PROYECTO FIN DE MASTER

Customer Lifetime Value y Segmentación de clientes de aerolíneas

Josefa Calero

Siena Martinez

Contents

Introducción

El objetivo principal de este proyecto es poder entender mejor las características de los clientes de las aerolíneas a partir del análisis de los movimientos de los pasajeros provistos en nuestra base de datos. Descubrir patrones a partir de las transacciones de los pasajeros nos permitirá realizar una segmentación apropiada de nuestros clientes a los que podremos ofrecer billetes con precios mejor ajustados, aumentar su retención y poder estimar los ingresos potenciales de cada uno de esos clientes. Estos tres resultados son de gran relevancia para poder por ejemplo mejorar el retorno y eficacia de las campañas de marketing, aumentar los ingresos por cada uno de los pasajeros y disminuir la tasa de pérdida de clientes.

El sector de los viajes y turismo cuenta con una larga historia en diferentes estrategias de segmentación de clientes buscando la agrupación de los usuarios con transacciones y características socioeconómicas similares que les permitan segmentar el mercado. Esos esfuerzos de segmentación se han realizado principalmente a partir de los distintos programas de fidelización de las aerolíneas con el fin de separar a los clientes con mayor valor potencias y garantizar su retención.

Sin embargo, investigaciones anteriores han demostrado que los programas de afiliación de las aerolíneas no son suficientes para poder identificar a todos sus pasajeros con gran valor potencial ya que no todos los clientes de que viajan por negocio poseen la tarjeta de afiliación a la compañía o la usan cuando reservan. Por otro lado, los pasajeros que vuelan en la clase *economy* representan un volumen importante y no están inscritos en sus programas de afiliación así que la influencia de la aerolínea en incentivar una compra futura es muy limitada sino se realiza un esfuerzo de análisis más complejo.

Para superar las limitaciones mencionadas, se propone la utilización de técnicas de machine learning destinadas a reconocer patrones en el comportamiento de los pasajeros de avión e identificar características demográficas comunes que nos permitan una óptima agrupación de los mismos. Para ello, se ha empleado K-means, algoritmo de aprendizaje no supervisado que a partir de las transacciones de los pasajeros realiza una separación de los datos en grupos que comparten características similares. A partir de los resultados de segmentación del K-means se emplean árboles de decisión, algoritmos de aprendizaje supervisado que nos permiten extraer reglas de clasificación y descubrir los distintos elementos que caracterizan cada segmento para separar a los clientes que mayor valor aportan de aquellos clientes eventuales que no son relevantes para la aerolínea.

Por último, para completar las conclusiones extraídas de los modelos anteriores, se estima el Customer Lifetime Value (CLV), métrica que estima el valor que un cliente puede aportar a una empresa durante el tiempo en que mantiene una relación con la misma. Conocer el CLV garantiza que el esfuerzo que una compañía realiza para captar clientes es rentable y se amortiza durante la relación que el cliente mantiene con la empresa. Nos permite identificar a los clientes más rentables y prevé el ROI de los programas de fidelización, asignando los recursos de una forma eficiente y rentable.

Datos utilizados

La tabla recibida contiene datos de las reservas aéreas realizadas, contiene 9.546.302 de registros. A continuación se describen las variables que se han utilizado a lo largo de todo el proceso:

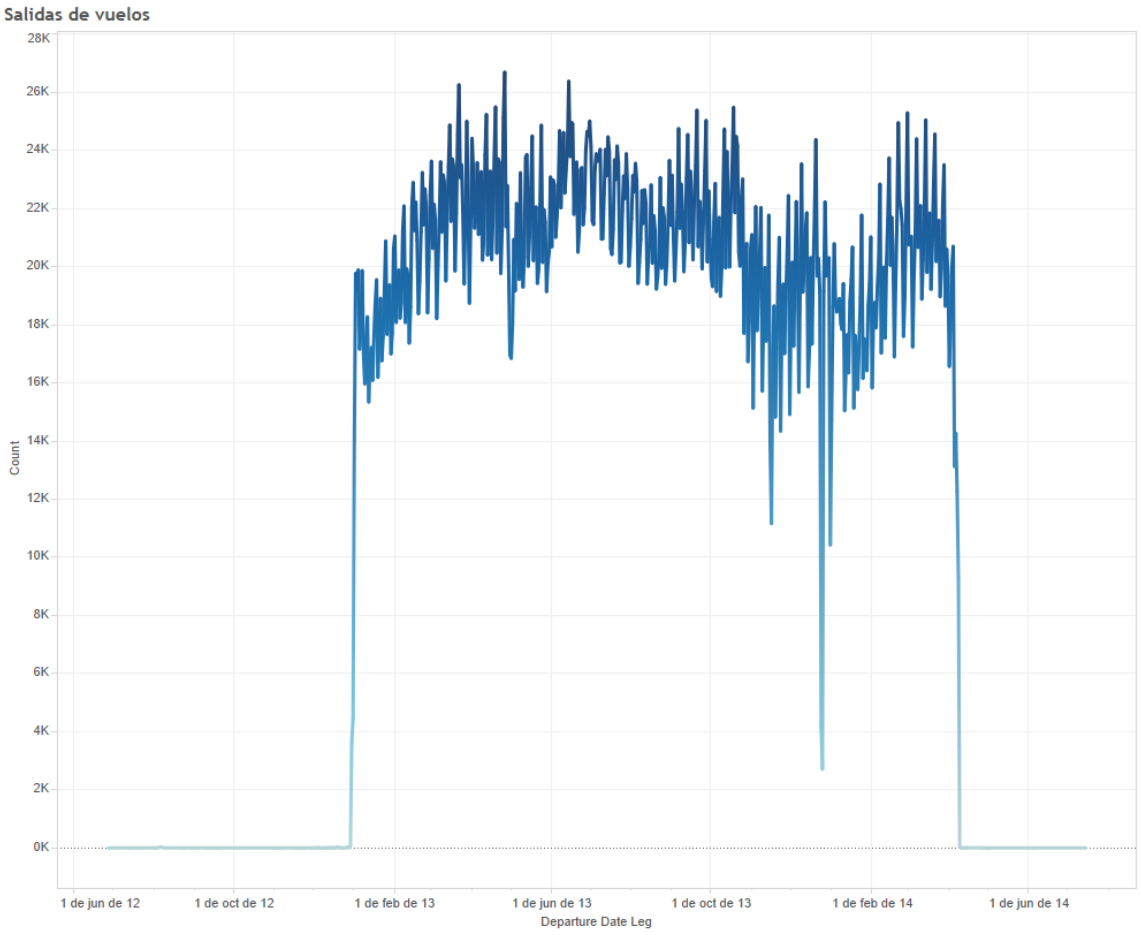
Nombre de la Variable	Descripción de la variable
Rloc	Código identificador del viaje
Gender	Género del pasajero
Age	Edad del pasajero
Date_of_birth	Fecha de nacimiento del pasajero
Document_number	Número del documento
Nationality	Nacionalidad del pasajero
Cabin_code	Clase del asiento del pasajero
Departure_date_leg	Fecha de salida del vuelo
Quality_index	Calidad del dato
Creation_date	Fecha de compra del billete
Advance_purchase	Tiempo transcurrido entre la fecha de reserva y del vuelo
Booking_status_code	Código del estado de la reserva
Board_point	Código de tres letras del aeropuerto de embarque
Board_lat	Latitud del punto de embarque
Board_lon	Longitud del punto de embarque
Board_country_code	Código del país de embarque
Board_continent_code	Código del continente de embarque
Off_point	Aeropuerto de final de trayecto
Off_lat	Latitud del punto de aterrizaje
Off_lon	Longitud del punto de aterrizaje
Off_country_code	Código del país de aterrizaje
Off_continent_code	Código del continente de aterrizaje
Distance_seg	Distancia en kms del segmento (puede ser diferente a la distancia real recorrida)
Revenue_amount_seg	Ingresos
Fuel_surcharge_amount_seg	Cargo por combustible
Route	Ruta

Carga y filtrado de datos

Para trabajar mejor con los datos recibidos, creamos una base de datos en PostgreSQL con la que podremos realizar validaciones y filtros de los datos. Generamos un script que crea la estructura de la tabla y, posteriormente, importamos los datos recibidos.

