# 1.Abstract

在數據驅動解決方案時代，客戶人口統計屬性（例如性別和年齡）起著核心作用，可以使公司增強服務質量並在正確的時間和地點定位合適的客戶。在市場營銷活動中，公司希望針對GSM（全球移動通信系統）的真正用戶而不是線路所有者。有時它們可能並不相同。這項工作提出了一種根據用戶的行為，服務和合同信息預測其性別和年齡的方法。我們使用呼叫詳細記錄（CDR），客戶關係管理（CRM）和計費信息作為數據源來分析電信客戶的行為，並應用不同類型的機器學習算法為營銷活動提供有關客戶人口統計屬性的更準確信息。該模型使用敘利亞電信公司提供的18,000個用戶的可靠數據集進行構建，用於培訓和測試。該模型通過使用大數據技術應用，在用戶性別預測方面達到了85.6％的準確性，在用戶年齡預測方面達到了65.5％的準確性。這項工作的主要貢獻在於，基於移動電話數據和端到端解決方案的用戶性別預測和用戶年齡預測的準確性得到了提高，該解決方案從電信領域的多個方面著手處理客戶數據。

# 2.Introduction

如今，手機已成為發展中國家發展最快的技術之一，全球普及率達到90％[1]。這使其成為存儲客戶數據的巨大倉庫。也就是說，客戶採取的每項操作（短消息服務（SMS），呼叫或Internet會話）都以所謂的（CDR）記錄在電信運營商內部。電信計費系統主要使用多種類型的CDR。

CDR包含許多信息（事件的類型，此事件涉及的人，日期時間，發生此事件的單元格標識符）。這些原始數據代表了分析人類和社會行為的寶貴來源[2]。在農業領域[3]，通過創建人口和細分的流動性概況，手機數據被用於分析與塞內加爾生計區有關的流動性和季節性活動模式。

在能源領域[4]時，此數據用於分析人類活動，促進農村地區人口增長估算並推斷電力需求。在衛生部門[5，6]中，使用手機數據來研究人類移動性與疾病流行率之間的關係。這些數據可用於分析人類行為併計算心理諮詢指標，以預測客戶的個性[7]。電信行業是眾多挑戰的沃土，這些挑戰使移動運營商受益，以改善他們在不同領域的業務和競爭優勢。電信領域有兩種主要的大數據使用案例：

## 2.1Network improvement

運營商一直關注網絡性能的提高。

使用大數據分析的結果是，運營商可以識別故障並結合根本原因分析[8]，提高體驗質量（QoE）[9]，執行實時故障排除並修復網絡性能問題。

各種自組織網絡（SON）自動化（例如，配置，配置和調試）可以根據大數據分析的敏銳度來方便流量請求和環境變化[10]。它允許操作員對警報進行優先級排序，這對於節省時間和防止服務故障非常有用[11]。

## 2.2Marketing, sales and customer loyalty

市場營銷和銷售被認為是電信行業中大數據使用的最大領域。適當的大數據分析使運營商可以基於客戶分析和細分[12]來創建更智能的營銷活動，並進行銷售分析以提高銷售額。它還可以用於從營銷促銷中獲得更好的結果，增加收入並實施地理營銷和實時營銷。

例如，Globe Telecom（菲律賓電信公司）使用大數據分析將促銷效果提高了600％[13]。

吸引新客戶的成本要比保持現有客戶高得多，因此，流失預測和管理已成為移動服務提供商關注的問題。希望保留其訂戶的移動服務提供商需要能夠預測其中的哪一位可能會改變運營商的風險[14]。

在這項研究中，我們提出了解決電信運營商中實際問題的解決方案。問題是，電信運營商有時會遭受其客戶的不可靠人口統計數據的困擾。這項研究介紹了一種解決方案，該解決方案使用了不同的領域，例如大數據科學，電信，用於性別和年齡預測的社會策略，以及機器學習方法和結果的比較。我們使用的端到端解決方案從以Web應用程序作為用戶界面關閉的數據採集開始，並通過信息圖表可視化（例如地圖）和通過查詢構建器界面進行的交互式查詢數據，包括所有相關數據處理，例如提取，加載和轉換（ELT）流程。

本文的其餘部分結構如下：在“相關工作”部分中，我們介紹有關用戶人口統計預測的相關工作。“方法”部分描述了我們在這項研究中使用的數據集，以及特徵提取方法和大數據生命週期。

