Санкт-Петербургский Государственный Университет»
Кафедра Системного Программирования

Сулягина Анастасия Александровна

Оптимизация предсказания оттока абонентов оператора сотовой связи

Курсовая работа

Научный руководитель:
ведущий разработчик ООО «НМТ» Константин Невоструев

Санкт Петербург, 2015

Содержание

Введение

Каждый год провайдеры телекоммуникационных услуг терпят убытки из-за оттока абонентов. Отрасль очень конкурентна, компании регулярно выдвигают все более выгодные предложения, и клиенты, предпочитающие высокое качество за низкую цену, переходят от одного оператора к другому. Годовой отток клиентов телекоммуникационных компаний в среднем составляет 25%, что немало.

Так как привлечение новых клиентов в несколько раз дороже, чем удержание старых, предотвращение оттока абонентов особо привлекательно для изучения. Точные предсказания оттока абонентов позволяют менеджерам телекоммуникационных компаний применять стратегии удержания клиентов, экономя большое количество средств.

С развитием методов машинного обучения точность прогнозов значительно возросла, что сделало возможным для компаний внедрять предсказывающие системы. Имея точные предсказания, оператор может своевременно предпринять необходимые шаги для удержания абонента: предложить лучший тариф, отправить спецпредложение или бонусы.

Для предсказания оттока надо решить задачу бинарной классификации абонентов, то есть разделения их на две группы: останутся они или уйдут. Наиболее хорошие результаты на данный момент показывают такие методы машинного обучения как градиентный бустинг, нейронные сети, случайный лес решающих деревьев.

В представленной работе описан процесс разработки классификатора, предсказывающего отток абонентов крупного российского мобильного оператора, также приведено его сравнение с другими существующими решениями и оценка его точности.

Предметная область

Машинное обучение

Классификация - разбиение множества объектов на несколько классов

Классификатор - модель машинного обучения, выполняющая задачу классификации

Выборка - набор данных с определенными классами

Ансамбль классификаторов - сложная модель машинного обучения, полученная комбинированием различных классификаторов

Оценка эффективности

Положительный класс - уходящие абоненты

Отрицательный класс - остающиеся абоненты

1. **True positive (TP)** - верно определенные классификатором в положительный класс
2. **True negative (TN)** - верно определенные классификатором в отрицательный класс
3. **False positive (FP)** - неверно определенные классификатором в положительный класс
4. **False negative (FN)** - неверно определенные классификатором в отрицательный класс

Precision - отношение (TP / (TP + FP))

Recall - отношение (TP / (TP + FN))

ROC кривая - кривая, показывающая отношение TP к FP

AUC - площадь под ROC кривой над прямой случайного угадывания (FP = TP)

Accuracy - процент верных предсказаний

Обзор

Предсказание оттока абонентов - актуальная задача, поэтому существует множество работ на эту тему. Для выбора наиболее перспективных моделей машинного обучения и оценки существующих решений мной были рассмотрены следующие работы:

Автор	**Использованные решения**
V. Umayaparvathi, K. Iyakutti [1]	Деревья решений, Нейронные сети
Mozer MC, Wolniewicz R [2]	Логистическая регрессия, Нейронные сети
Chih-Ping Wei, I-Tang Chiu [3]	Деревья решений
Hung, Shin-Yuan and Yen [4]	Нейронные сети, Деревья решений, K-Means – кластеризация
М.Корыстов [5]	Градиентный бустинг, ансамбли классификаторов

Работа Максима Корыстова *«Применение методов машинного обучения для предсказания поведения абонентов оператора сотовой связи»* заслуживает особого внимания, так как именно ее результаты необходимо улучшить.

Лучший результат прошлого года

	precision	**recall**	**AUC**
Ансамбль классификаторов	0.75	0.66	0.90

Данные

Работа была проведена с данными, предоставленными мобильным оператором «Мегафон». Были выгружены данные на 50000 абонентов в промежутке 15 месяцев, содержащие следующую информацию:

- количество минут входящих вызовов
- количество минут исходящих вызовов на городские номера в пределах области подключения
- количество минут исходящих вызовов на мобильные номера прочих мобильных операторов в за пределы области подключения
- количество минут исходящих вызовов на мобильные номера прочих мобильных операторов в пределах области подключения
- количество минут исходящих вызовов на данного оператора за пределы области подключения
- количество минут исходящих вызовов на данного оператора в пределах области подключения
- количество минут исходящих вызовов по междугородней связи
- количество минут исходящих вызовов по международной связи
- количество мегабайт потребленного интернет трафика
- количество отправленных СМС
- некоторые персональные данные

Данные для каждого пользователя были преобразованы во временные ряды по 3 месяца, были введены новые признаки, показывающие динамику изменения предпочтений пользователя.

