Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Санкт-Петербургский государственный университет»

Кафедра Системного Программирования

Курбанов Рауф Эльшад оглы

Модели машинного обучения для предсказания ухода пользователей

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:
д. ф.-м. н., профессор Терехов А. Н.

Научный руководитель:
д. ф.-м. н., профессор Терехов А. Н.

Рецензент:
технический директор ООО "Лаборатория Анализа Данных" Натёкин А. Г.

Санкт-Петербург
2015

SAINT-PETERSBURG STATE UNIVERSITY

Department of Software Engineering

Rauf Kurbanov

Machine learning models for user churn prediction

Bachelor's Thesis

Admitted for defence.
Head of the chair:
professor Andrey Terekhov

Scientific supervisor:
professor Andrey Terekhov

Reviewer:
CTO "Data Mining Labs .llc" Alexey Natekin

Saint-Petersburg
2015

Оглавление

Введение

В современном бизнесе растёт тенденция в сторону решений, ориентированных на данные. Такие решения существовали всегда, однако с течением времени рос объём накопленной информации в руках компаний и развивались направления исследований и подходы программной инженерии нацеленные на структурированные подходы к данным. Машинное обучение заняло место современной области знаний, выступающей в роли инструмента для создания новых решений старых проблем.

Машинное обучение (Machine Learning) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения.

Машинное обучение — не только математическая, но и практическая, инженерная дисциплина. Чистая теория, как правило, не приводит сразу к методам и алгоритмам, применимым на практике. Чтобы заставить их хорошо работать, приходится изобретать дополнительные эвристики, компенсирующие несоответствие сделанных в теории предположений условиям реальных задач. Практически ни одно исследование в машинном обучении не обходится без эксперимента на модельных или реальных данных, подтверждающего практическую работоспособность метода. [3]

Данная работа посвящена проблеме, надиктованной современными направлениями в бизнесе, которая успела стать классической в области машинного обучения: задаче предсказания ухода пользователей. Проблема предсказания оттока актуальна в широком спектре областей таких как телекоммуникации, маркетинг, финансы и даже медицина.

Мы рассмотрим существующие подходы к предсказанию оттока пользователей, пройдём полный путь от обработки данных до построения предсказывающей модели и реализуем прототип системы предсказания ухода на реальных данных по широкополосному доступу в интернет.

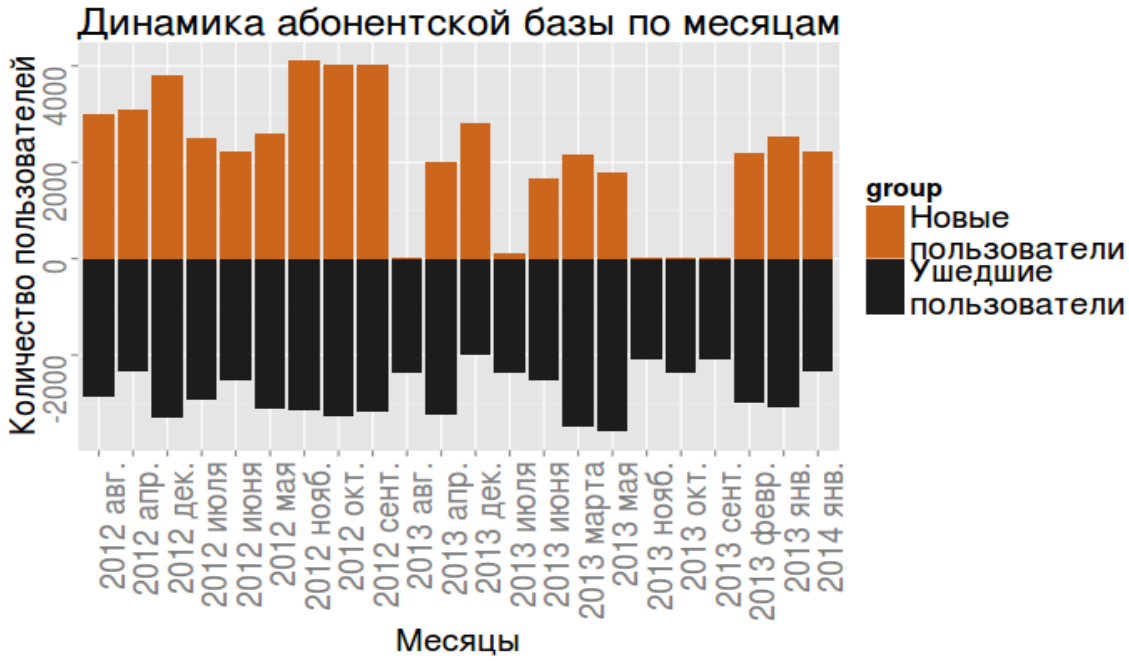

Рис. 1: Активность пользователей ШПД по месяцам.

