ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

###### Факультет бизнеса и менеджмента

Ляпина Екатерина Романовна

**Проектирование банковского рекомендательного сервиса с использованием методов машинного обучения**

Выпускная квалификационная работа - БАКАЛАВРСКАЯ РАБОТА / по направлению подготовки Бизнес-информатика

|  |  |
| --- | --- |
| Рецензент  преподаватель  Гоменюк К.С. | Научный руководитель  доцент.  Шевгунов Т.Я. |

Москва 2016

Содержание

[Введение 3](#_Toc451871708)

[Глава 1. Теоретические предпосылки исследования 6](#_Toc451871709)

[1.1 Анализ предметной области 6](#_Toc451871710)

[1.2 Анализ доступных данных и способов их использования 22](#_Toc451871711)

[1.3 Обзор имеющихся подходов выработки рекомендаций 26](#_Toc451871712)

[Глава 2. Архитектура разрабатываемой системы 30](#_Toc451871713)

[Глава 3. Практическая реализация сервиса 34](#_Toc451871714)

[1.1 Разработка рекомендательных алгоритмов 34](#_Toc451871715)

[1.2 Анализ эффективности различных рекомендательных алгоритмов 36](#_Toc451871716)

[1.3 Контентные методы 38](#_Toc451871717)

[Заключение 41](#_Toc451871718)

[Список литературы 42](#_Toc451871719)

[Приложения 46](#_Toc451871720)

[Приложение 1 Настройка программной среды 46](#_Toc451871721)

[Приложение 2 Коллаборативная фильтрация 48](#_Toc451871722)

[Приложение 3 Сравнение подходов 51](#_Toc451871723)

[Приложение 4 Алгоритм формирования дерева решений 52](#_Toc451871724)

[Приложение 5 Алгоритм формирования случайного леса 56](#_Toc451871725)

# Введение

Современный мир динамичен, и степень его динамичности со временем все возрастает. Для любого направления бизнеса чрезвычайно важно следить за возникающими изменениями во внешней среде. В сегодняшней реальности основные драйверы этих изменений – это проникновение технологий во все сферы жизни. Также важную роль играет увеличение общего количества людей на Земле и информации, которую они генерируют. Мы ищем способы анализировать большие объемы информации, которые не только не можем уже анализировать сами, но и даже силами 1 компьютера.

Систематический сбор, отражение, анализ данных чрезвычайно важен для фирм, смысл существования которых - удовлетворение потребностей потребителя. Увеличение количества обрабатываемых данных ведет к лучшему пониманию потребителя, его привычек и, как следствие, к разработке лучших продуктов, услуг и предложений. Последние тенденции в создании банковских продуктов направлены на увеличение интероперабельности систем и создание сервисов, которые бы соответствовали потребностям потребителей.

Одним из бурно развивающихся направлений совершенствования прикладных информационных технологий являются рекомендательные системы – инструменты автоматической генерации предложений по товарам и услугам, основанные на изучении данных о предыдущих действиях клиентов. Создание персонализированного банковского сервиса направлено на улучшение понимания банком потребностей и образа жизни своих клиентов для последующей выработки наиболее релевантных предложений банковских продуктов и предложений партнеров-ритейлеров.

Учитывая готовящийся законопроект о приведении чеков в электронный вид, потребности банков и их клиентов, актуально создание проектного решения, обеспечивающего реализацию совместного персонализированного сервиса для клиентов с привлечением данных оффлайн ритейлеров и банков.

Теоретическое значение работы состоит в том, что произведенный анализ определяет тенденции рыночных процессов, выявляет влияние готовящегося законодательного акта на банковский сектор и определяет возможные варианты его развития. Практическая значимость сервиса состоит в возможности улучшения пользовательского опыта, расширения знаний банка и торговца о клиенте, создания открытого сервиса для наилучшей утилизации данных сторонними разработчиками и экономии предприятиями на эмиссии банковских карт и карт лояльности. Полученные результаты исследования могут быть использованы для дальнейших прикладных разработок по данной и смежным темам.

Выработка рекомендаций – хорошо изученная научная область, которая имеет успешные приложения в индустрии. Рекомендательные системы уже стали неотъемлемой частью многих современных интернет-приложений, доказав свою эффективность в улучшении пользовательского опыта и повышении продаж фирм.

В работе рассмотрены алгоритмы подбора наилучшего типа предложения и товара, наиболее релевантных предложений, основываясь на методах коллаборативной фильтрации и контентном подходе. Данные алгоритмы реализованы для масштабируемой программной платформы Apache Spark на языке Scala и R. Эксперименты, направленные на оценивание качества и масштабируемости реализованных алгоритмов, проведены на облачном сервисе Microsoft Azure.

**Объектом** исследования является рынок банковских продуктов, возникающий после вступления в силу законодательного акта.

**Предметом** исследования являются варианты использования доступных данных о транзакциях пользователей, включая способы их хранения и анализа.

**Цель работы:** анализ рынка банковских услуг, потребностей банка и их клиентов. Последующее исследование существующих рекомендательных алгоритмов и разработка эффективного их сочетания для применения в системе рекомендаций программы лояльности, позволяющей предлагать нужные продукты с приемлемым уровнем релевантности, уровнем отклика и экономической целесообразностью в условиях большого числа пользователей при неполной или отсутствующей информации об их предпочтениях.

**Задачи:**

1. Проанализировать потребности банков, клиентов банков и ритейлеров.
2. Изучить готовящийся законопроект и требования регулятора.
3. Выработать решение по использованию доступных данных.
4. Определить основные параметры будущей системы и данные, которые необходимо использовать для создания сервиса.
5. Разработать программный продукт для хранения, трансформации и анализа большого количества данных о транзакциях, генерируемых будущим сервисом.
6. Осуществить обработку данных посредством использования методов машинного обучения для формирования рекомендаций по продуктам и выбора наилучшего способа предложения, используя коллаборативную фильтрацию, основанную на сходстве пользователей, коллаборативную фильтрацию, основанную на сходстве предметов, и алгоритмы классификации для подбора оптимального канала предложения.

**Методы**:

1. Анализ открытых источников, отражающих состояние и тенденции отрасли;
2. Анализ архивных материалов с тематических конференций;
3. Анализ общедоступного законопроекта;
4. Обозначение круга задач, которые могут быть решены с использованием новых данных при вступлении закона в силу;
5. Выделение наиболее перспективной задачи для разработки персонализированного сервиса;
6. Анализ доступной научно-методической литературы по архитектуре высоконагруженных систем и методам создания рекомендательных сервисов;
7. Обработка имеющихся данных математическими методами.

Данное исследование представляет собой работу, которая состоит из трёх глав: первая глава - это исследование предметной области, выявление потребностей сторон и формирование бизнес-требований к приложению; вторая глава - это описание структуры программного обеспечения; третья глава представляет собой реализацию программного продукта из имеющихся данных.

**Основными источниками информации являются:**

Статьи и книги по методикам проведения маркетинговых исследований; статьи и книги по архитектуре высоконагруженных систем и систем, работающих с большим объемом данных, прежде всего систем для массово-параллельной обработки данных; статьи и книги по методам создания рекомендательных сервисов.

# Глава 1. Теоретические предпосылки исследования

## 1.1 Анализ предметной области

Развитие банковской сферы в России во многом определяется государственной политикой. В феврале 2016 года утвердили проект изменений, вносимых в 54-ФЗ «О применении контрольно-кассовой техники при осуществлении наличных денежных расчетов и(или) расчётов с использованием платёжных карт» [45]. Основные положения предлагаемой системы – это:

* Передача информации о расчетах в электронном виде в адрес налоговых органов через оператора фискальных данных.
* Электронная регистрация контрольно-кассовой техники (ККТ) без посещения налогового органа и без физического предоставления ККТ.
* Построение системы гарантированного выявления нарушений на основе автоматизированного анализа информации о расчетах, выявления зон риска совершения правонарушений и проведения точечных результативных проверок.
* Вовлечение покупателей в гражданский контроль.

Следует отметить, что инициатива по внесению данной поправки была принята с учетом результатов внедрения Единой государственной автоматизированной информационной системы (ЕГАИС) — автоматизированной системы, предназначенной для государственного контроля над объёмом производства и оборота этилового спирта, алкогольной и спиртосодержащей продукции. По оценке специалистов, до начала внедрения ЕГАИС доля нелегальной продукции в общем объеме реализации алкогольных напитков на потребительском рынке России достигала 40 – 50%. Полагается, что данная система сделала оборот алкогольной продукции более прозрачным и за счет этого сократило потребление подакцизных товаров. В то же время, цены на акцизы не поднимались c 2014 года, а покупательская способность денег ухудшилась. Система имеет ряд недостатков, среди которых отсутствие балансировки нагрузки, медленный отклик и несовместимость системы с некоторым ПО, ведь данные отправляются в ФНС напрямую.

В рамках поправки ФЗ-54 о контрольно-кассовой технике, новая схема работы включает посредника в лице оператора фискальных данных, который будет накапливать, хранить и передавать эти данные в ФНС. Согласно новому постановлению, новый институт будет обязан:

1. передавать в контрольно-кассовую технику, осуществившую передачу ему в виде электронного документа кассового чека, отчета о фискализации, открытии и закрытии смены, закрытии фискального накопителя, подтверждение о получении указанного документа, подписанное фискальным признаком;
2. осуществлять ежедневную передачу в адрес налогового органа фискальных данных,
3. предоставлять налоговому органу доступ к фискальным данным в режиме реального времени, а также предоставлять эти данные налоговому органу;
4. обеспечивать возможность проверки факта применения контрольно-кассовой техники организацией через интернет, а также возможность проверки достоверности оформленного кассового чека в виде электронного документа;
5. осуществлять в случае, если это предусмотрено договором между оператором фискальных данных и пользователем, передачу копий кассовых чеков в виде электронных документов с указанием в качестве адреса электронной почты отправителя – адреса ОФД;
6. предоставлять по интернету на безвозмездной основе любому лицу, предоставившему номер кассового чека, реквизиты и ФИО, а также адрес места жительства, кассовый чек или бланк строгой отчетности в виде электронного документа, подписанного квалифицированной электронной подписью оператора фискальных данных;

Схема взаимодействия наглядным образом представлена на Рисунке 1.

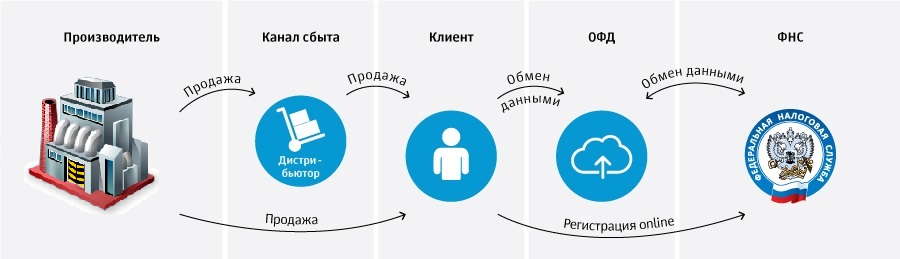


Рисунок 1 Схема обмена данными между контрагентами

Передача данных в ОФД будет производиться с использованием фискального признака (уникальной подписи каждого чека). При необходимости, передачу оператору фискальных данных электронных документов с фискальными данными можно осуществлять в зашифрованном виде. В обязанности ОФД также входит хранение данных на протяжении 5 лет.

По требованию покупателя организации и индивидуальные предприниматели дополнительно к кассовому чеку, отпечатанному контрольно-кассовой техникой на бумажном носителе, обязаны передать покупателю его копию в виде электронного документа в незашифрованном виде на предоставленный покупателем адрес электронной почты.

Организации и индивидуальные предприниматели вправе поручить передачу покупателям копий кассовых чеков в виде электронных документов оператору фискальных данных. В этом случае в качестве адреса электронной почты отправителя должен быть указан адрес электронной почты оператора фискальных данных.

Гражданину новая технология дает дополнительную защиту своих прав как потребителя за счет возможности:

* получить электронный чек у оператора фискальных данных и (или) в своей электронной почте;
* самостоятельно быстро и удобно проверить легальность кассового чека через бесплатное мобильное приложение и в случае возникновения вопросов тут же направить жалобу в ФНС России;
* Кассовый чек (бланк строгой отчетности) формируется контрольно-кассовой техникой в электронном виде (по требованию покупателя – на бумажном носителе) при осуществлении расчетов;

Плавный и поэтапный переход к новому порядку применения должен произойти до 1 февраля 2017 года. Законопроект никак не регламентирует что либо, кроме произведения налоговых расчетов по каждой позиции и хранения данных. Характер взаимодействия торгово-сервисного предприятия и оператора фискальных данных очень похож на процессинг платежных транзакций. Учитывая, что в банках сосредоточена и накоплена огромная экспертиза в области процессинга и обеспечения безопасности платежей, а также то, что многие финансовые организации предоставляют услуги по расчетно-кассовому обслуживанию, и юридические лица обязаны открывать счета в банке при регистрации, кажется вполне закономерным, что роль оператора фискальных данных будет занята банками.