Постановка задачи

Цель

Цель данной курсовой работы - повышения качества предсказаний классификатора по сравнению с предыдущей работой [5]. Для достижения данной цели необходимо решить следующие задачи:

- Провести анализ существующих решений
- Разработать классификатор, предсказывающий уход абонентов на основе выгруженных данных
- Настроить классификатор для улучшения результата
- Попробовать различные способы группировки данных
- Оценить точность разработанных моделей

Инструменты

- Язык программирования - Python
- Библиотеки
  - scikit learn и xgboost для построения и настройки классификаторов
  - theano и lasagne для написания нейронной сети
  - pandas для работы с данными

Обучение

Построение модели

Классификаторы, реализованные в процессе работы

1. **Решающее дерево**

   Дерево, в листах которого находятся атрибуты, а в остальных узлах - признаки, по которым классифицируется объект

2. **Логистическая регрессия**

   Простейшая модель машинного обучения, считает линейную функцию от признаков в класс объекта

3. **Случайный лес решающих деревьев**

   Ансамбль, состоящий из решающих деревьев, в узлах которых находятся случайные признаки. Предсказания деревьев комбинируются для получения более точного результата

4. **Метод k ближайших соседей**

   Модель машинного обучения, в которой класс объекта определяется голосованием k ближайших к нему объектов тренироврчной выборки

5. **Градиентный бустинг**

   Ансамбль, состоящий из решающих деревьев, в котором каждое следующее дерево добавляется с подсчетом градиента таким образом, чтобы скорректировать ошибку уже существующего ансамбля.

6. **Нейронная сеть**

   Модель машинного обучения, состоящая из соединенных и взаимодействующих перцептронов - искусственных нейронов. Перцептроны передают друг другу сигналы и обучаются на их основе

Для сравнения были выбраны градиентный бустинг и нейронная сеть, также были реализованы ансамбли классификаторов

1. **Ансамбль на основе случайных лесов и градиентного бустинга, объединенных логистической регрессией**

   Каждая модель обучается на части данных, делает свое предсказание, предсказания усредняются и на них обучается логистическая регрессия.

2. **Ансамбль на основе случайного леса и метода k ближайших соседей**

   Случайный лес обучается на исходных данных и передает свою конфигурацию в метод k ближайших соседей

Была проведена настройка выбранных классификаторов для улучшения результатов и кросс-валидация для избежания переобучения.

Настройка классификаторов

1. **Градиентный бустинг -** были настроены такие параметры как количество деревьев, количество признаков для обучения и вероятностный порог
2. **Нейронные сети -** настроены такие параметры как количество слоев, их порядок, для каждого слоя был определен тип и размер.
3. **Ансамбли**

   3.1. Основанный на деревьях решений - настроены такие параметры как тип взаимодействующих классификаторов, их количество, для каждого было определено количество деревьев, функция подсчета потерь.

   3.2. Основанный на методе ближайших соседей на основе случайного леса - настроено количество соседей и количество деревьев в лесе

Группировка данных

Пользователи были сгруппированы по возрасту, длительности пользования оператором, предпочитаемому типу связи, юридическое или физическое лицо. Наиболее хорошие результаты получились для пользователей следующих групп:

- 40+ лет
- юридические лица
- < двух лет с оператором
- пользующиеся интернетом

Имея такие результаты и большее количество данных, имело бы смысл реализовать классификатор, обучающийся на данных группах абонентов отдельно.

