الجمهورية العربية السورية
وزارة التعليم العالي
المعهد العالي لإدارة الأعمال
اختصاص إدارة العمليات والمعلومات

استخدام تقنيات التنقيب في البيانات للتنبؤ بتسرب الزبائن في شركة اتصالات
(حالة عملية: شركة X)

The implementation of data mining techniques to predict customer churn in a telecom company
(A case study X Company)

إعداد الطالبة
جودي صابوني

إشراف
د. ياسر رحال
م. نظرة رحمة

السنة الدراسية: الخامسة
العام الدراسي: 2020-2021

"مشروع أعد لنيل درجة البكالوريوس في إدارة الأعمال اختصاص إدارة العمليات والمعلومات"

الشكر و التقدير

أتقدم بالشكر الجزيل إلى أستاذي و مشرفي
الدكتور ياسر رحال
المتفاني و الداعم في طريق العلم ، على كل ما قدمه من نصح و مقترحات أدت إلى نجاح و رفع مستوى البحث.

شكرا جزيلا إلى مشرفتي و أستاذتي الداعمة و المساعدة و المتفانية ،
الأستاذة المهندسة نظرة رحمة
على فضلها الكبير جداً في الإشراف على مراحل إتمام هذا البحث و لم تبخل بتقديم النصح و العون في كل الأوقات .

لكل من علمني و كان جزءاً من هذه المسيرة الغنية بالمعرفة و العلم و المحبة إلى أساتذتي

د.وائل خنسة ، د. راضي خازم ، د.كادان جمعة ، د.محمد عنتور

إلى من كانوا لي خير سند ومساعد وداعم ، الذين ساهموا بكامل إمكانياتهم لإتمام هذا العمل

أخوتي و زملائي و شركائي الأعزاء

أخص بالشكر

روان الرفاعي، محمد براء قتابي، عبد الرحمن دياربكرلي

الإهداء

إلى من علمني جميع معاني الصبر و الإرادة و السعي ، من كان لي خير بيت و أمان و سلام ، إلى من له الفضل في جميع ما نجحت به ووصلت إليه ، إلى من كان قوتي في جميع مراحل حياتي ، إلى من أفتخر به أين ما كنت، إلى حبيبي الأول و الأخير ...**أبي**

إلى معلمة كلماتي الأولى ،ممسكتي في سقطات خطواتي الأولى، إلى ملجئي الأول دائماً، المرأة الحديدية، ملاكي في الحياة، التي علمتني أسمى معاني الحب الغير مشروط، إلى من صوتها يدفء قلبي، إلى أماني حناني و صديقتي، إلى الحاضرة دائماً فيَ كلَي، إلى من تعطي و تستمر في العطاء...**أمي**

إلى من يعتدل بهن ميل العالم، إلى من كانتا خير جليس وونيس و أنيس، إلى أحلى معاني المحبة بلا شر ولا مصالح، إلى قدوتي الناجحتي اللتين أفتخر بهما، كنتن خير قدوة لي ، إلى مساندتي بكل الأوقات، في ضعفي و في قوتي، توأمي الغالي، حبيبتيَ، أخوتيَ الكبار...**تالة و تسنيم**

إلى سندي وجبلي الحصين، حبيبي الصغير ،مستقبل عائلتنا و فردنا الصغير، صديقي و عزَي،
وليدي، أخي ...**وليد**

إلى صديقة طفولتي، توأم روحي، رفيقة أوقاتي، ملجئي في سعادتي و شقائي، سر ابتسامتي في حزني و انطفائي، مهما أقول و أتكلم فلن تكفي كلماتي، إلى أختي و رفيقتي و ابنة عمتي ...**روان**

إلى أصدقاء العمر، أحباب الروح، رفاق الدرب، أخوتي التي رزقتني بهم الحياة، لكم أنا فخورة بتعرفي عليكم، و لحظي لأقضي معكم أحلى سنوات شبابي، زملاء الشدائد و الضغوط كنتم لخير السند لي، قد يخونني التعبير و لكن لكل منكم مكان تعرفونه..

أماني، نجمتي، رفيقة العمر، صديقة الشدائد و اللحظات السعيدة،رفيقة روحي..**نسرين عودة**

إلى أصدقائي..إخوتي...

محمد يمان الخن – محمد بدر المبيض – أمينة أبو رشيد – نويل الباشا – محمد نزار تركماني – أنس النفاخ – راما دريج – بسمة أبو الذهب – كريم دقماق – عمر عرفة

إلى زملاء العمل و أصدقاء الأيام الصعبة، و الداعمين الأكبر في هذا البحث، إلى من أصبحوا في أفصر مدة من أقرب الأصدقاء و الأخوة لي

محمد براء قتابي – محمد شيخ أوغلي – عبد الرحمن دياربكرلي – محمد ناظم سبيعي – وسيم القاضي

إلى عائلتي الحبيبة، جدتي أطال الله في عمرها، جدي أطال الله في عمره

إلى روح جدتي رحم الله روحها و أسكنها في جناته

ملخص البحث

يعد تحليل وتوقع سلوك الزبون فيما إذا كان سيتسرب إلى المنافسين أم لا أحد أهم العمليات لمؤسسات الأعمال والشركات في مختلف المجالات، وخصيصاً شركات الاتصالات، نظراً لكون الزبون أهم الموارد بالنسبة لها، وكونه ذو أثر كبير على ربح هذه الشركات، ولهذا اتجهت الشركات لاستباق خطوات عن المنافسين وحل بعض من مشاكلها، والتي أهمها التنبؤ بتسرب الزبائن، ومعرفة الأسباب وراء التسرب وما أكثر العوامل المؤثرة في زيادته، واتخاذ الخطوات الاستباقية اللازمة للحد من هذه المشكلة.

يهدف ويسعى هذا البحث لبناء نموذج تصنيفي قادر على التنبؤ بتسرب الزبائن من شركة اتصالات، أولاً باتباع طرق عديدة لتحضير البيانات واختيار المتغيرات، ومن ثم الاستفادة من هذه البيانات في بناء المصنف باستخدام تقنيات التنقيب في المعطيات.

جرى هذا النظام على قاعدة بيانات زبائن تابعين لشركة اتصالات X ، وبعد تطبيق 5 خوارزميات (Logistic regression, SVM, KNN, J48, Random Forest)، و المقارنة بينهما حسب المعايير (Accuracy، Recall، Kappa statistic، Precision)، تم اختيار مصنف الغابة العشوائية Forest (Random)الذي أعطى أعلى دقة، حيث وصلت نسبة الدقة إلى 85%.

ومن خلال تحليل البيانات والنتائج والمعلومات المستخرجة، والتي طبقت عبر تصوير البيانات تمكن الباحث من معرفة بعض العوامل المؤثرة والأسباب وراء تسرب الزبائن من شركة اتصالات.

الكلمات المفتاحية: التنقيب في المعطيات، التصنيف، تصوير البيانات، تسرب الزبائن، اختبار الارتباط، الدقة، الغابة العشوائية.

Abstract

Analyzing and predicting the customer's behaviors, whether they will churn to competitors or not, is one of the most important operations for businesses and companies in various fields, especially telecommunications companies, given that the customer is the most important resource for them, and has a significant impact on the profit of these companies, and for these companies tended to anticipate steps from competitors and solve some of their problems, the most important of which is predicting churn, knowing the reasons behind the customer churn and what are the most influential factors in increasing it, and taking the necessary proactive steps to reduce this problem.

This research aims and seeks to build a classification model capable of predicting customer churn from a telecom company, first by following several methods for preparing data and selecting variables, and then benefiting from this data in building the classifier using data mining techniques.

This system was carried out on a customer database of X Telecom, and after applying 5 algorithms (Logistic regression, SVM, KNN, J48, Random Forest), and comparing them according to the criteria (Accuracy, Recall, Kappa statistic, Precision), the forest classifier was chosen. Randomization gave the highest accuracy, as the accuracy rate reached 85%.

And by analyzing the data, results, and information extracted which were applied through data imaging, the researcher was able to know some of the influencing factors and the reasons behind customer's churn out of a telecom company.

Key Words: Data Mining, Classifications, Data Visualize, Churn, correlation, Accuracy, Random Forest

الفهرس

جدول الأشكال والرسوم التوضيحية

الأشكال:

الرسوم التوضيحية:

جدول المصطلحات

Data Mining	التنقيب في المعطيات
Machine Learning	التعلم الآلي
Classifications	التصنيف
Cluster	العنقدة
Data Visualize	تصوير البيانات
Correlation	اختبار الارتباط
Attributes	واصفات
Churn	تسرب الزبائن
supervised learning	تعليم خاضع للإشراف
unsupervised learning	تعليم غير خاضع للإشراف
Accuracy	الدقة
Recall	الإستدعاء
Precision	الضبط
Kappa	معامل كابا
True Positive	إيجابية حقيقية
True negative	سلبية حقيقية
False Positive	إيجابية خاطئة
False negative	سلبية خاطئة
Oversampling	الإفراط في أخذ العينات
Imbalanced Data	بيانات غير متوازنة
Normalization	تقييس
Generalization	تعميم
Entropy information Gain	الربح في المعلومات
Random Forest	الغابة العشوائية
Logistic Regression	الانحدار اللوجيستي
Decision tree	شجرة قرار
k-nearest neighbors	خوارزمية الجوار الاقرب
Support vector machine	آلة المتجهات الداعمة

المقدمة

منذ بدايات القرن العشرين وحتى يومنا هذا، لاحظنا مزامنة التقدم العلمي الحديث للنمو الكبير في التقنيات العلمية بشكل عام وتلك الخاصة بالبيانات على وجه الخصوص.

الأمر الذي منح تحسينات نوعية في الأدوات الموجودة، بغية منحها القدرة على الإجابة عن الأسئلة المستجدة ذات الأطياف المتعددة، إلا أن الطيف الواسع لتعقيدها الناتج عن تعقيد وتطوير منظومة عالمنا، وبالأخص تعقيد فهمنا لهذا العالم جعل من تطويرها عمل غير كافي.

ليدفع العلماء (الرياضيين والمعلوماتيين منهم على وجه الخصوص) إلى استحداث عدد كبير متزايد من الأدوات والتقنيات الجديد المصقولة، التي تزيد قدرتنا على توسيع أفق تطبيقنا التي بلغت حدود غير مسبوقة، ما كنا نعتبره خيالا في يوم من الأيام.

نتيجة لتلك الثورة الرقمية فقد ظهرت تقنيات التنقيب عن البيانات في السياق كإضافة حديثة للقائمة الطويلة من المناهج والممارسات والأدوات العلمية، التي زودتنا بمنظور جديدة للكثير من مسائلنا التقليدية وسأهمت في زيادة فهمنا لمنظومات عالمنا المعاصر ومسائل هذه المنظومات بغية تمكننا من التحكم بها.

حيث تعتمد الشركات (تلك المعنية بالاتصالات على وجه الخصوص) على تقنيات التنقيب في البيانات للقيام بعملياتها وإدارتها، والتفاعل مع زبائنها ومورديها، والتنافس في السوق وإثراء جوانب البحوث العلمية.

فنجاح المنظمة يتوقف على كفاءة وفاعلية إدارتها في التنبؤ وصنع القرارات، منطلقين من مفهوم جوهري آخر كون أن البيانات هي الحجر الأساس الذي ترتكز عليه تلك القرارات في مختلف المستويات الإدارية ومختلف أقسام المنظمة.

في سياق حديثنا عن البيانات وتقنيات التنقيب نستعرض التعريف النظري لها رغم وجود العديد من وجهات النظر حول التعريف إلا أن الجميع يتفقون على أن التنقيب عن البيانات هي مجموعة فرعية متصلة بنظم المعلومات كحالها بالاتصال بذكاء الأعمال وعلم البيانات وتحليلات البيانات.

حيث يعد التنقيب عن البيانات عملية منهجية ومتسلسلة لتحديد واكتشاف الأنماط والمعلومات المخفية في مجموعة بيانات كبيرة، يُعرف أيضًا باسم اكتشاف المعرفة في قواعد البيانات.

ينطلق هذا البحث في البداية من دراسة قطاع صناعة الاتصالات والتسرب الوظيفي والقيمة التي سنكسبها له من خلال استخدام تقنيات التنقيب في البيانات.

حيث تعد صناعة الاتصالات المحرك الأساسي المهيمن في قطاع التكنولوجيا، البيئة الأكثر تنافساً في القرن الحادي والعشرين، بقيمة بلغت 1.4 تريليون دولار في عام 2019.

مستعينة تلك الأخيرة بتوجهات جديدة تتبنى من خلالها استراتيجيات فعالة ,للبقاء أطول فترة ممكن في سوق المنافسة، وزيادة الأرباح عن طريق اكتساب عملاء جدد والاحتفاظ بالعملاء الحاليين.

صنفت Harvard Business School استراتيجية الاحتفاظ بالعملاء الحاليين بأنها الاستراتيجية الأكثر نجاحاً في تحقيق ربح لشركات، حيث فإن زيادة معدلات الاحتفاظ بالعملاء بنسبة 5% ستؤدي إلى زيادة الأرباح بنسبة 25% إلى 95% كما و أن 65% من أعمال الشركة تأتي من العملاء الحاليين.

يتجلى نجاح هذه الاستراتيجية باستخدام تقنيات التنقيب في البيانات بالاعتماد على مجموعة من الخوارزميات الرياضية الشهيرة والعمليات الاحصائية والتحليلية، عبر جمع وتنظيف وتحليل بيانات العملاء والاحتفاظ بها وذلك للتوقع المسبق للزبائن المحتمل تسربهم والوقت المحتمل لهذا التسرب، وتكريس كافة الجهود للحفاظ على العملاء عبر اكتشاف المعرفة والتنبؤ بالمستقبل.

وعليه، إن التوجه السائد في العالم حالياً وخاصة في قطاع الاتصالات يتجه نحو استثمار تقنيات التنقيب في البيانات التي تسأهم في تحقيق أهدافها ,وهذا ما يقدمه البحث الحالي من خلال استخدام تقنيات التنقيب في البيانات للتنبؤ بتسرب الزبائن في شركة اتصالات.

الفصل الأول

الإطار التمهيدي

مشكلة البحث:

عادة ما تركز الشركات بشكل أكبر على اكتساب العملاء الجدد، بينما الاحتفاظ بالعملاء الحاليين كان يأتي دائماً كأولوية ثانوية، ومع ذلك يمكن أن يكلف جذب عميل جديد خمسة أضعاف تكلفة الاحتفاظ بعميل حالي .

لتعظيم عدد العملاء اتجهت الشركة للاحتفاظ بالعميل القديم بدلاً من جذب عملاء جدد للشركة، كما و في الاتجاه نحو العميل القديم، فسيكون لدينا البيانات اللازمة حول تفاعل العميل مع الخدمات، و بالتالي يمكننا التنبؤ بالعميل الذي سيغادر، و معرفة الرد المناسب لإبقاءه، بتقديم بعض العروض أو الباقات التي تهمه.

لهذا يسعى هذا البحث إلى بناء نموذج للتنبؤ بمشكلة تسرب الزبائن، حيث تعتبر هذه المشكلة من أخطر المشاكل التي تسعر الشركات لايجاد الحلول لها .

كما و تتطلب هذه المشكلة من الشركة جمع بيانات العملاء، و البحث فيها، لتحديد الأسباب المودية للتسرب، وتحليل البيانات بحيث يصبح لدينا القدرة على تحديد السمات المشتركة للزبائن المتسربين وسيتم ذلك من خلال الإجابة على التساؤلات الاتية:

1. هل يمكن الاستفادة من تقنيات التنقيب في المعطيات لبناء مصنف تنبؤي قادر على التنبؤ بالزبائن الذين سيجددوا عقودهم من الذين لن يجددوه؟
2. ماهي أبرز الأسباب التي تؤثر على الزبائن و تؤدي بهم لترك شركة الاتصالات و عدم تجديد عقودهم؟

دوافع اختيار البحث:

- الأهمية التي اكتسبتها المعلومات حيث اصبحت سلعة كغيرها من السلع لها قيمتها السوقية وتعتبر مورد إستراتيجي هام.
- لمساعدة شركات الاتصالات للحد من مشكلة تسرب الزبائن .
- لمساعدة قسم الأبحاث التسويقية لمعرفة الحزم التسويقية التي يمكن أن تقدم للزبائن لمحاولة الاحتفاظ بهم و الابقاء عليهم في الشركة .
- إثراء الأبحاث والدراسات العربية لقلة توجهها في هذا مجال.

أهداف البحث:

يهدف البحث إلى استخدام التنقيب في المعطيات لضبط تسرب الزبائن من خلال بناء نظام قادراً على التنبؤ بالزبائن المحتمل تسربهم من الشركة، ومحاولة اكتشاف السمات المشتركة بين الزبائن المتسربين للحد من الأسباب الداخلية في شركات الاتصالات التي تدعو الزبائن إلى ترك خدمة الشركة، مما يكبد هذه الشركات خسائر كبيرة، نظراً لأهمية العميل لهذا النوع من الشركات الخدمية.

وسيتم ذلك من خلال الإجابة على التساؤلات الموجودة في مشكلة البحث، ويمكن تلخيص الأهداف وفق التالي:

- تحديد أغلب حالات التسرب بدقة، و تقليل عدد حالات التوقع الخاطئ بعدم التسرب.
- التخفيض من التكلفة التي يمكن أن تنتج في حال جذب العملاء الجدد، و الاستعاضة عنها للابقاء على العملاء الحاليين.
- الاستفادة من تقنيات التنقيب في المعطيات لبناء مصنف تنبؤي قادر على التنبؤ بتسرب الزبائن ومعرفة فيما إذا كان العملاء سيجددون عقودهم أم لا.
- المقارنة بين الخوارزميات التصنيفية، و معرفة الخوارزمية الأفضل القادرة على إعطاء أفضل نتيجة.

- معرفة الأسباب التي تؤثر على الزبائن، و تؤدي لتركهم شركة الاتصالات.

أهمية الدراسة:

تكمن أهمية هذه الدارسة في كونها تقوم على دارسة أهم المواضيع في الموارد البشرية التسرب، الوظيفي داخل المنظمات، حيث أن التسرب الوظيفي يكلف المنظمات خسائر جسيمة نتيجة فقد الخبرات والكفاءات البشرية بالإضافة لإعادة تكوين وتدريب الموظفين الجدد وهذا بالتأكيد سيضعف نمو الإنتاجية.

يمكن توضيح أهمية البحث من خلال شقين أساسيين، الأهمية النظرية من البحث والأهمية التطبيقية منه وفق الآتي:

الأهمية النظرية

تكمن الأهمية النظرية للبحث من خلال توضيح المراحل التي مرت بها عملية بناء نموذج ضبط تسرب الزبائن المدروس بشفافية، فسوف يتم توضيح العديد من التعاريف والمصطلحات المتعلقة بهذا البحث، فيمكن أن يكون هذا البحث مرجعا للدراسات القادمة في هذا المجال لما يثيره من قضايا وتساؤلات يمكن أن تؤخذ بعين الاعتبار، حيث سيتم تقديم توصيات تسأهم في تحسين إدارة العلاقة مع العملاء في المنظمات بشكل عام، وفي قطاع الاتصالات بشكل خاص، حيث أن الباحث يحاول فهم الأسباب التي تدعو الزبائن إلى ترك شركة الاتصالات، و التخلي عن الخدمة، والتنبؤ بتسرب الزبائن حيث ان وجود هكذا نظام يساعد المؤسسات و الشركات على إدارة علاقاتها مع عملاءها، و الحفاظ على عملاءها الحاليين، التي تعد من أهم موارد الشركات الخدمية، كما و يساعد مسوقي الشركة لمعرفة الحزم التسويقية التي كم الممكن تقديمها للعملاء للإبقاء عليهم و الحفاظ عليهم، كما ان فهم هكذا ظاهرة والحد منها يوفر على الشركات تكاليف ضخمة كانت ممكن أن تتكبدها شركات الاتصالات في جذب عملاء جدد لها .

الأهمية التطبيقية

تكمن الأهمية التطبيقية للبحث من خلال شقين أساسين:

- التنبؤ: يسعى البحث إلى بناء عدة مصنفات قادرة ع التنبؤ بتسرب الزبائن واختيار الخوارزمية الأفضل القادرة على إعطاء أفضل نتائج على هذه الظاهرة، حيث ان ذلك سيساعد الشركات على فهم الأسباب الرئيسية التي تدعو الزبائن إلى ترك شركة الاتصالات.
- تحليل البيانات: يسعى البحث إلى تحليل البيانات الحالية للشركة للتوصل إلى معلومات وأسباب حول تسرب الزبائن لدى الشركة، واستعراضها وتقديم شرح عنه.

حدود البحث:

مكانية: تم تنفيذ الدراسة على بيانات مأخوذة تابعة لشركة اتصالات.

زمانية: تم اعداد البحث خلال المدة الممتدة من 2020/5/1 إلى 2021/6/30.

معوقات البحث:

- إن مشكلة التسرب يعتبر مسألة تصنيف ثنائية تعاني من مشكلة عدم توازن البيانات Imbalanced Data، حيث أن البيانات التي تعبر عن صنف الزبائن المتسربين يكون قليلة جداً مقارنة بالزبائن الغير متسربين.
- عدم توافر بيانات محلية لمثل هكذا حالات.
- عدم تعاون شركات محلية لإجراء مثل هكذا أبحاث لديها .

منهجية البحث:

اعتمد هذا البحث على المنهج التنبؤي وهو الذي يصطلع بدراسة "المستقبل" وفق برامج وآليات التنبؤ، من خلال دراسة التغيرات في مسلك الماضي وخط سيره، وإسقاطاتها على المقابل المستقبلي لها.

وتم ذلك من خلال دارسة وتحليل بعض سمات و اتجاهات عملاء، واستخدام أدوات التنقيب في المعطيات (التصنيف، تصوير البيانات، العنقدة) ومن ثم اقتراح النموذج الملائم بناءً على المعطيات الموجودة.

هيكلية البحث:

يتألف البحث من خمس فصول عرضنا في هذا الفصل مقدمة عن مشكلة تسرب الزبائن التي يعاني منها قطاع الاتصالات وأهمية معاجلتها والحد منها، من اجل تحقيق الربح لهذه الشركات، وتخفيض التكلفة التي قد تنتج في حال لم تحافظ الشركات على زبائنها الحاليين واتجهت لجذب زبائن جدد، كما تطرقنا إلى المشكلة الأساسية التي يعالجها هذا البحث، وتطبيق هكذا نموذج من أجل الوصول لنظام لتوقع تسرب الزبائن.

- **الفصل الثاني:** تحدثنا في هذا الفصل عن الدراسات السابقة التي تناولت مواضيع مشابهة عن دراستنا الحالية ووضحنا ملخص كل دراسة والنتائج التي توصل لها كل بحث كم تم التطرق إلى اوجه الاستفادة التي ساعدتنا في بحثنا الحالي.
- **الفصل الثالث:** شرحنا في هذا البحث عن الجوانب النظرية التي يتناولها البحث، و بدأنا بالشرح عن المفهوم العام للتنقيب في المعطيات وخطواته وانواعه، كما تم القاء الضوء على أهم التقنيات المستخدمة في هذا العلم، والتعريف بخوارزميات التصنيف المتبعة في البحث الحالي.

  كما و تم التطرق لنظرة عامة عن قطاع الاتصالات، كما تم تقديم بعض المفاهيم عن إدارة خدمة العملاء CRM و أهميتها، و أخيراً تم التطرق لشرح بعض المفاهيم التي تتعلق بالمشكلة الرئيسية و التي هي تسرب الزبائن .

- **الفصل الرابع:** يتكون هذا الفصل من 5 مباحث، المبحث الأول قدمنا شرحاً مفصلاً عن الحالة العملية في شركة اتصالات، و شرحاً عن البيانات التي استخدمت في هذا الدراسة وعرضنا سماتها ومعانيها، وعرضنا فيه الأدوات المستخدمة في عملية التنقيب في المعطيات التي تمت في هذا البحث.

  المبحث الثاني يحوي عملية التنقيب التي قمنا بها بهذه الدراسة من خلال عملية تصوير البيانات التي حاولنا فيها تسليط الضوء على أهم النقاط الواردة في البيانات، لمعرفة العوامل و الأسباب المؤثرة في عملية التسرب و تجديد العقود من عدم تجديدها.

  المبحث الثالث قمنا بعملية تحضير البيانات لجعل نتائج التصنيف غير مغلوطة.

  المبحث الرابع قمنا بعملية التصنيف وعرضنا نماذج و خوارزميات التصنيف المستخدمة في هذه الدراسة . المبحث الخامس قارنا بين الخوارزميات المستخدمة من اجل الوصول إلى أفضل مصنف يتنبأ لنا بمشكلة البحث.

- **الفصل الخامس:** تم تقديم النتائج والتوصيات التي توصلت اليها الدراسة والمراجع التي استخدمت في الدراسة.

الفصل الثاني

الدراسات السابقة

تمهيد:

تمثل الدراسات السابقة مصدراً مهماً من مصادر المعرفة، ومعيناً خصباً من المعلومات التي يحتاج إليها الباحثون والدارسون، إضافة لرفدها المجال البحثي بخلاصة نتائج وتوصيات أصحابها، وبذلك تستثمر الدراسات السابقة كأداة مهمة للتطور والتقدم.

كما أنها تساعد على توقير كم من المعلومات النظرية و النماذج الواقعية، حتى يمكننا الاستفادة منها في جميع مراحل البحث، و على أساسها يمكننا و نحاول البدء من حيث توقف الآخرون، للانطلاق في بحثنا ليكون تكملة للبحوث السابقة.

الدراسات السابقة:

الدراسة الأولى: عبد الرحيم قاسم احمد عام (2018).

"توقع تسرب الزبائن من شركات الاتصالات باستخدام تعلم الالة في بيئة البيانات الكبيرة"

تعتبر مشكلة تسرب الزبائن إلى المنافسين من المواضيع الهامة بالنسبة للشركات الكبرى، وخصوصاً شركات الاتصالات وذلك بسبب الأثر الكبير الذي ينعكس إلى ربح الشركات، لذلك تسعى الشركات لتطوير طرق للتنبؤ بتسرب الزبائن، ومعرفة العوامل التي تؤثر على زيادة التسرب، واتخاذ الإجراءات اللازمة للحد من هذه الظاهرة. نعرض في هذا البحث نظاماً يعتمد على خوارزميات تعلم الآلة للتنبؤ بالزبائن المحتمل تسربهم من شركات الاتصالات، حيث قمنا بتطوير طريقة جديدة لتحضير البيانات واختيار السمات.
جرى اختبار النظام في بيئة (Spark) المناسبة لمعالجة البيانات الكبيرة، على قاعدة بيانات عالمية من شركة (Orange) حيث تشير النتائج التي حصلنا عليها في هذا العمل إلى دقة عالية مقارنة بالأعمال الأخرى المنشورة باستخدام نفس قاعدة البيانات، حيث تشير النتائج التي حصلنا عليها في هذا العمل إلى دقة عالية مقارنة بالأعمال السابقة المنشورة باستخدام نفس قاعدة البيانات، حيث وصلت النسبة إلى 87.2%، وقمنا ايضاً بتجهيز قاعدة بيانات خاصة بشركة سيرتيل وتحتوي على كل المعلومات الخاصة بكل الزبائن لمدة 6 أشهر، بهدف تجريب واختبار النظام السابق، وقد أعطت نتائج جيدة بدقة وصلت ل 95%.
قمنا بتجريب عدة خوارزميات شجرية وهي (شجرة القرار والغابات العشوائية وخوارزمية شجرة التعزيز التدريجي وخوارزمية XGBoost)

نتائج الدراسة :

1- تكمن أهمية هذا النوع من الأبحاث في سوق الاتصالات في مساعدة الشركات على تحقيق مزيد من الربح، كما انه أصبح من المعروف أن التنبؤ بالتسرب الوظيفي يعد من أحد أهم المصادر التي تحقق عائداً للشركات من خلال الحفاظ على الزبائن.

2- إن عدم القدرة على منع الزبائن من إلغاء اشتراكاتهم مع الشركة يوجب على الشركة إعادة التفاوض على هذا العقد من أجل الحفاظ على الزبائن.

3- كانت البيانات المستخدمة للوصول إلى النموذج الأمثل على شقين القسم الأول من شركة (Orange) وتم تجريب هذا النموذج على شركة سيرتيل وتم وضع القاعدة التالية: إذا لم يقم أحد الزبائن بإجراء او استقبال أي مكالمات خلال فترة زمنية تبلغ 30 يوماً فقد تم اعتباره متسرباً.

4- تم اختيار 4 خوارزميات للتجريب على النموذج، حقق نموذج شجرة (XGBoost) أفضل النتائج حيث حقق Roc بنسبة 95.72% ومن ثم أتت خوارزمية شجرة التعزيز التدريجي ثم الغابات العشوائية واخيراً شجرة القرار.

❖ يعد هذا البحث من أقرب البحوث إلى بحثنا الحالي، حيث تم الاستفادة من المنهج البحثي المعتمد من قبل الباحث، الذي اعتمد على خوارزميات التصنيف للتوصل إلى نتيجة في التنبؤ بتسرب الزبائن في شركات الاتصالات .

تم الاستفادة من هذه الدراسة من خلال الاطلاع على مفاهيم خوارزميات تعلم الالة التي تم استخدامها في بحثنا، وكيفية معالجة قضية التسرب في شركة اتصالات باستخدام هذه الخوارزميات وكيفية المقارنة بينهم لاستخدام الخوارزمية الأفضل .

كما و تم الاستفادة منها من خلال المفاهيم التي قدمتها حول التنبؤ، تسرب الزبائن و غيرها.

و من خلال بحثنا العلمي الحالي، سنحاول استخدام خوارزميات تعلم الآلة على بيانات شركة أخرى، و محاولة الوصول إلى نموذج تنبؤي يساعدها في التنبؤ بمشكلة تسرب الزبائن.

الدراسة الثانية: Archana Singh & Ajay Rana

"Mining of Customer data in an Automobile Industry using Clustering Techniques"

إدارة علاقات العملاء (CRM) هي نهج رائد يتبعه المسوقون في عملية الاحتفاظ بعملائهم. لا يزال نهج CRM جديدًا بالنسبة لمديري الشركات، حيث لا يزال العملاء في الهند عند الطرف المتلقي فقط.

يمكن تطبيق تقنية التنقيب عن البيانات بقوة في الأبحاث متعددة التخصصات التي تبرز الأنماط والمعرفة من الحجم الضخم للبيانات والتي بدورها تساعد المنظمات في اتخاذ القرارات بكفاءة. تتعامل الصناعات القائمة على المنتجات والخدمات بشكل رئيسي مع العملاء بشكل مباشر أو غير مباشر. في هذه الدراسة، يركز على دراسة نوع العملاء المجمعين بناءً على أنماطهم في قطاع السيارات.

في ورقة البحث هذه، تُظهر الدراسة أن تقنيات التنقيب عن البيانات في بناء حلول تتمحور حول العملاء في المؤسسة الهندية مثل قطاع السيارات تكتسب شعبية ونتائج البيانات، ويمكن استخدام التحليل لمعرفة عميلك وتفضيلاته.

مع وجود مجموعة كبيرة ومتنوعة من المنتجات المتاحة في السوق، يُترك للعملاء العديد من الخيارات للاختيار من بينها بين العلامات التجارية وضمن مجموعة متنوعة من المنتجات من مختلف القطاعات. ن تحدد الدراسة المقترحة نوع العميل وستكون قادرة على فحص العلاقة بين إدارة علاقات العملاء وتقنيات التنقيب في البيانات. أيضًا، مع تقنيات استخراج بيانات التأثير في CRM، يُظهر أن التنقيب عن الأنماط، واستخراج الميزات من سيساعد الإدارة على بناء استراتيجية للاحتفاظ، وجذب عملاء نشطين محتملين جدد.