Чек – простейший документ купли-продажи, обеспечивающий защиту прав потребителя. Вместе с этим, данные, зарегистрированные в памяти ККТ, обеспечивают расчет налога по сумме проданных товаров за покупку. Однако, несмотря на всю ценность, люди хранят чеки только после покупки электротоваров, крупной и малой бытовой техники, что подтверждается в исследовании Национального агентства финансовых исследований [42].

В результате опроса было также выявлено, что в 36% случаев, чек при покупке люди не получали. Это значит, что торговцы осуществили продажу товаров без уплаты налогов. Такая ситуация складывается из-за невнимательности населения, непонимания назначения чека, отсутствия удобства хранения и учета. Помимо всего прочего, порядка 43% людей, забравших чек, сразу его выбрасывают.

31% респондентов готовы сохранить кассовый чек только при определенной сумме покупки. Приобретая тот или иной товар на сумму от 1000 руб. до 5000 руб. чек готовы сохранить - 23%, от 100 руб. до 500 руб. - 27%, от 500 до 1000 руб. - 37%, тогда как 46% участников опросов сохраняют кассовые чеки на любую покупку. В результате тестирования по поводу категорий товаров была составлена следующая статистика: на продукты питания сохраняют чеки 45% граждан, на товары высокой стоимости (драгоценности, автомобиль и пр.) - 53%, на обувь и одежду - 66%, на бытовую технику и электротовары чаще всего граждане нашей страны сохраняют кассовые чеки - 70%.

Люди хранят чеки разное количество времени. Это связано с тем, что возврат и обмен товара можно осуществить в течение 14 дней со дня покупки, гарантия на электронные товары обычно составляет около года. Кроме того, большинство чеков печатаются на термобумаге (термопринтер дешевле и быстрее), поэтому текст со временем выцветает.

Люди, которые хранят чеки длительное время, в основном страхуются от некачественных товаров, ведут домашнюю бухгалтерию, участвуют в акциях при покупке определенного товара, возвращают налог с некоторых операций или отчитываются о покупках.

Действие нового закона обеспечивает хранение электронного чека, но он подразумевает, что человек отсканирует его в официальном мобильном приложении, чтобы получить доступ к электронной копии. Возникновение программ по хранению и анализу номеров чеков неизбежно. Открытие данных для работы сторонних организаций с разрешения пользователя обеспечивает лучшее использование хранящихся данных.

Проверка на легальность – нужная функция при проверке алкоголя в системе ЕГАИС, потому как люди не хотят отравиться фальшивым товаром, интересуются подлинностью в случае покупки дорогой алкогольной продукции. Напротив, человеку нет нужны утруждать себя проверкой чеков на товары повседневного пользования в привычных местах, потому как он склонен считать, что вероятность форс-мажора крайне мала, а люди достаточно инертны в своем образе жизни. Кажется, что мотивация контроля за торговцами со стороны населения недостаточна потому как аудитория тех, кто хранит чеки крайне мала.

В то же время, монетизация банков до широкого проникновения IT состояла только в грамотном управлении финансами. Вместе с информационными технологиями появились карточки – уникальный банковский продукт. Банки стали получать дополнительный доход за счет комиссионных отчислений с размера транзакций в зависимости от денежного перевода или снятия наличных. Многочисленные исследования показывают, что современный потребитель зачастую ставит удобство использования на первое место, все меньше акцентируя внимание на цене. Такая ситуация происходит из-за смены мышления: все больше людей начинают понимать цену своего времени, что оттеняет желание экономить на всем подряд. К тому же общеизвестным является факт о том, что люди, использующие безналичный расчет склонны больше тратить, поэтому карточный эквайринг – востребованный банковский продукт, увеличивающий выручку и улучшающий сервис.

По данным Центрального банка, эмиссия карточек, выпущенных кредитными организациями, на 1 января 2016 года составляет 243 929 млн.ед, демонстрируя стабильные темпы выпуска [34]. Безусловно, такой рост складывается не только из новых обращений, но и из перевыпуска карт, у которых истек срок годности. Наряду с данной статистикой, нужно отметить тенденции в карточной эмиссии, связанные с выпуском виртуальных карточек, которые никак не учитываются статистикой ЦБ. Все меньше становится людей, которые используют карточки для снятия наличности в день выдачи зарплаты, благодаря внедрению программ банковской лояльности с крупными торговцами, проведению акций международными платежными системами для стимулирования пользования карточками и воспитания культуры использования этого продукта.

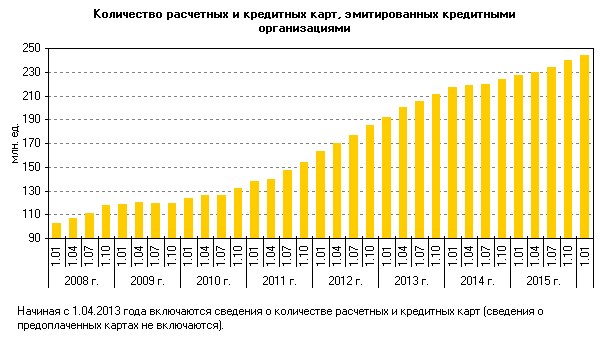


Рисунок 2 Статистика центрального банка по эмиссии банковских карт

Нельзя отрицать массовое распространение смартфонов, как самых персональных устройств и альтернативного способа оплаты. Таким образом, банки уже владеют системами по карточному процессингу и могут предоставлять различные способы оплаты товаров: карточкой, через электронный кошелек NFC или barcode. Государство активно продвигает систему безналичных платежей, потому что такой способ взаиморасчетов позволяет сделать систему прозрачной и лучше контролировать денежную массу и денежные потоки, что особенно актуально в связи принятием закона «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма». Государство всецело заинтересовано в переходе в цифру из-за прозрачности информационных и финансовых потоков, сокращения расхода бумаги на бланки. Эта ситуация предоставляет благоприятную почву для сопоставления данных о платежах с личностью клиента. Потоки безналичных платежей и обмен фискальной информацией будут происходить параллельно. Банки раньше всех начали заниматься анализом данных, но сейчас они знают, сколько тратят их клиенты и где, но не знают на что и зачем. Задачи, которые банки решают сейчас, используя продвинутую аналитику были выявлены после анализа открытых источников и наблюдений за тенденциями в отрасли [1, 5, 16, 24 – 27, 29 – 33, 35 – 37, 39, 41]. Они перечислены в Таблице 1.

Таблица 1

Решаемые задачи в области анализа данных в финансовом секторе

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Банк | Задачи | Данные | Планы | Технологии |
| Промсвязьбанк | Формирование наиболее актуальной задачи для клиентского менеджера. |  | Единое окно системы как фронт с логикой, но без процесса. |  |
| Уральский банк реконструкции и развития | Продажа данных другим организациям (hash). Используют соцсети как канал продаж. | Соцсети. | Полученную и проанализированную информацию из соцсетей о клиентах можно использовать для дальнейшего взаимодействия. | In-house решение. |
| Тинькофф Банк | Скоринг, продажа карт и привлечение ипотеки. | Данные социальных сетей для скоринга. |  | EMC Greenplum, SAS Visual Analytics и Hadoop. |
| Сбербанк | Риски, сегментация и скоринг, управление персоналом, прогнозирование очередей в отделениях, расчет бонусов сотрудников. Предсказание клиенто-потока в отделениях, предсказание банкротства фирмы. | Внутренние данные. | Банк намерен подключить к анализу больше типов своих внутренних данных и задействовать внешние источники (например, данные из соцсетей). В марте Сбербанк купил рекламную платформу Segmento, чтобы использовать ее данные для персонализации предложений своих клиентам и привлечения новых. Банк привлек «Яндекс» в качестве консультанта по анализу больших данных. | Teradata, Cloudera Hadoop, Impala, Zettaset, стек продуктов Apache (Hadoop, HBase, Hive, Mahout, Oozie, Zookeeper, Flume, Solr, Spark и пр.), специализированные базы данных (Neo4j, MongoDB и т.д.) и собственные решения в области data mining, predictive/prescriptive-аналитики, обработки естественного языка. |
| Альфа-банк | Предсказание банкротства. Анализ соцсетей и поведения пользователей сайта, скоринг, отток, персонализация контента и вторичные продажи. | Данные по отрасли, «Контур», ЕГРЮЛ и ФНС, данные о людях, которые являются собственниками бизнеса. | Возможности дополнительной монетизации своих массивов данных в рекомендательных системах, анализе линейки продуктов и предиктивном анализе поведения клиентов. | Oracle Exadata, Oracle Big Data Appliance и Hadoop. |
| Газпромбанк | Скоринг, антифрод, оперативная отчетность, персонализация предложений, доскоринговая проверка репутации потенциальных заемщиков, предоставление информации регуляторам. |  |  |  |
| ВТБ24 | Отток, формирование финансовой отчетности, анализ отзывов (соцсети, «Народный рейтинг» и «Банки.ру»). Выявление необоснованного cash back, ложные кредиты. |  |  | Teradata, SAS Visual Analytics и SAS Marketing Optimizer, CleverData. |
| Открытие | Анализ поведения клиентов и расчеты в реальном времени. |  |  | HP Vertica, Oracle GoldenGate. |
| Райффайзенбанк | Повышение скорости обработки данных. |  | Развитие аналитического CRM, процессов по распознаванию проблемных заемщиков, процессов предотвращения мошенничеств различного характера. | SAS, Oracle. |
| Ситибанк | Управление рисками, противодействие отмыванию доходов, полученных преступным путем, моделирование при разработке новых продуктов. |  |  | Hadoop, Teradata. |
| Уралсиб | Поиск математически оптимального распределения предложений и коммуникаций по портфелю банка, обеспечивающего максимизацию прибыли или другого целевого показателя по выбору пользователя системы. | Параметры предложений, контактная политика, склонность к приобретению того или иного продукта, предпочтения по каналам коммуникации, бюджетные ограничения, риски и прогноз утилизации. |  | IBM Contact Optimization. |
| ОТП Банк | Web-маркетинг и аналитика. Сегментация клиентов и подготовка данных для маркетинговых кампаний, целевые маркетинговые коммуникации, составление и управление многоканальными многоволновыми маркетинговыми кампаниями, расширенная аналитика о реакции клиентов на различные предложения. Автоматизация исходящих звонков в колл-центре. |  |  | IBM Campaign. |
| Всероссийский банк развития регионов | Антифрод. | Поведение владельцев карт, места, которые они посещают, привычки расходования средств. |  |  |
| Бинбанк | Отток, целевой маркетинг и cross-sale. |  |  | CleverDATA. |
| Deutsche Bank | Анализ неструктурированных данных. Переход от старой системы к новой, горизонтально масштабируемой. Сокращение времени на аудит и создание отчетов по риску. | Front end (trading data), the middle (operations data) and the back-end (finance data) | Привести в порядок свои данные. | Hadoop |
| Лето Банк (Поглощен банком ВТБ24) | Целевой маркетинг, слияние риск менеджмента и CRM, управление кредитными рисками. Модификация алгоритмов оценки клиентов и точный расчет целевых сегментов. Сокращение сроков запуска предодобренных кампаний, значительное сокращение времени обработки больших списков клиентов для коммуникации. |  | Концепция Next Best Action. | SAS Campaign Management, SAS RTDM |
| НБКИ | Повышение эффективности скоринговых моделей. | Социальные сети и другие открытые источники данных в интернете. |  | Double Data |

Сейчас ценовые войны за уменьшение комиссии затухают, банки оптимизируют внутренние процессы и стремятся к интероперабельности с внешними отраслями: об этом сигнализирует появление многочисленных кобрендинговых карт и программ лояльности с cashback.

Происходит смещение роли банка с хранилища денег на хранилище информации с персональными советами по тратам и сбережениям. В соответствии со сложившейся ситуацией о принимаемом законе, банкам выгодно утилизировать имеющиеся данные при согласии сторон и предоставлять на их основе сервисы для торговцев и конечных клиентов, которые надолго привяжут потребителя к банку. К таким сервисам прежде всего можно отнести персональных финансовых ассистентов, аналитические панели и программы лояльности.

