Universidad Autónoma Nacional de México

Facultad de Estudios Superiores Acatlán

Diplomado en Ciencia de Datos

4ª Generación

Efraín Ismael Flores Hernández

Volaris' Forecast of Revenue (VFR)

(Pronóstico del Ingreso de Volaris)

enero 2021

Índice

“The world is one big data problem”

Andrew McAfee

Introducción

¿Pronosticar venta para una aerolínea?

Parece una tarea imposible, es decir, ¿Cómo saber cuántos pasajeros van a comprar hoy un vuelo desde CDMX a Cancún para volar dentro de seis meses? Y si aún no parece complicado, ¿A qué precio pagarán? Para una aerolínea comercial, el inventario principal son los asientos disponibles para todos los vuelos: desde el momento analizado, hasta el último día que tenga publicado en sus canales de venta. Entonces, cada asiento puede asumirse como un producto con caducidad porque a partir del despegue, todos los asientos de ese vuelo ya no pueden venderse.

El sentido de negocio yace en incrementar la demanda para incrementar la oferta y consecuentemente (con el trabajo adecuado y continuo de todas las áreas de la empresa), mantener buenos índices de rentabilidad cada cierre de año. ¿Por qué no cada cierre de mes? Dada la naturaleza del negocio, existe una temporalidad muy clara en la demanda, generalmente definida por el periodo vacacional común: verano e invierno son temporadas altas para el turismo y, por ejemplo, febrero y agosto son meses conocidos por su bajo nivel turístico. Así que el reto para las aerolíneas se divide en dos:

1. Mantener niveles sanos de factor de ocupación, métrica que se define como la razón de los asientos ocupados entre los asientos totales
2. A través del dinamismo en los precios ofertados, aprovechar la demanda dispuesta a pagar más por un asiento, sobre todo en temporada alta

Entonces, tiene sentido que el pronóstico cuente con componentes como tendencia y estacionalidad, pero es necesario también agrupar los diferentes comportamientos entre las rutas, porque no es la misma anticipación de compra para unas vacaciones en Puerto Vallarta, que para una junta de negocios en Tijuana. En el siguiente trabajo, se intentará modelar la venta diaria de pasajeros e ingreso por ruta para cada mes de vuelo disponible desde el 01/01/2019 hasta el 19/09/2020 (por ahora, los registros se actualizarán hasta completar 2020).

Objetivo principal

Modelar un pronóstico de venta a nivel mes y segmento (la direccionalidad de una ruta, de ida o vuelta).

Objetivos secundarios

- Alta precisión en el pronóstico de pasajeros y ventas mensuales, tanto por compra de lugares como por productos adicionales (cuotas por prioridad de abordaje, asiento premium, perrito a bordo, etc.)
- Presentar una diferencia natural en el pronóstico para los meses de vuelo correspondientes a temporada alta, aunado a la diferencia de venta por la lejanía entre el mes de vuelo y el mes de venta
- Agrupar rutas por comportamiento de anticipación junto con el nivel de pasajeros e ingreso vendido, tomando en cuenta la capacidad (número de asientos disponibles)
- Pronosticar los pasajeros y el ingreso total esperado para cada mes de vuelo, esto servirá de apoyo para el establecimiento de metas a cierre de mes

Descripción de los datos

Capacidad

La capacidad de una aerolínea está definida por el número de asientos disponibles, la cantidad de vuelos no necesariamente es la mejor referencia dado que la misma aerolínea puede tener aviones de diferente tipo, con capacidad desde 144 hasta 230 asientos. Si bien el máximo número de pasajeros a bordo de un vuelo no pueden exceder el número de asientos, en muchas aerolíneas se practica la sobreventa, que significa vender (de manera adicional a la capacidad del avión) el porcentaje de asientos que se irán vacíos, estadísticamente hablando.

La cantidad de asientos también tiene variaciones según la demanda observada en cada ruta y mes de vuelo. Puede haber rutas con un vuelo redondo (frecuencia) a la semana con tipo de avión de 144 asientos y también puede haber otras rutas con nueve o diez frecuencias diarias con aviones de hasta 230 asientos.

Base y Ancillaries

En los objetivos secundarios, se habla sobre la venta por compra de lugares como por productos adicionales. El primero se conoce técnicamente como ingreso base, se refiere simplemente al pase de abordar, compras un sólo lugar en un sólo asiento para un sólo vuelo (los vuelos redondos o con conexión cuentan con ingreso base para cada tramo, es decir para cada segmento, porque en cada vuelo ocupas sólo un asiento y diferente).

Ahora, los productos adicionales o ancillaries que es como se conocen técnicamente, contemplan las cuotas con servicios-extra que el cliente decide (o no) adquirir. Entre las principales se encuentran: la maleta documentada, prioridad de abordaje, seguro de puntualidad, perrito a bordo, etc. Es importante tener claro tres aspectos de este tipo de ingreso:

1. No se puede adquirir un producto adicional sin antes adquirir un pase de abordar
2. El cliente puede elegir los productos adicionales que desee para el mismo vuelo

3. El momento de compra puede ocurrir desde la compra del pase de abordar, hasta el momento del despegue

Venta vs Ingreso

Existen dos dimensiones de tiempo importantes para una aerolínea, se trata de la fecha de venta y la fecha de vuelo. Interactúan porque la venta de pases de abordar y productos adicionales pueden ocurrir en cualquier momento desde la publicación del vuelo hasta la fecha de despegue. Explicado ello, la diferencia entre venta e ingreso radica en la dimensión analizada, la venta de hoy es para todos los meses de vuelo disponibles, pero si el análisis se enfoca por mes de vuelo, cada día de venta aporta al ingreso. Así, sólo en cada cierre de mes se conoce el ingreso total y final de dicho mes, porque ya no hay venta que siga aportando (porque no es posible vender para volar en el pasado).

Catálogo

En el objetivo principal se menciona que un segmento es tanto la ruta de ida como de vuelta, se explica con más detalle a continuación:

Cada aeropuerto tiene un código brindado por la IATA (asociación mundial que define estándares para la seguridad, eficiencia y sustentabilidad de la aviación). Dicho código consta de tres letras, por ejemplo:

- MEX para el Aeropuerto Internacional de la Ciudad de México (AICM)
- SJO para San José en Costa Rica, etc.

Entonces, el segmento de un vuelo está definido por la unión del código IATA origen y el de destino:

- DENGDL para el vuelo que sale de Denver hacia Guadalajara
- TLC-SJD-BJX para el vuelo que sale de Toluca y hace conexión en Los Cabos para llegar a León, etc.

Ahora, la ruta se refiere a la ida y vuelta del segmento, está definida como la unión por orden alfabético del código IATA de los dos aeropuertos:

- MTYSAT contempla los vuelos desde Monterrey a San Antonio, Texas y al revés
- TIJZIH contempla los vuelos de Tijuana a Ixtapa Zihuatanejo y al revés

Si ambos aeropuertos están en México el mercado es nacional, de otro modo es internacional. El mercado nacional se expresa como región MX y en el mercado internacional hay una división, si algún aeropuerto de la ruta toca Centroamérica, la región será CAM, de otro modo será US.

La unión de rutas con un aeropuerto principal en la misma región se denomina Hub, por ejemplo: en la región MX si la ruta toca Cancún o Cozumel entonces el Hub de esa ruta será CUN, si no pertenece a CUN y la ruta toca Tijuana o Mexicali entonces el Hub será TIJ, si no pertenece a CUN o TIJ y la ruta toca el ACIM o Toluca entonces el Hub será MEX y así sucesivamente. Entonces la ruta CUNTIJ pertenece al Hub CUN y la ruta MEXTIJ al Hub TIJ, por ejemplo.

Es importante mencionar que para la región US, un aeropuerto del segmento debe tocar territorio mexicano, ya sea el de origen o de destino. Por ley, una aerolínea mexicana no puede operar segmentos con origen y destino en otro país que no sea México, la nacionalidad de la compañía. En el caso de la región CAM, existe “otra aerolínea” (con nacionalidad costarricense) perteneciente a la principal (con nacionalidad mexicana), por ello, con el mismo nombre comercial pero diferente código de aerolínea (también brindado por la IATA y de sólo dos caracteres) se pueden volar segmentos con origen o destino Costa Rica. Lo importante es que Volaris tiene la posibilidad de volar SJOSAT (San José, Costa Rica a San Antonio, Texas) pero no LAXSAT (Los Ángeles a Texas), por ejemplo.

Ahora, la direccionalidad se refiere a la entrada o salida del segmento respecto al Hub, por ejemplo: el segmento GDLMEX pertenece al Hub MEX y como el aeropuerto MEX está posicionado al final del segmento significa que es el destino, está entrando al AICM por lo tanto su direccionalidad es “in”. Otro ejemplo para dejar clara la definición de direccionalidad: el segmento GDLMTY pertenece al Hub GDL y como el aeropuerto GDL está posicionado al inicio significa que es el origen, está saliendo de Guadalajara por lo tanto su direccionalidad es “out”.

Los Hub y rutas por mercado y región que alguna vez se han volado o se volarán por parte de la compañía son:

Tabla 1. Rutas y aeropuerto principal por Hub

Mercado	Región	Hub	Rutas	Aeropuerto principal
Nacional	MX	BJX	9	Léon, Guanajuato
		CUN	15	Cancún, Quintana Roo
		CUU	6	Chihuaha, Chihuaha
		GDL	22	Guadalajara, Jalisco
		MEX	32	Ciudad de México
		MID	3	Mérida, Yucatán
		MTY	8	Monterrey, Nuevo León
		TIJ	43	Tijuana, Baja California Norte
	MX Total		**138**	
Nacional Total			**138**	
Internacional	US	BAY	17	California (excepto LA)
		LAX	11	Los Ángeles
		MDW	14	Chicago
		O BIZ	10	Otros
		O LEI	13	Destinos turísticos de placer (Miami, Orlando, Denver, etc)
		O TX	10	Texas
	US Total		**75**	
	CAM	Intra CAM	3	Origen y destino en Centroamérica
		MX CAM	7	Desde México hacia Centroamérica y al revés
		US CAM	4	Desde USA hacia Centroamérica y al revés
	CAM Total		**14**	
Internacional Total			**89**	
TOTAL			**227**	

Fuente: Elaboración propia con datos de Volaris

Finalmente, en el catálogo también están denotadas las rutas competidas con los 3 competidores principales del país, para cada ruta y competidor, la ausencia de este último está definida con la letra “N”, su presencia con la letra “C” y alguna estrategia frontal con otros caracteres como “SBX”:

