УДК 004.62

ИСПОЛЬЗОВАНИЕ ДЕРЕВЬЕВ КЛАССИФИКАЦИИ В СТРАТЕГИЯХ ПРОДВИЖЕНИЯ ТОВАРОВ НА ПРИМЕРЕ ДАННЫХ О ПРОДАЖЕ ВЕЛОСИПЕДОВ

АНАНЬЕВА ДАРЬЯ ДМИТРИЕВНА

Кафедра цифровой экономики, Пензенский государственный университет, 440026, г. Пенза, ул. Красная, 40, e-mail: daf.raf.01@mail.ru

РЫНДИНА СВЕТЛАНА ВАЛЕНТИНОВНА

Кафедра цифровой экономики, Пензенский государственный университет, 440026, г. Пенза, ул. Красная, 40, e-mail: svetlanar2004@yandex.ru

Аннотация

Использование данных для разработки управленческих решений или data-driven (управляемый данными) подход – это базис для современного менеджмента. Он пронизывает все аспекты деятельности компании, но особенно заметны его преимущества в маркетинге. Аналитика данных в компании обычно подчиняется определенной иерархии: начальный уровень связан с описательной аналитикой, которая фиксирует положение компании в данный момент (отчеты по текущим показателям бизнеса), более продвинутый уровень – это попытка найти в данных объяснение текущему положению вещей и на верхнем уровне иерархии находятся прогностическое моделирование, машинное обучение – как инструмент изменения ситуации в выгодном для компании направлении. Так деревья классификации можно использовать для более качественной сегментации клиентской базы, что позволяет выявить профили клиентов, заинтересованных в продукте. Целью проведения исследования является описание перехода от отдельных инсайтов, выявленных в ходе интерпретации деревьев классификации, построенных на основе данных о покупках товара, к разработке стратегии продвижения. Цели исследования достигаются за счет обогащения исходных данных информацией о рыночной ситуации, о бизнес-окружении, о знаниях предметной области, которые используются в разработке стратегии продвижения и в выдвижении гипотез для дальнейшего уточнения пространства бизнес-решений в маркетинге. В работе использованы следующие методы исследования: классификация и системный анализ. В результате проведенного исследования представлен алгоритм перехода от результатов анализа данных о покупках на основе использования деревьев классификации к разработке стратегии продвижения товаров.

Ключевые слова: деревья классификации, бизнес-аналитика, Python, управление на основе данных, продвижение

Введение

Персонализация является мощным и популярным инструментом современного маркетинга, потому что показывает хорошие результаты. Но для того, чтобы правильно выстроить персональную тактику ведения взаимоотношения с клиентом - необходимо собрать и обработать его данные. Чем данных будет получено, тем точнее можно построить модель поведения с конкретным клиентом. Однако данные без обработки малоинформативны. И использование алгоритмов машинного обучения, таких как, например, построение деревьев классификации, позволяет сегментировать клиентов, и предлагать им то, что они хотят на самом деле. Модели машинного обучения используются во многих современных сервисах, таких как рекомендательная система Netflix или Youtube. Во многом именно способность персонализироваться делает данные сервисы настолько привлекательными для клиентов. Так что данная тема актуальна как в цифровом, так и в офлайн мире.

Основная часть

Деревья решений (DT) - это непараметрический контролируемый метод обучения, используемый для классификации. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных [1].

Некоторые преимущества деревьев решений:

- Достаточно просто понять и интерпретировать. Для визуализации деревьев решений существуют готовые методы.
- Требуется небольшая подготовка данных. Однако данный модуль не поддерживает отсутствующие значения, пропуски в данных нужно или удалить при их небольшом количестве и готовности потерять небольшое количество данных, или заполнить их, используя среднее, медиану или моду. Возможны и другие варианты заполнения пропущенных данных.
- Стоимость использования дерева является логарифмической по количеству точек данных, используемых для обучения дерева.
- Может обрабатывать как числовые, так и категориальные данные. Однако реализация scikit-learn не поддерживает категориальные переменные. Так что в данной статье исходные категориальные строковые значения будут приведены к числовым.