Создание персонализированного сервиса подразумевает пользу одновременно банкам, торговцам и клиентам. Всем было бы полезно знать больше о клиенте, включая клиента. Особенно остро эта потребность проявляется в период кризиса. Нужно проводить up-sale, cross-sale не только банковских продуктов, но и продуктов клиентов банка.

Торговцы хотят лучше прогнозировать спрос и знать больше о клиенте (включая анализ соцсетей). Сейчас эти потребности закрываются картами лояльности и онлайн-каналами у развитых фирм. Но торговец не знает, сколько денег у клиента и может ли он купить у него что-то еще; возможно, у торговца нет любимого товара постоянных клиентов, вопрос первичного привлечения также актуален. Интересным представляется сегмент маленьких торговцев без уникальных товаров, потому что все товары примерно однородны и оцифрованы (штрих-коды, QR-коды и RFID). В случае операторов фискальных данных, данные чистые (из-за того, что по ним платятся налоги), записываются 1 раз, а читаются много раз, что идеально соответствует идеологии Больших данных.

50% клиентов хотят, чтобы у банков был сервис для контроля бюджета или сервис, помогающий им экономить. Неплохо это работает в Америке. Американцы серьезно относятся к домашней бухгалтерии: начиналось все с записи чеков в тетрадь, потом был переход на электронные программы, затем на сервисы. Интересные продукты, такие как Mint и Quicken, агрегируют данные пользователя из всех банков и всех транзакций, делают персональные рекомендации и позволяют выставлять цели (данные по категориям покупок берут из MCC – merchant category code). Это работает хорошо, потому что в Америке почти все пользуются карточками и почти все банки поддерживают OFX формат. За пределами Америки все иначе - нет такого просто решения. Большинство людей финансово не грамотны и вести самому домашнюю бухгалтерию сложно: нужно время выделить, монотонно, не все аналитики, есть человеческий фактор. В этом свете представляется идеальным решение, которое бы незаметно для пользователя сопоставляло детальную информацию по тратам и раскладывала информацию по категориям (потому как разные люди воспринимают траты по-разному), советовала и знакомила с новыми товарами.

Существует огромная пропасть между покупателем и продавцом в сделке по купле-продаже, и сейчас потребитель действует методом проб и ошибок. Товаров и услуг стало слишком много, их сложно все попробовать и выбрать для себя оптимальный. Люди могут помочь друг другу обменом опыта в выборе подходящих товаров через посредника. Сейчас мало только произвести товар, нужно еще наладить связь с потребителем. Торговцы стремятся удержать клиентов, так как привлечение новых обходится гораздо дороже [31, 44]. Поэтому существуют различные маркетинговые кампании и программы лояльности:

* распродажа товара
* акции, направленные на привлечение новых покупателей
* запуск собственной дисконтной программы

К долгосрочным программам относятся:

1. Дисконтная программа, по которой покупатель получает скидки — на весь ассортимент или на отдельные группы товаров.
2. Накопительная дисконтная программа, где покупатель стимулируется на более дорогостоящие покупки, чтобы получать высокие скидки на все последующие покупки.
3. Организация объединенных магазинов с созданием единой карты лояльности нескольких магазинов, которые не конкурируют друг с другом напрямую. Магазины привлекают новых покупателей и повышают свою узнаваемость
4. Кобрендинговая программа с банком.

Краткосрочные акции:

1. Упаковка дополнительного товара к уже купленным.
2. Купон со скидкой на отдельный товар или весь чек.
3. Бонусы - внутренняя валюта, которая используется для оплаты товаров наряду с обычными деньгами.
4. Скидки, которые привлекают внимание покупателя и создают у него ощущение, что он не только покупает, но при этом еще и экономит.
5. Подарки при покупке определенного товара.

Учитывая все открывшиеся возможности и потребности сторон, необходимо создание сервиса, который бы максимально соответствовал ожиданиям.

* Предлагал выгодные и экономически обоснованные предложения для всех сторон;
* Уменьшал шаги аутентификации (NFC, barcode) и оплаты пользователя при покупке;
* Позволял однозначно идентифицировать пользователя в CRM;
* Позволял получить верифицированные социально-демографические данные о пользователе без заполнения лишних анкет;
* Расширял пользовательский опыт конечных клиентов;
* Формировал индивидуальные предложения;
* Учитывал контекст и индивидуальную воспринимаемость предложений пользователями;
* Давал торговцам возможность взаимодействовать с пользователями, экономить на SMS-информировании;
* Позволял торговцам экономить на эмиссии карт;
* Обеспечивал возможность участия в программе малого и среднего бизнеса;
* Предоставлял простой способ идентификации пользователя на подходе в магазин с персональным информированием с помощью iBeacon;
* Проверял данные дополнительным анкетированием при установке приложения;
* Давал возможность оплачивать покупки через HCE;
* Был хорошо интегрирован с уже имеющейся программой лояльности;
* Предоставлял различные виды аналитических панелей и системы поддержки принятия решений для торговцев;
* Привлекал новых клиентов;
* Обеспечивал возможность работы с данными сторонним разработчикам;

Мобильный девайс – наиболее персональное устройство, которое всегда под рукой, что делает рекомендации действительно персональными. Нужно стимулировать клиента идентифицироваться сразу на кассе: сделать это можно посредством безналичного расчета (тогда данные о клиенте сразу сопоставляются с его покупкой) или же посредством печати на чеке уникального идентификатора, который клиент должен будет считать через приложение после. В любом случае мотивацией клиента будет выступать желание получить скидку, бонусы при покупке, cashback или дополнительные услуги. По сути, теперь клиенту нет необходимости иметь при себе карточки программы лояльности или какие-то физические купоны.

На сегодняшний день среднестатистический потребитель имеет около 30 различных карт в своем кошельке и их количество постоянно растет. Неудобство хранения приводит к выводу карт лояльности из оборота. В тоже время, еще не охвачены регионы, где нет активной конкуренции за клиента и технической возможности поддерживать такие программы. Доля таких клиентов составляет около 30%, но активно снижается из-за развития регионов.

Передача данных от банка к продавцу и насыщение данных не могут происходить без согласия клиента, нужно учитывать требования закона «О персональных данных», спрашивая разрешение или настраивая безакцептный обмен данными по факту пользования инструментом.

## 1.2 Анализ доступных данных и способов их использования

Сведения о расчетах, которые должны содержаться в фискальных документах, хранящихся у операторов фискальных данных, указываются в следующем виде (форме):

1. адрес банковского агента (субагента)
2. адрес оператора по переводу денежных средств
3. адрес платежного агента (субагента)
4. адрес электронной почты покупателя (клиента) в случае передачи ему кассового чека (бланка строгой отчетности) в электронной форме (за исключением контрольно-кассовой техники, применяемой в автономном режиме)
5. место (адрес) осуществления расчета
6. дата, время осуществления расчета
7. идентификационный номер налогоплательщика оператора по переводу денежных средств
8. идентификационный номер налогоплательщика ОФД (за исключением случаев применения контрольно-кассовой техники в автономном режиме)
9. идентификационный номер налогоплательщика пользователя
10. сумма расчета
11. фамилия, имя, отчество (при наличии)
12. идентификационный номер налогоплательщика
13. наименование и место нахождения оператора по переводу денежных средств
14. наименование операции банковского агента (субагента)
15. наименование фискального документа
16. налоги
17. сокращенное наименование налога;
18. ставка налога
19. итоговая сумма налога для всех товаров (работ, услуг, платежей, выплат), указанных в кассовом чеке (бланке строгой отчетности)
20. номер автоматического устройства для расчетов
21. регистрационные номер контрольно-кассовой техники
22. номер смены
23. номер фискального документа
24. номер фискального накопителя
25. порядковый номер кассового чека (бланка строгой отчетности)
26. номер и дата документа-основания корректировки
27. наименование организации
28. признак расчета
29. размер вознаграждения, уплачиваемого плательщиком платежному агенту
30. размер вознаграждения, уплачиваемого физическим лицом в виде общей суммы, включающей в том числе вознаграждение банковского платежного агента (субагента) в случае его взимания
31. реквизиты товара
32. наименование товара (работ, услуг, платежа, выплаты)
33. количество товара (работ, услуг, платежа, выплаты)
34. цена за единицу товара (работы, услуги, платежа, выплаты)
35. общая стоимость товара (работы, услуги, платежа, выплаты) с учетом скидок (наценок)
36. сокращенное наименование налога
37. ставка налога
38. размер налога
39. применяемая система налогообложения при расчете
40. сообщение оператора
41. сумма корректировки
42. сумма корректировки налога
43. сумма расчетов, зафиксированных фискальным накопителем
44. сумма налогов, зафиксированных фискальным накопителем
45. сокращенное наименование налога
46. ставка налога
47. итоговая сумма налогов для всех товаров (работ, услуг, платежей, выплат), указанных в кассовых чеках (бланках строгой отчетности) с момента формирования отчета о фискализации фискального накопителя
48. сумма расчетов за смену
49. сумма налогов за смену
50. сокращенное наименование налога
51. ставка налога
52. итоговая сумма налогов для всех товаров (работ, услуг, платежей, выплат), указанных в кассовых чеках (бланках строгой отчетности) в течение смены
53. номера контактных телефонов платежного агента, поставщика и оператора по приему платежей, а также платежного субагента в случае приема платежа платежным субагентом
54. номера телефонов оператора по переводу денежных средств, банковского платежного агента и банковского платежного субагента (в случае его привлечения банковским платежным агентом)
55. номера телефонов оператора по переводу денежных средств, банковского платежного агента и банковского платежного субагента (в случае его привлечения банковским платежным агентом)
56. тип корректировки
57. фискальный признак документа
58. фискальный признак сообщения
59. фискальный признак подтверждения
60. фискальный признак оператора

Новый законопроект ориентирован на русских потребителей, поэтому крайне важно учитывать особенности менталитета и предпочтения наших граждан. Понимание мотивов потенциального покупателя имеет огромное значение при составлении рекомендации и предложении товаров, которые бы в наибольшей степени соответствовали потребительским ожиданиям касательно баланса стоимости и качества продукции.

С связи с данной интеграцией, следует также учитывать то, что непосредственным потребителем не всегда является тот, на кого оформлена карточка или аккаунт в приложении. К примеру, к одному банковскому счету может быть выпущено несколько карточек. Данные можно частично обновлять за счет того, что в России готовится единая база данных номеров мобильных телефонов, по которой банки смогут установить принадлежность номера мобильного телефона гражданину с соответствующими именем и фамилией.

Мотивов, которыми руководствуются люди при покупке, много, причем, у человека всегда присутствует целая гамма побудительных мотивов, из которых лишь определенные могут иметь существенную значимость и влиять на механизм и результат принятия окончательного решения. Выявив мотивы, можно вырабатывать эффективные приемы, способствующие привлечению пользователя и придания ему статуса постоянного, что, в конечном итоге, прямым или косвенным образом может повлиять на увеличение продаж.

Общий интерес в ряде случаев отличается от потребностей и желаний каждого индивидуума (члена группы). Мотивация покупателей отличается в зависимости от времени, когда принимается решение о покупке и места нахождения человека, поэтому учитывать контекст для формирования предложения крайне важно.

Ритейлеры традиционно занимаются различными видами анализа:

* Анализ динамики товарооборота, прибыли, средней суммы покупки и количества покупок по дням недели
* **Анализ сумм и структуры чеков**
* **Анализ структуры товарооборота и прибыли** **(АВС – анализ)**
* **Анализ эластичности товарооборота**
* XYZ-анализ
* RFM-анализ
* **Анализ эффективности использования торговых площадей**
* Анализ эластичности площади

Наибольший интерес представляет использование алгоритмов продвинутой аналитики на большом массиве данных так как он позволяет выявлять скрытые закономерности при проведении исследований.

## 1.3 Обзор имеющихся подходов выработки рекомендаций

Математические методы, которые применяются при разработке рекомендательных систем, можно разбить на две группы: методы коллаборативной фильтрации (collaborative filtering) и контентные методы (content-based, information filtering). Естественно, возможно одновременное использование методов двух групп (hybrid prediction).

Обычно польза от товара или услуги представлена определенным количеством очков, отражающих, насколько конкретному пользователю понравился конкретный объект. В самом общем виде проблема сводится к присвоению той или иной оценки товару, еще не известного покупателю. Очевидно, такая оценка дается исходя из анализа предшествующих предпочтений данного покупателя или любой другой информации о нем. После этого система предсказывает оценки для еще не известных потребителю товаров: те из них, которые получают наивысшие оценки рекомендуются потенциальному потребителю.

