras.

Como objetivo secundario se pretende establecer un modelo predictivo para el abastecimiento de inventario de productos específicos

Entregables principales

Como resultado del análisis se deberá crear un reporte que incluya tablas resumen, gráficos, mapas y/o diagramas para presentar la información a los encargados de la toma de decisiones.

En base a los resultados obtenidos se hará una propuesta de los tipos de modelos predictivos que mejor se ajuste a los datos explorados.

Roles y responsabilidades

Científico de datos: a los fines del presente proyecto el especialista en ciencia de datos deberá encargarse del proceso completo de extracción, limpieza, ingeniería y análisis de los datos, así como de la presentación de los resultados del análisis.

Características principales de los datos

En el caso del DataFrame de ventas se cuenta con una tabla de 188.809 filas x 28 columnas. El nombre y tipo de dato de cada columna se muestra en la imagen *"Información de DataFrame de Ventas"*.

Para describir mejor el comportamiento de los valores numéricos se agregó una tabla con estadísticas básicas para columnas numéricas, tal como se muestra en la imagen *"Descripción de DataFrame Ventas"*. Como se observan valores números de magnitud -1 en ciertos campos, será importante tratarlos de forma adecuada a la hora de estimar otras métricas estadísticas. Se deberán evitar cantidades con valores negativos. O asumir que se trata de devoluciones o compras.

La representación por histograma de los valores numéricos nos ayuda a entender que hay ciertas variables muy pocos valores probables, tal y como es de esperarse en el campo de "Cantidad". Muchas veces la cantidad será simplemente 1. El resumen de los histogramas se presenta en la imagen *"Histogramas de DataFrame de Ventas".* De este gráfico se puede decir que:

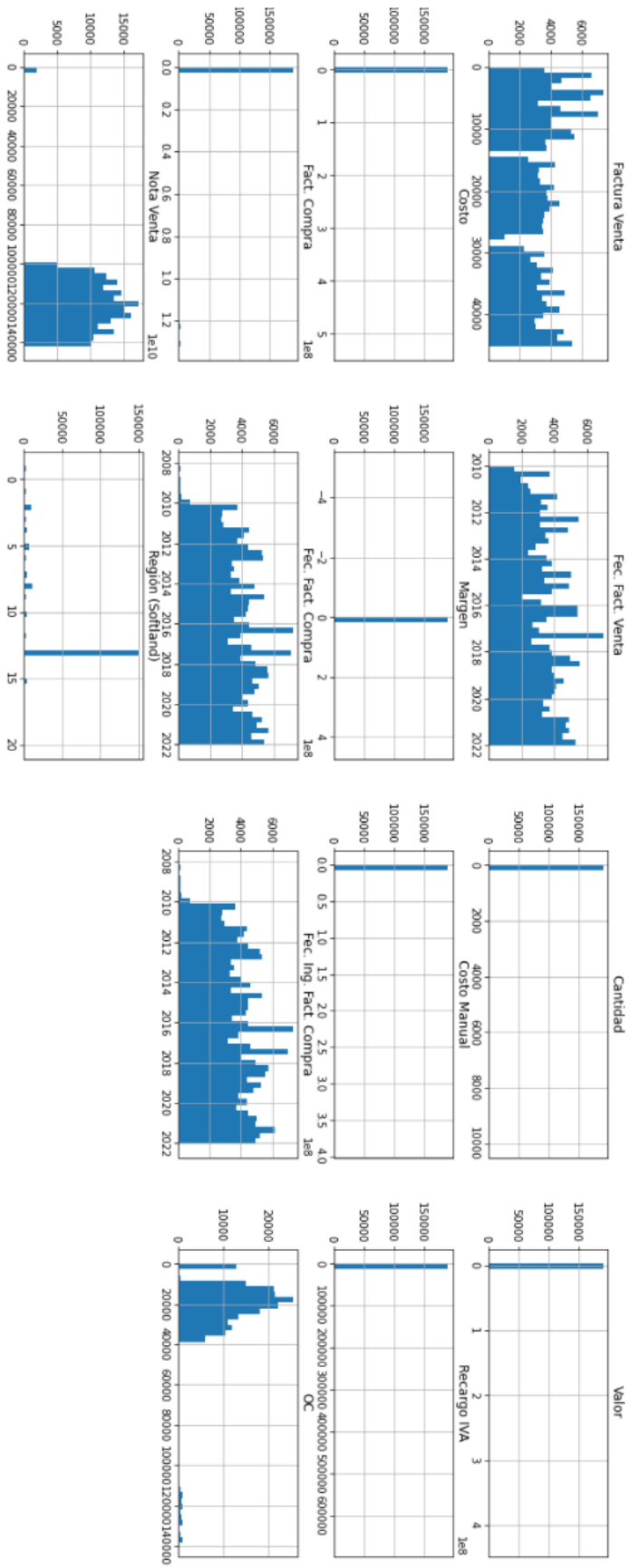
- Al visualizar la distribución de "Fec. Fact. Venta" se puede observar un máximo significativo en el 2017 y que en general la tendencia es al alza. También, se pueden observar ciclos repetitivos de ventas en períodos de 3 años. Esto parece corresponder a las suscripciones de ciertos productos Autodesk que se renuevan cada 3 años.
- La distribución de valores de Región Softland parece indicar que la zona con mayor proporción de ventas es la región 13.

```
Data columns (total 28 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Factura Venta           188809 non-null  int64
 1   Fec. Fact. Venta        188809 non-null  datetime64[ns]
 2   Cod. Producto           188809 non-null  object
 3   Descripción             188809 non-null  object
 4   Grupo                   188809 non-null  object
 5   Subgrupo                188809 non-null  object
 6   Cantidad                188809 non-null  float64
 7   Valor                   188809 non-null  float64
 8   Costo                   188809 non-null  float64
 9   Margen                  188809 non-null  float64
 10  Costo Manual            188809 non-null  float64
 11  Recargo IVA             188809 non-null  float64
 12  Rut                     188809 non-null  object
 13  Cliente                 188809 non-null  object
 14  Industria               188809 non-null  object
 15  Vendedor                188809 non-null  object
 16  Tipo Vendedor           188809 non-null  object
 17  Fact. Compra            188809 non-null  int64
 18  Fec. Fact. Compra       188809 non-null  datetime64[ns]
 19  Fec. Ing. Fact. Compra  188809 non-null  datetime64[ns]
 20  Proveedor               188809 non-null  object
 21  OC                      188809 non-null  int64
 22  Nota Venta              188809 non-null  int64
 23  Región (Softland)       188809 non-null  int64
 24  Comuna (Softland)       188809 non-null  object
 25  Región (CRM)            188809 non-null  object
 26  Comuna (CRM)            188809 non-null  object
 27  Se calculó el costo     188809 non-null  object
dtypes: datetime64[ns](3), float64(6), int64(5), object(14)
```

Información de DataFrame de Ventas

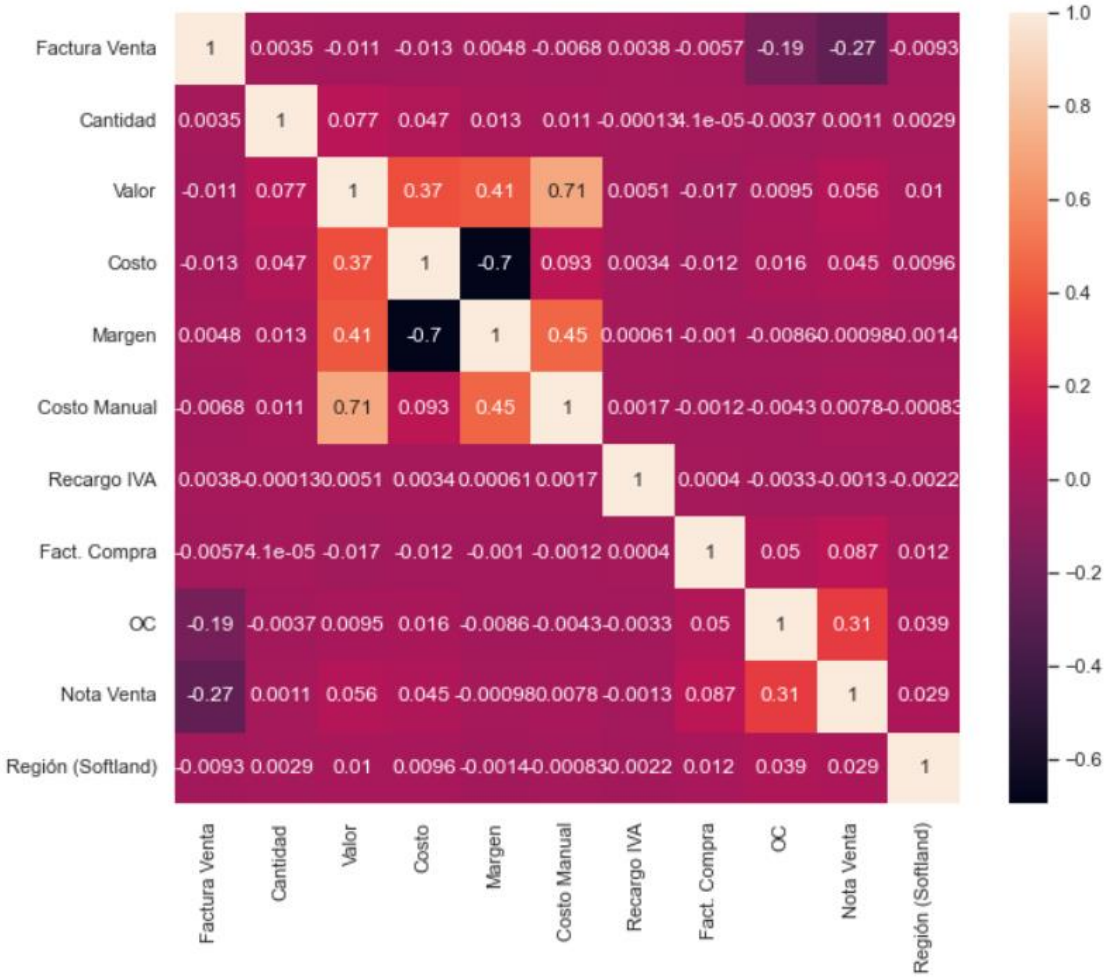
	count	mean	std	min	25%	50%	75%	max
Factura Venta	188809.0	2.118504e+04	1.386383e+04	1.0	8236.0	20238.0	34165.0	4.510800e+04
Cantidad	188809.0	1.624027e+00	3.698262e+01	-4.0	1.0	1.0	1.0	1.000000e+04
Valor	188809.0	3.391423e+05	1.551688e+06	-2813720.0	33500.0	105600.0	348565.0	4.269105e+08
Costo	188809.0	2.859021e+05	1.978028e+06	-4462014.0	25812.0	84662.0	302094.0	5.231724e+08
Margen	188809.0	5.324020e+04	2.008717e+06	-506072400.0	4200.0	13873.0	43020.0	4.269105e+08
Costo Manual	188809.0	1.739602e+04	1.008210e+06	0.0	0.0	0.0	0.0	3.826196e+08
Recargo IVA	188809.0	1.651243e+01	2.115936e+03	0.0	0.0	0.0	0.0	6.650000e+05
Fact. Compra	188809.0	1.161233e+08	1.188755e+09	-1.0	42889.0	249143.0	716063.0	1.320001e+10
OC	188809.0	2.172444e+04	1.719592e+04	-1.0	13623.0	19390.0	26342.0	1.383150e+05
Nota Venta	188809.0	1.200500e+05	1.586118e+04	0.0	111525.0	120927.0	129958.0	1.417950e+05
Región (Softland)	188809.0	1.159853e+01	3.257191e+00	-1.0	13.0	13.0	13.0	2.000000e+01

Descripción de DataFrame Ventas



Histogramas de DataFrame de Ventas

Siguiendo con la exploración de los datos, se realiza a continuación un ploteo de la matriz de correlación con un mapa de calor, que puede verse en la imagen *"Matriz de Correlación para DataFrame Ventas"*.



