Using Machine Learning on Big Data for Cross-Selling Prediction & Statistics

October 2017

بحث مقدم كمطلوب تكميلي لنيل درجة بكالوريس الشرف في علوم الحاسوب

بسم الله الرحمن الرحيم

جامعة السودان للعلوم والتكنولوجيا

كلية علوم الحاسوب و تقانة المعلومات

Using Machine Learning on Big Data for Cross-Selling Prediction & Statistics

إعداد :

مازن موسى محمد ابراهيم

معاذ الصديق عيسى النور

ياسر عبدالرحمن ابراهيم محمد الامين

بحث مقدم كمطلوب تكميلي لنيل درجة بكالوريس الشرف في علوم الحاسوب

إشراف :

د.هشام عبدالله منصور

الايـة

فَإِنَّ مَعَ الْعُسْرِ يُسْرًا

سورة الشرح الايـة(٥)

الحمدلله

الحمد لك يا ربي أن جعلت لنا من كل هم فرجا،ومن كل ضيق مخرجا، وسخرت لنا من عبادك المسارعين لقضاء الحاجات ،وتفريج الكرب.

شكر و عرفان

إلى ذلك الرجل الذي جعلنا أعيينا تبصر على المجال العملاق ، وقادنا خطوة بخطوة حتى وقفنا على أشدنا ومازال ذلك المرشد الذي لم يتوانى للحظة عن مد يد العون ...إليك يا من جعلت قلوبنا مليئة بالإمتنان والتقدير لك

.....المهندس عبد الحليم محمد سر الختم

إلى تلك الشابة الطموحة التي لم تنس يوما من ذلك الدرب الذي مرت به وساعدت غيرها على عبوره بكل اخلاص وتجرد لا لشيء إلا رؤية النجاح في غيرها

.....المهندسة نداء محي الدين

ولن ننسى أبدا تلك الروح المرحة ، العالم و الأب العطوف ،الذي يكاد تواضعه وبشاشته أن تنسيك عظم مكانته و عبقريته الفذة، الذي نعده كصديق أكثر من أستاذ و مرشد

.....د.هشام عبد الله منصور

ولن ننسى تلك المرأة العظيمة ، اللتي لم تتوانى يوما عن ارشادنا و تصحيحنا،وبذل النصح عند الحاحة ، التي كانت مثالا للأم لطلابها ،وخير معلم للأجيال

.....د.وفاء فيصل

إلى أولئك الزملاء و الزميلات ، الذين اصبحوا أكثر من إخوة لنا، إلى من علمونا معاني الإيثار و التكاتف و التضحية،إليكم

راسمي البسمة على شفاهنا،وناصحينا وقت زلاتنا

.....طلاب كلية علوم الحاسوب

إهداء

إلى من ساندتني دوما ،ووثقت في حين شك الأخرون ،إلى من وقفت بجانبي في أحلك الظروف،إلى من تحملت الألام لكي تمنحني الأمل...تعجز كلمات شكري عن جزاءك حقك

.....أمي العزيزة\عائشة ابراهيم

إلى ذلك الرجل الذي لولا مساندته و توجيهاته المستمرة ،ونصائحه الجميلة دوما ما كنت لأحلم بالوصول لهذه النقطة

.....د.عز الدين موسى الامام بله

إلى أولئك الذين ضحوا بمتاع الحياة و جميلها لا لشيء إلا ليروا تقدم أبنائهم وبناتهم

.....أساتذة كلية علوم الحاسوب و تقانة المعلومات

إلى أولئك الإخوة و الأخوات ، الذين لا تقدر كلماتي وصف مقامكم إليكم أنتم راسمي بسمتي و مضيئي دربي

.....طلاب السنة الرابعة

مازن

إلى روح الوالدة العزيزة الذي أسأل الله أن يتغمدها برحمته و يحيطها بعفوه .....

إلى الأب العزيز الذي كان بمثابة الصديق و الأب و الاخ

.....الأب\الصديق عيسى النور محمد

إلى العم الوقور ، الذي عاملني دوما كأحد أعز أبناءه ، ولم يقصر في حقي ابدا

.....العم\النور عيسى النور محمد

إلى أخي الصغير الذي ساعدني كثيرا للتفرغ لواجباتي الدراسية،وكان دوما العضد و السند و الانيس

.....الأخ\محمود الصديق عيسى

إلى جميع زملائي و زميلاتي الذين هم نعم الإخوة

.....طلاب كلية علوم الحاسوب و تقانة المعلومات

معاذ

Dedicated to my uncle Abdelrazig Elaffendi may he rest in peace and may Allah remit his sins and may paradise be his last fate

To my parents, my role model, my guide in this life, whom I dream to be one of them someday; the greatest father: E. Abdelrhman Ibrahim, and the majestic mother: Dr. Alwiya Elaffendi

To my real mother the honest one, my dear grandmother: Om-Elhessin, to my wonderful family: Mohammed, Manal, Eithar, Ibrahim, Ahmed, Eman, to my uncles.

To close friends, family, FIRE BITS, and colleagues to all who had supported us by a pray or encouraging speech

ياسر

المستخلص

ان المنافسة الشرسة بين الشركات العملاقة على استقطاب الزبائن تستنزف تكاليف ضخمة ، وتجذب في الغالب ذلك النوع من الزبائن الذي لايستهلك منتجات مؤسسة واحدة لفترة طويلة،ففي الغالب تجذبه العروض الموسمية لا أكثر.

لذلك انتبهت الشركات الكبرى إلى أنها بدلا عن التركيز على استقطاب زبائن المنافسين فإنها ستكسب أرباحا أكبر عبر زيادة مبيعاتها لزبائنها الحاليين،نسبة لوجود الكثير من البيانات المخزنة لدى المؤسسة عنهم مثل أنماط الشراء و البيانات الشخصية التي قدموها بناء على موافقتهم.ولكن بعض الشركات قد أفرطت في زيادة مبيعاتها لزبائنها المخلصين للدرجة التي جعلت نظرها ينصب على تحصيل الأرباح و العمولات بدلا عن التساؤل عما إذا كان هذا المنتج المباع سيفيد الزبون أم لا.

في هذه الدراسة محاولة لإيجاد توازن بين عملية البيع للزبائن الحاليين أو ما يعرف بالـCross Selling و رضا الزبائن عن المؤسسة ومنتجاتها بشكل كلي،وذلك لتحقيق مزيد من الأرباح على المدى الطويل.

ولإيجاد هذا التوازن استخدمت أدوات تنقيب البيانات الموجودة في ضمن أنظمة معالجة البيانات الضخمة(Big Data) وذلك لحصد الفوائد المعروفة لهذه الأدوات الحديثة.مما أدى إلى الوصول إلى نتائج جيدة.

ABSTRACT

The fierce competition between big corporations for customer acquisition consumes huge sums of money, it also attracts swinging customers who don't care about long relationships or brand loyalty, whom were attracted by seasonal offers.

Corporations have found it beneficial to focus on their customer base rather than trying to acquire competition's customers, they reasoned that they will gain more revenue by increasing sales to their customers, because they have huge amount of customers' data stored, such as transactions' history and the personal information provided by customers themselves willingly. But some corporations have engaged into excessive sales practices focusing only on increasing their profit margins and collecting commissions, rather than questioning whether these sold products will be beneficial to their customers or not.

This study is a trial to make a balance between cross selling practices and customers' satisfaction with companies and their products in general, to generate more profits for the long term.

To make this balance the study will use data mining tools integrated into big data processing systems to harness its full potential. Thus aiming to reach better results.

f

قائمة المصطلحات

المصطلح	الوصف
KDD	Knowledge Discovery in Database
RFM	Recency , Frequency , Monetary
NPS	Net Promoter Score
SQL	Structural Query language
CLV	Customer Lifetime Value
HDFS	Hadoop Distributed File System
SAS	Single Attachment Station
RDD	Resilient Distributed Dataset

WSSE	Within Set of Squared Errors
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
RDBMS	Relational Database Management System
PHP	Hypertext Preprocessor
MVC	Model View Controller
MYSQL	MY Structured Query Language

الفهرس

قائمة الصور

قائمة الجداول

الباب الاول

المقدمة

المقدمة:

تحسين استراتيجيات التسويق عبر تحليل قواعد البيانات هو عبارة عن منهجية كلاسيكية لزيادة المبيعات في المؤسسات،وهذه الاستراتيجية تعتبر ناجحة ، ولكنها لا تحسن المبيعات بالصورة المثلى.

في البدء كان الموظفون يقومون بمقارنة جداول البيانات تقليديا محاولين العثور على نمط ما بإستخدام الطرق الاحصائية التقليدية ،ثم أتت قواعد البيانات مع أدوات تحليلية مبسطة[1] ،ثم ظهرت خوارزميات استكشاف المعرفة داخل قواعد البيانات التي مكنت المدراء من اكتشاف معارف جديدة عن عملياتهم.

المنهج الأخير كان كافيا حتى فترة قريبة ،وذلك عندما اكتشف المدراء أن لديهم كميات كبيرة جدا من البيانات التاريخية التي لا يعرفون كيف يستفيدون منها،اضافة إلى ذلك تدني أسعار العتاد الحوسبي عموما و ذواكر الوصول العشوائي (RAM) على وجه الخصوص[2] ،وتواصل العملاء الدائم مع مؤسساتهم ،و المنافسة المحمومة بين الشركات العملاقة ذات الحلول التقليدية المعهودة،والشركات الناشئة ذات الحلول المبتكرة.

هذا البحث ينوي الإستفادة من القوة العظيمة التي توفرها أدوات معالجة البيانات الضخمة مدمجة مع أدوات تنقيب البيانات للعثور على أنواع مختلفة من المعارف التي لم يكن من الممكن استخلاصها سابقا.

مشكلة البحث:

إن استراتيجية الـــCross Selling كجزء من العملية التسويقية تقوم بزيادة الأرباح عبر تقليل التكاليف بشكل أساسي ، و تقوم معظم ادارات المؤسسات المالية الساعية وراء زيادة الأرباح الفصلية لإرضاء مساهميها بزيادة كمية الـــCross selling دون النظر إلى عواقب الاسراف في هذه العملية،حيث يكون الأثر القصير زيادة الأرباح الفصلية ،ولكن على المدى الطويل فقدان المؤسسة جزءا مقدرا من عملائها نسبة لعدم مناسبة المنتجات المباعة لهم و تخييب أمالهم بإستمرار ، الشيء الذي يعني عدم استرداد الأموال المنفقة على عملية استقطاب هؤلاء العملاء،وبالتالي الخسارة على المدى الطويل.

أهمية البحث:

إن عملية الـــCross Selling المنفذة بوضع رضا العميل في المقام الأول تؤدي إلى تقليل الأرباح على المدى القصـــير و لكن زيادتها على المدى الطويل ، اضـــافة إلى زيادة ثقة العميل في المؤســـســـة عبر طرحها للخيارات التي تحقق فائدته في المقام الأول.

أهداف البحث:

-تحسين الـCross Selling .

-تقليل التكاليف التشغيلية بالنسبة للمؤسسة المالية.

-زيادة رضا العملاء ومعدل بقائهم.

-زيادة هامش الربحية على المدى الطويل.

حدود البحث:

ولـدت البيـانـات بـإســتخـدام datanamic data generator ، ثم نقلـت إلى داخـل Hadoop عبر Apache Sqoop،ومن ثم أجريت المعالجات الأولية و التحليل بإستخدام Apache Spark و Hive.

منـهجية البحث:

لحل مشكلة البحث اتبعت الخطوات الأتية:

-اختيار المعمارية الملائمة لحل المشكلة

-تصميم المعمارية.

-توليد البيانات.

-ادخال البيانات إلى نظام التحليل.

-معالجة البيانات و استخلاص النتائج.

فرضية البحث:

الربحية الكلية للمؤسسة المالية على المدى الطويل تتناسب طرديا وكمية البيانات المتوفرة للتحليل.

هيكلية البحث

يتكون من خمسة أبواب مقسمة على النحو الأتي :

الباب الأول : ويتناول مقدمة عن البحث و المشكلة التي يعالجها ومنهجية الحل.

الباب الثاني : يتألف من أربعة فصـــول أولها يتناول مفاهيم التســـويق المتعلقة بالدراســـة، والفصـــل الثاني يستعرض مفاهيم و خطوات عملية تنقيب البيانات،أما الفصل الثالث ففيه شرح عام لأدوات معالجة البيانات الضخمة المستخدمة، و الفصل الرابع فيه شرح للدراسات السابقة التي قامت هذه الدراسة على ضوئها.

الباب الثالث : موضح فيه طريقة جمع البيانات المستخدمة ووصف هيكلتها والبيانات الموجودة داخلها.

الباب الرابع : وفيه شرح مفصل لعملية معالجة و تحليل البيانات كما تمت للحصول على النتائج.

الباب الخامس : ويحوي عرض النتائج المتحصل عليها و شرح معناها.

الباب الثاني
الخلفية النظرية
و
الدراسات السابقة

الفصل الاول
:Cross Selling

تعتبر عملية التسويق عملية مهمة في زيادة أرباح المؤسسات ، حيث أنها تعمل على تعريف الزبائن المحتملين و الحاليين بمنتجات المؤسسة ، و تساهم في دخول منتجاتها الى أسواق جديدة ، و تبيين فوائدها بالنسبة للزبون و المشكلة التي تساهم بحلها بالنسبة له ، وكذلك تبيين فرقها من منتجات المنافسين و الميزات التي تتفوق بها عليها بصورة ضمنية.

احدى أساليب التسويق هو الـCross Selling ، وهو عملية بيع منتجات أو خدمات اضافية للزبائن الحاليين، [3] .و يستخدم الـــCross Selling لتحقيق عدة أهداف تجارية منها :أولا زيادة عائدات المنشأة وهذا شيء واضح حيث أن زيادة المبيعات تعني زيادة عوائد المؤسسة ، ولكنها لاتعني بالضرورة زيادة صافي الأرباح(wells Fargo). وثانيا تقليل التكاليف حيث أن عملية البيع ترافقها حملات تسويقية مصاحبة لمنتج المؤســـســـة المعروض للزبائن الجدد و المحتملين مما يعني انفاق مبالغ مقدرة من المال ، و لكن هذه الحملات غير ضرورية للشركة بالنسبة للزبائن القدامى ، إذ لا توجد حوجة لتلك الحملات الاستقطابية،مما يعني عدم وجود تكلفة امتلاك الزبون في فاتورة المنشأة أو ما تعرف بالـacquisition cost .

ثالثا زيادة ارتباط العميل بالمؤسسة عبر اعتماده عليها في تزويده بعدة خدمات أو منتجات مما يعني زيادة تكلفة الانتقال إلى مؤسسة أخرى أو ما يعرف بالـــSwitching cost ،ورابعا تقليل الجهد المبذول من قبل موظفي المبيعات حيث أن الزمن المتوفر يستفاد منه في استقطاب زبائن جدد للمؤسسة. خامسا زيادة رضا الزبون من المؤسسة عن طريق تحقيق و اشباع رغباته[3] . ويعتبر الـCross Selling صفقة رابحة لكلا الطرفين إذا تم تنفيذه بالصورة الصحيحة، حيث أنه يعطي العميل احساسا بالأهمية و الاهتمام برغباته واحتياجاته.

ولكن ليست كل عملية بيع اضافية مربحة بالنسبة للمؤسسة،وذلك لأن عملية البيع لبعض العملاء غير مجدية ماليا نسبة لأن تكلفتها أعلى من الربح المتوقع منها ،ولكي تتفادى الشركات هذه المبيعات الضارة ، تقوم بإستخدام بعض المؤشرات المالية مثل الـــCLV(Customer Lifetime Value) الذي يستخدم الـModel RFM .

: Customer Lifetime Value (CLV)

وهي النقدية الحالية للمعاملات المالية المستقبلية للعميل مع المؤسسة ، وهو يساعد المؤسسات على وضع سقف أعلى لإنفاقها على عمليات الاستحواذ على الزبائن.

:RFM Model

وهي اختصـــار لـــــRecency , Frequency , and Monetary Model ، وهوأحد النماذج المستخدمة لايجاد قيمة الـCLV لأي عميل مع المؤسسة، حيث يتم ذلك عبر ايجاد قيم المتغيرات الثلاثة عبر تقســـيمها إلى تدريج معين (من 1 إلى 10 في الغالب)، وتحلل تلك النتائج لتحديد الاســـتراتيجية المناســـبة لإستهداف كل فئة على حدة. [4]

ولقياس رضا الزبون تم استخدام أداة مشهورة وفعالة تعرف بالـNPS،وهي مستخدمة من قبل كبرى الشركات كما سيوضح في الفقرة التالية:

: NPS (Net Promoter Score)

هو عبارة عن أداة تسـويقية لقياس مقدار رضـا الزبون عن طريق سـؤال واحد فقط وهو : " ما هي احتمالية أن تقوم بإقتراح هذا المنتج لصديق" يتبعها رقم من 1 إلى 10 [5]، حيث أن الأرقام بين 1 و 6 تمثل زبونا منفرا من المؤســـســـة ، ومن 7 إلى 8 تمثل زبونا ســـلبيا (غير نشـــط) ، أما 9 و 10 فهم الزبائن الايجابيون.وتم اختراع هذا المقياس كبديل عن الاســـتبيانات التقليدية نســـبة لإزعاجها للزبائن و عدم اكمال معظمهم لها ، اضافة الى أن الNPS مقياس بسيط و مفهوم بالنسبة لمدراء المؤسسات ، و يمثل هدفا واضحا بالنسبة لتحسينه للمؤسسة المعنية.