كان الهدف من النموذج التحليلي المتمحور حول العميل هو تحديد العملاء النشطين والمخلصين الذين يشترون المنتج مباشرةً وكذلك يسأهمون في المؤسسة من خلال الإشارة إلى العملاء الآخرين، عن طريق نشر ردودهم الإيجابية / السلبية حول المنتج والخدمة على مختلف المواقع الاجتماعية. يتم حساب التكلفة التراكمية للعميل من النموذج المقترح (المنتج، الخدمة، التوصية، التكاليف المخفية الأخرى) ويمكن أيضًا استخدام قيمة الاحتفاظ بالعملاء بهذا المعدل لمزيد من عملية اتخاذ القرار فيما يتعلق بالعملاء.

بالنسبة لتجزئة العملاء، فإن اعتماد المعلمات (العمر، اسم المتغير، التوقعات، التبادل، حالة الاستعلام، القيمة مقابل المال، مصدر الاستفسار، المهنة) يصنف العملاء إلى عملاء راضون جدًا، فوق المتوسط ، متوسط ، وسلبي. يمكن أن يؤدي هذا التقسيم إلى معدل تحويل أفضل من العملاء السلبيين إلى العملاء النشطين.

حيث يدلني مصدر الاستفسار إذا كان ماروتي في أي وقت وحالة الاستفسار نشطة، فهو العميل الأكثر ولاءً والذي يتم الحكم عليه كمقياس لعدد المنشورات والتعليقات على مواقع التواصل الاجتماعي حول المنتج والخدمة.

وأهم ما توصلت الية الدراسة:

- لوحظ ومن الحقائق المعروفة أن البحث وتحليل البيانات الذي يتم باستخدام تقنيات التنقيب عن البيانات يساعد المنظمة على اتخاذ قرارات أفضل وتخطيط الإستراتيجية لنمو أي صناعة موجهة نحو المنتجات والخدمات، من المهم الاستمرار في جلب سياسات جديدة لجذب عملاء جدد والاحتفاظ بالعملاء الحاليين هو التحدي الكبير الذي لا ينتهي أبدًا.
- في هذه الورقة البحثية، تم اقتراح نموذج تحليلي جديد يركز على العميل باستخدام التنقيب عن البيانات، وحساب قيمة الاحتفاظ النشطة بالعملاء والمعلمات التي تؤثر على هذه القيمة.
- تمت مناقشة دراسة تجريبية حول صناعة السيارات وتحليل البيانات وتفسير النتائج وبالتالي توفير حل الأعمال للمؤسسة لتحسين العملاء المحتملين.

❖ تم الاستفادة من هذه الدراسة من خلال ما قدمته من مفاهيم عن إدارة علاقة العملاء (CRM) و هو النهج الذي يتبعه المسوقين للحفاظ على عملائهم، أنه الأمر الذي يستفاد منه في توصيات بحثنا، للحد من تسرب الزبائن.

الدراسة الثالثة: Sajedun Nahar (2018).

Segmentation of Mobile Customers using Data Mining Techniques

في عالم اليوم التنافسي، للحفاظ على العملاء الرضا هو مفتاح النجاح لشركات الاتصالات .
تعد تقنيات التنقيب عن البيانات أكثر تفضيلاً لاكتشاف سمات العميل بالإضافة إلى احتياجاته التي يمكن تحقيقها من خلال تقسيم سلوكياتهم.
و التجزئة هي عملية تطوير ذات معنى لمجموعات العملاء المتشابهين ولهم نفس الخصائص بالحسابات الفردية والسلوكيات باستخدام كمينز التجميع .
تقترح الدراسة حلاً بتجزئة عملاء شركة الاتصالات.
وكان الهدف الاساسي هي تجميع العملاء باستخدام خاصية سلوكهم وتقديم الخدمات حسب المجموعة.
في الآونة الأخيرة، أصبح سوق الاتصالات المتنقلة متنافس بشكل كبير حيث زيادة عدد العملاء هو الأساس التحدي في صناعة الاتصالات الحديثة.

وفي هذه الدراسة، أظهرنا ذلك من خلال استخدام تجزئة العميل و يمكن لشركة الاتصالات أن تجتذب بسهولة عملائها بالمنتجات والخدمات المناسبة , هذا ايضا يساعد في تقديم الباقات والعروض والحزم للعملاء.
لذلك، يجب على الشركات أن تدرك عظمة تجزئة العميل وتنميط سلوك العميل لتحقيق نتائج أفضل من خلال تضييق شرائح العملاء وان تحليل الكتلة قادر على حل مشكلة تجزئة العملاء.

تعتمد هذه الدراسة على مجموعة K-mean clustering طريقة لحل تحليل تجزئة عملاء الاتصالات.
النتائج العملية التي توصلت اليها الدراسة:

- تشير إلى أن تحليل تقسيم العملاء لقطاع الاتصالات فعال وناجح وكان الهدف التجاري هو تجميع العملاء من حيث الخصائص السلوكية , وبعبارة أخرى، نحن ناقشنا بشكل نظري حول استخدام خوارزميات التنقيب عن البيانات لتقديم العروض المناسبة للعملاء.

- ❖ تم الاستفادة من هذه الدراسة من خلال الإطلاع على مفاهيم الحفاظ على العملاء من خلال تقسيمهم إلى شرائح، لتحدد لاحقاً لكل شريحة حزمة تسويقية من المنتجات و الخدمات، و باقات و عروض تساعد في اجتذاب العملاء الجدد و الحفاظ على العملاء الحاليين .

الدراسة الرابعة: Abba Suganda Girsang & Andri Wijaya (2018).

USE OF DATA MINING FOR PREDICTION OF CUSTOMER LOYALTY.

ستناقش هذه الأطروحة حول توقع ولاء العملاء في شركة PT. XYZ، باستخدام ثلاث خوارزميات في طريقة التنقيب عن البيانات. سيناقش هذا التحليل أيضًا 4 سيناريوهات للمقارنة في تحديد أجزاء مجموعة البيانات وتحليل السمات التي تؤثر ولا تؤثر على دقة نتائج النموذج أو الخوارزمية المستخدمة. من خلال المقارنة والتحليل، حصلت خوارزمية C4.5 مع جزء مجموعة البيانات الخاص بها (80٪ لبيانات التدريب و 20٪ لبيانات الاختبار) على أعلى نتائج دقة بلغت 81.02٪، مقارنة بالخوارزميات الأخرى وأجزاء مجموعة البيانات. في تحليل السمات، حصلت على سمة سبب الانفصال (السمة التي يتم تفسيرها على أنها سبب طلب العملاء للتوقف) باعتبارها السمة الأكثر تأثيرًا على دقة النتائج في عملية استخراج البيانات للتنبؤ بولاء العميل.

أهم ما توصلت اليه الدراسة:

- يمكن أن يؤثر تحديد مجموعة البيانات التي استندت إلى جزء من تدريب البيانات واختبار البيانات على دقة نتائج الخوارزمية المستخدمة, كما هو موضح في الفصل الثالث، في اختبار مجموعة البيانات باستخدام أجزاء مختلفة، ينتج عن ذلك دقة مختلفة .
- هناك أكثر السمات تأثيرًا على نتائج دقة النماذج أو الخوارزميات المستخدمة كما هو موضح في الفصل الثالث، فإن سمة سبب الانفصال هي السمة الأكثر تأثيراً في نتائج الدقة وسمة القرص على الإطلاق هي السمة المقلقة في جميع النماذج أو الخوارزميات المستخدمة في عمليات التنقيب عن البيانات التي يتم إجراؤها للتنبؤ بولاء العملاء.

- من تلك الخوارزميات الثلاثة (C4.5، Naïve bayes & Nearest Neighbour) التي تُستخدم لإجراء اختبار في أجزاء مجموعة البيانات المختلفة، خوارزمية C4.5 مع جزء مجموعة البيانات الخاص بها (80٪ لبيانات التدريب و 20٪ لـ نتائج الاختبار) أعلى نتائج دقة بنسبة 81.02٪. يمكن رؤيته في المقارنة في الجدول 16.

❖ تم الاستفادة من هذه الدراسة من استخدام أدوات التنقيب في المعطيات للتنبؤ، و تجربة الباحث على برنامج weka و بتجريب 3 خوارزميات تنبؤية أيضا كما في بحثنا، و المقارنة في نتائجها لاختبار المصنف الأفضل .

الفصل الثالث

الإطار النظري

المبحث الأول

التنقيب في المعطيات

1. التنقيب في المعطيات

إن التنقيب في البيانات يهدف إلى استخلاص المعلومات المخبأة في كتل البيانات الكبيرة، وهي تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية، واستخدامها يوفر للشركات والمؤسسات في جميع المجالات التجارية، الحكومية، القدرة على الاستكشاف والتركيز على أهم المعلومات في كتل البيانات الكبيرة، كما تركز تقنيات التنقيب على الاستشعار وبناء التنبؤات المستقبلية واستكشاف الأنماط والارتباطات والسلوك والاتجاهات مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب ووضع الحلول المناسبة للمشكلات والتخطيط والتطوير والتحديث في جميع المجالات.

وتجيب تقنيات التنقيب على العديد من الأسئلة في وقت قياسي بخاصة تلك النوعية من الأسئلة التي كان من الصعب الإجابة عليها إن لم يكن مستحيلاً باستخدام تقنيات التحليل الاحصائي الكلاسيكية، والتي كانت حتى وان وجدت فإنها تستغرق وقتاً طويلاً والعديد من إجراءات التحليل.

سيتم التطرق في هذا البحث على تعريف علم البيانات، خطواته، أنواعه، وأهم تقنياته وأهم الخوارزميات المستخدمة في تلك التقنيات.

2. مفهوم التنقيب في البيانات

حظي استخراج المعارف المخبأة في البيانات قدراً كبيراً من الاهتمام لدى العاملين في مجال تكنولوجيا المعلومات والاتصالات في السنوات الأخيرة نظراً لتوفر كميات ضخمة من البيانات والحاجة الملحة لتحويل هذه البيانات إلى معلومات ومعارف مفيدة وذات قيمة.

ويمكن استخدام هذه المعلومات والمعارف المكتشفة في طيف واسع من التطبيقات مثل تحليل السوق، كشف محاولات الغش، المحافظة على الزبائن وعلى العاملين، مراقبة الإنتاج الخ..

يعرف تنقيب البيانات على إنه "استخراج المعلومات التنبؤية المخفية من قواعد البيانات الكبيرة"

كما يمكن تعريفه على انه هي عملية تحليل البيانات لتحديد العلاقات التي لم تكتشفها التحليلات السابقة من قبل، كما أنها تحليل البيانات لإقامة علاقات وتحديد أنماط.

وتتم عملية التنقيب في البيانات من خلال الخطوات التالية:

1- تنظيف البيانات (Data cleaning): إزالة الضجيج ومعالجة البيانات الغير متناسقة.
2- تكامل البيانات (Data integration): توحيد شكل البيانات ومعانيها بحالة ورودها من مصادر متعددة.
3- اختيار البيانات (Data selection): تحديد البيانات ذات الصلة بموضوع البحث أو التطبيق.
4- تحويل البيانات (Data transformation): تحويل البيانات من صيغة إلى صيغة مناسبة لعملية التنقيب.
5- تنفيذ عملية التنقيب في البيانات (Data mining): تحديد طريقة أو طرق التنقيب المناسبة والخوارزمية أو الخوارزميات المناسبة لتطبيقها ومن ثم تطبيق هذه الخوارزمية أو الخوارزميات.
6- تقييم النماذج (Pattern evaluation): تصنيف النماذج المستخرجة وفق أهميتها.
7- عرض النتائج (Knowledge presentation): استخدام وسائل الاظهار وتمثيل النتائج للمستخدم[1].

3. أنواع تنقيب في البيانات

هناك العديد من وجهات النظر في تصنيف انواع تنقيب البيانات فمن وجهة النظر الموجهة بالعملية تم تحديد ثلاثة انواع هي:

1- الاكتشاف Discovery

وهو عملية النظر في قاعدة البيانات لإيجاد النماذج من دون ان تكون هناك فكرة محددة عما قد تكون عليه.

2- النمذجة التنبؤية Predictive Modeling

[1] (رزوق، 2013)

فيه تستخدم النماذج المكتشفة من قاعدة البيانات للتنبؤ بالمستقبل أي تخمين القيم غير المعروفة بالاعتماد على نماذج سابقة مكتشفة من قاعدة البيانات.

3- التحليل المبرهن Forensic Analysis:

وهو عملية تطبيق النماذج المستخلصة لإيجاد عناصر البيانات الشاذة او غير العادية أي انه يبحث في حالات محددة وغير عادية[2]..

اما من حيث وجهة النظر المرتبطة بطبيعية نشاط التنقيب في البيانات فتكون على ثلاثة أنواع:

4- التنقيب العارض Mining Episodic:

أي النظر إلى البيانات من منظور معين من اجل الوصول إلى نتيجة محددة وينجز هذا النوع من قبل المحللين.

5- التنقيب الاستراتيجي Strategic Mining:

وفيه يكون النظر إلى مجموعات أكبر من البيانات الكلية للحصول على فهم شامل لمقاييس مثل الربحية.

6- التنقيب المستمر Continuous Mining:

وهو يحاول فهم كيفية تغير العالم ضمن مدة محددة من الزمن ومحاولة فهم العوامل التي سببت التغير[3].

4. التصنيف (Classification)

تستخدم عملية تنقيب البيانات تقنيات عديدة تتمكن من خلالها اكتشاف الاتجاهات والنماذج الخفية في مقادير كبيرة من البيانات، ويعتبر التصنيف أحد هذه التقنيات.

[2] (Ahola & Rinta, 2001)
[3] (الدوري و احمد، 2007)

عندما يريد مدير التسويق في شركة الكترونيات تحليل لبياناته ليخمن ما إذا كان زبون معين سيشتري منتجاً معيناً ام لا تكون عملية تحليل البيانات المطلوب اجرائها هي التصنيف.

يعرف التصنيف على انه: وظيفة استخراج البيانات من خلال تعين العناصر في مجموعة للفئات أو الفئات المستهدفة. الهدف من التصنيف هو التنبؤ بدقة بالفئة المستهدفة لكل حالة في البيانات. على سبيل المثال، يمكن استخدام نموذج تصنيف لتحديد مقدمي طلبات القروض على أنهم مخاطر ائتمانية منخفضة أو متوسطة أو عالية. [4]

سهل تصنيف البيانات دراسة كل مجموعة على حدة ومعرفة المميزات المشتركة والغير ظاهرة (Hidden). كما يسهل عمليات التنبؤ (Predictions).

يعد التصنيف أحد أهم أساليب (Supervised Learning) وهذا يعني أننا نحتاج لبناء النموذج Model إلى تقسيم البيانات إلى عينتين، عينة لتدريب النموذج وأخرى لاختباره والتحقق من كفاءته.

1.4 تحضير البيانات

من اجل زيادة دقة عملية التصنيف وكفايتها لابد من تحضير البيانات وفق صيغة معيارية ويشمل تحضير البيانات العمليات التالية:

1- **تنظيف البيانات Data cleaning**

يشير هذا المصطلح إلى عملية معالجة أولية للبيانات بهدف إزالة أو تقليل الضجيج ومعالجة القيم المفقودة، على الرغم من أن معظم خوارزميات التصنيف تمتلك بعض آليات معالجة البيانات المشوشة أو المفقودة فإن هذه الخطوة تساعد من تخفيض مدى الالتباس والغموض خلال عملية التعلم.

2- **تحليل مدى الارتباط Relevance analysis**

ان العديد من واصفات البيانات قد تكون زائدة ومن ثم ان تحليل الارتباط (correlation analysis) يستخدم لتحديد ما إذا كان الواصفتان x1 وx2 مترابطين احصائياً،

[4] (Mills, 2011)

فإذا كان هنالك ارتباط وثيق بين الواصفتين فهذا يعني امكانية الاستغناء عن احدأهما في عمليات التحليل.

3- تحويل البيانات Data transformation

تستدعي عملية نقل البيانات تطبيق عملية (تقييس normalization) في عملية التقييس يتم جعل قيم واصفة ما تقع ضمن مجال رقمي مجدد، وعادة ما يكون مجال صغيراً مثل {0.1} ويمكن أيضاً عند نقل البيانات تطبيق عملية (تعميم generalization) على البيانات لتوليد سويات مفاهيمية أعلى، ويستخدم هنا (هرميات المفاهيم concepts hierarchies) وهذه العملية مفيدة بالتحديد من أجل الواصفات ذات القيم المستمرة، فعلى سبيل المثال يمكن تعميم القيم الرقمية للواصفة (Income) على شكل مجال ذي قيم متقطعة مثل: {Low, Medium, High}.[5]

2.4 تقنيات التصنيف

1.2.4 شجرة القرار (Decision Tree)

تعد من أبرز خوارزميات التصنيف وهي نموذج استكشافي يظهر على شكل شجرة كما يعبر اسمها، وبشكل دقيق يمثل كل فرع من فروع الشجرة سؤالاً تصنيفياً، وتمثل أوراقها أجزاء من قاعدة البيانات تنتمي للتصنيفات التي تم بنائها. .[6]

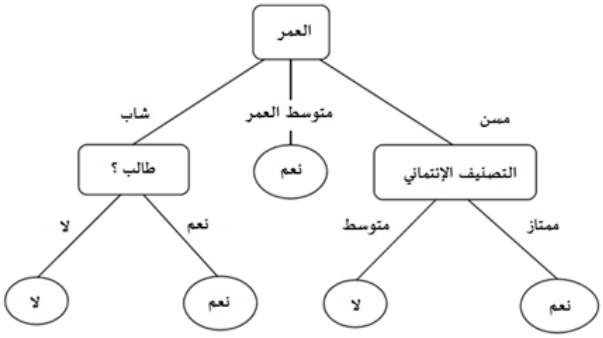
تعتبر خوارزمية شجرة القرار واحدة من الخوارزميات المشهورة في التصنيف، وهي مكونة من مجموعة من العقد والوصلات، وكل عقدة يمكن ان تعبر عن سمة من السمات، وكل وصلة تعبر عن قيمة ما لهذه السمة، أما عقد الأوراق فإنها تعبر عن الهدف المتنبأ به وهو حالة التسرب الوظيفي في هذا البحث.

[5] (رزوق، 2013)
[6] (Dean, 2014)

تكمن شهرة هذه الخوارزميات بسبب سهولة تفسيرها وكونها قادرة على التعامل مع السمات الفئوية ويمكن توسيعها لتتعامل مع أكثر من صنفين. .[7]

يوضح الشكل الاتي شجرة تمثل مفهوم "يشتري حاسوب" موضحاً التصنيف المناسب لهذه العملية.

بعض خوارزميات بناء أشجار القرار تولد اشجاراً ثنائية "حيث كل عقدة داخلية يتفرغ منها عقدتان فقط" وبعض الخوارزميات الأخرى تولد اشجاراً غير ذلك.



رسم توضيحي 1 . شجرة القرار

2.2.4 مصنف j48

يندرج هذا المصنف ضمن خوارزميات أشجار القرار والتي على اختلاف أنواعها تشابه إلى حد ما خوارزمية التصنيف (Bayes) من حيث اعتمادها على الاحتمالات الشرطية مع اختلاف رئيسي يكمن في هذه الخوارزمية تقوم بتوليد قواعد (Rules) لاستخدامها كجمل شرطية لتحديد السجلات والأحداث الاحتمالية بشكل عبارة شرطية (IF.... THEN).

[7] (احمد، 2018)

تعتمد هذه الخوارزمية إلى تقسيم مجموعة بيانات التدريب المراد تصنيفها إلى مجالات متقاطعة (Mutual Exclusive) ذات تسمية أو قيمة أو عملية لتوضيح وشرح البيانات داخل هذا المجال وذلك بالاعتماد على معيار يستخدم لحساب أو تعين أفضل المعاير لتجزئة هذا المجال من البيانات التي يتم تدريبها والذي يدعى التابع الإحصائي (Information Gain) والمعروف بالمعادلة التالية:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

بحيث:

- مجموعة بيانات التدريب (S) ومجموعة المعايير (A)
- Values (A): جميع القيم الممكنة للمعيار A
- (Sv) مجموعة جزئية من المجموعة (S) المنتمية للمعيار (A) وذات قيمة (V)
- تابع العشوائية (Entropy): يعبر هذا التابع عن عشوائية المعطيات وتتراوح قيمته بين (0-1)

ويعبر عنه بالمعادلة التالية:

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

حيث يعبر المتغير (Pi) عن احتمالية انتماء مجموعة البيانات (S) إلى الفئة (i)[8]

3.2.4 الغابة العشوائية (Random Forests)

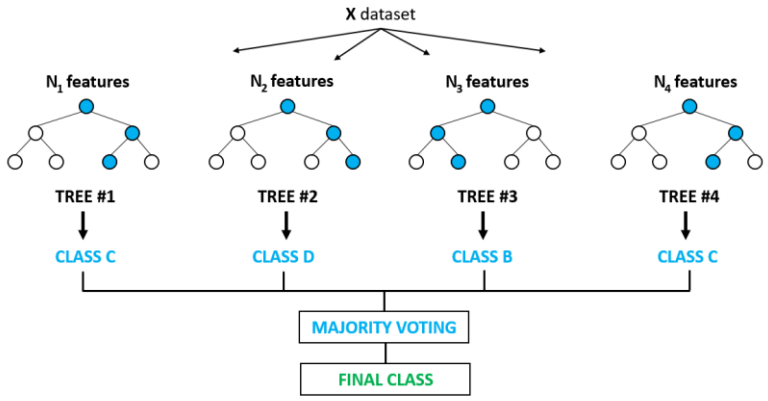
الغابة العشوائية عبارة عن خوارزمية تعليمية متعددة الاستخدامات قادرة على أداء مهام الانحدار والتصنيف. في الوقت نفسه، إنها أيضًا طريقة للحد من أبعاد البيانات، وتستخدم للتعامل مع القيم المفقودة، والقيم المتطرفة، وغيرها من الخطوات المهمة في استكشاف البيانات، وحققت نتائج جيدة.

[8] (الضاهر، 2014)

بالإضافة إلى ذلك، تعمل أيضًا كطريقة هامة في التعلم الجماعي، حيث تُظهر قوتها عند دمج عدة نماذج غير فعالة في نموذج واحد فعال.

في الغابات العشوائية، سنقوم بإنشاء الكثير من الأشجار القرار، وليس فقط توليد الأشجار الفريدة كما هو الحال في نموذج (CART)

عند تصنيف كائن جديد وتمييزه بناءً على سمات معينة، ستعطي كل شجرة في الغابة العشوائية خيار التصنيف الخاص بها و(التصويت) وفقًا لذلك، وسيكون الناتج الكلي للغابة هو أكبر عدد من الأصوات. خيار التصنيف في مشكلة الانحدار سيكون ناتج الغابة العشوائية هو متوسط جميع مخرجات شجرة القرار[9]



رسم توضيحي 2 . الغابة العشوائية

4.2.4 الانحدار اللوجيستي (Logistic regression)

هو نموذج احصائي ينتمي لنماذج الانحدار الخطي يمكن من نمذجة متغير ثنائي الحد بدلالة مجموعة من المتغيرات العشوائية المتوقعة، رقمية كانت او فئوية، يستخدم الانحدار اللوجيستي للتنبؤ باحتمالية وقوع حدث ما بمعرفة إضافية لقيم متغيرات يمكن ان تكون مفسرة أو مرتبطة بهذا الحدث، يشتهر الانحدار اللوجيستي ايضاً بتسميات نموذج لوجيت أو المصنف العام للانتروبية.[10]

[9] (Smith، 2017)
[10] (رزوق، 2013)

يمكن القول ايضاً ان الانحدار اللوجستي هو تصنيف وليس خوارزمية انحدار، ويقدر القيم المنفصلة (القيم الثنائية مثل 1/0، نعم / لا، صواب / خطأ) بناءً على مجموعة معينة من المتغيرات (المتغيرات) المستقلة. ببساطة يتنبأ بشكل أساسي باحتمال وقوع حدث عن طريق ملاءمة البيانات لدالة تسجيل الدخول. القيم التي تم الحصول عليها تقع دائمًا في حدود 0 و1 لأنها تتوقع الاحتمال، يُعرف أيضًا باسم تحويلة لوجت (Logit Transformation).[11]

ويقسم الانحدار اللوجيستي إلى ثلاثة اقسام:

الانحدار اللوجيستي الترتيبي (Ordinal logistic regression)

حيث يتميز هذا النوع بتفسير أثر المتغيرات المنبئة (المستقلة) باختلاف مستويات قياسها على الاستجابات الرتبية بمعنى أن يكون المتغير التابع متغيرا ترتيبيا[12].

الانحدار اللوجيستي الثنائي (Binomial logistic regression)

يعتبر أشهر أنواع الانحدارات اللوجستية، ويستخدم الانحدار اللوجستي الثنائي في تفسير أثر المتغيرات المفسرة على الاستجابات الثنائية، بمعنى تفسير قدرة مجموعة من المتغيرات المستقلة 'المنبئة' ذات المستويات المختلفة على التنبؤ بمتغير واحد تابع يكون ثنائي التفرع (dichotomous) مثل (ذكر/أنثى، مريض/معافى)...، لذلك فالقيم أما أن تكون موجودة (إيجابية) وتأخذ القيمة 1 أو غير موجودة (سلبية) وتأخذ القيمة 0 ،ولكون الباحث الإحصائي يهتم بوجود(تحقق) الاستجابات عند كل مستوى من مستويات المتغير المستقل الذي يعتقد بعلاقته بها، لذلك فإن متغير نسبة الاستجابة المشاهدة هو نسبة الاستجابات المتحققة(الإيجابية)عند كل مستوى من مستويات المتغير التوضيحي، وهي نسبة متغيره من مستوى إلى آخر ويقترب توزيعها كثيرا من التوزيع الطبيعي عندما يكون حجم العينة كبيرا [13]. وتأتي معادلة الانحدار اللوجستي ثنائي الحدين كالتالي:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

[11] (غانم و الجاعوني، 2011)
[12] (دعيش و ساري، 2017)
[13] (Maroof, 2012)

حيث أن:

- ❖ Y تمثل المتغير التابع (الثنائي).
- ❖ B0 تمثل الحد الثابت في معادلة الانحدار اللوجيستي ثنائي الحدين.
- ❖ BnXn تمثل قيمة اللوجبت بالنسبة للمتغيرات المستقلة.

الانحدار اللوجيستي متعدد الحدود (Multinomial logistic regression)

وهو أحد أنواع الانحدار اللوجستي، حيث يعتبر امتداد بسيطا للانحدار اللوجستي الثنائي، يتم استخدامه في حالة كان المتغير التابع يتكون من أكثر من فئتين تصنيفيتين أو اسميتين "مثال في تحديد المتغيرات المنبئة التي تتعلق بمجموعة من الأمراض عن غيرها[14].

5.2.4 الجار الأقرب (Nearest Neighbor Algorithm)

من خوارزميات التنصيف والتنبؤ الشهيرة في التعلم الالي وتعتبر من ضمن مجموعة التعلم المراقب (supervised learning) والتي تهدف للتنبؤ عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ له تتلخص فكرة هذه الخوارزمية على الشكل التالي: إذا كانت معظم عينات k الأكثر تشابهًا (أي أقرب جيران في مساحة الميزة) لعينة تنتمي إلى فئة معينة، فإن العينة تنتمي أيضًا إلى هذه الفئة. يتم إعطاء خوارزمية الجوار-k أقرب ما يسمى مجموعة بيانات التدريب، وعلى سبيل المثال إدخال جديد، ابحث عن مثيلات k الأقرب إلى المثيل في مجموعة بيانات التدريب (أي، المذكورة أعلاه K المجاورة، معظم هذه الحالات k تنتمي إلى فئة معينة) ويتم تصنيف مثيل الإدخال في هذه الفئة. 15

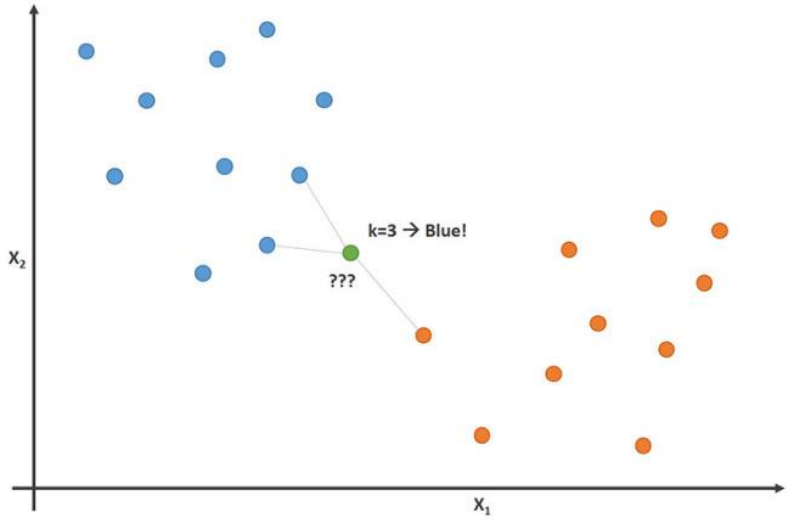
تعتمد الخوارزمية في عملها على قياس المسافة الإقليدية بين كل نقطة والنقطة الأقرب إليها وعندما تكون البيانات قريبة من بعضها تكون المسافة الإقليدية قليلة جدا بين كل نقطة والنقطة المجاورة لها ولكن كلما تباعدت قيم البيانات وتبعثرت أصبحت المسافات بين النقاط كبيرة ومنها جاء عنوان الخوارزمية، إذ يشير

[14] (Bayaga, 2010)
[15] (عبيد م.، 2019)

الحرف K إلى، الحالات التي سيتم تصنيفها بناء على المسافات بينه، يتم حسب المسافة من المعادلة التالية:

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(Xi - Yi)^2}$$

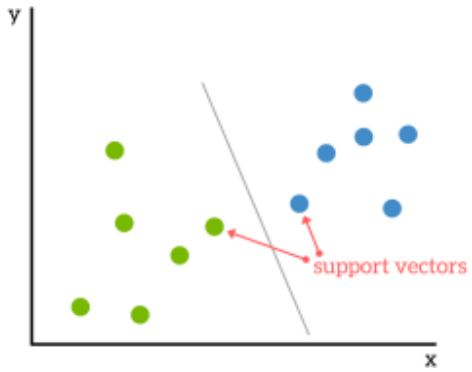
حيث أن: (d: المسافة بين أي نقطتين) و(Xj:Xi نقاط تموضع البيانات في فضاء البحث)



رسم توضيحي 3 . خوارزمية KNN

آلة المتجهات الداعمة (Support vector machine): هي خوارزمية تعلم آلي خاضعة للإشراف (Supervised Learning) يمكن استخدامها في مسائل التصنيف أو الانحدار ومع ذلك، فإنها تستخدم في الغالب في مسائل التصنيف. في خوارزمية (SVM)كما تعتبر طريقة تصنيف للبيانات خطية وهي طريقة غير إحصائية، نرسم كل عنصر من عناصر البيانات كنقطة في الفضاء ذو بعد n حيث n هو عدد الخصائص لديك مع قيمة كل خاصية هي قيمة إحداثيات معينة. ثم، نقوم بإجراء التصنيف من خلال إيجاد المستو فائق (Hyper-plane) الذي يميز الفئتين جيدًا.[16]

[16] (Dean, 2014)



رسم توضيحي 4 . خوارزمية SVM

المتجهات الداعمة (Support Vectors) هي ببساطة إحداثيات المراقبة الفردية. مصنف له المتجهات الداعمة (SVM Classifier) هو الحد الفاصل بين الفئتين(hyper-line / line) .

ومن الضروري توضيح مفهومين اساسين تم التطرق لهما من خلال التعريف وهما:

❖ المتجهات الداعمة

المتجهات الداعمة هي نقاط البيانات الأقرب للمستوي الفائق وهي النقاط التي إن تم إزالتها من مجموعة البيانات ستغير من موقع المستوي الفائق الذي يقسم البيانات. لذلك يمكن اعتبار هذه النقاط أنها العناصر المهمة في مجموعة البيانات[17].