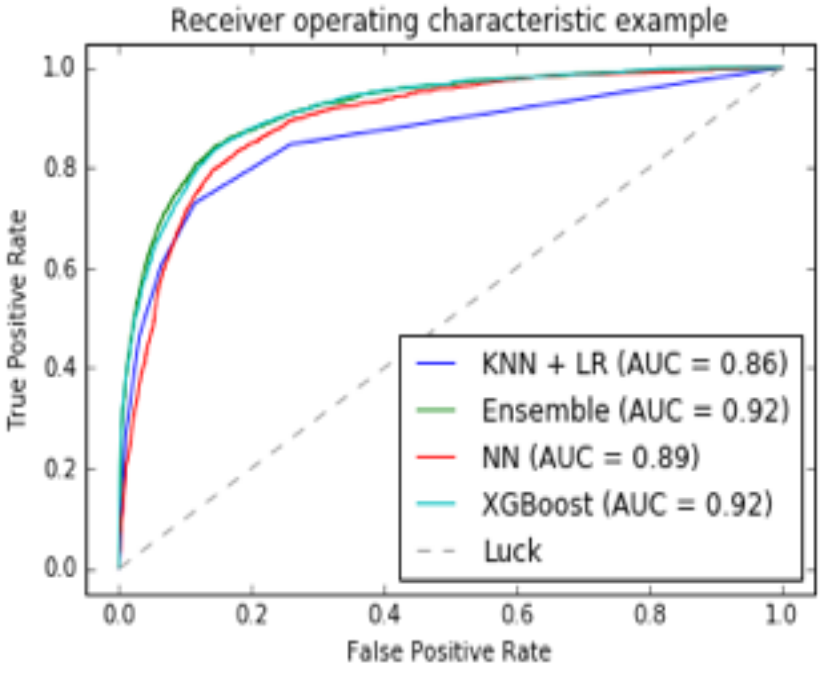
Результаты

Классификация

Сравнение классификаторов было проведено при помощи представленных в таблице метрик и на основании графика ROC. Для данной задачи наиболее важные метрики - precision, recall и AUC, так как они в большей степени, чем accuracy учитывают правильность отнесения объекта к положительному классу, в то время как доминирующий в выборке класс - отрицательный.

Модель	**precision**	**recall**	**AUC**	**accuracy**
Ансамбль*	0.75	0.72	0.92	0.88
Градиентный бустинг(XGB)	0.74	0.69	0.92	0.87
Нейронная сеть	0.70	0.62	0.89	0.86
Ближайшие соседи + случайный лес	0.74	0.60	0.86	0.87

График ROC иллюстрирует метрику AUC, видно, что лучшие результаты показывает ансамбль и XGBoost



* *Состоит из двух RandomForest, двух ExtraTrees с разными функциями качества и одного бустинга*

Улучшение результатов

С помощью ансамбля классификаторов, состоящего из случайных лесов и градиентного бустинга, улучшен результат прошлого года:

	precision	**recall**	**AUC**
Моя работа	0.75	0.72	0.92
Работа [5]	0.75	0.66	0.90

Группировка

Группа	**precision**	**recall**	**AUC**	**accuracy**
юр.лицо	0.81	0.93	0.87	0.81
физ.лицо	0.70	0.59	0.90	0.90
> 2 лет с оператором	0.75	0.58	0.86	0.94
< 2 лет с оператором	0.74	0.79	0.86	0.78
< 40 лет	0.71	0.63	0.86	0.90
> 40 лет	0.79	0.77	0.93	0.89
пользуются интернетом	0.74	0.73	0.92	0.88
пользуются только связью	0.75	0.69	0.91	0.89

Итоги

В рамках данной работы были изучены существующие решения в области предсказания оттока пользователей, выбраны и реализованы классификаторы, показывающие наибольшую точность при решении данной задачи. Был проведен сравнительный анализ реализованных моделей с помощью метрик precision, recall, ROC AUC и accuracy. Параметры моделей были настроены для улучшения результатов. Также были проведены эксперименты по группировке данных, которые показали перспективность реализации классификатора, обучающегося на разных группах абонентов по отдельности и комбинирующего предсказания при наличии большего объема данных.

Наилучший результат был получен с помощью ансамбля, основанного на комбинации различных случайных лесов, градиентного бустинга и линейной регрессии.

precision	**recall**	**AUC**
0.75	0.72	0.92

Результат прошлого года улучшен.

Список литературы

[1] V Umayaparvathi and K Iyakutti. Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications, 42(20):5–9, 2012.

[2] Michael C Mozer, Richard Wolniewicz, David B Grimes, Eric Johnson, and Howard Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. Neural Networks, IEEE Transactions on, 11(3):690–696, 2000.

[3] Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. Expert systems with applications, 23(2):103–112, 2002.

[4] Shin-Yuan Hung, David C Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. Expert Systems with Applications, 31(3):515–524, 2006.

[5] Максим Корыстов. Применение методов машинного обучения для предсказания поведения абонентов сотовой связи. 2015