Una vez cargados los datos, comenzamos el proceso de comprensión de los mismos. Vemos que la tabla cargada tiene 9.546.302 de registros, sin embargo, no todos estos registros tienen identificado el campo document_number y, como vamos a realizar una segmentación de

clientes, necesitamos que este esté completo ya que será uno de los campos clave. Además, revisando las fechas de salida de vuelos, se puede ver que no siguen una distribución constante:



Por lo que también filtramos los datos para que las fechas de salida de los vuelos estén entre el 01/01/2013 y el 07/04/0214. Siendo esta última fecha la que tomaremos como referencia a la hora de realizar la segmentación.

Además nos quedamos con aquellos registros cuyo estado de reserva sea HK (confirmado) y, finalmente, para tener todos los datos personales de los clientes y así poder perfilarlos a través de las variables sociodemográficas, aplicaremos el filtro quality_index = 1.

Una vez aplicados estos filtros, nos quedamos con 777.978 registros, de los cuales hay 235.331 documentos identificadores de clientes distintos.

Realizamos nuevas comprobaciones sobre esta tabla que hemos generado y vemos que hay varias personas asignadas al mismo documento de identidad por lo que procedemos a eliminar a todos ellos. Como criterio elegido, una persona será única si los campos document_number, gender, date_of_birth y nationality no están duplicados. Tras esta limpieza nos quedamos con 738.533 registros y 231.973 clientes únicos. Con lo que ya tenemos una tabla depurada de clientes a los que aplicarles la segmentación.

Metodología

La segmentación del mercado es una estrategia de dividir un mercado objetivo amplio en subconjuntos de los clientes que tienen unas características comunes y necesidades (Haley, 1968). El propósito de la segmentación del mercado es ser capaz adaptar la inversión realizada en marketing con ofertas de productos que se adaptan mejor a los clientes potenciales y por tanto, aumentar los ingresos gracias a la personalización.

Para realizar esa segmentación se ha partido del modelo de RFM (Recency, Frequency and Monetary) utilizado comúnmente para predecir el comportamiento de los clientes. Esto se consigue examinando lo que el cliente ha comprado utilizando tres factores: (R) Recencia de compra, (F) Frecuencia de compra y (M) Valor de la compra en términos monetarios. El análisis RFM se basa en la conocida "Ley de Pareto" o del 80/20. En el caso del análisis RFM se diría que el "80% de la compras provienen de 20% de los clientes" o "que el 20% de los clientes genera el 80% de las ventas".

Las variables del RFM van a ser los datos de entrada para el algoritmo de segmentación que vamos emplear, el K-Means. El algoritmo K-means, es uno de los métodos de clustering (aprendizaje no supervisado) más usados. Es destinado a situaciones en las cuales todas las variables son de tipo cuantitativo, y la distancia euclidiana es generalmente escogida como medida de disimilitud. La idea principal es definir k centroides (uno para cada grupo) y luego tomar cada punto de la base de datos y situarlo en la clase de su centroide más cercano. El próximo paso es recalcular el centroide de cada grupo y volver a distribuir todos los objetos según el centroide más cercano. El proceso se repite hasta que ya no hay cambio en los grupos de un paso al siguiente. El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro o cuando los grupos tienen diferentes tamaños y formas.

Después de la aplicación de los dos algoritmos de agrupamiento y etiquetado de lealtad a cada grupo, la siguiente etapa es conectar las características demográficas de los clientes a los resultados de la agrupación. Con el objetivo de que los clientes con el mismo grupo tendrán las mismas características. Por lo tanto en esta etapa se aplican los modelos de Decision Tree y Random Forest para predecir la cantidad de lealtad basada en las variables demográficas.

Consideramos que los árboles de decisión es un método apropiado para nuestro estudio, ya que potencialmente pueden proporcionar información de mayor valor de negocio que otros métodos de clasificación (redes neuronales, regresión logística) mediante la generación de reglas que identifican claramente los clientes de mayor valor y los criterios y elementos que más influyen la generación de ese valor.

