Customer Journey-Based Segmentation for Marketplaces -- Ebay

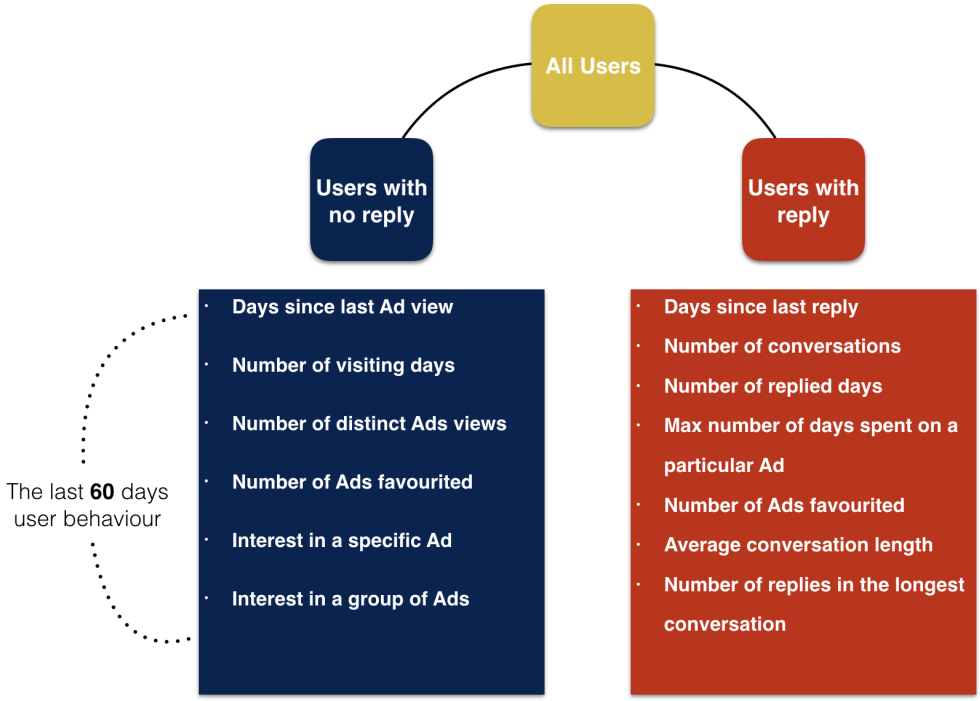
TLDR

- 我們希望根據用戶的當前客戶旅程階段將他們分為較小的組。因此，要回答的主要問題是：“我們能否在我們的市場平台上清楚地確定完整用戶體驗的有意義的旅程階段？” 這樣，我們將能夠定義一些與我們平台互動程度不同的用戶集群。
- 主要目的是什麼？
    - 該項目的最終目標是能夠根據平台上的當前狀態以不同的方式針對我們的用戶。另一種說法是：“個性化定位”。在此模型的幫助下，我們將有機會相應地區分客戶定位策略，以更有效地利用用戶並增加平台上的活躍用戶數量。
- 該分析針對哪個用戶組？C2C
    - 我們在這一階段問自己的問題是我們應該去買還是賣。買主是在平台上尋找要購買產品的人，而賣主是通過發布新的商品清單來出售他們的東西。由於買家在市場平台上創建的用戶操作（例如查看商品，保存商品或向用戶發送消息）的數量要比賣家多得多，因此，我們決定僅對買家運行此客戶旅程分析。也有一種方法可以對賣方進行類似的分析，但是目前，我們僅出於簡單和高效的考慮而堅持與買方合作。因此，在本文的其餘部分中，有時我們會使用“買方旅程”一詞來指代該問題。
    - 在市場平台上，主要有2種不同的用戶類型：B2C（經銷商）和C2C（個人）
        - 經銷商是發佈在我們平台上的廣告的主要來源，占我們賣方的很大一部分。因此，我們決定從此分析中排除所有經銷商，因為將他們分配到那些買方旅程階段中沒有任何意義。為了保護分析免受經銷商數據的誤導影響，我們僅對C2C用戶集中並進行整個分析。
- 該模型是針對哪個eCG的市場品牌開發的？