[17] (Kelleherd & Tierney, 2018)

❖ المستوي الفائق

كمثال بسيط، لنأخذ عملية التصنيف لمجموعة بيانات ذات خاصيتين فقط يمكنك اعتبار المستوي الفائق أنه الخط الذي يفصل خطياً ويصنف مجموعة البيانات.

بديهياً كلما ابتعدت النقاط عن المستوي الفائق كلما ازدادت ثقتنا بأنه تم تصنيف النقاط بشكل صحيح. لذلك نود أن تكون النقاط بعيدة عن المستوي الفائق قدر المستطاع مع مراعاة بقاءها على الجانب الصحيح من الخط الفاصل.

وبذلك عندما نضيف بيانات جديدة لاختبارها سيتم تصنيفها بناءاً على الجانب الذي تقع فيه بالنسبة للمستوي الفائق[18].

[18] (Dean, 2014)

المبحث الثاني

قطاع الاتصالات Telecom Sector

1. مفهوم قطاع الاتصالات

يتكون قطاع الاتصالات من الشركات التي تجعل الاتصال ممكنًا على نطاق عالمي، سواء كان ذلك من خلال الهاتف أو الإنترنت، من خلال الموجات الهوائية أو الكابلات، من خلال الأسلاك أو لاسلكيًا. أنشأت هذه الشركات البنية التحتية التي تسمح بإرسال البيانات بالكلمات أو الصوت أو الصوت أو الفيديو إلى أي مكان في العالم. أكبر الشركات في القطاع هي مشغلي الهاتف (السلكية واللاسلكية) وشركات الأقمار الصناعية وشركات الكابلات ومقدمي خدمات الإنترنت.

منذ وقت ليس ببعيد، كان قطاع الاتصالات يتألف من نادٍ من كبار المشغلين الوطنيين والإقليميين. منذ أوائل العقد الأول من القرن الحادي والعشرين، انجرفت الصناعة في سرعة إلغاء التنظيم والابتكار. في العديد من البلدان حول العالم، يتم الآن خصخصة الاحتكارات الحكومية وهي تواجه عددًا كبيرًا من المنافسين الجدد. انقلبت الأسواق التقليدية رأساً على عقب، حيث تفوق النمو في خدمات الهاتف المحمول على الخطوط الثابتة وبدأت الإنترنت في استبدال الصوت باعتباره الأعمال الأساسية.

2. نظرة عامة على قطاع الاتصالات

يتوقع المحللون أن ابتكار المنتجات وزيادة عمليات الدمج والاستحواذ لن يؤدي إلا إلى استمرار النمو والنجاح في صناعة الاتصالات. هناك العديد من الفرص للمستثمرين، ولن تؤدي زيادة المستثمرين إلا إلى إفادة القطاع بشكل أكبر.

إن استقرار نمو القطاع، حتى خلال فترات الركود، يعني أنه يعتبر استثمارًا دفاعيًا قويًا مع الحفاظ على جاذبيته للمستثمرين الذين ينمون. حتى في الأوقات الاقتصادية غير المستقرة والمتقلبة، فإن الطلب الثابت على خدمات الصوت والبيانات، إلى جانب خطط الاشتراك الواسعة، يضمن مصدرًا ثابتًا للإيرادات لشركات الاتصالات الكبرى.

أصبحت الاتصالات صناعة أساسية ذات أهمية متزايدة، مما يبشر بالخير لآفاقها المستقبلية ونموها المستمر. إن التقدم المستمر في خدمات الهاتف المحمول عالية السرعة والاتصال بالإنترنت بين الأجهزة يحافظ على دفع الابتكار والمنافسة داخل هذا القطاع. ينصب جزء كبير من تركيز الصناعة على توفير خدمات بيانات أسرع، لا سيما في مجال الفيديو عالي الدقة. بشكل أساسي، تتجه القوى الدافعة نحو خدمات أسرع وأكثر وضوحًا وزيادة الاتصال واستخدام التطبيقات المتعددة.

تستمر اقتصادات السوق الناشئة في كونها نعمة للصناعة، مع معدل نمو صناعة الهواتف المحمولة في بلدان مثل الصين والهند يدفع قدرات منتجي الأجهزة لمواكبة مستوى الطلب.

في الولايات المتحدة، يولي المحللون اهتمامًا وثيقًا للقضايا المتعلقة بحيادية الشبكة حيث يستمر الطلب على خدمات البيانات والفيديو في الزيادة في المستقبل. لا يزال هناك طلب قوي على حقوق الطيف اللاسلكي، ناهيك عن الاتجاه المتزايد نحو التوحيد من خلال عمليات الدمج والاستحواذ.

المبحث الثالث

إدارة علاقة الزبائن CRM

1. مفهوم إدارة علاقات الزبائن

هي الفلسفة التي تضع الزبائن في نقطة تصميم المنتجات من أجل توجيه موارد وجهود المنظمة لتقديم أفضل الخدمات وتعزيز ولاء الزبائن لها. [19]

وهي إستراتيجية أعمال محور اهتمامها هو العميل بالدرجة الأولى للحصول على رضاه والمحافظة عليه والاستحواذ على ولائه عن طريق تقديم خدمة مميزة له.[20]

وهي فلسفة أعمال تسمح للمنظمة فهم تفكير وتصرفات الزبائن وتحليل احتياجاتهم ومعرفة تطلعاتهم من خلال المعلومات المخزنة في قواعد البيانات، حتى تتمكن الشركات التوصل لما يرغبون به وأيضاً التنبؤ بسلوكهم مستقبلاً واتخاذ قرارات تسويقية صائبة من حيث التوقيت والنوعية، للمحافظة على مستوى ربحية أعمالها وتنميتها.[21]

إن تحديد نوعيات الزبائن، وفئاتهم، وما يرغبون به من منتجات وما يعانونه من مشاكل، خاصة على صعيد الخدمات والذي يعتبر من أهم عوامل المحافظة على ولاء الزبائن، حيث تعاني الشركات الكبرى من تسرب الزبائن، وهناك دراسات إحصائية تشير إلى أن معدل فقدان الزبائن لدى الشركات قد يبلغ 20 % من عدد العملاء الكلي كل عام، وأن كلفة عملية اكتساب عميل جديد قد تبلغ ستة أضعاف كلفة المحافظة على العميل الموجود. وأن 68 % من العملاء يغيرون الشركات والمؤسسات التي يتعاملون معها بسبب الخدمات، والملاحظ أن 4 %فقط من هؤلاء العملاء الذين توقفوا عن التعامل سبق لهم أن اشتكوا من سوء في الخدمات. بينما 90 % تركوا بدون سابق إشعار. وأن 82 %من العملاء الذين تم حل مشكلاتهم عاودوا إلى التعامل مع نفس الشركات. [22]

لذلك تكمن أهمية التعرف بشكل أكبر على حاجات العملاء ورغباتهم وأهم خصائص المنتجات التي يحتاجها الزبائن ويرغبون بها لكسب الزبائن والحفاظ عليهم وتعزيز ولائهم بأنها جهد. كما أنه لتمييز متكامل وإبقاء وتعزيز العلاقة مع الزبائن وتقوية العلاقة معهم بشكل مستمر، لتبادل المنفعة من كل الجوانب وإضافة قيمة لهم.

---

[19] (morse, 2004)
[20] (Greenberg, 2002)
[21] (Bygstad, 2003)
[22] (العنزي، 2010)

2. أهمية ادارة علاقة الزبائن

1.2 الوصول للعملاء المحتملين

"بدون نظام CRM جيد، لن تتمكن من تحويل حوالي 97% من العملاء المحتملين إلى مبيعات حقيقية."

الآن بعد أن استهلكت وقت وموارد في جذب العملاء المحتملين الجدد، ما الخطوة التالية؟ هل تقوم بتحويلهم لفريق المبيعات الخاص بك مباشرة، وإن كانت الإجابة نعم، فما هي الفرص الحقيقية بأن يتم تحويلهم لعملاء دائمين؟ هنا أنت بحاجة للاستفادة من أدوات التسويق الخاصة بك كالبريد الإلكتروني ووسائل التواصل الاجتماعي وأتمتة التسويق من خلال ربطها بمنصة CRM الخاصة بك.

في الوقت الذي تصبح بيانات عملائك كلها متاحة على منصة CRM سيكون لكلًا من المبيعات والتسويق رؤية كاملة عن هؤلاء العلماء المحتملين والتوقعات الخاصة بحالتهم ليتمكنوا من إنشاء واستهداف اتصالات جذابة لتحويلهم إلى عملاء حقيقين الوصل إلى صانعي القرار بشكل أسرع.

2.2 تعزيز واستدامة العلاقة مع عملائك

"46% من قادة المبيعات يؤكدون أن العلاقات الأعمق مع العملاء هي هدف رئيسي للحفاظ على النجاح"

من خلال نظام إدارة علاقات العملاء CRM الجيد يمكنك تطوير الفهم العميق لكل ما يهم عملائك وبناء علاقة قوية مبنية على الثقة والنجاح المتبادل. كما يساعدك نظام CRM على:

استكشاف تحدياتهم: اكتشف كل ما يهم عملائك من أهداف وتحديات وتفضيلات، وتابعهم باستمرار. قم بتسجيل كل الملاحظات في نظام إدارة علاقات العملاء CRM الخاص بك بالتفصيل حتى تتمكن في المرة القادمة من المتابعة من حيث توقفت بمراجعة سريعة.

تقديم التوصيات الأنسب: بعد فهمك للتحديات والأهداف التجارية لعميلك، وقتها فقط يمكنك تقديم التوصيات والعروض الترويجية الأنسب، أو أي محتوى له صلة بهذه الاهتمامات. كما يمكنكم ارشاد عميلك إلى كيفية استخدام منتجك أو خدماتك من خلال منصة إدارة علاقات العملاء لتبسيط عملية الاستخدام والمتابعة.

3.2 تقليل تكلفة المبيعات

"إن كانت احتمالية البيع لعميل محتمل جديد تقع ما بين 5% – 20%، فاحتمالية البيع للعميل الحالي تقع ما بين 60% – 70%."

يمثل العملاء الجدد عنصرًا رئيسيًا لاستمرار النمو، لكن ليس من السهل الحصول عليهم والوصول معهم لمرحلة البيع. الأمر السار في الموضوع أنه يمكنك تعويض التكاليف المدفوعة على اكتساب عملاء جدد من خلال المبيعات لقاعدة العملاء الحاليين.

4.2 تقديم أفضل خدمة/ منتج للعميل

"نسبة 55% من المستهلكين مستعدون لدفع المزيد مقابل تجربة خدمة عملاء أفضل"

من أحد أهم الفوائد الرئيسية لأنظمة إدارة علاقات العملاء CRM هي مساعدة مندوبي المبيعات على البيع أكثر وأسرع، فالوصول إلى سجل تفاعل العملاء من خلال رحلة العميل الكاملة يسمح للمندوب توقع حاجة العميل وتقديم أفضل الخدمات في الوقت المناسب.

عندما يتمتع فريقك بإمكانية الوصول الفوري إلى السجل الكامل للعميل، يمكن للجميع تقديم رسائل وحلول مخصصة بسرعة باستخدام الموارد المناسبة.

5.2 الاحتفاظ بالعملاء وضمان ولائهم

"خفض نسبة فقد العملاء بنسبة 5% يمكن أن يتسبب في زيادة الأرباح بنسبة من 25% وحتى 85%"

حسن رؤية فريقك لعلاقة مؤسستك مع جميع العملاء تساعدك على التنبؤ بالحسابات المعرضة للخطر والتعامل معها بشكل مناسب، بجانب تقديم فرص جديدة للعملاء الراضيين. من خلال الشفافية في تاريخ العملاء والحملات النشطة او الحالات المفتوحة، يمكنك توفير عمليات شراء وتجارب خدمة مرضية لعملائك تجعلهم يختاروا التعامل معك دائمًا ودفع المزيد.

لذلك كل ما تستثمره في توفير خدمة عملاء جيدة الآن يعود إليك في أرباح بصور مختلفة في المستقبل.

3. أ نواع البيانات التي تحتاجها إدارة العلاقة مع الزبائن

- بيانات ديموغرافية وتتضمن: السن، النوع، الوظيفة، والمكانة الاجتماعية لكل زبون من زبائن الموسسة.
- بيانات اتصال وتتضمن كلا من: عنوان السكن، أرقام الهاتف، وأماكن وجوده ووسائل الاتصال المفضلة والموظف المختص بالاتصال لكل زبون من زبائن المؤسسة.
- بيانات الدخل والاستهلاك وتتضمن كلا من: القدرة الشرائية، السلع المشتريات وكمياتها والسلع المفضلة، وعلامتها التجارية وأنماط الاستهلاك الخاصة بكل زبون من زبائن المؤسسة.
- بيانات أخرى وتتضمن كلا ما يلي: الأفراد المؤثريين على قرار الشراء ونسبة الانفاق على السلع المنافسة وأسباب تعامل كل زبون من زبائن المؤسسة معها ومقترحاتهم ورأيهم الشخصي في المنافسين للمؤسسة.

4. المنافع المحققة بالاعتماد على ادارة علاقة الزبون

تعمل ادارة علاقة الزبائن على تحقيق المنافع التالية:

1. التوجه نحو الزبون.
2. تمكين الزبون من المشاركة في دعم مسيرة المؤوسسة وتعزيز موقعها التنافسيفي السوق (شركاء لا زبائن).
3. منهجية ادارية مدركة للحاجات الانسانية مما يؤدي إلى ضرورة احداث التمييز في الخدمات المقدمة بتطلعات الأفراد وتوقعاتهم حول أنفسهم والمحيطين بهم.
4. تبني مفهوم المؤوسسة الحاضنة لزبائنها واحلاله مكان المؤسسة الجاذبة لهم.
5. توكيد مستمر على معايير الجودة والالتزام الدائم نحو الزبون .
6. التطور والابتكار وذلك بفتح آفاق البحث والتطوير للوصول إلى مستويات أعلى من الرضا لدى الزبائن مما يقدم لهم.
7. تحسين مشتمر للخدمة والحرص على تخفيض التكاليف.

8. تبني معلومات فعالة قادرة على تكوين قواعد معلومات إستراتيجية لدعم مراكز صناعة القرار في الوقت المناسب.
9. التمكين الموظيفي أي الاهتمام بالزبائن ورعايتهم من جانب موظفي المؤسسة لا يمكن أن يتحقق طالما أن هؤلاء الموظفين بعيدين عن ادراك دورهم ومشاركتهم , فإن لم تتوفر الحافزية الكافية للأداء والرغبة الأكيدة في العمل لا يستطيع تكوين اتجاهات ايجابية نحو الزبائن ولهذا لابد على المؤسسة أن توفر لموظفيها مساحات أوسع من التصرف وتدريبهم على أن يكونو شركاء في مسؤولية خدمة الزبائن .
10. خفض نسبة تسرب الزبائن وزيادة معدل الاحتفاظ بالزبائن المربحين.

ونرى أنه من أكبر المشاكل التي تتعرض لها الشركات الخدمية نتيجة إضطراب في إدارة علاقة الزبون هي تسرب الزبائن.

المبحث الرابع

تسرب الزبائن Customer Churn

1. تعريف تسرب الزبائن (Customer churn)

اليوم في القرن الحادي والعشرين، تعد شركات الاتصالات ذات الحصة السوقية الأكبر في مجال التنافس، يعد التنافس الأول لهذه الشركات هو المعني بالمصادر الربحية القادمة من العملاء.

حيث تعمل إدارة علاقات العملاء (CRM) على أن العملاء الحاليين هم المصدر الأكثر خصوبة للربح ،وتعد البيانات الخاصة بهم هي الأفضل بعملية اتخاذ القرارات على المستوى الاستراتيجي والتكتيكي.

نستنتج من ذلك، أن المنظمات شديدة التنافس قد أدركت أن الاحتفاظ بالعملاء هو أقل تكلفة من القيام بالبحث والتسويق للحصول على عملاء جدد.

وفي سياق ما سبق، برز نهج استخدام تقنيات التنقيب في البيانات منهج قوي على مدار السنوات الأخيرة باعتباره نهج لعملية اتخاذ القرارات من قواعد البيانات ومستودعات البيانات ذات المغزى من عملية قياس التسرب الوظيفي الحاصل في تلك المنظمات.

والمقصود بالتسرب في هذا السياق هو ميل العملاء إلى التخلي عن علامة تجارية وذلك بالتوقف عن كونهم عملاء يدفعون لشركة معينة. حيث يُطلق على النسبة المئوية للعملاء الذين توقفوا عن استخدام منتجات الشركة أو خدماتها خلال فترة زمنية معينة بما يسمى (معدل تضاؤل العميل – الاستنزاف).

إن توقع العملاء الذين من المحتمل أن يغادروا الشركة سيمثل مصدر إيرادات إضافيًا لتلك الشركات وهذا في سياق إذا تم في مرحلة زمنية مبكرة نظرًا للتأثير المباشر على عائدات الشركات لا سيما في مجال الاتصالات، حيث تسعى الشركات إلى تطوير وسائل للتنبؤ بالعملاء المحتمل تخليهم عن الخدمة، لذلك من المهم العثور على العوامل التي تزيد من تضخم العملاء لاتخاذ الإجراءات اللازمة لتقليل هذا التسرب.

تتمثل المسأهمة الرئيسية لعمل الباحث في هذا البحث على تطوير نموذج للتنبؤ بتسرب الزبائن الذي يساعد مشغلي الاتصالات على التنبؤ بالعملاء المرجح أن يكونوا عرضة للتسرب.

يستخدم النموذج الذي تم تطويره في هذا العمل تقنيات التنقيب في المعطيات على مجموعة البيانات الضخمة.

2. تأثير تسرب الزبائن على الشركات الخدمية

يمكننا القول إن تجربة العملاء هي التي تحدد تصور العلامة التجارية وتؤثر على كيفية إدراك العملاء القيمة مقابل المال للمنتجات أو الخدمات التي يستخدمونها.

يقول مايكل ريدبورد، المدير العام لـ (Hubspot Servuce)

معدل صغير من التسرب الشهري / ربع السنوي سوف يتراكم بسرعة بمرور الوقت.

1% فقط من التسرب الشهري يترجم إلى ما يقرب من 12 في المائة سنويًا ,نظرًا لأن الحصول على عميل جديد يعد أكثر تكلفة بكثير من الاحتفاظ بعميل حالي ،فإن الشركات ذات معدلات التسرب المرتفعة ستجد نفسها بسرعة في فجوة مالية حيث يتعين عليها تخصيص المزيد والمزيد من الموارد لإكتساب عملاء جدد".

يلخص الخبير إلى أن الشركات ذات معدلات التسرب المرتفعة لا تفشل فقط في تحقيق علاقاتها مع العملاء السابقين ولكنها تضر أيضًا بجهود الاستحواذ المستقبلية من خلال إنشاء كلام سلبي حول منتجاتها.

وعليه نستنتج أنه على الشركات الاستثمار في اكتساب عملاء جدد، في كل مرة يغادر فيها العميل، فإنه يمثل خسارة استثمار كبيرة. يجب تخصيص الوقت والجهد لاستبدالها. يمكن أن توفر القدرة على التنبؤ بالموعد المحتمل لمغادرة العميل، ومنحهم حوافز للبقاء، مدخرات ضخمة للأعمال.

3. الحد من تسرب الزبائن (Customer churn)

كسب العملاء الجدد يعتبر تحديًا كبيرًا، حيث يجهل العديد من أصحاب الشركات أن الحفاظ على الزبائن الحاليين هو مكسب مهم، بل هدف أعلى أولوية من جذب عملاء جدد.

تحاول الإدارة فهم سلوكيات العملاء الدقيقة والسمات التي تشير إلى مخاطر وتوقيت مغادرة العملاء مستعينين في أربع نقاط أساسية:

- وضع خطة استباقية من قبل الإدارة، سيكون لها تأثير كبير على الأحتفاظ بالعملاء.
- تقديم عروض أو حوافز خاصة تركز على الاحتفاظ بالعملاء.
- التفاعل مع العملاء، وذلك يتم عن طريق اشراك العملاء بالنشاطات والقرارات حتى تلك التي تعتبر إدارية

- تحديد العملاء الأكثر قيمة، حيث لايمكن الجزم بأن جميع العملاء لهم نفس القيمة المقدمة للشركة، وهو في سياق وضع الخطة الاستباقية بشكل متسلسل تبدأ من العملاء الأكثر أهمية

بجمع كل ذلك معًا، والتنبؤ بتسرب العميل أمر مهم. يمكن اتخاذ إجراءات فعالة للاحتفاظ بالعميل قبل فوات الأوان، تمثل القدرة على التنبؤ بأن العميل معرضًا لخطر كبير بالتسرب بينما لا يزال هناك متسع من الوقت للقيام بشيء حيال ذلك.

4. التنبؤ بتسرب العملاء باستخدام تقنيات التنقيب في البيانات

الشركات التي تراقب باستمرار كيفية تفاعل الأشخاص مع الخدمات، وتشجع العملاء على مشاركة الآراء، وحل مشكلاتهم على الفور لديها فرص أكبر للحفاظ على علاقات العملاء ذات المنفعة المتبادلة، كل ذلك يعتبر بيانات يتم جمعها من قبل الشركات لفترة زمنية.

حتى تتمكن من استخدامها لتحديد أنماط سلوك العملاء المحتملين، وتقسيم هؤلاء العملاء المتوقع تخليهم عن الخدمة، واتخاذ الإجراءات المناسبة لاستعادة ثقتهم.

حيث يتم اتباع نهجًا استباقيًا لإدارة تسرب العملاء عبر استخدم التحليلات التنبؤية باحتمالية النتائج أو الأحداث أو القيم المستقبلية من خلال البيانات الحالية والتاريخية.

تستخدم تقنيات التنقيب في المعطيات للتنبؤ خوارزميات مختلفة لعرض النتائج المراد التنبؤ بها بناء على ماهو نوع التنبؤ المراد والبيانات المتوفرة مثال:

Classification - clustering - neural networks

السمة الرئيسية للتنقيب في البيانات هي بناء أنظمة قادرة على إيجاد أنماط في البيانات مخفية والتعلم منها بدون برمجة واضحة.

كما هو الحال مع أي مهمة للتنقيب في المعطيات، يحتاج متخصصو علوم البيانات أولاً إلى البيانات للعمل معها. اعتمادًا على الهدف، يحدد الباحثون البيانات التي يجب عليهم جمعها. بعد ذلك، يتم إعداد البيانات المحددة ومعالجتها مسبقًا وتنظيفها وتحويلها في شكل مناسب لبناء نماذج التنبؤ. بمجرد اختيار النموذج الذي يقوم بالتنبؤات بأعلى دقة، يمكن وضعه قيد الإنتاج.

قد يبدو النطاق العام لبيانات العمل الذي ينفذه العلماء لبناء أنظمة تعمل بنظام التنقيب في المعطيات قادرة على التنبؤ بتناقص العملاء كما يلي:

1) فهم المشكلة والهدف النهائي
2) جمع البيانات
3) إعداد البيانات والمعالجة المسبقة
4) النمذجة والاختبار
5) نشر النموذج والمراقبة

الفصل الرابع

الإطار العملي

تمهيد

يتناول هذا الفصل المنحى العملي والذي يتمثل ببناء نموذج قادر على التنبؤ بتسرب الزبائن من شركة اتصالات، وقد تم تقسيم هذا الفصل إلى ثلاثة مباحث،

المبحث الأول: يهدف للتعريف بالحالة العملية في شركة الاتصالات، وتجميع البيانات، ويعطي وصفاً موجزاً للبيانات المستخدمة في البحث مع تفسير معانيها، وتعريف بالأدوات المستخدمة في الجانب العملي للمشروع.

المبحث الثاني: سيكون عبارة عن استعراض visualize للبيانات وشرح مختصر لأهم المعلومات المستخلصة من البيانات، لمحاولة معرفة الأسباب التي تدفع العميل للتخلي عن الخدمة، وللتعرف على أهم العوامل المؤثرة في قرار العميل بالتخلي أو بالبقاء.

المبحث الثالث: يتضمن تحضير البيانات وتنظيفها بالشكل الذي يوصلنا إلى بيانات سليمة قادرة على إعطائنا المعلومات المطلوبة.

المبحث الرابع: يحتوي التصنيف، لمحاولة معرفة الأسباب التي تدفع الزبون للتسرب من شركة اتصالات من خلال اختبار خوارزميات تصنيفية تسمح لنا ببناء نموذج قادر على التنبؤ بتسرب الزبائن.

المبحث الخامس: يقوم على المقارنة الخوارزميات السابقة، واختيار الخوارزمية التصنيفية الأفضل لهذا التنبؤ.

المبحث الأول

الحالة العملية: شركة اتصالات X

مجموعة بيانات من شركة اتصالات X، تحتوي البيانات على معلومات حول ما يقرب ل 6 آلاف مستخدم، خصائصهم الديموغرافية، الخدمات التي يستخدمونها، مدة استخدام خدمات المشغل، طريقة الدفع.

تم تحميل البيانات من موقع Kaggle [23] الخاص بنشر مجموعات البيانات التابعة للمنظمات والجامعات والشركات، و التي نشرت بغرض طرح بعض المقترحات من قبل المستخدمين لحل هذه المشكلة، و بغرض البحث العلمي .

المهمة هي تحليل البيانات والتنبؤ بتغيير المستخدمين (بتحديد الأشخاص الذين سيجددوا عقودهم والذين لن يجددون) أو هي التنبؤ بتسرب الزبائن وتركهم لشركة الاتصالات.

لتعظيم عدد العملاء اتجهت الشركة للاحتفاظ بعميل الشركة القديم بدلاً من جذب عميل جديد، حيث سيكلف الاحتفاظ بعميل الشركة أقل من جذب عميل ذلك.

في حال الاتجاه نحو العميل القديم، فسيكون لدينا بالفعل البيانات اللازمة حول التفاعل مع الخدمة، وعليه فإنه عند التنبؤ في حدوث اضطراب، سيكون الرد في الوقت المناسب وأسرع لمحاولة إبقاء العميل الذي يريد الاستغناء عن الخدمة.

بناءاً على البيانات الخاصة بالخدمات التي يستخدمها العميل، يمكننا أن نعرف بسهولة ما العرض الخاص الذي يمكن أن يقدم إليه، لمحاولة تغيير قراره بترك المشغل.

بهذا ستكون مهمة الحفاظ أسهل من جذب عملاء جدد لا نعرف عن سلوكياتهم أي شيء.

[23] الموقع الرسمي لموقع كيغل(https://www.kaggle.com/)

شرح البيانات

تتكون البيانات من 22 واصفة Attributes، 13 من هذه الواصفات هي واصفات ذات قيم صريحة (مباشرة بشكل لا لبس فيه)، 5 واصفات ذات قيم منطقية (متغير ثنائي له قيمتان محتملتان تسمى " صواب " و "خطأ") و3 واصفات هي واصفات ذات قيم رقمية.

المتغير "CHURN" هو المتغير المستهدف Target Attribute.

Attribute	Content	Information
Unnamed	5986	غير معرَف.
Customer ID	5986	معرف العميل.
Gender	Male Female	جنس العميل.
Senior Citizen	No Yes	سواء كان العميل متقاعداً أم لا.
Partner	No Yes	ما إذا كان العميل متزوجاً.
Dependents	0 1	هل العملاء مستقلين مادياً.
Tenure	From 0 To 72	مدة الخدمة: وهي عدد الأشهر التي كان بها العميل عميلاً للشركة.
Phone Service	No Yes	هل الخدمة الهاتفية مفعلة.
Multiple Lines	No No Phone Service	ما إذا كانت خطوط الهاتف المتعددة متصلة.

	Yes	
Internet service	DSL Fiber Optic No	مزود الإنترنت للعميل.
Online Security	No No Internet Service Yes	هل تم تمكين خدمة الأمان عبر الإنترنت.
Online Backup	No No Internet Service Yes	هل خدمة النسخ الاحتياطي عبر الإنترنت مفعلة.
Device Protection	No No Internet Service Yes	هل لدى العميل تأمين على المعدات.
Tech Support	No No Internet Service Yes	هل خدمة الدعم الفني مفعلة.
Steaming TV	No No Internet Service Yes	هل خدمة البث التلفزيوني مفعلة.
Steaming Movies	No No Internet Service Yes	هل تم تنشيط خدمة السينما والأفلام.
Contract	Month-to-month One year Two years	نوع عقد العميل.

Paperless Billing	No Yes	ما إذا كان العميل يستخدم الفواتير غير الورقية.
Payment Method	Bank Transfer (Automatic) Credit Card (Automatic) Electronic Check Mailed Check	طريقة الدفع.
Monthly Charges	From 18.25 To 118.75	الدفعة الشهرية الحالية.
Total Charges	From 18.8 To 8684.8	المبلغ الإجمالي الذي دفعه العميل مقابل الخدمات طوال الوقت.
CHURN	No Yes	هل هناك تسرب بالزبائن أم لا.

الأدوات المستخدمة في البحث

تمهيد

أداة البحث العلمي هي الطريقة التي يستخدمها الباحث في جمع المعلومات والبيانات من أجل تقديم إجابات قوية مصحوبة بالأدلة لأسئلة البحث العلمي، حيث تتحكم طبيعة الفروض في الأدوات المستخدمة في البحث، وبالتالي ليس هناك تصنيف يلزم الباحث باستخدام نوع محدد من أدوات البحث العلمي.

ونظراً لندرة البيانات في بعض المواضيع، يتوجب على الباحث أن يلم بعدة طرق، وأدوات تمكنه من الحصول على بياناته ودراسة مشكلة البحث، الأمر الذي يسأهم في حصوله على بيانات دقيقة وموضوعية، وتجنب البيانات المغلوطة، التي تعتبر أحد نقاط ضعف البحث العلمي.

وفيما يلي نوضح الأدوات والبرامج التي وجدت ملائمة للاستخدام في هذا البحث العلمي:

- **Weka**

مجموعة من الأدوات والبرمجيات المتقدمة التي تساعد المستخدم في التنقيب عن البيانات وتجميعها وتحليلها باحترافية من خلال خوارزميات معقدة كتعلم الآلة والتعدين وغيرها وكل هذا يساعد على معالجة البيانات بشكل قوي وتجربة القواعد والخوارزميات عليها وحتى يمكن تحرير البيانات وتغيير سماتها وتصنيفها دون استهلاك الكثير من موارد جهاز الكمبيوتر.

وحزمة ويكا Weka توفر للمستخدمين مجموعة من الأدوات ومخططات التعلم (خوارزميات متقدمة) والتي تساعد على التنقيب عن البيانات واستخراجها بكل سهولة ويمكن الاستعانة بتلك الخوارزميات وتطبيقها بشكل مباشر على مجموعة من البيانات أو استخدامها بواسطة الجافا كود الخاصة بك بدون مشكلة.

حزمة Weka تضم 4 أدوات متاحة ويمكن الوصول إليها وهم(Explorer, Experimenter , Simple CLI , Knowledge Flow)، وتسمح لك تلك الأدوات المهمة فتح مجموعة من البيانات وتحريرها كما يحلو لك بكل سلاسة، كما يمكن تغيير محتويات البيانات وتغيير السمات وتصنيف البيانات المتاحة وفقا لمجموعة محددة مسبقا من القواعد.

علاوة على إمكانية إجراء تحليل التكلفة والمنفعة واستعراض مصفوفة التكلفة ومنحني العتبات، وعبر الأدوات التي يوفرها برنامج Weka يمكن تجميع البيانات واستخدام الخوارزميات والقواعد وخصائص التقييم لتخطيط البيانات وعرض وتحليل الرسوم والمخططات البيانية بكل سهولة ودون صعوبات.

