Применение алгоритма «деревья решений» для анализа персональных данных потенциальных клиентов банка

И. А. Задворная, О. М. Ромакина

Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского
410012, Саратов, ул. Астраханская, 83

e-mail: zadvornayaia@gmail.com

Аннотация. Статья посвящена применению алгоритма Data Mining «Деревья решений» для анализа персональных данных потенциальных клиентов банка. Авторами ставится задача по выявлению персональных характеристик, влияющих на желание человека воспользоваться предложенной услугой и стать клиентом банка. Для решения поставленной задачи создается многомерная структура, к данным которой применяется алгоритм «Деревья решений», анализируются возможность и особенности его применения к изучению персональных данных.
Ключевые слова: интеллектуальный анализ данных, дерево решений, банковские данные, анализ данных, многомерная структура данных, Алгоритм CART, хранилище данных, данные большого объема, Microsoft SQL Server with Analysis Services.

1. Введение

В настоящее время наблюдается резкое увеличение объемов информации, хранящейся в базах данных. На сегодняшний день компании все чаще обращают внимание на возможность оптимизации затрат на маркетинговые мероприятия. В частности, во многих компаниях появляется все больший интерес к обнаружению закономерностей в базах данных, которые содержат информацию о клиентах. Для выявления таких взаимосвязей чаще всего используются системы, реализующие методы Data Mining, предназначенные для обработки и последующей интерпретации данных. Такие методы применяются для выявления текущих тенденций и разработки оптимальных стратегий в будущем. Они позволяют обнаруживать ранее неизвестные и нетривиальные закономерности в данных, собранных в компании. Получение таких знаний призвано обеспечить конкурентное преимущество за счет более глубокого изучения процессов, тенденций и рисков в бизнесе компании.

Обработка информации о взаимодействии клиентов с банком с целью побуждения их к длительному сотрудничеству — важная задача для максимально продуктивной работы финансовой компании. Для решения этой задачи прежде всего

необходимо классифицировать клиентов по различным группам в соответствии с их персональными характеристиками. При этом разделение производится по параметру успешности проведения текущей маркетинговой кампании. Выстраиваемая в рамках данной статьи модель используется для прогнозирования вероятности успешности работы с конкретным потенциальным клиентом в зависимости от известных персональных данных (данные в наборе обезличены). В процессе анализа также выявляются группы риска, для членов которых существует вероятность отказа от предложенной услуги. Такая информация позволяет заранее выработать рекомендации по работе с данными клиентами.

2. Об анализе данных

Разработка стратегий анализа данных, связанных с персональной информацией пользователей или клиентов услуг, особенно важна в банковской сфере. Это связано сразу с несколькими направлениями исследований: обнаружение структуры расходов клиентов, определение структуры транзакционных операций, сегментация клиентов по различным признакам, продажи товаров на основе сегментации клиентов, оценка рисков, предотвращение мошенничества и другие. Банкам приходится иметь дело с огромным количеством различных типов данных: от подробностей и истории транзакций до кредитных балансов и отчетов об оценке рисков. Всестороннее изучение данных в различных областях применения дает наиболее показательный результат для компании. Например, в работе [1] рассматриваются возможности комплексного применения интеллектуального анализа данных к исследованию клиентских банковских данных в сфере использования кредитных карт. В данной же статье рассматривается применение анализа данных в области маркетинга.

Потеря существующих или потенциальных клиентов — это потеря прибыли, которая могла быть получена компанией. Один из способов управления поведением клиентов — побуждение их к действию на основании их персональных данных. В ходе исследования будут найдены основные характеристики, дающие наиболее точный ответ на вопрос: «Какие группы людей для банка являются наиболее перспективными в рамках данной маркетинговой кампании?». Также необходимо узнать, для каких групп необходимо введение дополнительных стимулирующих мероприятий для успешного взаимодействия. Прогнозирование поведения клиентов — важный шаг для поиска ответов на эти вопросы, так как построение долгосрочных отношений с клиентами в ходе взаимовыгодного сотрудничества является основополагающей целью финансовой компании.

3. Описание и подготовка данных

Чтобы эффективно управлять клиентским поведением, внутри компании важно создать эффективную и точную модель взаимодействия с клиентами. Для достижения этой цели существует множество методов интеллектуального анализа данных и моделирования. В рамках решения данной задачи можно использовать как классические статистические модели, так и методы Data Mining (интеллектуальный анализ данных) [2]. Дерево решений — один из основных алгоритмов Data Mining. Он может быть хорошим способом моделирования клиентского поведения [3]. Первоначальный анализ персональных данных показывает возможность предопределенного разделения клиентов на различные группы. В данной работе применяется метод интеллектуального анализа данных для формирования набора групп потенциальных клиентов банка, чтобы смоделировать в дальнейшем особенности взаимодействия с каждой из групп. Полученная информация впоследствии может быть использована для модификации будущих маркетинговых кампаний.