    - Ebay分類集團（eCG）是一家傘形公司，目前在全球範圍內運行14個不同的分類平台。eCG品牌有兩種不同的平台類型：水平和垂直市場。橫向市場意味著平台由許多不同的類別組成，例如我們的荷蘭租戶“ Marktplaats”。而eCG中的Vertical是指用於特定產品類別的平台：汽車。該項目是為eCG的一個垂直平台開發的：Kijiji Autos，這是加拿大領先的市場品牌，人們可以在這裡買賣汽車。

資料探索客戶旅程概述

- 在我們了解用戶交易或購買的電子商務平台（例如eBay）中，客戶旅程主要從註冊到購買產品結束。因此，這些流程通常被這些企業稱為“購買渠道”或“轉化渠道”。但是，就我們的機密業務而言，定義並不那麼明確。因此，我們需要更精細地解決問題，尤其是在項目的第一步中：定義因素—特徵—在項目的後續步驟中將用於形成不同的客戶旅程階段。
- 在“探索步驟”期間，我們發現市場平台上的C2C用戶主要有兩個不同的組：查看者和復制者。
    - **查看者**：最近未回复列表但僅在我們的平台上瀏覽的用戶。
    - **回复者**：最近回復了至少1個列表的用戶。
    - 我們首先將所有用戶劃分為兩個用戶子組的主要原因是它們表現出不同的客戶行為特徵。對於查看者，我們只能在平台上跟踪他們的瀏覽操作，而對於回覆者我們更喜歡將重點放在他們過去的消息傳遞操作上
- 以下是分別為查看者和復制者提取的特徵列表：