ويعتبر هذا المقياس غير دقيق تحليليا من البداية حتى بالنســـبة لمخترعه ، و لكنه مفضـــل نســـبة لبساطته للزبائن و الاداريين.حيث يتم استخدامه اما على شكل نجوم او على شكل أرقام مدرجة.

يتم حسابه بالنسبة للمنشأة على النحو الاتي :

نسبة الـــPromoters – نسبة الـــDetractors ، حيث الـــPromoters هم الفئة أصحاب القيم 9 و 10 ، أما الـــــDetractors فهم من أبدوا تقييما من 6 وما دون ، وتعتبر درجة 50 و ما فوق ممتازة في عالم الأعمال،أما الاستراتيجية الأساسية لزيادة الـــNPS فهي تقليل الـــDetractors عبر تحويلهم الى Passive وهم أصـــحاب التقييم 7 و 8 ، وهذا أســهل مذهب لزيادة الـــــNPS ، حيث أنه من الصـــعوبة بمكان تحويل الـDetractors إلىPromoters ، و لكن من السهل نسبيا تحويلهم الى زبائن سلبيين.

الجدير بالذكر أن هذا المقياس مســـتخدم من قبل العديد من الشـــركات العملاقة كمعيار ناجح لقياس رضا الزبائن ، ومن بعض الشركات المعروفة التي تطبقه على سبيل المثال لا الحصر : (شركة ديل و اتش بي و فيسبوك وسوني).

الفصل الثاني
تنقيب البيانات :

ظهر مصطلح تنقيب البيانات في منتصف التسعينات في الولايات المتحدة الأمريكية وهو يجمع ما بين الاحصاء و عدة مجالات أخرى كقواعد البيانات ، الذكاء الاصطناعي ، و الـــMachine Learning .وتوجد عدة تعريفات لهذا المفهوم منها : أنه عبارة عن تحليلات لكمية كبيرة من البيانات بغرض ايجاد قواعد و أمثلة و نماذج يمكن أن ترشد أصحاب القرار و تتنبأ بالسلوك المستقبلي .[6]

من خلال التعريف السابق يمكن القول أن تنقيب البيانات عبارة عن تطبيق لإستخراج أو اكتشاف معرفة مفيدة و قابلة للاستغلال من خلال مجموعة كبيرة من البيانات ، حيث يساعد في اكتشاف المعرفة المخفية و النماذج غير المتوقعة إضافة إلى اكتشاف قواعد جديدة موجودة في قواعد البيانات الضخمة.

اكتشاف المعرفة الموجودة داخل قواعد البيانات (KDD) :

يعتبر تنقيب البيانات جزءا مكملا من عملية اكتشاف المعرفة داخل قواعد البيانات،وال(KDD) هو العملية الكلية التي يتم عبرها تحويل البيانات الخام الى معلومات مفيدة ، و عملية اكتشاف المعرفة في قواعد البيانات [6]، وهي تتضمن عددا من المراحل :

الرسم الأتي يوضح مراحل عملية اكتشاف المعرفة في قواعد البيانات :

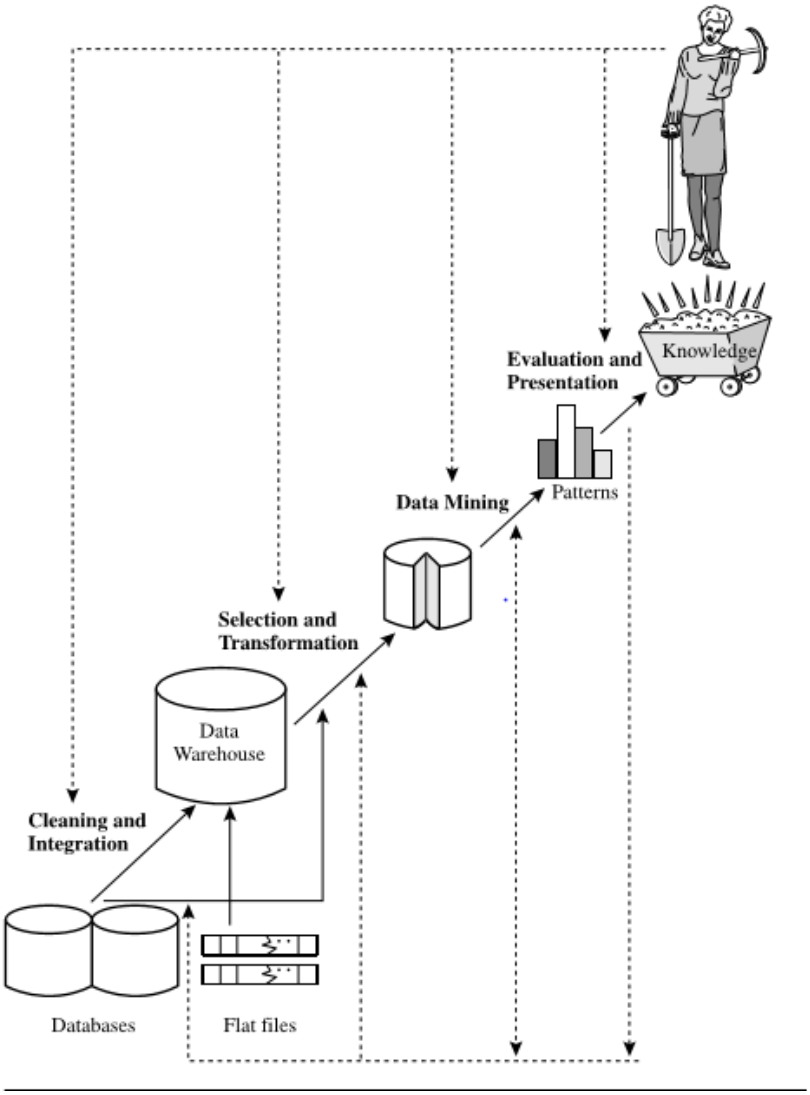


Figure 2-1 Data mining as a step in the process of knowledge

: Data cleansing.1

وهي مرحلة الاكتشاف والتصحيح الشكلي أي معالجة البيانات الناقصة ، وعزل البيانات التي تحوي تشويشا (Noise) من مجموعة البيانات.

: Data integration.2

وهي مرحلة جلب البيانات من مصادر متعددة و دمجها معا وذلك بغرض وضع البيانات بصورة موحدة

: Data selection.3

وهي المرحلة التي يتم فيها تحديد و استرجاع البيانات المناسبة المتعلقة بالتحليل من مجموعة البيانات .

: Data transformation.4

و هي عملية تحويل البيانات التي يتم اختيارها إلى شـكل ملائم للتحليل من حيث مناسـبتها لعمليات البحث و الاسترجاع بإستخدام أساليب مثل performance summaries و aggregation operations .

: Data mining.5

وهي المرحلة الأساسية التي يتم فيها تطبيق أساليب ذكية و خوارزميات لإستخراج نماذج مفيدة قدر الإمكان.

: Pattern evaluation.6

وفي هذه المرحلة يتم تقييم النماذج المهمة التي تمثل المعرفة استنادا إلى مقاييس محددة.

: Knowledge presentation.7

وهي المرحلة الأخيرة من مراحل اكتشـــاف المعرفة في قواعد البيانات ، وهي المرحلة التي يراها المســتفيد من عملية التحليل ، وهي تستخدم الأسلوب المرئي لمساعدة المستفيد في فهم و تفسير نتائج استخراج البيانات.

المراحل من الأولى إلى الرابعة هي أشــكال مختلفة للـــــdata preprocessing لكي تكون البيانات مهيأة للتحليل و تعطي نتائج واضــحة،ويمكن أن تقســم إلى المرحلتين في ان واحد و على ســبيل المثال انجاز الـــــdata cleansing و الــdata integration مع بعضها ، والــdata selection و الــdata transformation في مرحلة أخرى.

أساليب تنقيب البيانات(Data mining techniques):

يتضمن تنقيب البيانات عدد من الأساليب الرئيسية التي يمكن من خلالها استخدامها للوصول إلى الهدف من التحليل و منها :

1.قواعد الارتباط (Association rules) :

وهي أداة من أدوات اكتشاف المعرفة (KDD) وتسمح بالتعرف على كل القوانين الممكنة التي تشرح بعض الصفات الموجودة اعتمادا على وجود الصفات الأخرى الموجودة في مجموعة هائلة من البيانات ، و بمعنى أخر هي قواعد ارتباطية بين بعض البيانات داخل قواعد البيانات.[6]

و يمكن ان تستخدم قواعد الارتباط في ما يعرف بتحليل سلة السوق او بالـــmarket basket analysis ، حيث ان البيانات التي يتم تحليلها تتكون من معلومات حول العناصـــر (Items)التي يشـــتريها العميل وذلك لإيجاد العلاقة بين اكبر كمية من العناصر في سجلات المعاملات التجارية للمساعدة في عمليات اتخاذ القرار من خلال تحليل سلوك العملاء[7] و من أحد امثلتها :

Apriori Algorithm :

هي من الخوارزميات المعروفة في اكتشـاف قواعد الارتباط لايجاد العناصـر المتكررة وتسـتخدم خاصـية downward closure و ant monotonicity من اجل معرفة الدعم support وذلك لعرض العناصر التي تتجاوز نسبة ظهورها الـsupport المحدد.[7]

Clustering.2 (التجميع):

هو عملية تقسيم أو تجميع البيانات إلى مجموعات فرعية بحيث تكون عناصر كل واحدة منها مشابهة لبعضها أكبر ما يمكن ، والاختلاف بين سمات بيانات كل مجموعة و الأخرى أشد ما يمكن.[6]

فكرة تجميع البيانات بســيطة في طبيعتها و قريبة جدا من طريقة تفكير الإنســان ، حيث أننا كلما تعاملنا مع كمية كبيرة من البيانات نميل إلى أن تلخيص ذلك الكم من البيانات بناء على تشـابه الخصـائص بينها إلى عدد قليل من المجموعات و ذلك من أجل تسهيل عملية التحليل ، حيث يتم التقسيم في الــClustering بإستخدام عدة خوارزميات ، وهو غير موجه عموما Unsupervised Learning عكس التصنيف (Classification) و هو ما سنتناوله لاحقا.

كما أن الـــــClustering يســاعد المحلل على فهم التركيب الطبيعي لمجموعات البيانات وأحد خوارزميات التجميع المعروفة :

K-means algorithm:

هي من أشــهر خوارزميات التجميع . تســتخدم لتجميع عدة بيانات اعتماداً على خصــائصــها وتتم عملية التجميع من خلال تقليل المســافات بين البيانات ومراكز التجمع cluster center ،و لتوضــيح عملها نفرض ان المدخلات هي K الذي يمثل اعداد التجمعات (clusters) المراد الحصــول عليها وD الذي يمثل مجموعة البيانات المراد تقسـيمها ،حيث يتم الاختيار عشــوائيا لـــــK مع تحديد المركز المتوســط لمجموعة البيانات داخل التجمع باعتبارها القيمة المتوسـطة للبيانات داخل التجمع الذي تم اختياره مع تكرار عملية الاختيار يتم تعيين البيانات الي التجمعات الأكثر مماثلة وذلك اســتنادا إلي المســافة بين البيانات ومركز التجمع حتي لايوجد اختلاف داخل التجمع حيث تكون البيانات متشابة داخل التجمع مع الاختلاف مع البيانات في التجمعات الاخرى وهو ناتج الخوارزمية.[6]

Classification.3 (التصنيف) :

وهو تفســير أو التنبؤ بخاصــية فرد ما من خلال خصــائص أخرى، أي يقوم بتحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها لتصنيف البيانات ، وهنا يكمن الفرق بين التصنيف و التجميع . [6]

هنالك عدد من الطرق التي يتم استخدامها في تصنيف البيانات وهي الخوارزميات الإحصائية و خوارزميات الشبكات العصبية وشجرة القرار.

شجرة القرار: (Decision Tree)

وهي احدى خوارزميات التصـنيف و تعمل على استخدام مبدأ التقسيم والحل (divide and conquer) ،وهو تقسـيم المشــكلة إلى أجزاء وحل كل جزء على حدة،وينتج عنها مصــنف (Classifier) على شــكل شــجرة (Tree) يمكن تحويلها إلى مجموعة من القواعد تسمى بالـ(Decision rules).

.

الفصل الثالث
الأدوات و التقنيات :

في هذا الفصل سنستعرض الأدوات المستخدمة في هذا البحث عبر اعطاء تعريف مختصر لها و توضيح موقعها من عملية التحليل :

البيانات الضخمة(Big Data):

هي البيانات المشتقة من عدة مصادر أغلبها غير تقليدي ، مثل معاملات نقاط المبيعات،وسجلات معاملات بطاقات الائتمان ، وسجلات الويب.وتشمل أيضا البيانات شبه المهيكلة وهي البيانات التي تحوي هيكلية ليس بالضرورة للبيانات المخزنة إتباعها،حيث أنها موجودة كإشارة لكيفية تخزين البيانات فقط.

وتختلف البيانات الضخمة عن البيانات التقليدية في عدة سمات أبرزها ثلاثة :

السرعة التي تصل عبرها البيانات إلى المستودعات،وتعدد أنواع البيانات، وحجم البيانات المخزنة[2] ،والفقرات التالية توضح تقنيات البيانات الضخمة المستخدمة:

1. Apache Hadoop :

هو نظام تخزين و معالجة موزعة مفتوح المصدر عبر مؤسسة Apache ، وقد تم استلهام نموذج التخزين و المعالجة من الأوراق البيضاء التي نشرتها Google في عام 2003 لتوضح فيها نموذج عمل محرك البحث الخاص بها.[2]

وأهم ميزاته عن أنظمة المعالجة الموزعة الأخرى أنه يعمل على أجهزة تجارية ، وهي الأجهزة ذات امكانات المعالجة الضعيفة و المصنعة للاستخدام التجاري حيث لا مستخدم محدد في الاعتبار ، حيث السعر أهم من الجودة ، وبالتالي تكون نسبة الفشل فيها و الأعطال غير متوقعة على عكس تلك الخوادم عالية الجودة التي تطلبها المؤسسات العملاقة لتلبية احتياجاتها المعالجية.ويمكنه أيضا ادارة عدد لامحدود من الأجهزة ، أي أنه أبعد مشاكل التوسعية نهائيا في الـclusters .

بالاضافة إلى هاتين الميزتين يوفر Hadoop أيضا ميزة الـــFault Tolerance حيث أن الفلسفة الأساسية التي بني عليها هي أن فشــل أحد الأجهزة شـــيء عادي ومتوقع نســبة لأنه يعمل على أجهزة رخيصـــة ، حيث تمت معالجتها في الناحية التخزينية عن التخزين المكرر ، وفي مرحلة المعالجة عبر Job Trackers .ويقوم Hadoop بمعالجة البيانات بصـورة تختلف تماما عن الأنظمة التقليدية ، حيث أنه يقوم بنقل البرنامج إلى الجهاز التي تتواجد فيه البيانات ، بدلا عن نقل البيانات إلى الجهاز التي يوجد فيه البرنامج ، وهذا أقل حجما بالنسـبة للشـبكة التي تنقل عبرها البيانات ، و أسرع في الحالات التي توجد فيها البيانات خارج الclusterالمحلي.

ويعمل Hadoop بصورة أفضل مع البيانات شبه المهيكلة و غير المهيكلة كليا ،لأنه يتعرف على بنية و نوع البيانات عند قرائتها Schema On Read ، حيث تخزن كلها في صورة نصية Text.[2]

: Hadoop Distributed File System (HDFS).2

نظام ملفات هادوب الموزع مبني بحيث يخزن كميات ضـخمة من البيانات في أجهزة تجارية موزعة ، وهذه البيانات في الغالب تكتب مرة واحدة و يتم قرائتها عدة مرات ، وهو ما يعرف ب(WORM) أو Write Once – Read Many data ، وحجم وحدة التخزين الافتراضـــية(block) هو 128 MB ، وعلى عكس أنظمة الملفات التقليدية اذا تم القيام بتخزين بيانات أقل من حجم الـblock فإن بقيته تكون متاحة للاستخدام.[2]

هذا النظام مصـمم بحيث تكون سـرعة الوصـول إلى كل البيانات أعلى من سـرعة الوصـول إلى أول سـجل ، وهذا مهم لأن معظم تطبيقات الحوســبة الموزعة تتضـــمن معالجة معظم أو كل البيانات في كل عملية.وهنالك بعض التطبيقات التي لا يعمل فيها الـــHDFS بكفاءة مثل أنظمة الوصول و الاستجابة السريعة ، و الأنظمة التي تكتب في عدة أماكن في الملف الواحد و تقوم بتعديله لاحقا ، وأيضا تلك التطبيقات التي تستخدم ملفات صغيرة جدا.[2]

MapReduce.3 :

هو عبارة عن نموذج برمجي لكتابة البرامج التي تقوم بمعالجة البيانات الموزعة داخل الـــHDFS ، ويمكن كتابة هذه البرامج بعدة لغات منها : Java و Scala و Ruby on Rails و Python ، و هذه البرامج بطبيعتها مصممة للمعالجة الموزعة.[2]

برامج MapReduce مكونة من جزئين أســـاســـيين : Mapper وهو الجزء الذي يقوم بقراءة البيانات من الـــHDFS و يقوم بإجراء المعالجة الأولية عليها بحيث يستقبل متغيرين هما key و value ، أما الجزء الأخر فهو الـــReducer الذي يقوم بإستقبال مخرجات الـــMappers كمدخلات ،واختلافه الأساسي عن الـــMapper هو أن الـــMappers تقوم بالمعالجة بصـورة موزعة بحيث يعالج كل منها كمية متسـاوية من البيانات و يشـرف عليها ما يعرف بالـــــJob Tracker حتى انجاز عملها ،أما الـــــReducer فهو يقوم بمعالجة مركزية حيث كل نواتج الـMappers تتجمع في الألة التي يوجد بها الـReducer ليقوم بإعطاء الناتج النهائي.