Matriz de Correlación para DataFrame Ventas

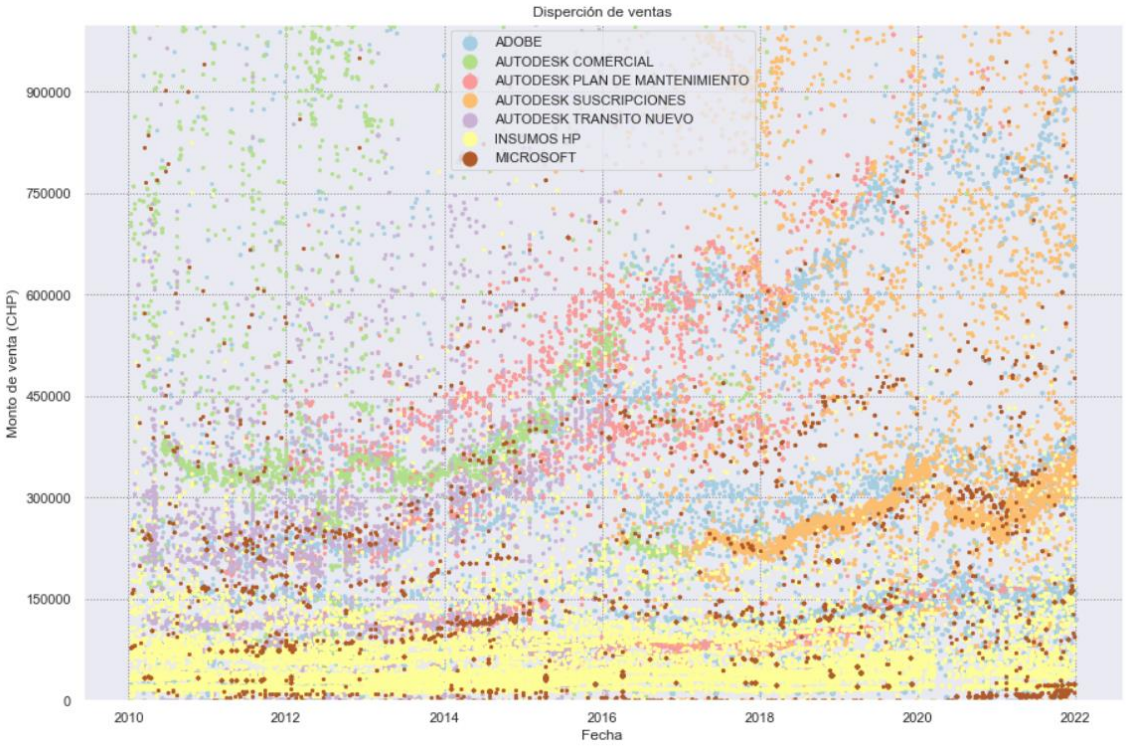
En esta matriz, se resalta una alta correlación entre los campos "Valor", "Costo", "Margen" y "Costo Manual". Resulta interesante notar que los campos "Costo" y "Margen" tienen una relación inversamente proporcional de 0.7, lo que naturalmente indica que a mayores costos hay menor margen y viceversa.

Ventas en el tiempo según grupo

Es de esperar que las agrupaciones y sub-agrupaciones de ventas resulten ser una forma natural de aglomerar conversiones de ventas con características similares. Es por ello que se analizó el comportamiento de los grupos más importantes con su evolución en el tiempo.

Debido a que la cantidad de grupos es extensa y heterogénea, fue necesario filtrar los datos y solo utilizar los grupos de ventas más significativos. El criterio utilizado fue el seleccionar grupos con más de 10.000 ocurrencias de ventas individuales. Acto seguido

se visualizaron las ocurrencias de ventas con colores correspondientes a estos grupos principales, según puede verse en la imagen “Dispersión de Ventas”.



Dispersión de Ventas

De este gráfico podemos obtener información valiosa sobre el comportamiento de las ventas en los últimos años. En general, se puede observar el efecto del comienzo de la pandemia al final del primer trimestre y comienzo del segundo trimestre de 2020. Seguido de una tendencia a la baja que termina a principio del tercer trimestre, donde se revierte la tendencia. Dicho crecimiento se mantiene hasta la actualidad.

Las ventas en la base del gráfico que tienen un mayor volumen, pero menor monto de venta, corresponden a "INSUMOS HP". El volumen de ventas de este grupo ha permanecido estable desde el inicio del registro en 2010. Estos productos, aunque de menor costo, constituyen una buena base en el flujo de caja, que ha probado ser estable durante más de una década.

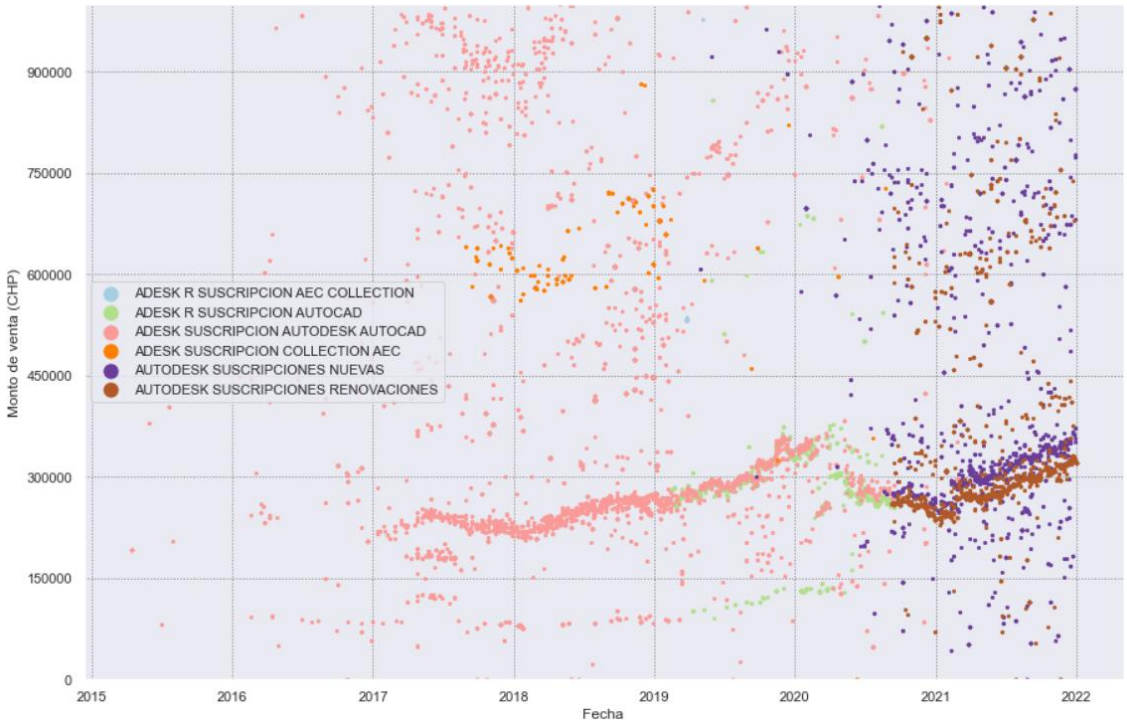
Otro grupo que llama la atención es la evolución de ventas en suscripciones Autodesk. Con color naranja se puede observar la concentración de ventas en años más recientes. Estas ventas en suscripciones son una línea fuerte en el esquema de negocio de la empresa. Lo que puede inferirse por la densidad de los puntos en el gráfico y su ubicación promedio en $300.000 (CLP).

Por último, es interesante notar que el grupo "ADOBE", en color azul claro, parece estar dividido en tres curvas evolutivas de ventas: la primera a una altura de 150.000 CLP, la segunda a 300.000 CLP y la tercera en un franco crecimiento, aunque menos densa,

llegando a montos de 900.000 CLP en la actualidad. Esto está de acuerdo con tres líneas de clientes para la marca ADOBE: 1) Línea Comercial, 2) Línea Educacional y 3) Línea Gobierno.

Ventas de Autodesk

Uno de los mayores Partners de la empresa en Autodesk, con quien se efectúan en alianza la mayor parte de la facturación anual. En vista de lo estratégico de esta alianza se estudió por separado la dispersión de ventas de este grupo. Y se utilizaron subgrupos para la leyenda gráfica. Esto puede verse en el gráfico "*Dispersión de Ventas Autodesk"*.



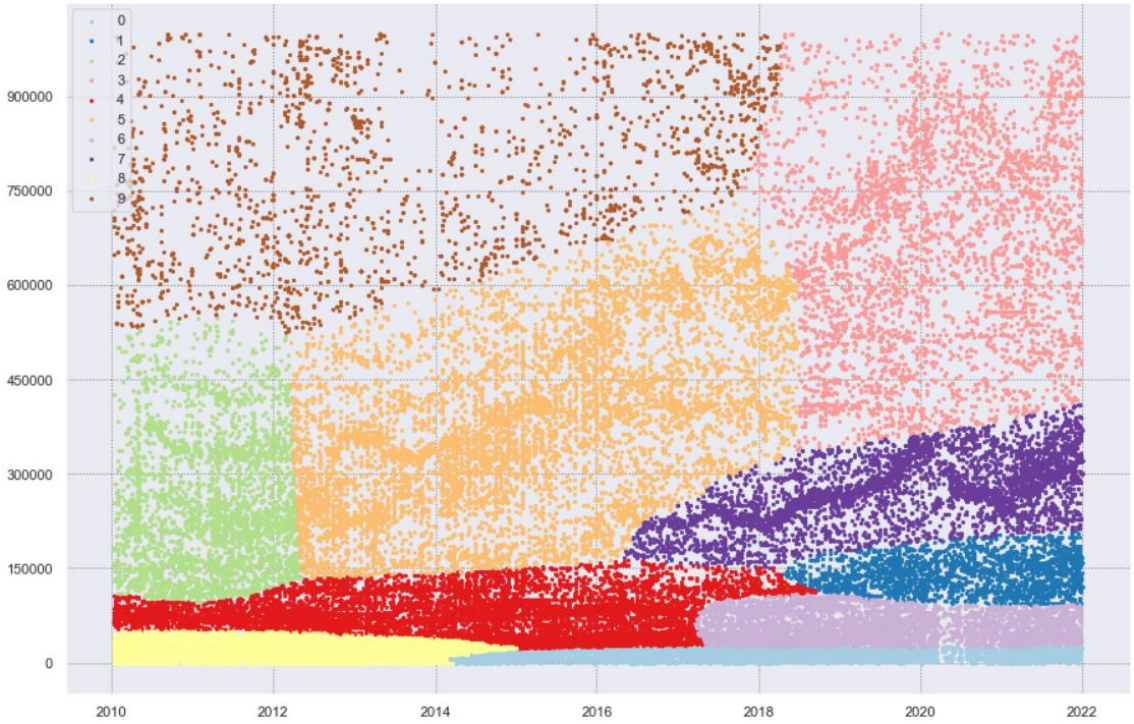
Luego de la parálisis temporal de ventas en el segundo trimestre de 2020 debido a la pandemia, las ventas por suscripciones de Autodesk tienden a la baja. Sin embargo, hay un claro cambio de tendencia desde inicios de 2021. Lo anterior está de acuerdo con la reactivación del mercado luego del primer año de pandemia. También se pueden ver dos líneas paralelas que diferencian las suscripciones de las renovaciones. Esto no se ve previo al evento de pandemia, donde los nuevos clientes y los clientes recurrentes parecen estar unidos en la misma línea de tendencia.

