経営課題にAIを!ビジネスインフォマティクス研究会 (第8回)
JSAI Special Interest Group on Business Informatics (SIG-BI #8)

クラウド会計ソフトにおける顧客離脱予測モデルの構築

Churn prediction on cloud accounting software

竹内亮介[1] *
Ryosuke TAKUECHI[1]

[1] freee 株式会社
[1] feeee K.K.

Abstract: freee K.K. provides a cloud accounting software service as subscription model under which a customer can use software over a period. Preventing customer churn is important in subscription business. This paper presents a prediction model which predicts whether each customer will quit the service or not, taking into account our insight that customers who don't use functions much tend to quit the service. The model's inputs are whether the customers have used each function or not. A classifier is ensemble SVM. Experimental results are that precision, recall and F-measure are approximately 80%. The high accuracy of the prediction model enables us to identify customers who might quit the service in 70 days and take emergency measures to beforehand.

1 はじめに

freee 株式会社はクラウド会計ソフトの会計 freee をサブスクリプション方式で提供している。サブスクリプション方式とは、ソフトウェアの利用形態の一つで、利用者は 1 ヶ月や 1 年など一定期間サービスを利用できる権利を購入し、一定期間経過後に顧客が利用契約を更新する方式である。顧客が利用を中止することを離脱と言う。サブスクリプション方式では一般に顧客獲得コストを顧客が支払う使用料から月々のサービス提供コストを引いた額で埋め合わせていくため、顧客が離脱せずに長い期間使用していくことがサービス提供者にとって重要である（図 1)。

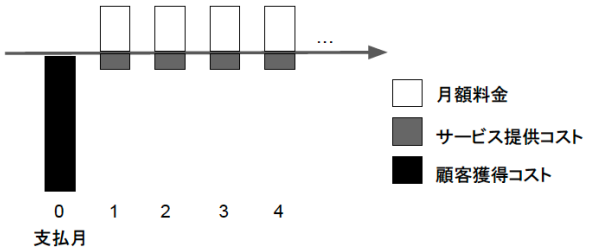

図 1: サブスクリプション方式のビジネスモデル

顧客の離脱は売り上げの低下など経営に大きな影響を及ぼすため、離脱を防ぐために各業界について予測モデルの研究が行われている。Hung [1] は台湾の通信電話業界における顧客離脱についてデータマイニングにより予測している。離脱する顧客は離脱しない顧客に比べて少ないことを考慮して、Yu [2] は One-Class SVM により通信会社の顧客離脱を分析し、Yaya [3] は balanced random forests と weighed random forests を組み合わせた Improved balanced random forests を提案し銀行業界の顧客離脱について分析している。Wouter [4] はルールを推論する AntMiner+と, ブラックボックスである SVM からわかりやすいルールを導出していく Active Learning Based Approach(ALBA) を用いて、KDD library にある通信業界に関するデータセットについて分析している。Li [5] は通信会社のオペレーターの品質向上のため、データ分析基盤を整え、顧客離脱モデルを構築し、実データを用いて評価をしており、分析手法でだけでなく予測基盤全体について報告している。Tsai [6] は方法について、予測モデル構築前に相関ルールによって入力を選択する手法を提案し、Multimedia on demand ビジネスに対して適用している。入力としてログイン時間、支払い料金、誕生日などを挙げている。

このように分析対象、および手法について様々な研究が行われており、一般的に難しい問題である。ここでは、クラウド会計ソフトにおける顧客離脱の予測モデルについて報告する。

*連絡先：freee 株式会社
〒 141-0031 東京都品川区西五反田 2-8-1 五反田ファーストビル
E-mail: takeuchi@freee.co.jp

2 予測モデル

高精度な予測モデルを構築するには、適切な入力を選択することが重要である。顧客は会計 freee をうまく使えていないと離脱しやすい、との定性的な知見があるため、本報告ではこの知見を考慮した上で入力を選択して顧客離脱モデルを構築する。

予測内容は、各顧客について基準日から M 日時点で N 日までの離脱是非である。今回は基準日を顧客が料金を支払った日、支払い後 30 日時点で支払い後 100 日までの離脱是非を予測する。理由は、離脱する顧客は比較的支払後早い時期に離脱しているため、30 日間で予測に使用できる十分なデータが集まるためである。

予測モデルについて述べる。入力は主に顧客の機能利用是非である。会計 freee はクラウドサービスで、各顧客のブラウザ上での機能の利用履歴は記録されている。初めてその機能を利用した日を初利用日とする。各機能について、初利用日が M 日時点より前ならばその機能を利用しており、初利用日が M 日以降もしくは利用していないならば機能を利用していない、の 2 値を割り当てた。