تم تطوير تطبيق weka في مخابر جامعة (Waikato) في نيوزيلندا بالاعتماد على لغة البرمجة جافا وكانت النسخة الأولى في عام 1996، ويعتبر أحد أفضل برمجيات التنقيب في البيانات حيث يعمل على كافة أنظمة التشغيل (Windows, Linux, Macintosh) كما يمكن تحميل البيانات إلى التطبيق من عدة مصادر تتضمن:

1. الملفات ذات الصيغ المعروفة لهذه التطبيق (arff, xrff, data, csv, bsi).
2. مختلف قواعد البيانات: (SqlServer, MySQL, Oracle).
3. مسارات ونطاقات عناوين الحواسب ضمن الشبكات الحاسوبية (URLs).[24]

يضم هذا التطبيق مجموعة كبيرة من أدوات التنقيب في المعطيات والتعلم الآلي مجانية المصدر وتتضمن الواجهة الرسومية العمليات الرئيسية التالية:

1. تصوير البيانات. (Data Visualize).
2. التصنيف. (Classification).
3. العنقدة. (Clustering).
4. قواعد الارتباط. (Association Rules).
5. اختيار المعايير. (Select Attributes).

[24] الموقع الرسمي لتطبق ويكا(https://www.cs.waikato.ac.nz/ml/weka/)

- **Data Hero**

عبارة عن برمجيات للبيانات التصورية والبيانات ذاتية الخدمة تسمح لأي مستخدم بالاتصال السريع بالخدمات السحابية دون مساعدة فريق تقنية المعلومات.

تسهّل المخططات المقترحة إنشاء المخططات واكتشاف الإحصاءات داخل بياناتك وتوصيلها بلوحات التحكم.

كما تمكنك من إنشاء مخططات ولوحات بيانات KPI يمكن للجميع فهمها ومشاركتها بسهولة مع الفرق والعملاء.[25]

تم استخدام المنصة للقيام بتصوير البيانات Data Visualize.

[25] الموقع الرسمي لداتا هيرو https://datahero.com/

المبحث الثاني

تصوير البيانات Data Visualize

تمهيد

أضحى مجال علم البيانات من المجالات ذات الطلب العالي عالمياً لأهميتها وتأثيرها في عالم الأعمال والأبحاث، وأحد المجالات التي تتفرع من علم البيانات وتحليل البيانات، مجال التصوير البياني أو تصوير البيانات Data Visualize.

يعرف تصوير البيانات بأنه تقديم البيانات بأسلوب فني جميل الشكل ومنسق اللون وواضح المعالم، بخلاف وسائل التقديم العلمية التي تهتم بالمحتوى أكثر من المظهر.

والغرض من التعبير البصري للبيانات هو توضيحها، كما وتنبع أهمية التصوير البياني من كونه أداة فعالة لتسهيل قراءة التقارير في خضم تراكم البيانات الضخمة، وكذلك يساعد متخذي القرارات برؤية أوضح للبيانات وصورة أشمل بهدف تسهيل اتخاذ القرارات.

تعد عملية تصوير البيانات اول التقنيات التي سنستخدمها في تحليل البيانات الموجودة لدينا من اجل القاء الضوء على أهم النقاط الموجودة لدينا سنحاول الإجابة على عدة تساؤلات.

ونظراً لأننا نتوقع أنماط التسرب (Churn) لمستخدمي الاتصالات، فلنكتشف العلاقة بين الأعمدة المختلفة، أو المتغيرات المختلفة، مع متغير الهدف Churn لمعرفة الأسباب التي تؤدي إلى تسرب الزبائن، ولمعرفة العوامل التي قد تؤثر في قرار الزبون في ترك شركة الاتصالات أو في البقاء فيها، و ما هي الخصائص، الخدمات، طرق الدفع أو الدفعات التي يمكن أن تحول رأي الزبون و قراره للتخلي عن هذه الشركة و ما تقدمه.

وسيتم تقسيم تصوير البيانات على النحو التالي:

1. المتغيرات الفئوية.
   1.1. تصوير المتغيرات الفئوية كل على حدا.
   1.2. تصوير علاقات المتغيرات الفئوية مع بعضها البعض.
   1.3. تصوير علاقات المتغيرات الفئوية مع المتغير الهدف.
2. المتغيرات الرقمية.
   2.1. تصوير المتغيرات الرقمية كل على حدا.
   2.2. تصوير علاقات المتغيرات الرقمية مع بعضها البعض.
   2.3. تصوير علاقات المتغيرات الرقمية مع المتغير الهدف.

1. المتغيرات الفئوية:

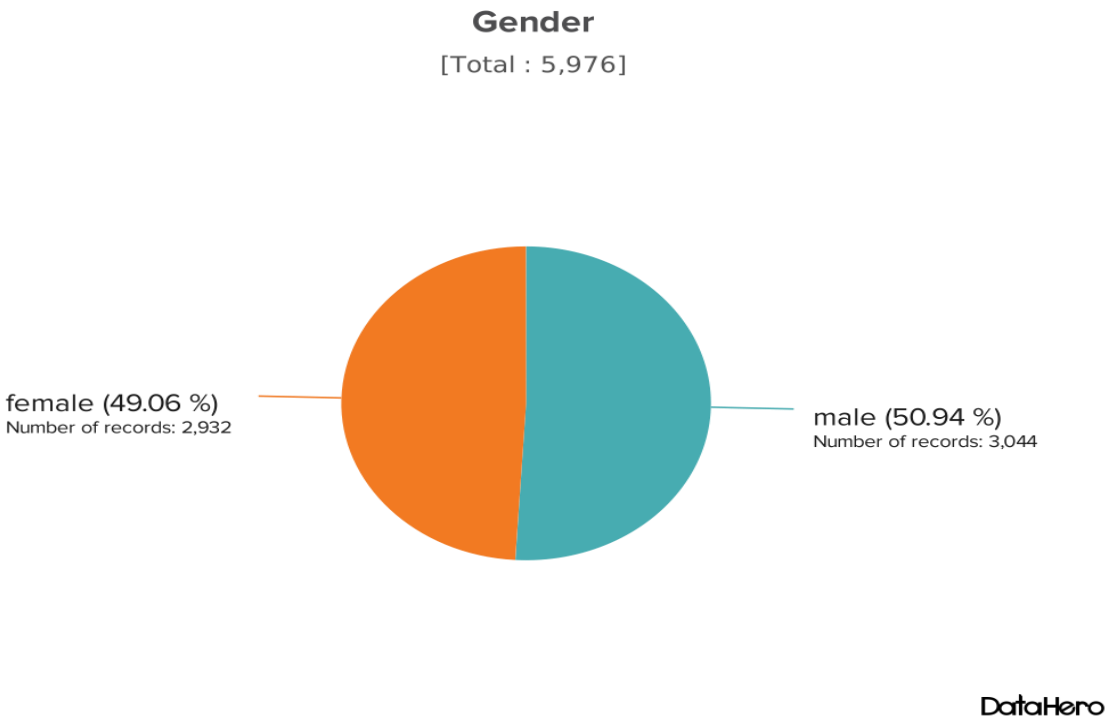
في هذا القسم سنحاول معرفة الاختلافات بين المتغيرات الفئوية وما هي المعلومات الأساسية لكل متغير على حدا، ومن ثم معرفة العلاقات بين المتغيرات الفئوية مع بعضها البعض، وبعدها معرفة علاقتها مع المتغير المستهدف وما الأثر التي تتركه للوصول إلى قرار التخلي عن الخدمة أم لا ولماذا.

وسنحاول أن نتوصل لإجابة عن التساؤلات التالية:

- ما هي أغلب الخدمات المستخدمة لدى العملاء؟
- ماهي طريقة الدفع المرجحة لدى العملاء، وما هي نوعية الفواتير المنتشرة للقيام بالدفع؟
- ماهي نوعية العقود المفضلة لدى العملاء عند الاشتراك بشركة الاتصالات؟
- ما هي المتغيرات الفئوية المرتبطة ببعضها؟ وهل يمكن التخلي عن أحد من هذه المتغيرات بناء على اختبار الارتباط؟
- ما هي خصائص العملاء الأكثر عرضة للتسرب، والتخلي عن الشركة؟
- ما هي الخدمات التي قد يشكل غيابها مشكلة تسرب لدى الزبائن أو العملاء؟
- ما هي نوعية العقود التي يمكن أن ينبؤنا توجه الزبون إليها في احتمالية تسربه؟
- ما طرق الدفع والفوترة التي يمكن أن يؤدي استعمال الزبائن لها لتسربهم؟

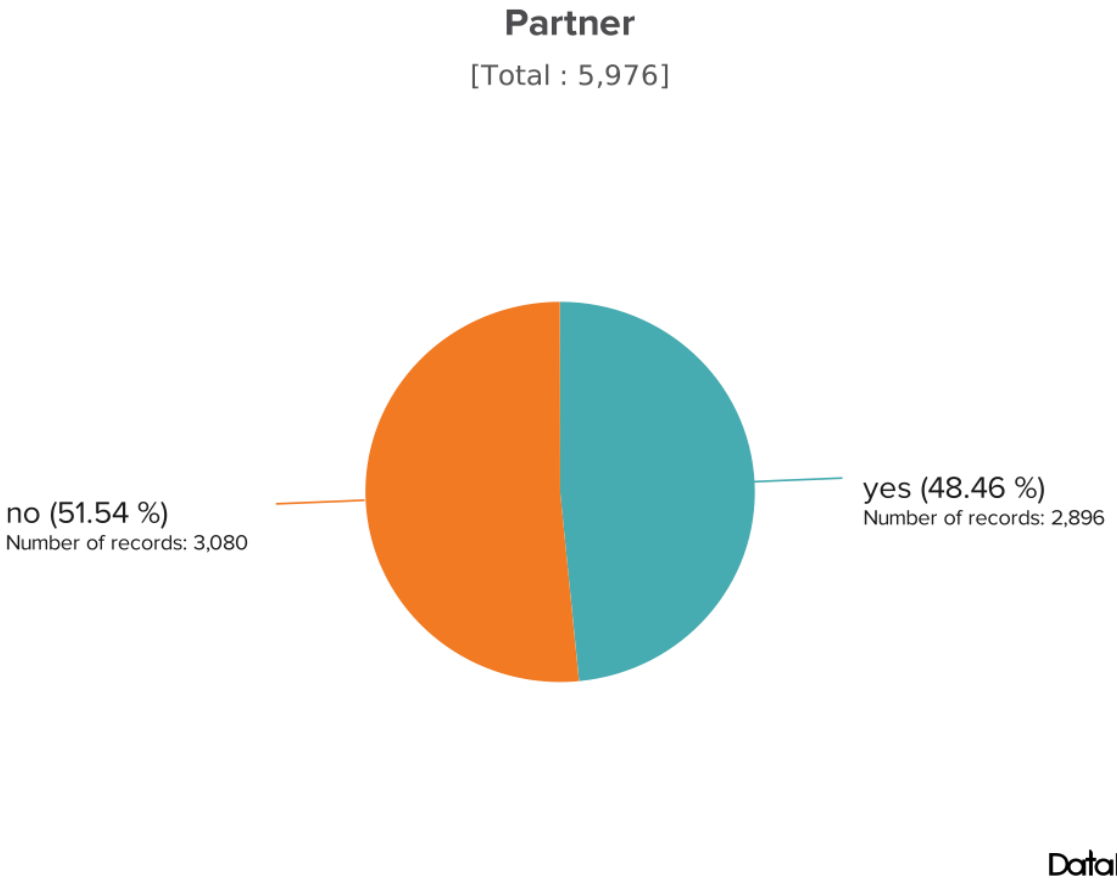
1.1 تصوير المتغيرات الفئوية كل على حدا

فيما يتعلق بخصائص العميل:



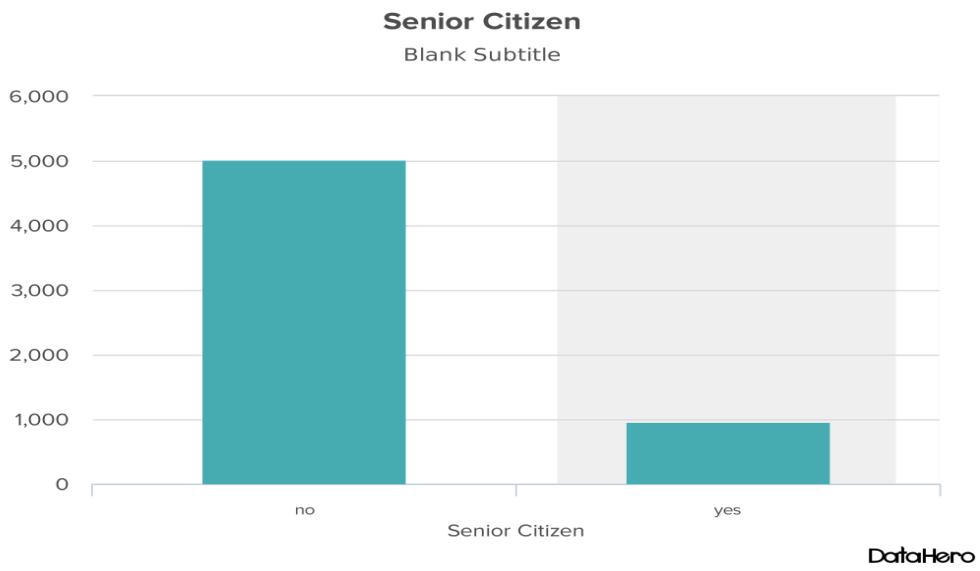
الشكل 1 . Gender

نلاحظ أن توزيـع الجنس كان بالتساوي تقريبا، بنسبة 50.94% للذكور، و 49.06% للإناث.



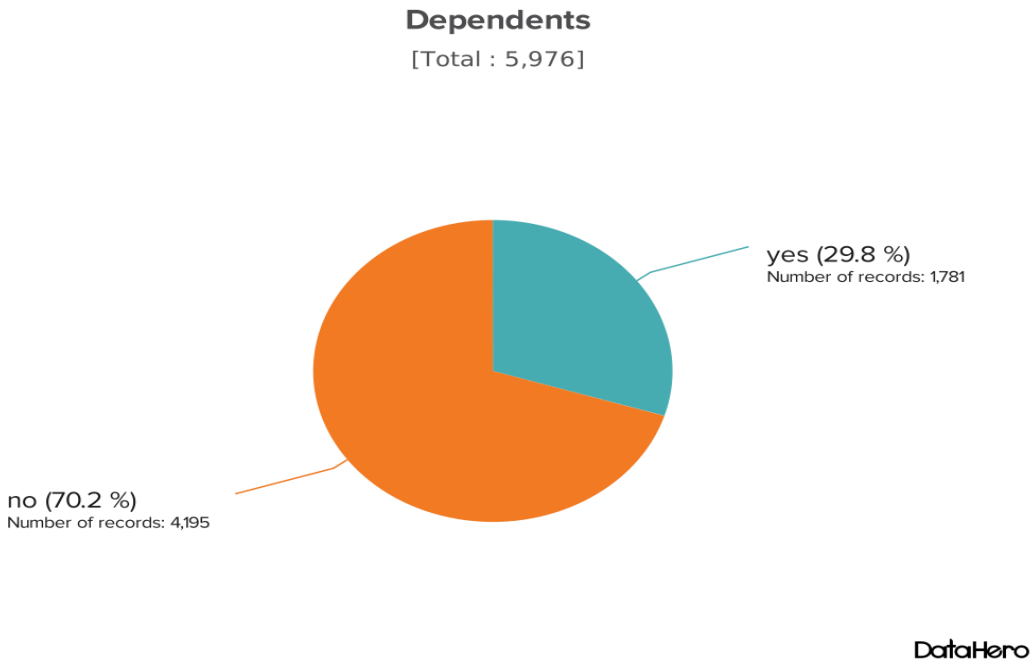
الشكل 2 . Partner

كما و نلاحظ أن توزع العملاء ذو الشركاء (اي المتزوجين) أيضاً كان تقريباً بالتساوي، بنسبة 48.46% للمتزوجين، و 51.54% لغير المتزوجين .



الشكل 3 . Senior Citizen

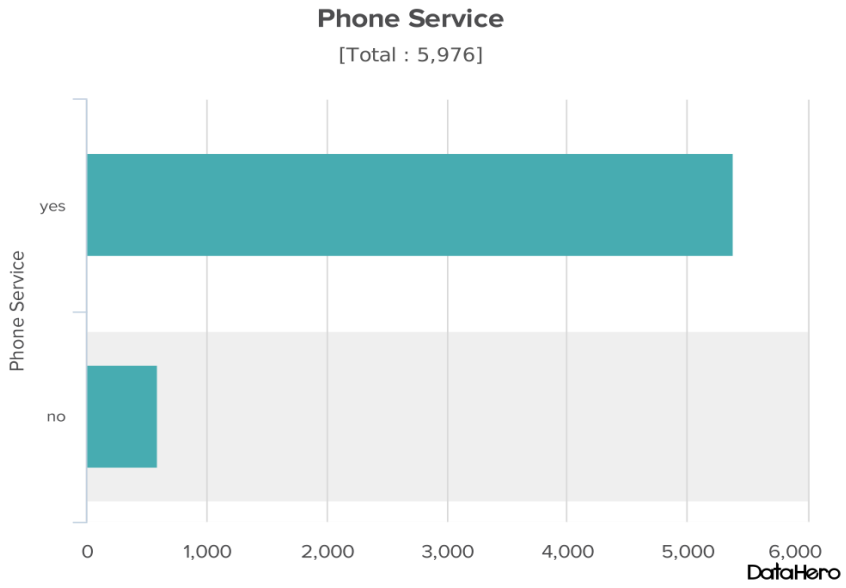
نلاحظ من هذا التصوير، أن معظم العملاء ليسوا مواطنين كبار السن (متقاعدين).



الشكل 4. Dependents

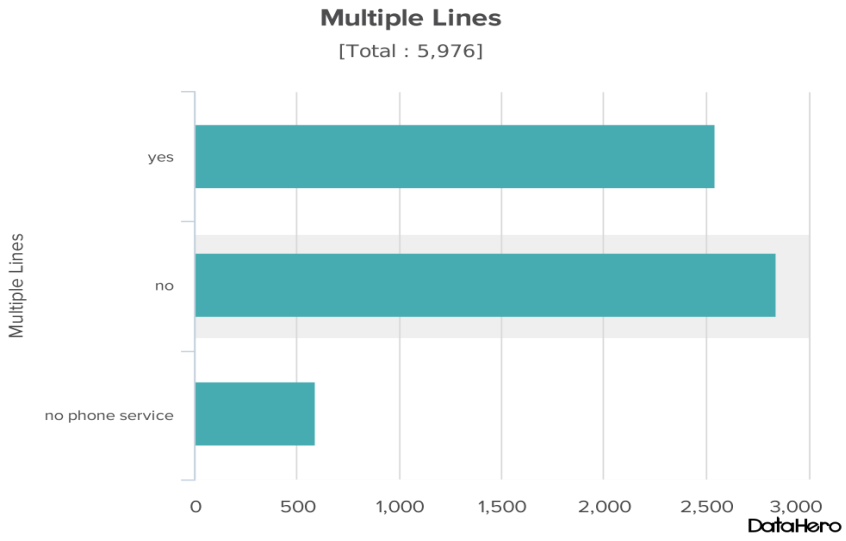
كما نلاحظ أيضاً أن 70% أي الحد الأقصى من العملاء هم لا معيل مادي لهم، أو قد تكون بمعنى أنه ليس لديهم عمل .

فيما يتعلق بالخدمات:



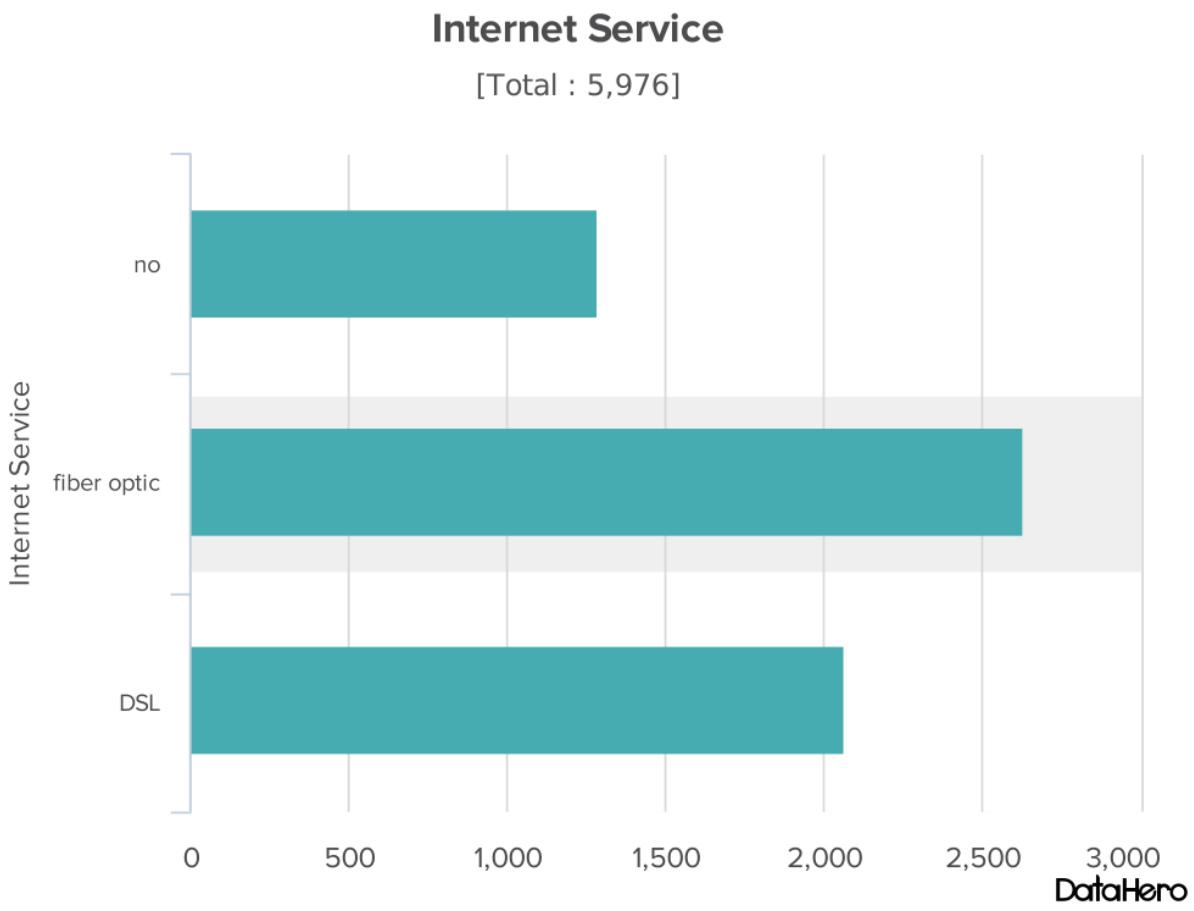
الشكل 5. Phone Service

نلاحظ من التصوير التالي، أن الغالبية العظمى من العملاء لديهم خدمة الهاتف .



الشكل 6. Multiple Lines

كما و نلاحظ أيضاً، أن نصف العملاء التي لديها خدمة الهاتف تملك خطوط متعددة، كما نلاحظ أيضاً أن معظم العملاء ليس لديهم خطوط متعددة.

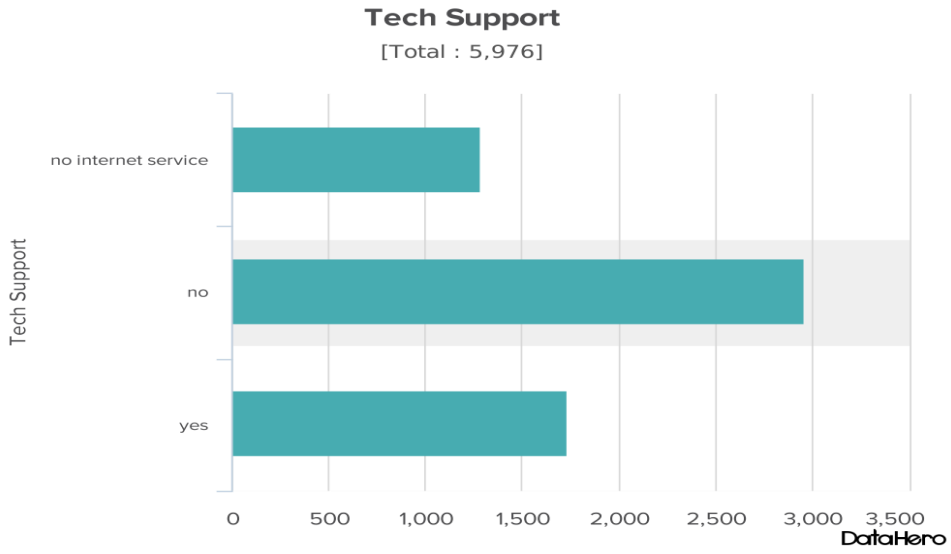


الشكل 7. Internet Service

يبين هذا التصوير أن، 22% من العملاء لا يملكون خدمة إنترنت، و الباقي الغالب يملك هذه الخدمة، كما نلاحظ أن معظم العملاء يختارون الألياف البصرية Fiber Optic.

فيما يتعلق بالأمن و الدعم:

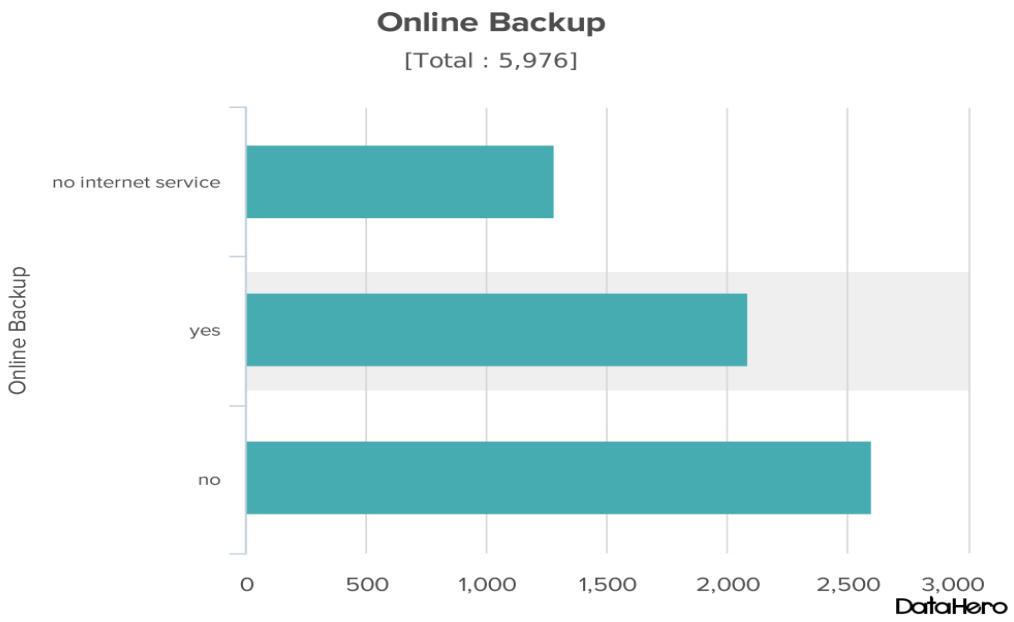
الشكل 8. Online Security



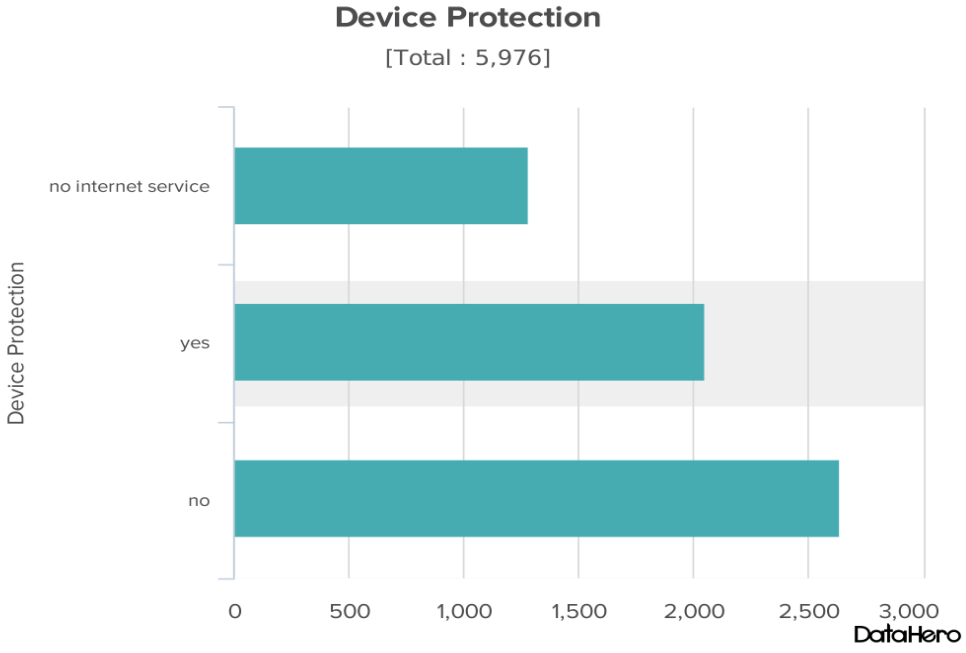
الشكل 9. Tech Support

نلاحظ هنا، أنه بالنسبة للعملاء الذين يستخدمون خدمة الإنترنت، فإن هناك نسبة تبلغ حوالي اثنين إلى واحد من العملاء الذين ليس لديهم أمان عبر الإنترنت، كما هو الحال بالنسبة للدعم الفني.

أي أن معظم العملاء لا يتمتعون بأمان عبر الإنترنت، كما أن معظمهم ليس لديهم دعم فني .



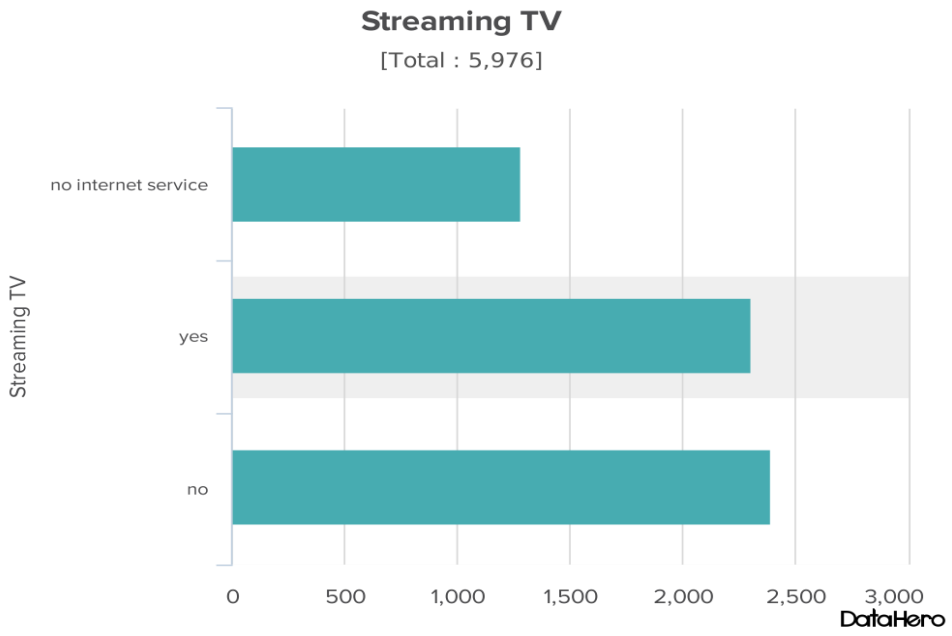
الشكل 10. Online Backup



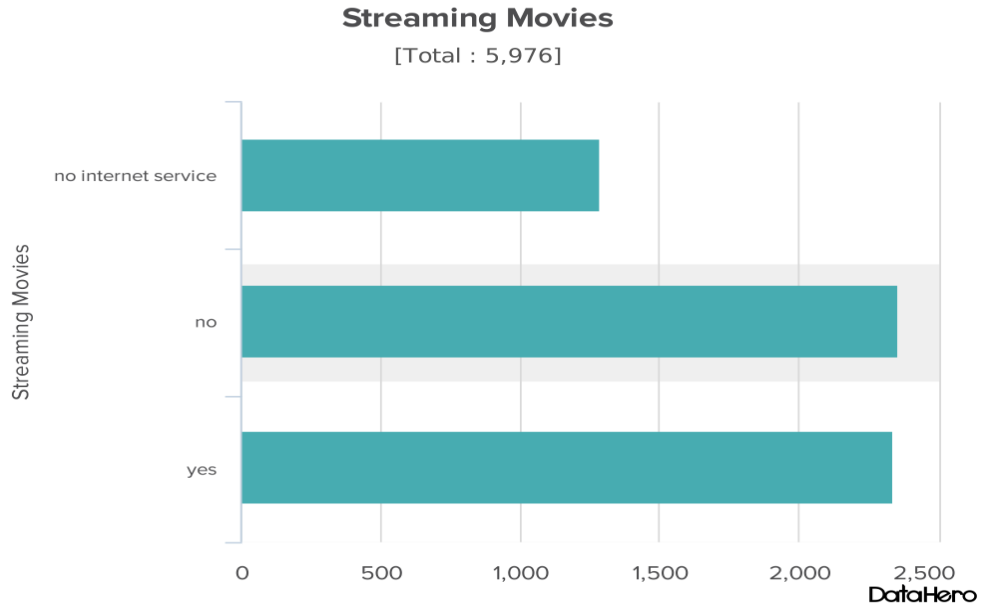
الشكل 11. Device Protection

و نلاحظ هنا، أن هناك العديد من العملاء الذين ليس لديهم نسخ احتياطي عبر الإنترنت، و حماية الجهاز ب35% للذين لديهم نسخ و حماية، و 44% للذين ليس لديهم نسخ و حماية .