Прогнозирование обычно производится 1) оптимизацией функции полезности, и эмпирическим обоснованием её поведения, или 2) нахождением функции полезности, оптимизирующей заданные параметры поведения, такие как среднеквадратическое отклонение. Потребитель должен получить в качестве рекомендаций товары с самыми высокими оценками из анализированных.

Контентные рекомендательные системы пытаются по истории транзакций предсказать наиболее релевантные результаты (товары, на которые пользователь купит с большей вероятностью). И только товары, обладающие высокой степенью общности с предпочтениями потребителя, будут рекомендованы. Такой подход к получению информации требует создания профилей потребителей со вкусами, предпочтениями, нуждами.

Информация для профилей может быть получена путем анализа совершенных потребителем действий и его метаданных. Последние методики позволяют учесть интересы и предпочтения пользователей и тем самым расширить пользовательский профиль. Необходимо понимать, какое минимальное количество данных требуется получить от пользователя, чтобы этого было достаточно для выработки точных рекомендаций. В начале построение хорошей модели по такому подходу затруднительно, но крайне желательно в будущем, так как оно решает проблему холодного старта, когда нельзя ничего рекомендовать новому пользователю из-за отсутствия истории.

Контентные методы в большинстве случаев основаны на Байесовом классификаторе и методах машинного обучения, включающих деревья решений, искусственные нейронные сети. Однако, контентные рекомендательные системы мыслят слишком узко. Пользователь получает рекомендации только тех товаров, которые были куплены ранее.

Альтернативой контентному подходу может служить коллаборативная фильтрация. Это метод рекомендации, при котором анализируется только реакция пользователей на объекты. Конечной целью метода является как можно более точное предсказание реакция на новые рекомендуемые товары. Чем больше оценок собирается, тем точнее получаются рекомендации. Иными словами, пользователи помогают друг другу в фильтрации объектов. Поэтому такой метод называется также совместной фильтрацией. В случае анализа транзакционных данных применима бинарная оценка: человек купил (или не купил) определенный товар.

Методы коллаборативной фильтрации сталкиваются с проблемой “холодного старта”. Пользователю необходимо показать интерес к довольно большому количеству разных товаров, прежде чем система сумеет правильно понять его предпочтения и дать подходящие рекомендации. Поэтому система не сможет давать точные рекомендации новому потребителю, произведшему очень мало действий.

Также в любой рекомендательной системе, количество оценок, которые необходимо предсказать, обычно намного превышает количество имеющихся оценок. Важно, чтобы система умела эффективно предвидеть оценки, исходя из небольшого количества примеров. Преодолеть проблему разреженности оценок рекомендательной системы можно, если при поиске похожих пользователей использовать информацию о пользователе, содержащуюся в его профиле.

Более того, необходимо наличие критического количества пользователей. Например, товар может не пользоваться популярностью, и тогда он будет рекомендоваться очень редко. Малое же количество рекомендаций может быть сделано обладателям необычных вкусов по сравнению со вкусами большинства, для которых в системе не найдется похожих пользователей. Таким образом, переменная “количество покупок” (которая связана с размером магазина и оборотом товара, которые он предоставляет) не имеет значение, нужно по ней формировать гибридные переменные.

Большая часть рекомендательных методик способом коллаборативной фильтрации основывается на ограниченном понимании пользователей и товаров. За пределом анализа остается профиль пользователя, данные о товарах и другая доступная информация. Например, традиционные коллаборативные алгоритмы ограничиваются только информацией о сделанных покупках. Сами профили все еще остаются слишком примитивными.

Нынешние коллаборативные рекомендательные системы оперируют в двумерном пространстве Пользователь-Товар. Это значит, что они выдают рекомендации, основываясь исключительно на информации о пользователе или о товаре и обходят стороной контекст, который первостепенно важен в нашей предметной области. Например, во многих случаях, полезность товара может зависеть от того, когда происходит потребление.

В таких случаях, простая рекомендация клиенту недостаточна; при выработке рекомендации система должна обратиться к дополнительной контекстуальной информации о времени и обстоятельствах предполагаемого потребления.

В литературе предлагались и другие основанные на различных моделях подходы коллаборативной фильтрации, включающие Байесов анализ, вероятностную релятивистскую модель, модель линейной регрессии, модель максимальной энтропии и графовые рекомендательные системы (Kibeom Leea, Kyogu Leea, b, 2014). Недавно большое количество работ было посвящено поискам более сложных вероятностных моделей коллаборативной фильтрации.

Исследования также рассматривают возможности составления рекомендаций, основанных на геолокации (Bongseok Choia, Ohbyung Kwonb, Bongsik Shinb, 2015), где общее повышение эффективности происходит за счет того, что пользователь обычно пользуется мобильным сервисом на бегу. Ряд работ показывают повышение удовлетворенности пользователей от персонализации (Liang et al, 2007; Tam and Ho, 2005), в то время, как другие работы говорят об обратном эффекте (Lavie et al., 2010). Эффект от внедрения рекомендательной системы индивидуален для каждого внедрения и зависит от структуры данных. Также в одном исследовании (Zhijie Lin, 2014) отмечалось, что рекомендации, выработанные системой, работают хуже, чем отзывы, сформированные пользователями.

Общей проблемой всех рекомендательных систем является то, что они недостаточно быстро реагируют на постоянно изменяющийся поток входных данных. Количество данных увеличивается, матрица становится все более разреженной, а контентная модель не актуальной. Необходимо постоянно перестраивать рекомендации, а для некоторых предметных областей актуальна работа в режиме, близкому к режиму реального времени.  Работа над решением данной проблемы должна вестись путем анализа трафика конкретной системы.

Рекомендательная система работает на увеличение верхней части воронки продаж за счет предоставления скидки на товары, которые потенциально могут заинтересовать человека. Поэтому крайне важно хранить, уметь анализировать, использовать при выработке рекомендаций пользователям и сегментации потребителей данные о контексте потребления. Именно благодаря ему можно понять мотивацию и потребности клиента. Такой подход устраняет основную слабость рекомендательных систем - проблему холодного старта. Но сейчас эти данные хранятся лишь частично. К примеру, состав корзины в контексте с днем недели может определить семейные поездки за продуктами в выходные от покупки на бегу.

Однако, как уже было сказано выше в обзоре подходов, рекомендации по косвенным данным истории транзакций не точны. Для построения действительно хорошей рекомендательной сети нужно привлекать данные по контексту потребления. Сегментирование по нескольким критериям позволяет более точно учесть различные запросы и мотивы потребителей к покупке. Поэтому важно хранить как можно больше данных.  Основной недостаток любой аналитической мотивационной теории является продолжением ее же достоинств: чем подробнее перечень потребностей, тем сложнее провести между ними границы. Таким образом, подобные классификации потребностей нуждаются в некотором обобщении. Генерализацию обычно производят по степени значимости тех или иных потребностей для индивидуума. В этом и есть персонализация, когда адресность услуги является залогом ее продажи.

# Глава 2. Архитектура разрабатываемой системы

Наиболее жизнеспособный подход строится на сборе данных транзакций и их последующей пакетной обработке большими частями. Имея срез данных о действиях покупателя за определённый период, возможно строить сложные модели поведения и применять их для выдачи рекомендаций.

Суть алгоритма заключается в том, что все пользователи уникально определяются по идентификаторам. Эти идентификаторы связываются с товаром из покупки. Таким образом поведение пользователя характеризуется данными о покупках. Подсчитав схожесть каждого подмножества со всеми другими, можно объединить пользователей в группы с похожими предпочтениями.

В условиях нынешней архитектуры основной системы и имеющихся данных, рекомендательную систему следует строить вокруг коллаборативной фильтрации, сочетающей в себе item-based и user-based подходы в оптимальной пропорции. В ритейле, товаров гораздо меньше, чем пользователей, новые поступления происходят не часто. Поэтому больший вес следует давать user-based рекомендациям.

Предположим, что показывается 4 рекомендации на экране мобильного телефона. Тогда, когда пользователь заходит в приложение, он сразу же определяется системой. Если такой пользователь уже зарегистрирован, то в системе есть готовые рекомендации для него, сформированные двумя подходами. На стороне клиента есть данные фильтрации, они должны отправиться на сервер системе, чтобы она могла отфильтровать предполагаемые к показу товары (подготовлено ли для данного пользователя предложение, учитывает ли оно дату показа и местоположение). Таким образом, сформированы два новых списка от двух методов коллаборативной фильтрации. К примеру, первые 3 ячейки заполняются товарами, подобранными user-based системой, оставшиеся - item-based (подбор пропорций будет показан далее). Случай, когда не подобрались рекомендации будет рассмотрен отдельно. Структура программного обеспечения и схема взаимодействия с пользователей представлены ниже на Рисунке 3.

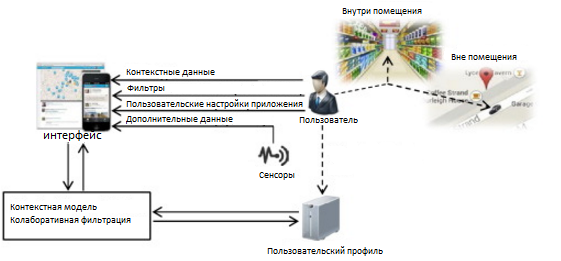


Рисунок 3 Схема взаимодействия пользователя с системой

Диаграмма компонентов в нотации UML 2.0 сервиса представлена на Рисунке 4.

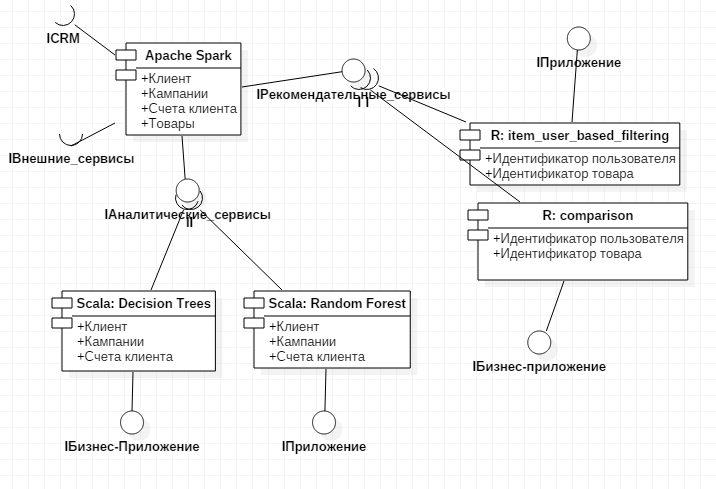


Рисунок 4 Диаграмма компонент

Нужно составлять выборку по логам рекомендаций. То есть нас интересует, какие акции были показаны, когда, что на них было показано, какими характеристиками обладал пользователь и как он реагировал (купил или нет за время действия акции). Из профиля пользователя нужен идентификатор, регион, устройство, браузер и операционная система, дата, результат, а также данные объявления (которые он видит), частота контакта с акцией (гибридная искусственная переменная). Обычно пользователи, которые видят объявление 100-й раз за день, реагируют на него не так же, как те, кто видит в 1-й. Поэтому нужно считать частоту каждого показа для каждого идентификатора пользователя. По этим данным нужно строить модель бинарного выбора, которая бы предсказывала вероятность отклика на акцию при данных характеристиках. Такой подход задействует переменные, доступные сейчас.

При входе в приложение, если клиент не идентифицируется, нужно предложить ему заполнить анкету, чтобы сформировать первичный профиль, который можно задействовать в контентном подходе. При построении контентной модели следует начать с популярных алгоритмов классификации: логистической регрессии, решающих деревьев. После необходимо сравнить эффективность предсказания различных подходов между собой и подобрать оптимальное количество параметров.

Такой подход страхует нас от случаев, когда пользователь первый раз зашел в систему и у нас еще нет рекомендаций от коллаборативной фильтрации или же по результатам коллаборативной фильтрации нам просто нечего рекомендовать (контентный подход не основной). Процесс выглядит следующим образом. Пользователь заходит в приложение, затем предложения ищутся по базе. Каждый найденный товар и его характеристики (которые показываются в объявлении и которые есть в модели), вместе с характеристиками пользователя должны быть обработаны системой. В результате у нас будет вероятность отклика на каждое объявление, товары можно ранжировать по ней, по цене и включать в показ пользователю.

