Modélisiation de la valeur de la longévité de vie des clients

Abstract

Les entreprises voient leur clients de plus en plus en tant que leurs durée de vie comme un nombre de clients présents net de profits calculés sur un certain nombre de mois.

La valeur de durée de vie est un indicateur puissant et clair qui mesure synthétiquement la profitabilité et le risque d’arrêt d’achats au niveau individuel de la profitabilité des clients . Pour les clients existants , la durée de vie attendue peut permettre les entreprise à développer la loyaleté de sa clientèle et le traitement de stratégies maximisant la valeur de leur clientèle. Pour les clients nouvellement acquis ,cette durée de vie aide les entreprise à choisir les bons clients à cibler pour le futur .

Le calcul de la durée de vie varie selon l’industrie . dans la télécommunication , la valeur marginale d’un client chaque mois et la courbe de survie du clients sont les 2 majeurs composants pour calculer la valeur de la durée de vie d’un client .Comme la valeur marginale est déduite des modèles de comptabilités ce qui reste est l’estimation de leur courbe de survavibilité . dans cette étude , l’analyse de la survie est appliquée pour estimer la courbe de survie des clients et d’où leur durée de vie.

Introduction

Dans le domaine de télécommunication les clients peuvent choisir entre différents fournisseurs des services multiples

et exerce activement leurs droits de changer d’un fournisseur à un autre . cette féroce compétition , ou les clients demandes des produits sur mesures pour des prix mois chers ,en parallèle avec des fournisseurs qui visent l’acquisition de leur objectifs d’entreprise .Etant donné le fait que la télécommunication comme industrie admet 30-35% de risque d’arrêt d’achat et sa coute 5-10 fois plus de recruter un nouveau client que de les maintenir , la rétention des clients est devenu plus importante que l’acquisition de nouvelle clientèle. Pour plusieurs opérateurs incubant la rétention de clients hautement profitables est une difficulté de l’entreprise .Plusieurs fournisseurs déploient des stratégies de rétention et synchronisent les programmes et processus pour maintenir leur clientèle un peu plus long en leur donnant des produits et services personnalisés à la demande .Avec des stratégies de rétention en place, plusieurs entreprises commence à inclure la réduction du risque d’arrêt d’affiliation(churn) comme un de leur objectives .

Avec un marché de plus en plus mur , les entreprise de télécommunication ne se satisfirent pas soit même en prédisant le churn ,cependant, ils commencent à voir leur clients en tant que des durées de vie . Non seulement , ils font la différence entre les clients qui durent plus et ceux qui ne le font pas , ils distinguent aussi entre les clients hautement profitables et ceux qui ne le sont pas. La durée de vie et donc développée pour satisfaire les entreprise de télécommunication qui ont besoin d’évaluer la valeur de leurs clients .

Les méthodes statistiques conventionnelles (régression logistiques , arbre de décision , etc …) sont très puissant dans la prédiction du churn .ces méthodes ne peuvent pas prédire la sortie ou la durée d’affiliation mais l’analyse de survie était en premier lieu faite pour décrire les données de survie , et donc un efficace et puissant outil est de prédire leurs survie/arrêt d’affiliation.

Le but de cette étude est de calculer la valeur de durée de vie en estimant la courbe de survie . les résultats aident les fournisseurs de télécommunications à développer la loyaleté et à développer des stratégies maximisant le valeur de leur clientèle . c’est aussi utile de développer des stratégies pour augmenter les « bons» clients .

Objectifs :

Les 2 objectifs de cette étude sont de développer le concept de durée de vie dans la télécommunication et le second et de démontrer comment les techniques d’analyse de survie sont utilisés pour estimer la valeur d’un client durant sa durée de vie

Définition et exclusions

Survie/arrêt d’affiliation : En télécommunications , le churn est l’action d’arrêter l’affiliation . dans cette étude l’arrêt forcé par le fournisseur (par manque de paiement par exemple) ou par le choix du client sont choisit(plus compliqué). La survie d’un client et l’opposé de son arrêt d’affiliation.

Active : les clients qui ont eu leurs services suspendus involontairement et qui sont dans l’étape de collection(la récolte) ne sont pas « actives »

Granularité : cette étude examine la Survie/Churn dans le niveau des comptes.