فيما يتعلق بقدرات التدفق:



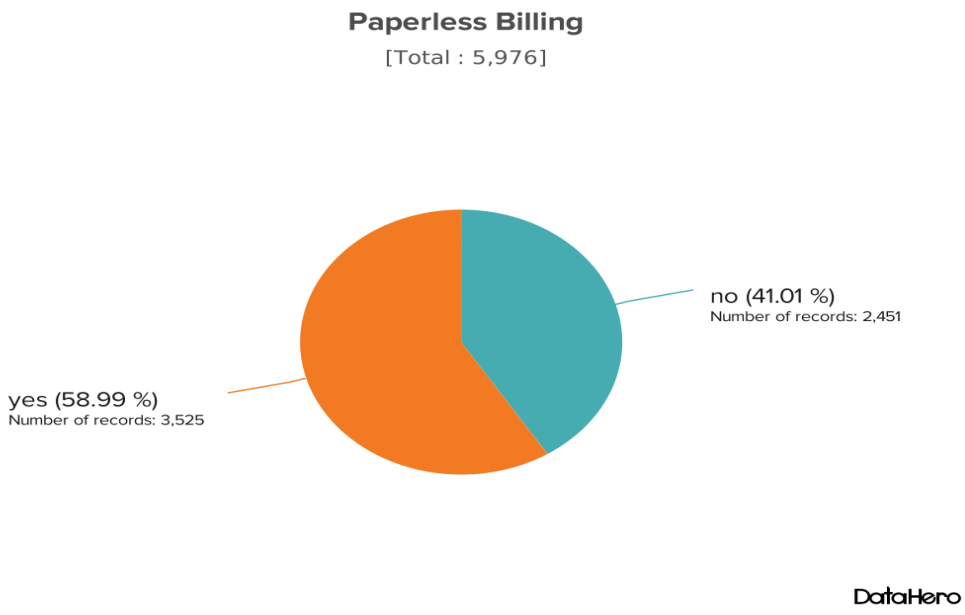
الشكل 12.Streaming TV



الشكل 13.Streaming Movies

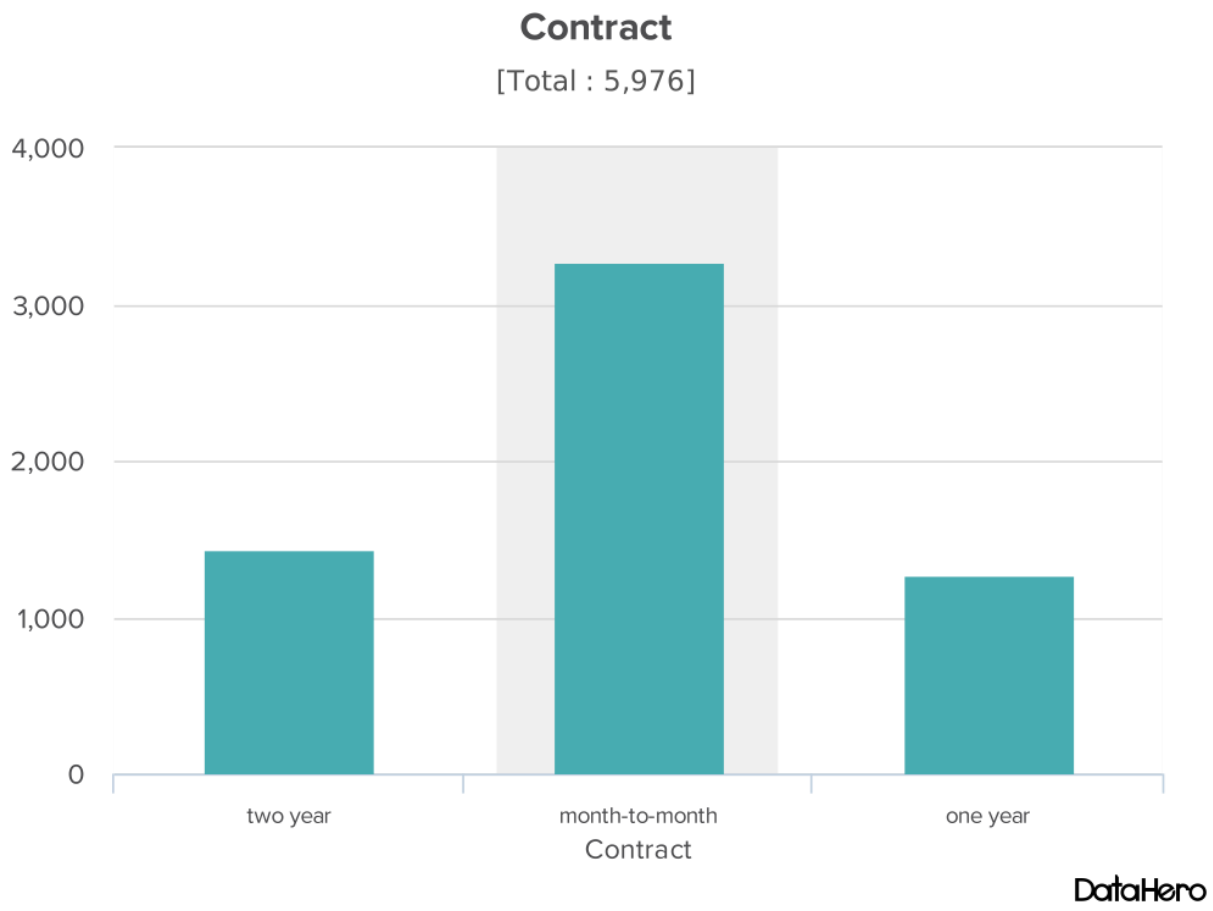
كما و من هذا التصوير ،يبدو أن العملاء متشابهين جداً بنسب متساوية للذين يستخدمون خدمة بث الأفلام، و خدمة البث التلفزيوني.

فيما يتعلق بخصوص المدفوعات:



الشكل 14. Payment Method

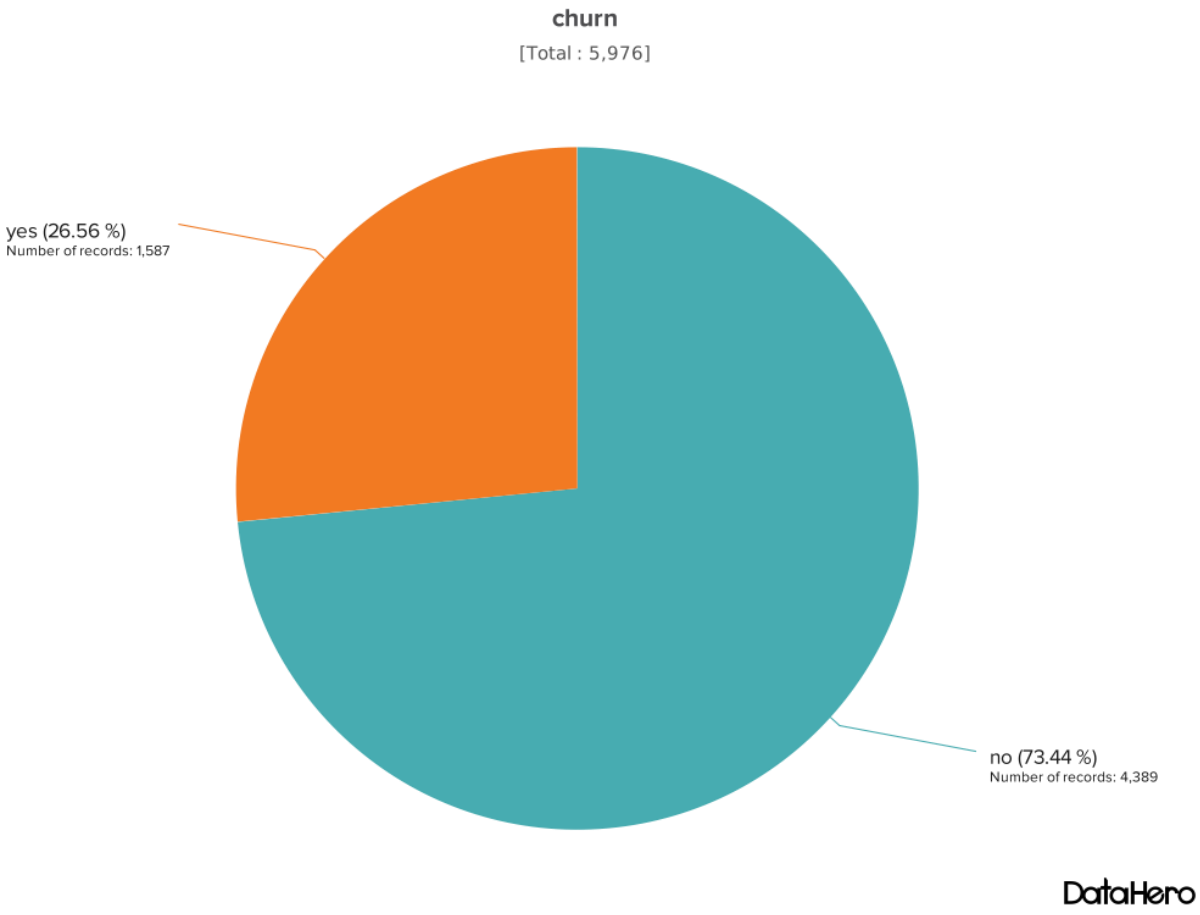
ومن التصوير التالي نجد أن، 60% من العملاء يختارون الفواتير غير الورقية، وكانت طريقة الدفع الأكثر شيوعاً هي، الشيك الإلكتروني.



الشكل Contract.15

كما ونجد أن، ما يقرب من نصف العملاء يشتركون بعقد شهري أو كما يشاع عنه عقد شامل (End-to-End Contract)، والنصف الآخر مقسم بين عقد عام واحد أو عامين.

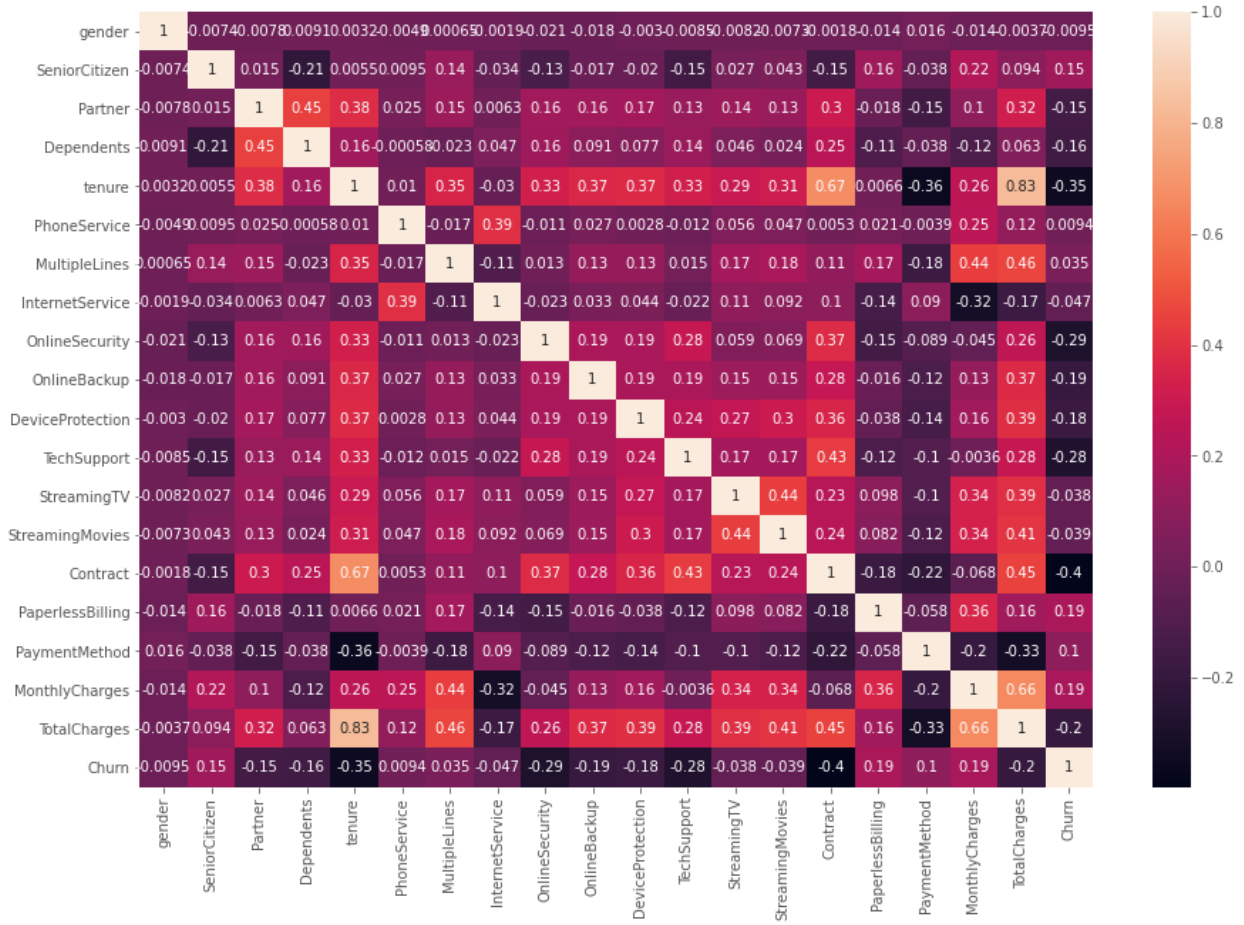
فيما يتعلق بخصوص المتغير المستهدف Target Variable:



الشكل Churn.16

هناك معدل تسرب بالزبائن قدره 27%، و هذا يبدو مرتفعاً جداً بالنسبة لخدمة الاتصالات.

1.2تصوير المتغيرات الفئوية مع بعضها البعض



الشكل 1.17 correlation analysis

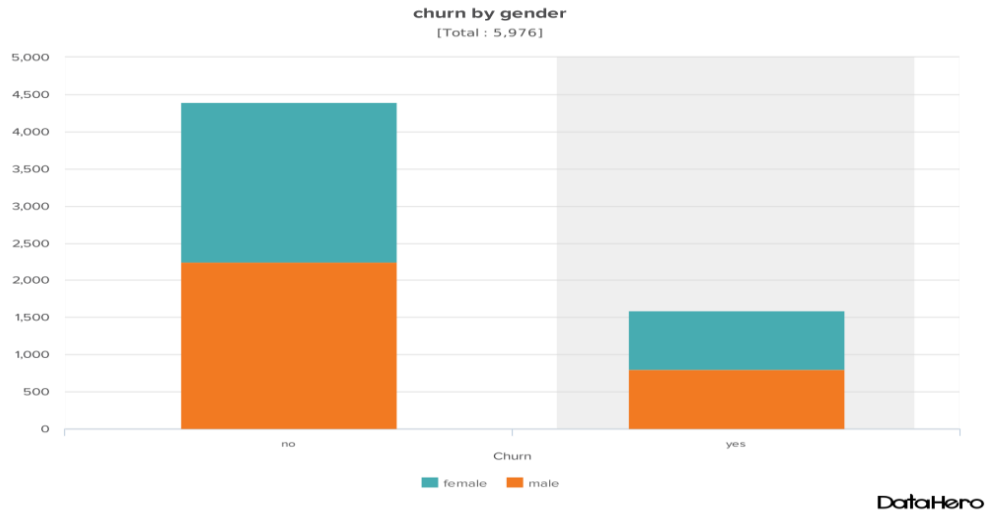
يتمثل تصوير هذه العلاقات ب اختبار الارتباط Correlation، و من خلاله نلاحظ علاقة قوية للعقد Contract مع مدة العقد Tenure .

كما نلاحظ أيضاً، وكما هو متوقع، ليس للجنس أي ارتباط مع أي من المتغيرات الفئوية الأخرى .

كما و يبدو أن معظم المتغيرات الفئوية الأخرى لديها نوع من العلاقة مع المتغيرات الفئوية الأخرى، على الرغم من ضعف هذه العلاقات.

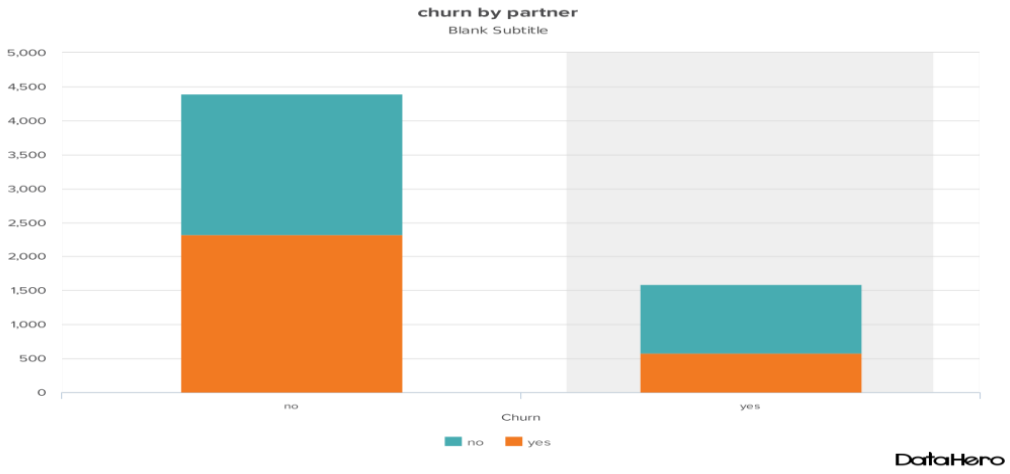
1.3تصوير علاقات المتغيرات الفئوية مع المتغير الهدف

فيما يتعلق بخصائص العميل:



الشكل 18.Churn By Gender

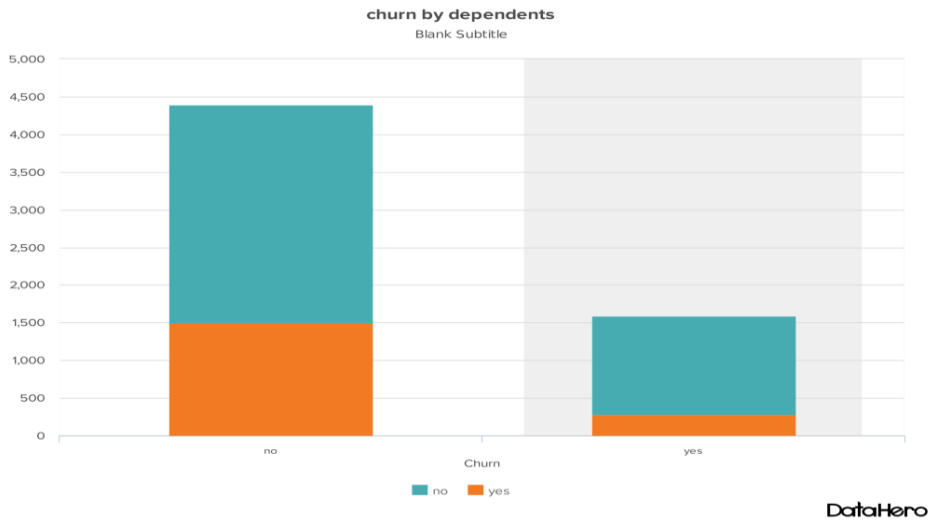
نلاحظ أن نسبة التسرب تتساوى عند الذكور 26% وعند الإناث 27%.



الشكل 19.Churn By Partner

ونلاحظ أن نسبة التسرب للذين لا يملكون شركاء تبلغ 33%، بينما الذين يملكون شركاء قد بلغ معدل التسرب لديهم 20%.

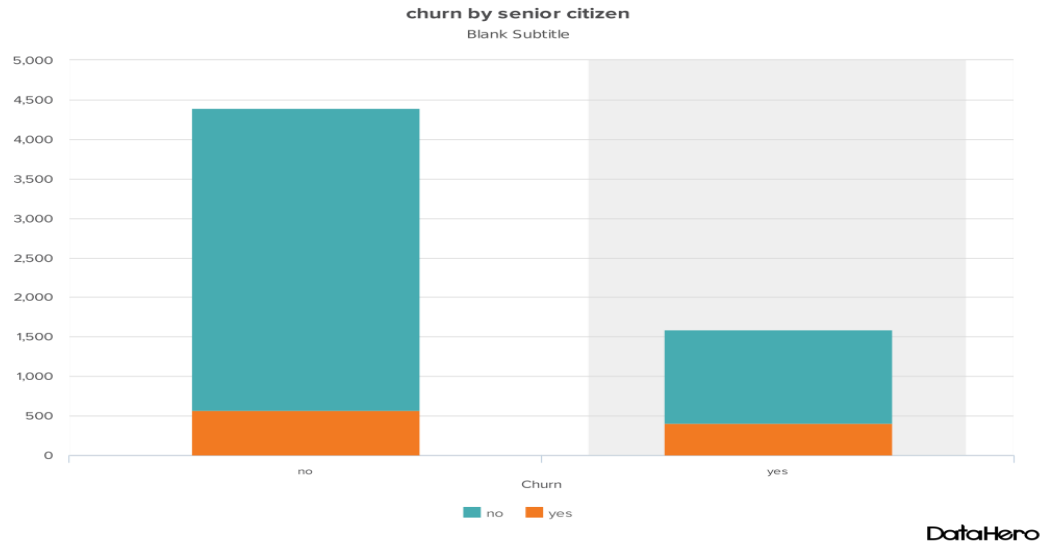
العملاء الذين لديهم شركاء أكثر عرضة للتسرب، إذ لا يشارك أي شخص أخر في اتخاذ القرار.



الشكل 20.Churn By Dependents

بلغت نسبة تسرب المعالون 16%، بينما كانت نسبة تسرب الغير معالون 31% .

العملاء الذين لا يوجد لديهم معيل هم أكثر عرضة للتسرب .

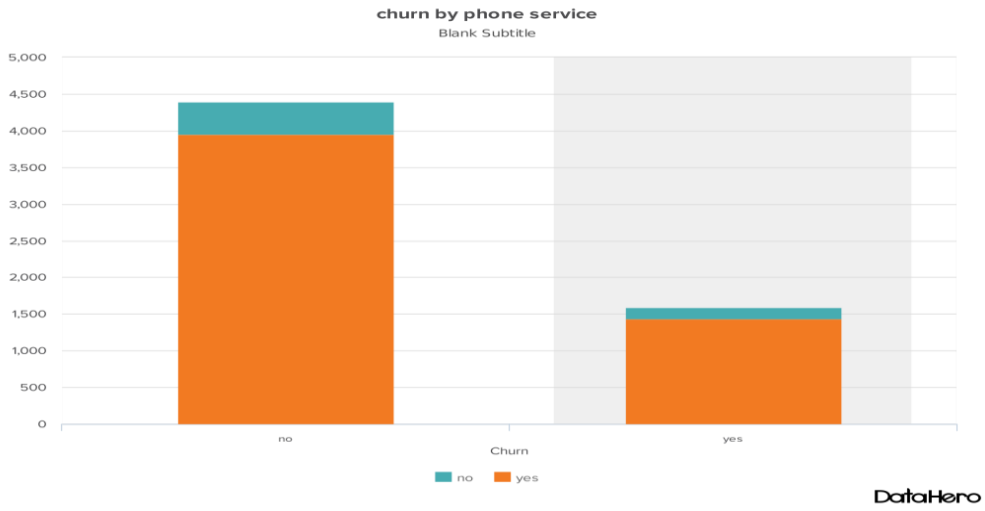


الشكل 21.Churn By Senior Citizen

بلغت نسبة تسرب الغير متقاعدين 24%، في حين أن نسبة تسرب كبيري السن المتقاعدين كانت 42%.

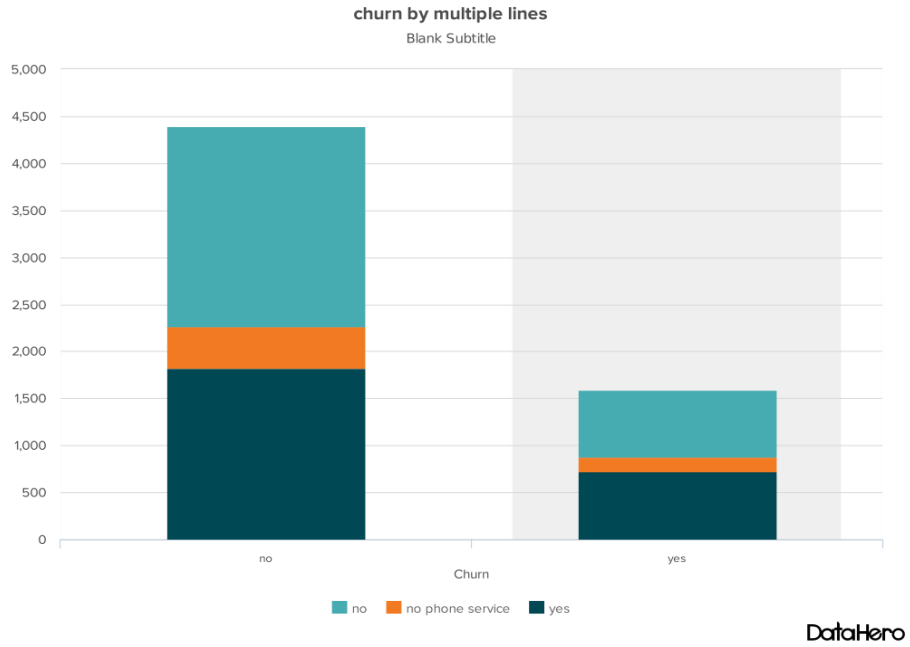
نجد أن المواطنون الأكبر سناً هي أكثر عرضة للتسرب .

فيما يتعلق بالخدمات:



الشكل 22.Churn By Phone Service

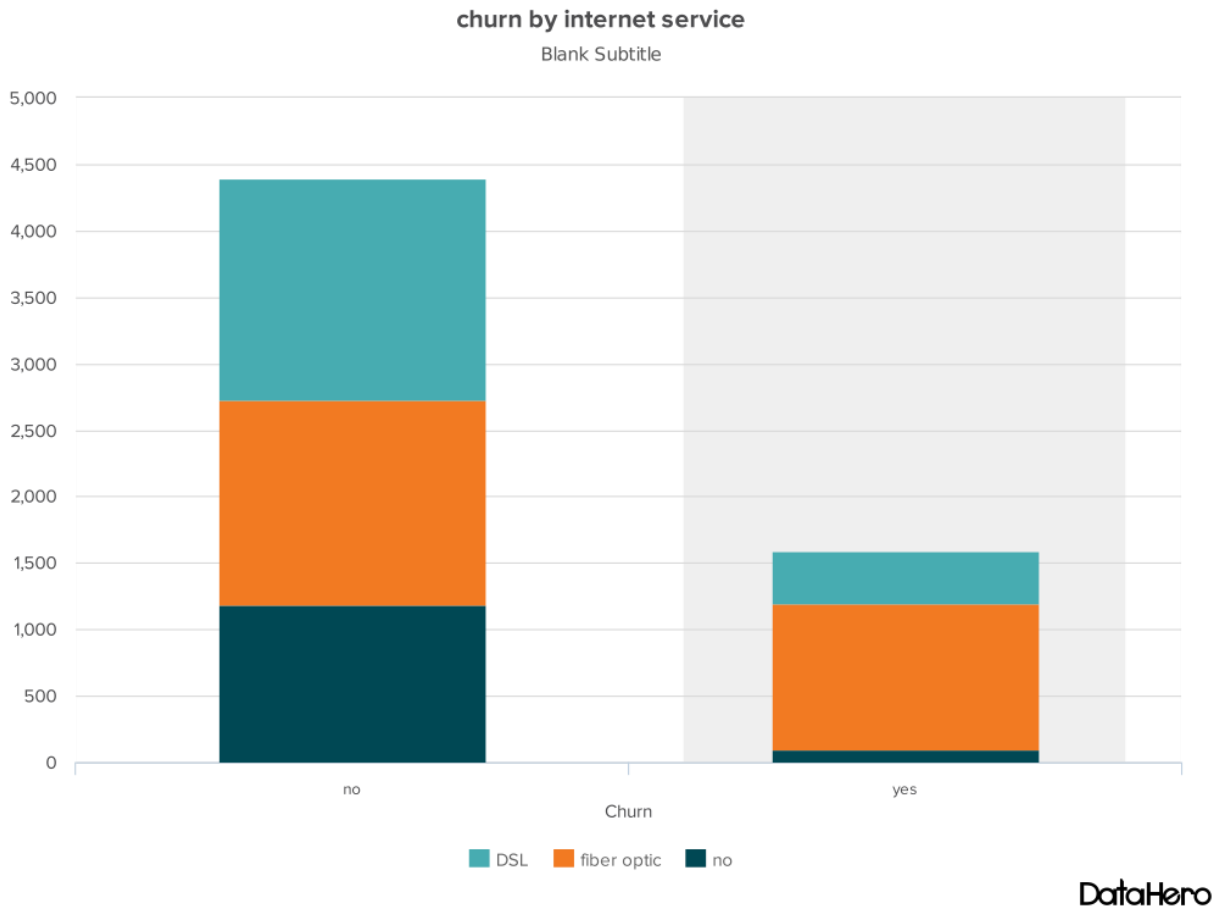
بلغت نسبة تسرب الذين يستخدمون خدمة الهاتف 27%، و نسبة تسرب الذين لا يستخدمون خدمة الهاتف 25%، حيث كان بدى كل من العملاء الذين يستخدمون أو الذين لا يستخدمون خدمة الهاتف معدل متطابق في التسرب.



الشكل 23.Churn By Multiple Lines

بلغت نسبة تسرب الذين لا يملكون خطوط متعددة 25%، و الذين يملكون 28%، و الذين لا يملكون خدمة الهاتف 25% .

بمعدلات متطابقة تقريبا للفئات الثلاثة للخطوط المتعددة.