Apache Sqoop.4 :

هو أداة تقوم بـإســـتخراج البيانـات المهيكلـة من وحدات تخزينهـا مثل قواعد البيانـات الموزعـة إلى نظـام Hadoop لإجراء المعالجات لاحقا [2]، ويمكن بعد ذلك إرجاع نتائج هذه المعالجات من نظام Hadoop إلى مصادر البيانات الأصلية لتخزينها.

Apache Hive.5 :

هو عبارة عن نظام مســتودع بيانات واســتعلامات يعمل على نظام Hadoop ، تم انشــاؤه في facebook لتلبية احتياجاته التحليلية على بياناته الضخمة التي تتجاوز عدة petabytes .

و أكثر ما يميز نظام Hive هو استخدامه لـــ(Hive query language(HQL وهي لغة شبيهة ب SQL تتيح لمستخدمها عمل استعلامات على البيانات المخزنة داخل نظام Hadoop [2]، حيث يتيح لمحللي قواعد البيانات التقليدية ســـهولة الانتقال إلى أنظمة البيانات شـــبه المهيكلة ، ولديه ثلاث بنائيات بيانات : tables و partitions و buckets.

: Apache Spark

هو منصة مفتوحة المصدر مخصصة لتشغيل برامج الحوسبة الموزعة إضافة لتوفير نموذج لكتابة البرامج الموزعة عليه ، يتميز Spark عن النماذج السابقة بإتاحته للبرامج التي تستخدم الـــIteration [8] ، الذي تستفيد منه برامج تنقيب البيانات على وجه الخصوص وذلك لإستخدامه الـــــRDD التي هي عبارة عن بنائية بيانات مخزنة في الذاكرة فقط ولا تتعامل في معالجتها مع الـــــdisk إلا عبر أوامر صريحة،مما يتيح سرعة أكبر في معالجة البيانات. [2]

:Apache Zeppelin

هو أداة ويب مفتوحة المصـدر تسـتخدم لتصـوير البيانات الضـخمة في صـورة رسـومات سـهلة الفهم والمقارنة بين عناصرها،ويستخدم للتحليل والتمثيل التفاعلي للبيانات[9] .

:Scala

وهي لغة برمجة بدأ تصميمها في عام 2001 على يد العالم الألماني مارتن أوردسكي ، وكان أول ظهور لها في عام 2003 ،واسم Scala مشتق من صفتها Scalable language ،وتتميز بتعدد استخداماتها نسبة لقلة الذاكرة المستخدمة و سرعة التنفيذ وهي مبنية على Java Virtual Machine مما يتيح لها الوصول إلى كافة مكتبات الجافا دون الحاجة لإعادة كتابتها مرة أخرى،وتتميز برامجها بأنها Statically compiled مما يتيح اكتشـاف الأخطاء فيها بسـهولة،كما يمكن اسـتخدامها كبرامج مفسـرة مما يتيح كتابتها في ملف منفصـل أو في محرر الأوامر Command line ،وتعتبر لغة أساسية في كتابة البرامج الموزعة التي تعمل على Map Reduce وApache Spark .

ومما وجب ذكره أن هذه الدراسة قامت بإستخدام بعض تقنيات الويب server side مثل Laravel framework [10] و هو اطار عمل مبني بلغة PHP و هي لغة برمجة مفتوحة المصدر تستخدم لتطوير مواقع الويب من خلال الخادم Server [11] ،وMySQL و هو عبارة عن نظام مفتوح المصـدر لإدارة قواعد البيانات العلائقية (RDBMS) [12]،وشـائع الاسـتخدام مع تطبيقات الويب لتطوير وبناء مواقع وتطبيقات الويب علي اساس إستخدام Model –View –Controller (MVC) لتميزه بالبساطة والقوة الأمنية العالية [10].

واستخدمت أيضا بعض تقنيات الـClient side مثل HTML وCSS وBootstrap وJavaScript [13].

الفصل الرابع
الدراسات السابقة:

في هذا الجزء سنمر على الطرق و التقنيات التي استخدمت من قبل الباحثين السابقين ، و ذلك لتبيين ميزاتها ،و نوع البيانات المستخدمة فيها، و العوائق التي واجهت أولئك الباحثين ، و الجوانب التي لم تشملها . ومن ثم سنوضح في ضوء هذه الدراسات الجوانب التي تعتبر فيها هذه الدراسة امتدادا لتلك الدراسات.

أول محاولة بحثية لتحسين الـCross Selling ترجع ل[1] حيث استخدم طريقة الـLatent trait لاقتراح المنتجات للزبائن الجدد بناء على تشابه أنماط معاملاتهم مع الزبائن الحاليين ، و لكن أخذ عليها أنها لم تغطي احتمالية ما إذا كان هؤلاء الزبائن الجدد يقومون بشراء خدمات أخرى تشمل المنتجات المقترحة من مزودين أخرين . و توصلت الدراسة التي جرت على 18 خدمة مالية إلى طريقة لتقدير احتمالية تملك العميل لخدمة موجودة غير مستخدمة من قبله أو لخدمة جديدة مطروحة من قبل المنشأة ، حيث يتم ترتيب هذه الخدمات حسب ترتيب تملكها ، وأيضا يتم ترتيب الزبائن بنفس طريقة ترتيب الخدمات حسب ترتيبهم في تملك الخدمات مربوطة بديموغرافياتهم و أهدافهم المالية ، وتقوم الخوارزمية بالتنبؤ بهذا الترتيب . و اقترح الباحث توسعة السوق عن طريق زيادة رأس المال المنفق في كل خدمة عوضا عن زيادة المنتجات المملوكة كطريقة أخرى لتحقيق أهداف الـCross Selling ، كما أشار أيضا إلى فائدة أخرى وهي زيادة نسبة الاحتفاظ بالزبائن ، و علل ذلك بأنه كلما كان عن الخدمات التي يستعملها العميل كثيرا كلما زادت كمية الخدمات التي يجب أن يلغيها حتى يذهب الى مؤسسة منافسة.

ومن القصور الذي واجه .Kamakura et al أنه في حالة شراء العميل لكل الخدمات أو عدم شرائه أيا منها فإن هذا النموذج لن يتمكن من التنبؤ بدقة ، أما وجه القصور الأخر الذي واجهته هذه الدراسة هو أن هنالك بعض العملاء يقومون برفض بعض الخدمات بالرغم من أن احتمال تملكهم لها الذي يظهره النظام كبير و هذا يرجع الى أحد ثلاث أشياء : إما أنا العميل ليس لديه معرفة كافية بالخدمة ، أو أن مستوى المخاطرة في هذا المنتج غير متناسب و أهداف العميل المالية ، أو أن هنالك خدمة أو منتجا أخر له احتمال تملك أكبر من هذا المنتج المرفوض . و لكن عدم تقدير المنتجات او الخدمات التي يستهلكها العميل من منافس أخر جعل الاقتراحات التي يقوم بها نظام Kamakura et al. غير دقيقة و مزعجة بالنسبة للزبائن كما سنوضح في الدراسات التي سنتناولها لاحقا.

لكن Kamakura et al. أدركوا العيب المنهجي الذي تحتويه دراستهم السابقة ، لذلك قاموا بمحاولة لمعالجة هذا القصور في دراسة جديدة عن طريق اضافة نتائج استبيان عينة من الزبائن عن كمية الاحتياجات التي يقومون بتغطيتها من المؤسسات الأخرى المنافسة [14]، و من ثم تخزين ما تم التوصل اليه من كمية المنتجات الأخرى التي يستهلكها الزبائن من المنافسين في قاعدة بيانات المؤسسة المالية [15]، حيث أن نموذجهم الحسابي الجديد يقوم بتقدير الفجوة التي لم يتم تغطيتها من قبل أي مزود لكي يتم تغطيتها باقتراحات الـــCross Selling ، وابعاده لمنتجات وخدمات المؤسسة التي تقوم المنافسة بتغطيتها . أحد أهم ميزات هذا النموذج الجديد هو أنه التعامل معه من مرحلة تمثيل العلاقات حتى عرض النتائج النهائية كلها تتم بصورة رسومية ، وهو ما يعني سهولة في التعامل و فهم النتائج من قبل المدراء ، حيث أنهم سوف يتمكنون من معرفة من هم العملاء الذين يفترض عرض خدمة معينة لهم حتى يحصلوا على أكبر نسبة عروض مقبولة على سبيل المثال ، و هذا يعد تحسينا ملحوظا لمنهجيته في تحسين الـــCross Selling . و لكنهم أوضحوا بعض أوجه القصور في هذا النظام ، حيث أنه معقد حسابيا بسبب استخدامه للنمذجة الرسومية في كافة مراحله ، و هو ما يعني احتياجه الى امكانات حسابية هائلة لم تكن متوافرة للكل في ذلك الوقت ، و هو ما اضطرهم أيضا إلى أخذ عينات من العملاء من قاعدة البيانات المذكورة سابقا و ليس كل العملاء وهو ما يعني اعطاء النظام لنتائج تقريبية.

الدراستان السابقتان اعتمدتا في اقتراح المنتجات الجديدة على تشابه أنماط الشراء عند الزبائن الجدد و الحاليين ، بغض النظر عن اختلاف الشخصيات و العوامل المؤثرة على قرارات الزبون ، و هو ما ركز عليه [16] حيث لاحظوا أن العملاء يقومون في العادة بشراء المنتجات و الخدمات متبعين ترتيبا تلقائيا حسب وظيفتها و تعقيدها (و هذا ينطبق على المنتجات المالية على وجه الخصوص ، حيث أن هنالك منتجات لا تناسب العملاء من فئات معينة ، إذ قد ينطوي على شرائها اثار عكسية مثل زيادة مستوى المخاطرة و تعرض العميل الى خسائر لا تتناسب و نسبة مخاطرته) .وافترضوا أيضا أن هذا الترتيب مستقل عن العملية التسويقية للمنشأة . و لاحظوا أيضا أن شراء العميل لمنتج أو خدمة معينة ينشئ حاجة مستقلبية لمنتج أو خدمة أخرى ، و هو ما اصطلح الباحث على تسميته بال "الترتيب الطبيعي"، وهو مهم في الأسواق التي تتولد فيها احتياجات جديدة بعد عملية الشراء الأساسية ، كتلك التي لم يثق فيها الزبون بعد بمزود المنتج أو الخدمة ، أو تلك المنتجات التي تحتاج من العميل فهما أفضل لمعرفة مدى توافقها مع المنتج الحالي أنهم يعرضون المنتجات بصورة تسلسلية حسب مرحلة النضج الطلبي للزبون ، كلما اقتربت مرحلة نضجه من منتج معين كلما كان ذلك المنتج أو الخدمة أوفى تلبية لاحتياجات العميل مما يعني احتمالا أكبر لشرائه لهذا المنتج .

استخدم الباحثون في دراستهم Ideal point model لتمثيل المنتجات و العملاء ، وقاموا بتحديد المتغيرات التي سيقومون بإجراء التحليل عليها وهي : المنافسة و تعني ما إذا كان لدى الزبون حساب مع مؤسسة أخرى ، و الرضا و يقاس عن طريق المسوحات التي يجريها البنك بصفة دورية ، و تكلفة انتقال عميل الى منافس أخر. كما عرف الباحثون أيضا أن النضج المالي أو ما اصطلحوا عليه سابقا بالطلبي يقاس عبر عدة عوامل متحدة وهي : الأملاك المتراكمة ، والأرصدة الشهرية ، و الفترة الزمنية للحسابات حسب أهميتها. البيانات التي تم اختبارالنموذج بها هي بيانات ل 1201 عميل لبنك معين تم أخذها بناء على رضائهم و موافقتهم ، و تمثل كمية معاملاتهم لمدة تقارب العام منذ يوليو 1997 الى يونيو 1998.

ثم قام الباحثون بمقارنة نموذجهم بأربعة نماذج ، حيث أن كلا منها يعتبر تحسينا للسابق وصولا الى نموذجهم الحالي ، وهي : سلاسل ماركوف مستخدمة احتمالات الشراء المشروطة ، بإفتراض أن المنتجات مستقلة ، يقوم هذا النموذج بالتنبؤ بعمليات الشراء المستقبلية بناء على المنتجات أو الخدمات المملوكة حاليا.أما النموذج الثاني فهو يقوم أيضا بإفتراض أن المنتجات مستقلة ، و هو يقوم بإستخدام binary probit model لكل منتج أو خدمة لكل فئة على حدة ، ولكن هذا النموذج يتجاهل العلاقات غير المنظورة بين الفئات ، و تخصيص الموجودات ، و تكاليف انتقال العملاء الى المنافسين ، والتجانس بين المنتجات ، ولكن الباحث أورده نسبة لشيوع استخدامه ضمن المؤسسات للتنبؤ بإحتمالية الـــcross selling ، وهو يشبه نماذج أخرى تقوم بإفتراض استقلالية قرار العميل عند شراءه لمنتجات مختلفة.

أما النموذج الثالث فهو multivariate probit model ، و هذا النموذج يقوم بالتعرف على العلاقات اللحظية أو الحالية بين المنتجات خلال عملية شراء معينة ، وليس العلاقات التاريخية بينها . أما النموذج الرابع أضاف تكاليف الانتقال الى النموذج السابق ، فيما أضاف نموذجهم النضج الطلبي الى النموذج الرابع ، وهو ما اعتبروا أنه يساعدهم في أخذ الاحتياجات المالية التي تم اشباعها في الاعتبار.

تمت محاكاة الاستقلالية باستخدام اجراءات الـــprobit الموجودة في SAS ، و كذلك تمت محاكاة النماذج الأربعة السابقة بإستخدام سلاسل ماركوف و طريقة مونتي كارلو، واستخدم معامل Bayes و قيم الانحرافات لتقييم النماذج الســابقة، حيث أثبتوا أن نموذجهم هو الأنسب للبيانات ، وتوصلوا إلى ترتيب للمنتجات يطابق ما تنبؤا به مسبقا، وهو ما أكده اداريو البنوك.

و أوضحوا بعض العوامل التي تؤثر على هذا النضج كمستوى التعليم و الدخل الشخصي و الحالة الاجتماعية ، و هم بذلك قد تفوقوا على الأخرين عبر تقدير الاحتياجات لكل زبون على حسب خلفيته ، و لكنهم لم

يراعوا ما قام به .Kamakura et al بتقدير ما تغطيه المنافســـة من خدمات ومنتجات للزبون الحالي والجديد للمؤسسة .

ومن بعض النتائج التي توصــلوا في دراســتهم الى أن الرضــا مهم عن كبار الســن و الاناث ، كما أن أرباب المنازل كلما زاد مســتواهم التعليمي و كانوا من الذكور كلما زادت ســرعة نضــجهم الطلبي بغض النظر عن العوامل الأخرى المؤثرة على هؤلاء الأشخاص، وأوضحوا أن تكاليف الانتقال إلى المنافس هي أهم عامل في تزويد المؤسسة بفرص Cross Selling ، أما المنتجات عالية المخاطر يكون اكتفاء العميل أهم عامل فيها ، إذ أنها تحتاج الى قدر عال من الثقة بين العميل و المؤسسة.

و في عام 2016 قام [4] قاموا بإعداد دراسة عن نظام استحسان لرواد المكتبات باستخدام Collaborative User based filtering ، و قد أثنوا فيه على لغة R كلغة مهمة و فعالة لتنقيب البيانات ، و أختلفوا عن سابقيهم بأنهم أخذوا البيانات الضخمة بالاعتبار عبر استخدامهم لــHadoop و Apache Mahout كنظم معالجة البيانات الضخمة ، وهو ما أوضح لنا الأهمية العملية التي اكتسبتها البيانات الضخمة خارج المجلات العلمية و تزايد اهتمام المؤسسات بها كأداة تجارية مهمة .وأوضحوا أن نظام الاستحسان مصمم على خيارين : إما أن تكون التفضيلات مطروحة حسب المستخدم ، أو أن تكون حسب المؤلف المختار. لكنها لم تراعي الاختلافات الشخصية بالنسبة للزبائن صراحة ، و لكن على عكس الـــLatent trait يمكن اضافة هذه التفضيلات لتناسب الاختلافات الشخصية في الـــScoring Matrix لنظام الاستحسان.

أما الورقتان الاتيتان فقد أولياهما الكثير من الوقت و الاهتمام لســببين هما : تشـــابه أهدافهما الى حد كبير بهدف الدراسة و هو تحسين الأرباح مع وضع الزبون و رضاه في الإعتبار كأولوية لا غنى عنها ، و السبب الأخر هو توضـــيحهما بدقة ما قاموا به من جمع للبيانات و معالجتها ، مرورا بالشـــروط التي تم عبرها اختيار الخوارزمية المناسبة للتحليل ، و المصاعب التي واجهتهم حتى الوصول للنتائج.