Clústeres de Ventas según Densidad

En los gráficos anteriores podemos observar que hay agrupaciones de ventas que tienden a ser más densas. Estas agrupaciones tienen tendencias al alta y a la baja. Parece ser una buena idea agrupar estos grupos de ventas por densidad. Este tipo de agrupación o "clustering" permitiría definir líneas de negocio en crecimiento y tomar decisiones más acertadas para potenciarlas, dividirlas o anularlas en función de su comportamiento.

Se empezó por acotar el dataset. Utilizar todo el dataset generaría demasiado ruido para el algoritmo de clustering. El algoritmo a utilizar es un modelo de Mezcla Gaussiana Bayesiana. Debido a que este algoritmo utiliza una función de densidad probabilística multivariable, resultó oportuno mantener zonas donde las ventas parecen ser más densas. Esto ocurre para montos inferiores a $1.000.000. También se descartaron valores atípicos negativos. Pues el enfoque está puesto en los montos de ventas reales. A fin de facilitar la agrupación fue necesario normalizar los datos de entrada. Si estos datos no son escalados la distancia entre los puntos producido por diferencia de escalas entre las características pudiera generar agrupaciones espúreas.

El resultado de este proceso de clustering se presenta en la gráfica *"Dispersión de Ventas con Clusters"*.



Dispersión de Ventas con Clusters

Hay cinco grupos de ventas que llegan a la actualidad:

- Grupo A - 1 (rosado)
- Grupo B - 2 (morado)
- Grupo C - 5 (azul oscuro)
- Grupo D - 0 (violeta)
- Grupo E - 4 (azul claro)

Con este clústering se puede filtrar los clientes de cada uno de los grupos y crear estrategias de marketing y ventas ajustadas a sus necesidades particulares.

Datos de Variables Macroeconómicas

Hasta ahora se han analizado datos internos a la empresa, provenientes del CRM. Ahora se procede a visualizar datos de variables macroeconómicas relacionadas al mercado de la empresa y ver su relación con los patrones de venta internos.

En primer lugar, se visualizarán las características básicas del conjunto de datos. Estos datos externos fueron tomados desde la página oficial del Banco Central de Chile. El DataFrame consta de las siguientes columnas:

- **USDCLP:** Cambio del peso chileno frente al dólar.
- **IPC**: Índice de Precios al Consumidor: es un indicador económico elaborado y publicado por el Instituto Nacional de Estadística (INE) de Chile, que mide mes a mes la variación conjunta de los precios de una canasta de bienes y servicios representativa del consumo de los hogares del país.
- **IMACEC**: Índice Mensual de Actividad Económica. Es una estimación que resume la actividad de los distintos sectores de la economía en un determinado mes, comparado con precios del año anterior; su variación interanual constituye una aproximación de la evolución del PIB. El cálculo del IMACEC se basa en múltiples indicadores de oferta que son ponderados por la participación de las actividades económicas dentro del PIB en el año anterior.
- **PIB**: Producto Interno Bruto. Indicador económico que refleja el valor monetario de todos los bienes y servicios finales producidos por un territorio en un determinado período de tiempo. Se usa para medir la riqueza que genera un país.
- **TCM**: Tipo de Cambio Multilateral. Es un indicador que mide el precio relativo de bienes y servicios de una economía con respecto a los de un grupo de países con los cuales se realizan transacciones comerciales.
- **TDN**: Tasa de Desocupación Nacional. Tasa de desempleo.
- **UF**: Unidad de Fomento. Su finalidad original era la revalorización de los ahorros de acuerdo con las variaciones de la inflación, permitiendo que el dinero ahorrado en bancos y cajas mantuvieran su poder adquisitivo.

A fin de estudiar la factibilidad de explicar el comportamiento de las ventas internas en función de variables macroeconómicas externas resulta útil crear una matriz de correlación entre los campos anteriores y las ventas directas dentro de la organización. Esta matriz se presenta en el gráfico “Matriz de Correlación de Variables Macroeconómicas”.

Hay una alta relación entre las ventas mensuales y las variables macroeconómicas. Sin embargo, hay que tener presente que el valor mensual de cada variable solo se tiene luego que dicho mes termina. Por lo tanto, es necesario hacer un desfase con el mes anterior para verificar su probabilidad de uso como valores predictivos.

Matriz de Correlación de Variables Macroeconómicas

ETAPA 4: INGENIERÍA DE CARACTERÍSTICAS Y MODELOS

Definición

La ingeniería de características se refiere al proceso de usar el conocimiento de dominio para seleccionar y transformar variables más relevantes a partir de datos sin procesar, creando así un modelo predictivo utilizando aprendizaje automático o modelado estadístico. El objetivo de la ingeniería y selección de características es mejorar el rendimiento de los algoritmos de aprendizaje automático (ML).

Se hizo uso de diversas técnicas para extraer características predictivas que mejoren el rendimiento de los algoritmos a proponer:

- **Creación de características**: La creación de características implica identificar las variables que serán más útiles en el modelo predictivo. Este es un proceso subjetivo que requiere intervención humana y creatividad. Las características existentes se mezclan a través de la suma, la resta, la multiplicación y la proporción para crear nuevas características derivadas que tienen un mayor poder predictivo.
- **Transformaciones**: La transformación implica manipular las variables predictoras para mejorar el rendimiento del modelo; por ejemplo, garantizar que el modelo sea flexible en la variedad de datos que puede ingerir; garantizar que las variables estén en la misma escala, haciendo que el modelo sea más fácil de entender; mejorar la precisión; y evitar errores computacionales asegurándose de que todas las características estén dentro de un rango aceptable para el modelo.

- **Extracción de características**: La extracción de características es la creación automática de nuevas variables a partir de datos sin procesar. El propósito de este paso es reducir automáticamente el volumen de datos en un conjunto más manejable para el modelado. Algunos métodos de extracción de características incluyen análisis de clústeres, análisis de texto, algoritmos de detección de bordes y análisis de componentes principales.
- **Selección de características**: Los algoritmos de selección de características esencialmente analizan, juzgan y clasifican varias características para determinar qué características son irrelevantes y deben eliminarse, qué características son redundantes y deben eliminarse, y qué características son más útiles para el modelo y deben priorizarse.

Objetivo

Lo que se busca es predecir las ventas de períodos futuros. Para ello se cuenta con los datos históricos de ventas (datos internos) y los datos históricos de variables macroeconómicas del mercado (datos externos). Este objetivo quedó finalmente automatizado y/o registrado en un cuaderno de Jupyter llamado *"03_feature_model_engineering.ipnb"*.

Creación de características

Ventas en un marco temporal común

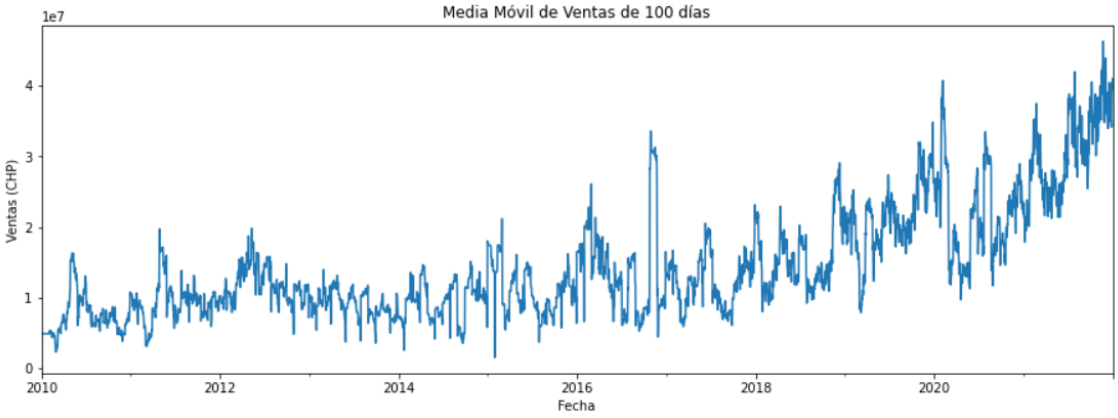
A fin de predecir ventas en función de otras variables fue necesario unificar las dos fuentes en un único DataFrame usando como base un mismo marco temporal. A este respecto se tomó cada instancia de venta y se unificaron por día. De esta manera fue posible crear un nuevo campo ventas en el dataframe de variables macroeconómicas representativo de ventas con resolución diaria.

Medias móviles

Debido a que no todos los días se producen eventos de ventas, se comenzó por crear medias móviles de distintos rangos temporales a fin de facilitar el entrenamiento. De introducir valores de venta diaria sin procesar, debido a que en su mayoría se tratará de valores iguales a cero, esto podría causar un sesgo en el proceso de entrenamiento. Al utilizar medias móviles en lugar de datos en bruto, se minimiza los períodos de tiempo sin actividad visible para las características de entrada.

Como las medias móviles tendrán valores null antes de completar la cantidad mínima de su ventana de promedio, se rellenaron los valores null con el siguiente valor distinto de null.

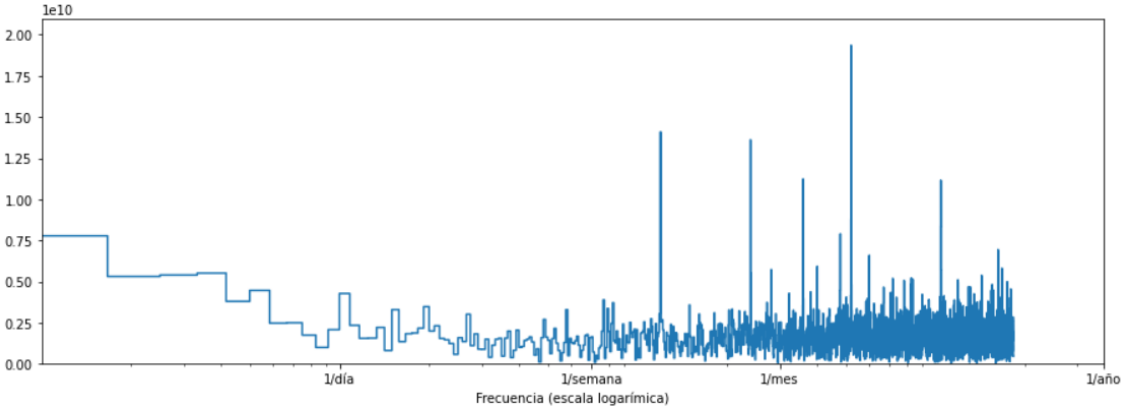
Se crearon medias móviles para los siguientes períodos en días: 5, 10, 15, 20, 30, 50, 100. Y dichas medias se ejecutaron sobre el campo "sell".