また、Li [5] では料金プランを予測モデルの入力としている。会計 freee では年額と月額の 2 つの料金体系があるため、この料金体系も入力とした。加えて、ユーザの属性も入力とし、計 256 の変数を入力とした。

例を表 1 に示す。顧客 2 は支払い後 30 日前に離脱してしまったため対象から除く。顧客 4 の機能 A 初利用日は支払い後 30 日より後のため利用していないとする。また、顧客 4 は 100 日より後に離脱しているためで予測モデルに与えるデータとしては離脱していないとみなす。

表 1: 機能の初利用是非の例

顧客 ID	初支払い日	離脱日	機能 A 初利用日
1	04/01	-	04/10
2	04/01	04/15	-
3	04/01	05/15	-
4	04/01	10/01	09/01

⇓

顧客 ID	離脱是非	機能 A 利用是非
1	F	T
3	T	F
4	F	F

識別器は SVM のアンサンブル学習である。まずデータセットから基準日から M 日以前に離脱した顧客を除く。訓練データのうち、90%をランダムにサンプリングする。会計 freee では離脱する顧客よりも離脱しない顧客の方がはるかに多いため、このまま予測モデルを構築すると、いかなる入力についても離脱しないと予測してしまう。そこで離脱した顧客数と同じだけ、離脱しない顧客をランダムに選択してダウンサイジングし、離脱した顧客と離脱していない顧客のサンプル数を同一としたデータセットを作成する。そのデータセットを用いて弱学習器として SVM を作成し、各サンプルの離脱是非を予測する。弱学習器数を L とし、(L+1)/2 回以上離脱と予測されれば離脱、(L+1)/2 回未満ならば離脱しない、とする。

予測モデルの評価基準は運用上の観点で重要な、適合率 (離脱と予測した顧客のうち実際に離脱した顧客の割合)、再現率 (実際に離脱した顧客のうち離脱と予測された顧客の割合)、F 値とする。何故ならば、離脱すると予測された顧客に対して freee が対応する場合、適合率が大きいならば離脱しない顧客に対して対応するコストが低くなる。再現率が大きいならば、離脱を防ぐことができる可能性が高まる。加えて、適合率と再現率のバランスを見るため F 値を見る。

3 実験

上記の予測モデルを使用して、今回行った実験方法について述べる。実験内容は (1)「利用」の定義の違いによる予測モデルの精度評価、(2) 実験に用いるデータの期間の長さによる予測モデルの精度評価、の 2 つである。

(1) の実験意図について述べる。機能の利用は、(a) 顧客が各機能のトップページ到達後にいずれかのサブ機能を実行した場合、(b) 顧客が各機能のトップページ到達後にその機能の典型的なサブ機能を利用、例えば登録、を行った際にその機能を利用したとする場合、の 2 つを定義とする。つまり、(b) の方が機能利用とされる条件が厳しい。(a) と (b) の結果を比較し、離脱する/しない顧客がどのように機能を利用しているのかについて明らかにする。

(2) の実験意図について述べる。会計 freee ではデータが日々追加されていく一方、マーケティングやカスタマーサポートなどのオペレーションも変化しているため、最新の顧客に対して過去のデータを利用して予測モデルを構築しても、現状の使い方に合っていない可能性がある。また、最新のデータのみを使用すると、サンプルが少なすぎて精度が下がる恐れがある。予測モデル構築に使用する適切なデータ期間を明らかにする。今回は 150 日から 30 日ごとに 360 日までをデータ期間とした。

評価方法について述べる。10 分割交差検証により評価する。まず対象となる全データを訓練データと試験データに分割する。訓練データで上記通り SVM のアンサンブル学習モデルを作成し、試験データについて

顧客の離脱是非を予測し、実結果と比較し、適合率、再現率、F 値を算出する。10 回の平均を精度とした。

識別器の各パラメータについて述べる。弱学習器の個数 L は、増えれば増えるほど計算時間が増加し、計算資源を使用してしまうため、予測精度と運用のコストを考慮して L を 31 とした。SVM のカーネルはガウシアンである。SVM のパラメータの γ とコストは、グリッドサーチによるパラメータチューニングの結果をもとに固定した。今回はパラメータを微調整しても大きく結果が変化せず、毎回パラメータチューニングを行うと運用コストが増大するためである。

4 結果

図 2 に実験結果を示す。塗りつぶしなしの点線が (a)、塗りつぶしありの実線が (b) で、黒色の丸、赤色の三角、緑色の四角がそれぞれ適合率、再現率、F 値である。再現率は 120 日を除き (a) と (b) で大きな差はなかった。適合率は多くの期間で (b) の方が 2 3%大きかった。F 値は適合率に影響され、(b) の方が大きくなった。

次に (b) の結果を使用データ期間の観点で見る。再現率は 210 日以降 80%を超えた一方、適合率は 300 日以降に低下した。適合率の低下に影響され、F 値も同期間に低下した。