В случае нового пользователя, у которого еще нет покупок, система должна предсказывать товары, которые скорее будут интересны пользователю. В случае наличия сопоставления, у нас больше данных, по которым можно строить контентный прогноз: идентификатор пользователя, идентификатор магазина, дата создания, сумма заказа, способ оплаты. Все эти данные помогают в определении мотива пользователя. Нужно не забывать актуализировать потребности и насыщать новыми данными профиль пользователя.

Следует отметить, что есть необходимость в создание отдельного key-value хранилища, которое будет содержать данные сопоставления клиента и чека потому, что человек может оплатить покупку наличными, а считать чек позже и получить cash back. Установление времени на получение остается на усмотрение банка или торговца.

Если оценивать характер возникновения данных и объемы операций, регулярно совершаемых в стране, становится понятно, что необходимо использовать инструменты, специально предназначенные для хранения и анализа большого количества данных. Классические реляционные базы данных в данным случае не подходят из-за ограничения по масштабированию, которое требует дорогого вертикального наращивания вычислительных мощностей. Для реализации анализа мною была использована платформа для массово-параллельной обработки данных с открытым исходным кодом Apache Spark. Преимущества использования этого фреймворка заключаются в том, что он:

* Оптимизирован для очень большого объема данных (TB, PB)
* Создан для задач, когда данные пишутся один раз, читаются много раз
* Отказоустойчив, не нужно создавать резервные копии
* Репликация данных происходит по мере их накопления
* Может исполняться на дешевом аппаратном обеспечении
* Поддерживает интерактивную, пакетную и потоковую обработку данных
* В несколько раз быстрее парадигмы Map Reduce за счет хранения данных в памяти
* Совместим с остальными элементами экосистемы Hadoop
* Поддерживает несколько языков с одинаковым функционалом
* Способен загружать данные из различных источников
* Есть встроенные библиотеки для различного рода анализа

Для анализа отклика на целевое предложение были использованы методы деревьев решений на основе алгоритма ID3 и использования уровня энтропии (см. Приложение 4) и случайные леса (см. Приложение 5). Такой подход выбран из-за того, что этот метод достаточно прост и хорошо интерпретируется, что полезно в выработке маркетинговой стратегии. Алгоритмы случайного леса обычно дают лучший результат и хорошо поддерживают параллельные вычисления в силу своей специфики. Рекомендации формируются на основе методов коллаборативной фильтрации.

# Глава 3. Практическая реализация сервиса

## 1.1 Разработка рекомендательных алгоритмов

Для демонстрации подхода использованы тестовые данные, предварительно матрица была составлена из имеющихся логов неизвестного сайта.

В item-based подходе задача формулируется следующим образом: Как понять, понравится ли объект данному пользователю? Нужно найти другие объекты, похожие на данный объект, и посмотреть, как пользователь на них реагировал. Вид исходных данных представлен на Рисунке 5.

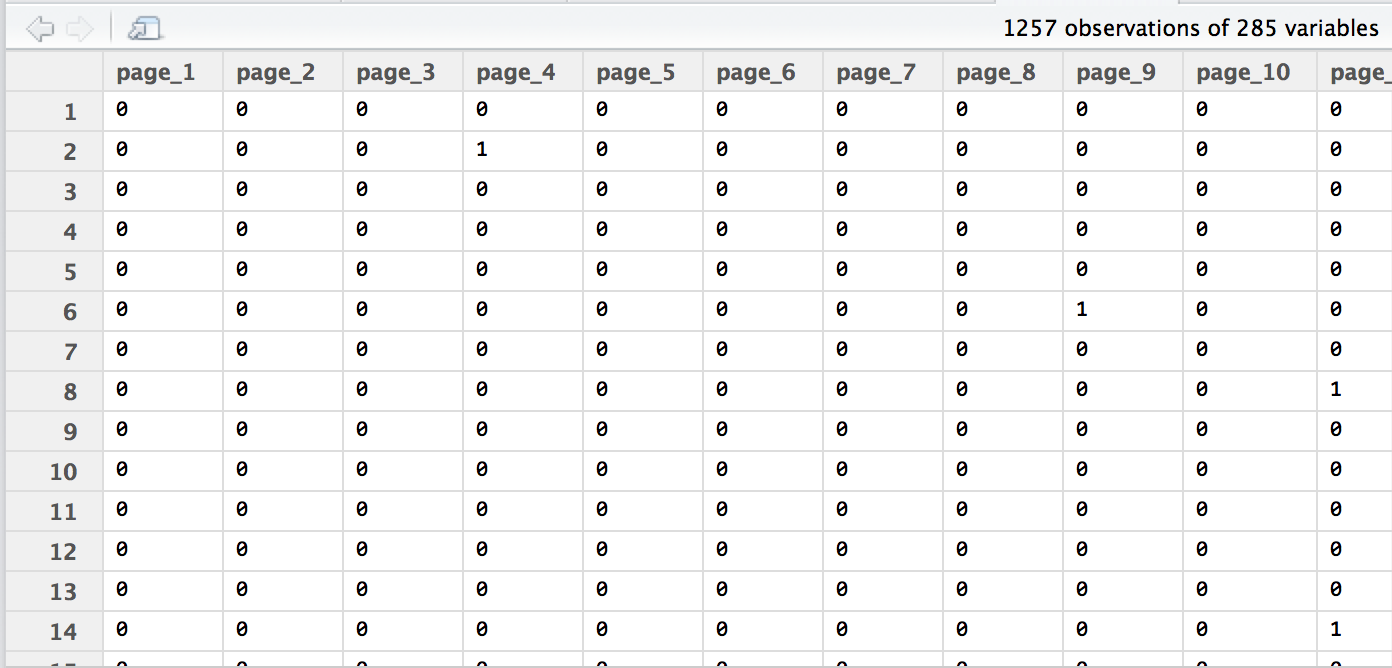


Рисунок 5 Пример матрицы Пользователь – Товар

Далее требуется определить коэффициент сходства объекта исследования (страницы) с остальными объектами для количественного определения их похожести. Каждая строка представляет вектор бинарных значений. Существует множество способов подсчета близости векторов. Один из самых простых - посчитать косинус между этими векторами. Результирующая величина покажет, насколько одна транзакция похожа на другую.

Затем нужно выбрать множество товаров, наиболее похожих на другие товары. Есть несколько способов выбора. Чаще всего фиксируется целая константа, например, значение, равное 10. Затем все товары сортируются по убыванию меры близости.

Итоговая матрица представлена на Рисунке 6.

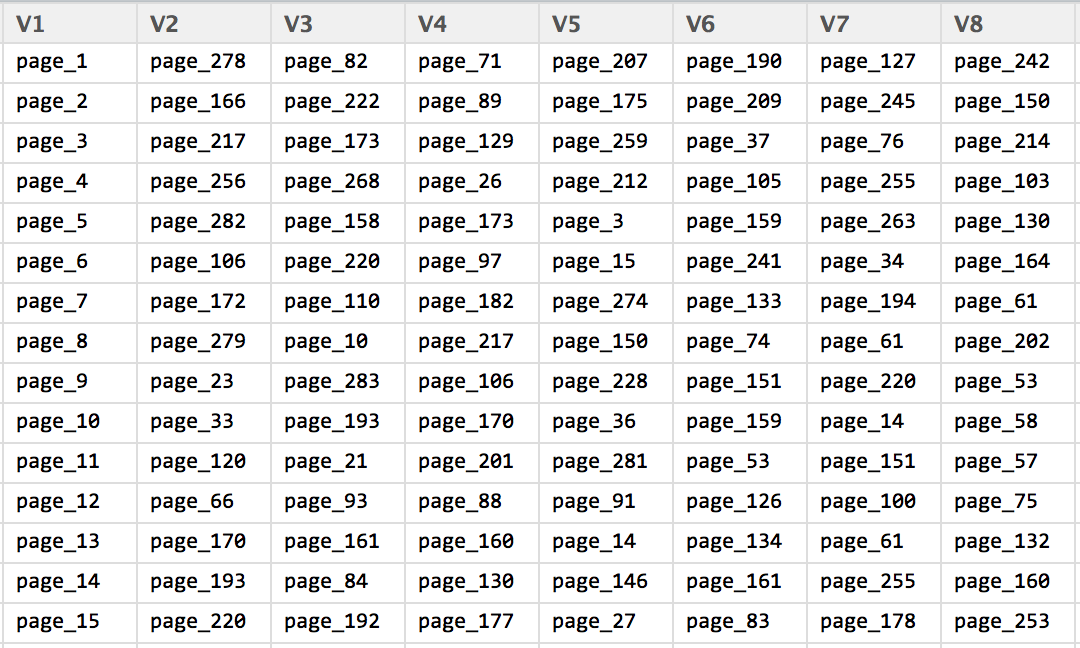


Рисунок 6 Пример составленных рекомендаций

В user-based подходе процесс похож. Задача формулируется следующим образом: Как понять, понравится ли новому пользователю определенный объект? Нужно просто найти других пользователей, похожих на данного, и посмотреть, как они восприняли этот объект.

Мы берем объект и смотрим, какие пользователи его выбрали, затем находим соседей-пользователей с такими же интересами. Затем мы создаем рейтинг объекта для пользователя (см. Приложение 2).

## 1.2 Анализ эффективности различных рекомендательных алгоритмов

Для демонстрации подхода использованы тестовые данные (см. Приложение 3). Предварительно матрица была составлена из имеющихся логов сайта. Для выявления рекомендаций на бинарной матрице, был применен ряд подходов:

* случайный выбор объекта (для сравнения со всеми другими)
* “популярные” объекты - на основании коэффициента корреляции Пирсона, исключая предметы, которые пользователь уже выбирал
* user-based коллаборативная фильтрация
* item-based коллаборативная фильтрация
* ассоциативные правила

ROC-анализ показал, что на уровне 5 ближайших объектов/соседей разницы между ведущими алгоритмами нет. На уровне анализа в 20 ближайших объектов/соседей чуть лучше ведет себя user-based коллаборативная фильтрация. Такие результаты могли получится вследствие эндогенности. То есть данные представляют собой не просто срез, а временные ряды, предпочтения людей меняются со временем и необходимо введение дополнительных переменных в анализ. Визуализация сравнения различных подходов представлена на Рисунке 7.

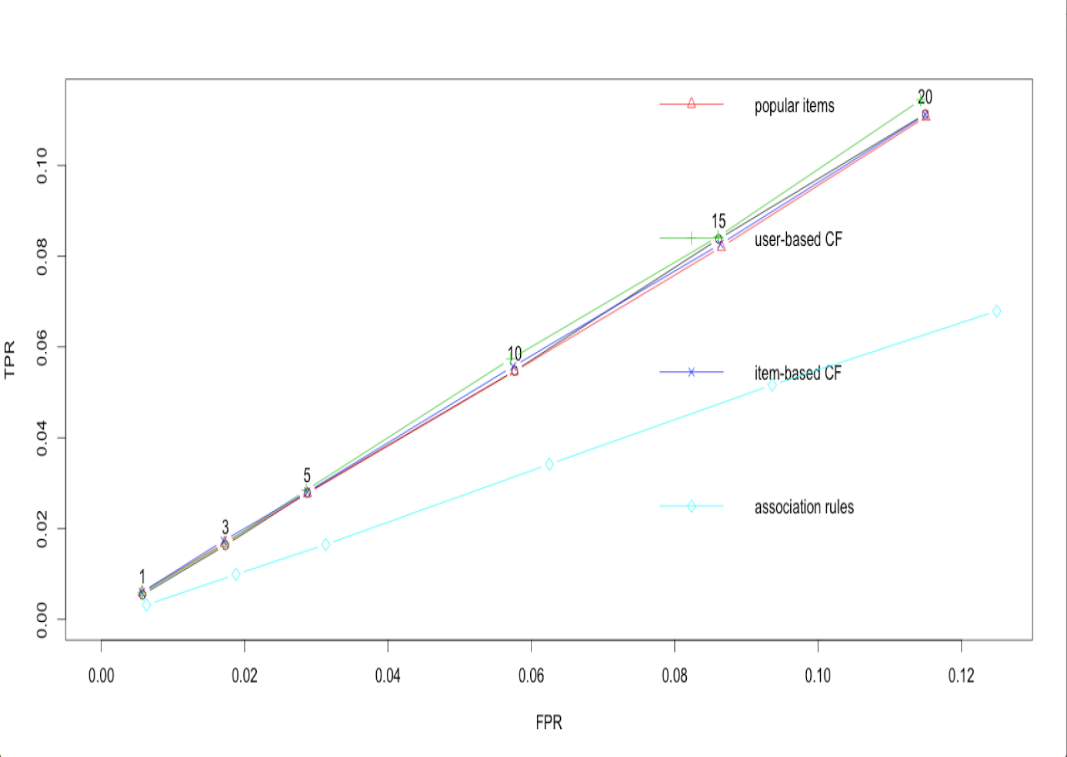


Рисунок 7 Сравнение результатов работы алгоритмов

Также стоит отметить, что по этим же причинам существенно проигрывают ассоциативные правила (которые в основном используются в задаче анализа рыночной корзины “market basket analysis”). Сравнение результатов работы алгоритмов по метрикам precision и recall представлен на Рисунке 8.