Media Móvil de Ventas de 100 días

Transformada Rápida de Fourier

Al no contar con información previa sobre las frecuencias más importantes es necesario identificarlas con alguna herramienta de procesado de señales. En este caso se optó por extraerlas haciendo uso de la transformada de Fourier. Al aplicar "Fast Fourier Transformation" sobre la serie de interés podemos observar picos en las frecuencias más significativas. Dichas frecuencias pueden ser utilizadas para crear características de periodicidad.

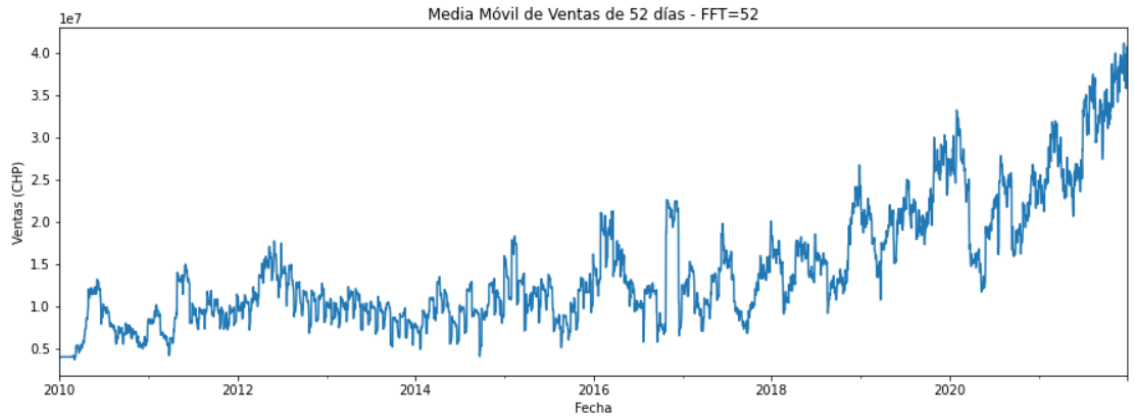


Gráfica de Valores Resultantes en la Transformada Rápida de Fourier

Haciendo uso de los valores resultantes de la aplicación de la Transformada Rápida de Fourier se pasó a seleccionar las frecuencias con mayor participación y evitando valores de frecuencia cero. Este algoritmo de selección de programó y ejecutó, dando como resultado las siguientes frecuencias:

1. 52.0836
2. 11.8334
3. 29.7501
4. 25.6669
5. 103.83

Con esas frecuencias se crearon nuevas medias móviles.



Media Móvil de 52 días (correspondiente a la frecuencia más participativa de la TRF)

Subdivisión de datos

De los datos disponibles 70% se utilizó para entrenamiento, 20% para validación y 10% para prueba. Además, no se tomaron valores de forma aleatoria. Esto es por dos razones:

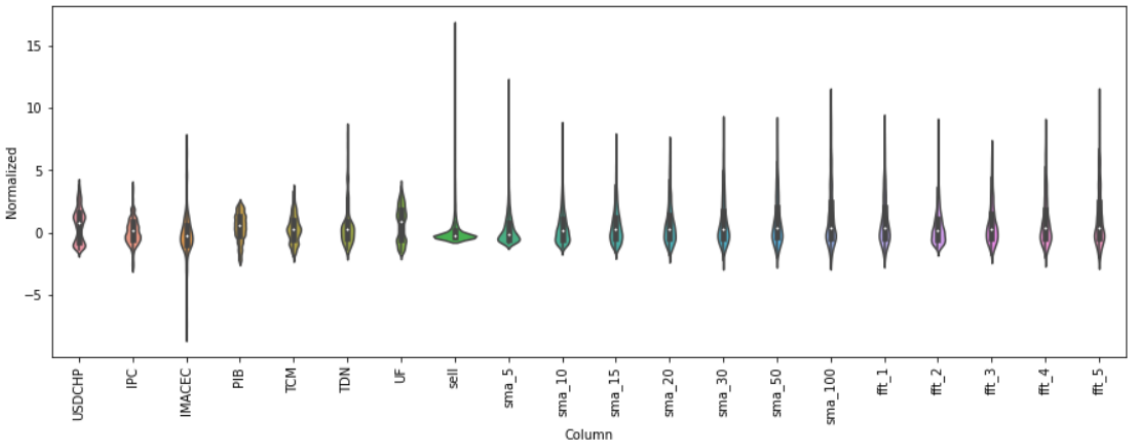
1. De esta forma se asegura la toma de datos en ventanas de ejemplares consecutivos.
2. Asegura que los resultados de validación y prueba sean más realistas al ser evaluados en los datos recolectados luego de que el modelo fue entrenado.

Normalización de datos

Es importante escalar las características antes de entrenar una red neuronal. La normalización es una forma común de hacer esta escala: restar la media y dividir por la desviación estándar de cada característica.

La media y la desviación estándar solo deben calcularse utilizando los datos de entrenamiento para que los modelos no tengan acceso a los valores en los conjuntos de validación y prueba.

También es discutible que el modelo no debería tener acceso a valores futuros en el conjunto de entrenamiento durante el entrenamiento, y que esta normalización debería hacerse usando promedios móviles. Sin embargo, en aras de la simplicidad, se usó un promedio simple.



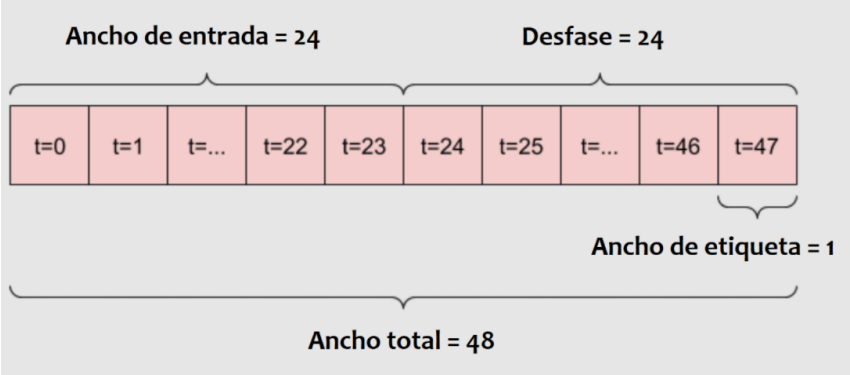
Columnas de dataframe con valores normalizados

Ventana de datos

Los modelos basados en redes neuronales realizan sus predicciones basados en una ventana de muestras consecutivas de los datos.

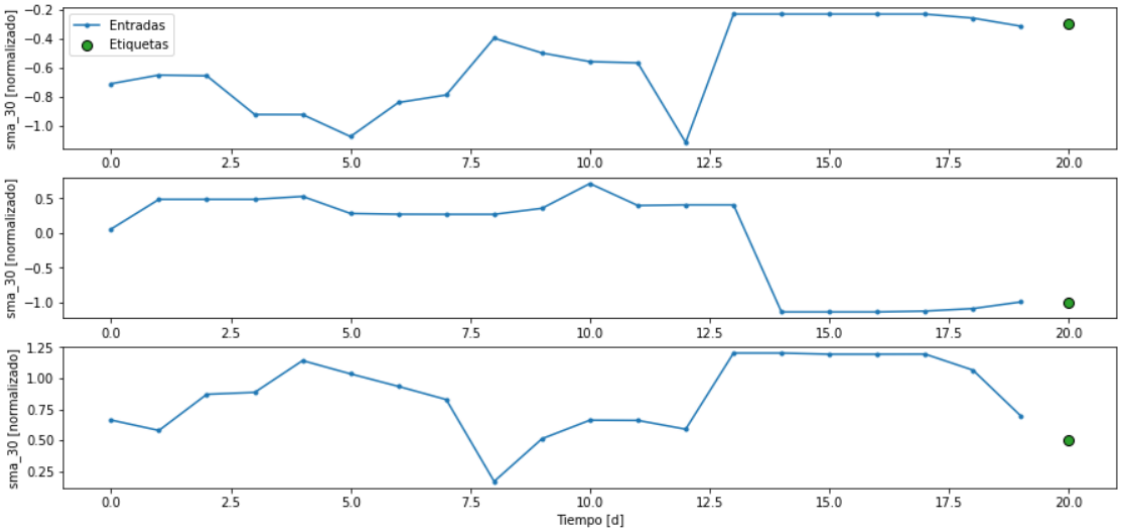
Las características principales de las ventanas de entrada son:

- El ancho (número de pasos de tiempo) de las ventanas de entrada y etiqueta.
- El tiempo desfasado entre ellos
- Qué características se utilizan como entradas y como etiqueta.



Configuración de ventana de datos (tomado y traducido de la página oficial de TensorFlow para series temporales)

En la siguiente imagen se muestra un ejemplo de ventanas de datos aplicando sobre la característica de media móvil de 30 días. La configuración de estas ventanas es como sigue: ancho total = 20 días, ancho de entrada = días, desfase = 1 día, ancho de etiqueta = 1 día.



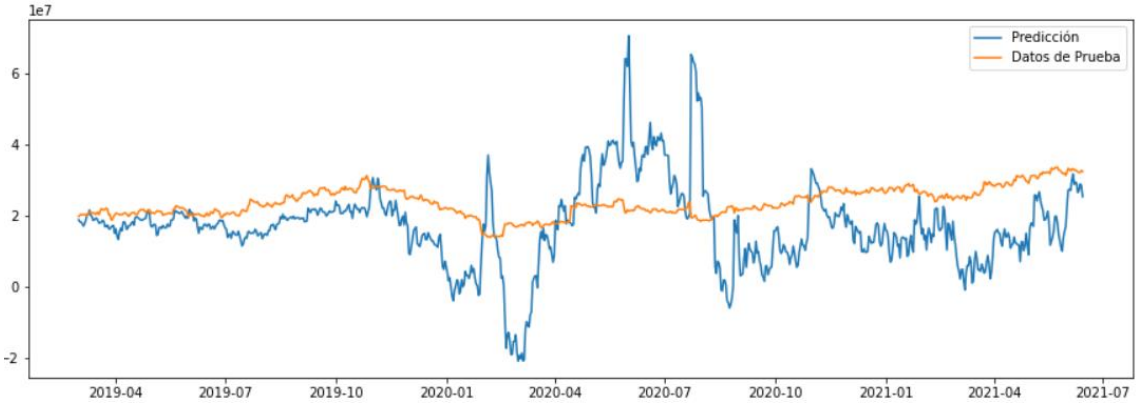
Ventanas de datos para la media móvil de 30 días: Ventana de 20 días, Ancho de Entrada de 19 días, Desfase de 1 un día, Ancho de etiqueta de un día.

