УДК 004.852

МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ ПРОГНОЗИРОВАНИЯ ПРОПУСКОВ ЗАНЯТИЙ СТУДЕНТАМИ

М. А. Третьяк, В. В. Фуников, О. А. Сафарьян, А. Е. Щекатурин, П. П. Орехов
Донской государственный технический университет (г. Ростов-на-Дону, Российская Федерация)

Рассматриваются вопросы программной реализации модели машинного обучения, способной составить прогноз посещаемости занятий студентом, основанный на заранее собранных реальных данных, сформированных путем опроса и анализа статистики. В данной работе был использован алгоритм обучения дерево принятия решений, основанный на использовании древовидного графа. Программная реализация алгоритма основана на расчете вероятности наступления того или иного события, преимущество заключается в структурировании и систематизации проблемы, а итоговое решение принимается на основе логических выводов. Основной задачей является поиск ответа на вопрос, возможно ли с использованием методов машинного обучения решить проблему прогнозирования посещаемости студентов, и если это возможно, то определить точность этого прогноза.

Ключевые слова: машинное обучение, вероятности, граф, прогнозирование, признаки, статистика, датафрейм, дерево принятия решений, классификация.

MACHINE LEARNING MODEL FOR PREDICTING STUDENTS' ABSENCES

M. A. Tretyak, V. V. Funikov, O. A. Safaryan, A. E. Schekaturin, P. P. Orekhov
Don State Technical University, (Rostov-on-Don, Russian Federation)

The paper considers the issues of software implementation of a machine learning model capable of predicting student attendance based on pre-collected real data generated by polling and statistical analysis. In this paper, a decision tree learning algorithm based on the use of a tree graph was used. The software implementation of the algorithm is based on calculating the probability of occurrence of an event, the advantage lies in structuring and systematizing the problem, and the final decision is made based on logical conclusions. The main task is to find an answer to the question whether it is possible to solve the problem of predicting student attendance using machine learning methods, and if possible, then determine the accuracy of this forecast.

***Keywords*:** machine learning, probabilities, graph, forecasting, signs, statistics, dataframe, decision tree, classification.

Введение. Технологии современного мира беспрерывно развиваются, появляются все более усовершенствованные средства и системы решения различного рода задач, но также возникают и новые проблемы, решение которых выходит на первый план. Примером такой проблемы является случай, когда человеку необходимо обрабатывать огромные потоки информации, анализировать их и совершать определенные умозаключения, но в связи с ограниченностью человеческих возможностей совершать подобного рода операции крайне неэффективно по времени и продуктивности. Одним из распространенных на сегодняшний день методов решения и оптимизации такой проблемы является машинное обучение.

Машинное обучение как класс методов является ответвлением от методов искусственного интеллекта, сама суть машинного обучения заключается не в прямом решении поставленной задачи, а в обучении, где применяется множество решений схожих задач средствами

математического анализа, численных методов, теории вероятности, математической статистики и других различных техник работы с данными в цифровом виде [1].

Данная проблема является актуальной, поскольку повсеместно появляются задачи, требующие обработки большого объема данных, но при этом требующие и большой отлаженности и структурированности, которые человек зачастую не в состоянии самостоятельно выполнить [1].

Представленная исследовательская работа заключается в разработке модели машинного обучения, способной составить прогноз посещаемости занятий студентом, основанный на заранее собранных реальных данных, сформированных путем опроса и анализа статистики.

Цель работы — реализовать модель машинного обучения с возможностью прогнозирования посещения занятий студентами.

Программная реализация модели машинного обучения. Анкетирование. Для корректной реализации модели машинного обучения прогнозирования пропусков занятий студентами был определен ряд полей, сформированный на основе анкетирования в течение двух недель студентов через сервис Google Forms, а также статистики посещения предприятия за один год, предоставленной компанией DBI. На рис. 1–3 представлены некоторые поля из анкеты.