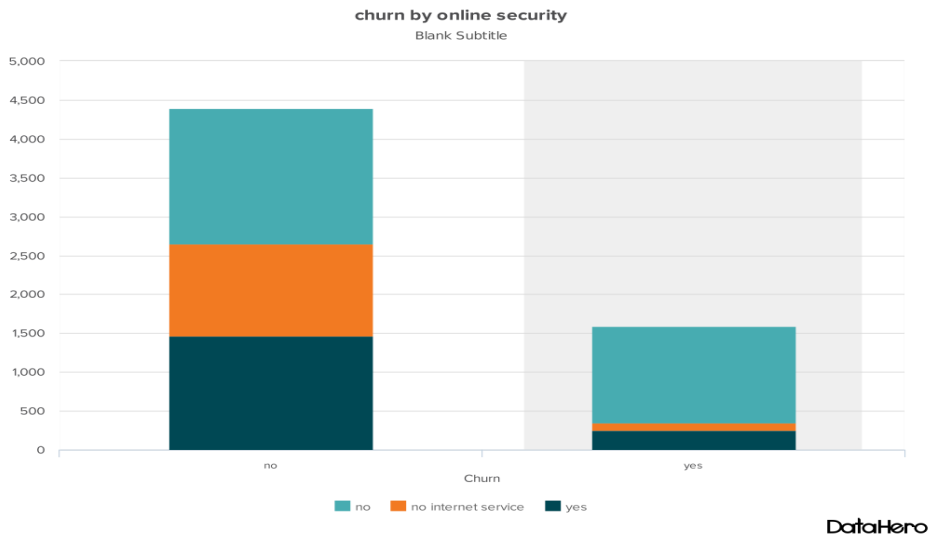


الشكل 24. Churn By Internet Service

بلغت نسبة تسرب الذين يستعملون الFiber Optic 42%، و مستعملي الDSL 19%، و الذين لا يملكون خدمة الإنترنت 8%.

و نجد أن العملاء الذين يستخدمون الألياف الضوئية Fiber Optic هم الأكثر عرضة للتسرب .

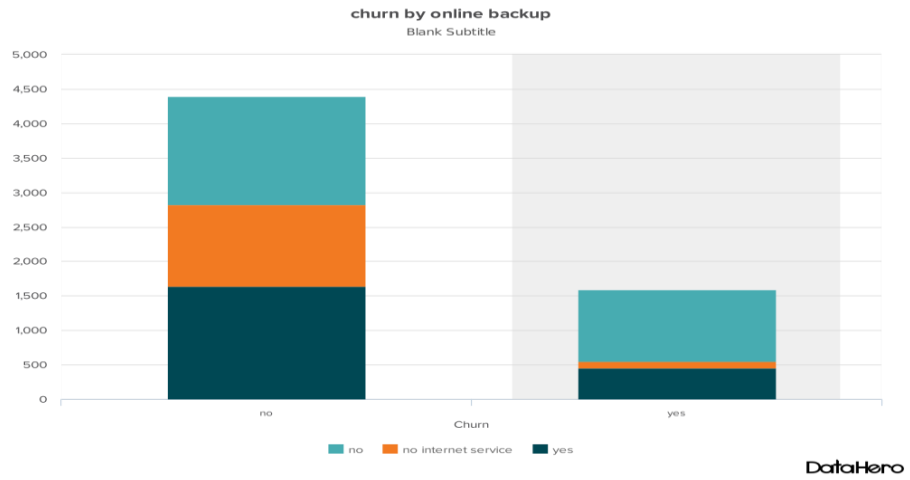
فيما يتعلق بالأمن والدعم:



الشكل 25.Churn By Online Secuurity

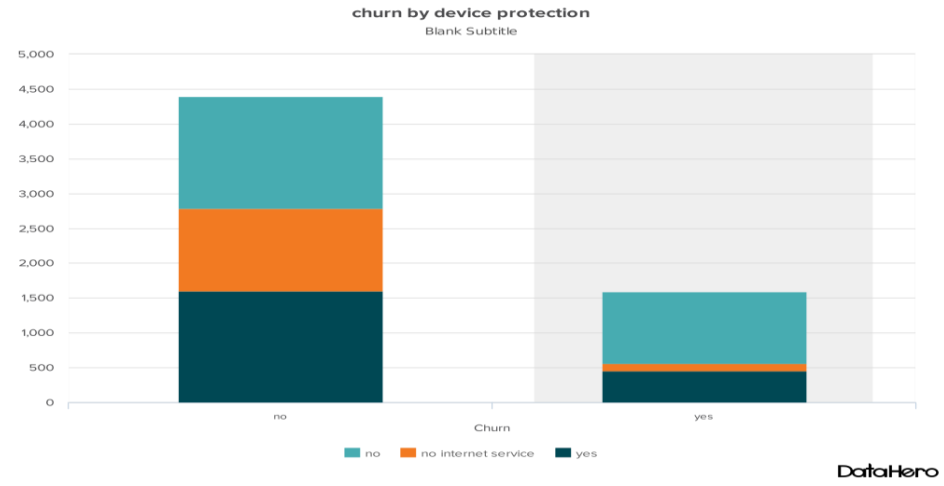
بلغت نسبة تسرب الذين ليس لديهم أمان 42%، والذين لديهم أمان 14%، والذين لا يملكون خدمة الأنترنت 8%.

نجد أن العملاء الذين ليس لديهم أمان، هم الأكثر عرضة للتسرب.



الشكل 26.Churn By Online Backup

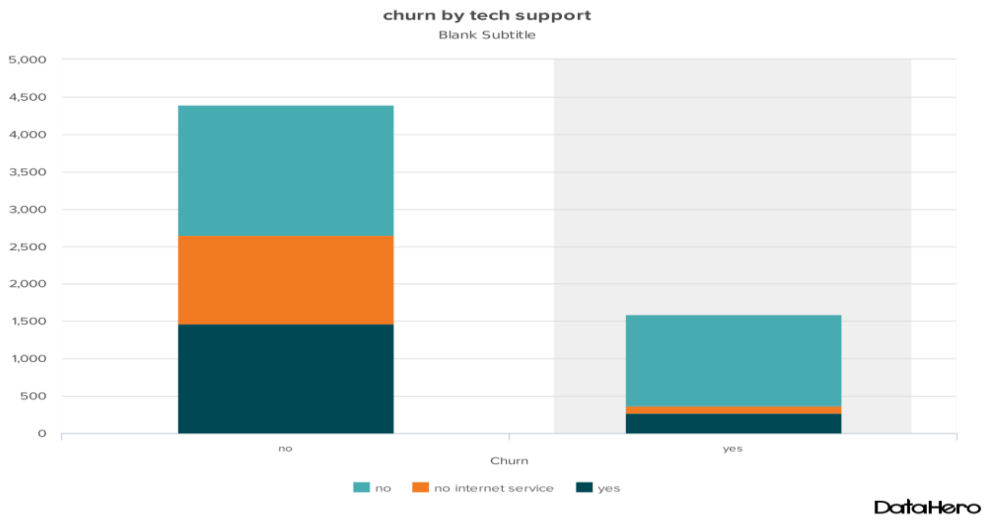
بلغت نسبة تسرب الذين لا يملكون النسخ الاحتياطي 40%، الذين يملكون 22%، الذين لا يملكون خدمة الأنترنت 8%.



الشكل 27.Churn By Device Protection

بلغت نسبة تسرب الذين لا يملكون حماية للجهاز 39%، الذين يملكون 22%، الذين لا يملكون خدمة الأنترنت 8%.

نجد أن معظم العملاء الذين ليس لديهم حماية للجهاز بتسربون.

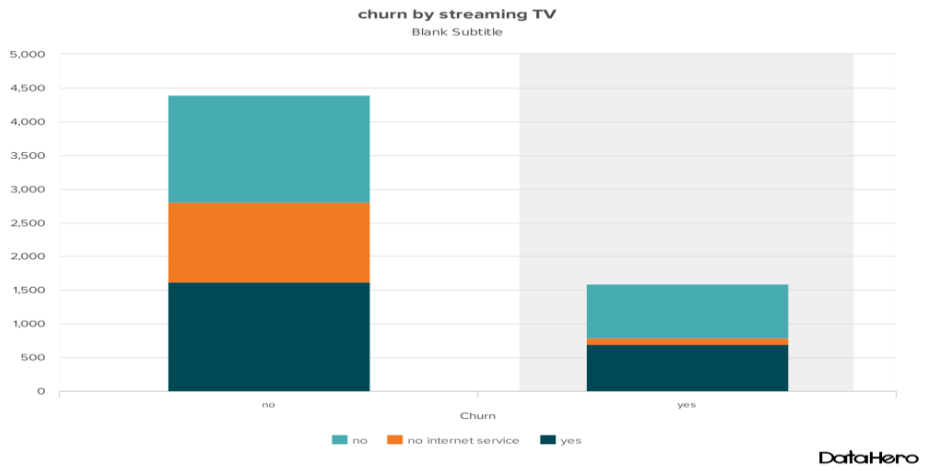


الشكل 28.Churn By Tech Support

بلغت نسبة تسرب الذين لا يملكون دعم فني 41%، الذين يملكون 15%، الذين لا يملكون خدمة الأنترنت 8%.

نجد أن معظم العملاء الذين ليس لديهم دعم فني بتسربون.

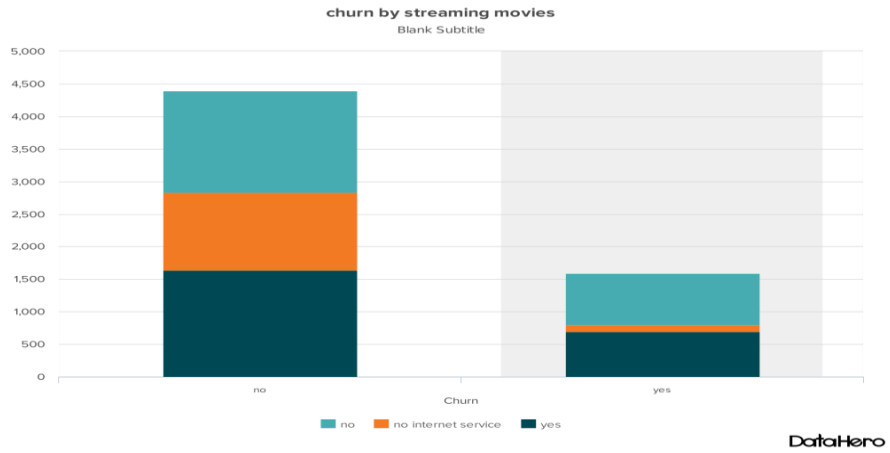
فيما يتعلق بقدرات التدفق:



الشكل 29.Churn By Streaming TV

بلغت نسبة تسرب الذين لا يملكون بث تلفزيوني 33%، الذين يملكون 30%، الذين لا يملكون خدمة الأنترنت 8%.

نجد أن معدل تسرب العملاء الذين يختارون أو لا يختارون البث التلفزيوني هي متساوياً تقريباً.

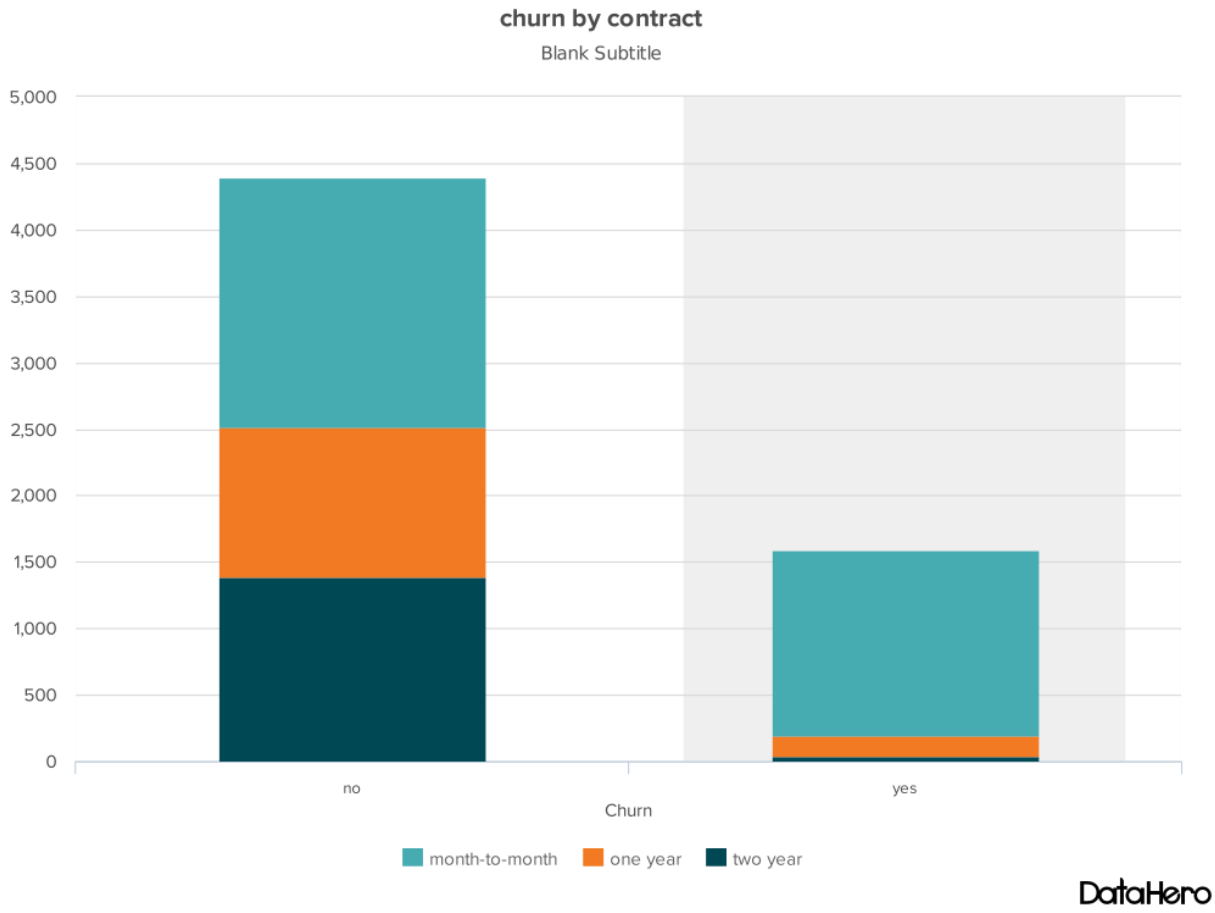


الشكل 30.Churn By Streaming Movies

بلغت نسبة تسرب الذين لا يملكون بث الأفلام 34%، الذين يملكون 30%، الذين لا يملكون خدمة الأنترنت 8%.

نجد أن معدل تسرب العملاء الذين يختارون أو لا يختارون بث الأفلام هي متساوياً تقريباً.

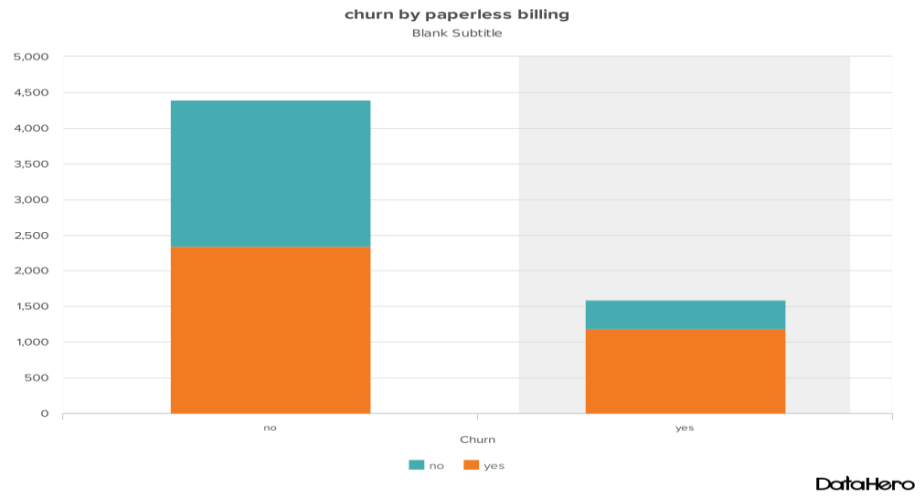
فيما يتعلق بخصوص المدفوعات:



الشكل 31.Churn By Contract

بلغت نسبة تسرب الذين يشتركون بعقد شهري 43%، الذين يشتركون بعقد سنة واحدة 12%، الذين يشتركون بعقد سنتين 3%.

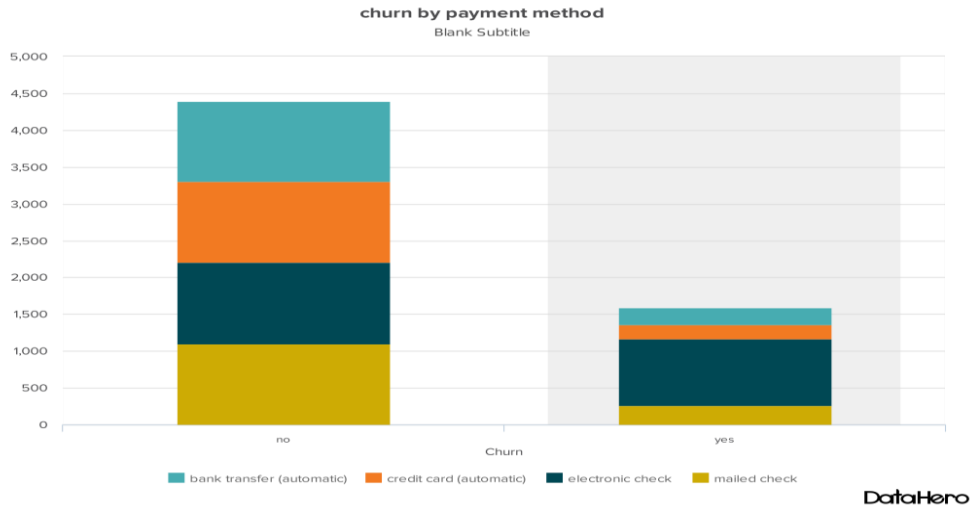
نجد أن العملاء الذين يتجهون إلى عقود شهرية هم الأكثر عرضة للتسرب، إذ لا يوجد التزام قوي.



الشكل 32.Churn By Paperless Billing

بلغت نسبة تسرب الذين يفضلون الفواتير غير الورقية 33%، الذين لا يفضلون الفواتير غير الورقية 17%.

نجد أن العملاء الذين يتجهون إلى عمليات الفوترة اللاورقية هم الأكثر عرضة للتسرب، إذ أن كل شيئ يمكن القيام به على الإنترنت أو عبر الهاتف.

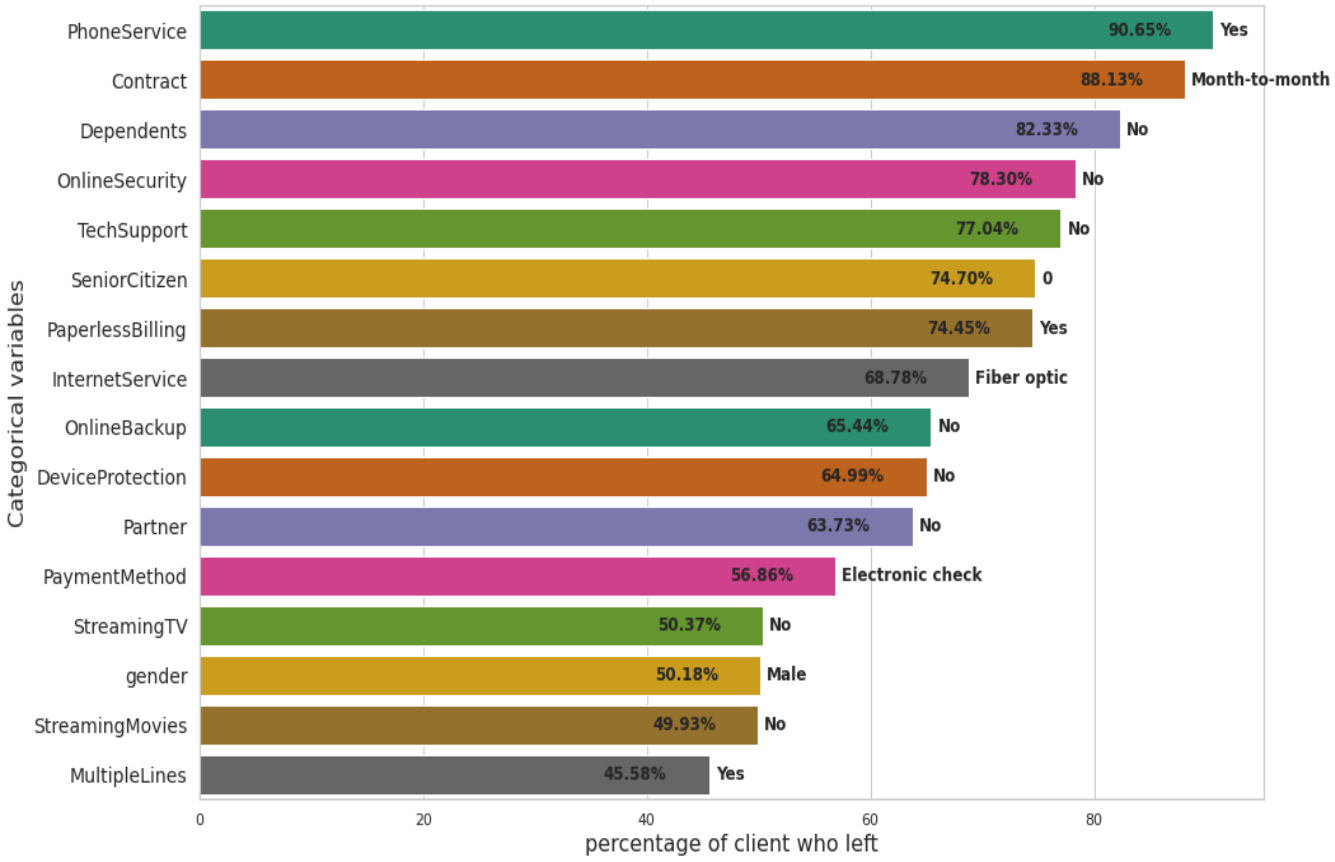


الشكل 33.Churn By Payment Method

بلغت نسبة تسرب الذين يدفعون عن طريق الشيك الالكتروني 45%، الذين يدفعون عن طريق الشيك البريدي 19%، الذين يدفعون عن طريق التحويل المصرفي 17%، والذين يدفعون عن طريق بطاقة الائتمان 15%.

نجد أن العملاء الذين يختاون إجراء الشيكات الالكترونية هم الأكثر عرضة للتسرب.

وفي النهاية تكون النسب المئوية للمتسربين، حسب جميع المتغيرات الفئوية، هي ما يلي:



الشكل 34.Percentage of client who left

2. المتغيرات الرقمية:

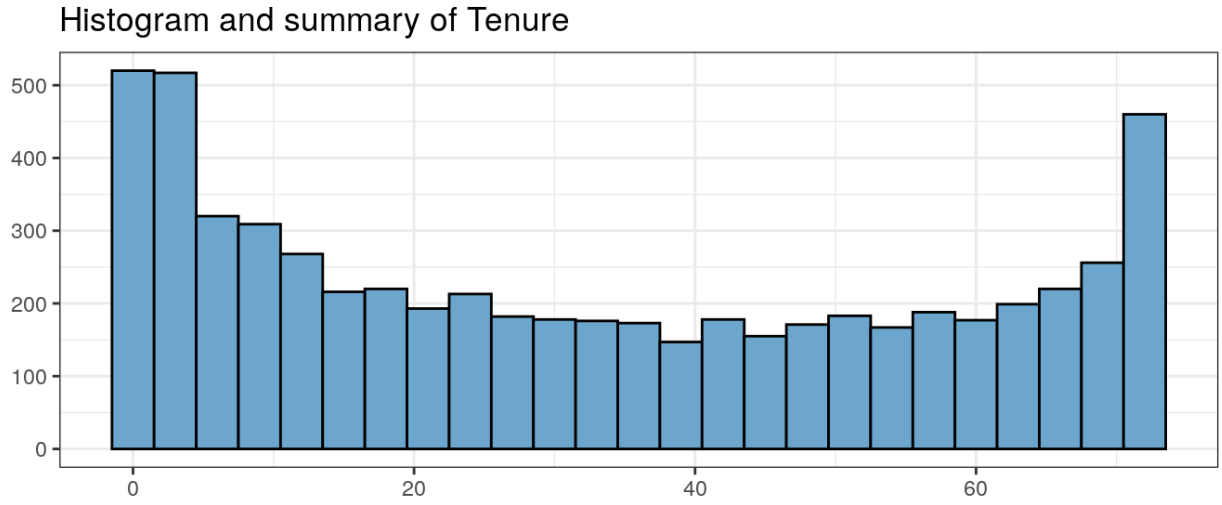
في هذا القسم سنحاول معرفة الاختلافات بين المتغيرات الرقمية وما هي المعلومات الأساسية التي تنتج عن كل متغير على حدا، ومن ثم معرفة العلاقات بين المتغيرات الرقمية مع بعضها البعض، وبعدها معرفة علاقتها مع المتغير المستهدف وما الأثر التي تتركه للوصول إلى قرار التخلي عن الخدمة أم لا ولماذا.

وسنحاول أن نتوصل لإجابة عن التساؤلات التالية:

- ما هي المدة الأرجح لبقاء الزبون في الشركة قبل الرحيل عنها؟
- ماهي قيمة الفواتير الشهرية الأكثر دفعاً من قبل العملاء، والتي تدل على كمية الخدمات التي يشترك بها العميل؟
- ماهي قيمة إجمالي الرسوم الأكثر دفعاً من قبل العملاء؟
- ما هي المتغيرات الرقمية المرتبطة ببعضها؟ وهل يمكن التخلي عن أحد من هذه المتغيرات بناء على اختبار الارتباط؟
- كم يبلغ متوسط مدة وجود العميل في الشركة الذي يمكن أن يتسرب عنده هذا العميل؟
- كم يبلغ متوسط الدفعات الشهرية الذي يمكن أن يتسرب عنده العميل؟
- كم يبلغ متوسط الرسوم الكلية الذي يمكن أن يتسرب عنده العميل؟

2.1تصوير المتغيرات الرقمية كل على حدا

فيما يتعلق بالمتغير Tenure (مدة وجود العميل في الشركة):



Min. : 0.0	1st Qu.: 9.0	Median :29.0	Mean :32.5	3rd Qu.:56.0	Max. :72.0

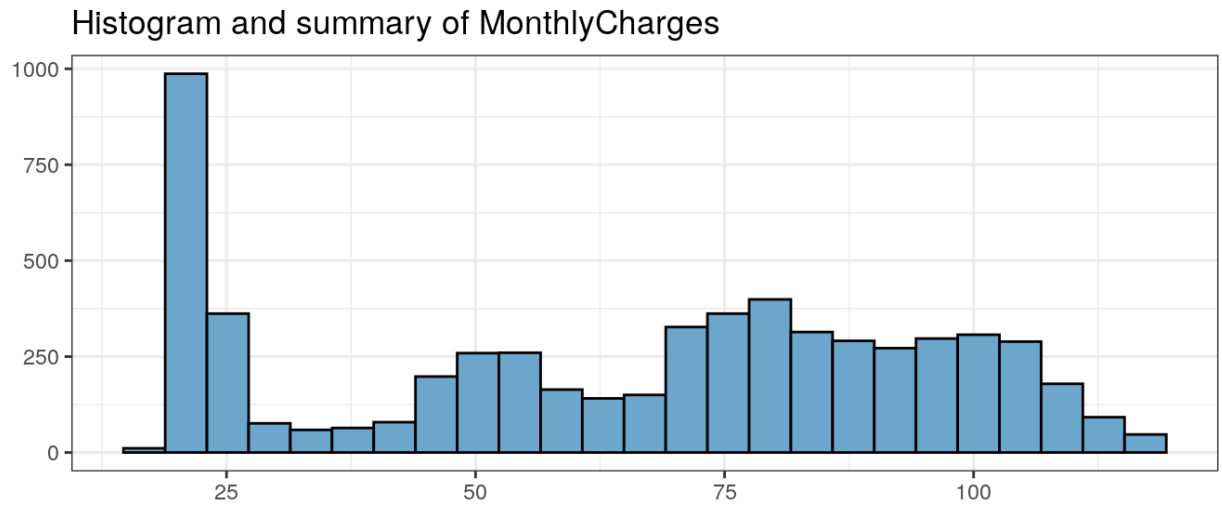
الشكل 35.Tenure

إن متوسط المدة هو 29 شهراً.

و نلاحظ أن التوزيع ثنائي النسق، أي أنه كان هنالك تركيز كبير على العملاء ذو الفترات القصيرة و ذلك أمر منطقي نظراً لمحاولة الإبقاء عليهم لمدة أطول.

كما أنع هناك تركيز كبير أيضاً على العملاء ذو الفترات الطويلة أيضاً، و لكنه أمر غير منطقي، و لكن ربما هؤلاء العملاء قد انضموا بعد أن حصولوا على عرض سخي مما جعل الشركة تستمر بالتركيز عليهم.

فيما يتعلق بالمتغير Monthly Charge (الرسوم الشهرية):

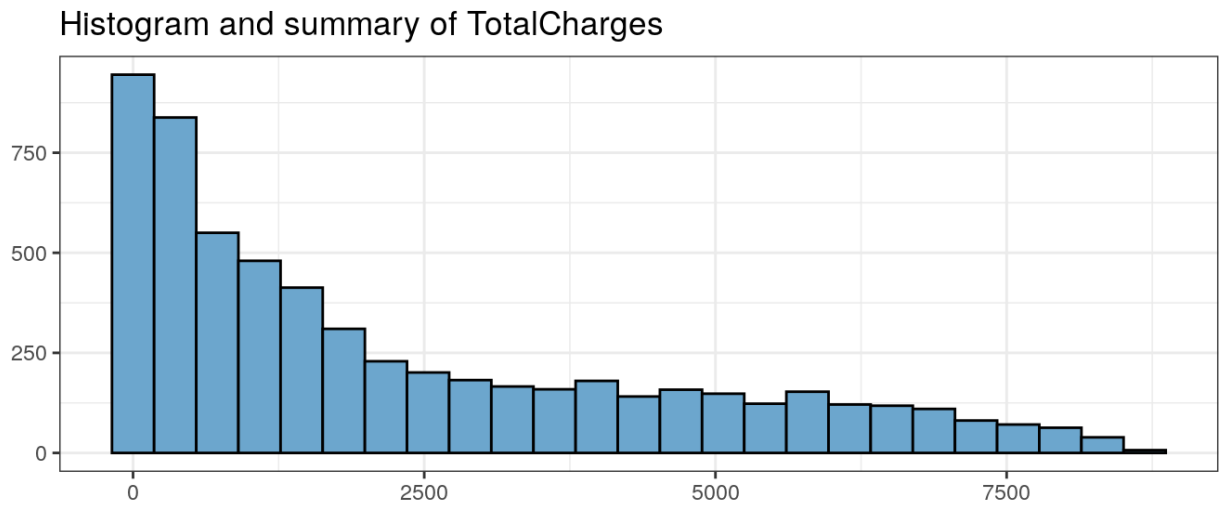


Min. : 18.2	1st Qu.: 35.6	Median : 70.4	Mean : 64.8	3rd Qu.: 89.9	Max. :118.8

الشكل 36. Monthly Charge

كان متوسط الفاتورة الشهرية حوالي 70$، كما أنه هناك نسبة كبيرة من العملاء الذين يدفعون مابين 18$ إلى 25$، والذي يدل على أن هذه هي الفاتورة الشهرية للعقد الأساسي من دون أي خدمات إضافية.

