Содержание

9 Тематический пример: отток клиентов

Есть только один босс — клиент. И он может уволить всех в компании, начиная с председателя, просто потратив деньги в другом месте.

Сэм Уолтон

Acme Telephonica (AT) — оператор мобильной связи, у которого есть клиенты во всех штатах США. Как и любая телекоммуникационная компания, AT борется с **оттоком** клиентов к другим операторам. Компания постоянно ищет новые способы решения проблемы **оттока** и в 2008 г. основала **службу удержания клиентов**, которая отслеживает количество звонков, поступающих в центр поддержки клиентов AT от каждого клиента, и выявляет клиентов, которые делают большое количество звонков, интерпретируя их как признак риска оттока. Служба удержания клиентов связывается с этими клиентами и делает им специальные предложения, стремясь сохранить их для компании AT. Однако этот подход оказался не особенно успешным, и за последние пять лет наблюдается неуклонное увеличение оттока.

В 2010 г. компания AT наняла Росса, специалиста по аналитическому прогнозированию, чтобы реализовать новый подход к снижению оттока клиентов. В этом тематическом исследовании описывается работа, выполняемая Россом для компании AT с помощью процесса CRISP-DM[1], чтобы выработать решение на основе интеллектуального анализа данных для этой бизнес-задачи. В настоящей главе мы обсудим каждый этап процесса CRISP-DM в этом проекте.

9.1. Понимание бизнеса

Как и в большинстве проектов по интеллектуальному анализу данных, компания AT обратилась к Россу не с готовым решением, а с бизнес-проблемами. Поэтому первой целью Росса было преобразование этой бизнес-задачи в конкретное аналитическое решение. Прежде чем выполнить это преобразование, Росс должен был полностью понять бизнес-цели компании AT. Это было достаточно просто,

[1] *См. раздел 1.5.*

поскольку руководство АТ заявляло, что их цель — снижение уровня оттока клиентов. Единственным фактором, который остался невыясненным, была величина этого сокращения. Основываясь на опыте предыдущих проектов, текущем подходе к удержанию клиентов, принятом в компании АТ, и ее исторических данных, Росс согласился с руководством АТ, что целевое сокращение с нынешнего максимума, равного 10%, до примерно 7,5% было бы реалистичным и, вероятно, достижимым. Росс указал руководству АТ, что, пока он фактически не изучил данные, он не может знать, насколько полезную модель сможет построить.

Следующей задачей Росса была полная оценка текущей ситуации в компании АТ. В частности, он должен был понять текущее состояние компании и ее готовность принять меры в ответ на идеи, которые он предложит. В компании АТ уже работала служба удержания клиентов, которая предпринимала активные меры для сокращения оттока клиентов. Кроме того, эта служба уже использовала данные, собранные в организации, чтобы выбирать целевых клиентов на основе моделей интеллектуального анализа данных.

Росс провел значительное время, беседуя с Кейт, руководителем службы удержания клиентов, чтобы понять, как они работают. Кейт объяснила, что в конце каждого месяца формировался список звонков, по которому выявлялись клиенты, сделавшие больше трех звонков в службу поддержки клиентов за предыдущие два месяца. Эти клиенты считались подверженными риску оттока в следующем месяце, поэтому служба удержания клиентов обращалась к ним со специальным предложением. Как правило, этим предложением было снижение тарифа в течение следующих трех месяцев, хотя сотрудники службы удержания клиентов могли свободно делать и другие предложения.

Росс также поговорил с Грейс, главным техническим директором АТ, чтобы выяснить имеющиеся ресурсы данных. Росс узнал, что в компании АТ существуют достаточно сложные транзакционные системы для регистрации последних звонков и биллинговой информации. Исторические записи о звонках и счетах, а также демографическая информация о клиентах содержались в хранилище данных. Грейс сыграла значительную роль в разработке процесса сбора информации о контактах клиентов со службой удержания. Росс надеялся, что это облегчит его задачу, потому что Грейс была главным хранителем всех ресурсов данных в АТ и ее поддержка проекта была бы важной. Кроме того, Росс провел довольно подробные беседы в отделах биллинга, продаж и маркетинга, а также сетевого управления.

На ранних этапах проекта Росс целенаправленно **вникал в ситуацию.** Беседуя с командой управления АТ, Кейт и Грейс, он много узнал о индустрии мо-

бильной связи. Основная структура бизнеса компании заключалась в том, что у клиентов был контракт на услуги, предоставляемые ею. Эти контракты не имели фиксированного срока действия и существенно изменялись каждый месяц, когда клиент платил ежемесячную **фиксированную регулярную плату**. Периодически внося плату, клиент получал **пакет минут** по сниженному тарифу. Делая разные регулярные платежи, клиенты получали пакеты разного объема. Если клиент использовал все время в своем пакете, в следующий раз ему предлагались **дополнительные минуты**, которые, как правило, были дороже, чем минуты, включенные в пакет. В компании AT все звонки разделялись на **пиковые** и **непиковые**. Пиковым было время с 8:00 до 18:00 с понедельника по пятницу, и звонки, сделанные в пиковое время, были дороже, чем остальные.

Основываясь на своей оценке текущей ситуации в компании AT, Росс разработал вопросы, ответы на которые могли бы помочь решить проблему оттока клиентов.

- **Какова общая стоимость клиента для компании**? Можно построить модель для прогнозирования общей стоимости, которую компания AT, скорее всего, получит от конкретного клиента за время его существования. Этот показатель может быть использован для идентификации клиентов, которые в настоящее время не выглядят ценными, но которые позже, вероятно, станут более ценными (часто в эту категорию попадают студенты). Предлагая этим клиентам стимулы для предотвращения их оттока, в будущем компания AT получила бы полную отдачу от них.
- **Какие клиенты, скорее всего, покинут компанию в ближайшем будущем**? Можно обучить модель прогнозирования выявлять клиентов из клиентской базы компании AT, которые, скорее всего, откажутся от ее услуг в ближайшем будущем. Служба удержания клиентов может сосредоточить свои усилия на этих клиентах. Процесс, который служба удержания клиентов AT использовала в начале проекта для идентификации клиентов, которые, вероятно, откажутся от услуг компании, был основан только на одном признаке — она просто подсчитывала, сколько звонков клиент сделал в службу поддержки клиентов AT. Вероятно, модель машинного обучения, которая рассматривала бы несколько признаков, могла бы лучше определить клиентов, которые могут покинуть компанию.
- **Какое предложение по удержанию будет лучше всего соответствовать конкретному клиенту**? Можно было бы построить систему для прогнозиро-

вания того, какое предложение из набора возможных предложений по удержанию лучше всего подходит конкретному клиенту. Это поможет службе удержания клиентов убедить больше клиентов оставаться с компанией АТ.