Проведение анализа данных состоит из предварительной подготовки данных, создания подходящей многомерной базы данных, построения модели интеллектуального анализа данных и обучения этой модели [4]. Для моделирования используется персональная информация о потенциальных клиентах банка, которой располагает менеджер во время проведения маркетинговой кампании. Для разделения клиентов на отдельные группы применяется алгоритм деревьев решений, описанный ниже. Затем сформированные группы используются для построения модели клиентского поведения, которая, в свою очередь, является базой для поиска оптимальной маркетинговой кампании для сохранения имеющихся клиентов банка и привлечения новых. Построенная модель позволит определять клиентов, которые более выгодны, чем другие, а не просто анализировать вероятность взаимодействия с ними.

Набор данных, используемый в этом исследовании, находится в UCI Machine Learning Repository (открытый источник данных) [5]. Данные в наборе обезличены и относятся к прямым маркетинговым кампаниям (проводимым с помощью телефонных звонков) португальского банковского учреждения. В ходе анализа рассматривались персональные данные клиентов и результативность взаимодействия в ходе текущей маркетинговой кампании. Основная задача данного исследования — отследить, какие персональные характеристики потенциальных клиентов влияют на их согласие/несогласие открыть депозитный счет в данном банке после предложения данной услуги.

Для оптимального использования алгоритмов Data Mining на основе описанного набора данных была построена многомерная база данных [6]. В составе системы реализованы следующие измерения [7]: Возрастная группа, Образование, Работа,

Семья. Все данные, представленные в системе, отражаются в расчете на разные возрастные группы. В измерении Образование содержится информация по всем уровням образовательной системы, включая начальное, среднее и высшее образование, а также о наличии профильной специализации. Информация о типе занятости и специализации позволяют анализировать влияние профессиональной деятельности клиентов на результат маркетинговой кампании. В измерении Семья содержится информация о семейном положении клиента, в том числе: разведен/разведена, женат/замужем, не женат/не замужем. Таблица фактов [8] непосредственно хранит итог проведения текущей маркетинговой кампании в разрезе указанных измерений.

Самым оптимальным решением для построения многомерной базы данных и дальнейшего анализа в данном случае является Microsoft SQL Server with Analysis Services [9]. Главные преимущества этого программного средства соответствуют требованиям к разрабатываемой системе: оно обеспечивает возможность визуализации данных из куба, имеет удобный интерфейс и возможность анализа данных с использованием технологий Data Mining.

4. Алгоритм «Деревья решений»

Для анализа данных будем использовать алгоритм Data Mining «Деревья решений». Это один из методов, применяемый для решения задач классификации. Она представляет собой процесс обучения модели, который отображает каждый элемент данных в один из классов.

Рассмотрим математическую модель классификации. Набор данных, состоящий из n элементов, представим как набор дискретных точек в n-мерном пространстве. Правило классификации представим в виде гиперкуба (обобщенное определение куба для пространства R_n) в данном пространстве, содержащим одну или несколько из этих точек.

Одним из распространенных алгоритмов, реализующих деревья решений, является CART. Данный алгоритм вырабатывает бинарные деревья и продолжает расщепление до тех пор, пока могут быть найдены новые, которые улучшают решение [10]. Алгоритм CART обеспечивает основу для работы других алгоритмов. Для анализа данных используется регрессионный подход к построению деревьев решений, базирующийся на алгоритме CART.

Каждый корневой узел модели CART представляет собой входную переменную x и точку разделения для этой переменной. Листовые узлы дерева содержат выходную переменную y, которая используется для прогноза значений. Создание

модели CART предполагает выбор входных переменных и разбиений для них до тех пор, пока не будет построено подходящее дерево.

Выбор входной переменной для использования в конкретном разбиении определяется с использованием жадного алгоритма. Построение дерева заканчивается с использованием предопределенного критерия остановки, такого, как минимальное количество экземпляров обучения, назначенных каждому листовому узлу дерева.

Для классификации используется индексная функция Gini, которая дает представление о том, какие узлы стоит выбрать. Обозначим набор данных *T*. В данном наборе содержится *n* классов. Долю обучающих экземпляров с классом *i* в интересующем разбиении обозначим как p_i. Тогда *G(T)* (индексная функция) вычисляется следующим образом (1).

$$G(T) = 1 - \sum_{i=1}^{n} p_i^2. \quad (1)$$

Индекс Gini для каждого узла в дальнейшем взвешивается по общему количеству экземпляров в родительском узле.