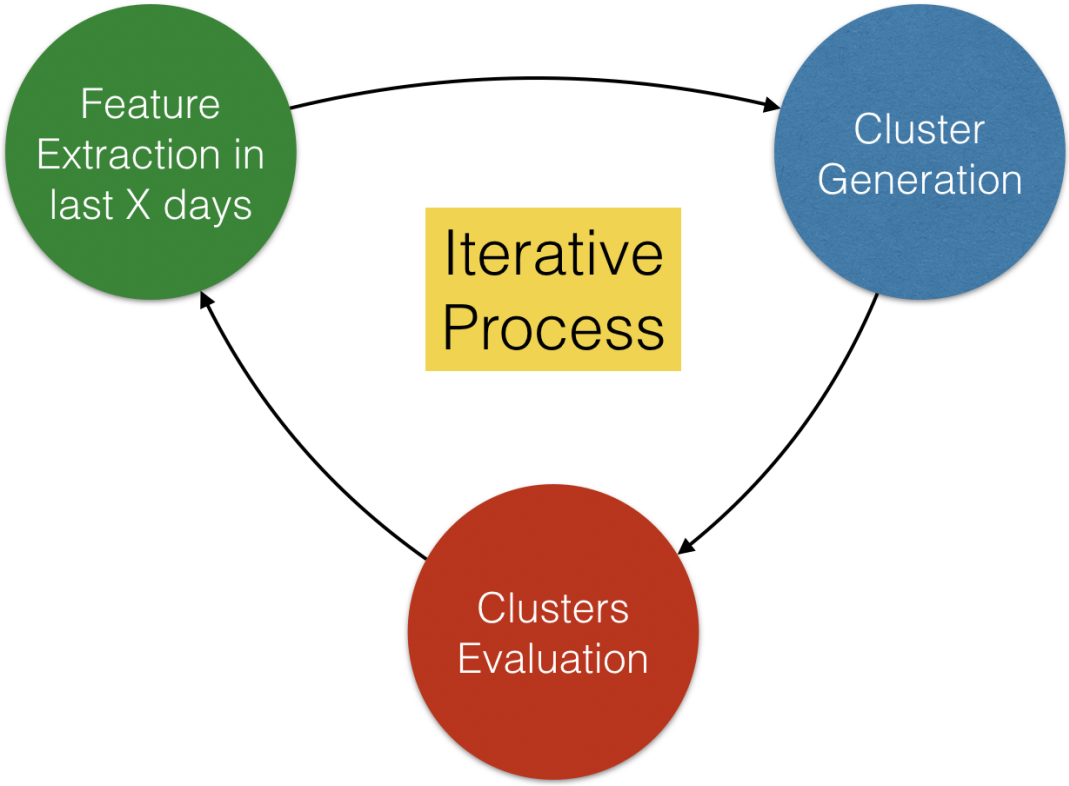


- Viewer
  - 自上次廣告視圖以來的天數：自上次用戶在平台上查看列表以來的天數。
  - 訪問天數：用戶訪問平台的不同天數。
  - 不重複廣告觀看次數：用戶觀看過的不同列表的次數
  - 收藏的廣告數量：用戶保存的列表數量
  - 對特定廣告的興趣：用戶訪問的觀看次數最多的列表所佔天數與該用戶訪問的總天數之比。
  - 對一組廣告的興趣：用戶查看的不重複列表數量與該用戶查看的列表總數之比。
- Replier
  - 自上次回復以來的天數：自上次用戶在平台上回复賣家以來的天數。
  - 對話次數：與不同賣家的對話次數。
  - 回复天數：平台上用戶回复賣家的天數。
  - 在特定廣告上花費的最大天數：用戶對同一商品的賣家回答的最大天數。
  - 收藏的廣告數量：用戶保存的列表數量

    - 平均對話時間：用戶的總回複數與該用戶進行的對話數之比。
    - 最長對話中的回複數：最長對話中發送給賣方的郵件數

確定分析的時間段

- 在從數據中提取任何用戶行為特徵之前，我們需要為此過程設置一個時間範圍。這是因為只有相對較新的用戶操作才能對用戶當前的客戶旅程階段產生影響。從上圖可以看出，此案例的時間段為60天，這是迭代過程的結果，如下圖所示。



- 整個過程從從用戶在平台上最近X天的操作中提取特徵開始，然後我們相應地生成集群，並評估最後一步中這些最終集群的質量。這個迭代過程一直持續到我們獲得最佳歷史時間段（在本例中為60天）。

方法

- 由於我們的數據中沒有任何預定義的用戶旅程階段，因此我們將細分問題歸結為無監督學習：聚類。
    - 我們僅了解我們的用戶過去在平台上的操作，但是此分析仍將涉及多少有意義的客戶旅程階段（集群）。實際上，這既是聚類分析的優點，也是挑戰。您會發現一些新事物，這些新事物可能會為您的業務提供有用的見解，但同時您也必須解決其中有許多未知數的問題。
- 考慮演算法背後的假設
    - 儘管大多數數據科學家可能會認為聚類分析沒有挑戰性，但實際上，如果您遺漏了幾個關鍵點，確實很容易犯很多基本錯誤。無論您使用哪種算法，都應該考慮集群算法背後的一些假設。由於我們選擇了“ K-Means”算法來解決我們的聚類問題，因此我將專門研究該算法的假設，並在本文的其餘部分中討論如何一一處理它們。
- 最後特徵提取→SparkSQL和Spark的DataFrame API。
    - 遵循步驟→通過Sparkling Water軟件包在Spark上運行不同的H2O算法。如果您還沒有嘗試過在Spark上使用H2O，則應該給它一個機會，因為我相信一旦您意識到這些H2O算法與SparkML軟件包中的H2O算法相比，它就會變得很討人喜歡。