Tabla 2. Rutas competidas por Hub

Mercado	Región	Hub	Sin competidor	Frontal VB	Un competidor	Dos competidores	Todos compiten	TOTAL
Nacional	MX	BJX	7	1	1			9
		CUN	4		8	1	2	15
		CUU	6					6
		GDL	11	5	5	1		22
		MEX	4	3	8	7	10	32
		MID	3					3
		MTY	1	3	3	1		8
		TIJ	29	5	7		2	43
	MX Total		**65**	**17**	**32**	**10**	**14**	**138**
Nacional Total			**65**	**17**	**32**	**10**	**14**	**138**
Internacional	US	BAY	14		3			17
		LAX	8		1	1	1	11
		MDW	12			2		14
		O BIZ	10					10
		O LEI	8		2	2	1	13
		O TX	7		1	1	1	10
	US Total		**59**		**7**	**6**	**3**	**75**
	CAM	Intra CAM	3					3
		MX CAM	6			1		7
		US CAM	4					4
	CAM Total		**13**			**1**		**14**
Internacional Total			**72**		**7**	**7**	**3**	**89**
TOTAL			**137**	**17**	**39**	**17**	**17**	**227**

Fuente: Elaboración propia con datos de Volaris

Ingeniería de variables

Unidad muestral

Existen 4 tablas que serán conectadas:

- Capacidad diaria por segmento y mes de vuelo
- Venta base diaria por segmento y mes de vuelo
- Venta ancillaries diaria por segmento y mes de vuelo
- Catálogo por segmento

Entonces, la tabla principal será capacidad, porque es la oferta que tienes presente en determinada fecha, los registros de esta tabla serán mayores que los de la tabla de venta base y la tabla de ancillaries porque no necesariamente vendes al menos un pasajero para todos los segmentos y meses de vuelo, aunque sí cuentes con asientos disponibles. La última tabla servirá para categorizar los segmentos dadas las reglas de negocio explicadas en la sección anterior.

La tabla de capacidad despliega el mes y año de vuelo en el formato “MMM. YYYY” y la tabla de base despliega registros con el nombre de mes completo, se unifican formatos y columnas para obtener la llave: “fecha de venta, segmento, año de vuelo, mes de vuelo”. La tabla conectada tiene la siguiente estructura:

Tabla 3. Estructura tabla inicial

Columna	Variable	Tipo
Fecha venta	fecha	fecha
Segmento	texto	cat nom
Año vuelo	texto/num	cat ord/entero
Mes vuelo	texto/num	cat ord/entero
Ruta	texto	cat nom
Mercado	texto	cat nom
Hub	texto	cat nom
Direcc	texto	cat nom
VB	texto	cat nom
4O	texto	cat nom
AM	texto	cat nom
Vuelos	número	entero
Asientos	número	entero
Pax	número	entero
Base MXN	número	continuo
Base USD	número	continuo
Ancillaries MXN	número	continuo
Ancillaries USD	número	continuo

Fuente: Elaboración propia

Y una pequeña muestra sería:

Tabla 4. Muestra del conjunto de datos

Fecha venta	Segmento	Año vuelo	Mes vuelo	Ruta	Mercado	Hub	Direcc	VB	4O	AM	Vuelos	Asientos	Pax	Base MXN	Base USD	Ancillaries MXN	Ancillaries USD
01/01/2019	OAXLAX	7	2019	LAXOAX	US	LAX	In	N	N	N	12	2,171	2	$3,897	$176	$3,320	$150
17/04/2019	TIJTGZ	5	2020	TGZTIJ	MX	TIJ	Out	N	N	N	14	2,467	5	$1,986	$102	$2,427	$125
15/06/2020	TGZMEX	6	2021	MEXTGZ	MX	MEX	In	C	C	C	52	10,720	21	$6,813	$361	$5,804	$308
01/01/2020	DGOLAX	11	2020	DGOLAX	US	LAX	In	N	N	N	6	953	3	$1,046	$54	$966	$50
17/05/2019	TIJDGO	5	2019	DGOTIJ	MX	TIJ	Out	N	N	C	37	6,880	3	$1,803	$89	$1,877	$93

Fuente: Elaboración propia con datos de Volaris

Dado que el nivel de capacidad desplegado es segmento, la venta de pasajeros o ingreso base y ancillaries están sesgados por la cantidad de vuelos o, mejor dicho, asientos en dicho segmento. Por ello, se decide crear una columna por pax, base y ancillaries en MXN dividida entre los asientos del registro respectivo (pax se multiplica por mil para que la escala no sea tan pequeña). Estas variables se llamarán “Pax_p”, “Base_p” y “Anc_p”, respectivamente. Se considerará que estas sean las variables dependientes para el pronóstico.

Nuevas variables

Así como se menciona en la introducción, el producto principal de una aerolínea es cada uno de los asientos disponibles en venta, los asientos de los vuelos que han despegado ya no pueden venderse, por ello es recomendable no contemplar registros de capacidad con la fecha de captura mayor al año-mes de vuelo.

Para evaluar la anticipación de la venta, es necesario crear las variables año y mes de venta. Se procede a crear la variable que llamaremos “Month to Departure” (MtD) que explicará la distancia en meses entre el vuelo y la compra, estará definida como: 12*(año vuelo – año venta) +mes vuelo – mes venta. Por ejemplo: si la venta ocurre el 03/09/2019 para volar en febrero de 2020, el MtD es igual a 5.

Otra variable que es importante para evaluar la condición de la venta es la tarifa vendida, se creará la variable obteniendo la razón de Base MXN entre Pax. La tarifa de ancillaries no se considera necesaria porque a nivel registro puede existir venta en ancillaries, pero no venta de pax, esto por las razones explicadas en la descripción de datos: un cliente puede comprar tantas veces quiera cualquier producto adicional desde que compra su pase de abordar hasta la fecha de despegue. Además, por experiencia se conoce que la variación entre la tarifa base vendida y la de ancillaries es de +-10%.

Análisis exploratorio

Univariado

Valores ausentes numéricos

Como se menciona en la sección de ingeniería de variables, no porque se oferten asientos disponibles significa que, en todas las fechas para todos los segmentos para todos los meses de vuelo, hay venta. Por ello, los valores ausentes en Pax, Base_MXN, Anc_MXN en realidad son 0.

Valores ausentes categóricos

Los segmentos sin clasificación de ruta, Hub, etc. corresponden a vuelos privados o conexiones de código compartido (acuerdo de conectividad con la aerolínea norteamericana llamada Frontier). Sin embargo, la venta no se contabiliza para la compañía, por lo tanto, no se tomarán en cuenta para el modelo de pronóstico, se omiten los registros sin ruta, Hub, región, etc.

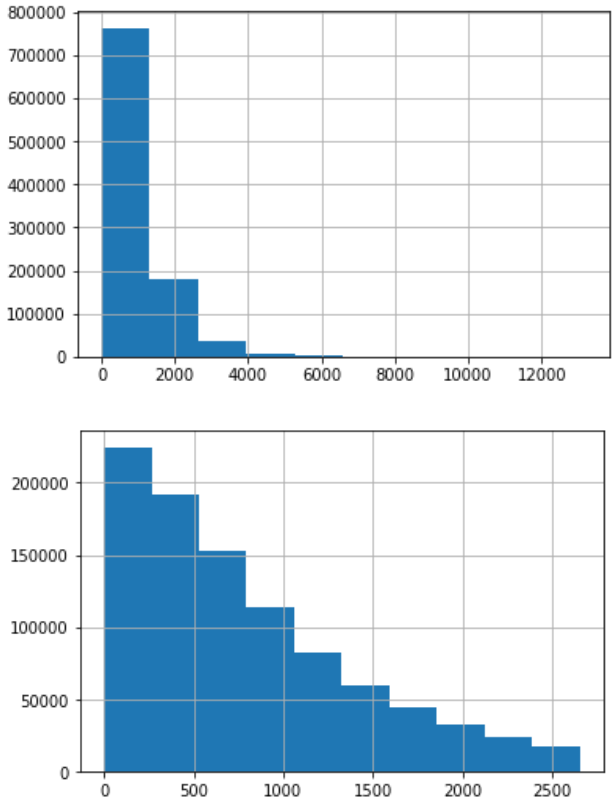
Outliers

Se omiten los outliers de la variable MtD porque no es común tener publicado más allá de 12 meses, se acota hasta 15 mediante el método de IQR.

Por parte de la capacidad, los valores máximos en asientos y vuelos no se consideran outliers porque son rutas con alta capacidad para temporadas altas, estos valores pueden repetirse.

En cambio, con la variable de tarifa base vendida. Es necesario acotar los outliers máximos, mediante el método IQR la distribución mejora:

Gráfica 1. Distribución tarifa base vendida con outliers y sin outliers



Fuente: Elaboración propia con datos de Volaris

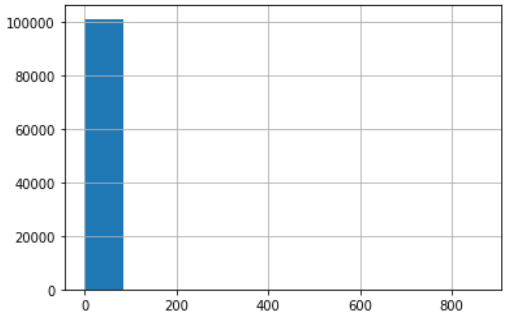
Bivariado

Outliers

Ahora, es importante saber que la distribución en la venta de pasajeros e ingreso depende de qué tan alejada esté la fecha de despegue, es decir, el MtD. Por ello, habrá outliers para cada MtD (donde sí hubo venta), mediante el método de IQR se omitirán los outliers.

