МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В НАУКЕ, УПРАВЛЕНИИ, СОЦИАЛЬНОЙ СФЕРЕ И МЕДИЦИНЕ

ЧАСТЬ II

Сборник научных трудов
III Международной научной конференции
«Информационные технологии в науке,
управлении, социальной сфере и медицине»

23–26 мая 2016 г.

Томск 2016

2. Интегрированные коммуникации: учебник для студ. учреждений высш. образования / под ред. О. В. Сагиновой. — М.: Издательский центр «Академия», 2014. — 320 с.

3. Бронникова Т.С. Маркетинг: теория, методика, практика: учебное пособие / Т.С. Бронникова. – 4-е изд., стер.М.: Кнорус, 2015. - 208 с.

ПРИМЕНЕНИЕ МЕТОДОВ DATA MINING ДЛЯ ПРОГНОЗИРОВАНИЯ СКОРОСТИ ПРОДАЖИ ПОДЕРЖАННЫХ АВТОМОБИЛЕЙ

О.Г. Вершинина
(г. Томск, Национальный исследовательский Томский государственный университет)
vershinina.olya@gmail.com

APPLICATION OF DATA MINING METHODS TO PREDICT USED CARS DEMAND

O.G. Vershinina
(Tomsk, National Research Tomsk State University)

Abstract. One of the popular web-service for selling of used cars provided a big database to create a new service for dealers. This service must predict period of time when a car can be sold. This task is divided into two parts: find an algorithm and create the software. In this paper we describe process of data preparing and finding the predictive model using Rapid Miner and R language.

Keywords. Data mining, Rapid Miner, R Studio, classification, decision trees.

Введение. Зарубежный онлайн-сервис более 10 лет предоставляет частным продавцам и крупным дилерам возможность размещать объявления о продаже подержанных автомобилей. Мы попытались использовать накопленные многотысячной аудиторией сайта данные для создания сервиса для дилеров, суть которого заключается в расчете скорости продажи автомобиля на основании его характеристик, предоставленных продавцом в объявлении. Данные об объявлениях, их история и статистика хранятся за последние три года. Такая база данных является достаточно репрезентативной для попытки поиска и построения некоторой предсказательной модели, на основе которой и будет работать создаваемый веб-сервис.

Постановка задачи. Для каждого объявления нужно определить время его активности на сайте. Фактически, время жизни объявления соответствует времени с момента его опубликования до момента продажи автомобиля.

Задача «минимум» состоит в том, чтобы определить некоторый промежуток времени, в который объявление будет активно. Временные периоды длиной от нескольких дней до нескольких месяцев в таком случае рассматриваются как категории, по которым можно классифицировать каждое объявление. Эта задача будет решена, если удастся получить настолько оптимальное деление прогнозируемого времени на периоды, чтобы предсказание времени активности для любого произвольного объявления максимально точно попадало в один из периодов.

Процесс решения задачи разбивается на несколько этапов [1].

Этап 1. Отбор и предварительная обработка данных. Поскольку данные никогда ранее не обрабатывались для подобного рода задач, то процесс обработки является важнейшим. Данные являются сильно зашумленными, с большим количеством пропусков и ошибок. Также для начальных экспериментов признаковый вектор сильно

редуцирован по длине, много признаков сознательно не включены для упрощения задачи.

Этап 2. Обучение и тестирование нескольких методов классификации с постепенным увеличением количества классов (то есть дроблением периодов времени до минимально возможных).

Этап 3. Оценка и сравнение точности полученных моделей.

Этап 4. Интерпретация результатов.

Исходные данные и предварительная обработка. Объявления на сайт поступают из разных источников: вводятся вручную, загружаются специальными скриптами с других сайтов или выгружаются из файлов, предоставляемых поставщиками. Данные для начальных экспериментов взяты от одного из таких поставщиков в формате *csv* файлов. Данные охватывают период с января 2013 года по февраль 2016 и состоят из объявлений от разных дилеров, что обеспечивает разнообразие признаков.

Для того, чтобы обеспечить репрезентативность данных, учитываются следующие факторы, диктуемые особенностями рынка и условиями задачи.

1. Сезонность: период времени, за который отбираются объявления, должен быть в рамках одного сезона, летнего или зимнего.
2. Конечный период активности в рамках исследуемого диапазона. Если рассматриваются объявления за период 2 месяца, то необходимо из них отобрать только те, время активности которых полностью входит в этот диапазон.