الورقة الأولى كانت ل [4]، حيث ركزت دراستهم على تحليل و تصنيف الزبائن في البنوك على مرحلتين: المرحلة الأولى و فيها يتم تصنيف الزبائن إلى ثلاث فئات عبر خوارزمية الـــK-means الموسعة،حيث أوضحوا أن النسخة الموسعة منها تشمل صفات اضافية مثل أقل سعة للـــcluster و دمج و فصل الـــclusters عن بعضها ، وتم تقسيم العملاء الى هذه الفئات بناء على سلوكهم و الــRFM ، ثم تلا ذلك استخدام الــAproiri association rule و ذلك لايجاد القوانين الضــمنية التي تربط عناصــر كل cluster مع بعضــهم ، وذلك لعمل profile لكل عميل يتم من خلالها تحديد أي استراتيجية تسويقية سيتم استهدافه عبرها تليها وضع ملف تفصيلي(profile) لكل زبون ، و ذلك لكي تتيح لمسؤولي المبيعات استهدافا أفضل للزبائن عبر الحملات الترويجية

البيانات المستخدمة في الدراسة تم أخذها من احدى شركات بطاقات الائتمان مسبقة الدفع الايرانية ، وهي عبارة عن جدولين يغطيان بيانات أكثر من 55000 عميل و 11 مليون معاملة في فترة عامين و نصف العام من مارس 2007 حتى أكتوبر 2009 ،وتم ربطها بإستخدام الـ customer ID ، و تم التخلص من الـ records غير المكتملة ، وتطبيق بعض العمليات الاحصائية عليها لاحتساب قيم بعض المتغيرات المساعدة في تعيين سلوك العميل.و نجح في نهاية الأمر في تقسيم العملاء الى مجموعات متجانسة حسب قيمتها المالية و سلوكياتها ، الشيء الذي أشاروا الى فائدته من ناحية الاستهداف الأفضل للحملات الترويجية و تحسين العلاقة مع الزبائن.

أما الدراسة الأخرى فهي ورقة ل [17] ، وهي عبارة عن ايجاد فرص cross selling في قطاع الاتصالات عن طريق الـclassifiers و عن طريق الـBayesian neural networks، حيث يقوم النظام بإعطاء بيانات الزبون للـ classifier المخصص لكل خدمة .الأول هو Bayesian neural networks ، وهي توضح أي عملاء لديهم أكبر قابلية لشراء المنتج أو الخدمة ، و تقوم أيضا بتوضيح سلوك العملاء و هو ما لا يتوفر في المنهجيات الأخرى.

في البداية تكون الشبكة خالية حيث تمثل المعلومات المتوفرة لها بالنسبة للعملاء ، وتكون الانماط المكتشفة في المحاولات الأولى مختلفة كثيرا عن الواقع و هذا ما يستلزم من مستخدمها شرح هذه الاختلافات للشبكة لكي يتحسن توقعها في المرات القادمة.

أما النموذج الثاني فهو مبني على تصميم classifier مستقل لكل خدمة ، ثم اعطاء بيانات العملاء له كمدخلات ، و يكون ناتجه احتمالية شراء كل عميل للخدمة ، و من ذلك تقوم المنشأة بإقتراحها للعملاء الذين تتخطى احتمالية شرائهم حدا معينا.الجدير بالذكر أن هذه الدراسة اشتملت على مقارنة لدقة نتائج الـ classifiers مقارنة مع سهولة استخدامها.

بالنسبة للبيانات قيد التحليل فقد أوضح الباحث أنه لم تتوفر لديه بيانات حقيقية ، ولكن تم تعويض ذلك بإستخدام مولد بيانات لدى المعهد القومي للاتصالات ببولندا ، كما أوضح أنه بذل جهدا مقدرا في سبيل أن تكون البيانات المولدة أقرب للواقعية،ولضمان موضوعيتها تم توليدها بواسطة شخص مختلف عن الشخص الذي يقوم بالتحليل.

تم وضع خمس فئات مختلفة للعملاء حسب مستويات انفاقهم للحفاظ على تنوع العملاء ، ثم تم اسناد واحد من ست باقات مكالمات الى العميل عشوائيا واضعين في الاعتبار فئته المختار فيها مسبقا ، تبع ذلك توليد بيانات المعاملات لكل زبون حسب الـ profile الخاص به.ثم تلا ذلك تحويل البيانات المولدة الى هيئة مناسبة للتخزين في مستودع بيانات ، ثم اسناد خدمة لكل زبون خاضعة للقواعد الاحتمالية.كمية البيانات المولدة هي 5000 سجل ، 4000 منها للتحليل و 1000 للاختبار.

أظهرت نتائج التحليل أن الـــBayesian neural network أقل دقة من الـــClassifier ، و لكنه تميز عن الـClassifier بنمذجته لسلوك العملاء و توضيحه لأنماط مهمة عنهم.

طريقة الـــBayesian neural networks مع الـــassociation rules أعطت نتائج أفضل من العروض العشـــوائية كما أوضـــحت التجربة ،و بالرغم من عدم دقتها الا أنها تعطي تفصـــيلا شـــاملا للعلاقة بين المتغيرات.أما طريقة الـــــClassifier فهي مفيدة وعملية جدا في تحديد ما هو المنتج الأكثر احتمالا للشـــراء لكل زبون على حدة ، ولكنها لا توضح لماذا أو كيف يتم اقتراحه.

و لا ننسى الدراسة القيمة التي أعدها [18]حيث احتوت دراستهم على تفصيل لبيئة الوحدات الاستثمارية في البنوك الســـودانية و طبيعة عملها ، و ركزت على تفاصـــيل النظام الاســـلامي للبنوك الســـودانية و أنواع المعاملات الربحية فيها.وعالجت الدراســـة تحويل التفضـــيلات و الاقتراحات التقليدية التي يجريها خبراء الاســـتثمار ذات الدقة الممتازة ، و لكنها تشـمل فقط العملاء المهمين للبنك ، الى تفضـيلات أقل جودة و أكثر شـمولا لكل العملاء كما توحي طبيعة نظامهم. و ركزت الدراسة على تحليل البيانات المولدة اليا عن طريق الـــClustering و Association rule algorithm .

البيانات قيد التحليل تم توليدها بإسـتخدام datanamic data generator ، وتم توليد 25000 معاملة ، وقد عللوا اسـتخدامهم لبيانات مولدة لشـدة صـعوبة الحصـول على بيانات من البنوك المحلية. أما الأداة المسـتخدمة لتحليل البيانات و تطبيق الخوارزميات السابقة فهي Rapid miner.

قسـم الباحث البيانات الى ثلاث فئات بناء على جدواهم الاقتصـادية بإسـتخدام خوارزمية K-means ، بعض ذلك قام بإشتقاق المنتجات المترافقة بإستخدام association rules الموجودة في الـــــRapid miner.الجدير بالذكر أنهم صـمموا نظامهم لتحسين الـــUpselling و الـــCross Selling معا، حيث اقترحوا الـــCross Selling للعملاء ذوي الجدوى الاقتصـادية الضـعيفة ، أما ذوي الجدوى الإقتصـادية العالية فتم اقتراح الـUp Selling لهم.

الباب الثالث

البيانات

البيانات:

البيانات هي الشكل الخام لأي محتوى يتم انتاجه و معالجته و تفسيره و نقله ،وتصنف البيانات حسب هيكليتها إلى ثلاثة أنواع : النوع الأول هو البيانات المهيكلة ، وهي تلك التي تتميز بوجود شـــكل محدد و ملزم لقراءاتها و تخزينها مثل قواعد البيانات العلائقية،أما النوع الثاني فهو البيانات شـبه المهيكلة ، وهي تلك التي تتسـم بوجود شـكل محدد لتخزينها ولكنه غير ملزم ويمكن تجاهله ، مما يجعل بعض البيانات مخزونة بغير النمط المحدد إلى جوار بيانات أخرى تتبع النمط المحدد لها مثل ملفات الـــــSpreadsheet،أما النوع الثالث فهو البيانات غير المهيكلة وهي تلك التي ليس لديها أي نمط محدد للتخزين والقراءة مثل الملفات النصية والصور.[19]

بيانات التحليل :

هذا البحث يحتاج إلى بيانات معاملات بنكية غير مهيكلة تحوي البيانات الشخصية كالبيانات الديموغرافية ،و تحوي أيضـــا أنماط تفاعل العملاء مع النظام البنكي التي تكون بدورها غير مهيكلة البتة ، وتكون ذات حجم ضـــخم يتناســـب و القوة التحليلية للنظام الذي يســـتخدم أدوات تخزين ومعالجة البيانات الموزعة كتلك الموجودة في مراكز البيانات،وهو مرتكز على Apache Hadoop والأدوات التابعة له،حيث كان المأمول ايجاد بيانات بحجم 100 غيغابايت كحد أدنى لتظهر قوة التحليل على بنية موزعة.

في البدء تم البحث عن بيانات ضـــخمة ذات بنية عشـــوائية (Unstructured data) فتم الذهاب إلى بعض البنوك الكبيرة و طلب بيانات المعاملات المالية محذوفة منها البيانات الشـخصـية احتراما لخصـوصـية العملاء مع توضـيحنا للغرض من اسـتخدامها ، و تم مقابلة ذلك بالرفض التام نسـبة للسـياسـات الداخلية التي تمنع ذلك حتى مع توضيح الباحثين بأن نتائج التحليل ستصب في مصلحة البنك وعملائه.

بعد ذلك تم القيام بالبحث في الانترنت عن مجموعة بيانات معنية بالمعاملات المالية ، و لكن هذه المحاولات لم تكلل بالنجاح أيضـا و ذلك لارتباط منح هذه البيانات بالانخراط في برامج تعليمية محددة ، أو لعدم مطابقتها للحوجة التحليلية من حيث نوعيتها التي هي في الغالب بيانات مهيكلة ، أو لصـــغرها البالغ حيث كانت في بعض الحالات لا تتعدى 5 ميغابايت.

بعد كل هذا السعي الحثيث و عدم وجود أي أمل لتلقي بيانات بنكية حقيقية و ضيق الزمن لم يكن لدى الباحثين بد إلا بناء عملهم على بيانات مولدة أليا و هي أســوأ خيار ممكن بالنســبة لمحللي البيانات، فقد كان من الممكن تغيير نطاقهم البحثي ليتناســب مع البيانات الموجودة ، و لكن ذلك لم يكن أحد الخيارات نســبة لاقتناعهم بأهمية هذا المنحى البحثي الذي تحتاج اليه كثير من البنوك عموما و المحلية خصوصا.

تم توليد البيانات بإستخدام Datanamic Data Generator MultiDB V6 وهي عبارة عن أداة تجارية لتوليد البيانات لاختبار البرامج و التطبيقات التي تعتمد على قواعد البيانات حيث يتم ملؤها عشـــوائيا بأي كمية من البيانات و أي عدد من الصفوف ، و لكن النسخة التجريبية منه تتيح للمستخدم توليد 25000 سجل بيانات في عملية التنفيذ الواحدة، أما النسـخة المرخصـة فتتيح عددا لا محدودا من السـجلات في عملية التوليد الواحدة ،ولكن ترخيص النسخة باهظ الثمن (499 دولار أمريكي)[20] و لا تتوفر بدائل أخرى له.

خطوات توليد البيانات :

1.إعداد قاعدة البيانات والعلاقات بين خاناتها :

في البداية أنشئت جداول البيانات التي سيتم توليد البيانات لها،وهي خمسة جداول مفصلة على النحو الأتي :

1.جدول بيانات العملاء (Customers) :

وهو يحتوي على كافة بيانات العميل الشخصية.

اسم الحقل	النوع	الوظيفة
customer_id	integer (10)	رقم العميل(مميز)
first_name	varchar (20)	الاسم الأول
last_name	varchar (20)	الاسم الأخير
Age	integer (3)	العمر (من 18 حتى 85 سنة)
Sex	varchar (6)	الجنس أو النوع (ذكر أو أنثى)
marital_status	varchar (20)	الحالة الإجتماعية (و تشـــمل خمس حالات : أعزب single – متزوج married – مطلق

		divorced – أرمل widowed – منفصـــل قانونيا legally separated)
Salary	integer (10)	المرتب (من 1200 إلى 29000 في الشهر)
job_type	varchar (20)	نوع الوظيفة
Education	varchar (30)	المســتوى التعليمي (وتشــمل القيم الأتية : أمي No degree ،تعليم ثـانوي High school ، دبلوم Diploma ،زمـالـة مهنيـة Associate degree ،بكـالريوس Bachelor ،مـاجســتير Master degree ، دكتوراه PhD ، مـا بعـد الـدكتوراه Post-Doctoral ،تـدريـب مهني Vocational training)
Residence	varchar (30)	السكن
account_type	varchar (50)	نوع الحســاب(احدى خمس حالات : 1.حســاب توفير أساسي Basic checking account 2.حساب توفير checking account 3.حساب ادخار Savings account 4.حســـاب ودائع أســـواق مـاليـة Money market deposit account 5.حســـاب شـــهادات ايداع Certificate of Deposits account)
account_status	varchar (10)	حالة الحساب (نشط active أو غير نشط Not active)

Table 3-1 customer table

2.جدول المعاملات(Transactions) :

يحتوي على سجل المعاملات المالية للعميل بتفاصيلها.

اسم الحقل	النوع	الوظيفة
transaction_id	integer (10)	رقم العملية المالية
transaction_date	Date	تاريخ القيام بالعملية (من 1\1\2000 إلى 5\9\2017)
transaction_type	varchar (20)	نوع العملية (واحد من سبعة أنواع 1.رهن عقاري mortgage 2.بطاقات ذات سيولة دائنة credit card 3.أسهم Stocks 4.سندات دين bonds 5.صناديق مالية mutual funds 6.أصول متداولة أليا ETF 7.قروض شخصية Personal loans 8.تمويل سيارات Automobile loans 9.قروض بضمان المنزل Home equity loans 10.مشتقات مالية (Derivatives
transaction_code	varchar (10)	نوع العملية من حيث كونها دائنة credit أو مدينة debit
transaction_amount	big integer (15)	المقدار المالي للعملية (قيمة عشوائية من 1000 إلى 1000000)

Maturity	integer (5)	أجل الدين أو القرض الذي ينبغي عنده اكتمال سداد الأصل و الفوائد التابعة له(احدى سبعة قيم : سنة واحدة 1 أو سنتان 2 أو خمس سنوات 5 أو عشر سنوات 10 أو خمسة عشرة سنة 15 أو ثلاثون سنة 30)
interest_or_commission	Float	سعر الفائدة بالنسبة للقروض أو قيمة العمولة بالنسبة لعمليات البيع و الشراء (قيمة عشوائية بين 0.025 إلى 0.15)
Satisfaction	integer (2)	رضا العميل (قيمة من 1 إلى 10 ، القيمة 0 تعني أن العميل لم يقم بإبداء رأيه)
customer_id	integer (10)	رقم العميل

table 3-2 transactions table

3.جدول ضمانات القروض (Collaterals):

وهو يشمل الضمانات التي وفرها العملية لعمليات اقتراضه المختلفة.

اسم الحقل	النوع	الوظيفة
collateral_id	integer (10)	رقم الضمانة
collateral_type	varchar (15)	نوع الضمانة (واحد من أربعة أنواع : سندات حكومية Government bonds أو سيارة automobile أو عقار Real estate أو بوليصة تأمين Insurance

collateral_amount	big integer (15)	المقدار المالي للضمانة (قيمة تتراوح بين 1000 إلى 1000000)
transaction_id	integer (10)	رقم العملية التي استوجبت هذه الضمانة

Table 3-3 Collaterals table

4.جدول الأقساط و الدفعات (Payments):

ويشمل عمليات دفع الأقساط المترتبة على عمليات اقتراض سابقة.

اسم الحقل	النوع	الوظيفة
payment_id	integer (10)	رقم الدفعية أو القسط
payment_amount	big integer (15)	كمية المال المدفوع(قيمة عشوائية بين 1000 و 1000000)
last_payment	Date	تاريخ اخر دفعية (من 1\1\2000 إلى 5\9\2017)
payments_status	varchar (15)	حالة الدفعيات أو الأقساط (احدى ثلاث حالات : إما مدفوعة بالكامل paid أو جاري دفعها partially paid أو غير مدفوعة unpaid
transaction_id	integer (10)	رقم العملية التي يجري دفع أقساطها

Table 3-4 Payments table

5.جدول حسابات العملاء(accounts) :

يضم حسابات بيانات العملاء.