- **Какие части сетевой инфраструктуры, скорее всего, выйдут из строя в ближайшем будущем**? Используя информацию о нагрузках на сеть, использовании сети и диагностике оборудования, можно было бы построить прогностическую модель, чтобы предсказать предстоящие сбои оборудования, чтобы можно было предпринять превентивные действия. Сетевые сбои являются фактором неудовлетворенности клиентов и в конечном счете приводят к их оттоку, поэтому снижение их количества может оказать положительное влияние на уровень оттока.

После совещания с руководством компании АТ было решено, что наиболее подходящим для анализа решением было бы предсказать, *какие клиенты, скорее всего, откажутся от услуг компании в ближайшем будущем*. Был установлен ряд причин, по которым был выбран этот проект.

- В ходе предыдущих бесед Росса с Грейс, техническим директором АТ, было установлено, что данные, необходимые для построения модели прогнозирования оттока, скорее всего, будут доступны.
- Модель прогнозирования может быть легко интегрирована с текущими бизнес-процессами АТ. У компании АТ уже была служба удержания клиентов, которая проводила активные действия, чтобы предотвратить отток, хотя и использовала очень простую систему для определения контактов с клиентами. Создав более сложную модель для идентификации этих клиентов, можно улучшить этот существующий процесс.
- Построение модели прогнозирования оттока заинтересовало руководителей АТ, поскольку они надеялись, что объяснение основных причин оттока клиентов поможет снизить его уровень. Кроме того, лучшее понимание основных факторов оттока клиентов было бы полезным для многих других подразделений АТ.

Напротив, для реализации других аналитических решений либо недоставало доступных данных (например, компания АТ не располагала данными об успехе или неудаче различных предложений по удержанию клиентов), либо требовались слишком значительные изменения бизнес-процессов (например, вычисление прогнозируемой общей стоимости клиента) или основные предположения

были недостаточно обоснованы (например, что отток клиентов сильно зависит от сетевых сбоев).

После того как было определено аналитическое решение, следующим шагом было согласование ожидаемой эффективности новой аналитической модели. Основываясь на оценке недавних исторических данных, руководство AT полагало, что на момент начала проекта их текущая система идентификации клиентов, которые могут отказаться от услуг компании, имела точность около 60%, поэтому любая вновь разработанная система должна была бы работать значительно лучше, чем существующая. По согласованию с руководством компании AT и службой удержания клиентов Росс согласился с тем, что его целью было создать систему прогнозирования оттока, которая обеспечивала бы точность прогнозирования более 75%.

9.2. Понимание данных

В процессе определения того, какое аналитическое решение было бы наиболее подходящим для текущей ситуации в компании AT, Росс уже начал анализировать доступные ресурсы данных. Его следующая задача заключалась в том, чтобы добавить к этому пониманию гораздо бо́льшую глубину, следуя процессу, описанному в разделе 2.3. Чтобы понять, какие данные были доступны, в каком формате они хранятся и где, требовалось очень тесное сотрудничество с Грейс. Это понимание послужило бы основой для разработки **понятий предметной области** (domain concepts) и **описательных признаков** (descriptive features), которые составили бы **базовую аналитическую таблицу** (ABT) для создания прогностической модели. Это был итеративный процесс, в ходе которого Росс переходил от Кейт, работавшей в службе удержания клиентов AT, к Грейс, техническому директору, и другим подразделениям, которые идентифицировали данные, связанные с оттоком клиентов. Вскоре стали очевидными ключевые ресурсы данных в компании AT, которые важны для этого проекта.

- Демографические данные о клиентах из хранилища данных AT.
- Отчеты о выставлении счетов клиентов, хранящиеся в базе данных биллинга AT, которые записываются в течение 5 лет.
- Транзакционная запись звонков, сделанных физическими лицами, в течение 18 месяцев.
- Транзакционная база данных отдела продаж, содержащая подробную информацию о телефонных трубках, выданных клиентам.

- Простая транзакционная база данных удержания, содержащая все контакты с клиентами и результаты этих контактов в течение 12 месяцев

Прежде чем двигаться дальше, Росс должен был определить **цель прогноза** (prediction subject) для ABT и целевого признака. Цель состояла в том, чтобы разработать модель, которая спрогнозировала бы, откажется ли клиент от услуг компании в ближайшие месяцы. Это означало, что предметом прогнозирования в этом случае был клиент, поэтому необходимо было построить таблицу ABT, строки которой содержали бы данные о каждом клиенте.

Предсказание оттока — это форма **моделирования склонности** (propensity modeling)[2], в которой событием является отказ клиента от услуг компании. Следовательно, Росс должен был согласовать свое определение оттока с бизнес-экспертами (в частности, службой удержания клиентов). Это определение будет использоваться для выявления событий оттока в исторических данных AT и, следовательно, имеет основополагающее значение для построения таблицы ABT для проекта. Бизнес-эксперты согласились с тем, что клиент, который был неактивен в течение одного месяца (т.е. не совершал никаких звонков или оплачивал счет) или который явно отказался от услуг или не возобновил контракт, считался бы отказавшимся от услуг. Росс также должен был определить продолжительность **периода наблюдения** (observation period) и **период результатов** (outcome period) для модели. Он решил, что **период наблюдения**, в течение которого он будет собирать данные о поведении клиентов, продлится 12 месяцев. Это решение было принято на основе имеющихся данных и предположений Росса о том, что более длительный период наблюдений, вероятно, мало повлияет на прогнозирование оттока. Что касается определения **периода результатов**, то компания согласилась с тем, что было бы наиболее полезным делать прогноз, что клиент, скорее всего, откажется от услуг за три месяца до того, как это произойдет на самом деле, поскольку это дало бы им время для принятия мер по удержанию клиента. Итак, продолжительность периода результатов была установлена равной трем месяцам[3].