Por ejemplo, para los registros donde efectivamente hay venta de pasajeros (Pax>0) y cuando la compra ocurre cinco meses antes del mes de despegue (MtD=5), la distribución de pasajeros unitarios (Pax_p) antes de quitar outliers tiene la siguiente forma:

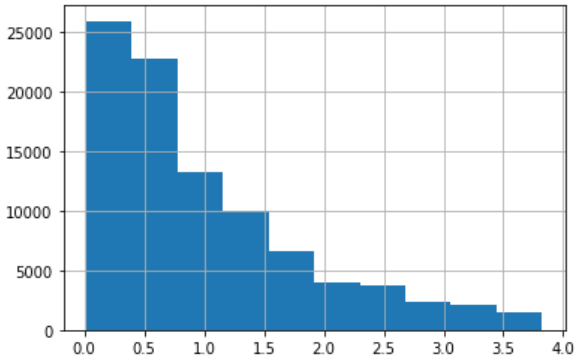
Gráfica 2. Distribución pax con outliers para MtD= 5



Fuente: Elaboración propia con datos de Volaris

En cambio, cuando se ejecuta el código sólo se omite el 3.5% del total de los registros originales y la distribución de los pasajeros unitarios para cada MtD mejora:

Gráfica 3. Distribución pax sin outliers para MtD= 5

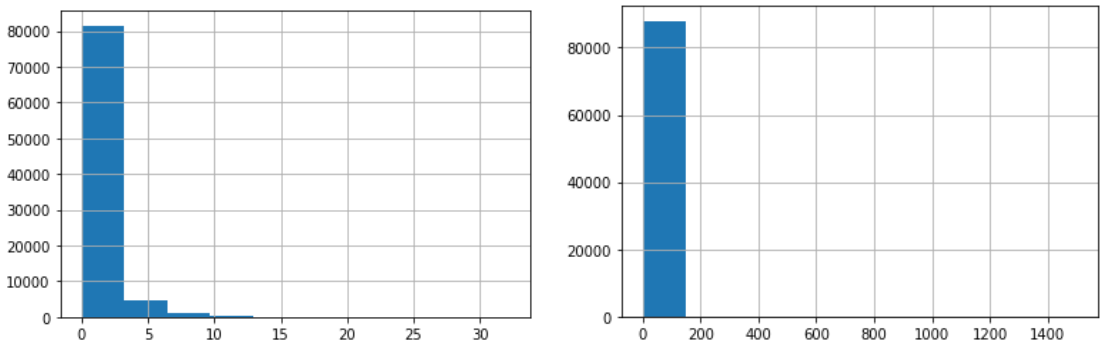


Fuente: Elaboración propia con datos de Volaris

Y aunque pareciera que en el panorama general (sin tomar en cuenta los MtD) sigue habiendo outliers máximos, estos se explican por rutas que venden en el propio MtD y ese mes de vuelo pertenece a la temporada alta, son comportamientos que se repiten por estacionalidad por lo que no se consideran outliers.

Lo que ocurre con la venta de pasajeros es un caso similar para la venta base y ancillaries, en ocasiones se cuenta con promociones fuertes que generan demanda inusualmente excesiva, pero depende del MtD. Continuando con el ejemplo del MTD=5, la distribución de ambos ingresos antes de remover outliers es:

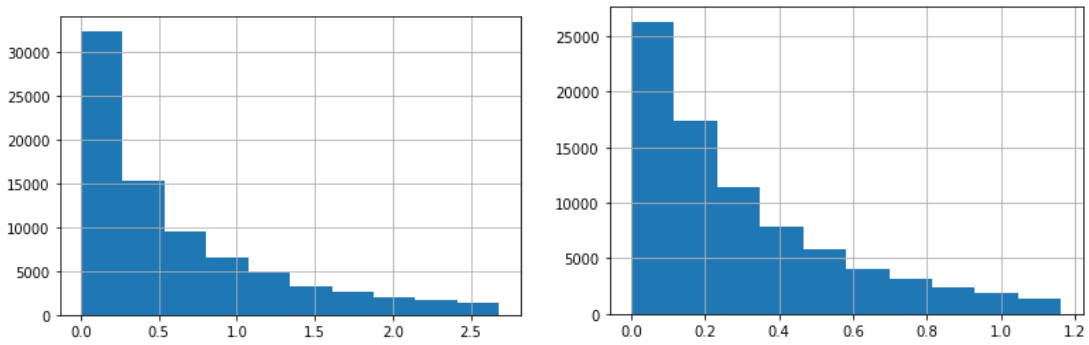
Gráfica 4. Distribución ingreso base y ancillaries con outliers para MtD= 5



Fuente: Elaboración propia con datos de Volaris

Y ahora, removiendo outliers con el mismo código de la imagen 1 pero con las variables correctas, sólo se omite el 6.7% de los registros anteriores (ahora -9.9% de los originales) y la distribución de ambos ingresos mejora considerablemente:

Gráfica 5. Distribución ingreso base y ancillaries sin outliers para MtD= 5



Fuente: Elaboración propia con datos de Volaris

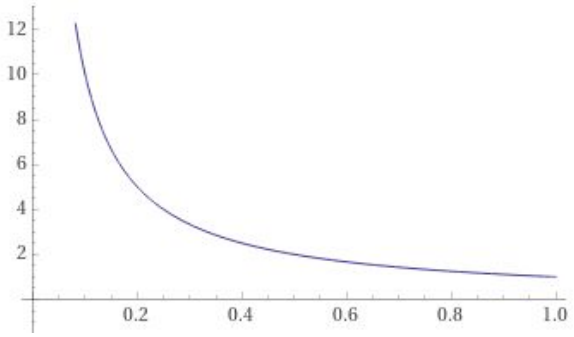
Al igual que con la venta de pasajeros, sin tomar en cuenta los MtD parece que existen outliers, pero se explican por la venta en el propio mes de vuelo y éste pertenece a temporada alta, son valores que por estacionalidad se repiten, no son outliers.

Curvas de maduración

¿Por qué hay venta mucho más alta en el propio mes de vuelo?

Para la mayoría de los vuelos, la demanda a lo largo de los días restantes para el despegue se comporta como la gráfica:

Gráfica 6. Gráfica $\frac{1}{x}$ con x [0,1]



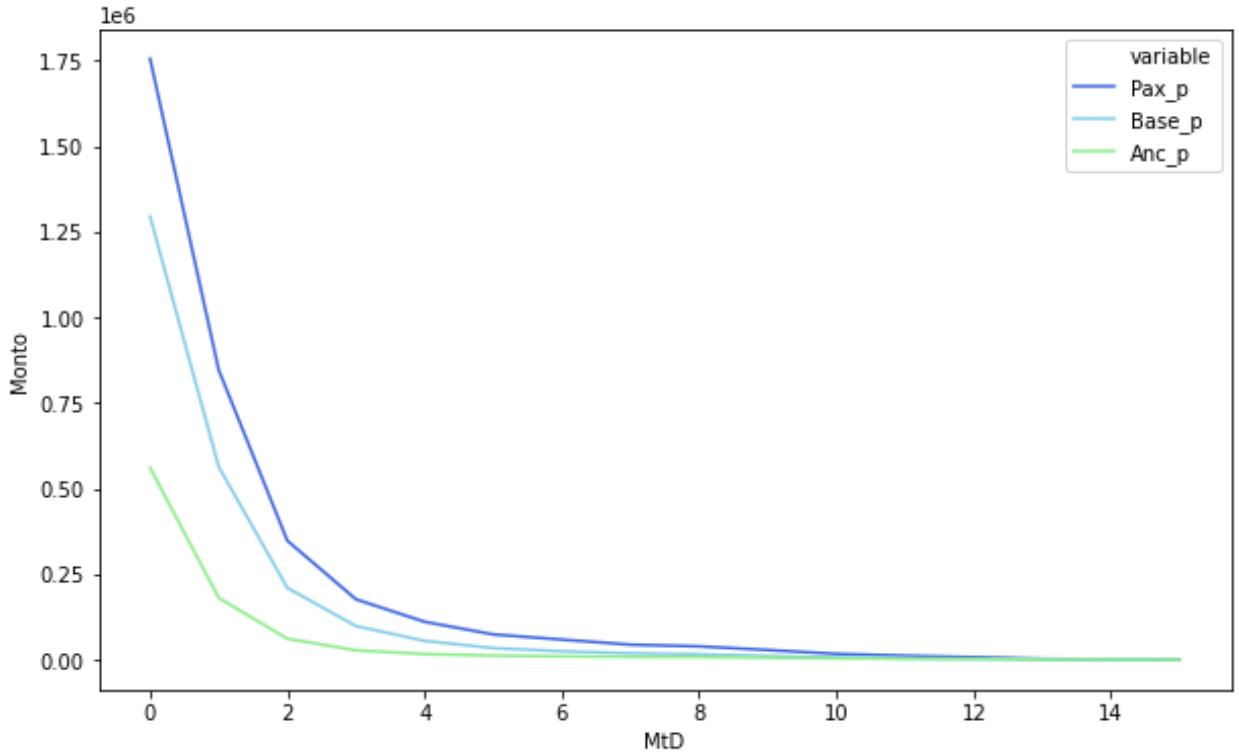
Fuente: Elaboración propia

Interpretando la gráfica desde la perspectiva de un vuelo, cuando éste es publicado (cuando comienza a estar disponible para la venta de lugares) tiene un factor de ocupación (LF por las siglas en inglés: Load Factor) de 0%, está vacío porque es

nuevo. Si el eje horizontal despliega los días restantes para el despegue (dtd por sus siglas en inglés: days to departure) y el eje vertical denota el LF, conforme nos acerquemos al dtd=0 (de derecha a izquierda) el factor de ocupación será igual o mayor al dtd anterior, porque el vuelo se va llenando. A esta representación gráfica a través de los dtd se conoce como **curva de maduración** de LF.

Al igual que la gráfica 5, el LF suele tener una pendiente más pronunciada en los LF cercanos a cero, es decir, no es tan común planear viajes con tanta anticipación (dependiendo la ruta). Hay que aclarar que el LF son los pasajeros acumulados entre los asientos del vuelo (o ruta, Hub, etc.) que se esté analizando, entonces el crecimiento de LF entre dos dtd multiplicado por los asientos será la venta de pasajeros en esa captura de dtd. La misma interpretación aplica para la variable MtD, así que es natural observar un comportamiento incremental de la venta conforme el MtD es menor:

Gráfica 7. Venta (ponderada por asientos) por MtD

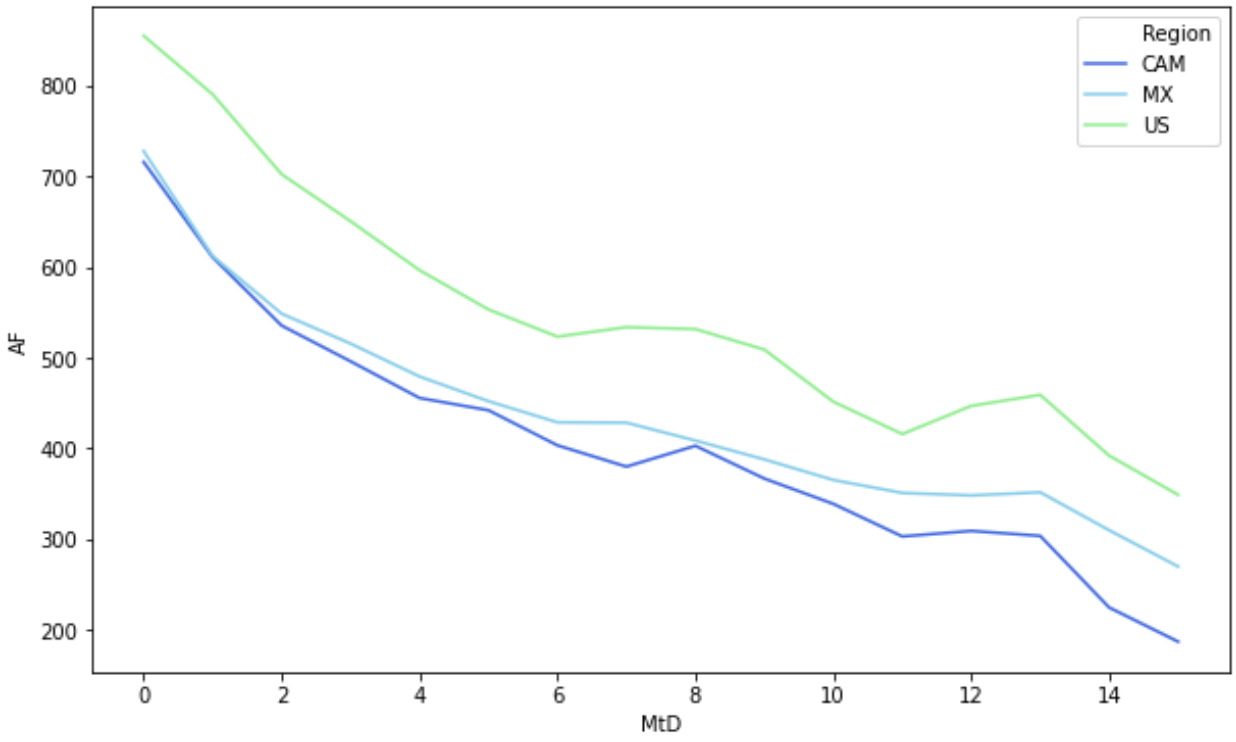


Fuente: Elaboración propia con datos de Volaris

Y como se presentó en la introducción, el área encargada del ingreso ocupa el dinamismo en precios para mantener niveles sanos de factor de ocupación en las temporadas bajas y aprovechar la demanda en las temporadas altas: donde la gente está dispuesta a pagar más por el mismo asiento. Véase como una tienda de frutas, en temporada de mangos el frutero/a (persona que vende frutas, para este ejemplo) sabe que hay gente dispuesta a pagar más por el mismo mango, así que, de un lote decide apartar cierta cantidad de mangos a un precio más elevado que el resto esperando que la demanda dispuesta a pagar esa tarifa llegue, debe acertar en la cantidad que aparta porque si esa demanda es menor a lo que pronosticaba, el mango se echará a perder, el vuelo despegará con asientos vacíos.

Explicado de manera general cómo funciona la administración de precios, es natural que conforme la fecha de despegue se acerque, el precio mínimo ofertado por asiento sea mayor, por eso la tarifa base promedio vendida por MtD tiene un comportamiento similar a las gráficas 6 y 7.

Gráfica 8. Tarifa base total por región y MtD



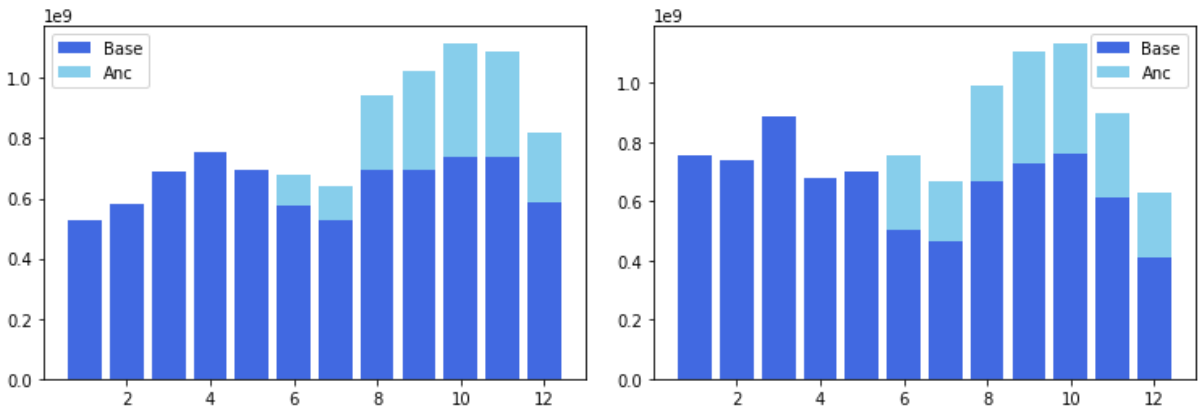
Fuente: Elaboración propia con datos de Volaris

Es común que la curva de maduración de tarifa, a diferencia de la de LF, no sea estrictamente creciente conforme el dtd disminuye, se explica por el mismo dinamismo de precios ejecutado por los estrategas de ingreso. Puede que el frutero se dé cuenta que apartó demasiados mangos en la temporada buena y decida abaratarlos antes de que se echen a perder. También tiene sentido que entre más larga la distancia en una ruta (mercado internacional), más cara será la tarifa promedio vendida.

Por otro lado, los precios de ancillaries no cuentan con el mismo dinamismo para el mismo producto, además que no sólo venden un producto sino múltiples y las veces que el cliente lo decida. Por esa razón no se analizará el comportamiento por MtD.

Para comprobar la temporalidad se muestra la agrupación por mes de vuelo y venta, cuando se modele el pronóstico esta agrupación servirá de apoyo para pronosticar el ingreso por mes de vuelo y así establecer metas mejor sustentadas por Hub.

Gráfica 9. Ingreso y venta total por mes para 2019



Fuente: Elaboración propia con datos de Volaris

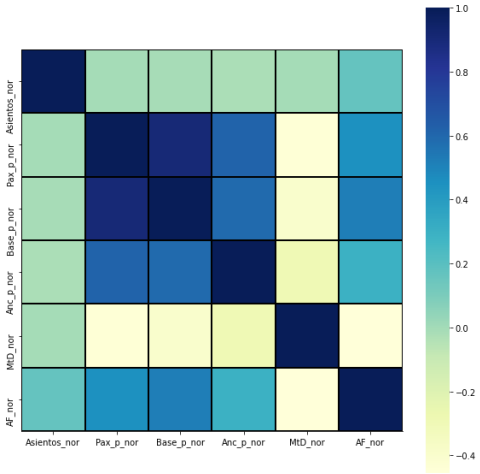
Al parecer, la venta de ancillaries tiene información faltante para los primero cinco meses de 2019, por ello el ingreso tampoco se despliega y en junio-julio parece ser

menor que los meses siguientes, porque sólo se despliega el ingreso por venta de los propios meses. Desafortunadamente, no es posible conseguir dicha información, sin embargo, si la agrupación del conjunto de datos para construir la Tabla Analítica de Datos (TAD, y es la estructura que recibe el modelo para desplegar resultados) es a nivel mes de venta, un cálculo porcentual respecto a la venta base pueda ser una solución.

Correlación

Es posible comprobar la relación entre las variables del conjunto de datos. Es importante normalizar estas variables, es decir, que tengan el mismo rango. En este trabajo, se opta por ocupar el método: $\frac{X - X_{min}}{X_{max} - X_{min}}$ y dado que se han trabajado los outliers, la distribución de las variables será de 0 a 1, ahora pueden compararse:

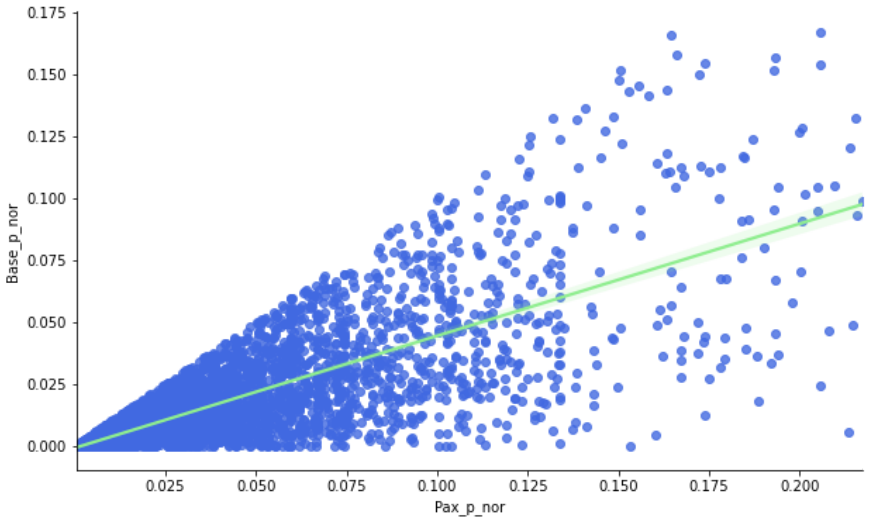
Imagen 1. Mapa de calor denotando correlación entre variables



Fuente: Elaboración propia con datos de Volaris

Sin embargo, no parece haber algo diferente a lo mencionado: entre más pequeño el MtD más venta hay. Tiene todo el sentido que, entre más venta de pasajeros, más venta base y entre más venta base, más venta de ancillaries.