اسم الحقل	النوع	الوظيفة
account_id	integer (10)	رقم حساب العميل (مميز)
Balance	big integer (15)	رصيد العميل(قيمة من 1000 إلى 1000000000)
customer_id	integer (10)	رقم صاحب الحساب

Table 3-5 accounts tables

الرسم التالي يوضح هيكلية البيانات داخل قاعدة البيانات generation المعدة داخل بواسطة برنامج XAMPP :

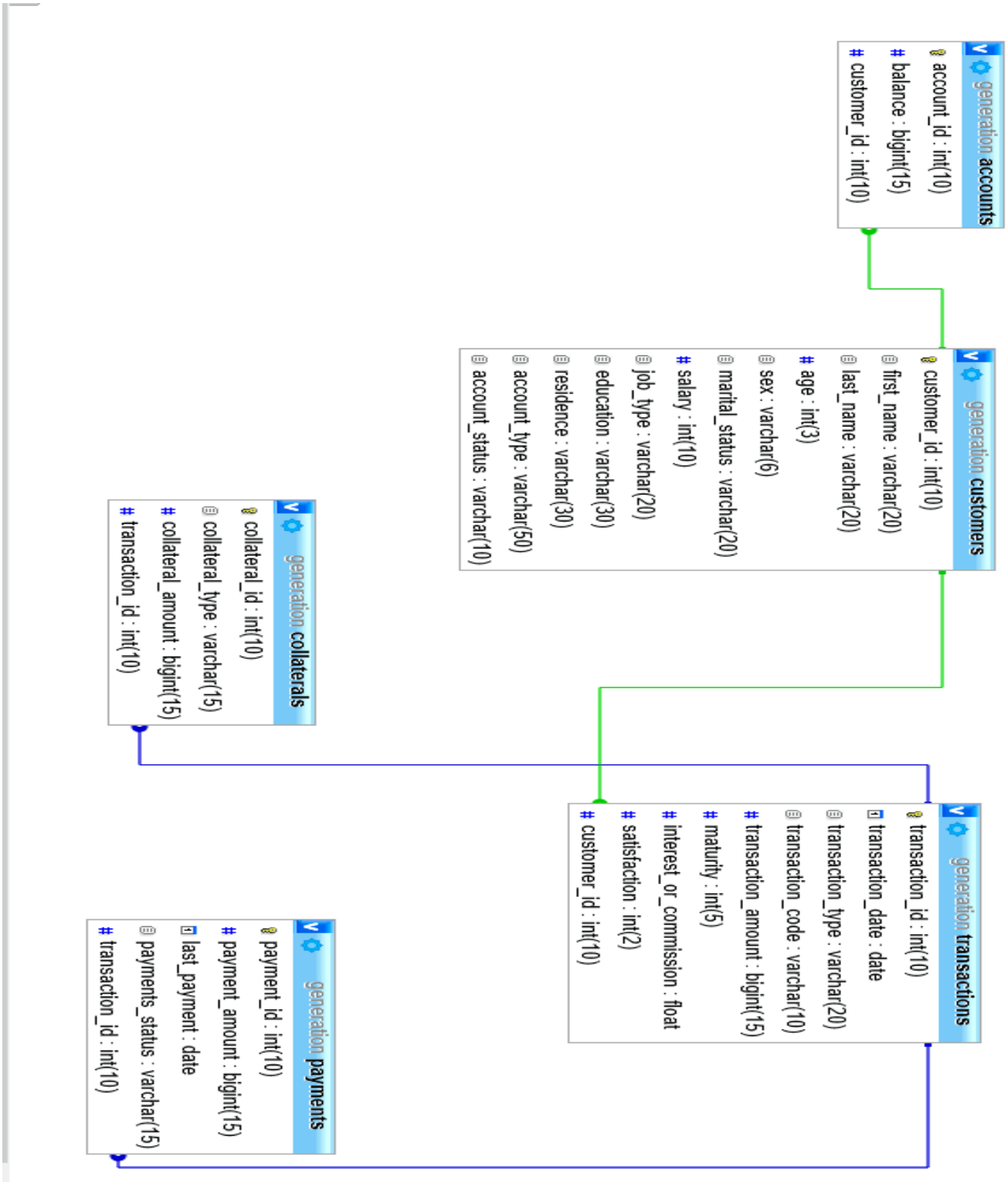


figure 3-1 bank schema

2.ربط قاعدة البيانات ببرنامج التوليد :

في الواجهة الرئيســـية لبرنامج Datanamic data generator عند خيار new project تم إختيار نوع قاعدة البيانات و ادخال اسمها و اسم المستخدم بدون كلمة مرور (وذلك لعدم وجود احداها).

ثم تلا ذلك اختيار جداول البيانات من قاعدة البيانات ، حيث تم اختيار جدول واحد في كل مرة نســـبة للحد المفروض للبيانات المولدة في كل مشروع جديد (25000 سجل)، وذلك ليتم توليد أقصى كمية ممكنة بالنسبة للجداول الخمسة .

في الخطوة السابقة تم القيام بوضع وصف للحقول الموجودة في جداول البيانات ،حيث وضحت القيم الممكنة لبعض الحقول، أما الحقول التي لم توضـح لها قيمها الممكنة مثل الاسـم الأول و الاسـم الأخير فذلك لأن قائمة الأســـماء التي يتم التوليد عبرها مدمجة مع البرنامج مع امكانية تعديلها و الاضـــافة لها،ولكن تم الاكتفاء بالقيم المزودة من قبل منتجي البرنامج نسبة لعدم تأثيرها في عملية التحليل لاحقا.

3.توليد البيانات :

كل العمليات اللازمة لجعل توليد البيانات ممكنا تمت في الخطوتين الســـابقين مثل تجهيز قاعدة البيانات وربطها ببرنامج التوليد.

ما تم في هذه الخطوة هو اختيار جدول واحد في كل مرة لكي يولد له 25000 ســجل للاســتفادة من البرنامج في أقصى صورة، وتم توليد البيانات مباشرة في جداولها.

الباب الرابع
التصميم
و
التطبيق

تصميم النظام :

الرسم التالي يوضح معمارية نظام التحليل والمعالجة و الادوات المستخدمة فيه، الذي سيتم توضيح وظائفه مفصلة في هذا الباب.

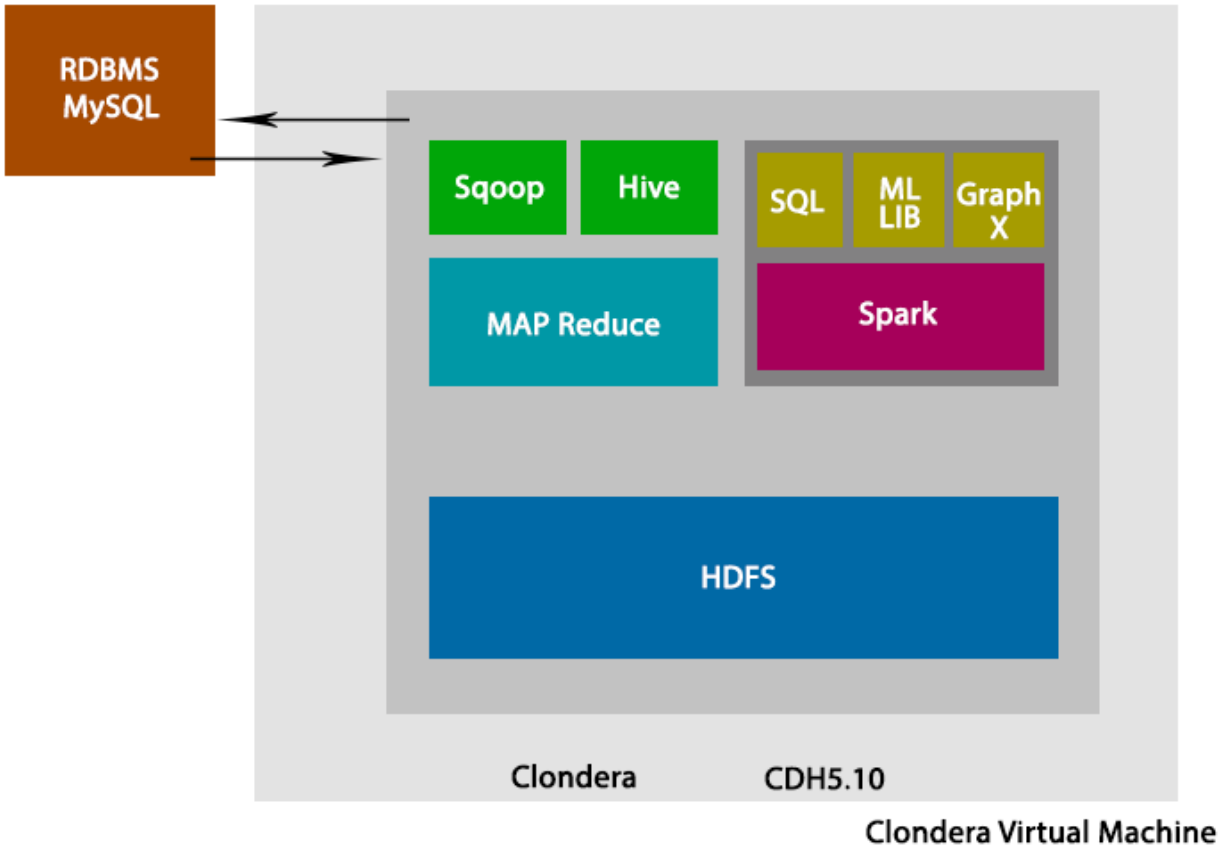


figure 4-1 system architecture

التطبيق:

كما ذكرنا في الباب السابق وأوضحنا عملية توليد البيانات ، سنوضح الان كيفية إدخال البيانات إلى نظام المعالجة Spark متبعة بكافة التفاصيل.

نقل البيانات:

في البدء تم تصـــدير قاعدة البيانات من برنامج MySQL إلى ملف ذي امتداد sql ، ثم تلا ذلك نقل قاعدة البيانات المسماة generation إلى داخل نظام Apache Hadoop عبر برنامج Apache Sqoop مع تفعيل الخيار الذي يسمح لقاعدة البيانات بالظهور في Apache Hive..

ثم تم استيراد الجداول مضغوطة بالصيغة snappy ، ولكنها لم تسمح لنوع البيانات date بالظهور بصورة صـــحيحة في الجدول transactions في العمود transactions_data،هذه المشـــكلة تمت معالجتها عبر اســـتيراد الجدول transactions والجداول الأخرى بنظام ضغط parquet file ،وهو نظام الضغط الأساسي لجداول Hive.

وهذه صور لجداول البيانات مدرجة داخل Apache Hive عبر Cloudera Hue:

جدول accounts :

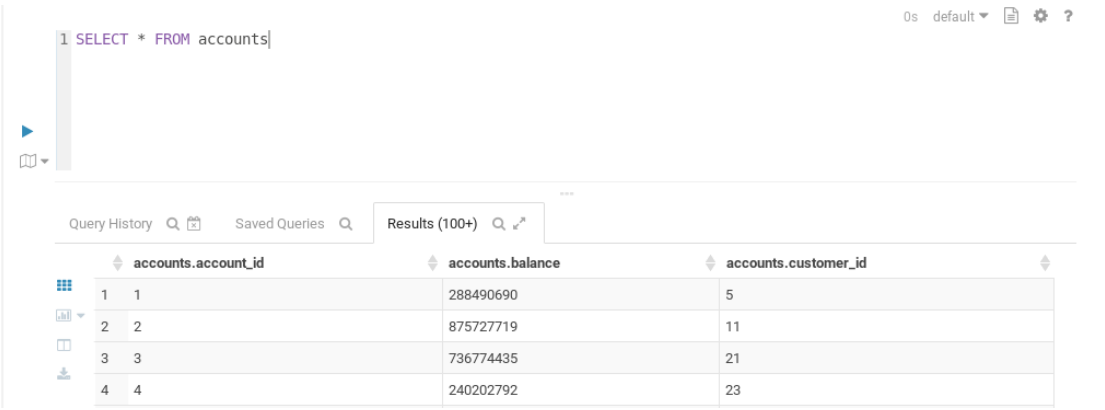
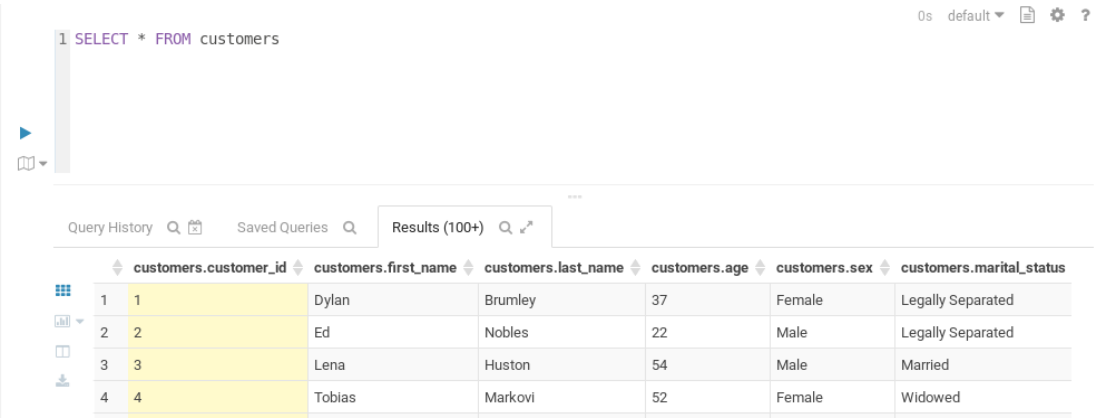


figure 4-2 account table

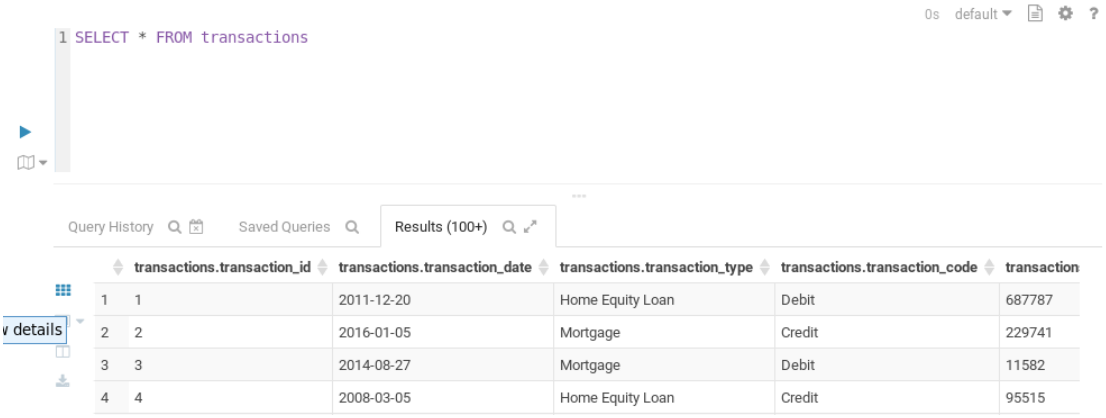
جدول customers :



	customers.customer_id	customers.first_name	customers.last_name	customers.age	customers.sex	customers.marital_status
1	1	Dylan	Brumley	37	Female	Legally Separated
2	2	Ed	Nobles	22	Male	Legally Separated
3	3	Lena	Huston	54	Male	Married
4	4	Tobias	Markovi	52	Female	Widowed

figure 4-3 customer table

جدول transactions :



	transactions.transaction_id	transactions.transaction_date	transactions.transaction_type	transactions.transaction_code	transaction
1	1	2011-12-20	Home Equity Loan	Debit	687787
2	2	2016-01-05	Mortgage	Credit	229741
3	3	2014-08-27	Mortgage	Debit	11582
4	4	2008-03-05	Home Equity Loan	Credit	95515

figure 4-4 transactions table

جدول collaterals :

```
1 SELECT * FROM collaterals
```

Query History | Saved Queries | Results (100+)

	collaterals.collateral_id	collaterals.collateral_type	collaterals.collateral_amount	collaterals.transaction_id
1	1	Government Bond	198641	8
2	2	Automobile	385837	15
3	3	Insurance	375978	17
4	4	Real Estate	559207	27

figure 4-5 collaterals table

جدول payments :

```
1 SELECT * FROM payments
```

Query History | Saved Queries | Results (100+)

	payments.payment_id	payments.payment_amount	payments.last_payment	payments.payments_status	payments.transaction_id
1	1	87945	2014-05-01	Partially Paid	6
2	2	78345	2000-12-13	DEFAULT	8
3	3	5359	2011-05-08	DEFAULT	9
4	4	29343	2017-02-06	Not Paid	13

figure 4-6 payments table

تنقيب البيانات:

ـالمعالجة الأولية:

البيانات المدرجة ليست بحوجة لعمليات معالجة أولية نسبة لكونها مولدة اليا وليست بيانات حقيقية،حيث لا توجد بها بيانات ناقصة(مفقودة)،ولكنها تحوي قيما شاذة وأخرى مكررة ، وهذه العيوب مما لا يمكن التحكم بها في عملية التوليد شبه العشوائي.

يوفر نظام Spark عدة خوارزميات لمعالجة القيم الشاذة مثل StandardScaler وNormalizer و MinMaxScaler.يقوم الStandardScaler عبر دالته setWithMean بتقريب البيانات من وسطها الحسابي عبر طرحه منها،أما الدالة الأخرى setWithStd فتقوم بتشتيت قيم البيانات المتقاربة من بعضها بمقدار الانحراف المعياري للوحدة،أما الدالة MinMaxScaler فتقوم بعمل Normalization للبيانات بين قيمة دنيا و قيمة عليا،

أما الدالة Normalizer فتقوم بمعالجة البيانات بناء على قيمة P معطاة بواسطة المستخدم.