Помимо фильтрации данных по критериям, описанным выше, предварительная обработка данных [2] также включает в себя:

1) удаление из признакового вектора некатегориальных и нечисловых признаков, таких как описание, комментарии, дополнительные маркировка (вариант) и т.п.;
2) удаление признаков, значения которых у большинства объектов имеют либо одинаковые, либо очень близкие значения, например, количество фотографий (у всех объявлений в выборке от 7 до 9 фотографий), включение налога в стоимость и т.п.;
3) категориальные признаки, такие как марка, модель, цвет автомобиля, обрабатываются с помощью словарей – таблиц соответствия различных написаний признака какому-то одному синониму, - которые позволяют свести разнообразие значений к конечному списку категорий;
4) по оставшимся признакам отфильтрованы объекты с пустыми значениями признаков или со значениями, представленными в выборке только один раз (например, какая-то определенная модель автомобиля или цвет).

Обучающая выборка после фильтрации состоит из 178 объектов, на которой предпринимается попытка обучить некоторые методы классификации [3].

Методы и инструменты. Для поиска классифицирующей модели на первом этапе используется два метода и два разных инструмента. Классификация методом *CART* (*Classification And Regression Tree*) [4] производится с использованием широко известной среды *RSudio* средствами языка *R* [5]. Метод рекурсивного построения деревьев решений (*Decision Tree*) применяется в данной работе в рамках другого удобного инструмента *Rapid Miner* [6]. В этой программе любой метод возможно настроить, выбрав различные критерии разбиения, а также визуально настроить процесс обработки данных и вычислений, поэтому он был выбран в качестве одного из инструментов.

Классификация. Признак, который определяется как классифицирующий это время активности объявления на сайте, измеряемое в днях. Поиск оптимальной модели производится итерациями. Каждая итерация включает в себя:

1) определение количества классов и их границ;
2) обучение метода;
3) анализ полученной модели;
4) применение модели на тестовой выборке;
5) вычисление ошибки;
6) сравнение ошибки с предыдущей итерацией.

Как только качество классификации на следующей итерации становится хуже, чем в предыдущей, это означает, что достигнуто максимальное деление на периоды, которое возможно с данной выборкой.

Выбор количества классов и их границы определяются визуально. Так, для первой итерации выборка разделена на два класса, условно названных *slow* и *fast*, которые обозначают период активности объявления больше 20 дней и меньше, соответственно.

Полученные деревья решений в обоих случаях дают точность классификации чуть больше 72%. Такая точность считается удовлетворительным для перехода к следующему этапу вычислений. Далее необходимо увеличить количество классов и обучить те же самые методы, сравнив затем их точность. Таким образом, должна быть выбрана наиболее оптимальное разбиение на классы и более эффективная модель, которую затем можно пытаться обучать на данных с более объемным признаковым вектором и более объемной обучающей выборке.

Заключение. Получены первые результаты вычислений, которые дают удовлетворительную точность классификации на тестовых данных, на основании чего можно утверждать о возможности успешного решения всей задачи в целом. Далее предполагается усложнение признакового вектора добавлением ранее удаленных полнотекстовых признаков и применение методов *NLP* к ним. Эти признаки также должны будут участвовать в классификации, и итеративная процедура поиска модели будет проводится аналогично описанному в данной работе сценарию. Кроме этого, предполагается сравнить эффективность и точность используемых методов классификации применительно к анализируемым данным.

ЛИТЕРАТУРА

1. Замятин А.В. Введение в интеллектуальный анализ данных: учеб. пособие. – Томск: Издательский Дом государственного университета, 2016. – 120 с.

2. Чубукова И. Data Mining. Лекция 18. Начальные этапы. Предварительная обработка данных [Электронный ресурс] // URL: http://www.intuit.ru/studies/courses/6/6/lecture/192?page=3 (дата обращения: 28.02.2016).

3. L. Rokach, O. Maimon. Data Mining with Decision Trees: Theory and Applications. – Singapore: World Scientific Publishing Co. Pte. Ltd., 2014. – 328 с.

4. L. Rutkowskia, M. Jaworskia, L. Pietruczuka, P. Dudaa. The CART decision tree for mining data streams // Information Sciences. – 2014. – Vol. 266. – P. 1-15

5. G. Grolemund. Hands-On Programming with R: Write Your Own Functions and Simulations. – USA: O'Reilly Media, Inc., 2014. – 250 с.

6. M. Hofmann, R. Klinkenberg. RapidMiner: Data Mining Use Cases and Business Analytics Applications. – USA: Taylor & Francis Group, LLC. – 2014. – 463 с.