Gráfica 10. Dispersión entre pasajeros y venta base

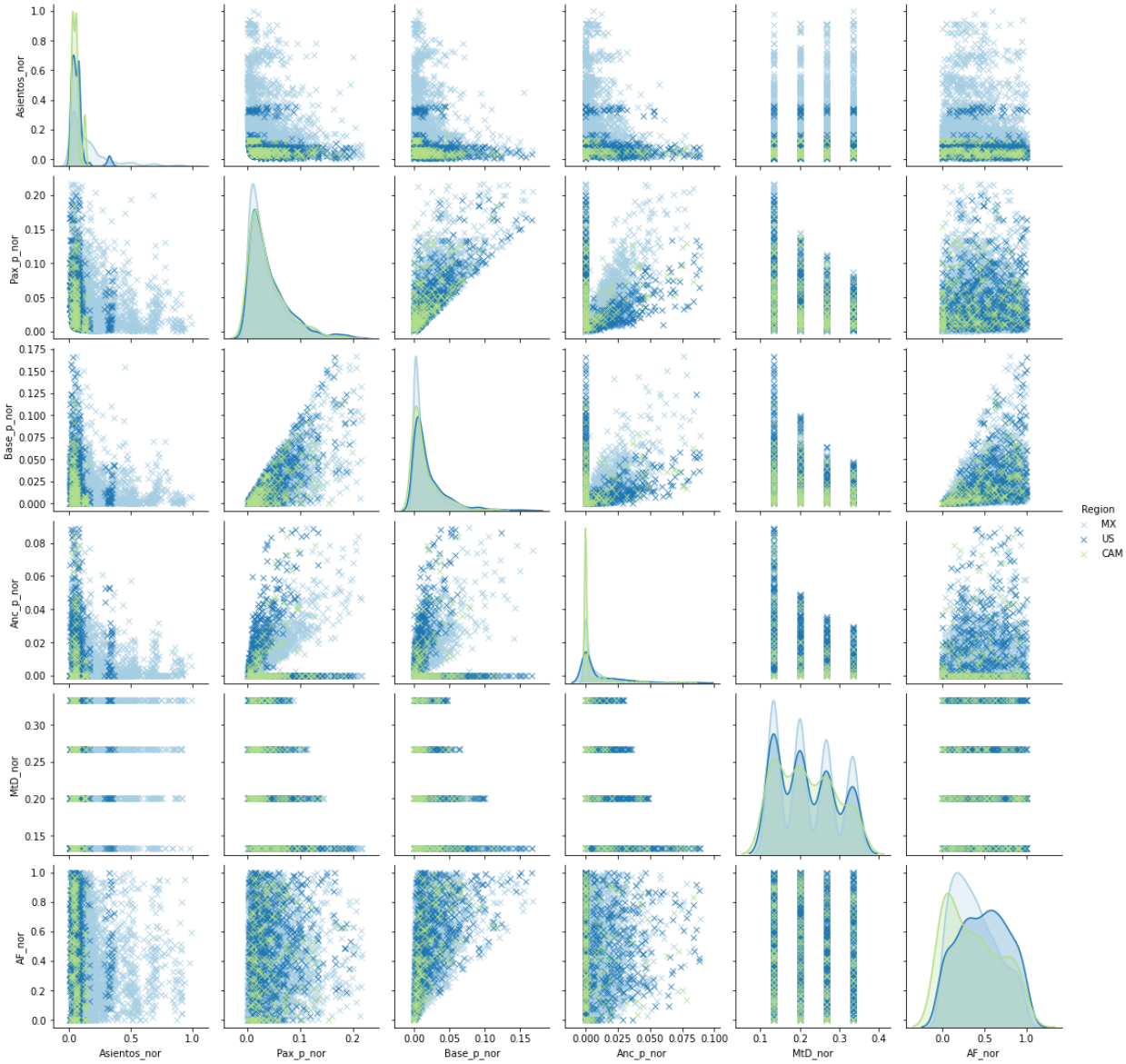


Fuente: Elaboración propia con datos de Volaris

Una manera práctica de encontrar correlación entre variables tomando en cuenta alguna categoría es a través de una pairplot:

Imagen 2. Pairplot entre variables normalizadas (MtD 2-5) coloreadas por región

(de registros con venta de pasajeros)



Fuente: Elaboración propia con datos de Volaris

Es notable que la capacidad de las rutas nacionales es mayor que en US y CAM. La anticipación (pax vs MtD) de Centroamérica es menor a que el resto de las regiones y la venta de ancillaries tiene una curtosis muy distinta a MX y se explica por dos razones: dadas las “cortas” distancias en las rutas nacionales, los clientes no optan por comprar productos adicionales, y los precios de dichos productos son mayores para rutas internacionales.

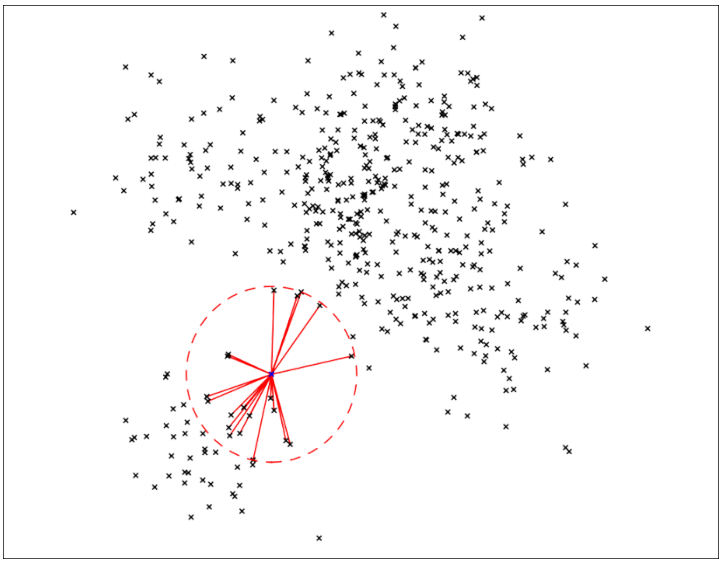
Multivariado

Imputar venta histórica

Como se ha mencionado en la sección anterior, no es posible contar con los registros de venta de ancillaries más atrás que mayo 2019, existe también la opción de imputar la venta de esos meses, aunque podría ser riesgoso por la cantidad de registros que se tienen que calcular. A continuación, se trabajará una metodología multivariante para imputar registros vacíos, en caso de no parecer exitosa, se recomienda omitir la variable de ancillaries en el conjunto de datos.

En pocas palabras, la metodología para imputar llamada “k vecinos más cercanos”, (KNN por sus siglas en inglés) radica en determinar el valor de un registro vacío según otros registros similares comparando otras variables numéricas del conjunto de datos.

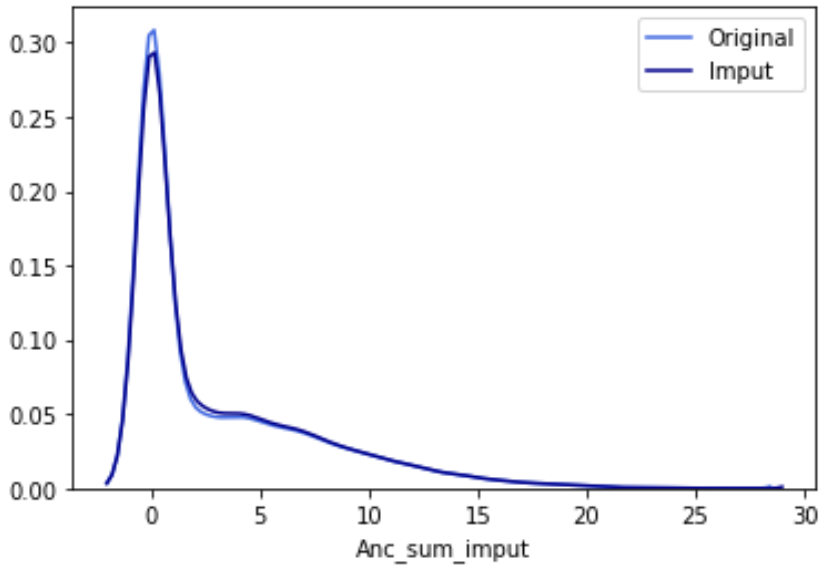
Imagen 3. Ejemplo KNN



Fuente: Nearest neighbor methods and vector models, Erik Bernhardsson (2015)

Así, utilizando KNN para calcular los registros de venta para mayo 2019, la distribución de la variable original contra la imputada no tiene diferencia notable:

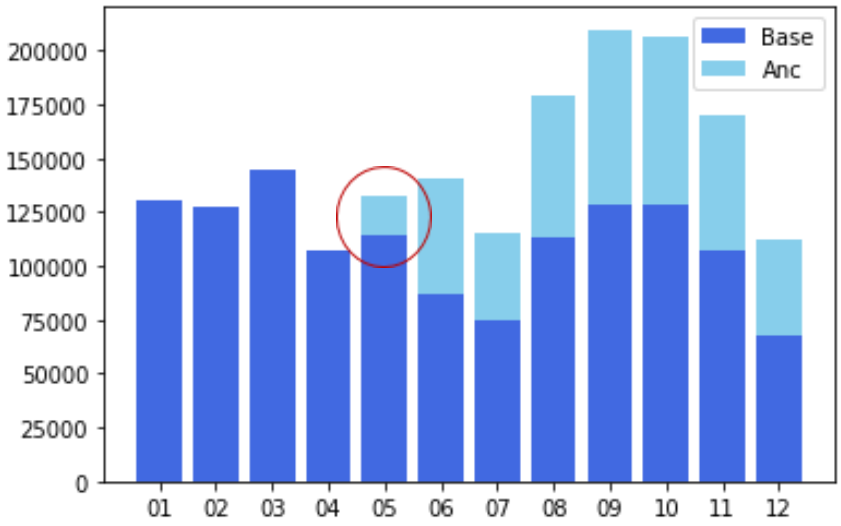
Gráfica 11. Imputar ausentes con KNN para venta de ancillaries en mayo 2019



Fuente: Elaboración propia con datos de Volaris

Sin embargo, al agrupar por mes de venta, el monto total no tiene un comportamiento similar contra los meses con valores correctos:

Gráfica 12. Venta de ancillaries por mes de vuelo en 2019



Fuente: Elaboración propia con datos de Volaris

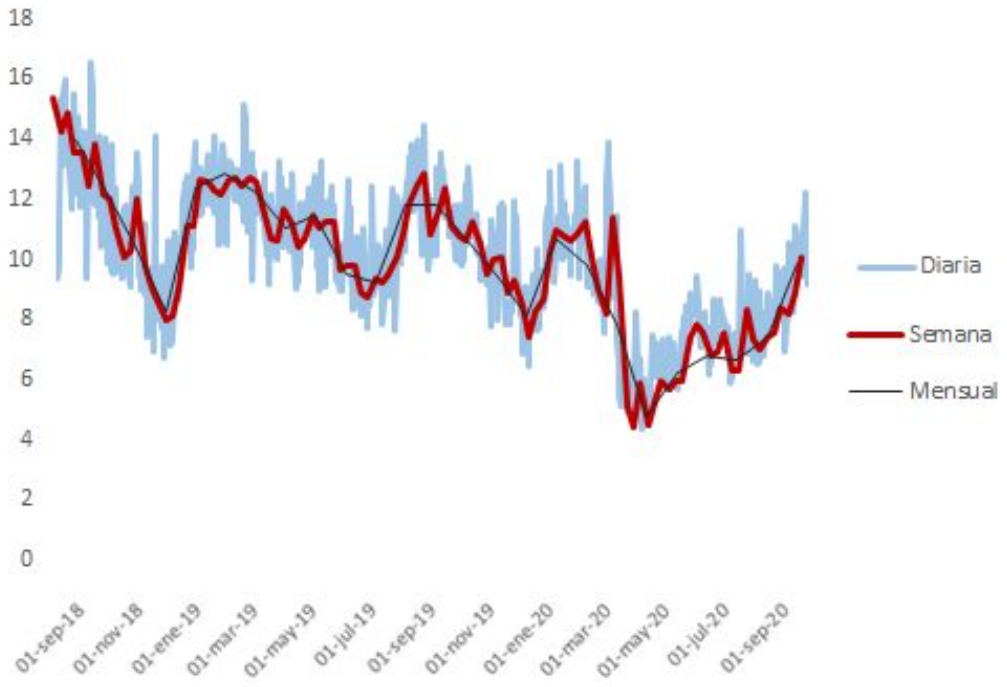
Dado que la venta de ancillaries depende directamente de la venta base, se opta por omitir la variable y sólo trabajar con venta de pasajeros e ingreso base de ahora en adelante.