К недостаткам деревьев решений можно отнести:

- Обучающиеся дереву решений могут создавать слишком сложные ветвистые деревья, которые плохо обобщают данные. Это называется переобучением. Чтобы избежать этой проблемы, необходимы такие механизмы, как обрезка, установка минимального количества выборок, необходимых для конечного узла, или установка максимальной глубины дерева. В данной статье мы ограничим глубину дерева решений до 5 узлов.
- Деревья решений могут быть нестабильными, поскольку небольшие изменения в данных могут привести к созданию совершенно другого дерева. Эта проблема смягчается за счет использования деревьев решений в ансамбле.
- Известно, что проблема обучения оптимальному дереву решений является NP-полной. Следовательно, практические алгоритмы обучения дереву решений основаны на использовании жадных алгоритмов, которые принимают локально оптимальные решения в каждом узле. Такие алгоритмы не могут гарантировать возврат глобального оптимального дерева решений.

Для иллюстрации использования деревьев решений в стратегиях продвижения товаров с сайта Kaggle был импортирован набор bike_buyers.csv, в котором собраны данные о 1000 покупателей велосипедов [2].

В наборе присутствуют следующие данные о клиентах:

- ID – идентификатор клиента,
- Marital Status – семейный статус,
- Gender – пол,
- Income – доход,
- Children – количество детей,
- Education – образование,
- Occupation – род деятельности,
- Home Owner – информация о том, является ли клиент владельцем дома,
- Cars – количество машин,
- Commute_Distance – расстояние до места работы,
- Region – регион проживания,
- Age – возраст,
- Purchased_Bike – информация о том, купил ли клиент велосипед.

Для построения деревьев в работе используется Anaconda (дистрибутив языка программирования Python и репозиторий библиотек, предназначенных для анализа данных и машинного обучения). Для того, чтобы извлечь полезную информацию из данных, необходимо записать их в удобную для обработки структуру данных и провести предварительную обработку. Для данной цели и построения дерева классификации нам понадобятся несколько библиотек: pandas для удобного представления и анализа данных [3], класс DecisionTreeClassifier из модуля sklearn.tree для построения дерева классификации, а так же matplotlib.pyplot и метод plot_tree для визуализации построенного экземпляра дерева для используемого набора данных.

Выполним импорт библиотек в Jupyter Notebook (Anaconda):

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier
from sklearn.tree import plot_tree
```

Проведем считывание данных из файла в переменную (тип DataFrame) библиотеки pandas.

```
bike_buyers_df = pd.read_csv("bike_buyers.csv")
```

Результат обращения к переменной, содержащей набор данных из айла представлен на рис. 1.

	ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute_Distance	Region	Age	Purchased_Bike
0	12496	Married	Female	40000.0	1.0	Bachelors	Skilled Manual	Yes	0.0	0-1 Miles	Europe	42.0	No
1	24107	Married	Male	30000.0	3.0	Partial College	Clerical	Yes	1.0	0-1 Miles	Europe	43.0	No
2	14177	Married	Male	80000.0	5.0	Partial College	Professional	No	2.0	2-5 Miles	Europe	60.0	No
3	24381	Single	NaN	70000.0	0.0	Bachelors	Professional	Yes	1.0	5-10 Miles	Pacific	41.0	Yes
4	25597	Single	Male	30000.0	0.0	Bachelors	Clerical	No	0.0	0-1 Miles	Europe	36.0	Yes

Рис. 1. Набор данных "bike_buyers.csv"

Произведем предварительную обработку данных. Проверим, есть ли пропущенные значения в данных. С помощью метода класса DataFrame isnull() определим массив той же размерности, что и исследуемый набор данных, со значениями True/False фиксирующими как True позиции, на которых находятся пропуски, посчитаем количество пропусков в каждом столбце методом sum():

```
bike_buyers_df.isnull().sum()
```

Результат выполнения операций над исходными данными представлен на рис. 2.

```
ID                  0
Marital_Status      7
Gender             11
Income              6
Children            8
Education           0
Occupation          0
Home_Owner          4
Cars                9
Commute_Distance    0
Region              0
Age                 8
Purchased_Bike      0
dtype: int64
```

Рис. 2. Нахождение пропущенных значений в наборе данных

Видим, что пропущенные значения есть, и их не очень много, так что исключим из набора данных строки, содержащие пропущенные значения, с помощью метода dropna():

```
bike_buyers_df_full = bike_buyers_df.dropna()
```

В переменную X сохраним исходный набор данных, исключив неинформативный в данном случае идентификатор клиента и целевой показатель Purchased_Bike, который регистрирует для наблюдения факт наличия покупки велосипеда, столбец с целевым показателем сохраним в переменную y.