Otras ventajas de los árboles de decisión se relacionan con su alta flexibilidad. En concreto, estas ventajas se derivan de cuatro propiedades. En primer lugar, los árboles de decisión son un método no paramétrico - significa que no hay distribuciones (por ejemplo de normalidad) o formas funcionales (por ejemplo, linealidad) necesita ser especificado. Además, son invariantes a las transformaciones (por ejemplo, no se requieren transformaciones logarítmicas). En segundo lugar, los árboles de decisión no requieren la preselección de variables, las variables con un alto poder explicativo pueden ser fácilmente separadas de los restantes variables, menos importantes. En tercer lugar, con los árboles de regresión se pueden utilizar tanto variables continuas y categóricas y (de forma automática) incluyen interacciones entre las variables. En

cuarto lugar, los árboles de decisión son robustos a los efectos de los valores extremos, y por último, los árboles de decisión pueden todavía generar resultados útiles. Esta cualidad es especialmente útil para los posibles clientes sobre los cuales es probable que sea limitada la información.

Por otro lado, aparte de los árboles de decisión, se ha empleado el método de Random forest, método que combina una cantidad grande de árboles de decisión independientes probados sobre conjuntos de datos aleatorios con igual distribución. Este método supone una modificación sustancial respecto a los árboles de decisión porque su construcción se base en el método *bagging*. La idea esencial del *bagging* es promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la variación. Para ello, se incluyen dos niveles de aleatoriedad. Primero se selecciona aleatoriamente con reemplazamiento un porcentaje de datos de la muestra total. El segundo elemento de aleatoriedad se da en cada nodo, al seleccionar la partición óptima teniendo en cuenta sólo una porción de los atributos, elegidos al azar en cada ocasión.

En relación al CLV se han empleado dos modelos distintos. Un primero modelo es el del CLV hemos utilizado el Beta Negative/NGB model que supone una superación del modelo de Pareto/NGB para estimar la probabilidad de una próxima compra para evaluar si cuantos de nuestros clientes están todavía activos dada su historial de compra. En este modelo se tiene en cuenta la edad actual del cliente, la edad en la realizó su primera compra, la recencia y la frecuencia. A partir de estas variables, como cualquier modelo de probabilidad bayesiana, estima la probabilidad de compra (p=1) o no (p=0).

Para completar el análisis del CLV hemos tenido también en cuenta aparte de si los clientes están activos o no, la predicción del valor de las futuras compras gracias a un modelo Gamma-Gamma. Este modelo parte de los siguientes tres principios:

El valor monetario de la transacción determinada de un cliente varía al azar alrededor del valor medio de las transacciones.

Los valores promedio de transacción varían entre los clientes pero a nivel de cada cliente este se mantiene estable a lo largo del tiempo.

La diferencia de la distribución de los valores medios de transacción de cada cliente es independiente de la transacción en sí.

Por último, en los valores monetarios de las futuras compras se ha tenido en cuenta como paso ulterior la tasa de coste de capital para ajustar los valores monetarios a las tasas de descuento por el paso del tiempo.

Resultados

En esta sección se van a exponer los principales resultados obtenidos para cada una de las técnicas de Machine Learning expuestas en el apartado de metodología. El lenguaje utilizado para todas ellas ha sido Python y el módulo que principalmente se ha utilizado ha sido sklearn.