Contrat client : Cette étude ne distingue pas les clients avec et sans contrats , Néanmoins des modèles séparés sont désirable pour chaque status de contrat.

Exclusion : cette étude n’inclue pas les comptes des employés.

**La valeur de vie d’un client**

Le calcul de la valeur de la durée de vie (LTV) varie selon les domaines en télécommunications , la valeur marginale d’un client par mois et la courbe de survie sont les composants les plus importants dans le calcul de la valeur du client durant sa durée de vie . cette valeur et la valeur net au présent d’un profit d’un client sur un certain nombre de mois . voilà la formule pour calculer la duré de vie .

LTV= MM\* somme (de i à T) p^i /(1+r/12)^i-1

Avec MM : la valeur marginale par mois durant les 3 derniers mois pour les clients existant ou la valeur du mois dernier pour les clients récemment acquis . MM est calculée soient avec de modéles de comptabilité ou estimé avec un ensemble de modèles de régressions . Le calcul de MM n’est pas le bout de cette étude .

T : est le nombre de mois en considération pour calculer la valeur de la durée de vie d’un client .elle peut être 24,36, ou un autre nombre décrit par un expert métier.

r : le taux de réduction

p’ =la séries de probabilité de servie d’un client d’un mois 1 à T , elle est estimée d’après un modèle de survie de clients.

**Analyse de survie et le Survie/Churn des clients**

la Survie est une famille de méthodes statistiques pour étudier l’occurrence et le temps des événements . Du début , l’analyse de survie a été faite pour les données longitudinaux sur l’occurrence d’événement. La suivie du churn d’un client est un bon exemple. Les données de survie ont 2 caractéristiques qui son t difficiles à manipuler conventionnellement avec des méthodes statistiques : <Censoring> et les covariables dépendantes du temps.

Généralement , la la fonction de survie et la fonction de danger(ou hazard) sont utilisée pour décrire le status de survie d’un client durant le tenure d’observations . la fonction de survie donne la probabilité de survie après un certain point temporelle t.

Cependant la fonction de danger décrit le risque d’un événement (dans ce cas , churn d’un client) dans un intervalle de temps après t , conditionnée sur le fait que le client a survécu jusqu’à t . et donc la fonction de danger est pls intuitive à utiliser que l’analyse de survie car elle essaie de quantifier le risque instantané d’arrêt d’affiliation d’un client dans un temps t donné la survie jusqu’à ce point.

Pour l’analyse de survie la meilleurs méthode d’observation est prospective. Nous commençons à observer les clients à un temps donné appelé (origine de temps)et puis on les suit pour une durée substantielle de temps en faisant le point sur les temps ou l’arrêt d’affiliation se produit . les clients qui n’ont pas encore quitté sont appelé ‘des cas censorée’ et ceux qui ont déjà quitté sont des ‘cas observés’. On ne prédit pas seulement le temps du churn mais aussi comment les covariables dépendantes du temps (appels des clients au centre de service , la migration d’un client à un autre type plan , le changement des options de méthode de paiement , etc) impactent les occurrence et le contexte temporelle de la Survie/Churn d’un client.

SAS/STAT à 2 procédure pour l’analyse de survie : PROC LIFEBREG et PORC PHREG. Le LIFEREG est une procédure qui produit des modèles de régressions paramétrés avec des données de vie ‘censorés’ en utilisant une estimation du maximum de vraisemblance (max likelihood). La procédure PHREG est une méthode semi-paramétré d’ analyse de régression utilisant le maximum de vraisemblance PARTIELLE .

PROC PHREG a gagné de popularité sur le PROC LIFEREG dans les 100 derniers années . car elle comprend les covariables dépendantes du temps . Cependant , si l’allure des distribution de survie et la fonction de danger sont connus , PROC LIFEREG donne des résultats plus efficaces et estimes avec un erreur standard inférieur à celui de PROC PHREG .