Modelos predictivos simples

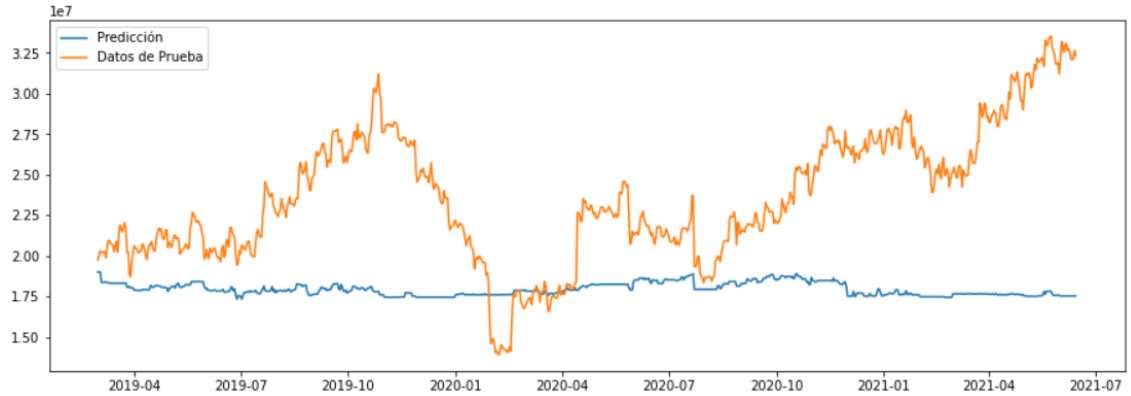
A fin de contar con una línea base de comparación para modelos más complejos se crearon modelos predictivos relativamente simples haciendo uso de regresiones de los siguientes tipos:

1. Regresión lineal
2. Regresión polinómica
3. Regresión de Bosques Aleatorios

Desempeño de regresión lineal



Desempeño de regresión polinómica



Desempeño de regresión con bosques aleatorios

Era de esperar que estos modelos no ofrecieran resultados aceptables. Pero sirvieron como punto de comparación para otros modelos de mejor desempeño.

Modelos predictivos basados en ventanas

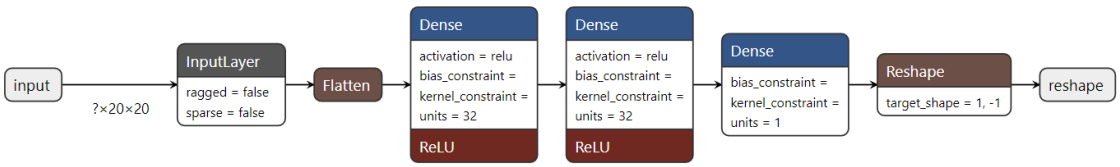
Los modelos anteriores no hacían uso de una ventana de datos subsecuentes al momento de entrenar el modelo, así que solo podían tomar en cuenta los valores actuales de las características. Esto parecer ser una pérdida de información en cierto sentido, pues es de esperar que los valores actuales guarden cierta relación, solo con el valor más reciente, sino también con valores previos. A este respecto resulta ventajoso hacer uso de algoritmos que puedan utilizar variables de entrada de una dimensión mayor a un simple vector. Es en este punto donde entran los modelos predictivos basados en ventanas. A fines de la presente investigación estos modelos constan de redes neuronales de diversos tipos. Desde redes neuronales profundas, pasando por redes neuronales recurrentes y terminando con redes neuronales convolucionales.

Red Neuronal Densa con un solo paso de salida

Los modelos más simples que se pueden construir son aquellos que predigan una única característica en el futuro, basado únicamente en las condiciones actuales. Esto implicaría una ventana de entrada = 1, y una ventana de etiqueta = 1. Sin embargo, se comenzó por crear un modelo de red neuronal densa con 20 pasos de ancho de ventana y un solo paso de salida o etiqueta.

La configuración de este modelo es como sigue:

1. Entrada de la forma (20, 20)
2. “Flatten” para convertir en vector la matriz de entrada.
3. Capa densa de 32 neuronas, activación relu.
4. Capa densa de 32 neuronas, activación relu.
5. Capa densa de 1 neurona
6. Cambio de forma para una sola variable de salida.



Arquitectura de Red Neuronal Densa con un solo paso de salida

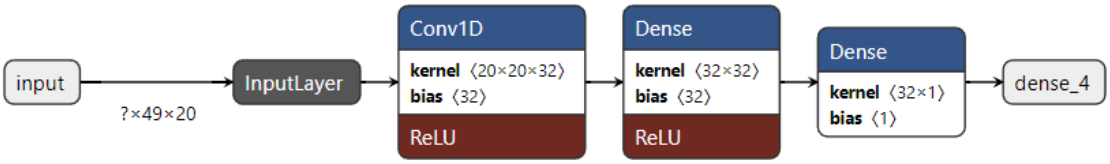
Red Neuronal Convolucional con un solo paso de salida

Las redes neuronales convolucionales nos permiten entrenar algoritmos que no los extraigan patrones a fuerza bruta sobre todo el conjunto de entrada, sino sobre porciones específicas dentro del mismo conjunto. Esto permitiría que se den pesos distintos a diversas porciones de la ventana de datos. Lo que resulta adecuado si pensamos que no todos los valores deberían

recibir la misma ponderación. Probablemente los valores más actuales deberían tener un mayor efecto en la predicción mientras que los valores más antiguos de la ventana un efecto menor.

La arquitectura de esta red es como sigue:

1. Convolución unidimensional. 32 filtros, activación relu.
2. Capa densa de 32 neuronas, activación relu.
3. Cada densa de 1 neurona



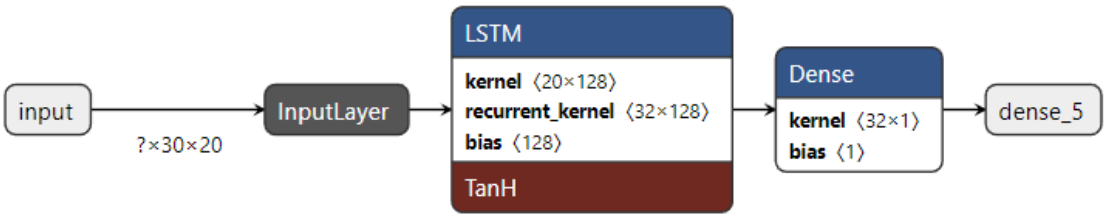
Arquitectura de red convolucional con un solo de salida

Red Neuronal Recurrente con un solo paso de salida

La característica principal de las redes recurrentes es que la información puede persistir introduciendo bucles en el diagrama de la red, por lo que, básicamente, pueden «recordar» estados previos y utilizar esta información para decidir cuál será el siguiente. Esta característica las hace muy adecuadas para manejar series cronológicas. Mientras las redes recurrentes estándar pueden modelar dependencias a corto plazo (es decir, relaciones cercanas en la serie cronológica), las LSTM pueden aprender dependencias largas, por lo que se podría decir que tienen una «memoria» a más largo plazo.

La arquitectura utilizada en esta red fue la siguiente:

1. Una capa LTSM de 32 unidades con retorno de secuencias
2. Una capa densa de 1 neurona



Arquitectura de Red Neuronal Recurrente LSTM con un solo paso de salida

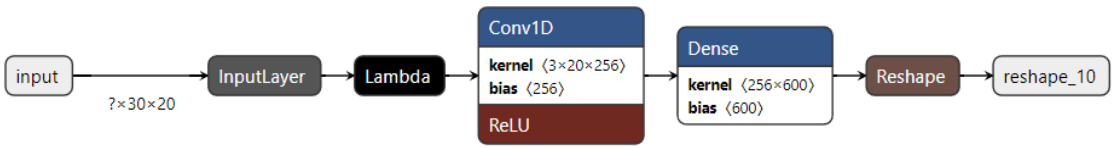
Red Neuronal Convolucional con múltiples pasos de salida

Este último tipo de red neuronal probada no hace solo una predicción de un solo paso a futuro, sino de múltiples pasos a futuro. De esta manera se pueden predecir períodos de tiempo superiores al marco temporal de un día.

La arquitectura de esta red es como sigue:

1. Capa Lambda con función de toma solo una parte del input
2. Convolución unidimensional. 256 filtros, activación relu.
3. Capa densa de neuronas = PASOS DE PREDICCIÓN x NÚM. CARACTERÍSTICAS, activación relu.

4. Reshape con la forma de salida igual la cantidad de pasos múltiples



Arquitectura de Red Neuronal Convolucional con múltiples pasos de salida

ETAPA 5: PREPARACIÓN DE DATOS PARA INTELIGENCIA DE NEGOCIO

Como parte de la etapa final de este proyecto se creó un tablero de Power BI para hacer semiento de las métricas e indicadores más importantes para el modelo de negocio. Para lograr extraer conocimiento realmente útil es necesario hacer cierto tratamiento a los datos disponibles. Si bien es cierto que este tratamiento puede llevarse a cabo directamente, hasta cierto punto, dentro de la herramienta de Business Intelligence, se optó por crear un cuaderno en Jupyter (*"04_data_preparation_for_bi.ipynb"*), específico para dicho tratamiento.

Esto tiene como ventaja que se sigue el mismo ambiente para el procesamiento de datos utilizados para la extracción, transformación y carga de datos del modelo predictivo.

Identificación de clientes 80/20

Tal como su nombre lo indica, los clientes 80/20 son aquellos clientes responsable por el 80% de las ventas (aproximadamente) y que suelen representar el 20% de la lista total de clientes (aproximadamente). No todos los años cada cliente compra la misma cantidad. Por ello se utiliza un histórico de los últimos 3 años para identificar a los clientes pertenecientes a esta categoría.

Con una función en Python, registrada en el cuaderno en archivo *"04_data_preparation_for_bi.ipynb"* se agrupó las ventas por cliente y se sumó el valor de ventas asociados a los mismos. Luego, se ordenó en orden descendente para iterar sobre los clientes hasta que la suma acumulada de ventas correspondiera al 80% del total. Finalmente, se agregó una columna booleana en el DataFrame llamada "80/20".

Renovaciones

Otro identificador útil para la presentación de los resultados es uno que permita determinar si la venta corresponde a una primera venta o una renovación. Para crear este identificador se utilizó el campo de descripción de la venta y se buscaron coincidencias de las siguientes cadenas de texto: "Renew", "RENOVA" y "RENEW". Finalmente, se agregó una nueva columna llamada "New/Renew" de tipo string que contendrá "Renew" o "New" según sea el caso.

Línea de producto

Siguiendo un procedimiento parecido al de renovaciones, se creó también una nueva columna llamada "Línea", donde se identificó la línea de producto en función del tipo de grupo o el valor de descripción.

Finalmente, el dataframe con las nuevas columnas creadas para la visualización se guardó en un archivo CSV que luego será cargado en Power BI.

ETAPA 6: CREACIÓN DE TABLERO DE BI PARA EQUIPO DE VENTAS

En esta última etapa se creó un tablero de inteligencia de negocios que facilitara la gestión del rendimiento por parte del equipo de ventas de la empresa. Esto quedó registrado en el archivo *"Presentación Cierre.pbix"*. Este tablero fue creado según las indicaciones del Gerente de Customer Success quién tenía la necesidad de contar un informe dinámico para el cierre de cada período. Los requerimientos recibidos fueron:

1. Visualizar las ventas en años anteriores y poder comparar las ventas de distintos años.
2. Visualizar las ventas adjudicadas por vendedor en los últimos 3 tres años que facilita la comparación de desempeño entre cada vendedor y diferenciando los montos por renovaciones de suscripción.
3. Visualizar la facturación en los últimos 3 años por parte de los consultores (llamados "V2") por cursos, consultorías o ventas de productos de software.
4. Poder identificar los clientes más importantes en función de la facturación asociada y hacerlo diferenciando según su pertenencia al grupo "80/20".
5. Una tabla con filtros dinámicos para la revisión del histórico de ventas.