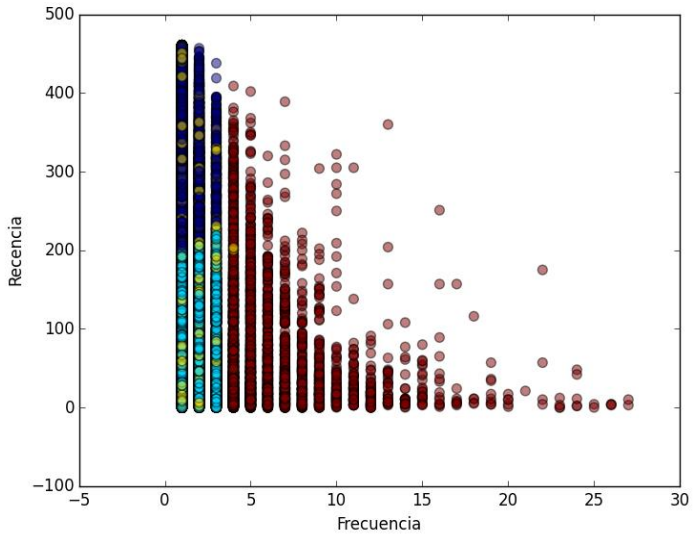
RFM y K-means

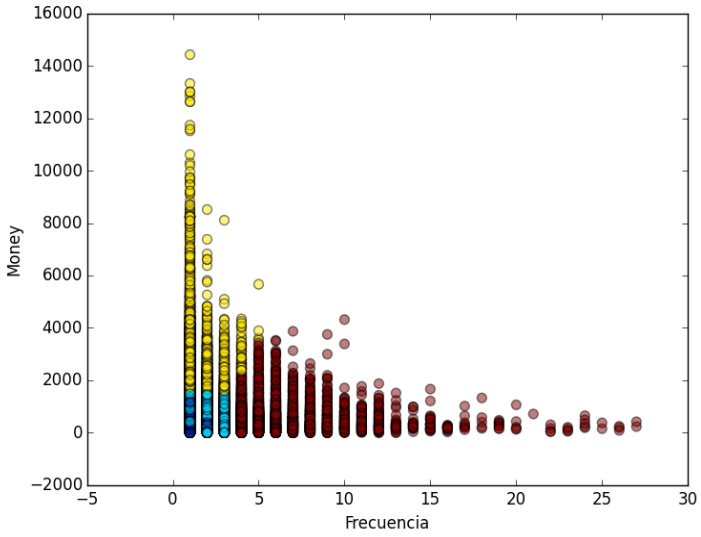
Partiendo de los datos filtrados de nuestra base de datos, generamos una tabla que contiene los siguientes campos:

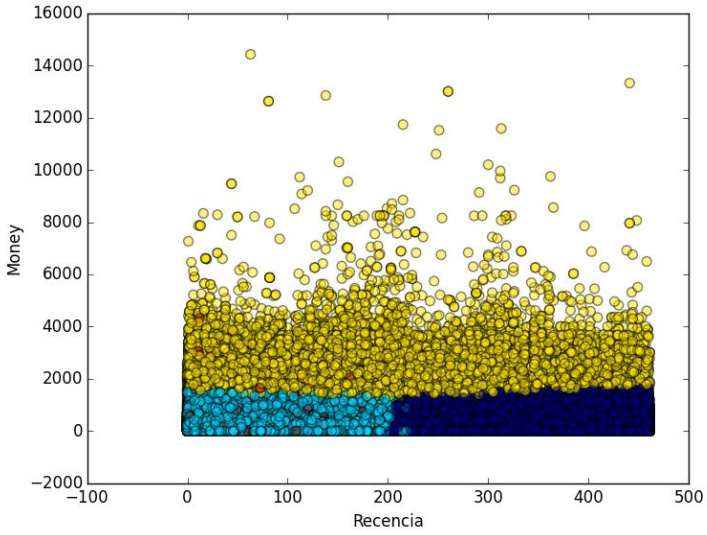
- Document_number → código único identificador del cliente
- Frecuencia → frecuencia de compra de billetes en el histórico que disponemos
- Recencia → tiempo transcurrido desde la compra del último billete
- Money → importe promedio gastado por viaje realizado

Para lanzar la segmentación utilizamos el paquete sklearn de Python. Este ejecuta una segmentación a partir de un número dado de clusters, por lo que generamos un bucle que lance segmentaciones de 1 a 8 grupos.

Revisamos cada una de las segmentaciones devueltas por el K-means y decidimos que la de 4 grupos era la óptima a la hora de segmentar los clientes ya que es la que presenta una distribución más heterogénea.







A partir de esta visualización y apoyándonos en los valores representativos de cada segmento, podemos obtener las siguientes conclusiones:

Segmento Azul Oscuro

Este segmento tiene 76.137 (33%) clientes, con una frecuencia media de 1, una recencia mínima de 202 días y 322 días de media y un gasto medio de 306€ por viaje. Dado que sólo ha comprado una vez y su recencia es tan alta, a este segmento pertenecerán los clientes *Desvinculados*.

Segmento Azul Claro

Este es el segmento más denso de los 4 que tenemos. Contiene al 59% de los clientes y, al igual que los clientes Desvinculados, sólo tienen una frecuencia de 1. Sin embargo, la recencia media de estos clientes es 4 veces menor. Son clientes que han realizado su primera compra hace poco. Los llamaremos clientes *Nuevos*.