---

[2] *См. раздел 2.4.3.*

[3] *Очевидно, что события оттока для разных клиентов будут происходить в разные сроки, поэтому для построения ABT необходимо выровнять периоды наблюдения и результатов для разных клиентов. Эта ситуация является примером сценария моделирования склонности, приведенного на рис. 2.6 в разделе 2.4.3.*

После определения целевого признака следующей задачей Росса было определение понятий предметной области, которые лежат в основе таблицы ABT. Понятия предметной области — это те факторы, которые, по мнению бизнеса, влияют на решение клиента об отказе. Понятия предметной области были выработаны с помощью серии семинаров с представителями различных подразделений AT, в частности службы удержания клиентов, а также отдела продаж и маркетинга и отдела выставления счетов. Эксперты компании AT полагали, что основными факторами, которые влияли на отток, были основные демографические данные клиентов (например, более молодые клиенты чаще склонны к оттоку); информация о платежах клиентов и, в частности, изменения в шаблонах выставления счетов (например, возможно, клиенты, чей счет внезапно увеличился, с большей вероятностью откажутся от услуг компании); информация о телефоне клиента (например, клиенты, долго пользующиеся одним и тем же телефоном, чаще отказываются от услуг компании); клиент имел опыт взаимодействия со службами AT (например, клиенты, которые часто звонят в службу поддержки, испытывают трудности с сетью AT и, следовательно, могут отказаться от ее услуг); и данные о фактических звонках, которые делал клиент, в частности, изменение шаблонов поведения (например, клиенты, которые начали звонить новым группам людей, чаще отказываются от услуг компании). Это множество понятий предметной области достаточно велико, чтобы охватить все характеристики, которые могли бы повлиять на вероятность оттока клиента (рис. 9.1).

Рис. 9.1. Множество понятий предметной области для прогнозирования оттока клиентов компании Acme Telephonica

Из этих понятий предметной области Росс выработал набор описательных признаков. Некоторые из описательных признаков были просто копиями доступных исходных данных. Например, столбцы Возраст, Пол, Кредитоспособность и Род занятий из набора демографических данных клиента могут быть непосредственно включены в качестве описательных признаков в таблицу ABT для отражения понятия предметной области Демография клиентов. Более интересными описательными признаками были те, которые должны были быть получены из исходных источников данных. Например, Росс узнал, что служба удержания клиентов считает, что одной из главных причин, по которой клиенты отказывались от услуг компании, было наличие новых высококачественных телефонов, которые предоставлялись другими сетями. Чтобы попытаться отразить понятие предметной области Информация о телефоне, были предложены три описательных признака.

- Смартфон. Признак, указывающий, является ли текущий телефон клиента смартфоном. Определяется по самой последней записи о телефоне клиента.
- КоличествоТелефонов. Количество разных телефонов, которые клиент имел за последние три года. Этот признак определяется путем подсчета всех записей о телефонах для конкретного клиента.
- ПродолжительностьИспользованияТелефона. На основании последней записи о телефоне этот признак фиксирует количество дней, в течение которых клиент пользуется своим текущим телефоном.

В анализе оттока и в любом виде моделирования склонности изменение обычно является ключевым фактором поведения клиентов. По этой причине, а также на основе обсуждений, проведенных со службой удержания клиентов, Росс включил понятия предметной области Изменение счета и Изменение социальной сети. Служба удержания клиентов выяснила, что клиенты часто принимали решение отказаться от услуг компании, если их счет значительно увеличивался из-за изменения шаблонов поведения или когда они начинали вести продолжительные разговоры по телефону с новыми друзьями или коллегами в других сетях. По этим причинам Росс разработал следующие описательные признаки.

- ИзменениеПродолжительностиЗвонков. Этот признак, получаемый из исходных данных о звонках, фиксирует величину, на которую изменилось количество минут, использованных клиентом в текущем месяце по сравнению с предыдущим.

- ИзменениеСуммыСчета. Этот признак, полученный из исходных данных о звонках, фиксировал сумму, на которую счет клиента изменился в текущем месяце по сравнению с предыдущим.
- Кол-воНовыхНомеров. Этот признак, полученный путем анализа фактических набранных номеров, зафиксированных в исходных данных о звонках, отражает, сколько новых номеров клиент стал набирать в текущем месяце. Номер считается набираемым часто, если количество его наборов превышает 15% от общего количества звонков клиента.

Часто описательные признаки, которые могут быть очень полезными, не могут быть реализованы из-за недоступности данных. Например, команда AT почувствовала, что клиент, начинающий часто звонить в другие сети, будет хорошим индикатором оттока, но подходящий признак извлечь не удалось. В своих записях звонков компания AT не хранила информацию о том, какие вызовы совершаются в сети. Кроме того, при свободном перемещении номеров между операторами сами номера больше не являются надежным индикатором сети.

Полный набор описательных признаков, разработанных Россом, а также краткое описание каждого из них приведен в табл. 9.1.

Таблица 9.1. Описательные признаки ABT, разработанные для задачи прогнозирования оттока клиентов компании Acme Telephonica

Признак	Описание
ИзменениеСуммыСчета	Процент, на который изменился текущий счет клиента по сравнению с предыдущим
ИзменениеКоличестваМинут	Процент, на который изменилось количество минут, использованных клиентом в текущем месяце по сравнению с предыдущим
СреднийСчет	Среднемесячная сумма счета
СредняяРегулярнаяПлата	Среднемесячная регулярная плата, оплачиваемая клиентом
СреднееСнижение КоличестваЗвонков	Среднее снижение количества звонков за месяц
ИзменениеОтношения ПиковыхНепиковыхЗвонков	Изменение отношения количества пиковых и непиковых звонков в текущем месяце по сравнению с предыдущим
СреднееКоличество ПолученныхМинут	Среднее количество минут, полученных клиентом за месяц
СреднееКоличествоМинут	Среднее количество минут, использованных клиентом за месяц