```
X = bike_buyers_df_full.drop(["ID", "Purchased_Bike"], axis = 1)
y = bike_buyers_df_full.Purchased_Bike
```

В исследуемом наборе данных присутствуют категориальные строковые значения, однако метод fit() может работать только с числовыми наборами данных. Это технически верно – не слишком понятно, по какому критерию сравнивать строки так, чтобы эту информацию можно было учесть при построении дерева и было логически оправдано во всех случаях применения данного алгоритма к произвольному набору данных. Для преобразования категориальных данных применим функцию get_dummies(), заменяющую исходную категориальную переменную на бинарные переменные (принимающие только два значения 0 и 1). Количество таких заместителей категориальной переменной на единицу меньше числа значений, которые переменная принимает в исходном формате: каждая бинарная переменная отвечает одному из значений и содержит единицы в строках с этим значением категориальной переменной. Одно из значений категориальной переменной не используется для образования бинарной переменной, ему соответствует одновременное равенству нулю всех бинарных переменных заместителей (рисунок 3).

```
X = pd.get_dummies(X, drop_first = True)
```

	Income	Children	Cars	Age	Marital_Status_Married	Marital_Status_Single	Gender_Female	Gender_Male	Education_Bachelors	Education_Graduate Degree	...
0	40000.0	1.0	0.0	42.0	1	0	1	0	1	0	...
1	30000.0	3.0	1.0	43.0	1	0	0	1	0	0	...
2	80000.0	5.0	2.0	60.0	1	0	0	1	0	0	...
4	30000.0	0.0	0.0	36.0	0	1	0	1	1	0	...
5	10000.0	2.0	0.0	50.0	1	0	1	0	0	0	...
7	40000.0	1.0	0.0	43.0	1	0	0	1	1	0	...
10	30000.0	3.0	2.0	54.0	1	0	1	0	0	0	...

Рис. 3. Набор данных, сохраненный в переменную X после преобразования категориальных переменных

Далее создаем экземпляр класса классификатора и выполняем построение дерева решений:

```
clf = DecisionTreeClassifier(criterion = "entropy", max_depth = 5)
clf.fit(X, y)
```

Визуализируем построенное дерево классификации (рисунок 4) с помощью метода plot_tree():