**Stratégie d’échantillonnage**

Un échantillon de 64 320 clients actives étaient choisies à partir d’une base de client complète d’un opérateur de télécommunications . Tout ces clients ont été active dés Janvier 16 2001 et le comportement Survie/Churn a été suivi pour les 20 prochains mois . Donc 16/01/2001 et l’origine de temps et Septembre 15 2002 et le temps d’arrêt de l’observation . Durant cette observation de 20 mois , les temps de sorties des clients à été enregistré .Pour chaque Client dans l’échantillon , une variable DUR est utilisé pour Indiquer le temps ou le client quittait , ou pour les cas « censoreés » la dernière observation durant laquelle les clients ont été observés. Les 2 calculés à partir d’un origine de temps Janvier 16 2001 et une autre variable Status qui distigue les cas censorés des cas observés . il est commun d’avoir Srarus =1 pour les cas observé et Status = 0 pour les cas censorés . dans cette étude , l’analyse de survie est

<Rightly censored>( signifiant que le client a quitté l’expériences avant sa terminaison ou que l’événement fondamentale se produit) donc tout les cas consorés ont une valeur de 21 mois pour la variable DUR.

**Sources de données**

Il existent 4 sources majeurs de données pour cette étude :  « census block level marketing » et informations financières, les informations sur le niveau démographique des clients , les données internes des clients et leur enregistrements de contact et appels . les 2 premières sources de données sont d’une partie tierce . une bref description suit

**Les données Internes des clients**

Ces données provienne du fournisseur télécom , elle consiste on 2 parties majeurs . la première est à propos les informations sur le client comme la chaine du marché , type de plan , paiement , type de facturation , status du contrat , code pour la segmentation des clients , être propriétaire d’autres produits , dispute , les frais de retards , les remises , promotions , dispute sur la facturation , et encore plus .

La 2éme parie est les information internes sur l’usage des données . Exemples de variables :

-Nombre Moyen d’appel comptés

-la longueur moyenne des appels

-le pourcentage d’appel change d’une période de temps une autre

-le partage de minute durant le temps de pointe/hors de pointe

-Usage des produits Add-on

**Les Enregistrement des Contacts de clients**

L’entreprise maintient un suivi des enregistrements détaillé des contacts des clients , ceci inclue leurs appels aux centre de service . leur appels au System IVR(serveur vocal interactif), l’accès web du client à son compte en ligne . les enregistrements sont ainsi classifié en catégories d’enregistrement .Parmi ces catégories :

Le questionnement générale provenant des clients, les demandes de changement de service , les questions à propose l’annulation.

**Processus de modélisation**

Le processus de modélisation inclut 5 3tapes Majeurs :

1.Exploration analytique des donnés (EDA) :

Cette exploration est conçu pour préparer les données pour l’analyse de survie . la fréquence univarié de l’analyse a été utilisé pour déterminer les valeurs de distributions , valeurs manquantes et valeurs aberrantes . La transformation des variables a été faite pour des raison comme la réduction de l’imbalance de données (skewness) car les transformation sont utiles pour amélioré l’apprentissage des données par le modèle . les valeurs aberrantes sont filtrées de façon à exclure les valeurs extrêmes qui suggérés de ne pas etre incluses. Ce filtrage tend à donner des modèles meilleures car les paramètres données sont plus stables . Variables avec des valeurs manquantes ne sont pas sévères . la démographie des variable avec plus de 25% de valeurs manquantes ont été remplacé par leurs indicateurs manquants .Pour les observations incomplètes , un choix est d’utiliser des données incomplétes mais ça peut mener à l’ignoration des données utiles des variables avec des valeurs non manquantes. Ça peut aussi perturber le biais des échantillons comme les observations n’ayant pas de valeurs manquantes. Donc , dans cette étude , les valeurs manquantes seront remplacés avec les méthodes appropriés . Pour les variables en intervalle , les valeurs remplaçantes seront calculés en se basant sur des quantiles de la distribution de la variable . d’où les valeurs qui ont été assignés de base dans la distribution de probabilité avec les valeurs non manquantes . Les valeurs Manquantes pour les variables de class sont remplacés avec les valeurs les plus fréquentes(nb occurrence (count)ou le mode)

2-La Réduction de variable – en commençant avec 328 variables dans le jeu de données originale et en utilisant PROC FREQ , une analyse univariée de toute les variables catégoriques croisée avec le status churn (variable STATUS) a été faite pour déterminer le signification statistique de l’exploration des données à inclure dans la prochaine étape de modélisation . Toute les variables catégoriques avec valeur chi² ou statique de test t de signification à 0.05 ou moins ont été retenus . cette étape réduit le nombre de variables à 225(&VARLIST1-je suppose que c’est l’adresse de la liste des variables-) . Ceci inclut les variables numériques et les variables catégoriques gardées de l’étape une .