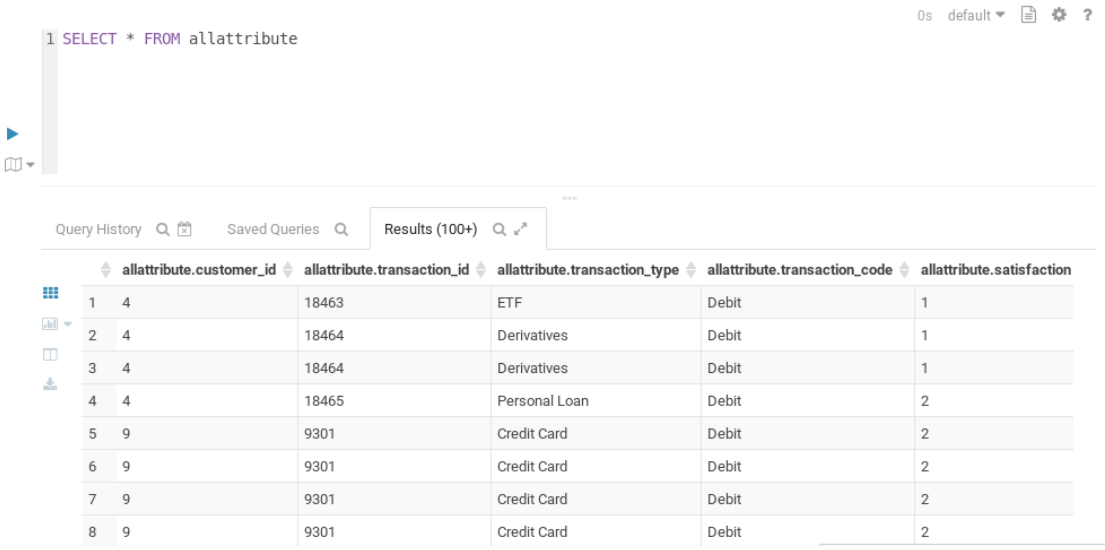
الأعمدة التي طبقت عملية الNormalization عليها هي balance في جدول account ، و transaction_amount في الجدول transactions ، ومع تجربة كافة طرق الNormalization المتاحة في Spark و مقارنة نتائجها وجدنا أن خوارزمية StandardScaler عبر دالتها Mean هي الأفضل لتجهيز البيانات، حيث جهزت البيانات بها وخزنت النتائج.

مما وجب ذكره أن عملية الNormalization قد استغرقت زمنا طويلا نسبة لعدم وجود طريقة واضحة لمعالجتها و ارجاعها إلى قاعدة البيانات،وتم ارتجال طريقة لمعالجتها من قبل الباحثين.

-تحويل البيانات:

في هذه الخطوة اختيرت الخصـــائص التي ســـيتم اجراء عمليات التجميع(Clustering) عليها ، وجمعت البيانات عن طريق (all attribute) و(CLV) تارة مع رضا العميل وتارة أخرى بدونه .

الطريقة الأولى تهتم بتجميع البيانات بناء على الخصـــائص التي يقوم معظم الباحثون بتقســيم العملاء عليها ، والطريقة الثانية كما في [18] فهي تقوم بأخذ الخصـــائص بناء على الCLV والRFM، ،وتم اســـتخدام طريقتين للتجميع لتبيين أن النموذج الجديد المقترح أدق من النماذج الســـابقة.ففي الطريقة الأولى يحوي الجدول الأول الأعمدة (customer_id,transaction_id,transaction_type,transaction_code,satisfaction) مـن الـجـدول transactions ، والأعـمـدة (account_status,job_type,account_type,age,residence) مـن الـجـدول customers ، مـع الأعـمـدة (recency,frequency,monetary) الـتـي تـم حســـابـهـا مـن الـجـدول transactions،وpayment_status من الجدول payments وbalance من الجدول accounts،



	allattribute.customer_id	allattribute.transaction_id	allattribute.transaction_type	allattribute.transaction_code	allattribute.satisfaction
1	4	18463	ETF	Debit	1
2	4	18464	Derivatives	Debit	1
3	4	18464	Derivatives	Debit	1
4	4	18465	Personal Loan	Debit	2
5	9	9301	Credit Card	Debit	2
6	9	9301	Credit Card	Debit	2
7	9	9301	Credit Card	Debit	2
8	9	9301	Credit Card	Debit	2

figure 4-7 all attribute table

والجدول الثاني يحتوي على الـCLV والـRFM ،ويحوي
(customer_id,recency,frequency,monetary,clv)

```
1 SELECT * FROM ourfeature
```

Query History | Saved Queries | Results (100+)

	ourfeature.customer_id	ourfeature.frequencey	ourfeature.monetary	ourfeature.recency
1	2	9	4546161	1933
2	4	9	5613927	1106
3	9	4	2501844	3607
4	11	9	5479155	2122
5	14	9	4580982	366
6	16	1	229741	645
7	18	36	7368936	1142
8	23	9	2317431	1879

Figure 4-8 clv without satisfaction

خوارزمية التحليل:

تجميع البيانات(Clustering) بإستخدام Kmeans :

اختيرت خوارزميات التجميع لتقسيم البيانات في هذه الدراسة لعدة أسباب: منها عدم توفر معرفة سابقة عن سلوك العميل،ومناسبة ومتطلبات المؤسسات المالية التي تسعى في الغالب إلى تقسيم عملائها إلى فئات كبيرة وذلك لطرح العروض والمنتجات التي تتناسب مع كل فئة،وكذلك مرونة هذه الخوارزميات التي تتيح تحديد عدد الفئات مسبقا،وهو ما يزيد من ملائمتها لهذا المجال على وجه الخصوص.

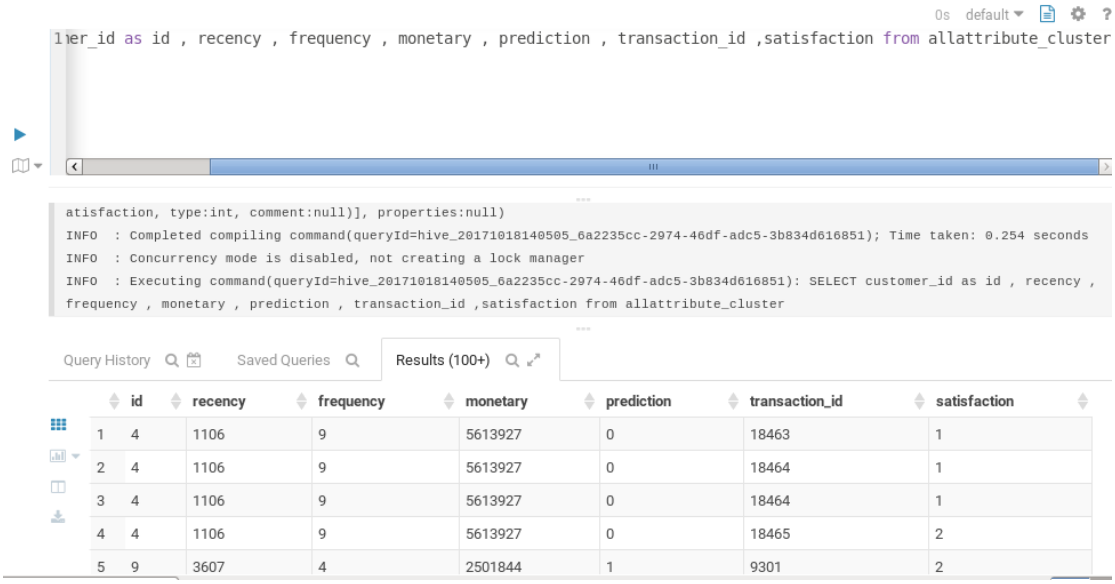
في البدء يُود توضيح قلة الخيارات المتوفرة من خوارزميات تجميع بيانات في Apache Spark 1.6.0 [21]،وهي النسخة المستخدمة لتحليل البيانات قيد الدراسة ، حيث تم اعتمادها كنسخة مستقرة من قبل شركة Cloudera المختصة في تحليل البيانات، وبالرغم من ظهور اصدارت جديدة من Spark فإنها غير مستقرة في الغالب و غير معتمدة للعمل مع الأدوات الأخرى.

معظم الخيارات المتوفرة على قلتها غير ملائمة للقطاع المستهدف (القطاع البنكي)،حيث أن الخوارزمية المناسبة لتصنيف زبائنه هي KMeans [22]، وعلى الرغم من توفر عدة نسخ منها تختلف في ألية التقسيم [21]، إلى أن النسخة التقليدية ملائمة للغرض ، ولم تذكر لها أي مشاكل من قبل السابقين،ولذا لم يكن مجال هذه الدراسة مقارنتها مع النسخ الأخرى منها،ولا مقارنتها مع خوارزميات التجميع الأخرى المتوفرة على هذا النظام .إذ أن المشكلة الأساسية هي الخصائص التي يتم تقسيم الزبائن بناء عليها ، إذ كان يتم تصنيفهم على خصائص كثيرة معممة كما هي موضحة في الجدول all_attribute الذي سيوضح لاحقا بالتفصيل، ولكن كما أوضح سابقون أن هذه الخصائص ليست دقيقة كفاية لتصنيف العملاء و إنما بناء على ما يعرف بالCLV .

لكن الـCLV يضع ربحية المؤسسة المالية في الإعتبار ويغض الطرف عن رضا العميل ، وهذا ما تم القيام بمعالجته في هذه الدراسة حيث أُضيف للـCLV متوسط رضا الزبون عن معاملاته مع المؤسسة المقاس بواسطة الـNPS ، وهو ما أظهر تحسنا ملحوظا عن التصنيف بناء على الطرق التقليدية (all attributes) كما سيوضح انفاً.

حددت طريقتان لعمل المجموعات تشترك في عدد المجموعات (k = 3) ، الأولى هي أن يتم التجميع بناء على كل الخصائص ،والتي هي
customer_id,transaction_id,transaction_type,transaction_code,satisfaction,account_statu):
s,job_type,account_type,age,residence,recency,frequency,monetary,payments_status,bala
.(nce

وبعد تطبيق خوارزمية الـ Kmeans عليها ظهرت البيانات كما هي موضحة في الشكل التالي :



	id	recency	frequency	monetary	prediction	transaction_id	satisfaction
1	4	1106	9	5613927	0	18463	1
2	4	1106	9	5613927	0	18464	1
3	4	1106	9	5613927	0	18464	1
4	4	1106	9	5613927	0	18465	2
5	9	3607	4	2501844	1	9301	2

figure 4-9 all attribute clustering output

وعند اجراء التجميع بها غير متضمنة رضا الزبون ظهرت نتائج التجميع كالأتي :

```
SELECT customer_id ,recency,frequency, monetary ,prediction,transaction_id FROM allattribute_without_satisfaction_cluster
```

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171031011414_679e6878-953b-4c31-b7b9-f61fbb5deacc): SELECT customer_id ,recency,frequency, monetary ,prediction,transaction_id FROM allattribute_without_satisfaction_cluster
INFO  : Completed executing command(queryId=hive_20171031011414_679e6878-953b-4c31-b7b9-f61fbb5deacc); Time taken: 0.002 seconds
INFO  : OK
```

Query History | Saved Queries | Results (100+)

	customer_id	recency	frequency	monetary	prediction	transaction_id
1	4	1106	9	5613927	0	18463
2	4	1106	9	5613927	0	18464
3	4	1106	9	5613927	0	18464
4	4	1106	9	5613927	0	18465

Figure 4-10 all attribute without satisfaction

أما الطريقة الثانية هي عبر الـــCLV و الـــRFM دون أخذ رضا الزبون في الاعتبار،وبعد تجميعها ظهرت البيانات كالأتي :

```
SELECT * FROM clv_without_satisfaction_cluster
```

```
INFO  : Completed compiling command(queryId=hive_20171030040000_f25b42f4-db53-4fe8-a43a-64c7162fba78); Time taken: 0.695 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171030040000_f25b42f4-db53-4fe8-a43a-64c7162fba78): SELECT * FROM clv_without_satisfaction_cluster
INFO  : Completed executing command(queryId=hive_20171030040000_f25b42f4-db53-4fe8-a43a-64c7162fba78); Time taken: 0.002 seconds
INFO  : OK
```

Query History | Saved Queries | Results (100+)

	clv_without_satisfaction_cluster.customer_id	clv_without_satisfaction_cluster.nr	clv_without_satisfaction_cluster.nf	clv_without_satisfaction_cluster.nm	clv_without_satisfaction
1	2	-0.20740357704321627	0.24473278142424049	0.24449197863961747	1
2	4	-0.71502006724985567	0.24473278142424049	0.50700863565772913	1
3	9	0.82010549745002959	-0.40926769064831003	-0.25811558212303565	2
4	11	-0.091394487987527209	0.24473278142424049	0.47387413343566465	2

Figure 4-11 CLV without satisfaction result

أما عند وضع رضا الزبون في الإعتبار أي التجميع بناء على الخصائص المقترحة بواسطة الباحثين وهي (recency, frequency, monetary, avg_satisfaction)

ظهرت النتائج كما هي موضحة في الجدول الأتي :

```
1 SELECT customer_id , nr , nf , nm ,avg_satisfaction , clv , prediction from clv_cluster
```

```
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171018215757_678a84bc-d0d9-4d6e-b861-fc0b397e9d33): SELECT customer_id , nr , nf , nm ,avg_satisfaction , clv , prediction
 from clv_cluster
INFO  : Completed executing command(queryId=hive_20171018215757_678a84bc-d0d9-4d6e-b861-fc0b397e9d33); Time taken: 0.01 seconds
INFO  : OK
```

Query History | Saved Queries | Results (100+)

	customer_id	nr	nf	nm	avg_satisfaction	clv	prediction
1	2	-0.20740357704321627	0.24473278142424049	0.24449197863961747	2.6666666666666665	0.20796477311203343	1
2	4	-0.71502006724985567	0.24473278142424049	0.50700863565772913	1.3333333333333333	0.32488286493019886	1
3	9	0.82010549745002959	-0.40926769064831003	-0.25811558212303565	3.5	-0.21869489308012935	1
4	11	-0.091394487987527209	0.24473278142424049	0.47387413343566465	4.333333333333333	0.3554495665127646	2
5	14	-1.1692355482086381	0.24473278142424049	0.25305292974128923	4.333333333333333	0.13521007601084067	2

figure 4-12 clv clustering output

النتائج الموضحة أعلاه لعملية التجميع سيتم مناقشتها و تناولها بإسهاب في الباب التالي.

استخلاص قواعد الارتباط:

أوجدت قواعد الارتباط (association rules) للمنتجات التي اشتراها كل العملاء ضمن كل فئة على حدة الذين حددوا في الخطوة السابقة، وذلك لأن سلوك العملاء داخل الفئة الواحدة متجانس نسبة لتقارب صفاتهم. والمدخلات كانت في الصورة الاتية:

```
1 SELECT * FROM association_solved_0
```

```
INFO  : Completed compiling command(queryId=hive_20171029093939_5e340e42-50c4-473c-9ed9-3ff837df4a23); Time taken: 0.225 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171029093939_5e340e42-50c4-473c-9ed9-3ff837df4a23): SELECT * FROM association_solved_0
INFO  : Completed executing command(queryId=hive_20171029093939_5e340e42-50c4-473c-9ed9-3ff837df4a23); Time taken: 0.0 seconds
INFO  : OK
```

Query History | Saved Queries | Results (100+)

	association_solved_0.customer_id	association_solved_0.mortgage	association_solved_0.credit_card	association_solved_0.stocks	association_solved_0.bonds	association_sol
1	3148	0	0	0	0	0
2	5109	0	0	0	0	0
3	3565	0	1	0	0	0
4	7817	0	0	0	0	0

Figure 4-13Association rule input table

لإيجاد قواعد الارتباط لأي فئة تم تحويل القيم المخزنة في جدول عملاء أي فئة من قيم numerical (صـــفر و واحد كانت ترمز لشـراء العميل لهذا المنتج من عدمه)إلى binominal حيث أن الخوارزمية تسـتقبل المنتجات في صـورة (true / false)،بعد ذلك تم ايجاد القيم الشـــائعة عبر خوارزمية FPGrowth عبر اعطائها (0.05) كأقل قيمة ممكنة للدعم (support)ومن ثم تخزين هذه النتائج واســـتخدامها كمدخل لخوارزمية association rule لإيجاد القواعد عبر اعطائها أقل حد للثقة (confidence) هو(0.8) ، وهذه القيم للدعم و الثقة استخدمت من قبل باحثين سابقين في نفس المجال [22]بعد ذلك تم تخزين نتائج كل فئة على حدة لإقتراح المنتجات المناسبة لها لاحقا.

الصور التالية توضح قواعد الارتباط الناتجة لكل فئة:

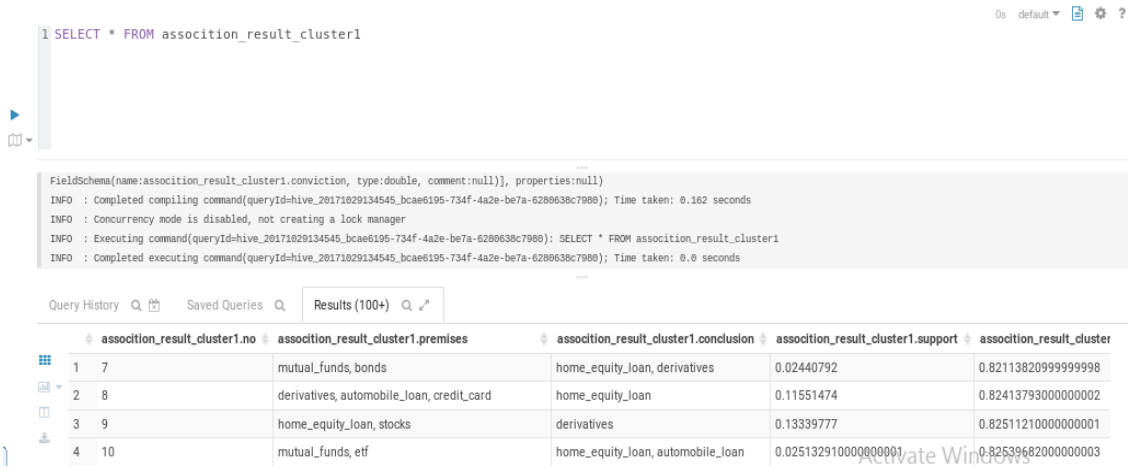
```
SELECT * FROM asscition_result_cluster0
```