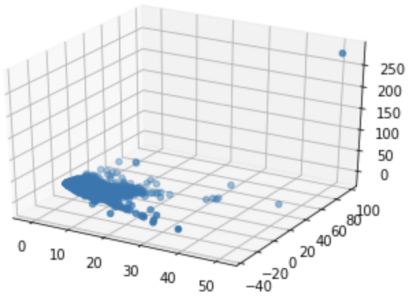
第一步：Isolation forest 移除離群值

- isolation forest算法已用於從數據中刪除異常值的目的。討論算法的細節不在本文討論範圍之內，但是我們想簡要地提及其背後的邏輯。
    - 隔離林基本上是具有隨機分割的隨機森林，而不是每個節點分割選擇最佳特徵候選者。在我們的案例中，這是一種無監

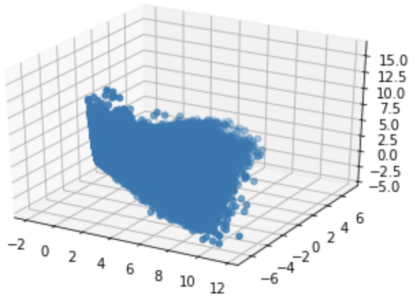
督的異常檢測方法。首先要構建多個決策樹，以便樹將葉子中的觀察結果隔離開來。由於離群值在某些特徵上具有極高的價值，因此與其他觀測值相比，離群值的隔離將更容易，更快捷。換句話說，它們在樹上的分支往往比其他的要短。因此，通過對每個觀測取樹上平均節點分裂數，我們得出分數，其中觀測需要的分裂數越少，異常的可能性就越大。與“隨機森林”相似，我們還有一些參數，例如樹木數量，樹木高度或要調整的採樣率。但是，我們還有一個針對“隔離森林”的額外參數，用於指示數據中異常值的總體比率：“污染率”。這確定了分數的分界點，以宣布觀察結果是否異常。

- 在下面的圖中，您將看到異常值移除過程對數據空間中觀察值分佈的影響。我們的數據空間是多維的，但是為了進行可視化，我們應用主成分分析（PCA）將維數減少為3，從而使我們能夠在笛卡爾坐標系中顯示數據點。在從數據中消除那些離群值之前，幾乎所有數據點都彼此緊密靠近，因此不適合進行聚類分析。但是，如您從右側圖可以看到的那樣，在過程觀察之後，坐標觀察更加均勻或更不密集。

PCA before outlier removal



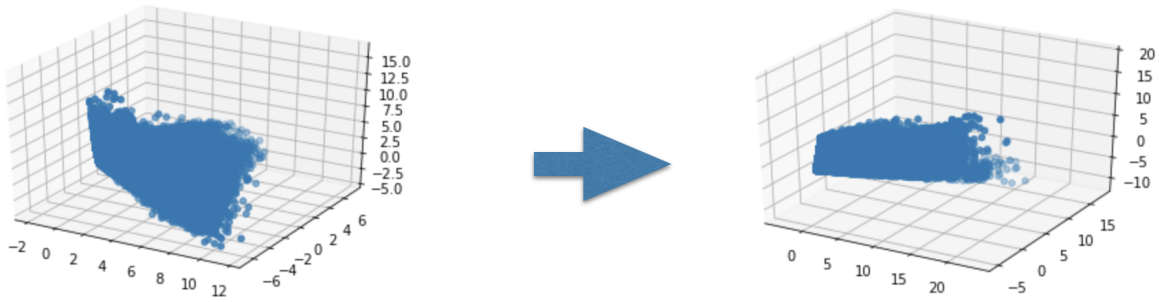
PCA after outlier removal



- 在解決任何集群問題時，在製定方法時應始終考慮到商業假設。對於聚類算法而言，事情並非總是那麼簡單，您可能需要根據自己的需求和未來的假設進行調整。K-Means算法的設計方式是，在生成聚類時所有維度均具有同等重要性，但並非一直如此。在我們的案例中，我們希望對某些維度賦予更多權重，以增加其重