Наиболее распространенной процедурой остановки является использование минимального количества экземпляров обучения, назначенных каждому листовому узлу. Если количество элементов в одном из узлов, получившихся в результате разбиения, меньше некоторого заранее заданного минимума, то разбиение дальше не производится, а этот узел принимается за конечный.

С помощью описанного алгоритма в системе проводится интеллектуальный анализ данных. На данном этапе определяются характеристики предыдущих клиентов, которые могут указать, захотят ли эти клиенты приобрести продукт в период текущей маркетинговой кампании.

Алгоритм деревьев решений Microsoft создает модель интеллектуального анализа данных, создавая серию разбиений в дереве. Для дискретных атрибутов алгоритм делает прогнозы на основе отношений между входными столбцами в наборе данных. Он использует значения, называемые состояниями, из этих столбцов, чтобы предсказать значение столбца, который обозначен как прогнозируемый. В частности, алгоритм идентифицирует входные столбцы, которые коррелируют с прогнозируемым [11].

5. Построение модели и анализ данных

Построим дерево решений, основанное на персональных характеристиках клиентов банка. Создание деревьев решений состоит из следующих этапов: создается новая структура анализа данных, выбирается метод/технология Data Mining (Microsoft Decision Trees) и представление источника данных (построенная многомерная база

данных). Затем указывается объем обучающей выборки и все столбцы, участвующие в построении дерева. В качестве ключевого поля будет использован уникальный идентификатор MarketingCampaign. В качестве выходного параметра рассматривается результат проведенной маркетинговой кампании.

Для создания дерева решений будем использовать следующие данные: информация о возрасте клиента, его образовании, наличии кредитов и типе работы. Результат работы алгоритма «Деревья решений» представлен на рис. 1.

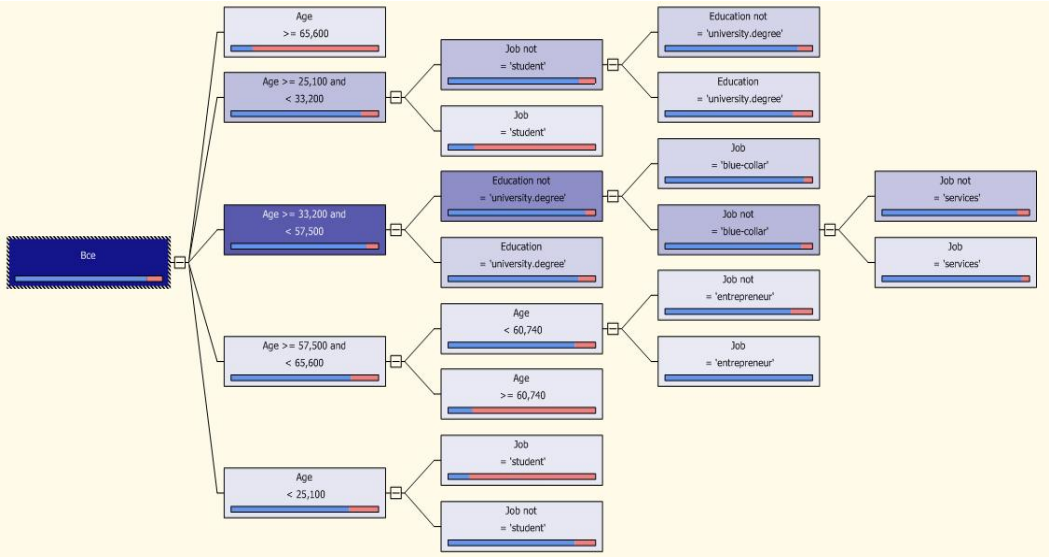

Рисунок 1. Результат работы алгоритма «Дерево решений»

По результатам данной модели наиболее перспективными для банка клиентами оказываются те, что подходят под одну из следующих характеристик: возраст старше 60 лет, студент в возрасте от 25 до 33 и студенты в возрасте до 25 лет. А наименее перспективными клиентами считаются предприниматели в возрасте до 60 лет.

Алгоритм построения деревьев решений позволяет определить набор значений характеристик, позволяющих отделить одну категорию данных от другой (в данном случае — успех или неудача проведения маркетинговой кампании для конкретного клиента). С помощью данного алгоритма можно выяснить, какие характеристики оказывают влияние на предсказываемый параметр и какова степень этого влияния.