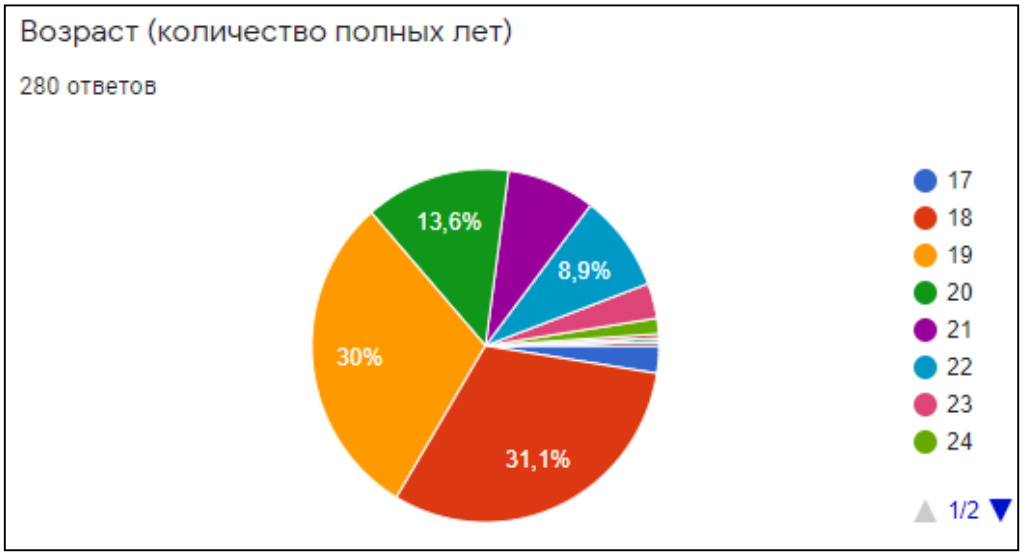

Рис. 1. Результат анкетирования по полю «Возраст»

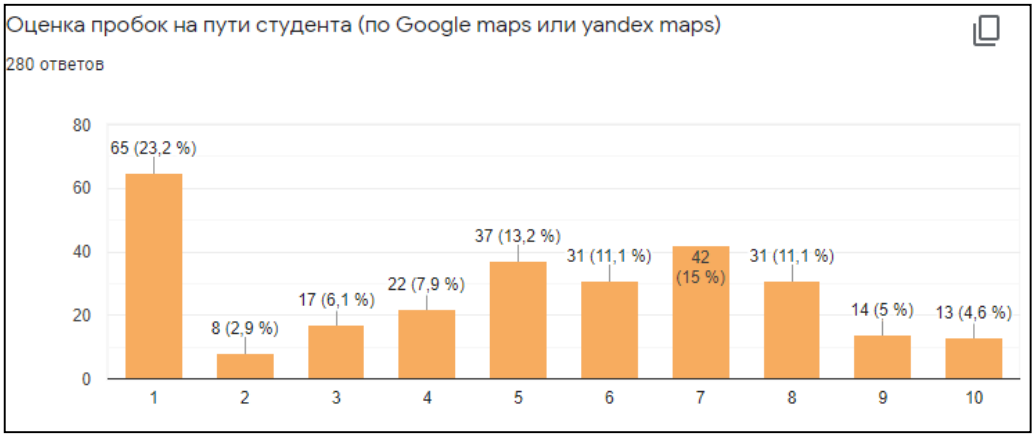

Рис. 2. Результат анкетирования по полю «Оценка пробок на пути студента»

Рис. 3. Рейтинг транспорта, используемого студентами

Описание программной реализации. В качестве языка программирования был выбран Python 3.8. При формировании датафреймов (выбранное содержимое таблицы) с использованием библиотеки pandas были отсеяны лишние признаки и отобраны те, которые представляют наибольшие зависимости [2]. На рис. 4 представлена датафрейм отобранных признаков в программе Jupyter.

```
[8]: table_personal.index = top_layer['ID']
[9]: table_personal
```

[9]:

ID	Причина пропуска	День недели	Траты на транспорт	Расстояние от дома до места обучения	Возраст	Количество детей	Выпивает	Курит	Домашние животные
15	28	5	4444	31	40	1	1	0	1
33	1	3	3787	25	47	2	0	0	1
32	28	1	4413	48	49	0	0	0	2
17	25	5	2733	22	40	2	0	1	0
20	23	5	3970	50	36	4	1	0	0
...	...	...	...	...	...	...	...	...	...
28	23	3	3436	26	28	1	0	0	2
3	23	1	2733	51	38	0	1	0	0
3	27	4	2733	51	38	0	1	0	0
18	0	3	5039	16	28	0	0	0	0
3	27	1	2733	51	38	0	1	0	0

700 rows × 9 columns

Рис. 4. Отобранные признаки

Полученные данные в дальнейшем будут использоваться для обучения модели. Данную задачу будем рассматривать как задачу классификации студентов по количеству пропусков, в качестве нормы было принято использовать среднее арифметическое пропусков студентов за данный период.