Окончание табл. 9.1

Признак	Описание
СреднееКоличествоМинутСверхПакета	Среднее количество минут, использованных клиентом за месяц сверх пакета
СреднееКоличествоЗвонковвРоуминге	Среднее количество звонков в роуминге, сделанных клиентом за месяц
ОтношениеПиковыхНепиковыхЗвонков	Соотношение между пиковыми и непиковыми вызовами, сделанными клиентом в этом месяце
КоличествоНовыхНомеров	Сколько новых номеров, на которые клиент часто звонит в этом месяце?
КоличествоОбращений	Количество обращений в службу поддержки, сделанных клиентом в прошлом месяце
КоличествоПредложений	Количество предложений, сделанных клиенту службой удержания
КоличествоПринятыхПредложений	Количество предложений службы удержания, которые принял клиент
Возраст	Возраст клиента
Кредитоспособность	Кредитный рейтинг клиента
Доход	Уровень доходов клиента
ПродолжительностьОбслуживания	Количество месяцев, в течение которых клиент пользовался услугами компании AT
РодЗанятий	Род занятий клиента
ТипРегиона	Тип региона, в котором живет клиент
СтоимостьТелефона	Стоимость текущего телефона клиента
ВозрастТелефона	Возраст текущего телефона клиента
КоличествоТелефонов	Количество телефонов, которые клиент имел за последние три года
Смартфон	Является ли текущий телефон клиента смартфоном?
Отток	Целевой признак

9.3. Подготовка данных

С помощью Грейс для реализации реальных сценариев обработки данных и интеграции данных с использованием инструментов, доступных в компании AT, Росс заполнил таблицу ABT, содержащую все признаки, перечисленные в табл. 9.1. Росс отобрал данные за период 2008–2013 гг. Используя определение оттока, в соответствии с которым клиент не совершал никаких звонков или не оплачивал счет за один месяц, Росс смог выявить случаи оттока в течение этого периода времени. Чтобы собрать экземпляры клиентов, которые не ушли от компании, Росс случайно опробовал клиентов, которые не соответствовали

определению оттока, но которые также могут считаться активными клиентами. Работая с Кейт, Росс определил активного клиента в качестве текущего клиента, который делал не менее пяти звонков в неделю и был клиентом в течение как минимум шести месяцев[4]. Это определение гарантировало, что к экземплярам без оттока в наборе данных будут относиться только клиенты с относительно нормальным профилем поведения с достаточно длинной историей, по которой для них могли быть рассчитаны реалистичные описательные признаки.

Заключительная таблица ABT содержала 10000 экземпляров, одинаково разделенных между клиентами, которые отказались от услуг компании, и клиентами, которые остались. В необработанных данных клиентов, которые остались с компанией, было больше, чем тех, которые ушли, в соотношении от 10 до 1. Это пример **несбалансированного набора данных** (imbalanced dataset), в котором различные уровни целевого признака, в данном случае *отток* и *не отток*, представлены в данных разным количеством экземпляров. Некоторые из подходов машинного обучения, которые мы обсуждали в предыдущих главах, лучше работают, если для их обучения используется **сбалансированная выборка** (balanced sample), и именно поэтому Росс создал таблицу ABT с равным количеством экземпляров для каждого целевого уровня[5].

Затем Росс разработал полный отчет о качестве данных в таблице ABT, включая диапазон визуализации данных. Таблицы отчетов о качестве данных показаны в табл. 9.2. Сначала Росс оценил уровень **пропущенных значений** (missing values). Среди непрерывных признаков выделялся только ВОЗРАСТ с отсутствующими 11,47% значений. Эту проблему можно было бы разумно решить с использованием подхода, основанного на восстановлении[6], но Росс воздержался от этого на данном этапе. Категориальные признаки ТИПРЕГИОНА и РОДЗАНЯТИЙ имели значительное количество пропущенных значений — 74 и 47,8% соответственно. Росс решительно решил полностью удалить эти признаки.

---

[4] Тот факт, что активные клиенты были определены как текущие, означает, что они были активны в одну и ту же дату — в день создания таблицы ABT. Это может быть проблематично: модель, обученная по этим данным, может игнорировать сезонные эффекты, такие как Рождество. Альтернативой является определение активного клиента как любого клиента, который проявлял активность в произвольно выбранный момент. Однако такое определение усложняет тот факт, что один и тот же клиент может оказаться как активным, так и пассивным, хотя, в принципе, описательные признаки для этих двух экземпляров должны вычисляться для разных периодов.

[5] Мы вернемся к этому обсуждению в разделах 9.5 и 10.4.1.

[6] См. раздел 3.4.

Таблица 9.2. Отчет о качестве данных для Acme Telephonic ABT

а) Отчет о качестве данных для непрерывных показателей

Признак	Кол-во	Процент пропусков	Мощность	Мин.	1-й квартиль	Среднее	Медиана	3-й квартиль.	Макс.	Станд. отклон.
Возраст	10000	11,47	40	0,00	0,00	30,32	34,00	48,00	98,00	22,16
Доход	10000	0,00	10	0,00	0,00	4,30	5,00	7,00	9,00	3,14
Количество телефонов	10000	0,00	19	1,00	1,00	1,81	1,00	2,00	21,00	1,35
Возраст Телефона	10000	0,00	1923	52,00	590,00	905,52	887,50	1198,00	2679,00	453,75
Стоимость Телефона	10000	0,00	16	0,00	0,00	35,73	0,00	59,99	499,99	57,07
Среднийсчет	10000	0,00	5588	0,00	33,33	58,93	49,21	71,76	584,23	43,89
СреднееКоличествоМинут	10000	0,00	4461	0,00	150,63	521,17	359,63	709,19	6336,25	540,44
СредняяРегулярнаяПлата	10000	0,00	1380	0,00	30,00	46,24	44,99	59,99	337,98	23,97
СреднееКоличествоЗвонковСверхПакета	10000	0,00	2808	0,00	0,00	40,65	0,00	37,73	513,84	81,12
СреднееКоличествоЗвонковРоуминге	10000	0,00	850	0,00	0,00	1,19	0,00	0,26	17,99	6,05
ИзменениеКоличестваМинут	10000	0,00	10000	–16,422	–1,49	0,76	0,50	2,74	19,28	3,86