要性以及對集群形成的影響。

- 對於觀看者，我們將以下尺寸的權重加倍：
  - 自上次廣告瀏覽以來的天數
  - 參觀天數
  - 不重複廣告觀看次數
- 對於復制器，我們將以下所列尺寸的權重加倍：
  - 自上次回復以來的天數
  - 對話次數
  - 最長對話中的回複數
- 在下面的曲線圖中，您可以看到在對維度應用不同權重的步驟之前和之後，二維坐標系中觀測值分佈之間的差異。顯然，該步驟對團簇的形成有很大影響。

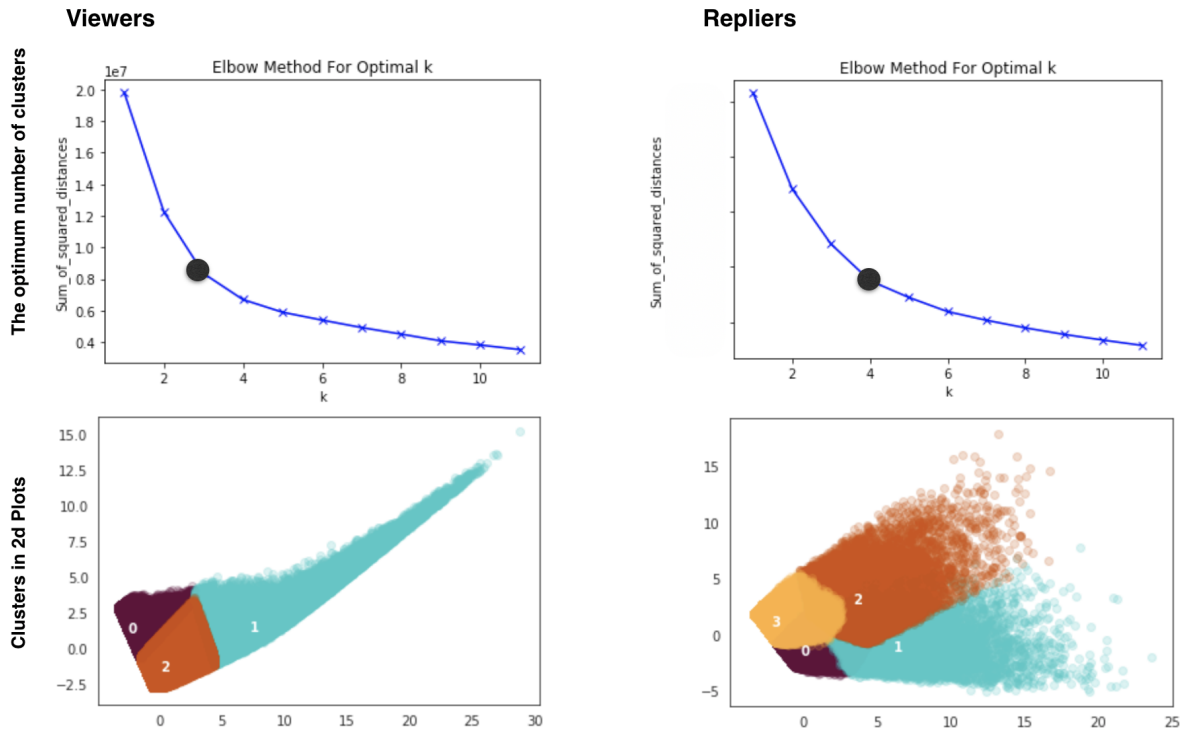


Modeling

- 在得出加權維數之後，下一步是在數據頂部應用k-means算法。由於我們對客戶旅程的可能階段還一無所知，因此在這一點上，我們應該確定最佳的集群數量，換句話說，就是確定不同的旅程階段。尋找最佳簇數的方法是肘分析。我們將K-Means算法用於一堆不同數量的聚類，並通過查看平方距離的聚類間和來評估那些聚類的質量。

我們針對兩個主要用戶組分別運行此分析：Vierwer & responser，如下圖所示。在第一個情節中，我們最終得到了3個類集-肘功能提示的Viewers旅程階段。然而，如右圖所示，根據對複制器的肘部分析，我