0.1. Актуальность проблемы

Актуальность проблемы оттока диктуется прежде всего требованиями бизнеса. В сфере хорошо сформировавшихся и высококонкурентых рынков, когда возможности привлечения новых клиентов начинают иссякать, вопрос удержания пользователей становится центральным. Также стоит учитывать тот факт, что удержание пользователя обходится для компании в разы дешевле привлечения нового.

Наиболее ярко проиллюстрировать актуальность задачи можно уже на первых этапах анализа предоставленных нам данных данных об абонентах. Можно наблюдать на Рис. 3 неутешительную тенденцию в динамике количества пользователей. Стоит также отметить, что со временем не только растёт масса пользователей покинувших провайдера, но и сокращается прирост новых, тем самым даже не компенсируя существующий отток. В таком положении дел компании задача предсказания оттока пользователей особо критична.

0.2. Постановка задачи

Сформулируем цели данной работы.

- Провести подготовку данных о пользователях широкополосного доступа в интернет. В этом пункте подразумевается очистка данных, построение множества признаков, выделение целевой переменной для задачи классификации.

- Сравнить наиболее подходящие модели для предсказания ухода и выбрать лучшую для нашей задачи. Модифицировать выбранную на этапе сравнения модель. Требуются как модификации на уровне модели, нацеленные на повышение точности и проверки корректности результата предсказаний, так и сугубо инженерные решение, нацеленные на оптимизацию работы конкретной системы.

- Реализовать прототип системы для предсказания ухода абонентов для провайдера широкополосного доступа в интернет. Получить предсказания ухода на текущих данных с достойными показателями результирующий метрик.

1. Обзор решений

Имеет место большое количество публикаций, посвящённых предсказанию ухода пользователей. Однако, при внимательном прочтении, становится ясно, что данные работы имеют некоторую специфику. Попробуем объединить существующие решения по некоторым признакам, чтобы показать, какая важная ниша образовалась в контексте существующих работ.

Существует некоторый пласт работ [5][2] в центре внимания которых стоит предсказание ухода именно в телекоммуникационной сфере. Фокус таких публикаций смещён в сторону аспектов специфичных для данной области, но не конкретной задачи. В этих работах акцентировано внимание на данных о пользователях известных в телекоммуникационной области, причём пространство признаков в выбранных работах достаточно мало, что сильно сказывается на общности предлагаемых решений.

Стоит упомянуть, что умеют место публикации, в которых напротив посвящены применения конкретной модели к задаче оттока, как например в работе [1], посвящённой применению методу опорных векторов в задаче предсказания ухода. Такая работа должна быть акцентирована, на описываемой модели и зачастую упускает специфику задачи, что может привести к снижению точности предсказаний.

В нашей работе мы займём пустую нишу в существующем спектре решений: подробно опишем цикл разработки решения для задачи ухода пользователей, подробно опишем шаги анализа данных и опишем путь подбора модели к данному конкретному решению, опирающийся на установленных особенностях в данных. Мы возьмём за основу подход задачи оттока методом классификации, однако постараемся нивелировать недостатки перед строгими статистическими подходами за счёт стадии предварительного анализа.

2. Обработка данных

В первоначальном виде данные представляют собой таблицу из 257 столбцов и 6500000 записей. Записей об абонентах широкополосного доступа в интернет 4500000. Данные представляют собой многомерный лог, фиксирующий данные о пользователе в момент оплаты счёта. Записи имеют вид:

Номер договора	Номер месяца	Количество дней с предыдущего платежа	Информация об абоненте	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Для каждого пользователя имеет место несколько таких записей. Мы будем работать с подмножеством исходного набора данных, удовлетворяющим условию репрезентативности.

Ввиду количественной избыточности и практической сложности интерпретации данных, используем лишь информацию о тех пользователях, для которых имеется возможность учесть временной контекст, уловив тем самым динамику в их поведении. Поскольку последние три месяца клиент может быть фактически неактивен, будем рассматривать клиентов с хотя бы шестимесячной историей, чтобы иметь по крайней мере три корректные точки для обучения.

