ПРОГНОЗНЫЕ ЗАДАЧИ В CRM-СИСТЕМАХ

Шуляк Е. В., Гуринович А. Б., Смирнов В. Л.
Кафедра ИТАС, Кафедра ВМиП, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: yegorshulak@gmail.com, gurinovich@bsuir.by

Оценка качества обслуживания клиентов, анализ предпочтений и прогнозирование представляет собой необходимую деятельность по созданию полноценно функционирующей и имеющей высокие показатели системы продаж. В данной работе была использована платформа Cloudera Hadoop и были определены оптимальные модели анализа данных.

Введение

Для исследования выбраны банковские маретиговые данные, используемые для решения прогнозных задач в телефонном маркетинге. Они подходят для машинного обучения путем создания единого классификатора. Фактически этот набор данных содержит 150 столбцов, но их количество сокращается до 22 в результате экспертного анализа.

I. Основные входные параметры

Информация представлена в CSV – текстовом формате для представления табличных данных. Основные входные параметры следующие.

- Данные банковских клиентов:
  - age_job: стаж работы;
  - categorical_job: тип работы;
  - marital: семейное положение;
  - education: образование;
  - default: наличие кредита в принципе;
  - housing: наличие кредита на жильё;
  - personal_loan: наличие личного кредита.
- Данные, связанные с последними контактами текущей кампании:
  - contact: тип связи (сотовая или стационарная);
  - month: последний месяц звонка;
  - day_of_week: последний день недели звонка;
  - duration: Продолжительность последнего звонка в секундах.
- Другие атрибуты:
  - campaign: количество контактов участников на данный момент;
  - pdays: количество дней, прошедших после последнего контакта с клиентом;
  - previous: количество контактов этого клиента, выполненных до данной кампании;
  - poutcome: результат предыдущей маркетинговой кампании (по категориям).
- Атрибуты социального и экономического контекста:
  - emp.var.rate: коэффициент изменения занятости – квартальный показатель;
  - cons.price.idx: индекс потребительских цен – ежемесячный индикатор;
  - cons.conf.idx: индекс доверия потребителей – ежемесячный индикатор;
  - euribor3m: Европейская межбанковская ставка предложения – ежедневный индикатор;
  - nr.employed: индекс занятости – квартальный показатель;
- Выходная переменная (желаемая цель):
  - y – Подписался ли человек на срочный депозит? ('да', 'нет')

Необходимо объяснить, что этот набор данных предоставлен исследованием Моро и др.[1] об успехах или неудачах при рекламе долгосрочных кредитов в банках. Он состоит из данных, собранных с 2008 по 2013 год. В нём был использован полуавтоматизированный выбор функций для изучения лучшего моделирования. «Полуавтоматизированным» его можно назвать потому, что на некоторых шагах были использованы экспертные суждения, основанные на интуитивном знании бизнеса, в частности, было использовано 14 вопросов для группировки важности параметров.

II. Анализ

Итак, после предварительной обработки данных, набор чист. Мной будут использованы 4 модели интеллектуального анализа данных:

Логическая регрессия (LR) – статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой.[2]

Дерево принятия решений (DT) – средство поддержки принятия решений, использующееся при анализе данных для прогнозных моделей. Структура дерева представляет собой «листья» и «ветки». На «ветках» записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи.

Нейронная сеть (NN) – математическая модель, а также её программное воплощение, построенная по принципу организации и функцио-

нирования биологических нейронных сетей – сетей нервных клеток живого организма.[3]

Метод опорных векторов (SVM) – набор схожих алгоритмов использующихся для задач классификации и регрессионного анализа. Идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющая гиперплоскость максимизирует расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.[4]

Для определения лучшей модели с лучшими параметрами применим два следующих подхода: ROC-анализ (англ. receiver operating characteristic, рабочая характеристика приёмника), использующий график, позволяющий оценить качество бинарной классификации, отображающий соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак. Количественную интерпретацию метода даёт показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) – площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций.[5] И LIFT-анализ (англ. lift, подъёмная сила), показывающий, насколько повышается вероятность нахождения T в анализируемом случае, если в нем уже имеется A. Количественную интерпретацию данного метода даёт показатель ALIFT – площадь, ограниченная LIFT-кривой и осью доли ложных положительных классификаций.[6]

III. Графики

На графиках показаны кривые относительно показателей Ложных положительных оценок (FPR) и Правдивых (TPR). Кривая логистической регрессии обладает лучшими параметрами. (см. рис. 1-2)

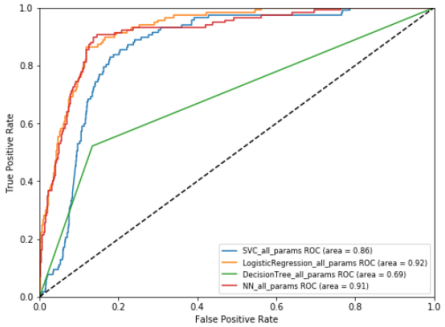

Рис. 1 – ROC-анализ

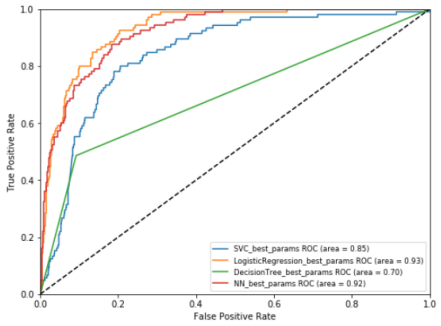

Рис. 2 – LIFT-анализ

«Продолжительность», «возраст», «Европейская межбанковская ставка предложения», «работа» и «образование» занимают первые четыре места в топе показателей важности. Одна из этих переменных – «Европейская межбанковская ставка предложения» за последние три месяц, банком не контролируется. «Продолжительность» сильно влияет на выходную цель (например, если продолжительность = 0, то ответ на вопрос, подписался ли человек на срочный депозит, очевидно, y = 'нет') но, тем не менее, продолжительность не известна до конца вызова, а кроме того, после окончания вызова ответ на вопрос уже известен. Таким образом, если намерение состоит в том, чтобы создать реалистичную прогнозирующую модель, этот параметр лучше вообще не учитывать. Следовательно, основными факторами являются социально-экономические.

IV. Выводы

В исследовании анализировались результаты работы нескольких моделей. До этого проводилась предварительная обработка и выбор функций. Расчёты показали, что оптимальными для решения прогнозных задач в CRM-системах являются логическая регрессия и нейронная сеть.

V. Список литературы

1. S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.
2. Harrell, Frank E. (2001). Regression Modeling Strategies (2nd ed.). Springer-Verlag.
3. Паклин Н.Б., Орешков В.И. Глава 9. // Бизнес-аналитика: от данных к знаниям(+CD): Учебное пособие. 2-е изд.. — СПб: Питер, 2013. — С. 428-472.
4. Владимир Вьюгин. Математические основы теории машинного обучения и прогнозирования. — МЦМНО, 2013. — 390 с.
5. «A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems». Machine Learning 45: 171-186.
6. Tufféry, Stéphane (2011); Data Mining and Statistics for Decision Making, Chichester, GB: John Wiley & Sons.