```
plot_tree(clf, fontsize=10, feature_names=list(X), class_names=["Purhased", "Didn't purchased"], filled=True)
plt.show()
```

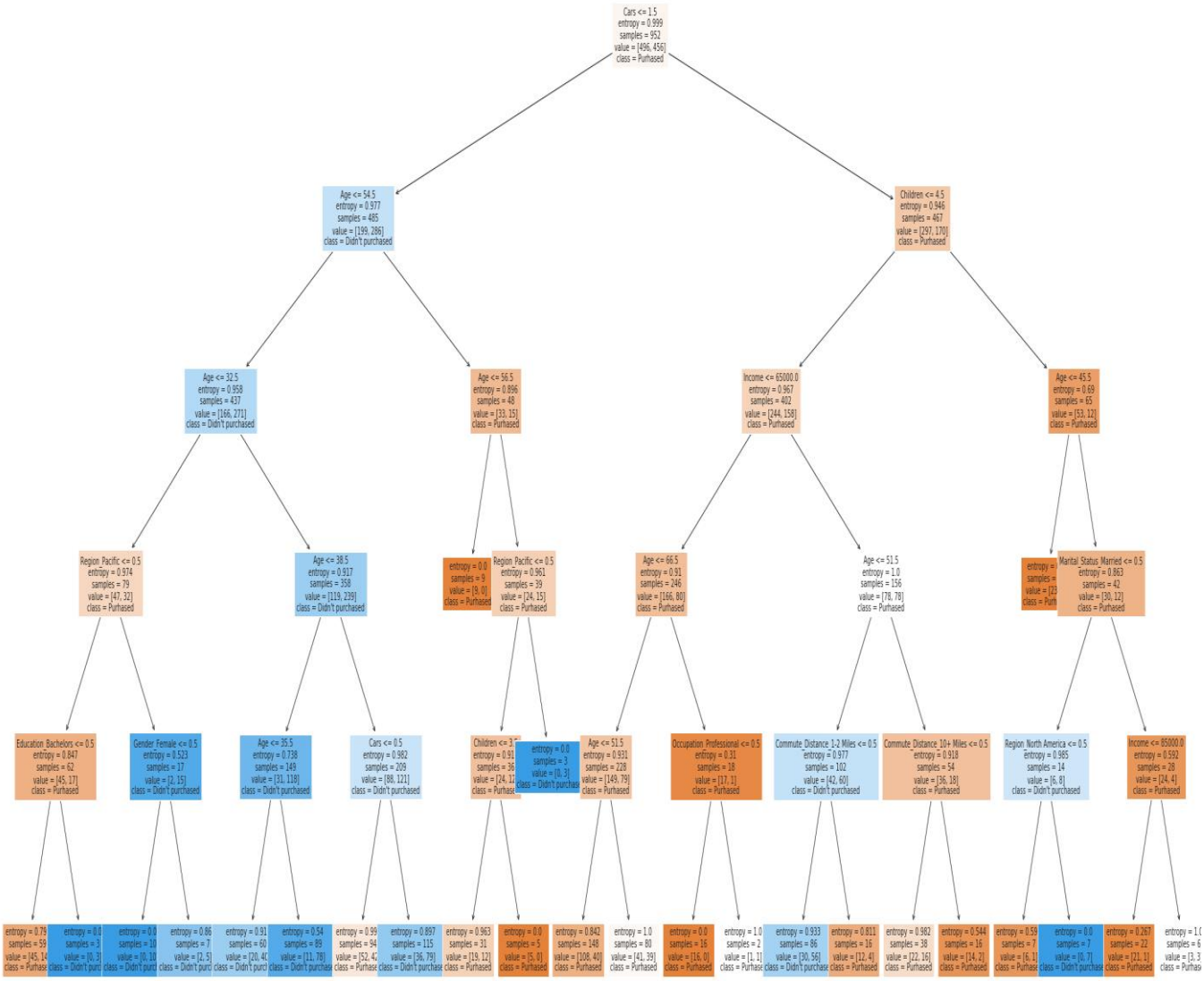

Рис. 4. Визуализация дерева решений

Проведем анализ полученных результатов:

- покупателями велосипедов чаще являются те, у кого, больше 2 автомобилей (297 людей к 199) и обратная ситуация в случае, если автомобилей меньше 2 (170 человек к 286);
- чаще всего те, у кого меньше 2 автомобилей и возраст в диапазоне от 35.5 до 38.5 лет велосипеды не покупают (11 человек к 78);
- среди людей в возрасте старше 54.5 лет, имеющих 1 автомобиль или не имеющих машин вовсе - количество покупок более, чем в 2 раза больше, чем отказов от покупки (33 человека к 15);
- среди людей, имеющих больше 2 автомобилей и более 4 детей - вероятность покупки больше, чем у тех, у кого детей меньше 4 (53/65 = 0.8% к 244/402 = 0.6%);
- при этом у людей, у которых больше 2 автомобилей, больше 4 детей и возрастом меньше 45.5 лет вероятность покупки равна 1.

Так же можно выделить несколько характерных групп, которые с очень большой вероятностью

- покупают:
  - клиенты, у которых количество машин больше 2, имеющих детей больше 4 и возрастом меньше 45.5 лет (особенно те, чей возраст не превышает 45.5 лет),

- клиенты с 1 автомобилем или неимеющие автомобиля вовсе в возрасте от 54.5 и до 56.5.

– не покупают:

- клиенты, у которых менее 2 машин, в возрасте от 38.5 лет и до 35.5 лет (особенно те, чей возраст находится в диапазоне от 32.5 лет и до 35.5 лет).

На основании данных результатов можно сделать несколько выводов о направлении продвижения. Данные показывают, что в основном велосипедный транспорт не является альтернативой автомобильному, что вполне могло бы быть, так как многие клиенты проживают в европейских странах, где созданы хорошие условия для использования велосипедного транспорта. Однако велосипеды чаще всего покупают те, у кого один или два автомобиля (рисунок 5). Это может говорить о том, что клиенты используют велосипедный транспорт для занятий спортом или для прогулок.

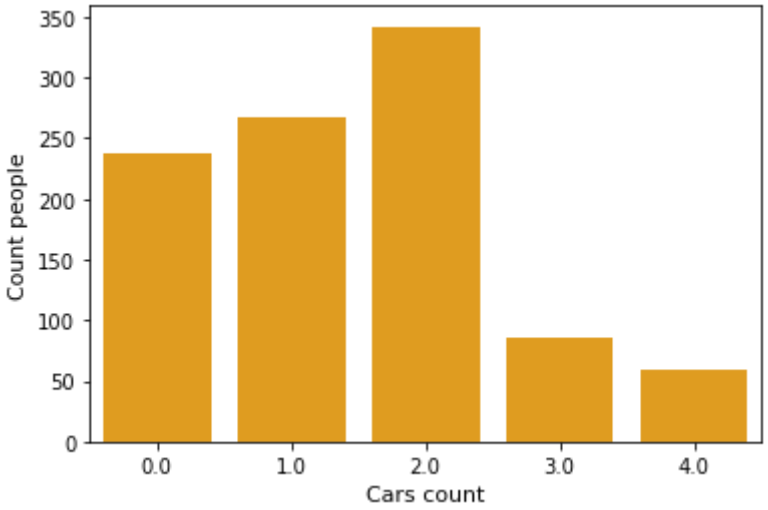

Рис. 5. Зависимость количества клиента от числа автомобилей

Перспективные группы потенциальных клиентов для проведения маркетинговых компаний – это люди старшей возрастной группы и многодетные. На эти сегменты необходимо обратить особое внимание.

Заключение

С помощью деревьев решения можно классифицировать людей на группы и выяснить, какие значения показателей соответствуют нашим целевым клиентам. Это помогает составить предсказательный прогноз на новых входных данных. А также выстраивать продвижение таким образом, чтобы основные усилия были сосредоточены на целевых сегментах, так как для них конвертация коммуникаций в покупку наибольшая.

Анализируя набор данных о 1000 покупателей велосипедов с помощью деревьев решений было выяснено, что на покупку велосипедов по большей части влияет количество автомобилей, которые есть у клиента, а так же его возраст и количество детей. Такие показатели, как образование и пол так же присутствуют в модели, но они скорее отсекают единичные экземпляры, которые попались в сформированную группу, и не являются показательными.