EVALUACIÓN

Básicamente se utilizaron dos métodos de evaluación para medir el desempeño de los algoritmos utilizados en la predicción de valores futuros.

- **Métricas para la cuantificación de error**: este primer criterio se siguió al utilizar fórmulas de medición de error bien reconocidas modelos estadísticos de regresión, que el problema que nos ocupa.
- **Curvas de pérdida**: aplicado únicamente en el caso de los modelos basados en redes neuronales.
- **Ajuste del punto o curva de predicción a los resultados esperados**: este segundo criterio es un poco más subjetivo, pues depende en mayor medida de la interpretación humana, tales como la ocurrencia de eventos con saltos importantes de venta o el acercamiento superior o inferior de las predicciones a las etiquetas. Decidimos incluirlo pues hay algunos factores que se observan mejor gráficamente.

Métricas para la cuantificación de error

Error Cuadrático Medio

El error cuadrático medio es una medida de calidad de un predictor. Si un vector de *n* predicciones es generado desde una muestra de n puntos en todas las variables, donde Yes el vector e valores observados de la variable a predecir, y $\hat{Y}$ es el vector de valores predichos, entonces el error cuadrático medio viene dado por:

$$ECM = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Una característica del error cuadrático medio es que reduce la complejidad al tratar con signos negativos. Para minimizar el ECM el modelo debe ser más preciso, lo cuál significa que debe estar más cerca de los datos verdaderos.

Esta métrica de error tiene la característica de que usa la misma escala de los datos medidos, razón por la que se le conoce como una medida de exactitud dependiente de la escala. Por lo tanto no puede ser utilizado para comparar series de valores con diferentes escalas.

Error Absoluto Medio

El Error Absoluto Medio es otra medida común de error de predicción con regresiones. Siguiendo la misma notación utilizada en el error cuadrático, el error absoluto medio puede definirse como:

$$EAM = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_i - Y_i|$$

Y al igual que la métrica de error cuadrático medio esta es una medida de exactitud dependiente de la escala.

Coeficiente de Determinación

El coeficiente de determinación, denotado comúnmente como R^2, puede verse como la porción de variación de la variable a estimar que es predecible a partir de las variables de entrada. Es una métrica de puntuación muy utilizada para evaluar modelos estadísticos de predicción. Siguiendo la misma notación utilizada para la definición de las métricas anteriores, el Coeficiente de Determinación se define como:

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{(\hat{Y}_i - \underline{Y})^2}$$

Donde la $\underline{Y}$ representa la media de los valores de la etiqueta, o valores reales.

Los valores de R^2para una regresión son una medida de qué tan bien las predicciones se aproximan a los puntos de datos reales. Un R^2 de 1 indica que las predicciones se ajustan perfectamente a los datos. Y cuanto R^2 está fuera del rango entre [0, 1] esto indica que el modelo se ajusta a los datos peor que el peor predictor de mínimos cuadrados posibles. Es decir, es peor que una simple regresión lineal.

Curvas de pérdida

Por curvas de pérdida nos referimos a la graficación de la métrica escogida como pérdida, en este caso el error cuadrático medio, para el proceso de entrenamiento de las redes neuronales. El gráfico de evolución de las curvas de pérdidas en los conjuntos de validación y prueba nos permite identificar sobreajuste del modelo a los datos de entrenamiento o divergencia de la solución.

Ajuste del punto o curva de predicción a los resultados esperados

Para cada modelo también se realizó una inspección visual de las predicciones obtenidas. Para ello se crearon gráficos que sobre la serie temporal mostraron los valores de las predicciones y las etiquetas de forma simultánea. De esta manera se pudo observar mejor los puntos específicos de mayor divergencia o convergencia.

RESULTADOS

Resultados asociados al algoritmo de predicción

Referente al objetivo primario de crear un algoritmo predictivo de ventas, se utilizan las métricas de calificación definidas en el apartado anterior para comparar los diversos tipos de algoritmos utilizados. Con dichas métricas se podrá escoger aquel con mejor rendimiento que luego será utilizado en la etapa de despliegue en producción (esta etapa no está contemplada como alcance del presente estudio).

Los resultados muestran que los modelos que mejor se acercan a los valores reales son aquellos creados con la premisa de utilizar una ventana de datos y donde igualmente se aplicó tecnología de redes neuronales. Los resultados de las tres métricas principales de evaluación se muestran en los siguientes gráficos. Sin embargo, es de notar que los modelos de regresión lineal, regresión polinómica y bosques aleatorios no se incluyen en las gráficas de Error Cuadrático Medio y Error Absoluto Medio, sino solo en la gráfica de Coeficiente de determinación. La razón es que, como se mencionó anteriormente, las primeras dos métricas son dependientes de la escala y como el error fue muy alto, en comparación, para los modelos más simples, la escala del gráfico no permitía apreciar los valores en los modelos basados en ventanas de datos.

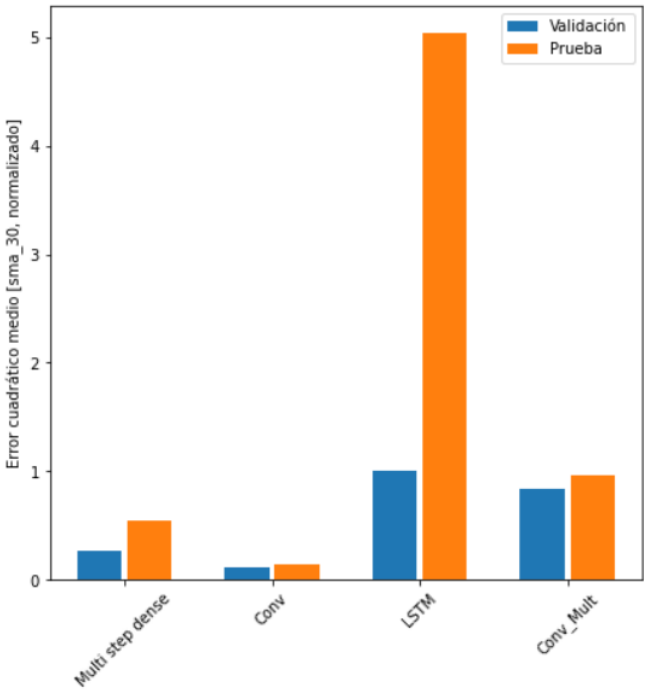


Gráfico comparativo de ECM para conjuntos de prueba y validación

En el gráfico anterior se puede apreciar que el modelo con menor error es el creado con una red convolucional que predice un único paso temporal a futuro, representado en este caso por un día. El modelo basado en una red neuronal recurrente de tipo LSTM tiene el peor desempeño en el conjunto de prueba. Sin embargo, es importante resaltar que el modelo basado en una red neuronal convolucional con predicción de múltiples períodos realmente tiene un buen desempeño. Pues hay que tomar en cuenta que este es el único modelo donde no realiza una predicción de un solo paso. Sino de 7 pasos a futuro. Es decir, este modelo predice una semana

completa de ventas a futuro. Y sin embargo, se comporta mejor que el modelo LSTM para un solo día de predicción a futuro.

Como es de esperarse, hay un mejor rendimiento cuando se trata de predicciones a un solo día a futuro. Pero con el presente rendimiento, parece adecuado incluso para producción el uso del algoritmo predictivo para una semana completa.

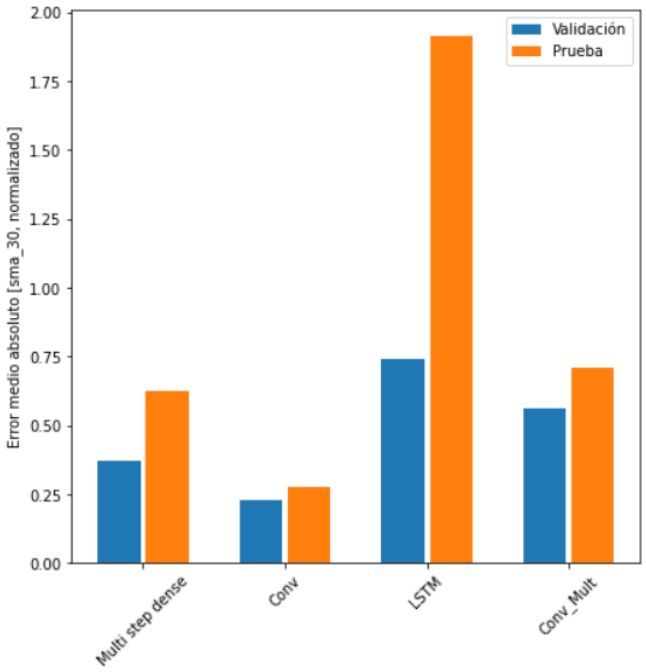


Gráfico comparativo de EAM para conjuntos de prueba y validación

La métrica de Error Absoluto Medio nos ofrece una interpretación muy similar a la del Error Cuadrático Medio. El modelo de Red Convolucional de un solo paso es el que cuenta con mejor desempeño.

Para el caso del coeficiente de determinación si fue posible mostrar el desempeño de todos los algoritmos. Como es de esperarse, las regresiones realizadas con regresión lineal, regresión polinómica y bosques aleatorios fueron las que estuvieron más lejos del valor óptimo, que para esta métrica es 1, tanto en los conjuntos de validación como en los de prueba.

Sin embargo, resulta interesante notar que, para el caso del coeficiente de determinación, es el modelo convolucional de regresión multipaso el que mejor desempeño ofrece en el conjunto de prueba. Lo que nos indica que el modelo que mejor provecho está sacando de las características de entrada para explicar el patrón subyacente en los datos para la predicción es precisamente el de la red neuronal de predicción múltiple. Decimos esto porque el coeficiente de determinación puede verse como la porción de variación de la variable a estimar que es predecible a partir de las variables de entrada.