La prochaine étape est d’utiliser PROC PHREG pour réduire encore le nombre de variables . une méthode de sélection par étape a été utilisé pour créer le modèle final avec signification statistique des effets de 42 variables des clients churn sur le temps.

PROC PHREG DATA = SASOUT2.ALL2 OUTEST =

SASOUT2.BETA;

MODEL DUR\*STATUS(0) = &VARLIST1

/ SELECTION = STEPWISE SLENTRY = 0.0025 SLSTAY

= 0.0025 DETAILS;

RUN;

**3-Estimation du modéle :**

Avec seulement 42 variables indépendantes , le jeu de données a un nombre raisonnable de variables pour faire l’analyse de survie avec des méthodes de régression. Avant d’appliquer les procédures aux données finales , la fonction de survie des clients et la fonction de dranger (hazard) ont été estimé en utilisant le code suivant . le but d’estimer la fonction de survie du client et la fonction de hazard (figure 1 et 2) est d’obtenir des connaissances des caractéristiques de survie et churn .

Pour la forme de la fonction de hazard , survie/churn du client dans cette étude démontre une fonction de hazard typique de modèle de type Log-Normal.

Comme discuté auparavant comme les formes des distributions de survie et fonctions de hazard étaient connus d’avance , PROC LIFEREG est attendu de produire des estimations plus efficaces ‘avec erreur standard plus petite’ que PROC PHREG.

PROC LIFETEST DATA = SASOUT2.ALL3 OUTSURV =

SASOUT2.OUTSURV

METHOD = LIFE PLOT = (S, H) WIDTH = 1 GRAPHICS;

TIME DUR\*STATUS(0);

RUN;

Fig 1 probabilité de survie

Fig 2

L’étape final est d’estimer la survie/churn des clients .PROC LIFEREG est utilisée pour calculer la probabilité de survie de clients . A cette étape, le jeu de données finale a été divisé en moitiés en 2 jeux de donnés : le jeu de donnée model et jeu de donnée de validation . La première est utilisé pour l’apprentissage et l’autre pour donner un score à la probabilité de survie de chaque client . Une variable Use est utilisée pour distinguer le modèle jeu de donnée ( mettre Use=0) et le set de validation (mettre Use =1). Dans le 2éme jeu de donnée , mettre DUR et STATUS comme manquants pour que les cas utilisés dans de validation du jeu de donnée sont à ne pas utilisé dans l’estimation du modèle et le jeu de donnée de validation .

IF RANUNI(0) < 0.5 THEN OUTPUT MODEL;

ELSE OUTPUT VALID;

DATA SASOUT2.\_MODEL;

SET MODEL;

USE = 0;

DATA SASOUT2.\_VALID;

SET VALID;

STATUS = .;

DUR = .;

USE = 1;

DATA SASOUT2.APPEND;

SET SASOUT2.\_MODEL

SASOUT2.\_VALID;

En Ajoutant(Appending) le jeu données de validation à celle de d’entrainement comme fait ci-dessus, le code SAS suivant permet de donner un score à la probabilité de survie de chaque client dans le jeu de donnée de validation.

Le macro PREDICT produit les probabilité de survie de chaque client dans un temps spécifié ( pour cet exemple le nombre de mois après l’origine du temps),en se basant sur le modèle entrainée par PROC LIFEREG . voir la référence « Allison> 1995 pour plus de détails sur ce macro .

PROC PHREG est aussi utilisée pour calculer la probabilité de survie prédite. En les comparant avec ceux de PROC LIFERE ,la procédure PHREG ne donne pas une qualité aussi bonne que celle LIFEREG

(voir cod)

%MEND PREDICT;

PROC LIFEREG DATA=SASOUT2.APPEND OUTEST=A;

MODEL DUR\*STATUS(0) = &VARLIST1 /

DIST=LNORMAL;

OUTPUT OUT=B XBETA=LP CONTROL =USE;

RUN;

%PREDICT(OUTEST=A,OUT=B,XBETA=LP,TIME=03)

4-Validation du modéle :

Avec chaque client dans la validation du jeu de donnée recevant un score pour la probabilité de survie durant une période spécifique d’intervalle de temps (le nombre de mois dés l’origine du mois- 16 Janvier 2001). Chaque client aura de P1 à P21 probabilités de survie prédites.