2.1. Выделение целевой переменной

Будем считать абонента ушедшим, если он не пополнял баланс своего счёта в течение более чем трёх месяцев. Однако, поиск месяца, когда абонент официально признан ушедшим не имеет смысла, так как фактически это значит, что пользователь уже 3 месяца неактивен, а значит возможности удержать его уже нет. Следовательно целевой переменной будем считать последний месяц, в который пользователь будет активен, то есть совершит последний платёж. Будем предсказывать уход на месяц вперёд, таким образом получаем следующие два целевых класса:

- 0: продолжит пользоваться услугами оператора.
- 1: покинет оператора в следующем месяце

Поскольку в нашем случае ушедших пользователей намного меньше и классы не уравновешены, то будем считать класс 1 целевым.

2.2. Построение признаков

Не смотря на то, что в данных имеется более 250 полей для каждого абонента, изначальное пространство было увеличено более чем в 5 раз.

К изначальному множеству переменных, содержащих информацию о пользователе следует добавить признаков, учитывающих динамику поведения абонента по времени. Для этого, по каждому оставшемуся полю в данных X построим следующие признаки:

- X.diff: плавающие разности соседних значений признака X.
- X.sum.3: плавающие суммы значений признака X за 3 дня.
- X.max.3 плавающий максимум значений признака X за 3 дня.
- X.mean.3: плавающее среднее значения признака X за 3 дня.
- X.mean3.d: плавающее среднее трёх соседних значений X.diff.
- X.max.3.d: плавающий максимум трёх соседних значений X.diff.

3. Метрики для классификации

3.1. Точность и полнота

Для задачи распознавания с неуравновешенными классами используются такие метрики как точность и полнота. Для описания метрик введём некоторые обозначения.

- true positive: количество истинно положительных результатов.
- true negative: количество истинно отрицательных результатов.
- false positive: количество ложно положительных результатов.

Точность (precision) определяется как отношение числа корректно классифицированных элементов к общему числу элементов:

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{1}$$

Полнота (recall) это отношение числа найденных элементов целевого класса, к общему числу элементов целевого класса:

$$\text{recall} = \frac{\text{true positive}}{\text{true positive } + \text{false negative}} \tag{2}$$

3.2. F-мера

F-мера ($F - measure$, $F_1 - score$) позволяет объединить точность и полноту в одной усреднённой величине. Для этой цели среднее арифметическое не подходит, так как, например, модели для классификации достаточно отнести все элементы к целевому классу, чтобы обеспечить равную единице полноту при близкой к нулю точности, и среднее арифметическое точности и полноты будет не меньше 1/2. Среднее гармоническое не обладает этим недостатком, поскольку при большом отличии усредняемых значений приближается к минимальному из них. Поэтому хорошей мерой для совместной оценки точности и полноты является

F-мера, которая определяется как взвешенное гармоническое среднее точности и полноты:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{3}$$

[6]

3.3. Площадь под ROC-кривой

ROC-кривая (англ. receiver operating characteristic, рабочая характеристика приёмника) — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных, как несущих признак, (англ. true positive rate, TPR, называемой чувствительностью алгоритма классификации) и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных, как несущих признак (англ. false positive rate, FPR, величина 1-FPR называется специфичностью алгоритма классификации) при варьировании порога решающего правила.

Количественную интерпретацию ROC даёт показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). [7]

4. Сравнение моделей

Были получены предсказания с помощью набора наиболее перспективных моделей для нашей задачи. Мы добились высокий результирующих метрик с помощью всех рассмотренных моделей, такой точности мы добились за счёт грамотно построенного пространства признаков. Однако лучше предсказания были получены с помощью модифицированной машины градиентного бустинга, детали реализации которого опишем в следующей главе.