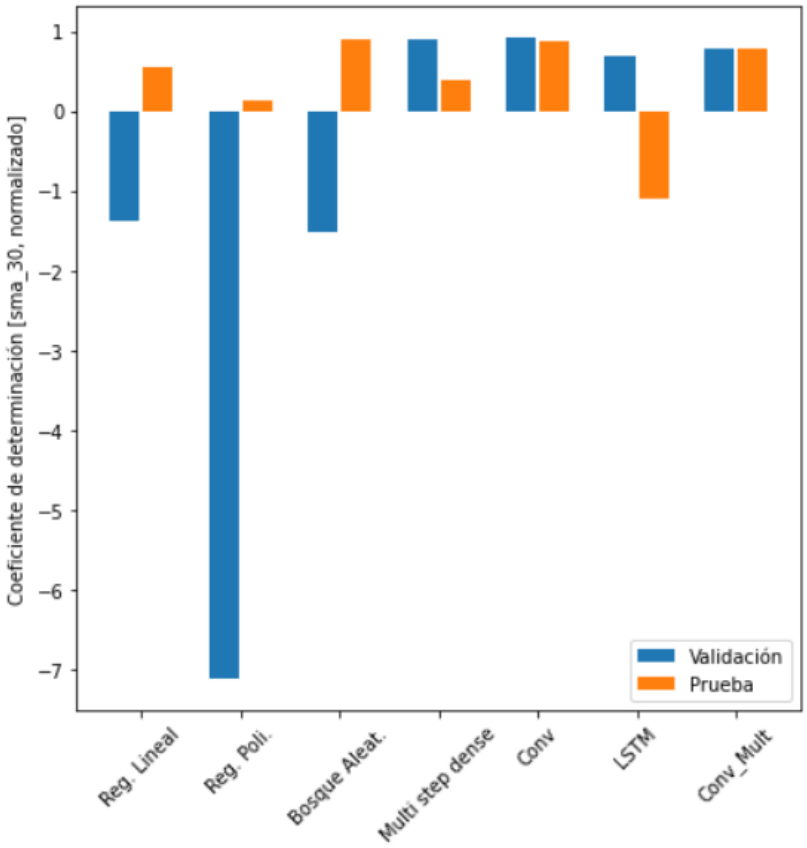
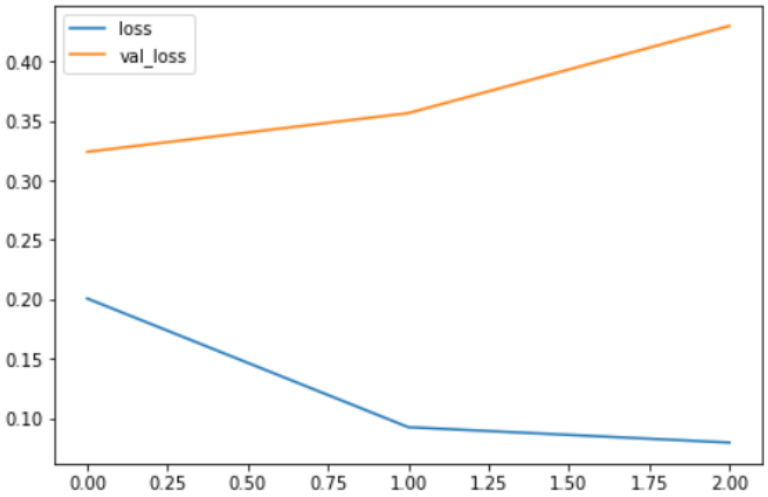
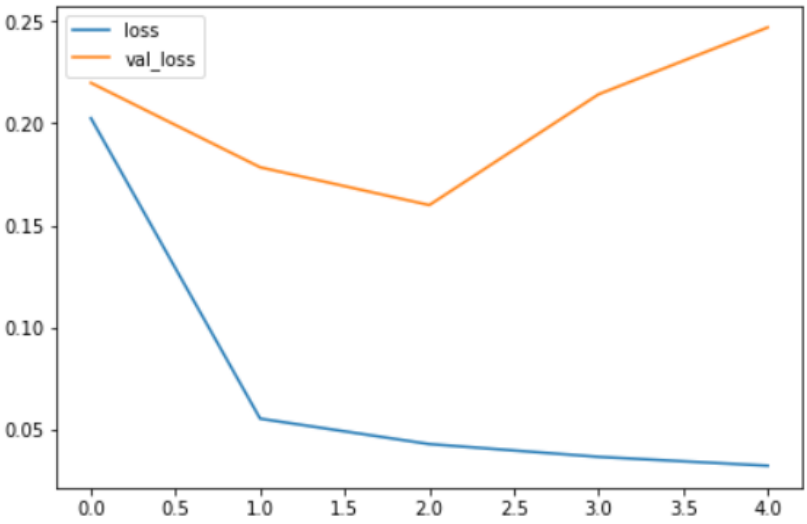


Gráfico comparativo de R^2 para conjuntos de prueba y validación en todos los modelos

Como se mencionó en el apartado de evaluación, las curvas de los entrenamientos para el caso de las redes neuronales también pueden ofrecernos información valiosa del producto obtenido. El entrenamiento del modelo denso termina divergiendo, pero como es de esperarse tiene un mejor rendimiento en entrenamiento que validación. Y lo mismo ocurre en el modelo convolucional de un paso.

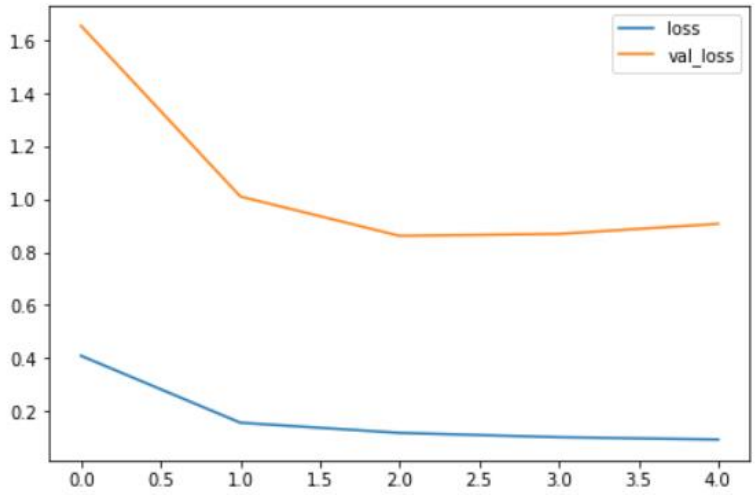


Entrenamiento de Modelo Denso

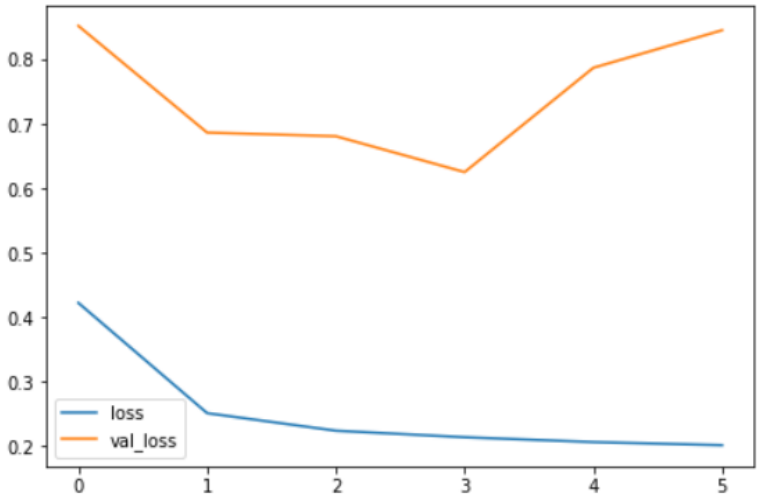


Entrenamiento de Modelo Convolucional de un paso

Se debe aclarar que para el entrenamiento de las redes neuronales se utilizó la opción de stop temprano en la biblioteca de Keras. Esta opción permite detener el proceso de entrenamiento cuando en cierta cantidad de épocas consecutivas no hay una disminución del valor de pérdida. Esta cantidad de épocas consecutivas se entiende como un parámetro de paciencia y en este caso fue configurado a un valor de 2. Esto puede verse en la gráfica de "*Entrenamiento de Modelo Convolucional de un paso*" y en la gráfica "*Entrenamiento de Modelo Recurrente LSTM de un paso*". y así también se explica por qué hay distinta cantidad de épocas para cada algoritmo.

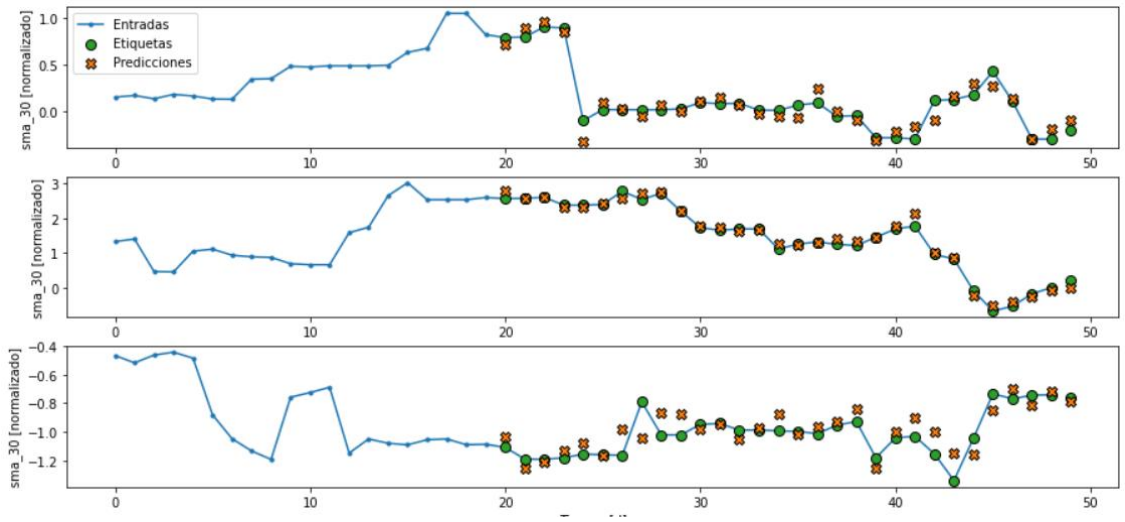


Entrenamiento de Modelo Recurrente LSTM de un paso



Entrenamiento de Modelo Convolucional de un paso

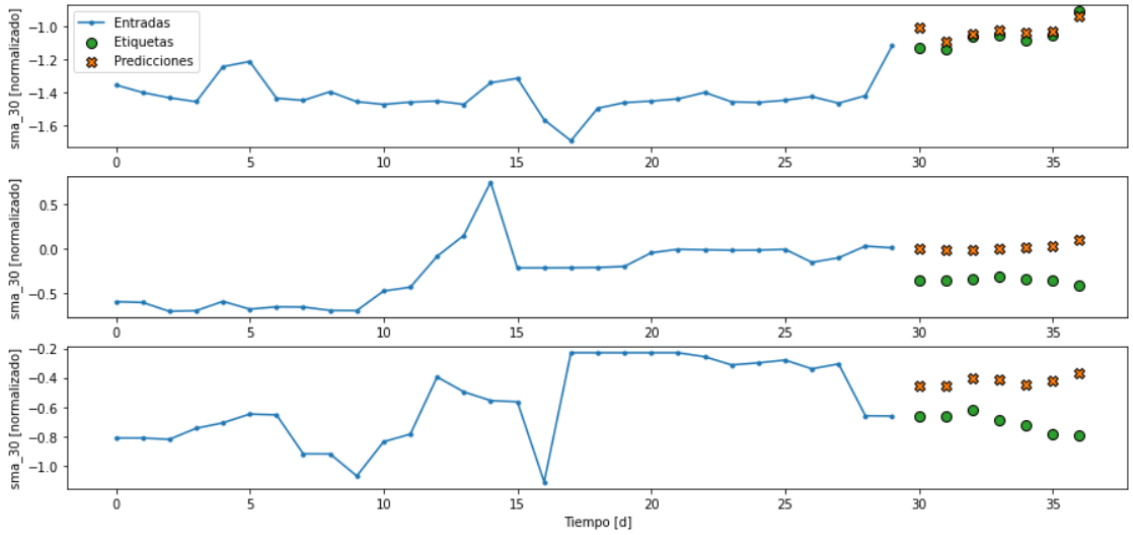
Finalmente se presenta el desempeño de forma gráfica de los algoritmos. En este caso nos quedamos únicamente con los modelos basados en redes convolucionales.