Modelado supervisado

Caso discreto

En primer lugar, se busca predecir si habrá venta de pasajeros para determinada fecha, segmento y año-mes de vuelo. Pero, consecuentemente al intentar predecir de manera generalizada la venta de pasajeros diaria (modelación continua), ésta puede resultar en una certeza muy baja dadas las variaciones de un día contra el anterior y el siguiente. Por esta razón, se recurre a la agrupación de venta promedio de pasajeros en un nivel superior de tiempo.

Gráfica 13. Línea de tiempo con venta agrupada de pasajeros promedio



Fuente: Elaboración propia con datos de Volaris

Existe la opción de agrupar a nivel mes de venta reduciendo así prácticamente todas las variaciones abruptas, sin embargo, también se reduce significativamente la dimensión del conjunto de datos. Por lo tanto, se opta por predecir la cantidad de pasajeros promedio por semana para cada ruta en el mes de vuelo correspondiente, a nivel sistema es denotado por la línea roja en la gráfica 15.

Ahora, dado que se cuenta con sólo un cuatrimestre de venta para 2018 y el año 2020 aún no está completo aunado a la irregularidad en venta ocasionada por la ya bien conocida pandemia mundial, se opta por tomar en cuenta la venta de pasajeros sólo de 2019.

Se define como variable objetivo una nueva columna, desplegando 1 si en el registro correspondiente ocurre venta de pasajeros y 0 en caso contrario. El conjunto de datos contempla un 78% de registros con venta de pasajeros, la razón es: no porque tengas publicado un vuelo para despegar el siguiente año significa que esta semana vas a vender al menos un pasajero para él.

Al tener una estructura dependiente a una fecha, es sensato asumir la posibilidad de una tendencia y estacionalidad, por lo que la creación de variables como: venta de la semana anterior, dos, tres, cuatro y cinco semanas anteriores, son una buena idea. Se genera este conjunto de variables y serán llamadas variables "shift". Después se generan las columnas para el promedio de las variables shift con la interpretación: promedio de venta de pasajeros para las últimas dos semanas, tres semanas, y así consecutivamente hasta el promedio de venta de pasajeros de las últimas 5 semanas.

Para la predicción se va a construir una función $\hat{y} = f(X)$ siendo $\hat{y}$ el valor estimado dependiente de X: la matriz de características con valores continuos no nulos. En el conjunto de datos se encuentran columnas no numéricas como Hub, direccionalidad, etc. por lo que el paso siguiente es convertir esas variables categóricas a numéricas, la manera de lograrlo es a través de una transformación a "dummies", por ejemplo:

Tabla 5. Ejemplo de variables dummy

Ruta	Region	MX	US	CAM
BJXMID	MX	1	0	0
GDLSAT	US	0	1	0
CUNTIJ	MX	1	0	0
MEXSJD	CAM	0	0	1

Fuente: Elaboración propia con datos de Volaris

Para la categoría Región existen tres posibilidades, o mejor dicho, tres clases: MX, US y CAM. Entonces, para la variable categórica Región se construyen tres columnas

dummy: una por clase, cada columna será dicotómica y tendrá el valor 1 si la ruta correspondiente corresponde a dicha clase, caso contrario tendrá el valor 0. Calculadas las variables dummy para cada columna categórica, se cuenta ahora con la matriz X de valores continuos no nulos con la que se predecirá si para determinada semana, segmento y año-mes de vuelo, habrá venta de pasajeros o no.

Para comprobar la efectividad de dicha predicción, el conjunto de datos se separa aleatoriamente en un conjunto de entrenamiento y uno de prueba, en donde el conjunto de entrenamiento será el de mayor tamaño y en él, se ajustará qué tanto afecta cada característica al resultado final (coeficientes de cada variable en X) tal que $\hat{y}$ sea lo más cercano al resultado real, es decir, a la variable objetivo que indica si ocurrió venta de pasajeros en esa semana, segmento y año-mes de vuelo.

Selección de variables

Debido a la creación de variables dummy, incluso cada clase de Hub posee una columna dicotómica, esto elevaría la complejidad de cálculo para definir a $f(X)$ (desde ahora, se llamará modelo) y eso no se traduce proporcionalmente en una mejor aproximación. Por lo tanto, es práctico elegir las variables que mejor explican el valor de la variable objetivo. Además, para hacer comparables todas las variables se procede a escalar cada una de manera que queden en el rango de 0 a 1, el método es el mismo utilizado para apreciar la correlación entre variables: $\frac{X - X_{min}}{X_{max} - X_{min}}$.

Gracias a la metodología estadística que despliega las variables que mejor explican la variable objetivo a través de un análisis de varianzas, se despliegan efectivamente, las variables shift, es decir, que dada la naturaleza de serie de tiempo, es útil predecir si en determinada semana habrá venta de pasajeros dado que hubo o no venta de pasajeros en una o más semanas anteriores. Adicional a estas variables, la metodología estadística también arroja las variables promedio, Vuelos y MtD, después de definir $f(X)$ encontraremos si la relación de estas variables es fuerte o débil y positiva o negativa respecto a la variable objetivo.

Resultado

En la construcción del modelo se ajustan los coeficientes de X_i, es decir, se establece si existe una relación positiva o negativa entre la variable analizada y la variable objetivo: una relación positiva indica que si una variable crece la otra también y una relación negativa significa que si la variable crece, la otra decrece. Y, además de ser definida el tipo de relación, también se despliega la fuerza de ésta. Entonces, interpretando los resultados expuestos por la tabla 6:

Tabla 6. Relación entre variables* y objetivo

Variables	Relación
Promedio de pasajeros vendidos en las últimas cinco, dos y cuatro semanas	Fuerte positiva
Cantidad de vuelos publicados	Débil positiva
Año de vuelo es 2020	Débil negativa
Los meses restantes para el despegue son menores a 4	Fuerte negativa

**extracto de las variables con más impacto*

Fuente: Elaboración propia con resultado del modelo discreto

- Entre mayor sea el promedio de pasajeros vendidos en las últimas cinco dos y cuatro semanas, mayor la probabilidad de que ocurra venta de pasajeros. Significa que el mes de vuelo correspondiente cuenta con demanda estable, al menos durante la última semana de venta
- Entre más vuelos publicados tenga el segmento para el mes de vuelo correspondiente, más probabilidad de venta de pasajeros. Habla de la alta demanda en las rutas con mayor capacidad, tiene todo el sentido
- Dado que se ocupan los registros de venta en 2019 y se conoce que las curvas de maduración indican baja demanda anticipada, el año de vuelo 2020 es poco probable a ser comprado

- Entre más meses falten para el despegue, menor la probabilidad de venta de pasajeros. Tiene sentido que, al menos en la cultura general, la planeación de un viaje no se haga con mucha anticipación

Se utilizó un ensamble paralelo constituido por un Bosque Aleatorio, un clasificador ADA-Boost y uno de los mejores modelos en la industria: XGBoost. Al utilizar un ensamble mediante el clasificador votante suave (pondera los resultados de cada modelo), se tiene ahora un ensamble robusto y "mejor informado" respecto al fenómeno dado que ataca el problema desde diferentes perspectivas. Pero la pregunta importante es: ¿Qué tan bien predice el modelo si habrá (o no) venta de pasajeros?

Una matriz de confusión explica qué tan acertado es un modelo tanto en las respuestas correctas como incorrectas, es decir, según las clases de la variable objetivo (en este caso sólo son dos: hubo venta de pasajeros o no), compara el porcentaje de registros en el conjunto de datos contra la clasificación que estima el modelo. Así, la suma de la diagonal principal de dicha matriz será el porcentaje de certeza para el modelo en cuestión. La ventaja de esta presentación radica en que muestra con claridad el porcentaje de predicciones incorrectas para cada una de las clases de la variable objetivo.

Tabla 7. Matriz de confusión para modelo discreto

		Estimado	
		Venta	Sin Venta
Real	Venta	73%	5%
	Sin Venta	7%	15%

Fuente: Elaboración propia con resultado del modelo discreto

La certeza del modelo es de 88% con un grado de Falsos Positivos de 7%, definido como las predicciones erróneas cuando no ocurre el evento, es decir, el modelo predecirá que hay venta, cuando no ocurre según el comportamiento histórico. Caso contrario para los Falsos Negativos, el modelo predice en el 5% de las ocasiones que no habrá venta, cuando efectivamente la hubo. Cabe recalcar que estos porcentajes

difícilmente serán cero dado que el modelo intenta generalizar el fenómeno: hay venta o no de pasajeros.

A continuación se analizarán posibles patrones en las predicciones erróneas tanto para los Falsos Positivos como los Falsos Negativos con el fin de proponer ajustes en la metodología que mejoren la certeza del modelo.

Imagen 4. Mapa de calor en los errores de predicción por MtD y temporada

Mes vuelo	Temporada
ene	Alta
feb	Baja
mar	
abr	Alta
may	Shoulder
jun	
jul	Alta
ago	
sep	Baja
oct	
nov	Shoulder
dic	Alta

		Falsos negativos			Falsos positivos		
Temporada	Rango MtD	MX	US	CAM	MX	US	CAM
Baja	00-02						
	02-04						
	04-06						
	06-08						
	08-10						
	>10						
Shoulder	00-02						
	02-04						
	04-06						
	06-08						
	08-10						
Alta	00-02						
	02-04						
	04-06						
	06-08						
	08-10						
	>10						

Fuente: Elaboración propia con resultado del modelo discreto

Es útil separar los meses de vuelo según la demanda conocida: las temporadas altas son sin duda verano, invierno y la semana santa, al menos en México. Por otro lado, los meses más difíciles para el turismo y aviación son febrero y septiembre junto con su respectivo mes siguiente, la razón es la entrada a clases y fin de vacaciones, una combinación conocida como la cuesta de enero/agosto. Finalmente, hay meses que están en un nivel intermedio de demanda, a este periodo se le denominará temporada shoulder, englobando los meses de mayo, junio y noviembre, que son meses predecesores a la temporada alta.

Al contar con anticipación de demanda diferente entre las tres regiones de Volaris, es comprensible que el modelo no logre generalizar a nivel total. Por un lado, parece que

en la temporada shoulder y alta, los errores de predicción están sesgados cuando faltan de cuatro a ocho meses para el despegue (hasta 10 para la temporada alta). Dado que este comportamiento se repite para ambos tipos de error en la predicción, lo más probable es que se expliquen por las reglas de negocio enfocadas en planes promocionales semanales en las cuales se busca incentivar demanda para ese periodo de MtD ya que el corto plazo siempre cuenta con demanda de último minuto.

Por otro lado, para la temporada baja los Falsos Positivos, definidos en este caso como las ocasiones donde el modelo predice que hay venta cuando no es así, comparados con los Falsos Negativos, tienen una distribución distinta. La explicación puede ser la anticipación de la demanda en cada región, e incluso dentro de cada una existe variedad en la anticipación, por ejemplo: no es lo mismo comprar un vuelo a Cancún, donde ya planeaste tu hospedaje e itinerario con anticipación, que comprar un vuelo a Guadalajara por un tema de negocios, o comprar un vuelo a Chicago para ver a tu familia la próxima Navidad. La propuesta para reducir este tipo de error en la predicción es agrupar rutas según su anticipación con algún método no supervisado de clústeres.