Модель	ACC	PREC	RECALL	F1	AUC
xgboost	0.99350	0.99893	0.99986	0.99652	0.99993
gbm	0.99757	0.98829	0.99038	0.98933	0.99757
random forest	0.99890	0.99839	0.99199	0.99518	0.99589

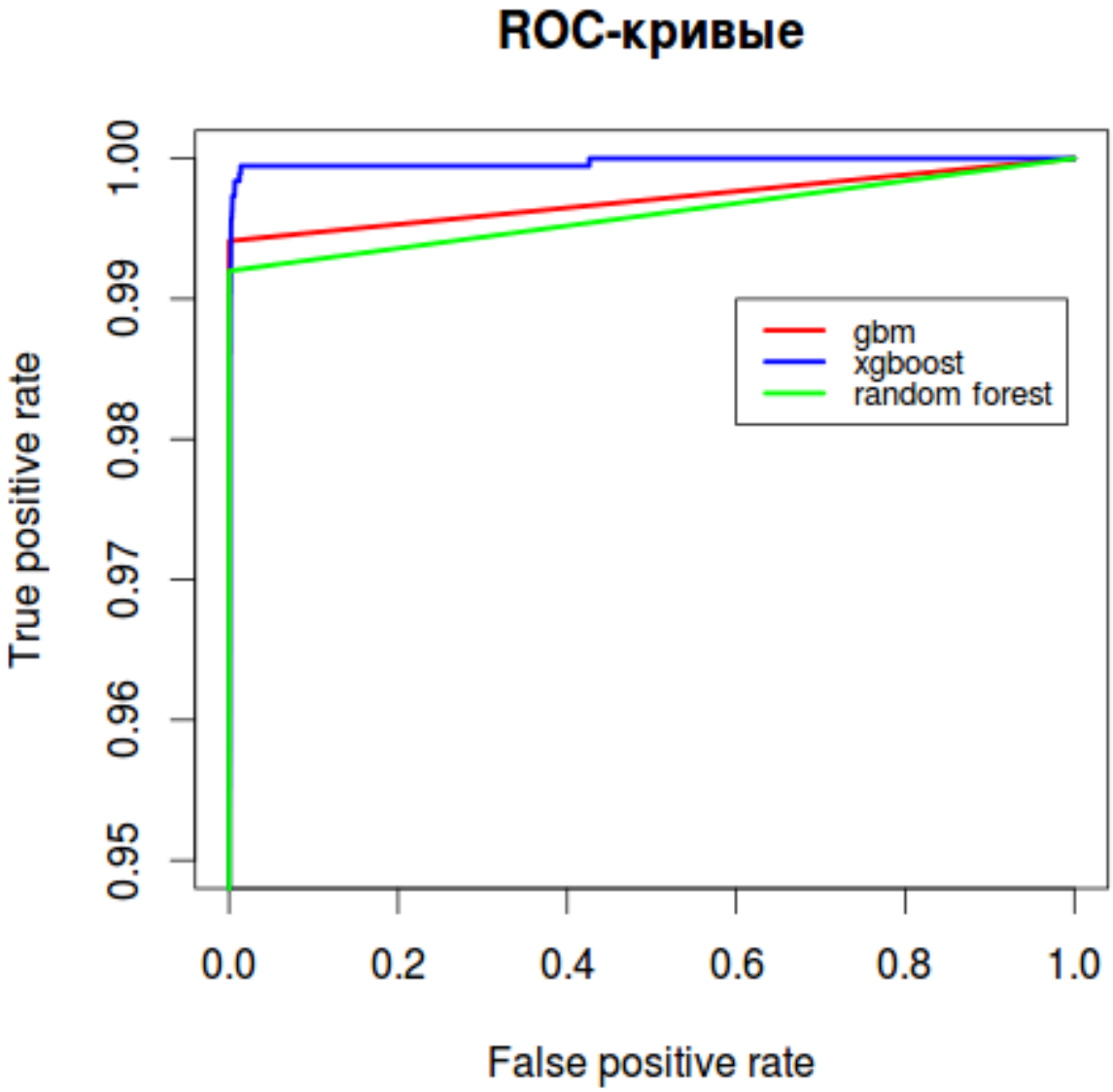

Рис. 2: Сравнение моделей.

5. Интерпретация модели

При решении задач предсказания на векторах с высокой размерностью и сложных зависимостях между параметрами мы как правило не может построить модель исходя из теории или каких-либо априорных предположений. Именно такие ситуации встречаются нам чаще всего в реальной жизни.

Однако в таких условиях нам приходят на помощь непараметрические модели машинного обучения, которые строятся исключительно по данным. При таком подходе от исследователя требуется заранее подготовить пространство параметров, с чем мы успешно справились на стадии обработки данных. Теперь следует преступить к интерпретации полученной модели и по возможности сократить количество параметров, не допустив при этом серьёзных потерь точности.