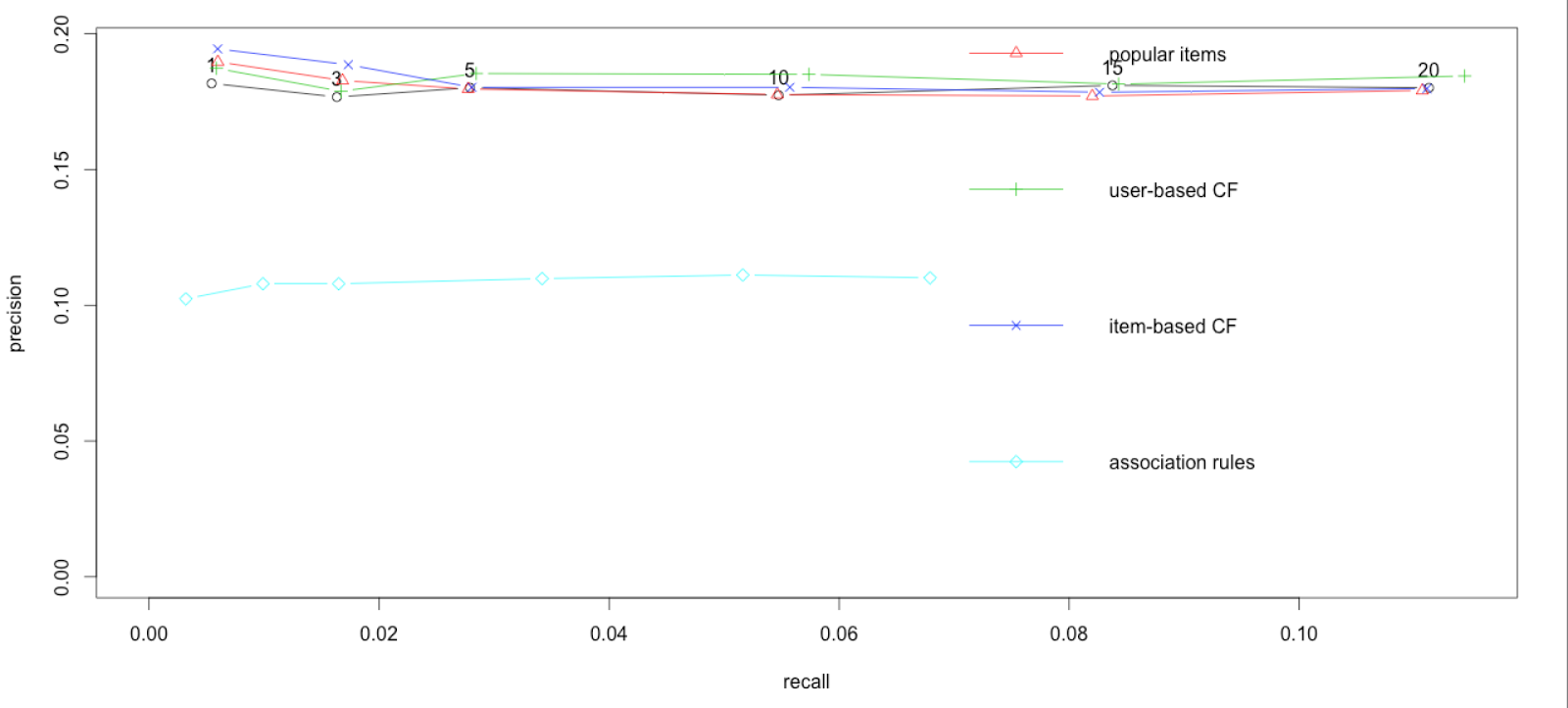
****

Рисунок 8 Сравнение результатов работы алгоритмов

## 1.3 Контентные методы

Для успешной работы с новыми покупателями, привлечения наиболее релевантной аудитории в магазин, необходимо дополнить рекомендательную систему, использующую методы коллаборативной фильтрации контентным подходом. Нужно задействовать имеющийся профиль пользователя и данные о покупках в других магазинах, чтобы формировать релевантные предложения новому покупателю. Однако, сегментация покупателей и создание профилей полезны не только для решения проблемы «холодного старта», но и для оценки лояльности пользователей, выявления склонных к оттоку постоянных клиентов.

Можно улучшать работу с клиентами, опираясь на анализ типов лояльности и восприимчивость к предложениям, корректируя их с точки зрения максимизации дохода и развития бизнеса. Угадывание потребности позволяет делать более выгодные предложения и получать большую прибыль. Нужно учитывать участие в маркетинговых мероприятиях, реакцию на информационные сообщения, контекст потребления. Однако, основная сложность связана с тем, что сегменты, выделенные на основе анализа прибыльности клиентов, не совпадают полностью с сегментами, сформированными по демографическим признакам.

Скидки нужно предоставлять клиентам, учитывая их ценность для компании, иначе затраты на их удержание могут превысить доходы от работы с ними. Тем более приверженность клиентов конкретной компании по факту определяется не только активными действиями торговых компаний по их удержанию и полезными персональными предложениями, но и совершенно другими факторами, среди которых можно отметить: местоположение магазина, цены товаров, их качество, уровень сервиса и др.

Решается задача выбора оптимального предложения путем определения характеристик потенциального предложения и выявления покупателей, которые соответствуют этим характеристикам. Классификационная модель определяет принадлежность покупателя по ряду признаков его поведения к одному из классов (откликнется или нет). В результате построения классификационных моделей на основе данных о пользователях и их реакциях на предыдущие предложения были определены характеристики покупателей, склонных к отклику. Традиционно, строится несколько моделей, после чего определяется наиболее точная. Для анализа мною было построено 2 модели при помощи различных методов на одинаковом наборе данных. Следующий этап после построения моделей - это оценка их точности на контрольных выборках по нескольким критериям.

Переменные, которые используются для создания выборки:

* Age – возраст клиента;
* Job – сфера деятельности;
* Marital – семейное положение;
* Education – уровень образования;
* Default – результат выдачи предыдущего банковского займа;
* Balance – количество денег на всех счетах
* Housing – тип жилья;
* Loan – есть ли сейчас долг;
* Contact – канал взаимодействия, по которому клиент узнал об акции;
* Day – день, в который произошел контакт;
* Month – месяц, в который произошел контакт;
* Duration – количество дней до конца предложения с момента контакта;
* Campaign – тип предложения, который использовался ;
* Pdays – сколько дней назад была проведена более ранняя кампания;
* Previous – тип более ранней кампании;
* Poutcome – результат проведения более ранней кампании;
* Y – результат кампании, зависимая переменная;

Для быстрой выработки основных правил использовался метод деревьев решений, основанный на алгоритме ID3 с использованием энтропии. Так как в наборе данных 17 переменных, не многие из которых коррелируют друг с другом, то оптимальная глубина дерева установилась на уровне 4, что обеспечивает приемлемую скорость выработки первичных результатов. Встроенный метод анализа сам делает решение о включении в модель наиболее важных предикторов. Созданная модель легко интерпретируется, хорошо обрабатывает пропущенные значения и быстро обучается. К недостаткам такого подхода можно отнести ограничения по точности и плохую работу при большом количестве предикторов.

Случайные леса относятся к числу наиболее точных и популярных алгоритмов, представляющий собой ансамбль моделей, основанный на деревьях решений с последующим голосованием за лучший результат. Такой метод очень эффективен при большом количестве предикторов, хорошо параллелится и работает с пропущенными данными. Однако, этот метод требует больших затрат ресурсов и времени.

Для того, чтобы уменьшить требования по памяти и мощностям процессора, мною учитывалась слабая корреляция, которая позволила отобрать на основе корреляции только необходимые независимые переменные. Сократив число предикторов, уровень корреляции остался тем же, обеспечивая такое же качество классификации. Оценка результатов работы указанных подходов представлена в Таблице 2.

Таблица 2.

Оценка результатов работы алгоритмов

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Экспертная оценка | | | |
| Оценка системы |  | Random forest | | Decision Trees | |
| + | - | + | - |
| + | 14 | 7 | 6 | 5 |
| - | 7 | 31 | 11 | 21 |
| Precision | | 0,6 | | 0,54 | |
| Recall | | 0,6 | | 0,35 | |
| F-мера | | 0,6 | | 0,42 | |
| Accuracy | | 0,76 | | 0,63 | |

Таким образом, алгоритм случайных деревьев показал наилучший баланс полноты и точности. Нужно отметить, что он достаточно жадный до ресурсов, но хорошо проявляет себя при параллельной обработке данных и масштабируется. Прикладная ценность данной работы заключается в возможность воздействовать на покупателей адресным предложением, не надоедая частотой. В итоге для покупателей с различными характеристиками были разработаны специальные предложения, которые имеют индивидуальный характер.

# Заключение

В результате автором была изучена рыночная ситуация, потребности сторон, выделены наиболее перспективные области развития банковских сервисов и продуктов с учетом требований регулятора, выраженных в готовящемся законопроекте. Был произведен анализ данных, к которым получит доступ новый агент и определены возможности их использования. После выделения основных параметров будущей системы и данных, которые необходимо использовать, была произведена настройка рабочих сред для реализации проекта: развертывание и настройка серверов для хранения данных и анализа, которые позволили разработать программный продукт для хранения, трансформации сырых данных и их обработки средствами интеллектуального анализа. Данные были обработаны алгоритмами интеллектуального анализа, после чего было произведено сравнение результатов их работы алгоритмов на тестовых данных и разработана контентная модель, основанная на банковских данных о пользователе и результате предыдущих кампаний.

Конкретная техническая реализация заключается в создании автором программы формирования рекомендаций на языке R и дополнении ее результатами контентной модели, реализованной на языке Scala на программной платформе Apache Spark. Практическая значимость такого сервиса состоит в возможности улучшения пользовательского опыта, расширения знаний о клиенте банка и торговцев, создания открытого сервиса для наилучшей утилизации данных сторонними разработками и экономии на эмиссии банковских карт и карт лояльности.

За последние годы был совершен значительный прогресс в развитии рекомендательных систем. Были предложены контентные, коллаборативные и гибридные алгоритмы выработки рекомендаций. Некоторые системы нашли практическое применение в коммерческой индустрии. Тем не менее, несмотря на прогресс, для более эффективной работы в большом списке приложений современное поколение рекомендательных систем требует дальнейших улучшений.

Все еще существуют различные ограничения, с которыми сталкиваются современные рекомендательные методы, для которых описаны необходимые улучшения, способные сделать работу банка и ритейлера более эффективной. К таким улучшениям относятся, в ряду прочих, усовершенствованное моделирование пользователей и товаров, включение контекстной информации в рекомендательный процесс.