在“結果與討論”部分，我們描述了我們在性別，年齡方面的研究結果，並在實際案例研究中評估了我們的方法，同時我們提出了擬議的框架。最後，我們總結我們的工作並在“結論”部分描述未來的工作。

# 3.Related Work

一些著作討論了使用不同方法在不同領域中的性別和年齡標識，例如：在[15]中，twitter數據用於根據用戶的名字或[16]中的twitter圖片或文本來預測用戶人口統計挖掘與分析[17]。而基於瀏覽行為創建的特徵可以預測用戶的人口統計數據[18]。但是，移動用戶的數量超過了社交網絡用戶的數量，因此許多研究都討論了基於用戶的智能手機應用程序預測用戶人口統計信息的方法[19，20]。根據移動電話數據預測用戶人口統計信息的大多數工作（也稱為CDR）都依賴大型訓練集來預測人口統計數據。Felbo [2]解決了機器學習中的高級方法，並使用深度學習算法來從其在大尺寸特徵中的效率中受益，並讓這些算法通過特徵工程代替手工工程來完成其工作。但是，使用大數據集來訓練模型，也就是說，該數據集依賴於15萬名客戶，包含超過2.5億條記錄，識別性別的準確度為79％，年齡的準確度為63.1％。Sarraute等。[21]使用包含500萬客戶的數據集來提取用戶的行為和社交網絡特徵，並使用主成分分析（PCA）方法選擇重要特徵。預測性別的最高準確度為81.4％，預測年齡為62.3％。[22]通過研究呼叫網絡和消息傳遞網絡，並使用雙相關因子圖模型來預測手機社交圖譜中的性別和年齡，從而依靠客戶的社交特徵（程度中心性，三合會封閉...）。

結果是：使用F1-Measure可以預測性別為0.8063，預測年齡為0.7132。Martinez使用了10,000個用戶的小規模可靠數據集，並且通過使用多種算法（SVM，Random Forest和K-means），當減少了預測實例的百分比時，獲得的準確性為80％[23]。在[1]中，bandicoot1工具用於提取1400多種具有不同類別的行為特徵，並使用不同算法（例如隨機森林，SVM，KNN）對這些特徵進行了測試，該模型的預測準確性最高為79.7％在預測南亞發展中國家的性別。

在這項工作中，我們集中精力為敘利亞社會提取合適的專用功能，我們從多種資源（如客戶服務和合同）中提取功能。

# 4.Methods

本節描述了這項工作中使用的數據集，以及特徵提取方法和大數據生命週期。

## 4.1Data description and preparation

問題理解和數據理解階段有助於我們確定數據源，並定義重要的數據源以提取相關特徵。Data Lake被開發為所有數據的單點。我們的預測模型選擇了5個數據源：

1.CDRs

CDR數據包含客戶及其所有屬性採取的所有操作。網絡交換機中的數據收集器以CDR形式捕獲使用情況。這些原始CDR進而由中介系統轉換為計費系統可以理解的格式。表1顯示了CDR字段的示例：

2.Customer’s services

客戶註冊的所有服務均已根據服務類型（例如與政治新聞，體育新聞和星座運勢等相關的服務）進行了手動收集和分類，這些類別被視為客戶功能。結果，我們得到了客戶服務表。表2是一個示例。

3.Customers’ contract information

客戶合同信息已從CRM系統中獲取，其中包含有關客戶的基本信息（性別，年齡，位置...）以及客戶訂閱的信息，因為一位客戶可能擁有多個訂閱（兩個或更多GSM））具有不同的訂閱類型：預付費，後付費，3G，4G ...訂閱。

4.Cells and sites database

電信公司站點相關的數據，其組件和運營商存儲在關係數據庫中。該數據用於提取空間特徵。表3說明了此數據源的匿名數據的形狀。

5.Reliable dataset

建立這樣的預測系統需要一個樣本，其中包含該樣本中每個gsm的真實人口統計信息，例如性別和年齡，無論gsm所有者的人口統計信息是什麼，因為有時實際用戶和gsm所有者並不相同。此可靠的樣本用於監督學習算法的階段培訓和測試，但是遵循定向方法，在大約6個月內隨機收集了大約18,000個客戶的數據。

用於構建模型的樣本數據包含64％的男性和36％的女性。數據已根據年齡分為幾組。每個年齡段代表不同的年齡階段。在進行模型訓練後，系統將預測客戶的年齡組。表4顯示了年齡組以及樣本中該類別的百分比密度：

## 4.2Feature extraction

基於我們的搜索和猜測，完成了特徵工程和提取的工作。因此在內存處理工具中用於處理和分析目的。但是，為每個客戶提取了220個功能。這些特徵屬於6個特徵類別，每個類別都提供了示例（請注意，有些特徵屬於多個類別）。