Predicciones hechas con modelo convolucional de un paso

En la gráfica anterior se observa como, siguiendo un paso temporal de un día, este algoritmo se apega bastante bien a los valores reales. Se muestran tres corridas del mismo algoritmo. En la primera corrida se puede apreciar que el modelo es capaz de detectar una caída abrupta en las ventas. Este es un comportamiento deseable y en particular útil cuando hay eventos que pueden alterar el patrón de ventas.

En la siguiente gráfica observamos el desempeño del modelo convolucional de múltiples pasos, con tres pruebas individuales. En la primera observamos un comportamiento bastante apegado a los valores reales. Sin embargo, para la segunda y tercera pruebas se evidencia un desfase de hasta el 50% del valor predicho.



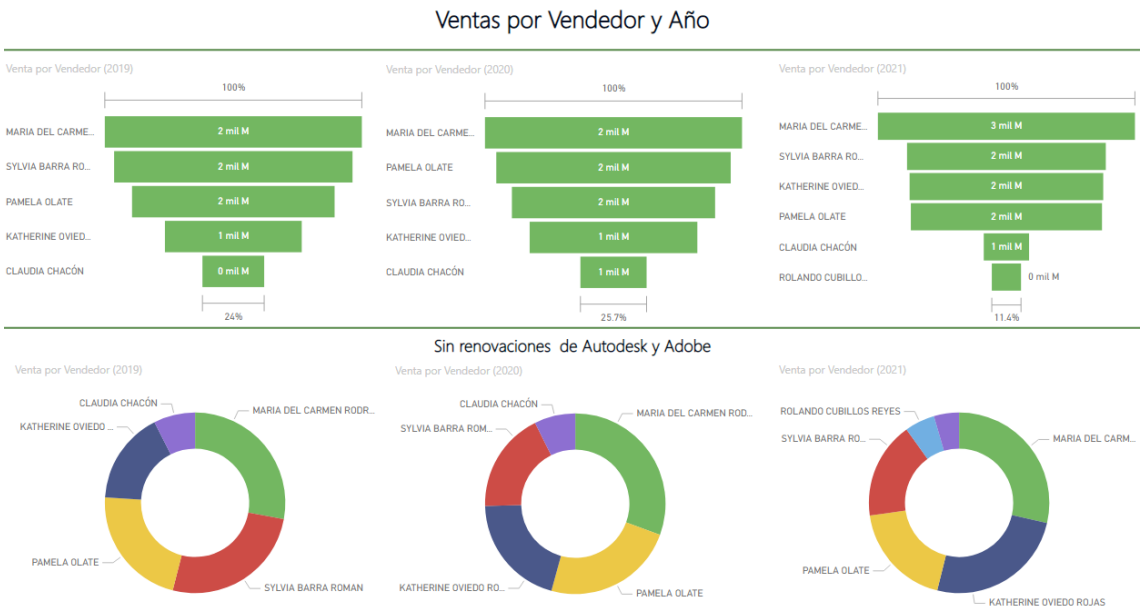
Predicciones hechas con modelo convolucional multipaso

Resultados del tablero de BI

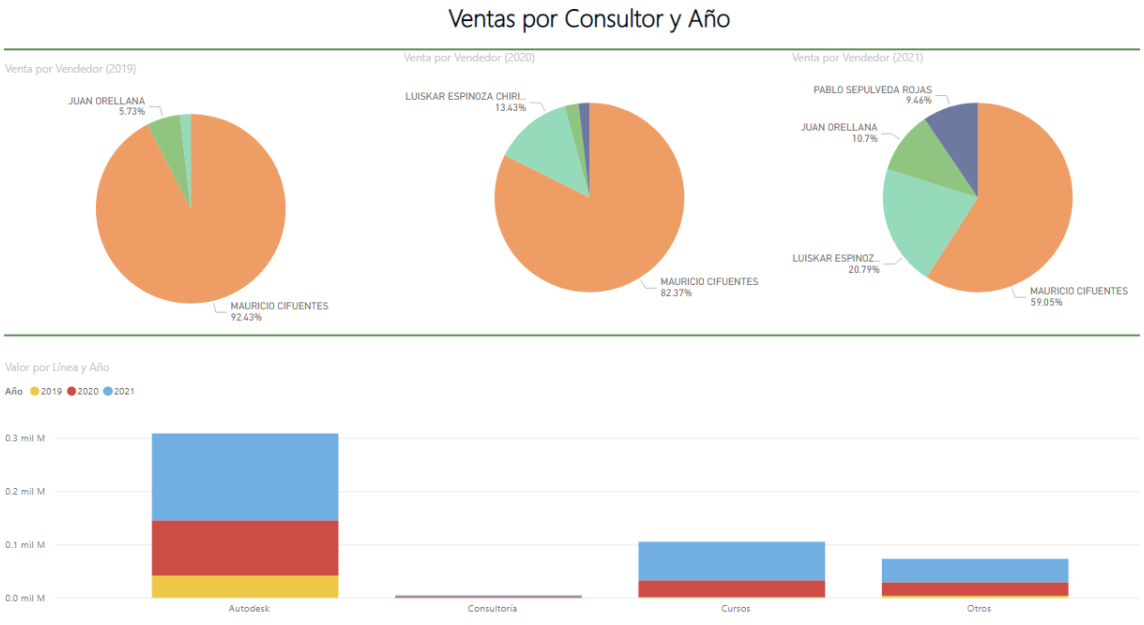
A este respecto se solicitó información al encargado de presentar el informe de cierre de venta por año. Quien admitió que con el tablero creado de forma dinámica resulta mucho más fácil hacer la presentación a los gerentes correspondientes. También ha resultado útil en reuniones de ventas semanales, donde se va actualizando el acumulado en tiempo real desde los valores obtenidos desde el LMS. Eso sí, aún es necesario ejecutar manualmente el cuaderno de Jupyter para el procesado de los datos en bruto una vez por semana para actualizar la información.

Las imágenes a continuación presentan la interfaz de cada página de la aplicación de visualización creada con Power BI.

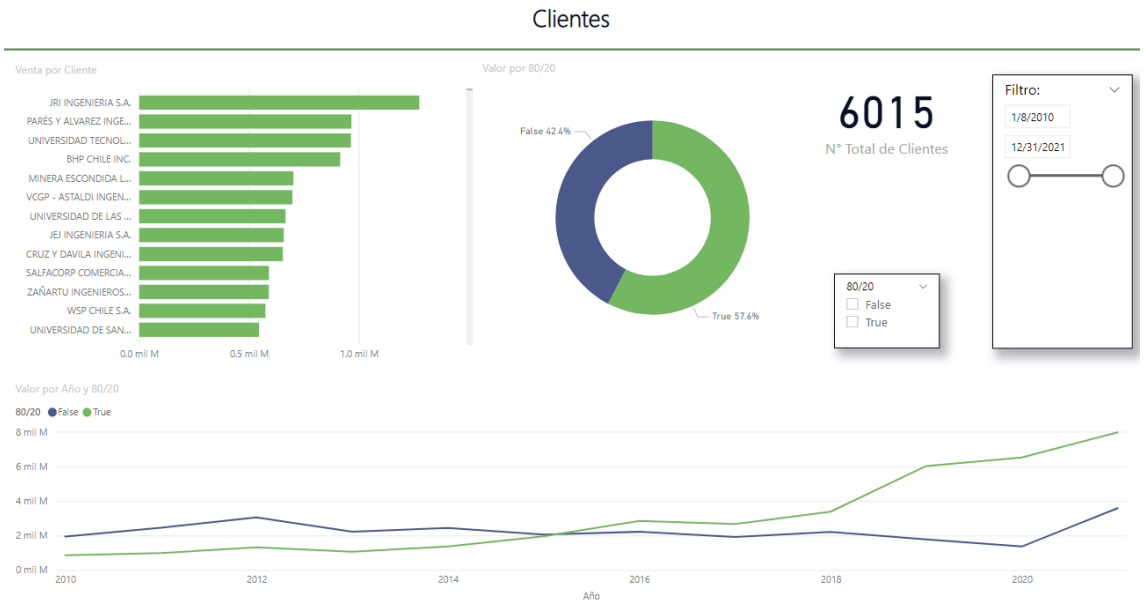
Dashboard principal de ventas por año



Página de ventas por vendedor



Página de ventas por consultor



Página de inspección de clientes

Página con tabla de dinámica de registros

CONCLUSIONES Y TRABAJOS FUTUROS

Los resultados obtenidos muestran que la predicción de ventas futuras es totalmente viable dentro de la organización en cuestión (COMGRAP). Un seguimiento diario rindió mejores resultados. Sin embargo, a los fines propuestos resulta insuficiente, pues la idea es contar con algoritmo predictivo que de soporte a la asignación de metas de ventas. Y estas asignaciones se hacen semanal o mensualmente.

El algoritmo de predicción múltiple es prometedor a este respecto, pues sí que facilita al menos predicción de ventas semanales. Sin embargo, su comportamiento es aún errático y debe ser mejorado antes de pasar a la etapa de despliegue y producción.

Algunas acciones futuras que podrían mejorar el desempeño son:

1. **Probar diversas arquitecturas**: el mundo de las redes neuronales es vasto y hay una gran cantidad de arquitecturas de redes neuronales profundas que pueden probarse para tratar de mejorar el rendimiento. Son de especial utilidad artículos de investigación en repositorios públicos donde hay una gran cantidad de propuestas de arquitectura para problemas de regresión. Muchos incluso con códigos o seudo-códigos ya desarrollados.
2. **Hacer una optimización de hiper-parámetros**: luego de escoger una mejor arquitectura se pueden utilizar métodos de optimización de hiper-parámetros, tales como búsqueda de rejilla, para conseguir las mejores configuraciones de ratas de aprendizaje, número de épocas, funciones de pérdida, entre otros.
3. **Utilizar métodos de ensamblaje**: se puede mejorar el rendimiento al combinar predicciones de múltiples modelos. Hay que recordar que las redes neuronales profundas son métodos no lineales. Lo cual ofrece una mayor flexibilidad ante grandes cantidades de datos. Sin embargo, como el proceso de aprendizaje es estocástico esto quiere decir que resultan sensibles a determinados conjuntos de entrenamiento. Lo que conduce a una varianza alta. Un buen acercamiento para reducir la varianza es entrenar múltiples modelos y combinarlos en los llamados ensamblajes.

Finalmente, algunas recomendaciones para trabajos futuros serían:

A. **Re-escalar los valores de predicción**: debido a que los datos fueron escalados para facilitar el proceso de entrenamiento, será necesario reescalar las predicciones para volverlas operativas.
B. **Despliegue en producción**: a fin de que el algoritmo predictivo sea realmente útil para la empresa, es necesario desplegarlo finalmente en un ambiente de producción donde se puedan enviar solicitudes de predicción y obtener de vuelta predicciones. Esto típicamente implica el desarrollo de una API que permita el enlace con otras soluciones empresariales.
C. **Promocionar el uso de tableros de inteligencia de negocio en otros departamentos**: si bien se obtuvieron buenos resultados para el tablero de informe de ventas. Se aconseja replicar la metodología a otros reportes y departamentos de la empresa para hacerlos más Data-Drive.