Segmento Amarillo

A este segmento pertenecen aquellos clientes con hasta frecuencia 5 y una recencia media de 172 días. Sin embargo, a diferencia de los otros dos segmentos, estos clientes tienen un gasto medio muy superior por viaje, este asciende a 2.456 €. A estos clientes los denominaremos *Turistas* ya que parece que viajan en determinadas épocas del año (vacaciones).

Segmento Rojo

Llama la atención en este segmento la elevada frecuencia media que tienen estos clientes, pues esta es de 6 viajes en el período estudiado. Además, tienen la menor recencia media de todos los clientes, 60 días y un gasto medio de 534€. Estos serán los clientes *Habituales*.

Segmento	Nº clientes	Frecuencia mínima	Frecuencia máxima	Frecuencia media	Recencia mínima	Recencia máxima	Recencia media	Money mínimo	Money máximo	Money medio
Azul oscuro	76.137	1	3	1	202	461	322	0	1.732	306
Azul claro	136.614	1	3	1	0	233	88	0	1.518	338
Amarillo	13.606	1	5	1	0	461	172	1.318	14.427	2.456
Rojo	5.616	4	27	6	0	409	60	0	4.317	534

Clasificación de usuarios

Para la estimación de valores futuros de los clientes como se ha comentado en el apartado anterior se han empleado dos modelos de clasificación distintos ambos basados en los árboles de decisión.

Las variables independientes utilizadas han sido: edad, género, distancia recorrida, antelación en la compra del billete, tipo de cliente (Business, Economy, Premium), pasajero con tarjeta de fidelización, tipo de vuelo (Internacional o Nacional) y canal de compra. La variable dependiente ha sido calculada a partir de los percentiles de la variable money obtenida en el modelo RFM.

Antes de estimar los modelos se ha realizado un análisis exploratorio a partir de gráficos que muestran la distribución de cada una de las variables independientes frente a la pertenencia de cada uno de los clusters resultantes del K-means. Los resultados más relevantes muestran que no hay diferencias significativas en cuanto a género y sexo, por lo que seguramente, estas dos variables no tengan un peso muy significativo en los modelos de clasificación. Donde sí se aprecian distribuciones distintas es en el tipo de cliente, en el canal de compra, en el tipo de vuelo, en la distancia y en la antelación de compra, siendo especialmente relevante en estas dos últimas.

Tanto para el árbol de decisión como para el random forest se muestra un ranking de la importancia de cada una de las variables en la clasificación. Tanto para el arboles de decisión como para el Random Forest la principal variable para clasificar a un pasajero en un segmento u otro es si vuela en Business o no, la distancia, la antelación el compra, si utiliza el canal de la aerolínea para la reserva del vuelo. El resto de variables no son relevantes. Por lo que a la hora de estimar cuanto se gasta cada pasajero hay que prestar atención es si viaje en Business, Economy o Premium, el canal que utiliza para la compra, los días de antelación de la compra y la distancia, es decir, el trayecto que escoge.