Так же с помощью дерева классификации стало возможно выявить несколько характерных групп покупателей, которые склонны покупать или воздержаться от покупки. С помощью этой информации можно формировать персонализированные предложения и концентрировать свои усилия на целевых группах клиентов.

Список литературы

[1] Документация scikit-learn. – URL: https://scikit-learn.ru/1-10-decision-trees/

[2] Kaggle. Bike Buyers 1000. – URL: https://www.kaggle.com/datasets/heeraldedhia/bike-buyers

[3] Документация pandas. – URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html?highlight=dataframe#pandas.DataFrame

USING CLASSIFICATION TREES IN PRODUCT PROMOTION STRATEGIES USING BICYCLE SALES DATA AS AN EXAMPLE

ANANYEVA DARYA DMITRIEVNA

Department of digital economy, Penza state University, 40 Krasnaya str., Penza, 440026, e-mail: daf.raf.01@mail.ru

RYNDINA SVETLANA VALENTINOVNA

Department of digital economy, Penza state University, 40 Krasnaya str., Penza, 440026, e-mail: svetlanar2004@yandex.ru

Annotation

The use of data for the development of management decisions or data-driven approach is the basis for modern management. It permeates all aspects of the company's activities, but its advantages in marketing are especially noticeable. Data analytics in a company usually follows a certain hierarchy: the initial level is associated with descriptive analytics that captures the company's current position (reports on current business indicators), the more advanced level is an attempt to find an explanation of the current state of things in the data and at the top level of the hierarchy are predictive modeling, machine learning – as a tool for changing the situation in a favorable direction for the company. So classification trees can be used for better segmentation of the customer base, which allows you to identify profiles of customers interested in the product. The purpose of the study is to describe the transition from individual insights identified during the interpretation of classification trees based on data on purchases of goods to the development of a promotion strategy. The objectives of the study are achieved by enriching the initial data with information about the market situation, the business environment, and domain knowledge, which are used in developing a promotion strategy and in putting forward hypotheses to further clarify the space of business solutions in marketing. The following research methods are used in the work: classification and system analysis. As a result of the conducted research, an algorithm for the transition from the results of the analysis of purchase data based on the use of classification trees to the development of a product promotion strategy is presented.

Keywords: classification trees, business analytics, Python, data-based management, promotion