Продолжение табл. 9.2

Признак	Кол-во	Процент пропусков	Мощность	Мин.	1-й квартиль	Среднее	Медиана	3-й квартиль.	Макс.	Станд. отклон.
ИзменениеСуммыСчета	10000	0,00	10000	–31,67	–2,63	2,96	1,96	7,56	42,89	8,51
СреднееКоличествоПолученныхМинут	10000	0,00	7103	0,00	7,69	115,27	52,54	154,38	2006,29	169,98
СреднееКоличествоВходящихЗвонков	10000	0,00	524	0,00	3,00	25,29	13,33	33,33	610,33	35,66
СреднееКоличествоИсходящихЗвонков	10000	0,00	310	0,00	0,00	8,37	2,00	9,00	304,00	17,68
ОтношениеПиковыхНепиковыхЗвонков	10000	0,00	8307	0,00	0,78	2,22	1,40	2,50	160,00	3,88
ИзменениеОтношенияПиковыхНепиковыхЗвонков	10000	0,00	10000	–41,32	–6,79	–0,05	0,01	6,50	37,78	9,97
СреднееСнижениеКоличестваЗвонков	10000	0,00	1479	0,00	0,00	0,50	0,00	0,00	9,89	1,41
ПродолжительностьОбслуживания	10000	0,00	56	6,00	11,00	18,84	17,00	24,00	61,00	9,61
КоличествоОбращений	10000	0,00	109	0,00	0,00	1,74	0,00	1,33	365,67	5,76
КоличествоПредложений	10000	0,00	5	0,00	0,00	0,05	0,00	0,00	4,00	0,23
КоличествоПринятыхНомеров	10000	0,00	5	0,00	0,00	0,02	0,00	0,00	4,00	0,155
КоличествоНовыхНомеров	10,000	0,00	4	0,00	0,00	0,20	0,00	0,00	3,00	0,64

Окончание табл. 9.2

б) Отчет о качестве данных для категориальных показателей

Признак	Кол-во	Процент пропусков	Мощность	Мода	Частота моды	Процент моды	2-я мода	Частота 2-й моды	Процент 2-й моды
РодЗанятий	10000	74,00	8	Служащий	1705	65,58	Рабочий	274	10,54
ТирРегиона	10000	47.80	8	Пригород	3085	59,05	Город	1483	28,39
СемейноеПоложение	10000	0,00	3	Неизвестно	3920	39,20	Да	3594	35,94
Дети	10000	0,00	2	false	7559	75,59	true	2441	24,41
Смартфон	10,000	0,00	2	true	9015	90,15	false	985	9,85
Кредитоспособность	10000	0,00	7	b	3785	37,85	c	1713	17,13
Домовладелец	10,000	0,00	2	false	6577	65,77	true	3423	34,23
КредитнаяКарта	10000	0,00	6	true	6537	65,37	false	3146	31,46
Отток	10,000	0,00	2	false	5000	50,00	true	5000	50,00

Вычисляя **мощность** (cardinality) признаков, Росс заметил, что несколько непрерывных признаков, например, Доход, Возраст, СтоимостьТелефона и КоличествоПредложений, имеют очень низкую мощность. В большинстве случаев данные оказались корректными, потому что диапазон значений, которые они могли принимать, был узким по естественным причинам. Например, СтоимостьТелефона может принимать лишь небольшое количество значений, например 59,99; 129,99; 499,99 и т.д. Признак Доход отличался тем, что у него было всего 10 разных значений (гистограмма этого признака подтвердила этот факт (рис. 9.2, *а*)). Грейс объяснила Россу, что доходы были фактически записаны по группам, а не как точные значения, так что на самом деле этот признак был категориальным. Мощность категориальных признаков КредитнаяКарта и ТипРегиона была выше, чем ожидалось (гистограммы этих признаков см. на рис. 9.2, *б* и *в*). Проблема заключалась в том, что некоторые уровни имели несколько представлений. Например, для признака ТипРегиона города были обозначены и как *город*, и как *г*. Росс легко исправил эту проблему, сопоставив уровням этого признака одну последовательную схему маркировки.

Четыре непрерывных признака выделялись тем, что, возможно, имели **выбросы** (outliers): СтоимостьТелефона с минимальным значением 0, что казалось необычным; СреднееКоличествоМинут с максимумом 6336,25, который сильно отличался от среднего и третьего квартилей этого признака; СреднееКоличествоПолученныхМинут с максимумом 2006,29, который также сильно отличался от среднего значения и третьего квартиля этого признака, и СреднееКоличествоМинутСверхПакета с нулевыми минимальным значением, 1-м квартилем и медианой по сравнению со средним значением, равным 40. Гистограммы этих признаков представлены на рис. 9.2. Росс уточнил у Грейс и Кейт, что это были действительно выбросы. Например, некоторые телефоны раздают бесплатно, а некоторые клиенты просто делают много звонков. Однако они провели некоторое время, обсуждая признак СреднееКоличествоМинутСверхПакета. Гистограмма для этого признака имеет необычную форму, которая приводит к необычным значениям минимума, 1-го квартиля и медианы (см. рис. 9.2, *з*). Более внимательно изучив данные для этого признака, они в конце концов объяснили эту форму тем фактом, что большинство клиентов не превышали количество минут в своем пакете, что объясняет большое количество нулей в этой гистограмме. Значения больше нуля, похоже, имеют почти нормальное распределение с большим разбросом, и большое количество корректных нулевых значений отражает необычные значения минимума, 1-го квартиля

и медианы. Пока Росс только отметил эти выбросы, как что-то, с чем ему, возможно, придется иметь дело на этапе моделирования.