فيما يتعلق بالمتغير Total Charge (إجمالي الرسوم):



Min. : 0	1st Qu.: 402	Median :1409	Mean :2294	3rd Qu.:3842	Max. :8685

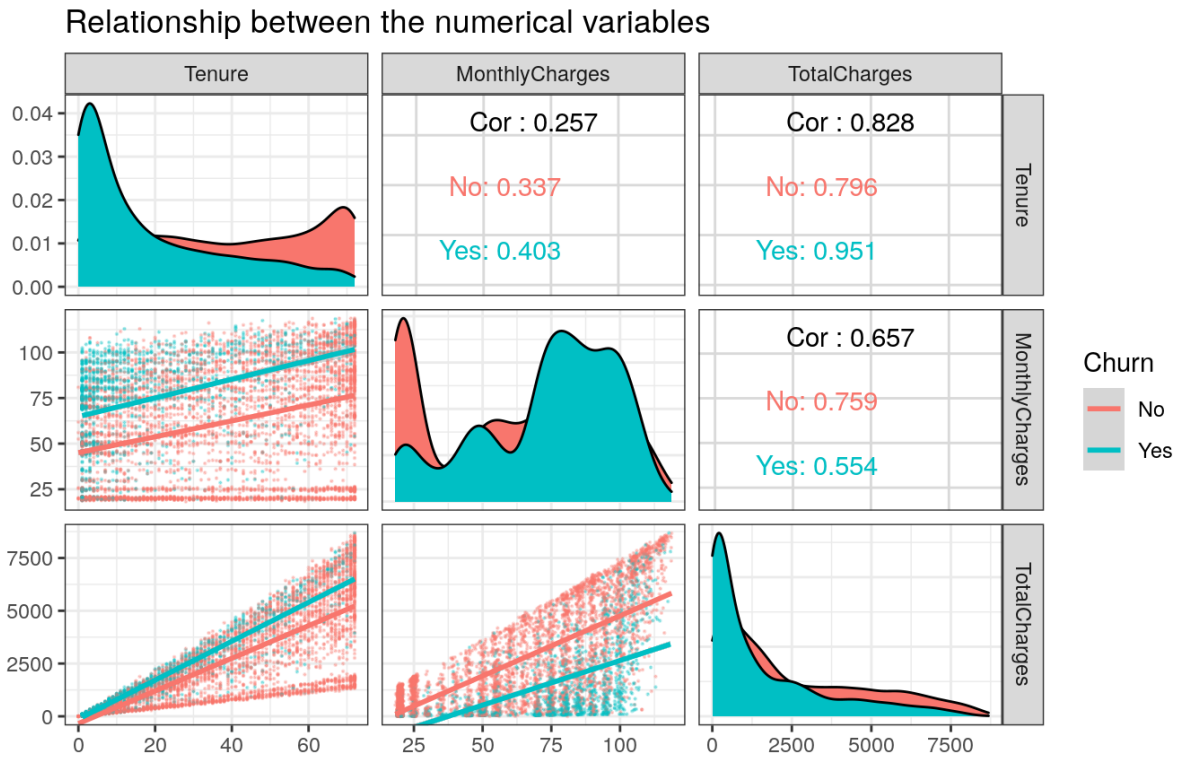
الشكل 37.Total Charge

نلاحظ ما يلي:

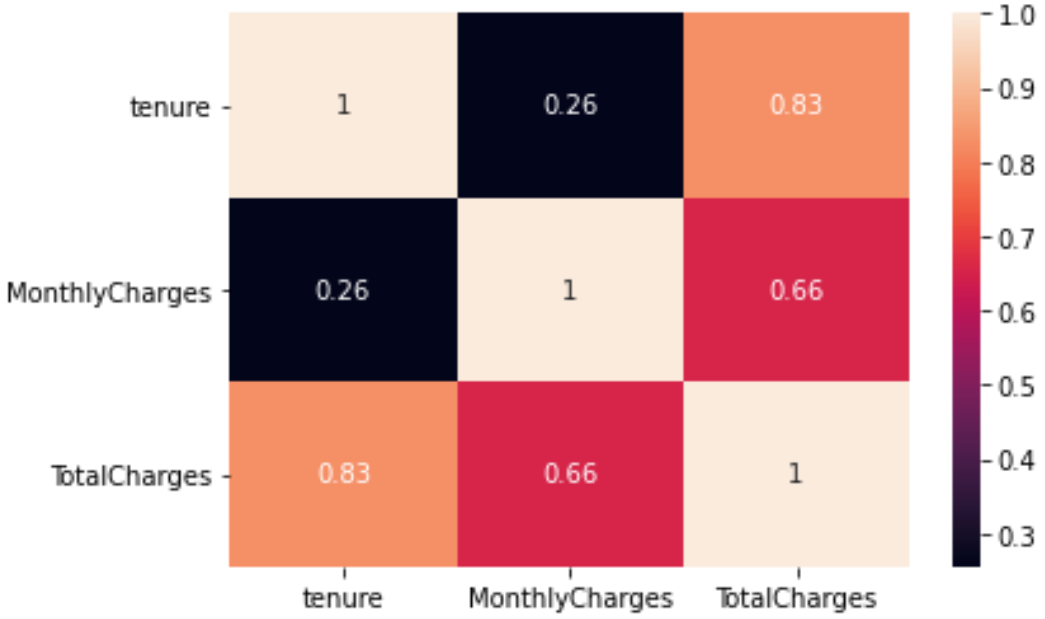
التوزيع منحرف إلى اليمين.

ونظراً لوجود نسب عالية جداً من العملاء ذو أعلى مدة (Tenure) التي بلغت 72$، وبما أن تكرار إجمالي الرسوم يكون أقل عندما يكون هناك رسوم إجمالية كبيرة، فنفترض هنا وجود عدد كبير، غير مكترث له، من العملاء الذين يمتلكون أعلى مدة في الشركة كما ويتمتعون برسوم شهرية منخفضة أو في الحد الأدنى.

2.2تصوير المتغيرات الرقمية مع بعضها البعض



الشكل 38.Relationship between numerical variables



الشكل 2.39 correlation analysis

من التصورين السابقين نلاحظ:

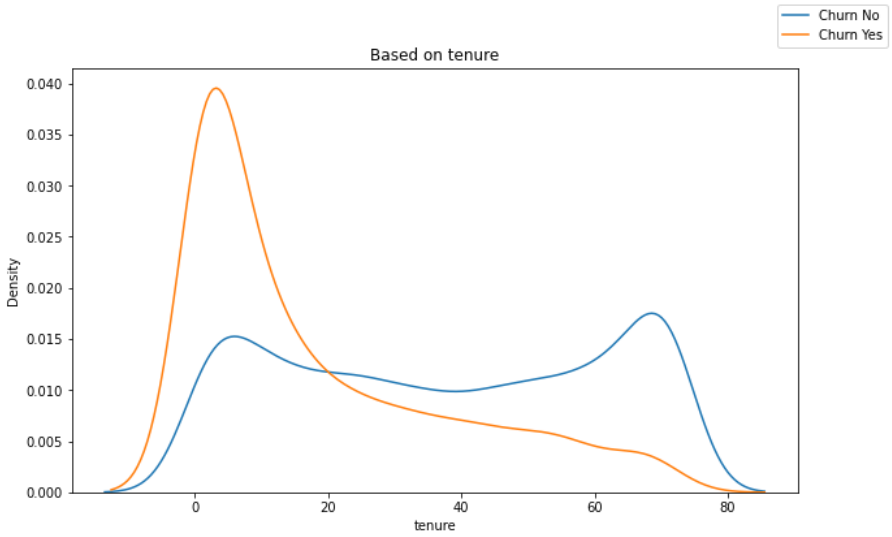
وجود ارتباط وثيق بين متغير Total Charge (الرسوم الإجمالية)، وبين Monthly Charge (الرسوم الشهرية) وTenure (المدة).

حيث أن الارتباط القوي منطقي جداً، لأن معظم العملاء يدفعون يدفعون نفس المبلغ كل شهر.

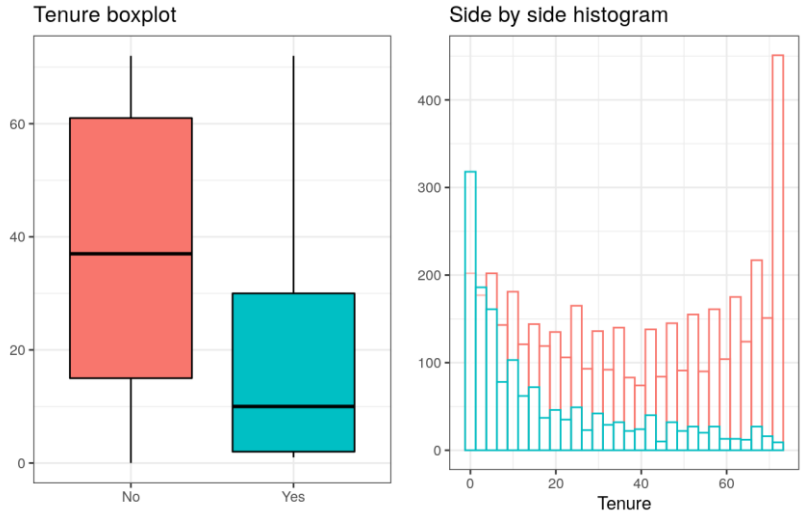
وأجل تجنب الترابط بين المتنبئين لدينا، يجب أن نسقط المتغير Total Charge من قائمة المتغيرات المتماثلة.

2.3تصوير المتغيرات الرقمية مع المتغير الهدف

فيما يتعلق بالمتغير Tenure (مدة وجود العميل في الشركة):



الشكل 1.40 Churn By Tenure

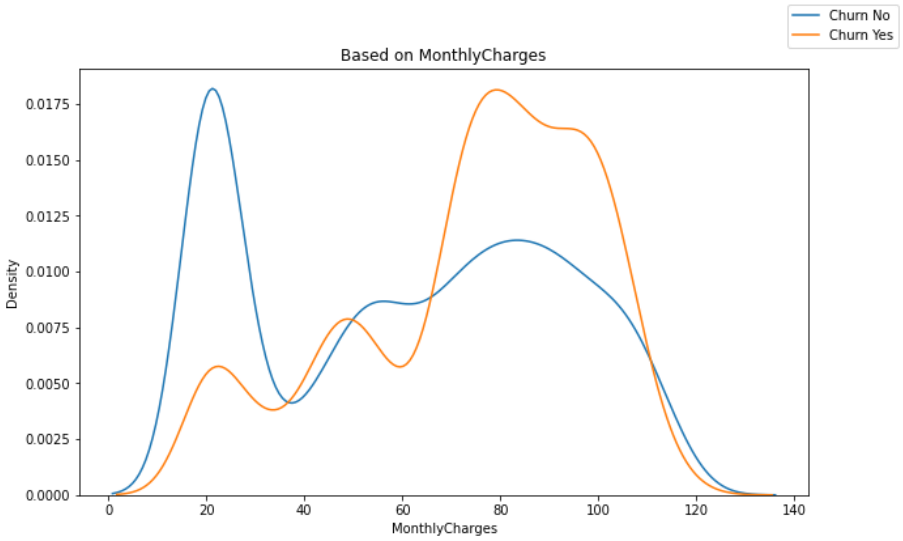


الشكل 2.41 Churn By Tenure

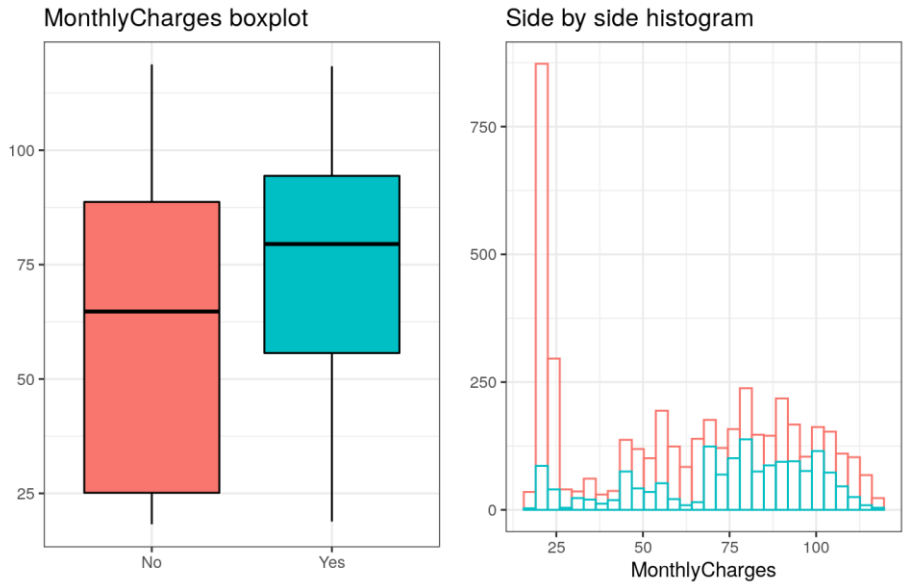
هنا يمكننا أن نرى بوضوح أنه، كل ما ازدادت مدة بقاء العميل في الشركة كلما ما انخفضت نسبة تسرب العميل من الشركة، أي أن العميل يصبح ملتزماً أكثر بالولاء لخدمات هذه الشركة.

كما ويبلغ متوسط مدة وجود العميل في الشركة للعملاء المتسربين حوالي 18 شهراً.

فيما يتعلق بالمتغير Monthly Charge (الرسوم الشهرية):



الشكل 1.42 Churn By Monthly Charge

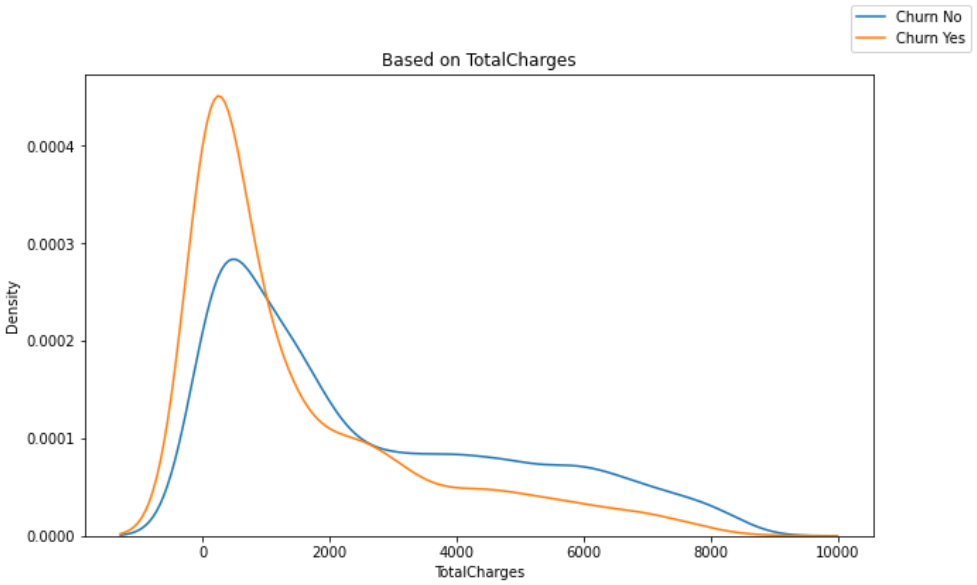


الشكل 2.43 Churn By Monthly Charge

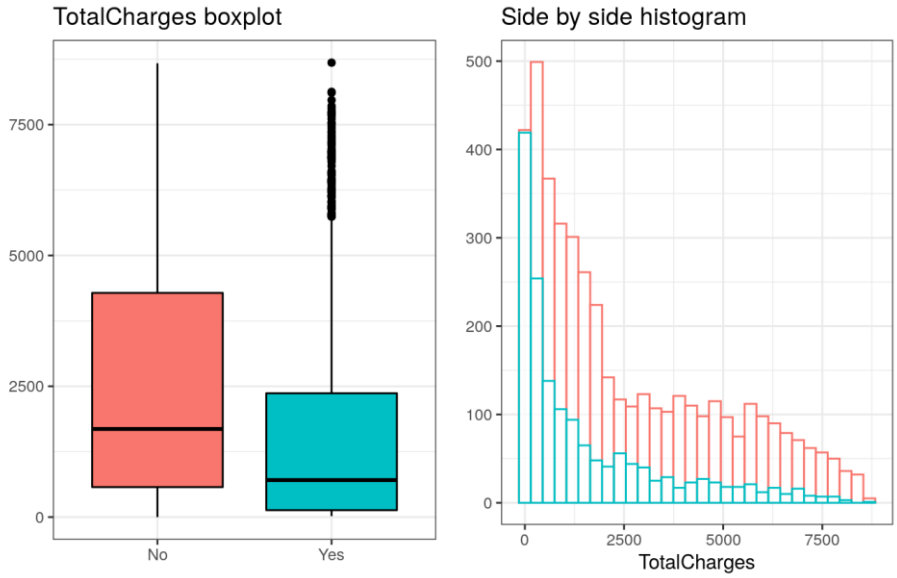
نلاحظ هنا أن العملاء المتسربين هم العملاء الذين يمتلكون رسوم شهرية أعلى من البقية، الأمر الذي يدل بدوره على أن هؤلاء العملاء هم الذين كانوا يمتلكون خدمات أكثر من غيرهم أيضاً.

كما ويبلغ متوسط الرسوم الشهرية للعملاء الذين تسربوا حوالي 74$.

فيما يتعلق بالمتغير Total Charge (إجمالي الرسوم):



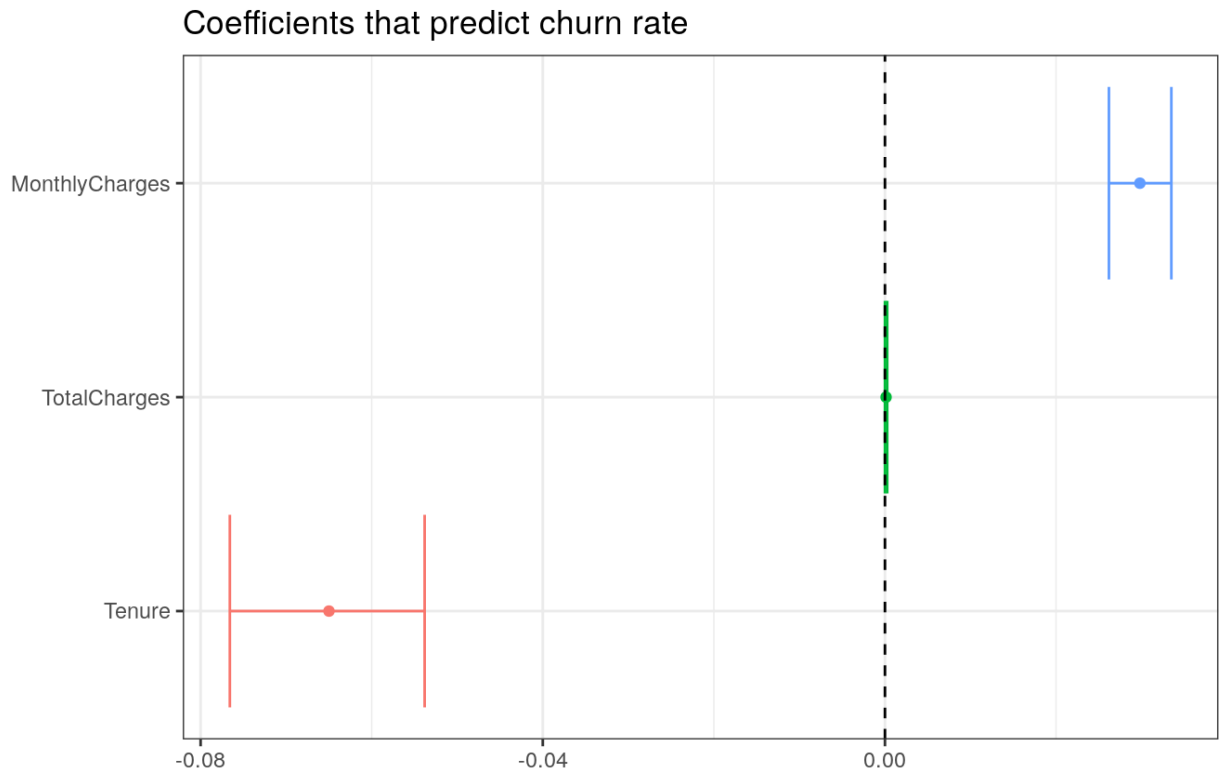
الشكل 1.44 Churn By Total Charge



الشكل 2.45 Churn By Monthly Charge

من التصوير السابق، فلا توجد علاقة ظاهرة يمكن الاعتماد عليها، ولكن وبالاستناد إلى المعلومة السابقة فيمكننا الاستنتاج أنه في حال كانت الرسوم الشهرية العملاء المتسربين هي الأعلى، ففي حالة الرسوم الإجمالية، فستكون إجمالي رسوم العملاء المتسربين هي الأقل.

فيما يتعلق بجميع المتغيرات الرقمية مع المتغير الهدف:



الشكل 46.Coefficients that prefict churn rate

في التصور السابق أخذت المتغيرات الرقمية الثلاثة وقمنا بمراجعتها مع المتغير الهدف Churn باستخدام Logistic Model، ونتج ما يأتي:

لاحظنا وجود علاقة سلبية بين متغير مدة وجود العميل في الشركة (Tenure) ومعدل التسرب (Churn).

وفيما يتعلق بمتغير الرسوم الشهرية (Monthly Charge) لاحظنا وجود علاقة إيجابية.

والعلاقة بين متغير إجمالي الرسوم (Total Charge) ومعدل التسرب (Churn) هي علاقة صغيرة جداً وقريبة جداً من الصفر 0، مما يعني عدم اعتبار هذا المتغير ذو دلالة.

المبحث الثالث

تحضير البيانات

تمهيد

يملك تحضير البيانات العديد من التقنيات التي تستخدم به قبل البدء بعملية التحليل الإحصائي وتنقيب البيانات.

والسبب في ذلك أن التضخم الكبير الذي أضحت عليه قواعد البيانات في هذا العصر يجعلها عرضة لاحتواء الكثير من البيانات المزعجة أو غير المتناسقة أو حتى فقدان بعض البيانات الموجودة فيها، وذلك بسبب ضخامتها وتدفقها من مصادر متعددة.

كما أن البيانات ذات الجودة المنخفضة سوف تؤدي بطبيعة الحال إلى نتائج بجودة منخفضة أيضاً عند تحليلها والتنقيب فيها.

ومن أجل ذلك ينبغي رفع جودة البيانات أولاً ومن قم يمكننا أن نتوقع ارتفاع كفاءة التحليل والتنقيب فيها وتيسير عملياتها لتكون بالشكل الأمثل.

كما تعتبر المعالجة المسبقة للبيانات خطوة مهمة في أي مشروع بهدف تصوير البيانات، والتنقيب في هذه البيانات، فبيانات العالم الحقيقي بشكل عام غير مكتملة وتحوي أخطاء، من أجل إنتاج نموذج جيد لحل هذه المشكلة يجب معالجة هذه القضايا واخذها بعين الاعتبار.

نستعرض فيما يلي عدة مسائل وتقنيا مستخدمة لتجهيز البيانات:

1. تنظيف البيانات

أولاً: احتوت البيانات على متغيرين (2 Attributes) كل قيمة فيهما هي تعتبر قيمة فريدة لا تعطي أي نوع من المعلومات، كما وأنهما لا تشكلان أي إضافة للمصنف، أو أي مساعدة مهمة في التنبؤ، بل والعكس، سيكونون بمثابة ضجيج لنتائج البيانات.

المتغير الأول (Unnamed)، والتي ربما تمثل أرقام الصفوف عند كل ملاحظة، والتي قد تكون مأخوذة من قاعدة بيانات أكبر.

المتغير الثاني (Customer ID).

ليصبح عدد متغيرات البيانات (Attribute) 20.

ثانياً: تم اختبار درجة اتساق البيانات داخل نفس مجموعة البيانات أو عبر مجموعات بيانات متعددة، للتأكد من خلوها من التعارض.

كما وتم التأكد من أن كل متغير لا يحتوي على أي قيم مشابهة أو متشابهة ولكن يمكن تصنيفها بشكل منفصل، والتي قد تكون بسبب أخطاء إملائية أو القيم التي تتم كتابتها بشكل مختلف.

و نتيجة لذلك رصد تعارض المتغير (Senior Citizen) مع اتساق باقي المتغيرات، و تم تحويلها من متغير ثنائي (1,0)، إلى متغير فئوي (Yes , No)، لتصبح باتساق المتغيرات الأخرى، و كتصنيف القيم في كل من المتغيرات المذكورة أعلاه.

ثالثاً: Missing values: تم التدقيق في مجموعات البيانات للتأكد من خلوها من القيم المفقودة وتم ملاحظة التالي:

تم اكتشاف 10 إدخالات للبيانات كانت مجرد مسافات فارغة، في قيم المتغير (Total Charge)، وبالتالي فإنها تشكل (بيانات مفقودة).

ولملئ القيم الفارغة نحتاج إلى معرفة الميزة (المتغير) التي توفر مزيداً من التفاصيل عن هذا المتغير، فنجد أن متغير ال (Total Charge) مرتبط بشكل كبير مع المتغير (Tenure)، والذي يرمز إلى عدد الأشهر المحققة لانتساب العميل للشركة.

ونلاحظ، أن مع الأسف قيم المتغير (Tenure) عند هذه القيم وفي هذه الأسطر تساوي (0 صفراً)، أي أن جميع العملاء العشرة الذين لديهم القيم الفارغة (Tenure = 0) يتمتعون بفترة صفرية، مما يعني أنهم عملاء جدد لم يدفعوا فواتير بعد.

ولهذا قررنا حذف هذه الأسطر وذلك للأسباب التالية:

1. ليس لدينا معلومات من المتغير (Tenure) كافية لهذه النقاط.
2. يوجد لدينا بالفعل بيانات كافية في فئة ال (No Churn) لذا يمكننا إسقاط (حذف) هذه الأسطر.
3. كما أنها كمية صغيرة جداً من القيم مقارنة بعدد الصفوف في مجموعة البيانات، وإزالة صفوف البيانات بأكملها بهذه القيم المفقودة لن يكون لها أي تأثير سلبي على التحليل.

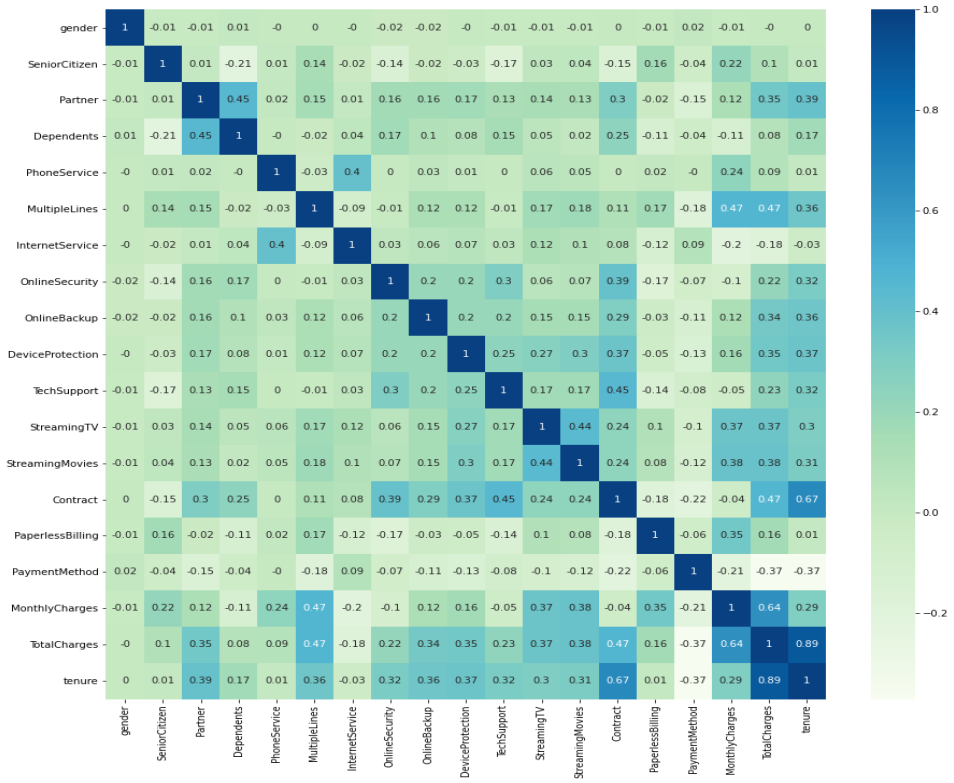
2. تحليل مدى الارتباط Correlation Analysis

كمرحلة ثانية من مراحل تحضير البيانات، نتجه وبمساعدة برنامج EXCEL لتطبيق تحليل لاختبار مدى الارتباط بين المتغيرات.

حيث أن العديد من متغيرات البيانات قد تكون زائدة، أو تعطي نفس كم المعلومات التي تعطيه متغيرات أخرى، فيمكننا بهذه الحالة الاستعاضة بمتغير واحد يعطي المعلومات المطلوبة، وهي طريقة تقلل من التشويش في البيانات، وتسمح بعد تطبيقها بإعطاء نتائج أكثر دقة من التي ممكن أن تعطيها في حال عدم تطبيقها.

وبشكل علمي أكثر، فإن (Correlation Analysis) يستخدم لتحديد ما إذا كانت الواصفتان X_1 و X_2 مترابطتين إحصائياً، فإذا كان هنالك ارتباط وثيق بينهما، فهذا يعني أن أحدهما يمكن إزالتها من أي عملية تحليل قد تجري لاحقاً.

تم تطبيق التحليل، باستخدام برنامج EXCEL، وبعد أن تم تحويل جميع البيانات من بيانات اسمية، إلى بيانات رقمية، وعبر استخدام دالة CORREL. وبعد تطبيق التحليل، كانت النتائج كما يلي:



الشكل 47.Correlation analysis

ومن خلال نتائج تحليل الارتباط أعلاه، نلاحظ ترابط بين العناصر الآتية:

X_1	X_2	درجة الارتباط
Tenure	Contract	0.67
Tenure	Total Charge	0.89
Monthly Charge	Total Charge	0.64

ومن خلال الارتباطات الناتجة، تم أخذ القرار بالاستغناء عن العناصر التالية:

- Tenure
- Total Charge

ليصبح عدد المتغيرات (Attribute) 18، بعد أن تم استبعاد متغيرين سابقاً، بالإضافة إلى هذين المتغيرين.

3. تحويل البيانات Data Transformation

في هذه المرحلة نقوم بإعادة المتغيرات جميعها إلى صيغتها الاسمية السابقة، ونقوم بتطبيق عملية (التعميم Generalization) على المتغيرات لتوليد سويات مفأهمية أعلى، كما ونقوم باستخدام (هرميات المفاهيم Concepts Hierarchies) حيث العملية هذه مفيدة تحديداً من أجل المتغيرات ذات القيم المستمرة، فعلى سبيل المثال يمكن تعميم القيم الرقمية للمتغير (Income) على شكل مجال ذي قيم متقطعة مثل: {low, medium, high}.

لم تظهر لدينا الحاجة لتطبيق هذه العملية في المتغيرات جميعها، نظراً لأن جميع البيانات ذو سوية مفأهمية متساوية.

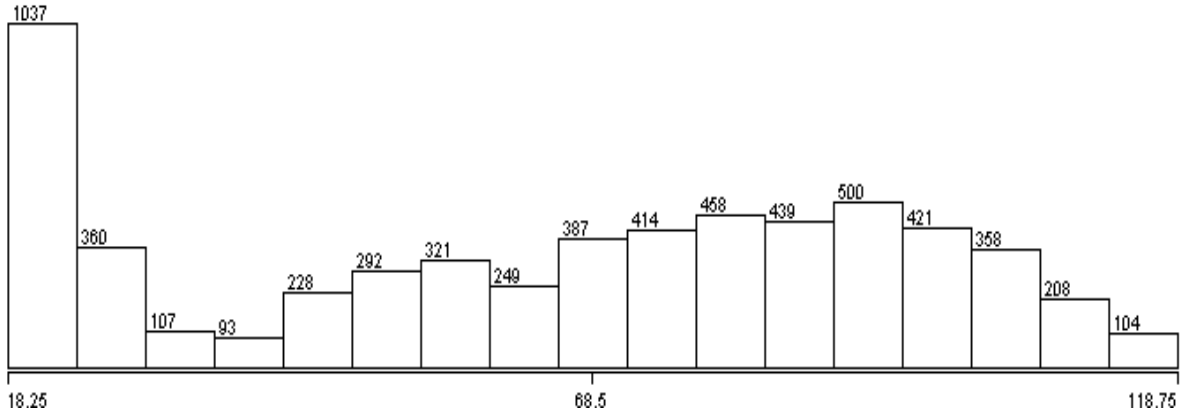
فيما عدا متغير وحيد بقي لدينا بعد تحليل مدى الارتباط، واستبعاد المتغيرات المرتبطة، وهو متغير (الدفعات الشهرية Monthly Charge).