Customer Lifetime Value

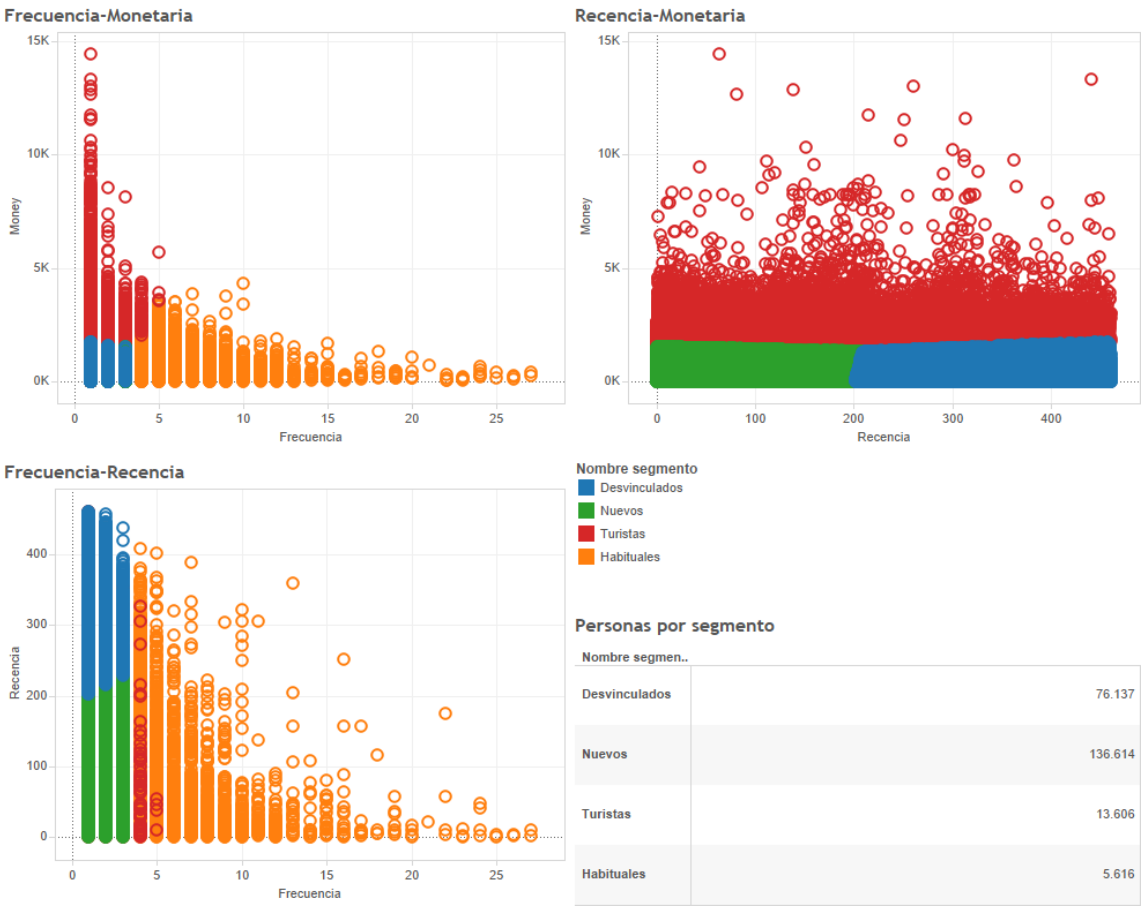
En el caso del Lifetime Value como se ha mencionado en la sección de metodología se ha estimado tanto la probabilidad de que el pasajero vuelva a realizar una comprar como una predicción del valor monetaria de una futura compra.

Vemos que claramente los usuarios que cuentan menor recencia y mayor frecuencia son los que tienen una probabilidad más alta en la repetición de una futura compra. Sin embargo, parece que el modelo del BG/NGB no es tan bueno en la estimación del número de transacciones en el periodo siguiente. Se ajusta bien en predecir un número de transacciones de 1 y 3 pero no tanto cuando el número de transacciones reales es 2 y 4. En cuanto a la predicción del valor monetario medio de futuras compras vemos que el modelo tiene más precisión

Visualización de datos – Tableau

Aprovechando la base de datos creada en PostgreSQL, conectamos Tableau de forma que podemos crear los reportes directamente sin necesidad de exportar los datos a un archivo Excel.

La primera representación que hacemos es la visualización de la segmentación obtenida tras lanzar K-means en Python. De esta manera podemos ver claramente cómo se relacionan los tres ejes de la segmentación RFM.