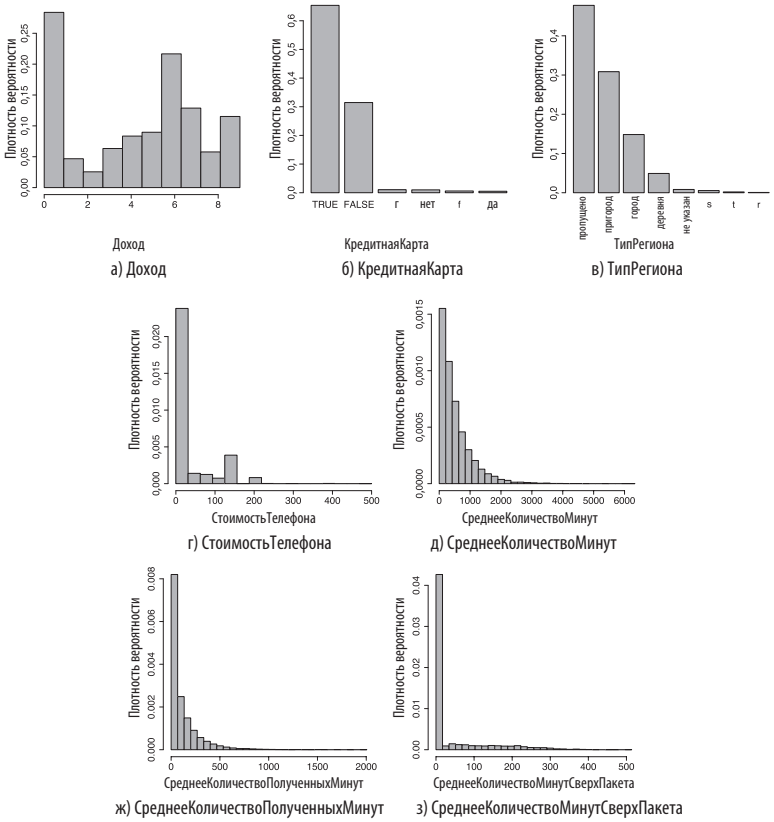

Рис. 9.2. Гистограммы для признаков из таблицы ABT с необычной мощностью (а–в); гистограммы для признаков из таблицы ABT, которые потенциально имеют выбросы (г–з)

Затем Росс обратил внимание на визуализацию взаимосвязи между каждым описательным и целевым признаками. Ни один отдельный признак не отличался очень сильной связью с целевым признаком, но можно было увидеть доказательства связей между описательными и целевым признаками. Например, на рис. 9.3, *а*, показана немного более высокая склонность сельских жителей к оттоку. Аналогичным образом на рис. 9.3, *б*, показано, что клиенты, которые отказались от услуг компании, имели тенденцию делать больше звонков сверх своего пакета, чем те, кто этого не сделал.

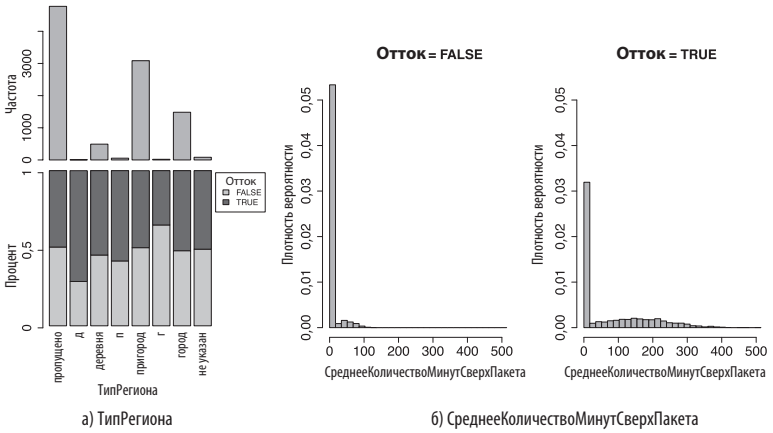

Рис. 9.3. Блочная диаграмма признака ТипРегиона (а); гистограммы признака СреднееКоличествоМинутСверхПакета, построенные по значениям целевого признака (б)

Подробно рассмотрев полный отчет о качестве данных, Росс принял следующие решения в отношении выявленных проблемных признаков. Во-первых, он решил удалить признаки Возраст и РодЗанятий из-за слишком высокого уровня пропущенных значений у каждого из них. Однако он решил сохранить признак ТипРегиона, потому что у него, похоже, были некоторые связи с целью. Он также применил запланированное сопоставление значений ТипРегиона с последовательной схемой маркировки: {*s*|*пригород*} → *пригород*; {*t*|*город*}→ *город*; {*пропущено*|*не указан*} → *пропущено*.

Росс разделил этот набор данных на три случайно распределенных части: обучающая выборка (50%), выборка для валидации (20%) и тестовая выборка (30%). Обучающая выборка использовалась в качестве основных данных для обучения построенных моделей прогнозирования. Выборка для валидации использовалась для настройки, а тестовая — для окончательной проверки эффективности модели.

9.4. Моделирование

Требования к этой модели заключались в том, чтобы она была точной, допускала интеграцию в более широкие бизнес-процессы в компании АТ и, возможно, позволяла понять причины, по которым люди могут отказаться от услуг компании. При выборе подходящего типа модели необходимо учитывать все эти аспекты, а также структуру данных. В этом случае таблица АВТ состояла из смеси непрерывных и категориальных описательных признаков и имела категориальный целевой

признак. В частности, категориальный целевой признак делает деревья решений подходящим выбором для этой задачи моделирования. Кроме того, алгоритмы дерева решений способны обрабатывать как категориальные, так и непрерывные описательные признаки, а также обрабатывать отсутствующие значения и выбросы без необходимости преобразовывать данные. Наконец, деревья решений относительно легко интерпретировать, и структура модели может дать некоторое представление о поведении клиентов. Все эти факторы, взятые вместе, показали, что деревья решений являются подходящим выбором для этой задачи.

Росс использовал таблицу ABT для обучения, настройки и тестирования ряда деревьев решений для прогнозирования оттока с учетом набора описательных признаков. Первое дерево, построенное Россом, в качестве критерия расщепления использовало прирост энтропии, ограничило непрерывное разделение до бинарных вариантов и не использовало усечения. Росс решил снова проконсультироваться с бизнес-экспертами и выяснил, что наиболее подходящей оценочной мерой является простая точность классификации. Первое построенное дерево на тестовом наборе достигало **средней точности классификации**[7], равной 74,873%, что было достаточно обнадеживающим.