В работе был использован алгоритм обучения дерево принятия решений, основанный на использовании древовидного графа, данный алгоритм основывается на расчете вероятности наступления того или иного события. Его преимущество заключается в том, что он структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов [3].

Основной вопрос, на который предстоит дать ответ: «Возможно ли с использованием методов машинного обучения решить проблему прогнозирования посещаемости студентов»? Если это возможно, то определить точность этого прогноза. На рис. 5 представлен результат процесса обучения модели.

```
       Причина пропуска  День недели  ...       Курит  Домашние животные
count        700.000000   700.000000  ...  700.000000         700.000000
mean          19.195714     2.887143  ...    0.072857           0.754286
std            8.471461     1.420313  ...    0.260088           1.310100
min            0.000000     1.000000  ...    0.000000           0.000000
25%           13.000000     2.000000  ...    0.000000           0.000000
50%           23.000000     3.000000  ...    0.000000           0.000000
75%           26.000000     4.000000  ...    0.000000           1.000000
max           28.000000     5.000000  ...    1.000000           8.000000

[8 rows x 9 columns]
Enter choice for Data: 2
To Learn model, we choice 2
700
Accuracy on training set: 0.939
Accuracy on test set: 0.810
Complete save model!
Open model!
```

Рис. 5. Результат обучения

Обученная модель прогнозирует посещение студента при определенных условиях (признаках) с точностью 0.93 на данных, использованных при ее обучении, и с точностью 0.81 на данных, которые не использовались при обучении, а были задействованы при тестировании и составляют 30 % от общего объема данных.

Полученный результат не дает исчерпывающего ответа на вопрос задачи прогнозирования, так как остаются неопределенность и вероятность возникновения непредвиденных событий (признаков), которые могут составлять зависимости, при которых точность прогноза будет меняться [4]. На рис. 6 представлен набор значений, описывающих посещаемость занятий каждого студента (прогноз). На рис. 7 отображена модель неполного дерева решения для данной задачи с учетом всех вышеописанных признаков. Полное дерево можно найти при переходе по ссылке проекта в библиографическом списке [5, 6].

```
Answer: [1 0 0 1 1 0 1 1 1 0 1 0 0 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 1 0 1 1 1 1 0
 1 0 0 1 1 1 1 1 1 0 1 1 0 0 1 0 0 1 1 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0
 1 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 1 0 0 0 1
 0 0 1 1 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 1 0 1 1 0 1 0 0 1 1 0 0 0 1 0 0 0 0
 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 1 1 0 1 0 1 0 1 0 0 1 0
 1 0 0 0 1 0 1 0 0 0 1 0 0 0 1 1 1 0 1 1 0 0 1 0 0]
     Answer
681       1
626       0
329       0
620       1
399       1
..      ...
160       0
442       0
611       1
578       0
103       0

[210 rows x 1 columns]
```