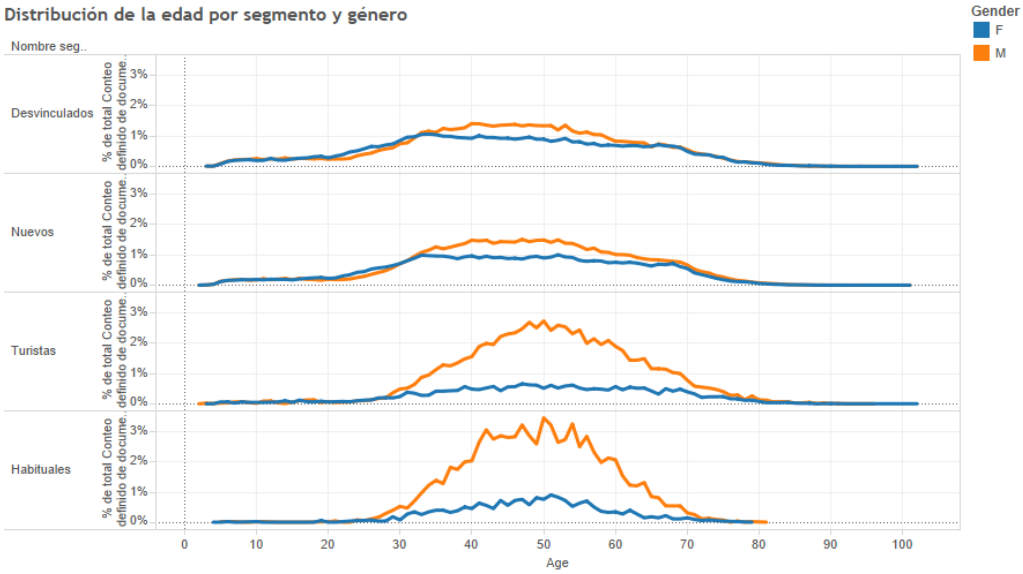
De esta manera, podemos ver que forman parte del segmento azul aquellos clientes que han comprado 3 o menos veces en el período estudiado y que, además hace más de 7 meses que no han comprado un billete. Estos son clientes Desvinculados.

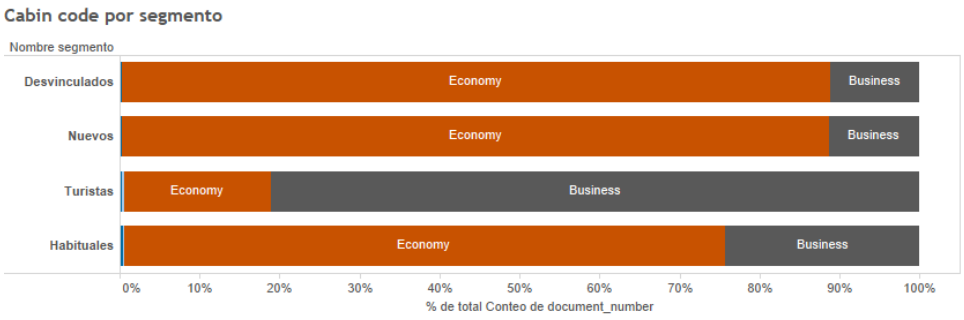
Los clientes Nuevos (verde) son aquellos con una recencia inferior a 7 meses.

Los clientes del segmento rojo tienen un amplio rango de gastos, con una frecuencia de compra entre 4 y 5 veces y la recencia abarca todo el rango que disponemos. Por lo que estos son clientes que viajan por Turismo ya que viajan en determinados momentos del año (vacaciones).

Los clientes Habituales (amarillo) son aquellos con una frecuencia muy superior al resto de clientes.

Pasamos a analizar la distribución de los clientes según el segmento al que pertenecen:





Preparamos una tabla en PostgreSQL que, para cada cliente y viaje, nos indica la duración del mismo (días de diferencia entre la ida y la vuelta) y el día de la semana que se realizó cada viaje.

Generamos un dashboard interactivo que nos muestra la frecuencia de vuelo para cada día de la semana y segmento. Además se puede ver la distribución de los días de duración de los viajes de los clientes.

El dashboard permite filtrar por día de la semana del trayecto de ida y de vuelta, la duración de los viajes o el segmento de los clientes:

Por último, a nivel global, generamos un dashboard que muestra los itinerarios entre los aeropuertos de origen y destino.

Se puede filtrar por la distancia recorrida durante el vuelo y los aeropuertos de origen y destino deseados:

Distancia del vuelo
375,801946131 to 37.197,799799998

Aeropuerto de origen
BRU

376 37.198

Origenes y destinos de los vuelos

Aeropuerto de destino

- [x] ALC
- [x] ALG
- [x] AMM
- [x] ATL
- [x] BDS
- [x] BES
- [x] BEW
- [x] BHX
- [x] BLQ
- [x] BOO
- [x] BRE
- [x] BRU
- [x] BVA
- [x] CBG
- [x] CIA
- [x] CUF
- [x] CZM
- [x] DAY
- [x] DEN
- [x] DOA
- [x] DTM
- [x] DYG