# Список литературы

1. Big Data Alchemy: How can Banks Maximize the Value of their Customer Data? / Capgemini Consulting, 2014 – 16 стр.
2. *D. Jannach, M. Zanker, A. Felfernig, G. Friedrich* Recommender Systems. An Introduction. New York: Cambridge University Press 32 Avenue of the Americas, 2011. 352 P.
3. *F. Ricci, L. Rokach, B. Shapira, P.B. Kantor* Recommender Systems Handbook. Springer, 2011. 842 P.
4. *G. Miner, R. Nisbet, J. Elder* IV Handbook of Statistical Analysis and Data Mining Applications. - 1 изд. Elsevier Publishing, 2009.
5. GlowByte Consulting внедрила IBM Campaign в ОТП Банке [Электронный ресурс] / Пресс-центр GlowByte Consulting. Режим доступа: http://gbconsulting.ru/about/press-center/company-news/2015/08/13/glowbyte-vnedril-ibm-campaign-otp-bank/, свободный. (дата обращения: 13.07.15).
6. *Joshua D. Suereth.* Scala in Depth// Manning, 2012.
7. *Kemp, Simon*. «Digital in 2016.» We Are Social. N.p., 27 Jan. 2016. Web. / http://wearesocial.com/uk/special-reports/digital-in-2016
8. *Kibeom Leea, Kyogu Leea*. Escaping your comfort zone: A graph-based recommender system for finding novel recommendations among relevant items // Expert Systems with Applications, 2015
9. *Limayem et al, M. Limayem, S.G. Hirt, C.M.K. Cheung.* How habit limits the predictive power of intention: the case of information systems continuance // MIS Q., 31 (4) (2007), pp. 705–737
10. *Lavie et al., T. Lavie, M. Sela, I. Oppenheim, O. Inbar, J. Meyer.* User attitudes towards news content personalization. Int. J. Hum. Comput. Stud., 68 (8) (2010), pp. 483–495
11. *Mike Frampton*. Mastering Apache Spark // Packt Publishing. 2015.
12. *P. Melville, V. Sindhwani* Recommender systems. Encyclopedia of Machine Learning. 2010.
13. R Graphics Cookbook / Winston Chang. – USA: O'Reilly, 2012 – 522 с.
14. *Cay Horstmann*. Scala for the Impatient, Basic Classifiers// Addison-Wesley 2012.
15. *Shalizi С*. Categorizing Data Vectors Types of Categorization, Basic Classifiers, Finding Simple Clusters in Data // Data Mining. 2009.
16. Social Link [Электронный ресурс] /НБКИ. Режим доступа: http://www.nbki.ru/servicescredit/social/connections/ свободный. (дата обращения: 13.05.15).
17. *T. Hastie, R. Tibshirani, J. Friedman* The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - Second Edition. Springer, 2009.
18. *U. von Luxburg, S. Ben-David* Towards a Statistical Theory of Clustering // School of Computer Science, University of Waterloo, Canada. - 2015.
19. *Y. Koren* Collaborative Filtering with Temporal Dynamics. KDD’09. 2009
20. *X. Su, T.M. Khoshgoftaar* Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence. 2009.
21. *W.P. Jones, G.W. Furnas* Pictures of Relevance: A Geometric Analysis of Similarity Measures. Journal of the American society for information science. 1987.
22. *Zhijie Lin.* An empirical investigation of user and system recommendations in e-commerce// Decision Support Systems. 2014.
23. *А. Гомзин, А. Коршунов* Системы рекомендаций: обзор современных подходов. Пре-принт. Москва: Труды Института системного программирования РАН. 2012. 20 С.
24. Банк УРАЛСИБ развивает маркетинговую платформу с помощью IBM Contact Optimization [Электронный ресурс] / Пресс-центр GlowByte Consulting. Режим доступа: http://gbconsulting.ru/about/press-center/company-news/2015/05/12/bank-uralsib-razvivaet-marketingovuyu-platformu-s-ibm-contact-optimization/, свободный. (дата обращения: 12.03.16).
25. Банк "Открытие" внедрил решение HP Vertica для анализа big data [Электронный ресурс] / Пресс-центр банка Открытиe. Режим доступа: https://www.openbank.ru/about/press/721/, свободный. (дата обращения: 12.03.16).
26. Бинбанк присоединился к числу пользователей биржи данных [Электронный ресурс] / CleverDATA. Режим доступа: http://cleverdata.ru/binbank-birzha/, свободный. (дата обращения: 12.03.16).
27. Большие данные не стали мэйнстримом в российских банках [Электронный ресурс] / Cnews. Режим доступа: http://www.cnews.ru/news/top/bolshie\_dannye\_ne\_stali\_mejnstrimom/, свободный. (дата обращения: 16.02.16).
28. Воронцов К. В. Методы кластеризации // 2015.
29. Заготовка килобайтов банками [Электронный ресурс] / Коммерсантъ. Режим доступа: http://www.kommersant.ru/doc/2609914, свободный. (дата обращения: 13.07.15).
30. Искусственный интеллект может не только экономить, но и зарабатывать [Электронный ресурс] / Bankir.ru. Режим доступа: http://bankir.ru/publikacii/20160422/iskusstvennyi-intellekt-mozhet-ne-tolko-ekonomit-no-i-zarabatyvat-10007471/, свободный. (дата обращения: 1.01.16).
31. Исследование рынка банковских программ лояльности 2015 / Frank Research Group, 2015 – 45 стр.
32. Как устроен рынок big data в России [Электронный ресурс] / Rusbase. Режим доступа: http://rusbase.com/howto/big-data-in-russia/, свободный. (дата обращения: 13.03.16).
33. Когда данных слишком много: банки на передовой Big Data [Электронный ресурс] / Slon. Режим доступа: https://slon.ru/specials/data-economics/articles/bankiri/, свободный. (дата обращения: 07.02.16).
34. Количество платежных карт, эмитированных кредитными организациями, по типам карт [Электронный ресурс] / Центральный банк РФ. Режим доступа: http://www.cbr.ru/statistics/p\_sys/print.aspx?file=sheet013.htm, свободный. (дата обращения: 10.01.16).
35. Конференция ICBDA [Электронный ресурс] / ICBDA. Режим доступа: http://2015.icbda2015.org/, свободный. (дата обращения: 13.07.15).
36. Конференция SAS Forum 2015, Retail Analytics trends [Электронный ресурс] / SAS. Режим доступа: http://www.sas.com/content/dam/SAS/ru\_ru/image/events/SAS-Forum-Russia/Presentation/Retail-Intelligence/06\_Retail-Analytics-Trends-(SFR2015).pdf, свободный. (дата обращения: 29.12.15).
37. Конференция SAS Forum 2015, Merchandise intlligence [Электронный ресурс] / SAS. Режим доступа: http://www.sas.com/content/dam/SAS/ru\_ru/image/events/SAS-Forum-Russia/Presentation/Retail-Intelligence/07\_Merchandise-Intelligence-(SFR2015).pdf, свободный. (дата обращения: 25.12.15).
38. Ляпина Екатерина. Разработка рекомендательного сервиса для СМИ на основе анализа больших данных, 2015 – 40 стр.
39. Отчет о конференции Big Data и BI Day [Электронный ресурс] / TAdviser. Режим доступа: http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:Big\_Data\_%D0%B8\_BI\_DAY\_2015-2, свободный. (дата обращения: 26.08.15).
40. Расстояние между объектами (кластерами) и мера близости // bibliotekar.ru URL: http://www.bibliotekar.ru/economicheskaya-statistika-2/11.htm (дата обращения: 20.05.2015).
41. Райффайзенбанк: Технологии Big Data помогают управлять клиентской информацией [Электронный ресурс] / Cnews. Режим доступа: http://www.cnews.ru/articles/rajffajzenbank\_tehnologii\_big\_data, свободный. (дата обращения: 15.02.16).
42. Свыше половины населения РФ после покупки хранят кассовые чеки [Электронный ресурс] / Сервис ЗаеМ. Режим доступа: http://zaew.ru/53%25-rossijan-hranjat-cheki, свободный. (дата обращения: 20.05.16).
43. *Сидоров, Илья*. «Покупки с мобильных устройств. Участие мобильных устройств в принятии решения о покупке.» День интернет-рекламы. Google Россия. Режим доступа: ttp://msk.advdays.ru/upload/iblock/852/Покупки%20с%20мобильных%20устройств%20(Google%20Russia).pdf, свободный. (дата обращения: 31.12.15).
44. Тенденции развития банковских программ лояльности [Электронный ресурс] / Bankir.ru. Режим доступа: http://bankir.ru/publikacii/20160519/tendentsii-razvitiya-bankovskikh-programm-loyalnosti-10007558/, свободный. (дата обращения: 19.05.16).
45. Федеральный закон от 22.05.2003 N 54-ФЗ (ред. от 08.03.2015) "О применении контрольно-кассовой техники при осуществлении наличных денежных расчетов и (или) расчетов с использованием платежных карт" [Электронный ресурс] / Консультант Плюс. Режим доступа: https://www.consultant.ru/document/Cons\_doc\_LAW\_42359/свободный. (дата обращения: 20.05.16).
46. Целевой маркетинг. 10 заметок финансового стартапа [Электронный ресурс] / Ре>банкинг. Режим доступа: http://rebanking.ru/tselevoy-marketing-10-zametok-finansovogo-startapa/ свободный. (дата обращения: 03.10.15).

Приложения

Приложение 1

Настройка программной среды

#Install Java 8

##########################################################################

sudo apt-add-repository ppa:webupd8team/java

sudo apt-get update

sudo apt-get install oracle-java8-installer

java -version

#Install Scala

##########################################################################

wget http://www.scala-lang.org/files/archive/scala-2.11.8.deb

sudo dpkg -i scala-2.11.8.deb

sudo apt-get install scala

cd ~

sudo cat >> .bashrc

export SCALA\_HOME=/usr/local/src/scala/scala-2.­11.8

export PATH=$SCALA\_HOME/bin:$PATH

#Install Git

##########################################################################

sudo apt-get install git

#Install Spark

##########################################################################

wget http://apache-mirror.rbc.ru/pub/apache/spark/spark-1.6.0/spark-1.6.0.tgz

tar -xvf spark-1.6.0.tgz

cd spark-1.6.0

sbt/sbt assembly

#Spark Web UI

#It is needed to open port 8080 via Azure VM endpoints for our VM

#Spark Standalone cluster comes with administrative web UI

#Our working environment: http://sparklyapina.cloudapp.net:8080

#FTP

##########################################################################

sudo apt-get install vsftpd

sudo vi /etc/vsftpd.conf

local\_enable=YES

write\_enable=YES

anonymous\_enabled=NO

pasv\_enable=YES

pasv\_address=[Your public address]

pasv\_min\_port=12000

pasv\_max\_port=12003

sudo service vsftpd restart

#R 3.2.1

##########################################################################

sudo sh -c 'echo "deb http://cran.rstudio.com/bin/linux/ubuntu trusty/" >> /etc/apt/sources.list'

gpg --keyserver keyserver.ubuntu.com --recv-key E084DAB9

gpg -a --export E084DAB9 | sudo apt-key add -

sudo apt-get update

sudo apt-get -y install r-base

sudo apt-get -y build-dep libcurl4-gnutls-dev

sudo apt-get -y install libcurl4-gnutls-dev

sudo apt-get install libxml2-dev

sudo apt-get install gdebi-core

sudo apt-get install libapparmor1 # Required only for Ubuntu, not Debian

wget http://download2.rstudio.org/rstudio-server-0.98.1103-amd64.deb

sudo gdebi rstudio-server-0.98.1103-amd64.deb

Приложение 2

Коллаборативная фильтрация

data <- read.csv(file="Data.csv")

data$X <- NULL

#Удаляем столбец с номерами пользователей

data.ibs <- (data[,!(names(data) %in% c("X", "user"))])

getCosine <- function(x,y)

{

this.cosine <- sum(x\*y) / (sqrt(sum(x\*x)) \* sqrt(sum(y\*y)))

return(this.cosine)

}

Сравним между собой сами страницы и составим матрицу их схожести:

data.ibs.similarity <- matrix(NA, nrow=ncol(data.ibs), ncol=ncol(data.ibs),

dimnames=list(colnames(data.ibs), colnames(data.ibs)))

for(i in 1:ncol(data.ibs)) {

# Loop through the columns for each column

for(j in 1:ncol(data.ibs)) {

# Fill in placeholder with cosine similarities

data.ibs.similarity[i,j] <- getCosine(as.matrix(data.ibs[i]),as.matrix(data.ibs[j])) }}

data.ibs.similarity <- as.data.frame(data.ibs.similarity)

data.neighbours<-matrix(NA,nrow=ncol(data.ibs.similarity),ncol=11,dimnames=list(colnames(data.ibs.similarity)))

for(i in 1:ncol(data.ibs)){

data.neighbours[i,]<-(t(head(n=11,rownames(data.ibs.similarity[order(data.ibs.similarity[,i],decreasing=TRUE),][i])))) }

sumproduct(purchaseHistory, similarities)/sum(similarities).

getScore <- function(history, similarities)

{

x <- sum(history\*similarities)/sum(similarities)

}

holder<-matrix(NA,nrow=nrow(data),ncol=ncol(data)-1,dimnames=list((data$user),colnames(data[-1])))

for(i in 1:nrow(holder))

{

# Пробегаемся по объектам-колонкам

for(j in 1:ncol(holder))

{

# Считываем из ячейки объект и пользователя

user <- rownames(holder)[i]

product <- colnames(holder)[j]