```
FieldSchema(name:asscition_result_cluster0.conviction, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20171029134545_ade5b615-e76f-4a32-be89-b72e201ea615); Time taken: 0.288 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171029134545_ade5b615-e76f-4a32-be89-b72e201ea615): SELECT * FROM asscition_result_cluster0
INFO  : Completed executing command(queryId=hive_20171029134545_ade5b615-e76f-4a32-be89-b72e201ea615); Time taken: 0.0 seconds
```

Query History | Saved Queries | Results (100+)

	asscition_result_cluster0.no	asscition_result_cluster0.premises	asscition_result_cluster0.conclusion	asscition_result_cluster0.support	asscition_result_cluster0.confi
1	8	personal_loan, stocks	etf	0.13918170399999999	0.81656184499999995
2	9	home_equity_loan, stocks	etf	0.13864570300000001	0.81770284500000001
3	10	personal_loan, automobile_loan	etf	0.14579238899999999	0.81845536600000002

Figure 4-14 Association result cluster 0



```
SELECT * FROM asscition_result_cluster1
```

```
FieldSchema(name:asscition_result_cluster1.conviction, type:double, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20171029134545_bcae6195-734f-4a2e-be7a-6280638c7980); Time taken: 0.162 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20171029134545_bcae6195-734f-4a2e-be7a-6280638c7980): SELECT * FROM asscition_result_cluster1
INFO  : Completed executing command(queryId=hive_20171029134545_bcae6195-734f-4a2e-be7a-6280638c7980); Time taken: 0.0 seconds
```

Query History | Saved Queries | Results (100+)

	asscition_result_cluster1.no	asscition_result_cluster1.premises	asscition_result_cluster1.conclusion	asscition_result_cluster1.support	asscition_result_cluster
1	7	mutual_funds, bonds	home_equity_loan, derivatives	0.02440792	0.82113820999999998
2	8	derivatives, automobile_loan, credit_card	home_equity_loan	0.11551474	0.82413793000000002
3	9	home_equity_loan, stocks	derivatives	0.13339777	0.82511210000000001
4	10	mutual_funds, etf	home_equity_loan, automobile_loan	0.025132910000000001	0.82539682000000003

Figure 4-15 Association result cluster 1

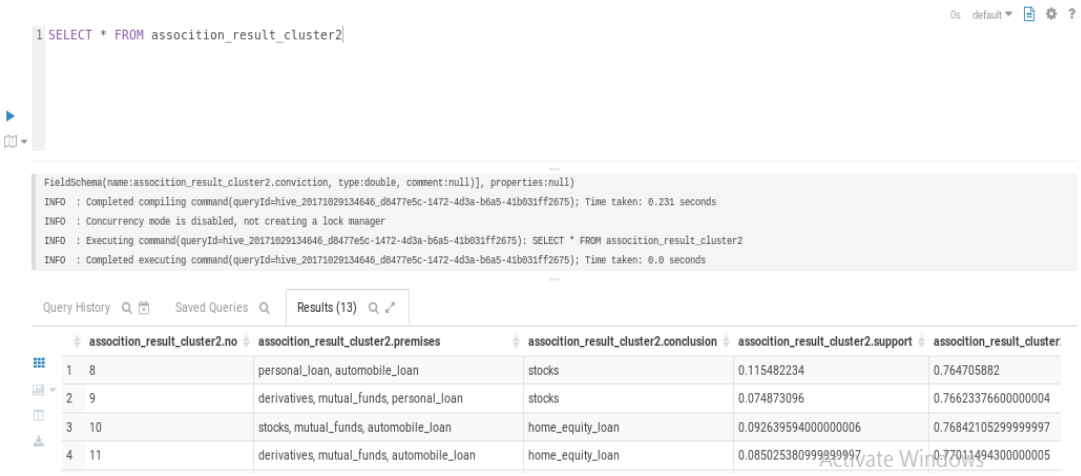


Figure 4-16 Association result cluster 2

وهذه القواعد الناتجة سيتم تقييمها في الباب التالي

استخدام النتائج لمسؤولي المبيعات في المؤسسات المالية :

تم اسـتخدام بعض من التقنيات المذكورة سـابقا في بناء موقع تفاعلي و ذلك لتمكين المسـتخدمين من سـهولة استعراض النتائج و تقديم الاقتراحات.الواجهات وطريقة استخدامه موضحة في الملاحق.

الباب الخامس

النتائج
و
المناقشة

في هذا الباب سيتم تناول النتائج التي تم التوصل إليها عبر أدوات تنقيب البيانات،وذلك بشرحها واستخدام طرق متعددة للتحقق منها :

تقييم نتائج Kmeans:

عرضت البيانات في صورة مخططات (Pie Chart) بإستخدام Apache Zeppelin وذلك لتوضيح نسبة العملاء في كل فئة عبر استخدام الطريقتين المذكورتين سابقا ،وذلك في حالة ادخال رضا الزبون في المعادلة و في حالة عدمه.

أولا:

بوجود رضا العميل(satisfaction) :

الرسم التالي يوضح كل تقسيم العملاء بإستخدام كل الخصائص مضافا إليها رضا العميل:

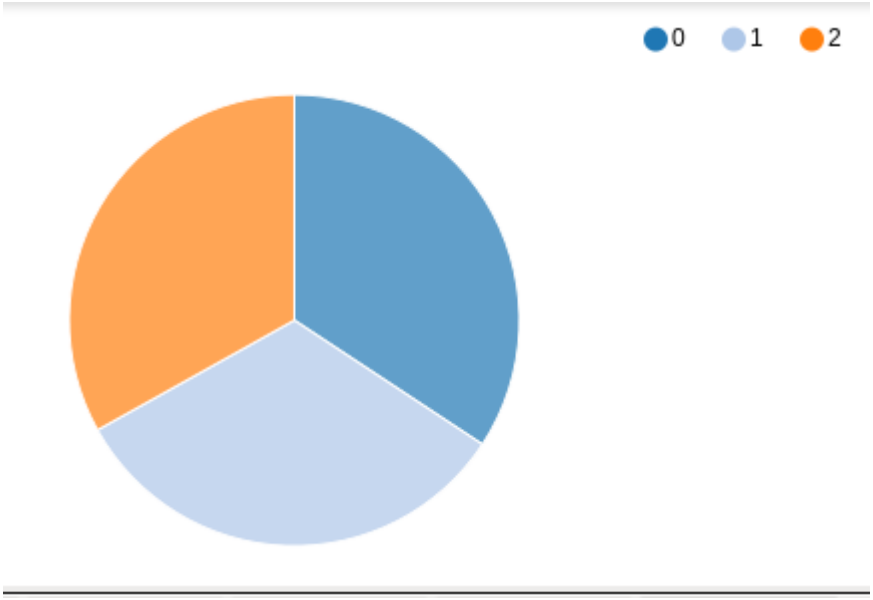


Figure 5-1 All attributes including satisfaction

الفئة (0) تحوي 34% من العملاء،أما الفئة (1) فتحوي 33% من العملاء ،بينما الفئة (2) تحوي 33% من العملاء.

أما الرسم الثاني فيوضح تقسيمهم بناء على الـCLV والـRFM مع أخذ رضا العميل في الاعتبار ،وهو النموذج المقترح:

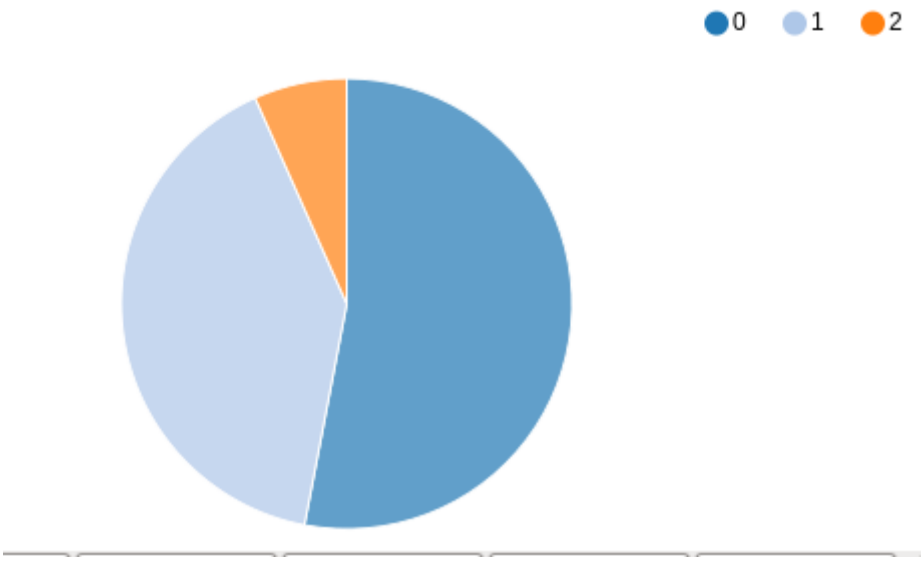


Figure 5-2 Proposed features (satisfaction in addition to CLV)

عند استخدام الصفات المقترحة لوحظ اختلاف في نسبة توزيع العملاء ، حيث احتوت الفئة (0) على 53% من العملاء،بينما احتوت الفئة (1) على 40% من العملاء ،أما الفئة (2) فاحتوت على 7% المتبقين من العملاء.

ثانيا :

بعدم وجود رضا العميل(satisfaction):

تقسيم الفئات بإستخدام all attributes:

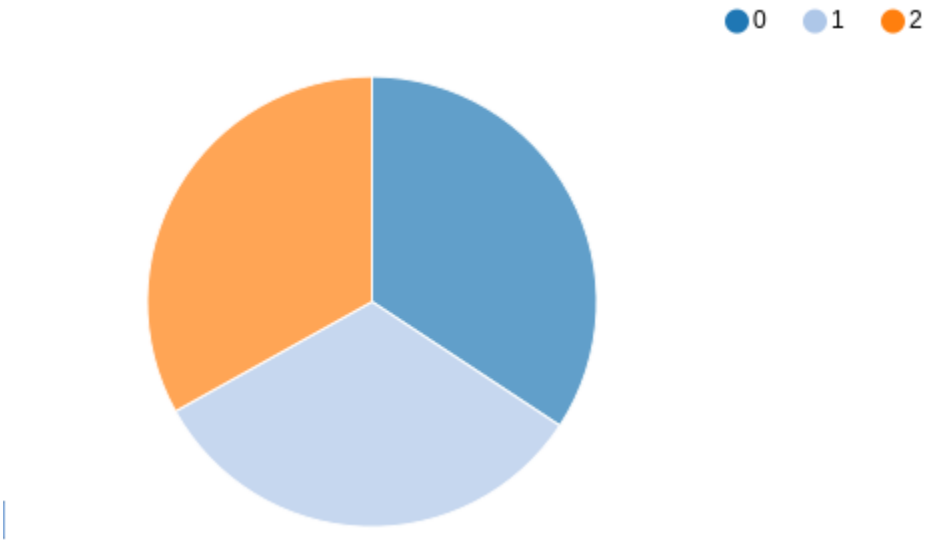


Figure 5-3 All attributes without satisfaction

حيث احتوت الفئة (0) على 34% ،أما الفئة (1) فتحوي 33% من العملاء،أما الفئة (2) فتحوي 33% من العملاء.

أما بإستخدام الCLV:

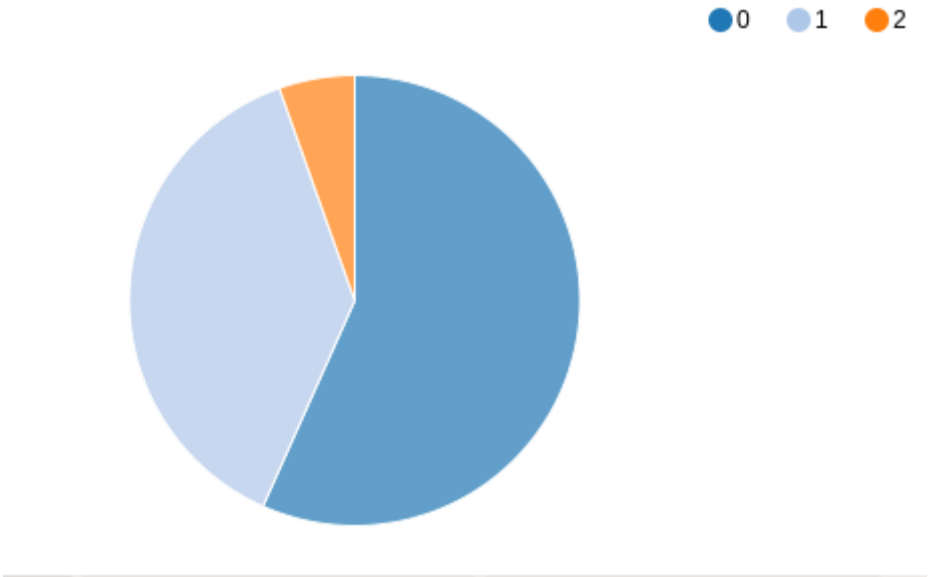


Figure 5-4 CLV without satisfaction

حيث احتوت الفئة (0) على 57% من العملاء،أما الفئة (1) فاحتوت على 38% من العملاء،بينما احتوت الفئة (2) على 5% من العملاء.

جدول مقارنة يوضح نسبة توزيع العملاء في كل فئة عند استخدام الsatisfaction وبدون استخدامه :

All attributes		CLV		رقم الفئة
With satisfaction	Without satisfaction	With satisfaction	Without satisfaction	
%34	%34	%53	%57	0
%33	%33	%40	%38	1
%33	%33	%7	%5	2

Table 5-1 comparison table

نلاحظ من الجدول السابق أنه ليس هناك فرق في النتائج عند التصنيف بإستخدام All attribute ، حيث ظهر فرق عددي بسيط غير مؤثر على النسبة الكلية ، أما عند استخدام CLV للتصنيف فالفرق كان ملحوظا عند ادخال رضا الزبون ضمن الصفات المؤثرة في عملية التجميع بإستخدام CLV ، حيث زاد العملاء المحتملون و الأساسيون بنسبة 2% وهي نسبة غير ضئيلة بالنسبة لقاعدة زبائن المؤسسات العملاقة ، حيث أن زيادة العملاء الأساسيين تعني أرباحا أكثر لم تكن مستهدفة بالطريقة المثلى،أما الزيادة في العملاء المحتملين فتعني تعديل الحملات الدعائية والعروض التشجيعية لتتناسب مع قدر أكبر من العملاء،وكلتا الفئتين توضح زيادة متوقعة في الربحية.

صحة التصنيف المتوصل إليه يمكن تقييمها من ناحيتين هما الناحية التجارية وذلك بتوضيح ما إذا كانت صفات و نسب العملاء في كل فئه تتناسب مع ما هو موجود في الواقع،أما الناحية الأخرى فهي الناحية الإحصائية لتوضيح جدوى الفئات المقسمة من الناحية الرياضية.

التحقق من ناتج عملية التجميع من الناحية التجارية:

وجد أن العملاء عموما مقسمون بناء على قيمة الـــCLV والـــRFM في نظر المؤسسات المالية على النحو الأتي:

-العملاء الأساسيون،وهم اولئك أصحاب القيمة المادية العالية والولاء التام للمؤسسة،وتكون قيمهم R⇩, F⇧ and M⇧ .

-العملاء المحتملون،وهم الذين من المتوقع ولاؤهم للمؤسسة عند تعاملهم معها، حيث تكون قيمهم R⇧, F⇩ and M⇧ .

-العملاء المفقودون،وهم الذين لا يمكن التأكد عما إذا كان ما يزالون عملاء للمؤسسة ، وقيمهم تكون R⇧, F⇩ and M⇩ .

الجدير بالذكر أيضا أن متوسط رضا العملاء المقاس بواسطة NPS يتناسب طرديا وولاء العميل ، حيث أنه من البديهي عند رضا العميل عن المنتج أو الخدمة طلب المزيد منها وهو السلوك المتمثل في العملاء الأساسيين، وتقل الكميات المشتراه لاحقا قليلا في العملاء المحتملين ، وذلك لقيام بعضهم بتلبية احتياجاته عبر منتجات مؤسسة أخرى،أما العملاء المفقودون فهم من لم تناسبهم خدمات المؤسسة بتاتا وغير راضين عنها،لذلك لا يمكن التأكد عما إذا كانوا سيستمرون في الشراء من المؤسسة نسبة لعدم وجود البديل،أو تركها من أجل بديل أفضل.

بناء على المعلومات الموضحة أعلاه،سيتم شرح النتائج التي توصل إليها،حيث تم تقسيم البيانات إلى ثلاثة تجميعات مميزة ، كل منها يحمل الخصائص الأتية:

-الفئة (0): وعملاء هذه الفئة يتسمون بخصائص العملاء المفقودين.

-الفئة (1): وهم من يتسمون بخصائص العملاء المحتملين.

-الفئة (2): وهم من يحملون خصائص العملاء الأساسيين.

الجدول التالي يوضح متوسط الـRFM ورضا العملاء المتحصل عليه لكل فئة:

Average (satisfaction)	Average (CLV)	Average (Monetary)	Average (frequency)	Average (recency)	الفئة
2.826	-0.179	-0.196(M⇩)	-0.201(F⇧)	0.029(R⇧)	0
7.985	-0.216	-0.239(M⇧)	-0.249(F⇩)	0.087(R⇧)	1
4.911	2.409	2.650(M⇧)	2.734(F⇧)	-0.663(R⇩)	2

table 5-2 result

الجدير بالذكر أن رضا العملاء في كل فئة لم يأتي كما هو متوقع بالنسبة للعملاء الأساسيين،حيث أن الزبائن الموالين للمؤسسة يتميزون و المعاملات ذات القيمة المادية الكبيرة و المتكررة،االشيء الذي يتناقض و متوسط رضا لزبون منفر(detractor) حيث أنه من البديهي أن يترك المؤسسة إلى أخرى إذا لم تناسبه خدماتها ، ولكن يمكن أن تفسر أيضا على أنهم لا يجدون البديل الذين يوفر لهم هذه الخدمات بالكمية التي يحتاجونها.حيث أن الفئتين 0 و 1يمكن أن يلبوا هذه الحوجة غير المشبعة من منافسين أخرين،ولكن لا يمكن لأصحاب الفئة 2 استخدامها.

أما في حالة وجود البديل المناسب لأصحاب الفئة 2 ، فلا يملك الباحثون إلا أن يقروا أن هذا ليس إلا خللا في جودة البيانات المولدة،حيث أنه لايمكن التحكم في علاقة المتغيرات مع بعضها،وهذا مما لم يمكن للباحثين فعل شيء حياله.

التحقق من نتائج عملية التجميع احصائيا:

للتحقق منها استخدمت (WSSE) وهي اختصار لـ(Within Set of Squared Errors) وهي عبارة عن الدالة المستخدمة في خوارزمية الـKMeans لتوزيع البيانات في الفئات، حيث أن الـKMeans تحاول تقليل الأخطاء (المسافات) عن مركز الفئة(*).

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x(j)-u(i))^{2}$$