1.Behavioral individual features

每天的平均呼叫持續時間，持續時間的熵，白天每天的出站呼叫持續時間的標準偏差，晚上每天每天的入站呼叫持續時間的平均值，每天在工作時間收到的SMS的標準偏差等。其數量約為（200特徵）。

2.Social behavior features

聯繫人數量，聯繫人熵，每個客戶收到的平均交易，每個聯繫人已發送交易的標準差等。（20個功能）。

3.Spatial and mobility features

移動性的平均值，假日中的移動性平均值，家庭區域類型，夜間天線熵，白天天線熵，工作日天線熵，假日天線熵等。它們的數量大約為（21個特徵）。

4.Temporal features

每個工作日（週日至週四）和節假日，或者每天（9–16）和晚上：在假期等期間，每天（17–8）平均收到的SMS數量（165個功能）。

5.Types of Services

參與：技術新聞服務，教育服務，體育新聞服務，政治新聞服務，娛樂服務等，（13個功能）。

6.Contract information tariff type, GSM type

提到的特徵總數為421，但是大約有201個特徵屬於一個以上的類別，因此特徵總數為220。

特徵提取是將原始數據轉換為適合建模的派生值。用於性別和年齡模型的所有提取特徵都屬於兩種特徵之一，即統計特徵或分類特徵。我們用來提取統計特徵的統計函數是：

1.Probability

2.Standard deviation

3.Percentage

4.Entropy

以上是統計特徵，接下來是分類特徵：

分類特徵：可以採用有限且固定數量的可能值之一，此類特徵的數量為15。例如，“具有經濟新聞服務”特徵的取值為0或1。

合同數據源的人口統計特徵（例如性別和年齡）被排除在外，因為它們不可靠，我們需要根據行為來預測客戶的性別和年齡。

這些特徵是分類算法的輸入。但是，在描述性分析階段研究了性別和年齡之間的特徵分佈，以更深入地了解客戶的行為。圖1和圖2示出了關於那些特徵的統計研究的樣本。

作為統計研究，某些特徵值隨男性和女性以及年齡組而異。

## 4.3Big Data Life Cycle

這項工作包括不同的階段（數據探索，特徵提取和選擇以及模型驗證），但是使用了不同的機器學習方法：

* 無監督學習用於檢測異常值和降維，例如k均值聚類和PCA。
* 監督學習主要用於預測客戶年齡和性別的分類方法，因此測試了12種分類算法。

我們已經測試了大約12種最受歡迎的分類算法。關於數據獲取和ELT流程，我們使用了多種工具（Apache Flume，2 Apache Nifi，3 Apache sqoop4）來移動數據，這些數據具有不同的來源和類型（結構化，半結構化，流式處理和批處理）。數據源和數據湖到大數據平台。數據以Apache Parquet存儲格式（具有Snappy壓縮）的形式存儲在hadoop分佈式文件系統（HDFS）[24]中，以進行有效的存儲和處理[25]。Apache Spark [26]處理了數十億個CDR記錄中的特徵提取，並將結果存儲在Hadoop數據庫HBASE5中，以便從Web應用程序中檢索。

實現的框架代表了數據生命週期階段和大數據管道，如圖3所示。

# 5. Results and discussion

分析數據集以通過提取的特徵識別客戶類型，使用k-Means聚類檢測到異常行為的客戶，例如平均每天有超過120分鐘的通話或平均有60個以上的通話的客戶每天（後來，我們知道他們的工作類型可以解釋他們的異常情況）。然後，使用提取的特徵測試了許多分類算法。R語言環境及其包（如caret和xgboost）用於預處理這些功能並進行建模。使用的預處理方法是：

* 測試了具有10和100個主要組件的PCA，儘管它由於減少了尺寸而加快了模型的執行，但是並沒有改善模型的結果。
* Z分數或標準分數：雖然它稍微改善了SVM模型，但並沒有改善我們獲得的最佳模型xgboost。

測試了以下分類算法：線性判別分析（LDA），支持向量機[（SVM（基於徑向）），極大梯度提升（XGBoost），隨機森林，邏輯回歸，GLMNET，KNN，樸素貝葉斯，CART，C5.0，梯度增強機（GBM）和袋裝CART，但是根據評估結果選擇了最佳模型。