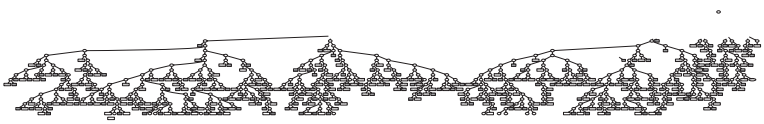

Рис. 9.4. Необработанное дерево решений, построенное для прогнозирования оттока клиентов компании AT (приведено только для демонстрации его размера и сложности). Чрезмерная сложность и глубина дерева свидетельствуют о том, что, вероятно, произошло переобучение

Это дерево показано на рис. 9.4, а отсутствие усечения очевидно по его сложности. Эта сложность и чрезмерная глубина дерева свидетельствуют о переобучении. Во втором дереве, которое построил Росс, использовалось **последующее усечение, уменьшающее частоту ошибок**[8], для которого использовалось тестовое множество, созданное из исходного набора данных. Достаточно большой набор данных, с которого Росс должен был начать, в свою очередь, позволил создать достаточно большое тестовое множество. Это означало, что

[7] *Средняя точность классификации в этом разделе использует* ***среднее гармоническое значение****.*

[8] *См. раздел 4.4.4.*

усечение, уменьшающее частоту ошибок, в данном случае было подходящим методом[9]. На рис. 9.5 показано дерево, полученное в результате этой итерации обучения. Совершенно очевидно, что это гораздо более простое дерево, чем предыдущее. Признаки, используемые на верхних уровнях обоих деревьев и считающиеся наиболее информативными, оказались одинаковыми: СРЕДНЕЕКОЛИЧЕСТВОМИНУТСВЕРХПАКЕТА, ИЗМЕНЕНИЕСУММЫСЧЕТА и ВОЗРАСТТЕЛЕФОНА.

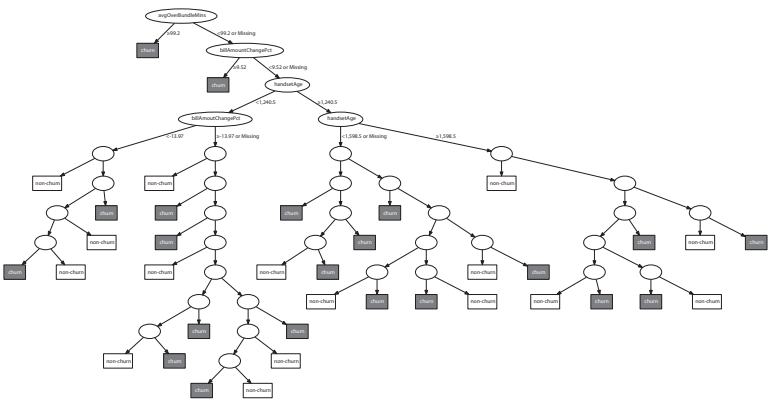

Рис. 9.5. Усеченное дерево решений, построенное для задачи прогнозирования оттока клиентов компании AT. Серые листья указывают на предсказание оттока, в то время как белые — на сохранение клиентов. Для экономии места мы показываем только те признаки, которые тестировались на узлах верхнего уровня

Таблица 9.3. Матрица ошибок для предсказания оттока на основе усеченного дерева и стратифицированного отложенного тестового множества (см. рис. 9.5)

		Прогноз		Полнота
		Отток	Нет оттока	
Цель	Отток	1058	442	70,53
	Нет оттока	152	1348	89,85

[9] *Если бы данные были более скудными, следовало бы применить усечение с использованием статистического критерия, такого как* χ^2.

Таблица 9.4. Матрица ошибок для предсказания оттока на основе нестратифицированного отложенного тестового множества

		Прогноз		Полнота
		Отток	Нет оттока	
Цель	Отток	1115	458	70,88
	Нет оттока	1439	12878	89,95

Используя усечение, Росс смог повысить среднюю точность классификации при тестировании на отложенном тестовом множестве до 79,03%, что значительно лучше по сравнению с предыдущей моделью. В табл. 9.3 показана матрица ошибок для этого теста. Матрица ошибок показывает, что эта модель была немного более точной при классификации экземпляров с целевым уровнем *нет оттока*, чем с целевым уровнем *отток*. Исходя из этого, результаты Росса уверенно свидетельствовали о том, что это дерево было хорошим решением задачи прогнозирования оттока клиентов в компании АТ.

9.5. Оценка

Оценки модели, основанные на доле ошибок классификации, описанные в предыдущем разделе, являются первым шагом в оценке эффективности созданной модели прогнозирования. Точность классификации, равная 79,03%, намного выше цели, согласованной с бизнесом. Однако это вводит в заблуждение. Эта эффективность основана на стратифицированном отложенном тестовом множестве, в котором содержится столько же случаев оттока, сколько случаев без оттока. Тем не менее распределение экземпляров с оттоком и без оттока в более широкой клиентской базе компании АТ существенно отличаются друг от друга. Вместо пропорции 50:50 фактическое базовое отношение, по сути, оказалось ближе к 10:90. По этой причине очень важно выполнить вторую оценку, в которой данные теста отражают фактическое распределение значений целевых признаков в бизнес-сценарии.

Росс имел вторую выборку (не перекрывающуюся с предыдущей), которая не была стратифицирована в соответствии со значениями целевых признаков. Матрица ошибок, иллюстрирующая эффективность модели прогнозирования на этом тестовом множестве, показана в табл. 9.4.

Средняя точность классификации на нестратифицированном тестовом множестве составила 79,284%. Росс также вычислил **кумулятивный прирост**,

подъем и **кумулятивный подъем**[10] (рис. 9.6). Диаграмма кумулятивного прироста, в частности, показывает, что если сотрудники компании АТ будут звонить только 40% своих клиентов, то они идентифицируют примерно 80% клиентов, которые могут отказаться от их услуг, что является убедительным доказательством того, что модель хорошо распознает различные типы клиентов.