أما هذه الدالة فتقوم بجمع مربعات أبعاد العناصر عن مركز الفئة،بحيث تقلل مجموع المسافات الكلي.وتعمل على حساب مركز كل فئة الذي يمثله متوسط العناصر في كل فئة،وتقوم بالعمل بناء على الخطوات الأتية:

1.اسناد كل نقطة لمركز الفئة الأقرب إليها.

2.حساب مراكز الفئات الجديدة الناتجة عن الخطوة السابقة.

ويتم تكرار هذه العملية حتى يتم الوصول إلى الحد الأقصى من الـــIterations أو عدم حصول تغير في الفئات،وتستخدم هذه الدالة لإيجاد أفضل قيمة رياضية لعدد الفئات ، حيث أن K المناسبة هي التي تعطي أقل قيمة ممكنة لWSSE.

وقبل تطبيق هذه الدالة على البيانات قيد الدراسة التي تم تجميعها في ثلاث فئات سابقا ، تم قسمة هذه البيانات بنسبة 80% للتدريب ، و20% للإختبار،بعدد Iterations يساوي 10 ،وفي كل مرة يتم تجميعها بإستخدام عدد فئات k مختلف ومن ثم حساب الWSSE بإستخدام الدالة computeCost ، وكانت النتائج كما هي موضحة في الجدول الأتي:

K	WSSE
2	10194.8431
3	8497.4810
4	6140.3612
5	5214.0403
6	4961.2353
10	3522.4329
15	2065.6829
20	2529.8937

Table 5-3 WSSE

من الجدول السابق نجد أن أنسب قيمة لK هي 15 ، وأغفلت النتائج السابقة و التالية لها نسبة لعدم وجود تغير في النمط التتابعي لها أي عند k > 15 نجد أن WSSE(k) أكبر من WSSE(15) ،وعندما تكون قيمة k < 15 نجد أن WSSE(k) أقل من WSSE(15) .

القيمة 15 مناسبة من الناحية الرياضية ،ولكنها غير مجدية في عالم الأعمال،إذ أنه من الكلفة بمكان تقديم 15 عرضا مختلفا للفئات المختلفة ،وصعوبة تعامل موظفي المبيعات مع هذا الكم الواسع من الخيارات ، أضف إلى ذلك أنه لا تتوفر ألية عملية لإستهدافهم ،إذ أن الغرض من تقسيم العملاء إلى فئات هو إيجاد صفات مشتركة عامة بين المندرجين داخل هذه الفئة ، لا تفصيل العروض لتناسبهم بنسبة 100% ،إذ أن هذا الاختيار يتم بأخذ الأرباح و التكاليف المحتملة في الاعتبار، وهو ما يختلف عن النماذج العلمية البحتة.

تقييم قواعد الارتباط الناتجة من الناحية التسويقية:

العملاء المفقودون ظهرت لهم 136 قاعدة ارتباطية ناتجية من دعم 5% وثقة 80% ، أما العملاء المحتملون فظهرت لهم 162 قاعدة ارتباطية لقيمة دعم 5% وثقة 80% ، بينما نتجت 13 قاعدة ارتباطية للعملاء الأساسيين لدعم 5% وثقة 75% ، وتم انقاص قيمة الثقة بالنسبة للعملاء الأساسيين نسبة لظهور قاعدة ارتباطية واحدة قيمة الثقة لها مساوية أو أكبر من 80% ،إذ أنها غير كافية بالنسبة للمؤسسات بغض النظر عن دقتها،لذلك لعدم توفيرها لخيارات مرنة لهذه الفئة القيمة للمؤسسة،الشيء الذي يتضارب مع المنتجات المتنوعة التي توفرها المؤسسة ، والمرونة التي يحتاجها العميل لتلبية احتياجاته المالية.

جدول يوضح عدد قواعد الارتباط الناتجة لكل فئة :

رقم الفئة	عدد العملاء	الدعم(Support)	الثقة(Confidence)	عدد القواعد الناتجة
0	5597	%5	%80	136
1	4138	%5	%80	162
2	788	%5	%75	13
المجموع	**10523**			**311**

Table 5-4 association rule

تقييم قواعد الارتباط الناتجة من الناحية الاحصائية:

لتقييم جودة القواعد المشتقة استخدم مؤشر يعرف بالـــlift ، وهو يستخدم لتحديد نوع العلاقة بين المدخلات و النواتج ، حيث ينتج رقما له ثلاثة تفسيرات ممكنة، فإذا كانت قيمة الـــlift الناتجة أقل من واحد فتعني أن العلاقة عكسية بين الطرفين أي وجود أحدهما يقتضي عدم وجود الأخر، أما إذا كانت القيمة مساوية للواحد فتعني أن القيم مستقلة،أما إذا كانت القيمة أكبر من واحد فتعني علاقة طردية بين المتغيرات مما يعني جودة القواعد الناتجة سابقا.

وعندما استعلم عن القواعد الارتباطية التي تفوق قيمة الـــــlift لها الواحد ، ظهرت 311 نتيجة ،وهي كل القواعد الناتجة،والصورة التالية توضح ذلك:

Figure 5-5 Lift result

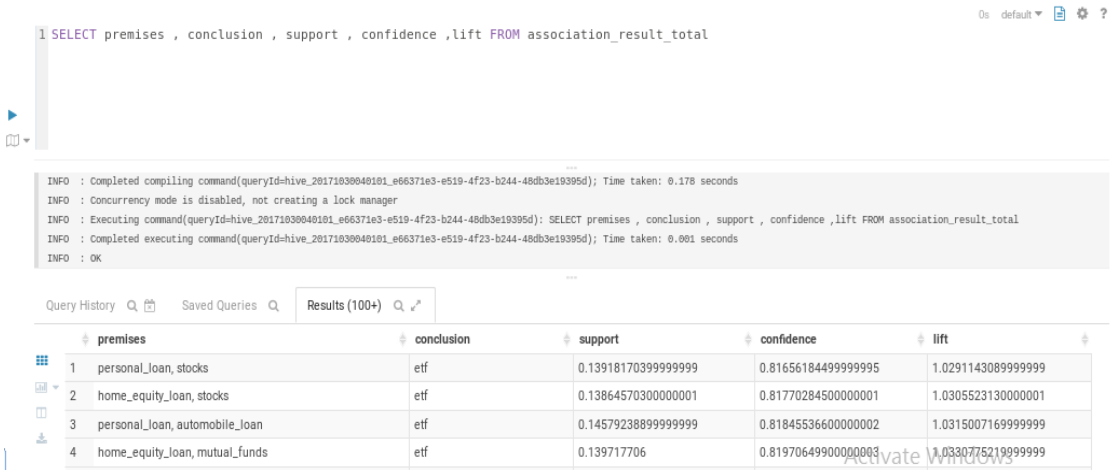


Figure 5-6 Association result

الباب السادس

الخلاصة

و

التوصيات

مقدمة :

هذا الباب يتناول ما تم الوصول إليه من خلال هذه الدراسة ، وتوضيح المشاكل و المعوقات التي واجهت هذا البحث،ثم توضيح بعض التوصياتب التي يمكن للباحثين القادمين مواصلة العمل بها.

الخلاصة:

حولت البيانات المخزونة من نظام قواعد بيانات تقليدية إلى نظام تخزين ومعالجة موزعة لتتناسب مع زيادة حجم البيانات المخزنة عن العملاء وسهولة التوسعة والمعالجة لهذه الكمية الضخمة من البيانات، حيث أديرت عملية تخزين البيانات عبر Hadoop والوصول إليها عبر Hive ،أما معالجة هذه البيانات فتمت عبر Apache Spark ومن ثم خزنت النتائج في Hadoop للوصول إليها لاحقا لعمل الإقتراحات أو ادخالها في معالجات لاحقة.

قسم العملاء إلى ثلاث فئات معلومة يتميز عملاء كل فئة منها بصفات تختلف عن الأخرين من الناحية المادية والسلوك الاستهلاكي والرضا العام للعملاء عن المنتجات المقدمة لهم ،وبناء على ذلك تم ايجاد قواعد ارتباط تساعد مسؤولي التسويق لدى المؤسسات المالية على اقتراح منتجات مناسبة لإحتياجات أمثاله.

المشاكل والمعوقات:

واجهت هذه الدراسة عدة مصاعب هي :

1. عدم توفر بيانات بنكية ومالية حقيقية للتحليل الأكاديمي،وعدم وجود سياسة منح بيانات للباحثين.
2. ضعف امكانات الأجهزة المتاحة للمعالجة.
3. التكلفة العالية لترخيص أدوات توليد البيانات وعدم وجود بدائل مجانية.
4. حداثة الأدوات المستخدمة التي توجد ببعضها أخطاء (bugs) ووجود العديد من مشاكل التوافق بين هذه الأدوات مما يعيق الإستخدام الفعال لها.
5. صغر المجتمع الالكتروني المتوفر لدعم هذه الأدوات،وعدم توفير الشركات للحلول التي تجدها في هذه الأدوات للمجتمع مفتوح المصدر.
6. ندرة الخبراء المحليين و الإقليميين في هذا المجال و صعوبة الوصول إليهم.
7. قلة المراجع التقنية المتوفرة وعدم تغطيتها لكافة المشاكل التي يمكن مواجهتها من قبل الباحثين الجدد.
8. عدم القدرة على دمج PHP وXAMPP مع centOS و Apache Hive.

التوصيات:

نأمل من الباحثين القادمين :

1. تصميم نظام دعم اتخاذ قرار بناء على النتائج المتوصل إليها.
2. تصميم نظام يعمل على معالجة البيانات اللحظية و دمجه بموقع تفاعلي لعرض النتائج.
3. استخدام هذا النظام على بيانات شبه مهيكلة وغير مهيكلة لإستخلاص المعرفة منها.

الخاتمة :

استخدمت تقنيات البيانات الضخمة وذلك وتقسيم العملاء إلى ثلاث فئات بناء على سلوكهم الإستهلاكي ،ثم اقتراح الخدمات الملائمة بناء على النتائج التي توصل إليها النظام ،ثم عرضت هذه الإقتراحات عبر واجهة تفاعلية.

المصادر و المراجع

[1] "Applying latent trait analysis in the evaluation of prospects for cross selling of ،W. A. K. e. al February 1991. "،financial services

[2] Hadoop the definitive guide 4th edition. ،T. White

[3] Cross selling of financial products - a study based on customers of Kerala"."

[4]

[5] .2016 "،customers' satisfaction in Banks " Techniques for measuring ،E. Lidia

[6] USA: Morgan Kaufmann ،Data Mining Concepts and Techniques ،M. K. a. J. P. Jiawei Han .2012 ،Publishers is an imprint of Elsevier

[7] Data Mining Techniques For Marketing, Sales and Customer Relationship ،M. J. B. a. G. S. Linoff .2004 ،Wiley Publishing, Inc Indiana: ،Management

[8] Advanced Analytics with Spark. ،U. L. S. O. a. J. W. Sandy Ryza

[9] .2016 ،Big Data Analytics ،V. Ankam

[10] .2012 ،Laravel Starter ،S. McCool

[11] .2015 ،Modern PHP : New Features and Good Practices ،J. Lockhart

[12] .2007 ،Learning MySQL ،S. M. ". T. a. H. E. Williams

[13] .2014 ،Learning Web Application Development ،S. Purewal

[14] " cross selling through database marketing : a mixed data factor analyzer for data ،W. A. K. e. al May 2002. "،augmentation and prediction

[15] "Cross Selling through database marketing-mixed data factor". ،Kamakura

[16] "cross-selling sequentially orderd products:an application to consumer banking ،b. s. ,. r. t. shibo li services".

[17] *Journal of* "،"Cross-selling models for telecommunication services ،S. Jaroszewiczx *telecommunication and information technology.*

[18] "،"Developing Up-Selling and Cross-Selling Data Mining Model forBanking Sector ،m. ,. n. Fatima .2013 ،*UNIVERSITY OF KHARTOUM*

[19] K. Gordon، "Principles of data management: Facilitating information sharing"، 2007.

[20] "Datanamic - main page"، [متصل]. Available: www.datanamic.com. [تاريخ الوصول 3 9 2017].

[21] A. H. M. Zekic-Susac، "Predictive analytics and strategies in Big Data platforms – comparison"، 2016.

[22] M. A. F. a. .. S. Mohammadi، "Mining the Banking Custome Behaviour Using Clustering and Association Rules Methods"، *International Journal of Industrial Engineering and Production Research, vol. 21, no. 2008-4889*، pp. 239-245، 2010.

الملاحق

في هذا الملحق عرض لواجهة الويب التي تم تصميمها بإستخدام Laravel framework لعرض النتائج التي تم التوصل إليها بإستخدام Apache Spark:

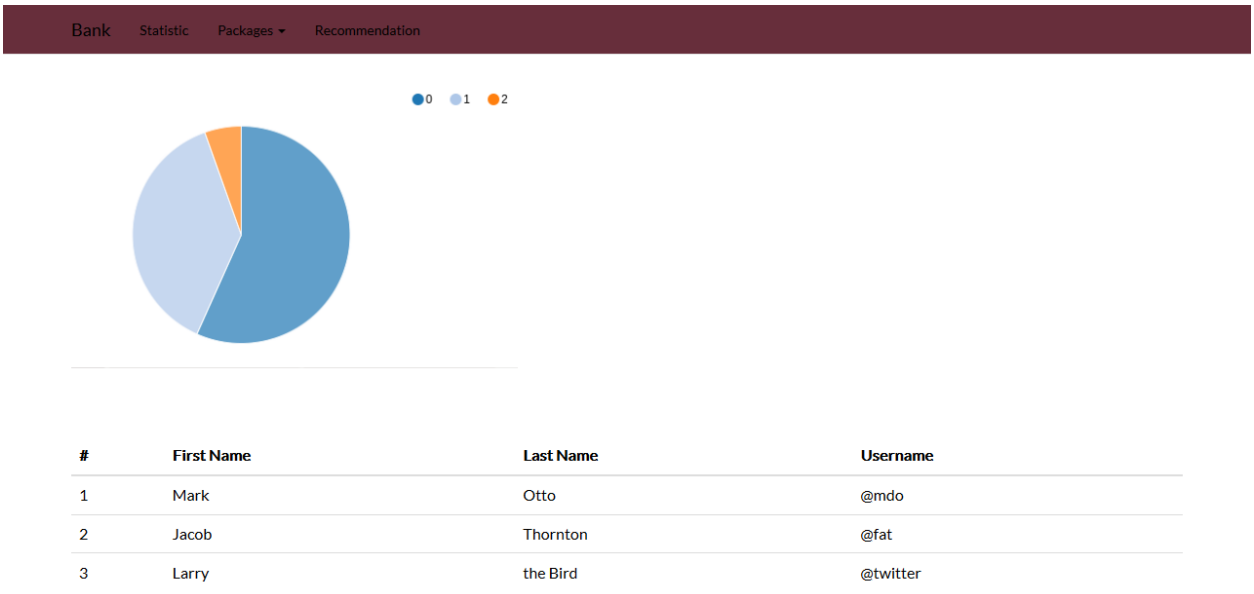


Figure A-1 Statistics Page

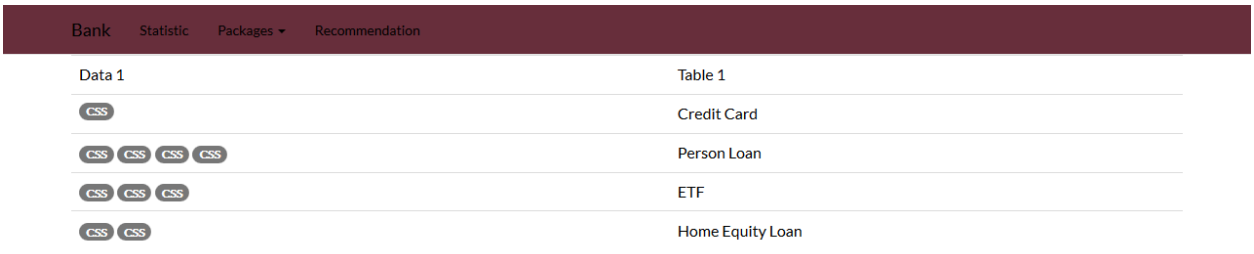


Figure A-2 Page Displaying Recommended Packages