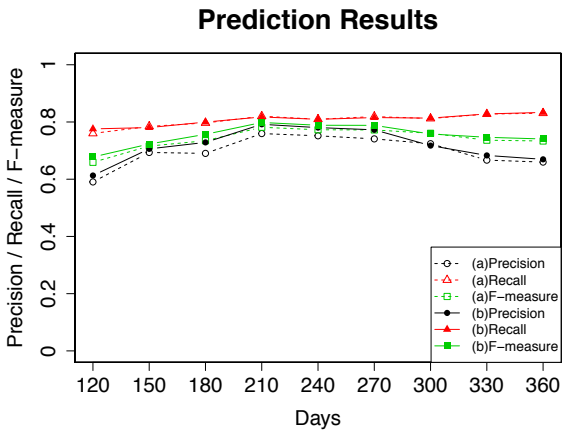

図 2: 実験結果

5 考察

まず、(1)「利用」の定義の違いによる予測モデルの精度評価の結果について考察する。各機能について典型的なサブ機能を利用している場合のみ利用したとみなす場合の方が、いずれかのサブ機能を利用したとみなす場合よりも、全体的に適合率が高くなった。(a) の方が「使用」した機能が多く、「機能をうまく使えていない」がモデルに反映されきれなかった。そのため、離脱すると予測しすぎたと考えられる。

次に、(2) 実験に用いるデータの期間の長さによる予測モデルの精度評価について述べる。会計 freee では機能が随時追加され、セールスやカスタマーサポートのオペレーションが随時変更されている。よって、古すぎるデータを使用すると、それらの変更が反映されないため、適合率が低下することを確認できた。

この予測モデルは一定期間のデータについて各機能の利用是非を 2 値で入力としている。より高精度な予測モデルとして、以下の改善が考えられる。まず、実験 (1) の結果から「典型的な」機能を定義する方がよいとの結論が得られたが、「典型的」の定義については改善の余地がある。次に、一度でも機能を利用した場合利用したと見なしているのを 1 回利用したかどうかの 2 値から、一定回数以上かどうかの 2 値に変更、もしくは使用回数を整数値として扱うようにすることが考えられる。何度も使用された機能があれば顧客は離脱しにくくなる可能性がある。3 つ目として、大幅に機能やオペレーションが変更された場合、適切な期間を見つける必要がある。

6 まとめ

クラウド会計ソフトの会計 freee はサブスクリプションモデル、つまり利用者が料金を支払うと一定期間利用できる形式で提供されている。サブスクリプションモデルでは支払い後の顧客離脱率をいかに低減させるかがビジネス上重要である。顧客が離脱するのかを予測するのかは一般的に難しい問題である。今回機能を使いこなしていない顧客は離脱しやすいとし、過去データを用いて、入力として顧客の機能の利用状況など、出力として期間内の離脱是非、SVM のアンサンブル学習による顧客離脱予測モデルを作成したところ、適合率、再現率、F 値がそれぞれ 80%程度となった。

経営指標には未来を予測する先行指標と、問題があったことを示す遅行指標がある [7]。一般に顧客の離脱は遅行指標であり対応ができなかった。しかし、この顧客離脱予測モデルを使用することで、離脱と予測された顧客について対応を取ることが可能となり、先行指標とすることができる。今後は顧客への対応を行い、実際に離脱を防ぐことができるようになったのかを調べていく。

参考文献

[1] Shin-Yuan Hung, David C. Yen, Hsiu-Yu Wang: Applying data mining to telecom churn management,

Expert Systems with Applications, Vol. 31, Issue 3, pp. 515–524 (2006)

[2] Yu Zhao, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren: Customer Churn Prediction Using Improved One-Class Support Vector Machine, *ADMA 2005*, Vol. 384, pp. 300–306 (2005)

[3] Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying: Customer churn prediction using improved balanced random forests, *Expert Systems with Applications*, Vol. 36, Issue , pp. 5445–5449 (2009)

[4] Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens: Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert Systems with Applications*, Vol. 38, Issue 3, pp. 2354–2364 (2011)

[5] Hui Li, Di Wu, Gao-Xiang Li et al.: Enhancing Telco Service Quality with Big Data Enabled Churn Analysis: Infrastructure, Model, and Deployment, *Journal of Computer Science and Technology*, Vol. 30, Issue 6, pp. 1201–1214 (2015)

[6] Chih-Fong Tsai, Mao-Yuan Chen: Variable selection by association rules for customer churn prediction of multimedia on demand, *Expert Systems with Applications*, Vol. 37, Issue 3, pp. 2006–2015 (2010)

[7] アステリア・クロール, ベンジャミン・ヨスコビッツ: Lean Analytics, pp.16 (2015)