們總共有4個簇。



- 在找到兩個用戶組（查看者和復制者）的最佳群集數量後，我們根據群集質心的不同維值為每個群集賦予了有意義的代表名稱。
    - 首先，讓我們仔細看一下下表中顯示的群集的質心。我們通過解釋不同維度上的簇質心的值來得出簇的名稱。我們通過在表格中的那些單元格周圍放置紅色邊框來指示群集中決定性的尺寸。在檢查了質心之後，強烈建議您閱讀以下部分中對群集的簡短描述和指示，以更清楚地了解這些群集名稱背後的基本原理

Users with no reply

	Days Since Last Ad View	Number of Visiting Days	Number of Distinct Ad Views	Number of Ads favourited	Interest In a Specific Ad	Interest In a Group of Ads
Churn	42	2	7	0	0.8	0.9
Browser	10	4	17	1	0.55	0.9
Prospective Replier	5	20	170	10	0.2	0.75

Users with reply

	Days Since Last Reply	Number of Conversations	Number of Replied Days	Average Conversation Length	Max number of days spent on a particular Ad	Number of Ads Favourited	Number of replies in the longest conversation
Losing Interest	43	3	2	2	1.5	13	4
Vanilla Buyer	12	5	4	2	2	20	5
Prospective Buyer	8	51	18	2.5	3.5	47	18
Bought	20	9	7	4	7	23	39