Рассмотрим подробнее взаимозависимость параметров (рис. 2). Для дерева решений из 4 (возраст, семейное положение, наличие кредитов и сфера деятельности) параметров, выбранных для анализа, на результат влияют только 3. При этом влияние возраста клиента на предсказываемое значение наиболее существенно.

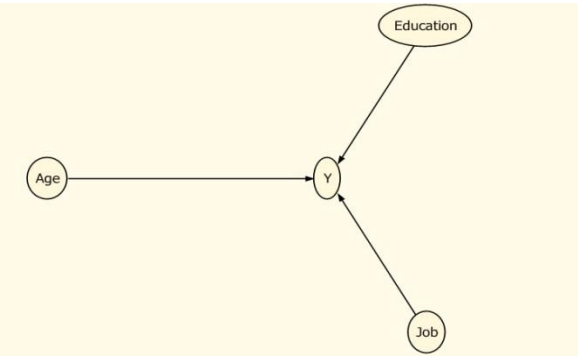

Рисунок 2. Влияние параметров на результат

Для наглядной оценки достоверности построенной модели воспользуемся графиком точности прогнозов [12]. Он имеет следующий вид: ось *X* представляет собой процентную долю из набора данных, выбранных для прогноза, а на оси *Y* указывается процент точных прогнозов для этих данных. На графике изображены две линии: синяя — характеризует стандарт идеальной модели (строится как прямая, исходящая из начала координат под углом 45^{o}), а красная — показывает поведение реальной построенной модели. По отклонению можно дать оценку достоверности построенной модели. В соответствии с графиком (рис. 3) отклонение по полученному дереву решений не превосходит 15%. Модель не является идеальной, но достаточно точно проводит прогноз основного результирующего показателя.

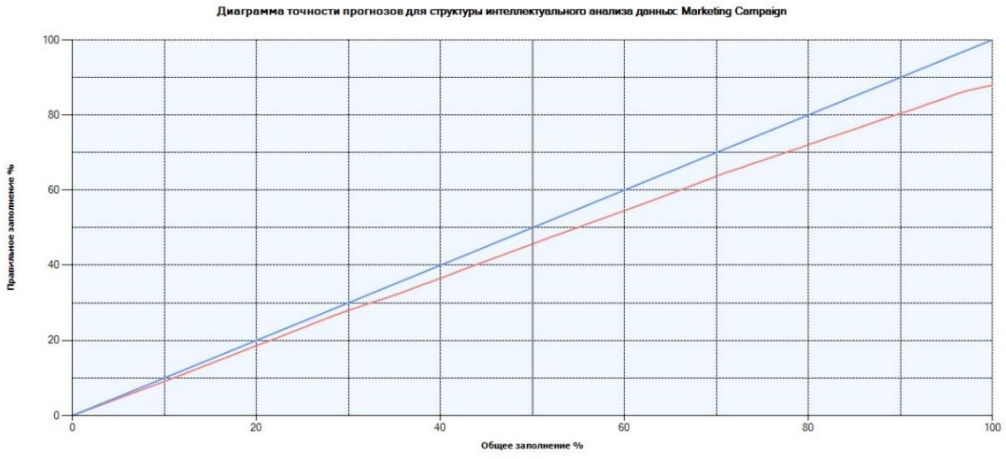

Рисунок 3. Диаграмма точности прогнозов

Необходимо отметить, что в рамках данного исследования предпринимались попытки использования других наборов входных параметров для классификации, а именно, помимо уже описанных характеристик также рассматривались количество произведенных звонков в период текущей маркетинговой кампании, время в днях с момента последнего обращения клиента в банк, результативность предыдущей кампании для конкретного клиента, количество звонков за весь период времени. Построенные деревья решений не привели к удовлетворительным результатам, так как порождали либо очевидные решения, либо слишком сложные конструкции, что, вероятно, связано с проблемой переобучения.

6. Заключение

Представлен алгоритм Data Mining «Деревья решений» и рассмотрены особенности его применения для анализа персональных данных потенциальных клиентов банка. Группировка клиентов была выполнена по сходным характеристикам: возрасту, семейному положению, образованию и работе. Предварительно обработанные данные были подвергнуты классификации с использованием алгоритма «Деревья решений». Результаты группировки потенциальных клиентов банка показывают основные характеристики клиентов, которые позволяют выделить наиболее вероятные группы для плодотворного сотрудничества. Данная информация может помочь банкам определить направление дальнейшего развития маркетинговых кампаний. Приведенная классификация также демонстрирует, на какие группы клиентов стоит обратить внимание в первую очередь, что предопределяет возможность узконаправленной работы с клиентами разных групп для наиболее оптимального функционирования компании при наименьших затратах.