В итоговой модели для предсказания ухода наиболее значимы следующие признаки:

- Максимум трёх соседних разностей "BIT_TOTAL_CHARGE"
- Среднее трёх соседних разностей "BIT_TOTAL_SPENT"
- Максимум трёх соседних значений "BIT_SPENT"
- Максимум трёх соседних значений "BIT_C_ABONPAY"
- Сумма трёх соседних значений "BIT_SPENT"
- Сумма трёх соседних значений "BIT_C_ABONPAY"
- Максимум трёх соседних значений "BIT_CHARGE"
- Среднее трёх соседних разностей "BIT_C_ABONPAY"
- Максимум трёх соседних разностей "BIT_C_ABONPAY"

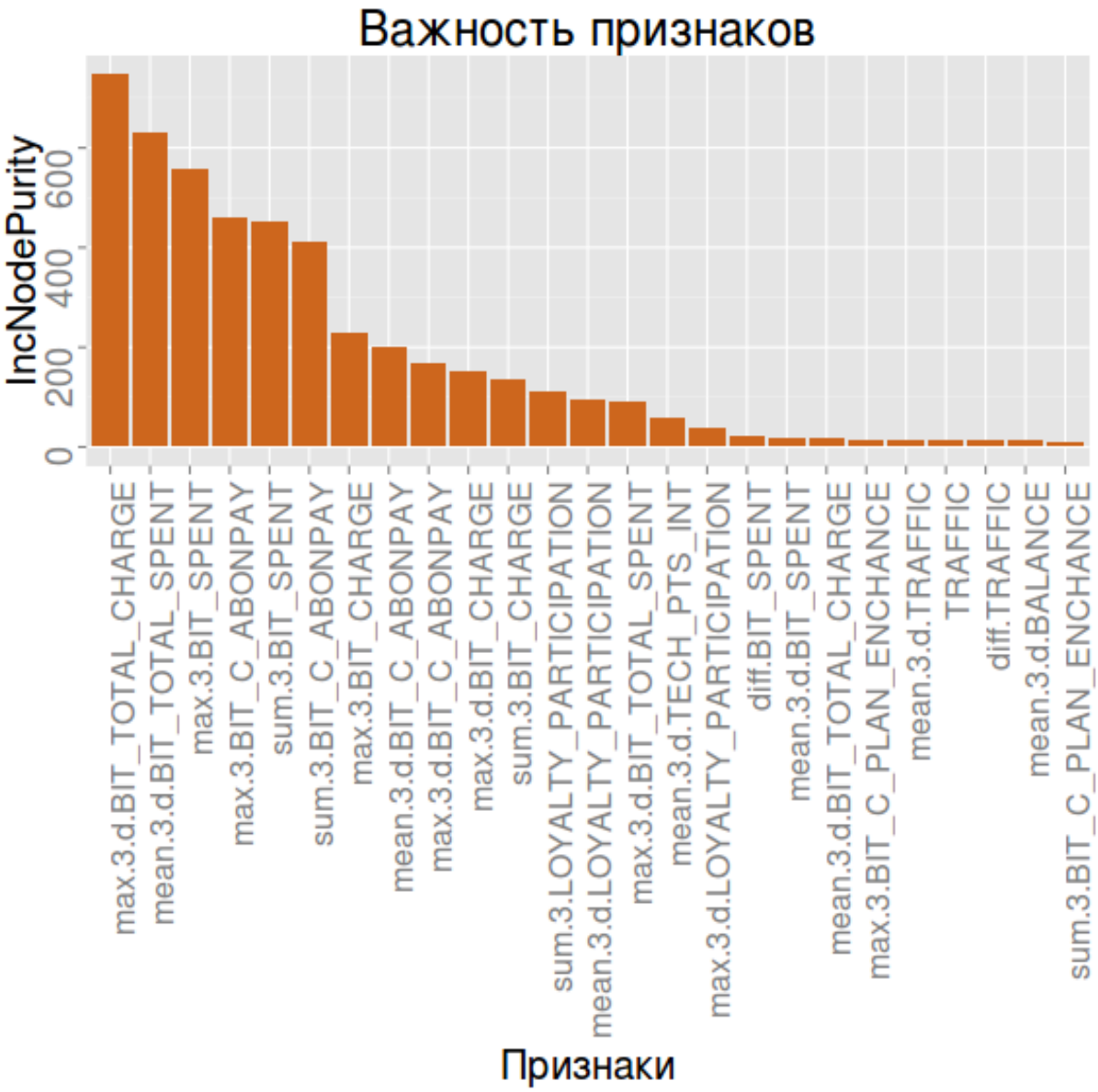

Рис. 3: Важность признаков.