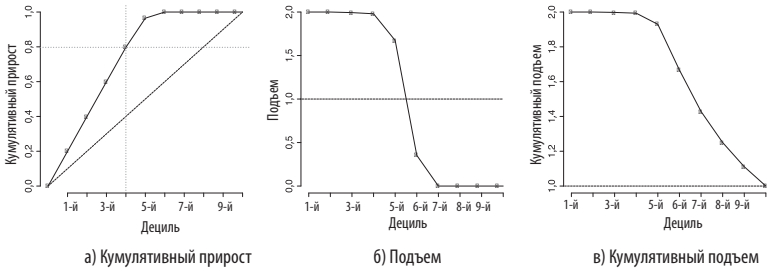

Рис. 9.6. Кумулятивный прирост (а); подъем (б) и кумулятивный подъем для прогнозов, сделанных на крупной тестовой выборке (в)

Учитывая эти хорошие результаты, Росс решил, что было бы целесообразно представить модель другим бизнес-подразделениям. Это был важный шаг в завоевании доверия к модели. Дерево, показанное на рис. 9.5, легко интерпретируется, но если его показать другим бизнес-подразделениям, людям может быть трудно справиться с таким большим объемом информации, поэтому Росс решил создать специальную ограниченную версию дерева решений с небольшим количеством уровней для представления модели бизнес-подразделениям (хотя для фактического развертывания он намеревался использовать большее усеченное дерево). Идея заключалась в том, чтобы упрощенность дерева сделала его более легко интерпретируемым. Тот факт, что наиболее информативные признаки занимают узлы ближе к вершине дерева, означает, что низкорослые деревья обычно отображают самую важную информацию. Многие средства машинного обучения позволяют задавать в качестве параметра максимальную глубину дерева, что позволяет создавать такие **низкорослые деревья** (stunted trees).

На рис. 9.7 показано низкорослое дерево Росса, сгенерированное для решения проблемы оттока, где глубина ограничена пятью уровнями. Это дерево дает немного более низкую точность классификации на тестовом множестве (78,5%), зато очень легко интерпретируется. Основные признаки при распознавании оттока — это, очевидно, СРЕДНЕЕКОЛИЧЕСТВОМИНУТСВЕРХПАКЕТА, ИЗМЕНЕНИЕСУММЫСЧЕТА и ВОЗРАСТТЕЛЕФОНА. Из этих данных следует, что клиенты, ско-

[10] ***Кумулятивный прирост***, ***подъем*** *и* ***кумулятивный подъем*** *введены в разделе 8.4.3.3.*

рее всего, откажутся от услуг компании, если их счет резко изменится, если они начнут превышать количество минут в своем пакете звонков или если они пользуются старым телефоном и рассматривают возможность купить что-то более новое. Это полезная информация, которую бизнес может использовать, чтобы попытаться разработать другие стратегии борьбы с оттоком параллельно с использованием этой модели для создания списков клиентов для службы удержания. Бизнес-экспертов заинтересовали признаки, которые были выбраны как важные для построения дерева, и было много обсуждений по поводу отсутствия признаков, описывающих взаимодействие клиентов со службами поддержки компании AT (это было основой предыдущей модели).

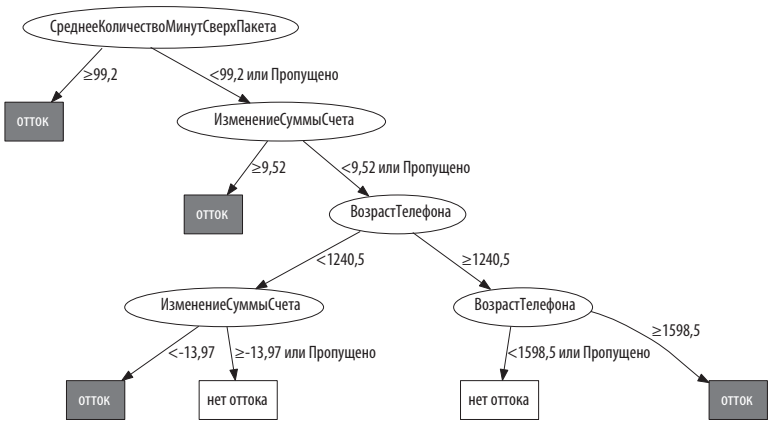

Рис. 9.7. Усеченное и низкорослое дерево решений, созданное для решения задачи прогнозирования оттока клиентов компании Acme Telephonica

Для дальнейшей поддержки своей модели Росс организовал тестирование с использованием контрольной группы (см. раздел 8.4.6), в котором в течение двух месяцев база клиентов компании AT была случайным образом разделена на две группы, а списки клиентов, которым должна была звонить служба удержания, были выбраны двумя способами: старым, основанным на звонках от клиентов в службу поддержки, и новым — на основе дерева решений. Через два месяца было показано, что отток в группе, для которой служба удержания клиентов использовала новую модель для составления своего списка звонов, составил примерно 7,4%, тогда как для группы, использующей старую модель, он составил более 10%. Этот эксперимент показал руководству компании AT, что новая модель дерева решений может значительно снизить уровень оттока в клиентской базе AT.

9.6. Развертывание

Поскольку компания АТ уже использовала процесс, в котором его служба удержания создавала списки звонков на основе собранных данных, развертывание новой модели дерева решений было достаточно простым. Основная проблема заключалась в возврате на этап подготовки данных, чтобы сделать подпрограммы, используемые для извлечения данных для таблицы ABT, устойчивыми и надежными, чтобы их каждый месяц можно было использовать для генерации новых тестовых экземпляров. Это означало работу с отделом информационных технологий для разработки готовых к развертыванию подпрограмм **извлечения–преобразования–загрузки** (extract-transform-load — ETL). Затем был написан код для замены предыдущего простого правила выбора клиентов на дерево решений.

Последний шаг в развертывании заключался в том, чтобы внедрить **постоянную валидацию модели**, чтобы вовремя распознать признаки, свидетельствующие о том, что развернутая модель устарела. В этом случае в течение достаточно короткого промежутка времени после того, как были сделаны предсказания, возникает обратная связь, дающая представление об эффективности модели. Прогнозы дня можно легко сопоставить с фактическим поведением клиентов (с учетом мер, предпринимаемых компанией). Система мониторинга, которую внедрил Росс, в конце каждого квартала составляла отчет, в котором оценивалась эффективность модели в предыдущем квартале, сравнивая, сколько людей, с которыми не связывалась служба удержания клиентов, фактически отказалось от услуг компании. Если это число значительно изменилось по сравнению с данными, используемыми для построения модели, то модель будет считаться устаревшей и потребуется ее перестройка.