Литература

[1] *Задворная И. А., Ромакина О. М.* Применение алгоритмов Data Mining для анализа данных в сфере кредитования // Математическое и компьютерное моделирование в экономике, страховании и управлении рисками: Материалы VII Международной молодежной научно-практической конф. — Саратов : Научная книга, 2018. С. 61–66.

[2] *Pyle D.* Business Modeling and Data Mining. — Waltham : Morgan Kaufmann Publishers, 2003.

[3] *Zaki M. J., Ho C.-T.* Large-Scale Parallel Data Mining. — Berlin : Springer, 2000.

[4] *Han J., Kamber M.* Data Mining: Concepts and Techniques. — Waltham : Elsevier, 2006.

[5] *Moro M., Cortez P., Rita P.* UCI Machine Learning Repository. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. — Elsevier, 2014 (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing).

[6] *Rafanelli M.* Multidimensional Databases: Problems and Solutions. — Hershey : Idea Group Inc, 2003.

[7] *Inmon W. H.* Building the Data Warehouse. — 3rd Edition. — N. Y. : Wiley Computer Publishing, 2002.

[8] *Adamson C.* Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance. — N. Y. : Wiley Computer Publishing, 2006.

[9] *MacLennan J., Crivat B.* Data Mining with Microsoft SQL Server 2008. — N. Y. : John Wiley and Sons, 2009.

[10] *Wang J.* Encyclopedia of data warehousing and mining. — Hershey : Idea Group Inc, 2005.

[11] Microsoft Decision Trees Algorithm [Электронный ресурс]. Режим доступа: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017

[12] Lift Chart (Analysis Services — Data Mining) [Электронный ресурс]. Режим доступа : https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=sql-server-2017

***Авторы*:**
Ирина Александровна Задворная — магистрант кафедры математического и компьютерного моделирования, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского

Оксана Михайловна Ромакина — кандидат физико-математических наук, доцент кафедры математического и компьютерного моделирования, Саратовский национальный исследовательский государственный университет имени Н. Г. Чернышевского

Application of the algorithm “decision trees” to analysis of personal information on potential bank clients

I. A. Zadvornaya, O. M. Romakina

Saratov State University, 83, Astrakhanskaya st., Saratov, Russia 410012
e-mail: *zadvornayaia@gmail.com*

Abstract. The article is devoted to the application of the Data Mining “Decision Trees” algorithm for analyzing personal data of potential bank customers. The main objective of the research is to obtain personal characteristics that most strongly influence a person’s desire to use the offered service and become a bank customer. In the course of the study, a multidimensional data structure is created for analyzing large data to solve the problem. The decision trees algorithm is applied to the constructed data structure. The article analyzes the possibility and features of the application of this algorithm to the analysis of personal data.
Keywords: data mining, decision tree, bank data, data analysis, multidimensional data structure, CART algorithm, data warehouse, large data, Microsoft SQL Server with Analysis Services.

References

[1] *Zadvornaya I. A., Romakina O. M.* (2018) Primeneniye algoritmov Data Mining dlya analiza dannykh v sfere kreditovaniya. In Matematicheskoye i komp'yuternoye modelirovaniye v ekonomike, strakhovanii i upravlenii riskami: Materialy VII Mezhdunarodnoy molodezhnoy nauchno-prakticheskoy konf. Saratov, Nauchnaya kniga. P. 61–66.
[2] *Pyle D.* (2003) *Business Modeling and Data Mining*. Waltham, Morgan Kaufmann Publishers.
[3] *Zaki M. J. & Ho C.-T.* (2000) *Large-Scale Parallel Data Mining*. Springer.
[4] *Han J. & Kamber M.* (2006) *Data Mining: Concepts and Techniques*. Elsevier.
[5] *Moro M., Cortez P. & Rita P.* (2014) *UCI Machine Learning Repository. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems*. Elsevier.
[6] *Rafanelli M.* (2003) *Multidimensional Databases: Problems and Solutions*. Hershey, Idea Group Inc.
[7] *Inmon W. H.* (2002) *Building the Data Warehouse*, 3rd ed. Wiley Computer Publish.
[8] *Adamson C.* (2006) *Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance*. Wiley Computer Publish.
[9] *MacLennan J. & Crivat B.* (2009) *Data Mining with Microsoft SQL Server 2008*. John Wiley and Sons.
[10] *Wang J.* (2005) *Encyclopedia of data warehousing and mining*. Idea Group Inc.
[11] Microsoft Decision Trees Algorithm (https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-decision-trees-algorithm?view=sql-server-2017).
[12] Lift Chart (Analysis Services — Data Mining) (https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=sql-server-2017).