6. Реализация

После того как мы выбрали модель для предсказаний и завершили обработка данных стоит упомянуть некоторые моменты реализации. При разработке подобной библиотеки стоит помнить, что она является прототипом будущей высоконагруженной системы, поэтому было решено использовать в реализации модифицированную реализацию модели.

Заказчик имеет большую базу данных, которая быстро обновляется, поэтому возникает желание улучшить производительность системы за счёт параллелизма. Для распараллеливания процесса обучения градиентного бустинга хорошо подходит библиотека с открытым исходным кодом XGBoost.

XGBoost: eXtreme Gradient Boosting - это оптимизированная реализация машины градиентного бустинга общего назначения. В существующей реализации представлено небольшое количество решающих классификаторов, однако, присутствует классификатор над деревьями решений, лучше всего подходящий для многомерных данных с аномалиями. Пакет XGBoost позволил нам добиться кроссплатформенного параллелизма, так как производит вычисления с помощью как на операционной системе Linux, так и на Windows с помощью инструментария OpenMP.

Также нам предоставляется возможность достичь выигрыша в использовании памяти за счёт информации о структуре данных. Мы знаем, что подающиеся на вход модели данные разрежены, то есть в них имеется большое количество нулей и пустых ячеек. Здесь нам опять же приходят на помощь возможности пакета XGboost, предоставляющего оптимизированные структуры для хранения разреженных данных.

6.1. Модуль кросс-валидации

Переобучение (переподгонка, пере- в значении «слишком», англ. overfitting) в машинном обучении и статистике — явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительно плохо работает на примерах, не участвовавших в обучении (на

примерах из тестовой выборки). [8]

Для того чтобы избежать переобучения модели, добавим функцию автоматической кросс-валидации. Скользящий контроль или кросс-проверка или кросс-валидация (cross-validation, CV) — процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам Фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. [4]

В реализуемом модуле автоматической кросс-валидации для нашей системы будем пересчитывать площадь под ROC-кривой метрику на тренировочном и тестовом множестве на каждой итерации градиентного бустинга. Таким образом мы сможем остановить процесс на своевременной итерации предотвратив переобучение, а также получить объективную оценку точности.

В результате всех модификаций нами разработан прототип системы, предоставляющий интерфейс для предсказания ухода пользователей на потоке данных нашего заказчика. Разработанный прототип эффективно удовлетворяет требованиям заказчика, оптимизирован по времени обучения и затратам памяти, безопасен с точки зрения объективности предсказаний.

Заключение

В результате данной работы достигнуты следующие результаты:

- Проведена обработка данных о пользователях широкополосного доступа в интернет.
- Выбрана и модифицирована эффективная модель для предсказания оттока.
- Реализован прототип системы для предсказания ухода абонентов для провайдера широкополосного доступа в интернет. Получены предсказания на реальных данных.

Список литературы

[1] Coussement Kristof, Van den Poel Dirk. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques // Expert systems with applications. — 2008. — Vol. 34, no. 1. — P. 313–327.

[2] Huang Bingquan, Kechadi Mohand Tahar, Buckley Brian. Customer churn prediction in telecommunications // Expert Systems with Applications. — 2012. — Vol. 39, no. 1. — P. 1414–1425.

[3] MachineLearning.ru. Машинное обучение // MachineLearning.ru. — 2015. — URL: `http://goo.gl/nGykSy` (дата обращения: 27.05.2015).

[4] MachineLearning.ru. Скользящий контроль // MachineLearning.ru. — 2015. — URL: `http://goo.gl/bOeaFr` (дата обращения: 27.05.2015).

[5] Wei Chih-Ping, Chiu I-Tang. Turning telecommunications call details to churn prediction: a data mining approach // Expert systems with applications. — 2002. — Vol. 23, no. 2. — P. 103–112.

[6] Wikipedia. Precision and recall // Wikipedia The Free Encyclopedia. — 2015. — URL: `http://goo.gl/3wL3xN` (дата обращения: 27.05.2015).

[7] Wikipedia. ROC-кривая // Википедия, свободная энциклопедия. — 2015. — URL: `https://goo.gl/eFWzdu` (дата обращения: 27.05.2015).

[8] Wikipedia. Переобучение // Википедия, свободная энциклопедия. — 2015. — URL: `https://goo.gl/RY655n` (дата обращения: 27.05.2015).