# Делаем пустой строкой те колонки, которые пользователь уже купил

if(as.integer(data[data$user==user,product]) == 1)

{

holder[i,j]<-""

} else {

# Используем уровень похожести в 10 пользователей

topN<-((head(n=11,(data.ibs.similarity[order(data.ibs.similarity[,product],decreasing=TRUE),][product]))))

topN.names <- as.character(rownames(topN))

topN.similarities <- as.numeric(topN[,1])

# Удаляем первого, потому что у него будет похожесть, равная 1

topN.similarities<-topN.similarities[-1]

topN.names<-topN.names[-1]

# Смотрим историю пользователя за 10 покупок

topN.purchases<- data[,c("user",topN.names)]

topN.userPurchases<-topN.purchases[topN.purchases$user==user,]

topN.userPurchases <- as.numeric(topN.userPurchases[!(names(topN.userPurchases) %in% c("user"))])

# Считаем рейтинг объекта и пользователя

holder[i,j]<-getScore(similarities=topN.similarities,history=topN.userPurchases)

} } }

data.user.scores <- holder

Проранжируем наши рекомендации:

data.germany.user.scores.holder<-matrix(NA,nrow=nrow(data.germany.user.scores),ncol=100,dimnames=list(rownames(data.germany.user.scores)))

for(i in 1:nrow(data.germany.user.scores))

{

data.germany.user.scores.holder[i,]<-names(head(n=100,(data.germany.user.scores[,order(data.germany.user.scores[i,],decreasing=TRUE)])[i,]))

}

  Приложение 3

Сравнение подходов

#install.packages("recommenderlab")

library(recommenderlab)

library(reshape2)

library(ggplot2)

#Загружаем данные в формат, удобный для пакета

data <- read.csv(file="Data.csv")

data$X <- NULL

data[,1]<-NULL

for (i in 1:ncol(data)) names(data)[i] <- i

#у нас 1257 строк, 285 колонок

data<-as.matrix(data)

m <- matrix(sample(c(0,1), 358245, replace=TRUE), nrow=nrow(data), ncol=ncol(data), dimnames=list(users=paste("u", 1:nrow(data), sep=''), items=paste("i", 1:ncol(data), sep='')))

matrix <- as(m, "binaryRatingMatrix")

matrix

scheme <- evaluationScheme(matrix[1:1257,], method="split", train=.9, k=1, given=110)

#scheme <- evaluationScheme(matrix[1:1257], method="cross", k=4, given=110, goodRating=1)

#algorithms <- list( "random items" = list(name="RANDOM", param=NULL), "popular items" = list(name="POPULAR", param=NULL),"user-based CF" = list(name="UBCF", param=list(method="Cosine",nn=5)))

algorithms <- list("random items" = list(name = "RANDOM", param = NULL), "popular items" = list(name = "POPULAR", param = NULL), "user-based CF" = list(name = "UBCF", param = list(method = "Jaccard", nn = 5)), "item-based CF" = list(name = "IBCF", param = list(method = "Jaccard", k = 50)), "association rules" = list(name = "AR",param = list(supp = 0.01, conf = 0.2, maxlen = 2)))

results <- evaluate(scheme, algorithms, n=c(1, 3, 5, 10, 15, 20))

results

getConfusionMatrix(results)[[1]]

plot(results, annotate=TRUE)

plot(results, "prec/rec", annotate=TRUE)

Приложение 4

Алгоритм формирования дерева решений

//Decision Trees

val datadir = "/home/azureuser/data"

//Create a SQL Context from Spark context

val sqlContext = new org.apache.spark.sql.SQLContext(sc)

import sqlContext.implicits.\_

//Load the CSV file into a RDD

val bankData = sc.textFile(datadir + "/bank.csv")

bankData.cache()

bankData.count()

//Remove the first line (contains headers)

val firstLine=bankData.first()

val dataLines = bankData.filter(x => x != firstLine)

dataLines.count()

//Convert the RDD into a Dense Vector

//Change labels to numeric ones

import org.apache.spark.mllib.linalg.{Vector, Vectors}

import org.apache.spark.mllib.regression.LabeledPoint

def transformToNumeric( inputStr : String) : Vector = {

val attList=inputStr.split(";")

val age=attList(0).toFloat

//convert outcome to float

val outcome = attList(16).contains("no") match {

case true => 1.0

case false => 0.0

}

//create indicator variables for single/married

val single = attList(2).contains("single") match {

case true => 1.0

case false => 0.0

}

val married = attList(2).contains("married") match {

case true => 1.0

case false => 0.0

}

val divorced = attList(2).contains("divorced") match {

case true => 1.0

case false => 0.0

}

//create indicator variables for education

val primary = attList(3).contains("primary") match {

case true => 1.0

case false => 0.0

}

val secondary = attList(3).contains("secondary") match {

case true => 1.0

case false => 0.0

}

val tertiary = attList(3).contains("tertiary") match {

case true => 1.0

case false => 0.0

}

//convert default to float

val default = attList(4).contains("no") match {

case true => 1.0

case false => 0.0

}

//convert balance amount to float

val balance = attList(5).contains("no") match {

case true => 1.0

case false => 0.0

}

//convert loan to float

val loan = attList(7).contains("no") match {

case true => 1.0

case false => 0.0

}

//Filter out columns not wanted at this stage

val values= Vectors.dense(outcome, age, single, married,

divorced, primary, secondary, tertiary,

default, balance, loan )

return values

}

//Change to a Vector

val bankVectors = dataLines.map(transformToNumeric)

bankVectors.cache()

bankVectors.collect()

//Statistical Analysis

import org.apache.spark.mllib.stat.{MultivariateStatisticalSummary, Statistics}

val bankStats=Statistics.colStats(bankVectors)

bankStats.min

bankStats.max

val colMeans=bankStats.mean

val colVariance=bankStats.variance

val colStdDev=colVariance.toArray.map( x => Math.sqrt(x))

Statistics.corr(bankVectors)

//Transform to a Data Frame

//Drop columns with low correlation

def transformToLabelVectors(inStr : Vector ) : (Float,Vector) = {

val values = ( inStr(0).toFloat,

Vectors.dense(inStr(1),inStr(2),inStr(3),

inStr(4),inStr(5),inStr(6),inStr(7),

inStr(8),inStr(9),inStr(10)))

return values

}

val bankLp = bankVectors.map(transformToLabelVectors)

bankLp.collect()

val bankDF = sqlContext.createDataFrame(bankLp).toDF("label","features")

bankDF.select("label","features").show(10)

//Indexing needed as pre-req for Decision Trees

import org.apache.spark.ml.feature.StringIndexer

val stringIndexer = new StringIndexer()

stringIndexer.setInputCol("label")

stringIndexer.setOutputCol("indexed")

val si\_model = stringIndexer.fit(bankDF)

val indexedBank = si\_model.transform(bankDF)

indexedBank.select("label","indexed","features").show()

indexedBank.groupBy("label","indexed").count().show()

//Split into training and testing data

val Array(trainingData, testData) = indexedBank.randomSplit(Array(0.9, 0.1))

trainingData.count()

testData.count()

import org.apache.spark.ml.classification.DecisionTreeClassifier

import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

//Create the model

val dtClassifier = new DecisionTreeClassifier()

dtClassifier.setMaxDepth(4)

dtClassifier.setLabelCol("indexed")

val dtModel = dtClassifier.fit(trainingData)

dtModel.numNodes

dtModel.depth

//Predict on the test data

val predictions = dtModel.transform(testData)

predictions.select("prediction","indexed","label","features").show()

val evaluator = new MulticlassClassificationEvaluator()

evaluator.setPredictionCol("prediction")

evaluator.setLabelCol("indexed")

evaluator.setMetricName("precision")

evaluator.evaluate(predictions)

//Draw a confusion matrix

predictions.groupBy("indexed","prediction").count().show()

Приложение 5

Алгоритм формирования случайного леса

//RandomForests

val datadir = "/home/azureuser/data"

//Create a SQL Context from Spark context

val sqlContext = new org.apache.spark.sql.SQLContext(sc)

import sqlContext.implicits.\_

//Load the CSV file into a RDD

val bankData = sc.textFile(datadir + "/bank.csv")

bankData.cache()

bankData.count()

//Remove the first line because of headers

val firstLine=bankData.first()

val dataLines = bankData.filter(x => x != firstLine)

dataLines.count()

//Convert the RDD into a Dense Vector

//Change labels to numeric ones

import org.apache.spark.mllib.linalg.{Vector, Vectors}

import org.apache.spark.mllib.regression.LabeledPoint

def transformToNumeric( inputStr : String) : Vector = {

val attList=inputStr.split(";")

val age=attList(0).toFloat

//convert outcome to float

val outcome = attList(16).contains("no") match {

case true => 1.0

case false => 0.0

}

//create indicator variables for single/married

val single = attList(2).contains("single") match {

case true => 1.0

case false => 0.0

}

val married = attList(2).contains("married") match {

case true => 1.0

case false => 0.0

}

val divorced = attList(2).contains("divorced") match {

case true => 1.0

case false => 0.0

}

//create indicator variables for education

val primary = attList(3).contains("primary") match {

case true => 1.0

case false => 0.0

}

val secondary = attList(3).contains("secondary") match {

case true => 1.0

case false => 0.0

}

val tertiary = attList(3).contains("tertiary") match {

case true => 1.0

case false => 0.0

}

//convert default to float

val default = attList(4).contains("no") match {

case true => 1.0

case false => 0.0

}

//convert balance amount to float

val balance = attList(5).contains("no") match {

case true => 1.0

case false => 0.0

}

//convert loan to float

val loan = attList(7).contains("no") match {

case true => 1.0

case false => 0.0

}

//Filter out columns not wanted at this stage

val values= Vectors.dense(outcome, age, single, married,

divorced, primary, secondary, tertiary,

default, balance, loan )

return values

}

//Change to a Vector

val bankVectors = dataLines.map(transformToNumeric)

bankVectors.cache()

bankVectors.collect()

//Statistical Analysis

import org.apache.spark.mllib.stat.{MultivariateStatisticalSummary, Statistics}

val bankStats=Statistics.colStats(bankVectors)

bankStats.min

bankStats.max

val colMeans=bankStats.mean

val colVariance=bankStats.variance

val colStdDev=colVariance.toArray.map( x => Math.sqrt(x))

Statistics.corr(bankVectors)

//Transform to a Data Frame

//Drop columns with low correlation

def transformToLabelVectors(inStr : Vector ) : (Float,Vector) = {

val values = ( inStr(0).toFloat,

Vectors.dense(inStr(1),inStr(2),inStr(3),

inStr(4),inStr(5),inStr(6),inStr(7),

inStr(8),inStr(9),inStr(10)))

return values

}

val bankLp = bankVectors.map(transformToLabelVectors)

bankLp.collect()

val bankDF = sqlContext.createDataFrame(bankLp).toDF("label","features")

bankDF.select("label","features").show(10)

//PCA

import org.apache.spark.ml.feature.PCA

val bankPCA = new PCA()

bankPCA.setK(3)

bankPCA.setInputCol("features")

bankPCA.setOutputCol("pcaFeatures")

val pcaModel = bankPCA.fit(bankDF)

val pcaResult = pcaModel.transform(bankDF).select("label","pcaFeatures")

pcaResult.show()

//Indexing for Decision Trees

import org.apache.spark.ml.feature.StringIndexer

val stringIndexer = new StringIndexer()

stringIndexer.setInputCol("label")

stringIndexer.setOutputCol("indexed")

val si\_model = stringIndexer.fit(pcaResult)

val indexedBank = si\_model.transform(pcaResult)

indexedBank.select("label","indexed","pcaFeatures").show()

//Split: training and testing data

val Array(trainingData, testData) = indexedBank.randomSplit(Array(0.9, 0.1))

trainingData.count()

testData.count()

import org.apache.spark.ml.classification.RandomForestClassifier

import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

//Create the model

val rmClassifier = new RandomForestClassifier()

rmClassifier.setLabelCol("indexed")

rmClassifier.setFeaturesCol("pcaFeatures")

val rmModel = rmClassifier.fit(trainingData)

//Predict on the test data

val predictions = rmModel.transform(testData)

predictions.select("prediction","indexed","label","pcaFeatures").show()

val evaluator = new MulticlassClassificationEvaluator()

evaluator.setPredictionCol("prediction")

evaluator.setLabelCol("indexed")

evaluator.setMetricName("precision")

evaluator.evaluate(predictions)

//Draw a confusion matrix

predictions.groupBy("indexed","prediction").count().show()