為了進行模型訓練和驗證，將可靠的數據集分為（80％–20％）。

所有分類算法均使用10倍交叉驗證進行了訓練，並依靠以下指標進行模型評估：

1.Accuracy

2.Area under the curve(AUC)

3.F1-measure

這些度量用於本研究中以評估測試集上的模型。

表5和圖4示出了關於性別預測的評估結果。表6和圖5顯示了使用大數據平台（6個節點，每個節點具有16核處理器和32 GB內存）的有關年齡預測的最佳4個評估結果。

結果，諸如GBM，xgboost [28]和隨機森林之類的集成學習算法在其他分類算法上具有更多優勢，並獲得了最佳的準確性（AUC和F1測度，即F1測度中的xgboost得分0.8903，用於性別預測））。

在性別預測模型上對xgboost進行調整是（使用xgboost軟件包）：max\_ depth = 10，eta = 0.1，gamma = 0，min\_child\_weight = 0.9，lambda = 0，alpha = 0.9，nrounds = 150，子樣本= 1。

用於年齡預測的xgboost調整為：max\_depth = 20，eta = 0.1，gamma = 0，min\_child\_weight = 0.9，lambda = 0，alpha = 0.9，nrounds = 50，子樣本= 1。

在性別預測中增加用於xgboost的樹木數量超過150棵並不能提高性別模型的準確性。同樣，將樹木的數量增加到50以上並不能提高年齡模型的準確性。也就是說，在模型學習過程中，每當我們將樹添加到xgboost模型中時，都會在訓練和測試集上測試錯誤率，如果測試集上的錯誤率沒有降低，那麼即使訓練集的錯誤率持續下降，因為該模型很可能會過擬合。

Gain度量被認為是特徵重要性的度量，因此它用於檢測性別和年齡模型的信息性特徵。

Gain表示通過獲取模型中每棵樹的每個特徵的貢獻而計算出的相應特徵對模型的相對貢獻。與其他功能相比，此指標的值越高，表示生成預測就越重要。它由公式[28]計算

基於分類方法結果的比較，選擇了建議的框架，如圖6所示。

根據xgboost模型，基於增益測度的性別模型的最高預測特徵是接聽電話的持續時間熵，面向年輕女孩的服務。圖7顯示了用於性別預測的前5個信息功能的增益度量。圖8顯示了用於年齡預測的前4個信息功能的增益度量，例如所有通話時長的熵，客戶發送的平均SMS等等。

這些結果反映了我們保守社會的性質，在這種社會中，男性通常要承擔比女性更多的責任，因此男性可以處理幾種類型的聯繫（業務，家庭，朋友等）。這證明了其電信行為的熵是合理的（圖1、5）。也;圖2、7中的年齡模型表明，由於接觸類型較少，包含大學年齡年輕人的（A）群體的熵較小。年齡越大，與他們的電信行為相比，其熵越大分組（B）和（C）。通過增加對更多的人和更大的家庭的承諾，可以證明每次聯繫的平均交易量減少是合理的，這在我們國家很常見。

這項工作的局限性在於，由於遵循直接方法和此過程的人力資源限制，從隨機客戶（僅約18000個客戶）收集可靠的數據（用於培訓和測試）花費了很多時間（約6個月）。

可以通過將其擴展到包括現在尚未包括的2個新的與年齡相關的組來改進這項工作，一個組用於18歲以下的人，另一個組用於60歲以上的人，並達到更大的目標性別比例要更均衡。

此外，這項工作僅針對兩種類型的CDR（分別稱為CDR和SMS CDR）進行，由於存儲和過程限制，我們無法處理其他類型的CDR。互聯網使用CDR被認為是提取更有價值的功能的另一個數據源，如果工作擺脫了前面提到的限制，那麼可靠的數據集將更大並且更適合於深度學習算法，並且模型將更加健壯和準確。

另一個限制是，這項工作已在敘利亞社會中應用，這可能與其他社會有所不同，因此本研究中的信息功能或多或少在其他社會中很重要。

# 6.Conclusion

這項工作需要大量工作來根據電信客戶的行為，服務和合同信息來對其進行分析。這項工作的主要貢獻是端到端解決方案，該解決方案從電信領域的多個方面著手處理客戶數據。從創建大量具有不同類型（行為，社會，空間，移動性，時間，服務和合同特徵）的特徵開始。這使我們有機會深入了解客戶。然後根據統計特性研究了年齡和性別之間具有多種信息屬性的關係。

另一個研究方向是探索與客戶相關的更多屬性，並使用這些結果來創建動態報價系統，該系統具有足夠高的智能來推薦和定制個性化報價。