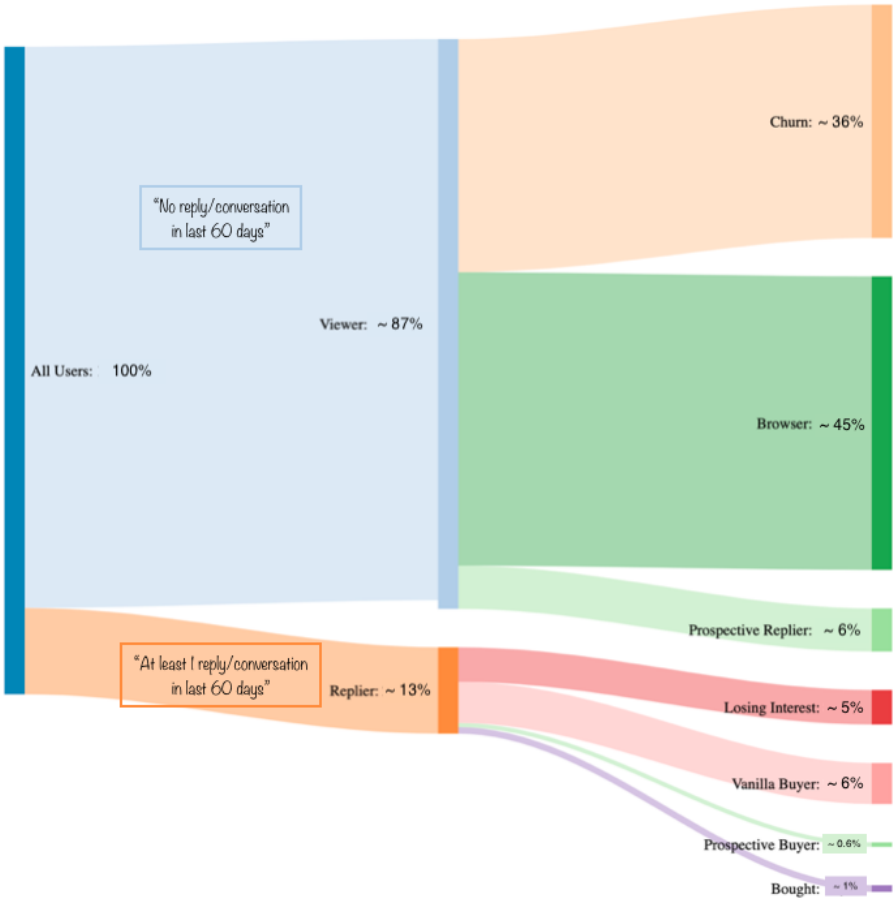
Viewer群集的簡短說明和指示符

- 1-Churn：他們很久以前就停止訪問我們的平台。
- 觀看者中的“自上次廣告觀看以來的天數”
- 2-Browser：他們仍然訪問我們的平台，但不那麼活躍。
    - 通常，任何維度上表現都沒有那麼明顯
- 3-To-conversion：目前，他們積極訪問我們的平台。這組用戶最有可能開始與賣家進行對話。
    - 觀看者中的最少“自上次廣告觀看以來的天數”
    - 觀看者之間的最大“不重複廣告觀看次數”
    - 觀看者中最大的“收藏的廣告數量”

中繼器群集的簡短說明和指示符

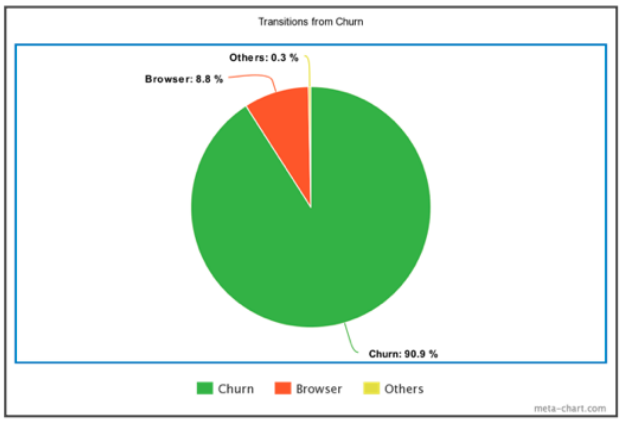
- 1-失去興趣：不久前，他們停止回复列表。他們很可能已經放棄了在平台上尋找汽車的搜索。
    - 回复者之間的最長“自上次回復以來的天數”
- 2-香草買家：目前，他們仍在與賣家進行一些對話，但並不那麼活躍。
    - 通常，任何尺寸都沒有極值

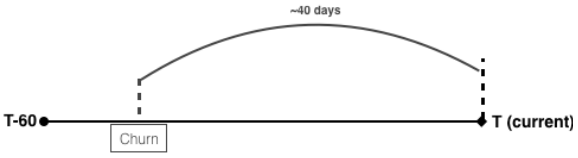
- 3-準買家：目前，他們正在積極地與不同的賣家進行對話。這組用戶最有可能購買汽車。
    - 回复者之間的最少“自上次回復以來的天數”
    - 複製者之間的最大“會話數”
    - 回复者之間的最大“回复天數”
    - 最高“收藏的廣告數量”
- 4-購買：最近，他們已經買了車，離開了我們的平台。
    - 複製方之間的最大“平均會話長度”
    - 重複者之間的最大“在特定廣告上花費的最大天數”
    - Repliers 之間最大的“最長對話長度”



在生成集群之後，我們還想知道7天之內集群之間的轉換。要回答的問題是，在接下來的7天中，每個集群中有多少百分比的用戶遷移到任何其他集群。換句話說，我們測量了趨向於保持穩定狀態的集群數量。創建這些過渡的另一個原因是對集群生成過程進行完整性檢查，以查看集

群之間是否發生任何意外的或不可思議的過渡。在下面的餅圖中，您將看到在7天內從該特定群集到其他群集的轉換百分比。





在本文中，我們向您介紹了我們在項目期間採取的步驟：基於客戶旅程的市場細分。首先，我們從定義尺寸（特徵）開始，在整個分析過程中都使用了尺寸。然後，我們談到了k均值聚類的一些關鍵假設，並討論了在我們的案例中我們如何處理這些問題。在那之後，我們還提到了我們如何通過解釋集群質心來提出這些集群名稱的。在本文的最後，我們深入研究了群集的詳細信息以及群集之間如何形成過渡。

本文的主要結論是解決聚類問題可能很棘手，因此，在進行分析時，您最好也要牢記一些關鍵點。我們希望您喜歡閱讀本文並發現它有用。