P1 est la probabilité de survie prédite à la fin du mois 1 , P2 ….. fin mois 2 et ainsi de suite. En triant les clients dans l’ordre descendants des probabilités de survie pour un interval de temps donnée.Si le modèle fonctionne , les clients avec la plus haute probabilité de survie auront la plus grande chance de survire à un temps spécifiques .

Pour Valider le succés du modèle , on trie les probabilités de survies pour une période donnée en deciles et en la comparant le nombre de clients survivants jusqu’à un temps spécifique pour chaque déciles .

Par exemple , on fait le trie des probabilité de survie jusqu’à P3 en déciles et compter le nombre de clients qui ont serveçus jusqu’au mois 3.Fig3 décrit le gain augmenté cumulatif (cumulative lift gains) . Seuls 3 périodes sont reportés à un intervalle de temps donné.Par exmple , les 3 premiers déciles capturent environ 74% des survivants jusqu’au mois 3 , tandis que les 5 top déciles capturent 92% des survivants durant le mois 3.

5-Estimation de la durée de vie d’un client :

Avec le modèle estimant avec succés la courbe de survie des clients ( une séries de probabilité de survie de P1 à P21 par exemple) la valeur de la durée de vie d’un client peut etres estié pour chaque client.

On suppose un client A sortant ayant la valeur marginale par mois MM pour les 3 derniers mois de $22.5 et un traux de réduction r de 8.5% . à partir du modèle , la série de probabilités de survie prédites : .000, 0.990, 0.972, 0.958,

0.939, 0.922, 0.906, 0.888, 0.871, 0.858, 0.841, 0.826,

0.812, 0.798, 0.792, 0.787, 0.782, 0.777,0.772, 0.767,

0.762, 0.757, 0.752, 0.747 et la valeur du client sur la durée de vie est 423.05

On suppose un autre client B avec MM de $45 pour les 3 derniers mois avec meme taux de réduction , la série de probabilité de survies est 1.000, 0.995, 0.982, 0.012, 0.002, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000,

0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000,

0.000, 0.000, 0.000, 0.000, 0.000. en utilisant la formule décrite auparavant .

D’où la valeur de la durée de vie de ce client est de $133.65

A partir de la valeur de la durée de vie, il est clair que le client A significativement plus de valeur pour le fournisseur que B . même si B se montre plus profitable que A

Fig 3

**Implémentation :**

la valeur de la durée de vie d’un client peut etres implimenté avec des façons multiples , elle peut etre utiliser comme un outil de la segmentation de la valeur d’un client . elle peut aussi etre utilisépour segmenter les clients par leurs comportement churn . D’où la valeur de la durée de vie d’un client peut etre utilisé pour mesurer l’efficacité d’une compagne de Marketing .

Les résultats de la courbe de survie des clients estime qu’on peut créer plusieurs stratégies pour étendre la durée de vie attendu de l’individu

Conclusion

Cette étude présente une application du modélisation de la valeur de la durée de vie d’un client . la valeur nette du profit des des clients calculée sur plusieurs mois en utilisant les techniques d’analyse de survie

[c’est un copier-coller de l’abstrait au début du document]

La valeur de durée de vie est un indicateur puissant et clair qui mesure synthétiquement la profitabilité et le risque d’arrêt d’achats au niveau individuel de la profitabilité des clients . Pour les clients existants , la durée de vie attendue peut permettre les entreprise à développer la loyaleté de sa clientèle et le traitement de stratégies maximisant la valeu de leur clientèle. Pour les clients nouvellement acquis ,cette durée de vie aide les entreprise à choisir les bons clients à cibler pour le futur .

La valeur de durée de vie dans cette étude est essentiellment basé sur la valeur à « un seul produit » . elle peut etre étendu pour inclure des probabilités de vente croisées pour estimer la valeur de la durée de vie d’un cleint.