Es evidente que el modelo acierta mejor en la venta para volar en los meses más cercanos (certeza del 99% para MtD de 0-2 y 96% para 2-4). Pero conforme el mes de venta está más alejado, más difícil es predecir el fenómeno y esto se debe a lo ya mencionado, la anticipación de diferentes mercados, por ejemplo: las denominadas rutas de placer como Miami, San Francisco o playas como Puerto Vallarta o Ixtapa Zihuatanejo.

Caso continuo

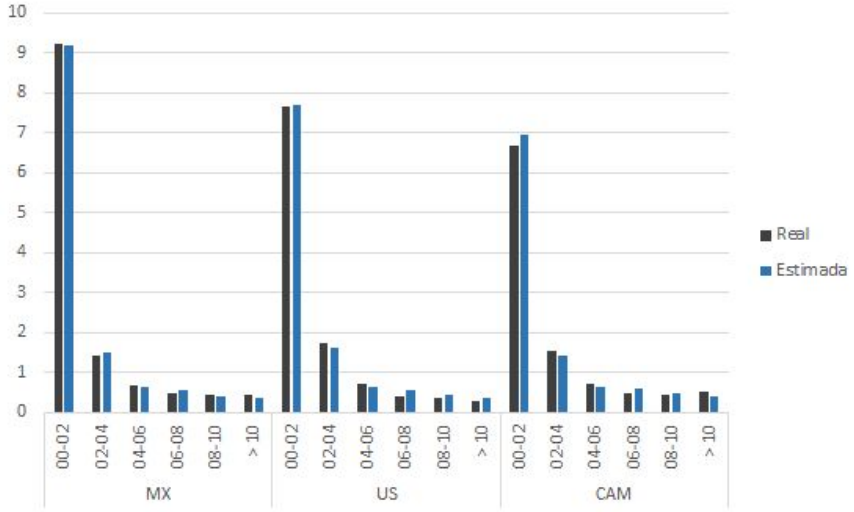
Ahora, se trabajará únicamente con los registros con venta de pasajeros, ya que para predecir en entorno de producción, el modelo de regresión logística servirá de apoyo: los registros que arroje con venta serán utilizados para predecir cuántos pasajeros se vendieron.

De igual forma que en el caso de modelación discreta, se trabajarán los datos en un nivel semanal, para disminuir diferencias entre lo real y lo estimado y así mejorar la certeza del modelo continuo.

Resultado

Se modela con el apoyo de un Árbol de Regresión hiperparametrizado, esto quiere decir que el algoritmo ha sido optimizado a fin de encontrar el mejor modelo posible, este proceso tomó casi 8 horas, pero los resultados valen la espera. En resumen, este modelo predice con base en las similitudes de todas las características analizadas, comparando variable por variable. De esta manera, por ejemplo: para predecir la cantidad de pasajeros promedio vendidos en una semana de mayo, para la ruta GDLMTY que pertenece al Hub GDL y a su vez a la región MX, que volará dentro de cuatro meses y ha tenido demanda estable en el último mes, se promediará la venta de las subdivisiones finales (hojas) más parecidas a estas características: venta en mayo, del Hub GDL, con MtD igual a cuatro, etc. De esta forma, el Árbol de Regresión logra ser un modelo poderoso y acertado en el 81.14% de los registros en el conjunto de prueba (para la región MX es ligeramente mejor: 83.5%). Las diferencias más notables ocurren en CAM cuando faltan menos de dos meses para el despegue, y hay dos interpretaciones válidas: el modelo sobreestima o dado que el mercado es relativamente “nuevo” (menor de 5 años), la demanda tiene un comportamiento diferente a la mexicana en cuanto a cultura de compra anticipada, tal vez los costarricenses saben que entre más cercana sea la fecha de vuelo, más caro será comprar y prefieren no hacerlo.

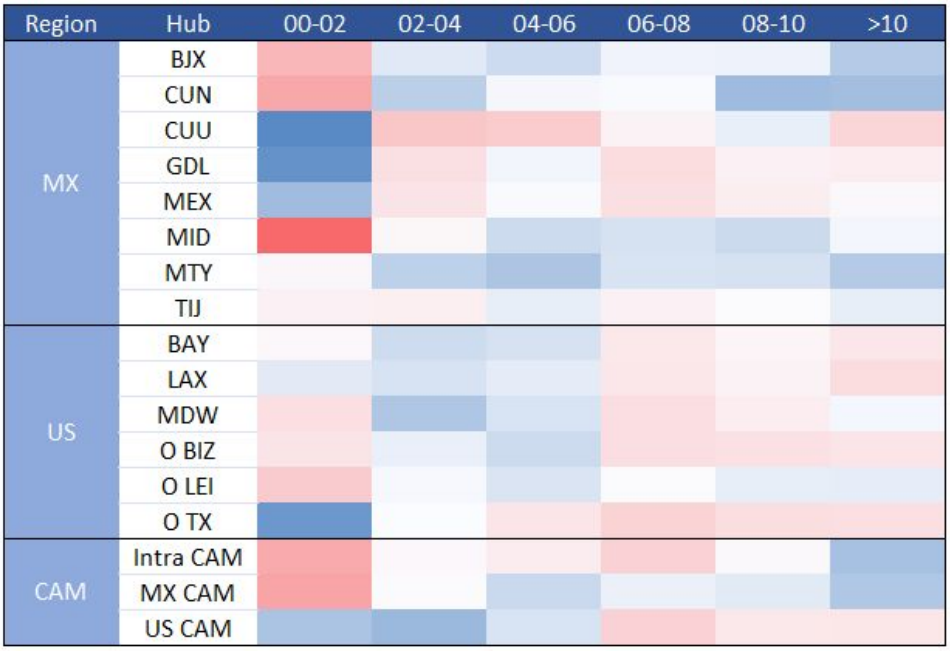
Gráfica 14. Pasajeros promedio vendidos semanalmente por Región y rango de MtD



Fuente: Elaboración propia con resultado de modelo continuo

Analizar la ubicación y distribución de los errores pueden darnos indicios sobre qué estrategia tomar para mejorar la certeza del modelo o incluso lograr que diferentes tipos de modelaje trabajen en conjunto para generalizar con mayor éxito el comportamiento real del fenómeno. En ese sentido, el comportamiento de las estimaciones por encima de lo real (azul) y debajo de ella, por Hub y meses restantes para la salida, se muestran a continuación:

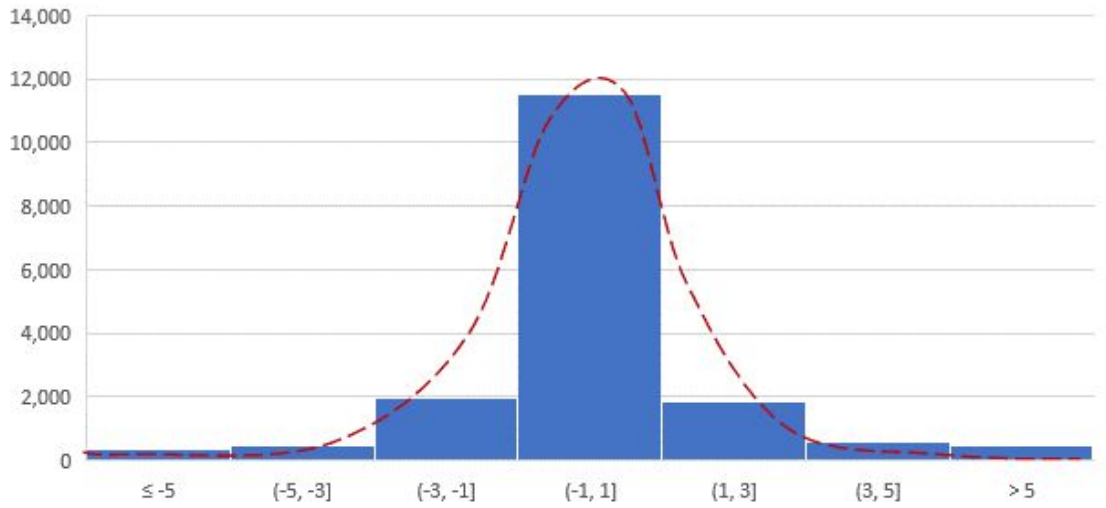
Imagen 6. Mapa de calor en los errores de predicción por Hub y MtD



Fuente: Elaboración propia con resultado del modelo continuo

Tomando en cuenta el rango de MtD cero a dos, es decir, la venta de pasajeros que ocurre cuando faltan dos meses o menos para el vuelo, es evidente que en los Hub con mayor anticipación como Cancún, Mérida y en general la región de Estados Unidos o Centroamérica, el modelo subestima la venta de pasajeros asumiendo que su curva de maduración no tiene el repunte de demanda en el último minuto, caso contrario con los meses con mejor demanda en el propio mes de vuelo, como el AICM, Chihuahua y Guadalajara donde el modelo sobreestima la demanda en el corto plazo pero subestima respecto al comportamiento generalizado en los MtD siguientes. Por ello, las propuestas hechas para mejorar la certeza del modelo discreto también aplican para mejorar la certeza del caso continuo: agrupar rutas por anticipación de venta.

Gráfica 15. Distribución de errores en predicción



Fuente: Elaboración propia con resultado de modelo continuo

Ahora, al graficar la distribución de errores en la predicción en la cantidad de pasajeros (ya sea positiva o negativamente) se despliega la famosa distribución denominada como “normal” con una varianza relativamente estable y pequeña. Es importante lograr que la frecuencia de las diferencias negativas y positivas extremas se reduzcan o eliminen, o dicho en términos matemáticos, la función de pérdida se minimice.

Conclusión preliminar

Como primer acercamiento ambas modelaciones: la discreta que pretende predecir si para determinada semana, segmento y año-mes de vuelo habrá venta de pasajeros o no y, la continua (que se basa en el resultado positivo de la primera) que intenta predecir la cantidad de pasajeros vendidos en promedio para determinada semana, ruta y año-mes de vuelo, cuentan con una certeza aceptable y resuelven (hasta ahora) parte del problema. La agrupación de rutas y meses de vuelo o venta, sustentada en resultados concluyentes de modelos de aprendizaje automático no supervisado, servirán para mejorar la certeza de ambas.