متغير (الدفعات الشهرية Monthly Charge):

تم جمع ورصد الدفعات الشهري للعملاء في مجموعة البيانات من 18.25 وحتى 118.75.

والذي وفق إجراء التحليل المطلوب تبين لنا بأن الدفعات الشهرية لا يتم توزريعها كتوزيع طبيعي، ويستنتج من ذلك بوجود دفعات شهري للعملاء بانتشار أكثر من باقي الفئات الأخرى، وعليه قام الباحث بتقسيم هذه الدفعات بطريقة متساوية للوصول إلى أعداد متقاربة من العملاء، وذلك لضمان أن تتناسب المدخلات من البيانات مع القدرة الحسابية لأدوات التنقيب في المعطيات.

وكان توزيع الدفعات الشهرية كما يلي:



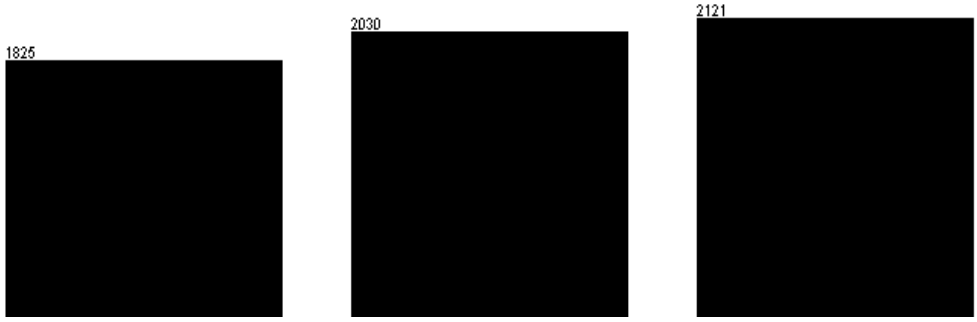
الشكل 48. توزيع الدفعات الشهرية

نلاحظ أن الدفعات الشهرية الأكثر دفعاً بين العملاء هي 18.25 بفرق كبير بعدد باقي الدفعات، حيق بلغ عدد العملاء نحو 1037 عميل. وعليه تم تقسيم البيانات على الشكل التالي:

- Low
- Med
- High

وبناءاً على هذا التقسيم، أصبح توزيعها على الشكل التالي:

No.	Label	Count	Weight
1	low	1825	1825.0
2	high	2030	2030.0
3	med	2121	2121.0



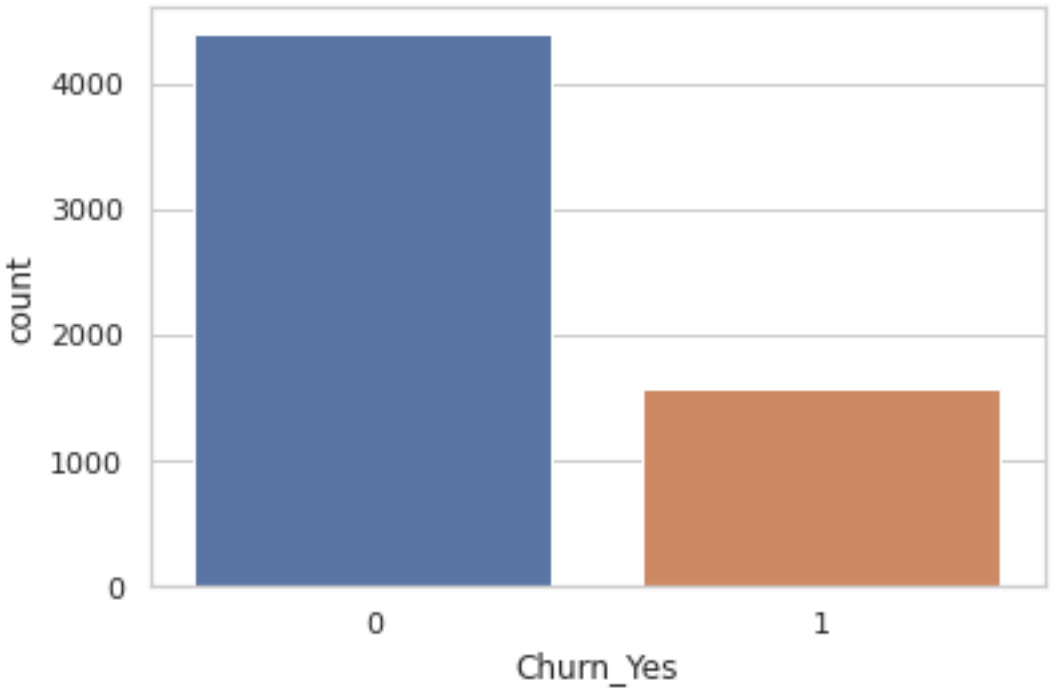
الشكل 49.توزيع الدفعات الشهرية الجديد

4. توازن البيانات Data Balance

تعتبر مجموعات البيانات الغير متوازنة، هي حالة خاصة لمشكلة التصنيف، حيث لا يكون توزيع الفئة متجانساً بين الفئات عادة، وغالباً في حالات التصنيف التي تتكون من فئتين، الفئة الغالبة (غير متسرب No Churn)، والأقلية (متسرب Yes Churn).

إن التعامل مع هذه المجموعات يعتبر مشكلة، حيث أن خوارزميات التصنيف القياسية عادة تعتبر أن مجموعة التدريب متوازنة، وهذا قد يفرض وجود تحيز تجاه الفئة الغالبة.

وفي بحثنا، نستعرض المتغير الهدف (Churn)، والذي يمثل بيانات المتسربين وغير المتسربين، ونلاحظ ما يلي:

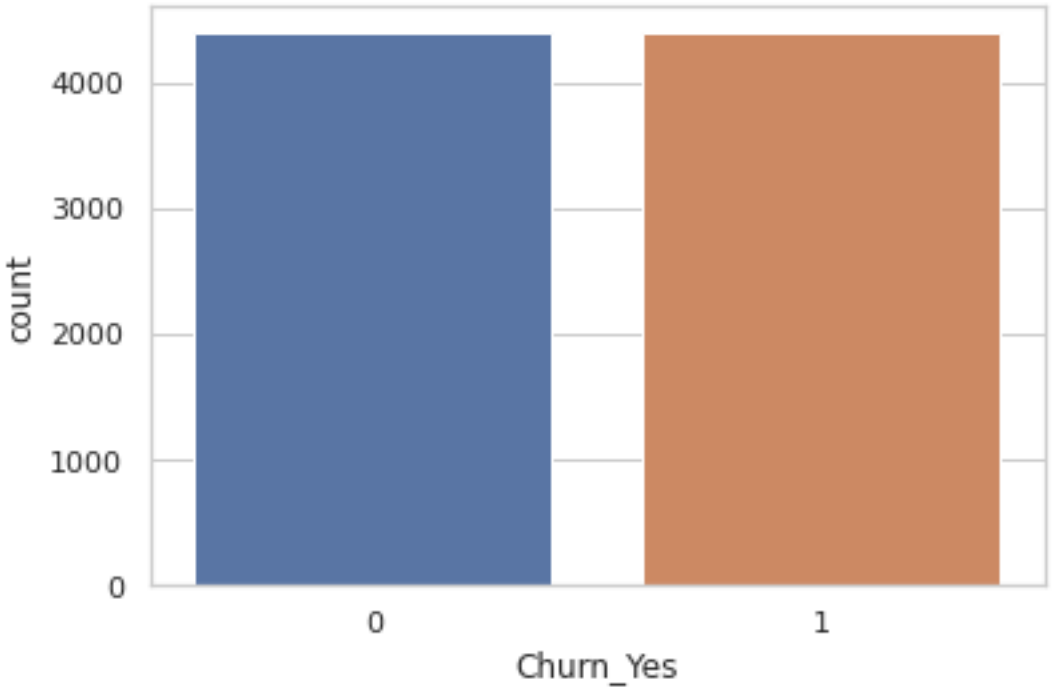


الشكل 50.Imbalanced Dataset

نلاحظ اختلال كبير في التوازن في الفئتين، حيث أن المتسربين يشكلون حوالي ال 26.5% من إجمالي بيانات المتغير (Churn).

هنالك تقنيات مختلفة للتعامل مع البيانات غير المتوازنة، وفي هذه الحالة خصيصاً، ونظراً لأن البيانات منحرفة، تم تطبيق خوارزمية SMOTE للتعامل معها، والتي تقتضي الإفراط (Oversampling)، والتي

تعني الزيادة العشوائية لفئة الأقلية والتي تشكل (المتسرب Yes Churn)، وتوزيع عدد الإدخالات بالتساوي. ليصبح توزع البيانات في المتغير Churn على الشكل التالي:



الشكل 51.Oversampling Technique

نلاحظ تساوي المتسربين مع الغير متسربين، وعليه تم تطبيق خوارزميات التصنيف.

المبحث الرابع

التصنيف Classification

1. بناء المصنف

تهدف هذه الخطوة إلى بناء نموذج تنبؤي من خلال استخدام تقنيات التصنيف التي تم توضيحها في القسم النظري. تم استخدام برنامج ويكا (Weka) لتنفيذ وبناء النموذج التنبؤي

من اجل تحديد دقة تصنيف البيانات التي تم تدريبها تم اعتماد أسلوب الاختبار (Cross Validation) بنسبة تقطيع متساوية 10 وذلك لجميع حالات التصنيف في هذا البحث ويمكن تلخيص عمل هذه الطريقة بالخطوات التالية:

1. تدريب بنسبة 100% من بيانات التدريب.
2. تقسيم بيانات التدريب إلى 10 اقسام بشكل عشوائي.
3. تنفيذ عملية تخمين المقادير لكل من هذه الأقسام العشرة وحساب نسبة تطابق القيم الحقيقية لهذه البيانات مع القيم التي تم تخمينها.
4. حساب هذه النسبة لكامل مجموعة البيانات لتكون نسبة الدقة المطلوبة.

سيتم تطبيق 5 مصنفات والمقارنة بينهم للوصول إلى أفضل خوارزمية تصنيف بناءً على البيانات الحالية، سيتم اختيار خوارزمية (الغابات العشوائية، شجرة القرار، الانحدار اللوجيستي، KNN، SVM)

تم اختيار المصنفات السابقة لان التصنيف في حالتنا هو تصنيف ثنائي (Yes, No).

وعلى اعتبار ان المصنفات السابقة جميعها خوارزميات تعلم آلي خاضعة للإشراف (Supervised learning)، فإنها مناسبة لطبيعية البيانات الخاصة بالبحث.

قبل البدء لا بد من التعريف بالمعايير التي سيتم استخدامها في مقارنة وتقييم المصنفات.

❖ **معايير المقارنة**

- **Accuracy**

دقة المصنف هي نسبة التوقعات التي حصل عليها نموذجنا بشكل صحيح بالنسبة للعدد الكلي، حيث أن:

الدقة = عدد التنبؤات الصحيحة العدد الإجمالي للتنبؤات.

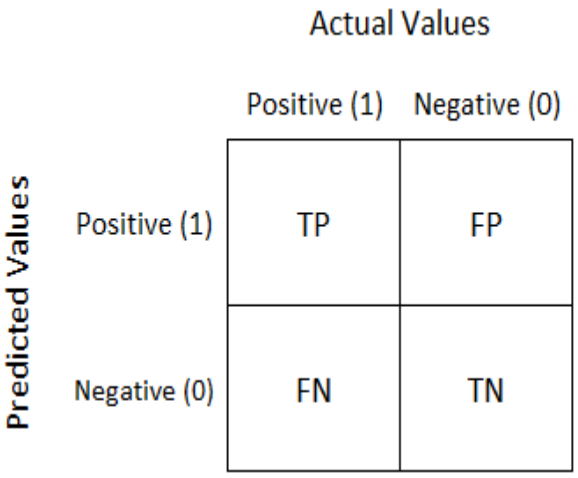
وتأخذ شكل المعادلة التالية:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

حيث أن:

- TP =True positive (توقع إيجابي صحيح)
- TN =True negative (توقع سلبي صحيح)
- FP =False positive (توقع إيجابي مغلوط)
- FN =False negative (توقع سلبي مغلوط)

ويتم تطبيق المعادلة السابقة على مصفوفة التشويش (confusion matrix).



رسم توضيحي 5. Confusion Matrix

- **Kappa statistic**

يعتبر هذا العامل أحد أهم المؤشرات لقوة التصنيف من خلال قراءته لمصفوفة الشك واختصار دقة التصنيف الموجود داخلها إلى رقم وحيد يتراوح بين 0 إلى 1 من خلال تطبيق المعادلة التالية:

$$k = \frac{P(a) - P(e)}{1 - P(e)}$$

حيث:

P (a) نسبة عدد حالات التطابق.

P (e) نسبة عدد الحالات التصادفية.

- **Recall**

هو أحد مقاييس التقييم الأكثر استخدامًا لمجموعة البيانات غير المتوازنة. يحسب عدد الإجابات الصحيحة الفعلية التي توقعها المصنف على أنها صحيحة.

يُعرف الاستدعاء أيضًا بالمعدل الإيجابي الحقيقي (TPR) أو الحساسية أو احتمال الاكتشاف.

ويطبق من خلال المعادلة التالية:

$$\text{Recall} = \frac{TP}{TP + FN}$$

حيث أن:

- TP =True positive (توقع إيجابي صحيح)
- FN =False negative (توقع سلبي مغلوط)

- Precision

يصف دقة نموذج استخراج البيانات لدينا. من بين تلك الحالات المتوقعة إيجابية، كم منها إيجابي بالفعل؟

تسمى الدقة أيضًا مقياس الدقة أو الجودة أو القيمة التنبؤية الإيجابية.

$$\text{Precision} = \frac{TP}{TP + FP}$$

حيث أن:

- TP =True positive (توقع إيجابي صحيح)
- FP =False positive (توقع إيجابي مغلوط)

1.1 الانحدار اللوجيستي Logistic regression

أعطى نموذج الانحدار اللوجيستي نتائج على الشكل التالي:

```
Time taken to build model: 0.89 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6552               74.6411 %
Incorrectly Classified Instances      2226               25.3589 %
Kappa statistic                          0.4928
Mean absolute error                      0.3365
Root mean squared error                  0.4115
Relative absolute error                 67.2995 %
Root relative squared error             82.2949 %
Total Number of Instances             8778

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.709    0.216    0.766      0.709    0.737      0.494    0.824     0.838     no
               0.784    0.291    0.729      0.784    0.756      0.494    0.824     0.791     yes
Weighted Avg.  0.746    0.254    0.748      0.746    0.746      0.494    0.824     0.814

=== Confusion Matrix ===

   a    b   <-- classified as
 3111 1278 |    a = no
  948 3441 |    b = yes
```

حيث أعطى دقة مصنف 74.64% وبلغ (Recall) بالنسبة ل (Yes) 0.709 وبالنسبة ل (No) 0.784،

كما أنه أعطى النتائج التالية:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
Logistic regression	74.64	0.49	0.784	0.709	0.729	0.766

1.2آلة المتجهات الداعمة SVM

عند تطبيق خوارزمية ال SVM كانت النتائج على الشكل التالي:

```
Time taken to build model: 76.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6468               73.6842 %
Incorrectly Classified Instances      2310               26.3158 %
Kappa statistic                          0.4737
Mean absolute error                      0.2632
Root mean squared error                  0.513
Relative absolute error                 52.6316 %
Root relative squared error            102.5978 %
Total Number of Instances             8778

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.654    0.180    0.784      0.654    0.713      0.480    0.737     0.686     no
               0.820    0.346    0.703      0.820    0.757      0.480    0.737     0.667     yes
Weighted Avg.  0.737    0.263    0.744      0.737    0.735      0.480    0.737     0.676

=== Confusion Matrix ===

   a    b   <-- classified as
 2870 1519 |    a = no
  791 3598 |    b = yes
```

حيث أعطى دقة مصنف 73.68% وبلغ (Recall) بالنسبة ل (Yes) 0.820 وبالنسبة ل (No) 0.654،
كما أنه أعطى النتائج التالية:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
SVM	73.68	0.47	0.820	0.654	0.703	0.784

1.3الجار الأقرب KNN

عند تطبيق خوارزمية ال KNN كانت النتائج على الشكل التالي:

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7045               80.2575 %
Incorrectly Classified Instances      1733               19.7425 %
Kappa statistic                          0.6051
Mean absolute error                      0.2135
Root mean squared error                  0.3769
Relative absolute error                 42.6904 %
Root relative squared error             75.3896 %
Total Number of Instances             8778

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.698    0.093    0.882      0.698    0.780      0.619    0.876     0.901     no
               0.907    0.302    0.750      0.907    0.821      0.619    0.876     0.814     yes
Weighted Avg.  0.803    0.197    0.816      0.803    0.800      0.619    0.876     0.858

=== Confusion Matrix ===

   a    b   <-- classified as
 3065 1324 |    a = no
  409 3980 |    b = yes
```

حيث أعطى دقة مصنف 80.25% وبلغ (Recall) بالنسبة ل (Yes) 0.907 وبالنسبة ل (No) 0.698،

كما أنه أعطى النتائج التالية:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
KNN	80.25	0.60	0.907	0.698	0.750	0.882

j48 1.4

عند تطبيق خوارزمية ال J48 كانت النتائج على الشكل التالي:

```
Time taken to build model: 0.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        6874               78.3094 %
Incorrectly Classified Instances      1904               21.6906 %
Kappa statistic                          0.5662
Mean absolute error                      0.285
Root mean squared error                  0.408
Relative absolute error                 56.9919 %
Root relative squared error             81.6055 %
Total Number of Instances             8778

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.705    0.139    0.836      0.705    0.765      0.573    0.818     0.843     no
               0.861    0.295    0.745      0.861    0.799      0.573    0.818     0.734     yes
Weighted Avg.  0.783    0.217    0.790      0.783    0.782      0.573    0.818     0.788

=== Confusion Matrix ===

    a    b   <-- classified as
 3094 1295 |    a = no
  609 3780 |    b = yes
```

حيث أعطى دقة مصنف 78.30% وبلغ (Recall) بالنسبة ل (Yes) 0.861 وبالنسبة ل (No) 0.705،
كما أنه أعطى النتائج التالية:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
J48	78.30	0.56	0.861	0.705	0.745	0.836

1.5 الغابات العشوائية Random Forest

عند تطبيق خوارزمية ال Random Forest كانت النتائج على الشكل التالي:

```
Time taken to build model: 2.34 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7467               85.0649 %
Incorrectly Classified Instances      1311               14.9351 %
Kappa statistic                          0.7013
Mean absolute error                      0.2283
Root mean squared error                  0.341
Relative absolute error                 45.6695 %
Root relative squared error             68.2055 %
Total Number of Instances             8778

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.771     0.069     0.917      0.771     0.838      0.710     0.910     0.928     no
               0.931     0.229     0.802      0.931     0.862      0.710     0.910     0.866     yes
Weighted Avg.  0.851     0.149     0.860      0.851     0.850      0.710     0.910     0.897

=== Confusion Matrix ===

    a    b   <-- classified as
 3383 1006 |    a = no
  305 4084 |    b = yes
```

حيث أعطى دقة مصنف 85.06% وبلغ (Recall) بالنسبة ل (Yes) 0.931 وبالنسبة ل (No) 0.771،

كما أنه أعطى النتائج التالية:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
Random Forest	85.06	0.70	0.931	0.771	0.802	0.917

المبحث الخامس

المقارنة واختيار المصنف

بعد أن تم سابقاً بناء مصنفات باستخدام عدة خوارزميات سيتم المقارنة بينهم لاختيار المصنف الأفضل النهائي، وسيتم ذلك من خلال مقارنة(Accuracy ، Recall، Kappa statistic، Precision) الخاصة بكل مصنف. ونتوصل إلى التالي:

	Accuracy	Kappa statistic	Recall yes	Recall No	Precision Yes	Precision No
Logistic regression	74.64	0.49	0.784	0.709	0.729	0.766
SVM	73.68	0.47	0.820	0.654	0.703	0.784
KNN	80.25	0.60	0.907	0.698	0.750	0.882
J48	78.30	0.56	0.861	0.705	0.745	0.836
Random Forest	85.06	0.70	0.931	0.771	0.802	0.917

كما هو واضح بالجدول فإن خوارزمية الغابة العشوائية Random Forest تعطي تفوقاً على جميع المقاييس ف لو أردنا المقارنة بناء على الدقة تكون هي الأعلى، كما هي الأعلى على معيار (Kappa)، وعلى اعتبار يهمنا معرفة الزبون المتسرب أكثر من الغير متسرب ف يمكن ملاحظة أن (Recall yes) هو الأعلى ايضاً، كما أنها الأعلى على المعيار الأخير (Precision)، لتكون هي المصنف الأفضل على جميع المقاييس والمعايير.

يليها مباشرة مصنف KNN، الذي يليه مصنف J48، والذي منه يمكننا استنتاج أن الخوارزميات الشجرية تعطي الدقات الأعلى على بياناتنا. وتأتي خوارزمية SVM أخيراً حيث انها أعطت اقل دقة (Accuracy) وهي الأقل على معيار (Kappa) ومعيار (Precision)، لكن على معيار (Recall)، يأتي الانحدار اللوجيستي Logistic أخيراً.

الفصل الخامس

النتائج والتوصيات

النتائج (Results)

شكلت هذه الدراسة رديفاً للعديد من الأبحاث الساعية للتأكيد على إمكانية الاستفادة من الأدوات المختلفة للتنقيب في المعطيات في تسرب العملاء في شركة اتصالات والحد منه، بالاستفادة من بيانات الشركة عن الخدمات المقدمة لهم.

تطابقت نتائج البحث مع الدراسات المرجعية التي أكدت على الدور الفعال للتنقيب في المعطيات على استكشاف المعرفة وخاصة مع وجود بيانات حقيقية، متوارنة و موثوقة.

وفيما يلي نعرض أهم ما توصلت إليه الدراسة من نتائج:

نتائج تطبيقية:

1. إن استخدام الخوارزميات الشجرية يعطي أفضل النتائج في التصنيف، والتي يمكن أن تطبق على البيانات، والتي كان أفضلها خوارزمية الغابة العشوائية.
2. التوصل إلى مصنف قادر على التنبؤ بتسرب الزبائن من شركة اتصالات، بدقة جيدة جداً وصلت إلى 85%.
3. التوصل إلى طريقة لحل مشكلة عدم توازن البيانات، والقيم المفقودة.
4. اعتماداً على تحليل البيانات، القائم على التصوير Data Visualize، تمكن الباحث من الوصول إلى النتائج التالية:

- بالنسبة للمتغيرات الفئوية:
  - نلاحظ أن المؤثر الأول بالتسرب بنسبة 88.13%، هو توجه العملاء إلى عقود شهرية، إذ لا يتجه العملاء لتشكيل التزام قوي تجاه الشركة.
  - يليه تأثير الذين لا يوجد لديهم معيل بنسبة 82%، إذ لربما لا يوجد لديهم مصدر مادي للاشتراك بالخدمات المختلفة
  - يليه على التتالي الأمان، والدعم الفني، إذ عدم وجود وتوافر هذه الخدمات، يدفع العميل للتخلي عن الخدمة جميعها.
  - كما أن يكون العميل من كبار السن المتقاعدين، وعدم استعمال الفواتير الورقية، واستعمال ال Fiber Optic هي كلها أسباب تؤثر وتجعل العملاء بتسربون.

- كما أن عدم توافر خدمة النسخ الاحتياطي، وحماية الجهاز، هي أيضاً أسباب تدفع العميل للتخلي عن الشركة.
- كما أنه ليس للجنس أي ارتباط مع أي من المتغيرات الفئوية الأخرى.

- بالنسبة للمتغيرات الرقمية:
  - إن متوسط مدة بقاء العميل في الشركة هو 29 شهراً.
  - متوسط الفاتورة الشهرية حوالي 70$، كما أنه هناك نسبة كبيرة من العملاء الذين يدفعون مابين 18$ إلى 25$.
  - إن متوسط مدة بقاء العميل في الشركة للعملاء المتسربين هو 18 شهراً.
  - متوسط الرسوم الشهرية للعملاء المتسربين هو 74$.
  - متوسط الرسوم الإجمالية للعملاء المتسربين هو 1550$.
  - يوجد ارتباط وثيق بين متغير إجمالي الرسوم، مع المتغيرين مدة بقاء العميل في الشركة والرسوم الشهرية.
  - يمكن إسقاط (حذف) متغير إجمالي الرسوم بناءاً على نتائج تحليل الارتباط.

نتائج نظرية:

1. تقدم أدوات التنقيب في المعطيات إمكانات واسعة لتطبيقات الذكاء الصنعي لشتى أنواع البيانات ومن بينها معطيات عن زبائن شركة اتصالات، مما يسمح لنا بفهم أوسع لسلوكيات الزبائن، و القيام بخطوات استباقية للحفاظ على الزبائن.
2. تعتمد دقة نتائج التنقيب في المعطيات بصورة أساسية على مدى دقة وشمولية البيانات المستخدمة.
3. يشكل (Weka) أداة فعالة للاستفادة من تقانة التنقيب في المعطيات، من خلال ما يمنحه التطبيق من خيارات وإمكانيات متنوعة، حيث يسمح التطبيق بتجهيز البيانات وفق صيغ متعددة واستخدامها وفق شتى تقانات التنقيب في المعطيات.
4. إن عملية تحضير البيانات قبل القيام بعملية التنبؤ، تعد خطوة جوهرية وأساسية في عملية التصنيف حيث أن بناء نموذج من دون تحضير البيانات يعطي نتائج مغلوطة.

التوصيات (Recommendations)

في ضوء النتائج التي تم عرضها سابقاً، وبهدف التطبيق العملي للنموذج المستنتج، المقترح، الذي يعمل كمساعد لشركات الاتصالات للتنبؤ بتسرب الزبائن، والحد منه، ومعرفة أهم العوامل المؤثرة، بالاعتماد على بيانات العملاء، وأهم الخدمات الذين يتوجهون إليها. نتوجه بالتوصيات التالية:

1. إعطاء التنقيب في البيانات أهمية أكبر داخل الشركات لما تحمله من أدوات تساعد على معرفة نتائج ومعلومات كان من الصعب الوصول اليها سابقاً.
2. الاهتمام باستمرارية السعي بجمع المعلومات والبيانات عن الزبائن لما له من أهمية للتنبؤ بتسربهم.
3. إشراك الزبون في التعبير عما يريد، سماع شكاويه، الأمر الذي يزيد من ارتباطه وانتماءه بالشركة.
4. تقديم المكافئات، بالعروض والباقات، للعملاء الملتزمين لدينا، ليزداد ولاء العملاء، الأمر الذي يسأهم في زيادة الأرباح.
5. تقديم خدمات إضافية بتكلفة منخفضة، لمساعدة الزبائن في الحصول على المزيد من الخدمات أيضاً.
6. الاهتمام أكثر بالتسويق لخدمة الأمان، النسخ الاحتياطي، حماية الجهاز، والدعم الفني، لما لهم من تأثير في تسرب الزبائن.
7. محاولة الشركة لخفض الرسوم الشهرية للعملاء، لجذبهم باستمرارهم في تجديد عقودهم.
8. التوسع في هذه الدارسة من خلال الاستفادة من أدوات وخوارزميات أخرى للتنقيب في المعطيات، لمعالجة هذا النوع من البيانات، وعلى وجوه الخصوص استخدام خوارزميات وأدوات أخرى غير التصنيف، مثل العنقدة وقواعد الترافق سعياً للوصول إلى نتائج أفضل.

المراجع

المراجع الأجنبية:

- Ahola, j., & Rinta, R. (2001). *Data mining case studies in coustomer profiling.*
- alex soft .(2019) .customer churn predicting using machine learning . *KDnuggets*.
- Bayaga, A. (2010). *multinomial logistic regression: usage and appmication in risk analysis.*
- Bendik Bygstad .(2003) .The implementation puzzle of CRM systems in knowledge based organizations .*Information Resources Management Journal*.
- brian beers .(2021) .what is the telecommunications sector .*investopedia*.
- Chris Smith .(2017) .*Decision Trees and Random Forests .*
- data mining tutorial .(2020) .*EDucba*.
- Dean, J. (2014). *Big Data, Data Mining, and Machine Learning.*
- deval shah .(2017) .datamining tools .*towards data science*.
- dionysios zelios .(2018) .predicting customer churn . *Avaus*.
- Greenberg .(2002) .Integrating an emotion-focused approach to treatment into psychotherapy integration .*Psychotherapy Integration*.189–154 ،
- ian witten .(2019) .data mining with weka .*Future learn*.

- janice m. morse .(2004) .Qualitative Significance .*SAGE*.
- Kelleherd, J., & Tierney, B. (2018). *DATA SCIENCE.*
- Maroof, D. A. (2012). *Statistical Methods in Neuropsychology.*
- Mills, P. (2011). *Efficient statistical classification of satellite measurements.*
- Rudhwan Sideek و ،Ghayda AL-Talib .(2015) .*Evaluation of clustering validity .*
- sharon lin 5 .(2018) .data scince models for predicting enterprise churn . *ReForge*.
- Yen, C., & Hen, H. (2006). *Applying data mining to telecom churn.*

المراجع العربية:

- اميرة خضير كاظم العنزي. (2010). *دور ابعاد ادارة علاقات الزبون والتفكير الابداعي في تحقيق النجاح الاستراتيجي.* جامعة الكوفة - كلية الادارة والاقتصاد - قسم ادارة الاعمال.
- أيوب العيسى . (2021). *data visualization.* Imam Mohammad Ibn Saud Islamic University.
- جلال الضاهر. (2014). *تصميم نموذج دعم لادارة الموارد البشرية بالاعتماد على تقنيات الذكاء الصنعي.*
- حنين محمد. (2015). *لماذا يتخلى عملاؤك عن منتجك و ما الذي يمكن فعله.* academy hsoub.
- د. محمد مصطفى عبيد. (2017). *كتاب التحليل المتقدم و تنقيب البيانات .* القاهرة: دار الفكر العربي.
- راكان رزوق. (2013). *التنقيب في البيانات الاسس النظرية والتطبيقية.*
- زكريا الدوري، و داليا احمد. (2007). *دور تنقيب البيانات في زيادة اداء المنظمة.*
- عبد الرحيم احمد. (2018). *توقع تسرب الزبائن في شركات الاتصالات باستخدام تعلم الالة في بيئة المعطيات الكبيرة.*
- عدنان غانم، و فريد الجاعوني. (2011). *استخدام تقنية الانحدار اللوجيستي ثنائي الاستجابة في دراسة اهم المحددات الاقتصادية والاجتماعية لكفاية لاسرة.* دمشق.
- غيداء الطالب، و رضوان الجوادي. (2008). *تقبم صحة العنقدة.*
- محسن حمود. (2020). *التمثيل المرئي للبيانات.* data science.
- محمد أبو خالد. (2020). *ما هي أسباب فشل إدارة علاقات العملاء.* Tijaratuna.
- محمد حجوز. (2014). *تحسين خوارزميات K-Means.*

- ayaz taher. (2020). *تهيئة و تنظيف البيانات.* Hasafa.
- محمد دعيش، و محمد ساري. (2017). *نموذج االنحدار اللوجستي: مفهومه، خصائصه، تطبيقاته.*
- مصطفى عبيد. (2019). *التحليل المتقدم وتنقيب البيانات.*
- مطلب سوزي الشبيل. (2012). *تطبيقات إدارة علاقات الزبائن في مراحل الشراء الالكتروني و أثرها في بناء القيمة للزبون.*