Рис. 6. Прогноз посещаемости студентов

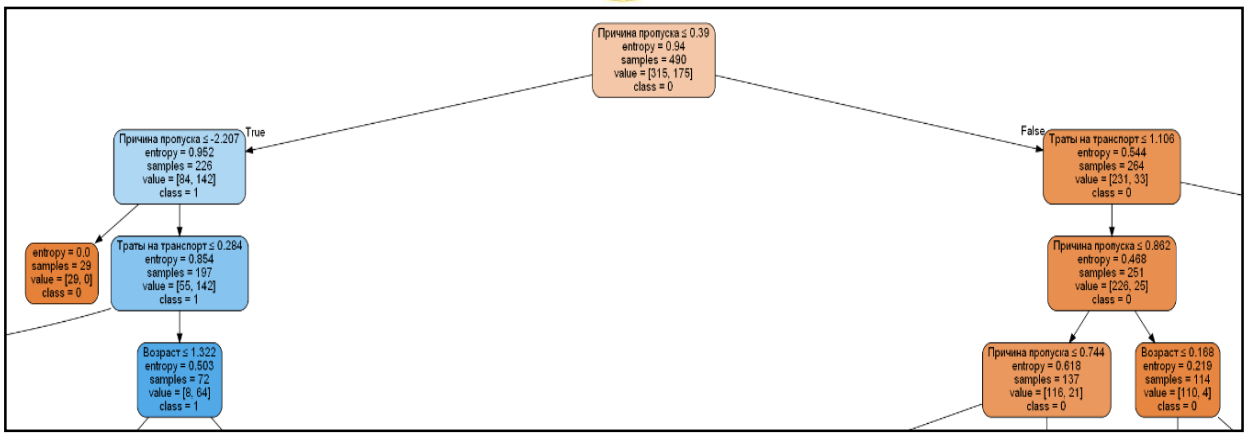

Рис. 7. Неполное дерево решения

Заключение. Таким образом, в ходе данной работы были рассмотрены основные понятия, задачи, виды машинного обучения, актуальные на сегодняшний день, также была реализована модель такого обучения на примере задачи прогнозирования посещаемости занятий студентом, основанной на заранее собранных реальных данных, сформированных путем опроса и анализа статистики.

Данная реализация потребовала собрать исходные данные путем анализа представленной статистики и анкетирования студентов через сервис Google Forms, затем были выбраны начальные признаки (поля анкеты) для обучения данной модели. В процессе анализа результатов анкетирования также были отброшены или переформированы некоторые признаки, которые в итоге представляли наибольшую зависимость. Разработанная модель составляет прогноз посещаемости занятий студентом и может быть использована в образовательных учреждениях, а также модифицирована для предприятий с целью отслеживания посещаемости сотрудников.

Библиографический список

1. Alekseev, Grigoriy. Введение в машинное обучение / Grigoriy Alekseev // Хабр : [сайт]. — URL: https://habr.com/ru/post/448892/ (дата обращения: 25.04.21).

2. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах ; [пер. с англ.]. — Москва : ДМК Пресс, 2015. — 400 с.

3. Шлезингер, М. Десять лекций по статистическому и структурному распознаванию / М. Шлезингер, В. Главач. — Киев : Наукова думка, 2004. — 545 с.

4. Айвазян, С. А. Прикладная статистика: исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — Москва : Финансы и статистика, 1985. — 487 с.

5. ModelPeople / Github.com : [сайт]. — URL: https://github.com/Ahostility/modelPeople (дата обращения: 25.04.2021).

6. Обухов, А. Д. Автоматизация распределения информации в адаптивных системах электронного документооборота с применением машинного обучения / А. Д. Обухов // Advanced Engineering Research. — 2020. — Т. 20, №4. — С. 430–436. https://doi.org/10.23947/2687-1653-2020-20-4-430-436

Об авторах:

Третьяк Михаил Андреевич, студент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344003, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), ale1431999@gmail.com

Фуников Владислав Витальевич, студент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344003, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), vlad@funikov.ru

Сафарьян Ольга Александровна, преподаватель кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344003, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), доцент, кандидат технических наук, safari_2006@mail.ru

Щекатурин Александр Евгеньевич, студент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344003, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), lizardwizardetofake@gmail.com

Орехов Павел Павлович, студент кафедры «Кибербезопасность информационных систем» Донского государственного технического университета (344003, РФ, г. Ростов-на-Дону, пл. Гагарина, 1), opekhovpavel@gmail.com

Authors:

Tretyak, Mikhail A., Student, Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, RF, 344003), ale1431999@gmail.com

Funikov, Vladislav V., Student, Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, RF, 344003),vlad@funikov.ru

Safaryan, Olga A., Lecturer, Student, Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, RF, 344003), Associate professor, Cand.Sci., safari_2006@mail.ru

Shchekaturin, Aleksandr E., Student, Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, RF, 344003), lizardwizardetofake@gmail.com

Orekhov, Pavel P., Student, Department of Cybersecurity of Information Systems, Don State Technical University (1, Gagarin sq., Rostov-on-Don, RF, 344003), opekhovpavel@gmail.com