Modelado no supervisado

El objetivo de esta sección es encontrar y separar los comportamientos que son notablemente diferentes entre sí. Se mencionó en el apartado que explica las curvas de maduración que el Load Factor (LF), métrica que mide qué tan lleno está un vuelo, tiene un comportamiento creciente conforme se acerca el día a la salida, sin embargo, como también se ha mencionado a lo largo del presente documento, la anticipación de la demanda suele ser diferente dependiendo el origen-destino, el mes de vuelo, entre otras características. A continuación, se detalla el proceso para aportar valor a los modelos supervisados gracias a la modelación no supervisada.

Tabla 8. Extracto de la Tabla Analítica de Datos (TAD) preliminar para Clustering

Segmento	Año	Mes	Dtd_000	Dtd_001	Dtd_002	...	Dtd_198	Dtd_199	Dtd_200
BJXPVR	2020	ene	88%	88%	87%	...	4%	4%	4%
LAXGDL	2019	feb	93%	91%	84%	...	1%	1%	1%
CZMMTY	2019	jul	95%	94%	94%	...	11%	11%	11%
OAXTIJ	2019	jul	90%	89%	88%	...	6%	6%	6%

Fuente: Elaboración propia con datos de Volaris

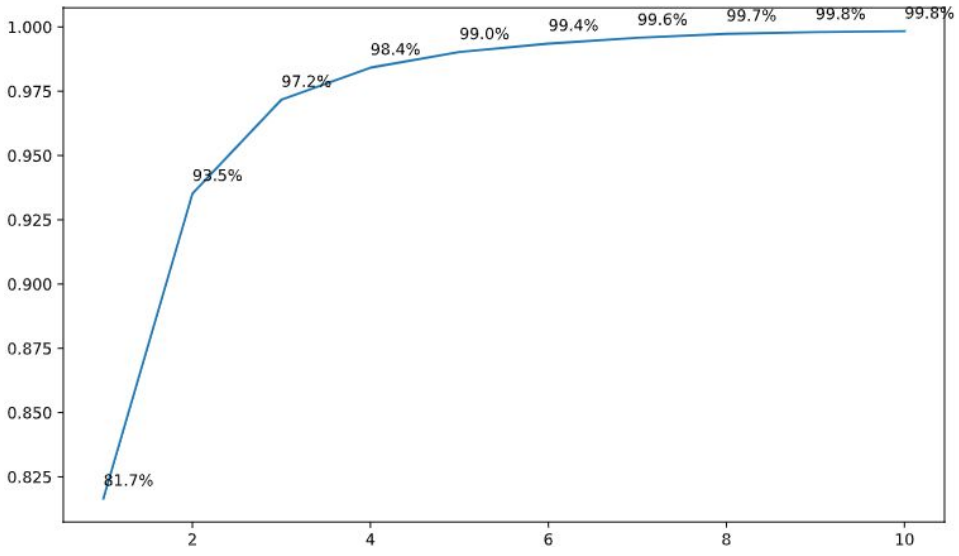
Al estructurar los datos con el formato presentado en la tabla 8, cada registro representará la curva de maduración de LF para un segmento-mes desde 200 días a la salida hasta el día de despegue. La TAD será quien “alimentará” a la modelación no supervisada para que gracias a su estructura, se puedan identificar los grupos de segmentos-mes que más se parecen entre sí. El siguiente paso es reducir la dimensión de nuestra TAD preliminar, no es indispensable contar con las 200 columnas de Dtd ya que cada día tiene una correlación muy fuerte con los días adyacentes.

Reducción de dimensiones

A través de un análisis de componentes principales (PCA por sus siglas en inglés), se aplicará una reducción de espacio, es decir, se reducirán las variables numéricas a través de una transformación de combinaciones lineales de factores no correlacionados entre sí, intentando minimizar la pérdida de varianza entre todas las variables. En otras palabras, se resumirán los 200 días restantes para el despegue en menos columnas pero que representen fidedignamente el total de las curvas de maduración.

Después de omitir el 1% de registros atípicos y escalar cada Dtd con la normalización estándar $\left(\frac{x-\mu}{\sigma}\right)$ se analiza a qué espacio reducido es conveniente llevar los datos:

Gráfica 16. Varianza explicada por PCA según número de componentes



Fuente: Elaboración propia con la reducción de espacio

Así, se opta por tres componentes dado que explican el 97.2% de la varianza calculada sobre el total de las curvas de maduración, además este número de componentes servirá de apoyo para diferenciar gráficamente los futuros grupos.

Clustering con K-means

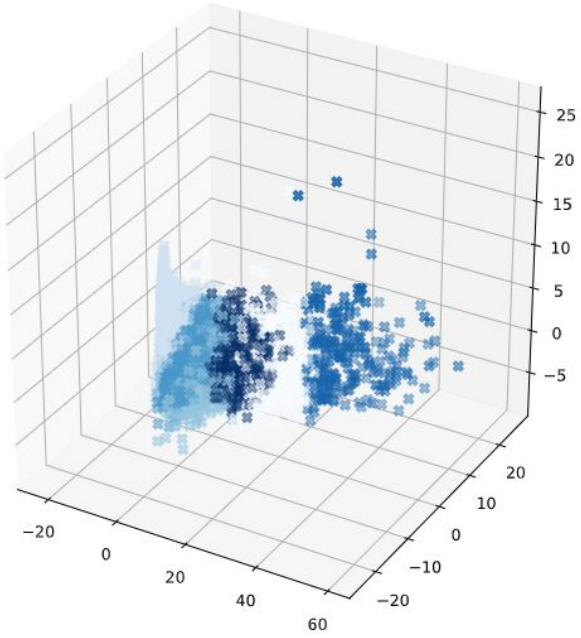
Este tipo de modelación no supervisada pretende clasificar datos en *k* diferentes grupos (clústeres) tal que los datos de un mismo clúster estén más cerca de su promedio (centroide) que el de algún otro grupo. El proceso para lograrlo se resume en:

1. Se eligen *k* promedios iniciales aleatoriamente
2. Se crean *k* clústeres a partir de esos puntos, asociando cada observación con el promedio más cercano
3. El centroide de cada clúster se convierte en el nuevo promedio
4. Se repite el paso 2 y 3 hasta converger, es decir, el centroide es igual al promedio

La pregunta para responder ahora es, ¿En cuántos clústeres es conveniente separar los datos? Afortunada o desafortunadamente, la respuesta es: “depende”, depende del problema a tratar.

En este caso, se conoce a fondo la industria y al contar con tres tipos de mercado distintos (LEI: Leisure, BIZ: Business y VFR: Visiting Family and Relatives) y dos regiones principales (MX: México y US: Estados Unidos) se opta por elegir k = 6 y mediante la gráfica de dispersión de los datos resultantes de PCA se logra apreciar que cada grupo está “unido” y se logra diferenciar respecto a otro.

Gráfica 17. Dispersión de grupos en espacio reducido

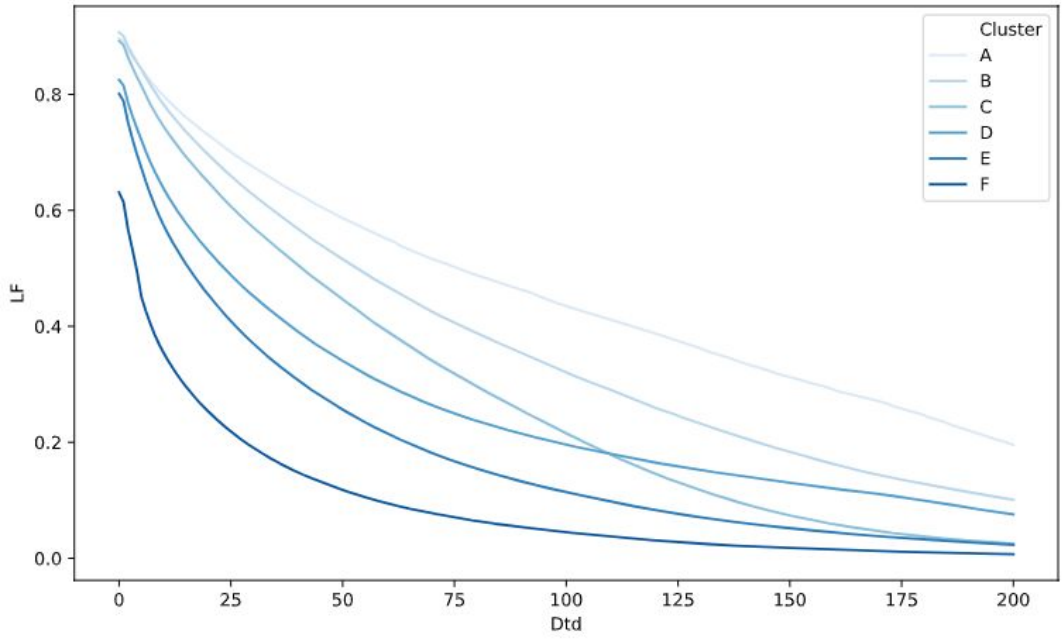


Fuente: Elaboración propia con la reducción de espacio

Resultado

La predicción del modelo K-means concatenado a la curva de maduración de LF original (sin transformación de PCA) evidencia los diferentes comportamientos, los seis más representativos o mejor dicho, el promedio de ellos.

Gráfica 18. Curva de LF para los distintos clústeres



Fuente: Elaboración propia con resultados del modelo K-means

Desde la perspectiva de la industria, las curvas superiores indican una anticipación de compra considerablemente alta, por ejemplo: los segmentos-mes que corresponden a la curva A tienen la mitad de cada vuelo ocupado (en promedio) cuando todavía faltan más de dos meses para el despegue (Dtd = 77) que corresponden en la generalidad a rutas LEI y VFR en temporadas altas, caso contrario a los segmentos-mes con comportamiento “de último minuto” que se agrupan en la curva E, que logran mismo nivel de LF pero hasta 23 días antes de la salida, que corresponden principalmente a las rutas BIZ, los viajes de negocios que se concretan de una semana para la siguiente.

Conclusión preliminar

El tratamiento correcto de datos arroja una característica valiosa que si bien es inherente al comportamiento de las rutas, permanecía “escondida” y la modelación no supervisada la “ha traído a flote”. Además de que muy probablemente ayude a los pronósticos de la modelación supervisada, también aporta valor al negocio, ya que ahora puede definirse una estrategia segmentada por tipo de comportamiento, por ejemplo: promociones enfocadas en la anticipación de rutas con curvas C y D, protección de fechas más demandadas para las curvas A y B y por último, un enfoque en la tarifa ofertada dada la inelasticidad de las curvas E y F, cada una de estas estrategias preliminares pueden ser ejecutadas en el rango de Dtd donde se obtengan mejores beneficios.

Como siguientes pasos se plantea:

- Utilizar redes neuronales enfocadas en series de tiempo o redes convolucionales para mejorar la certeza de la modelación supervisada
- Estructurar un pipeline ensamblando los modelos necesarios tal que pueda predecir nuevos datos con facilidad

To be continued..