Предварительная оценка успешности учащихся на курсе

АВТОР: ДЕЙКИНА СОФЬЯ
НАСТАВНИК: СВЕТЛОВА ВИКТОРИЯ СЕРГЕЕВНА

Проблема и актуальность

ПРОБЛЕМА

проблема нахождения оптимального метода оценки успешности учащихся на курсе

АКТУАЛЬНОСТЬ

прогноз результативности выполнения курса может помочь автору уже на небольшом сроке увидеть
тех людей, которые в будущем смогут закончить курс успешно

Цель и задачи

Цель: создать программу, которая будет успешно прогнозировать результативность выполнения курса

предобработка данных

выбор оптимального алгоритма для обучения

реализация программного кода и обучения на выборке

анализ полученных результатов

ПРЕДОБРАБОТКА ДАННЫХ

01

Предоставленные данные

Данные о решениях
Данные о комментариях
Данные о списывании
Данные о сертификатах

Итоговые данные

Единый датасет с отобранными значениями

	user_id	last_timestamp	correct	wrong	day	certificate_url	cheating	comments
0	127890	1.589470e+09	15.0	1.0	1.0	0.0	6.0	0.0
1	220676	1.589193e+09	9.0	28.0	1.0	1.0	26.0	0.0
2	326719	1.589318e+09	75.0	15.0	1.0	0.0	51.0	0.0
3	1635722	1.588280e+09	15.0	4.0	1.0	1.0	20.0	0.0
4	2258383	1.588067e+09	7.0	0.0	1.0	1.0	83.0	0.0
...	...	...	...	...	...	...	...	...
829	251229774	1.591801e+09	39.0	33.0	2.0	1.0	45.0	0.0
830	251294630	1.591733e+09	7.0	0.0	1.0	1.0	4.0	0.0
831	251664717	1.591797e+09	68.0	76.0	1.0	0.0	63.0	0.0
832	252308518	1.591875e+09	136.0	5.0	1.0	0.0	71.0	0.0
833	253108923	1.591967e+09	4.0	0.0	1.0	1.0	4.0	0.0

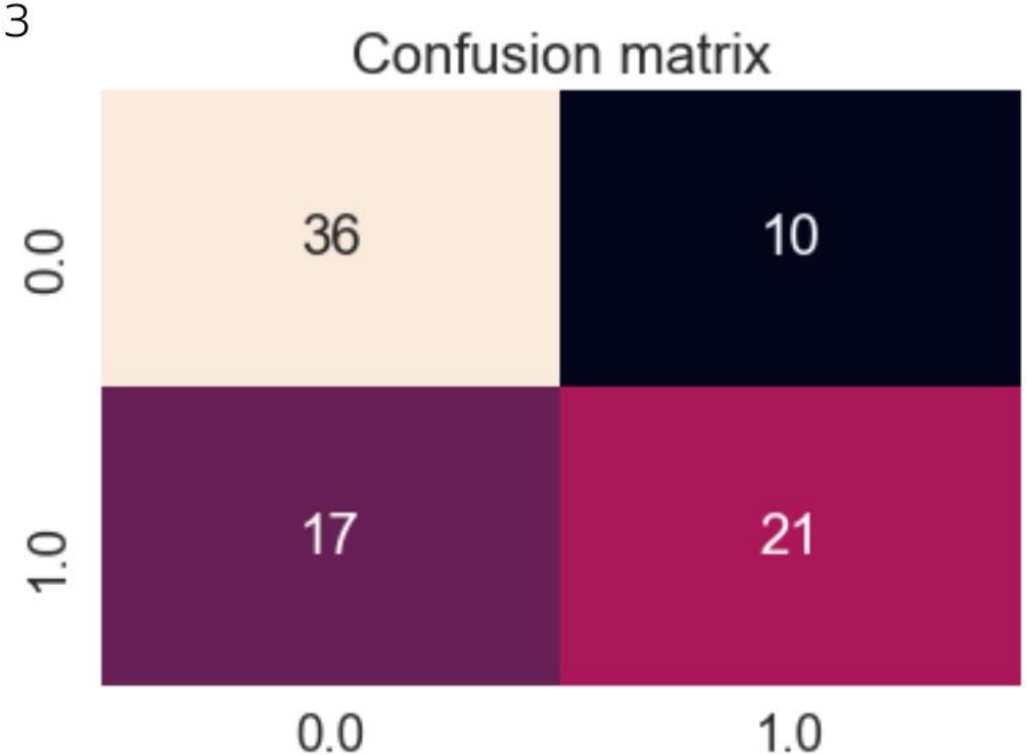
ВЫБОР АЛГОРИТМА

02

RandomForestClassifier

Random forests - это алгоритм, составленный из множества деревьев решений

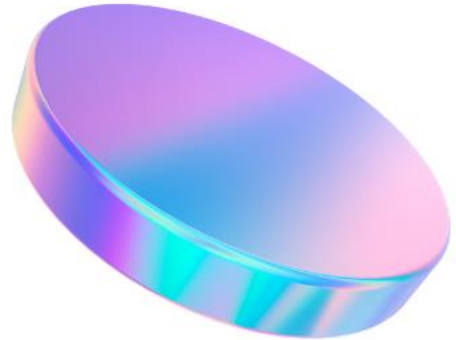
	Model	Accuracy
0	RandomForestClassifier	0.68
1	SGDClassifier	0.55
2	SVC	0.55
3	DecisionTreeClassifier	0.67
4	ExtraTreeClassifier	0.61
5	LogisticRegressionCV	0.55
6	LogisticRegression	0.55



ОБУЧЕНИЕ И РЕАЛИЗАЦИЯ

03

Язык программирования	**Python**
Библиотека для обучения модели	**RandomForest Classifier**
Библиотека для работы с данными	**Pandas**
	Numpy

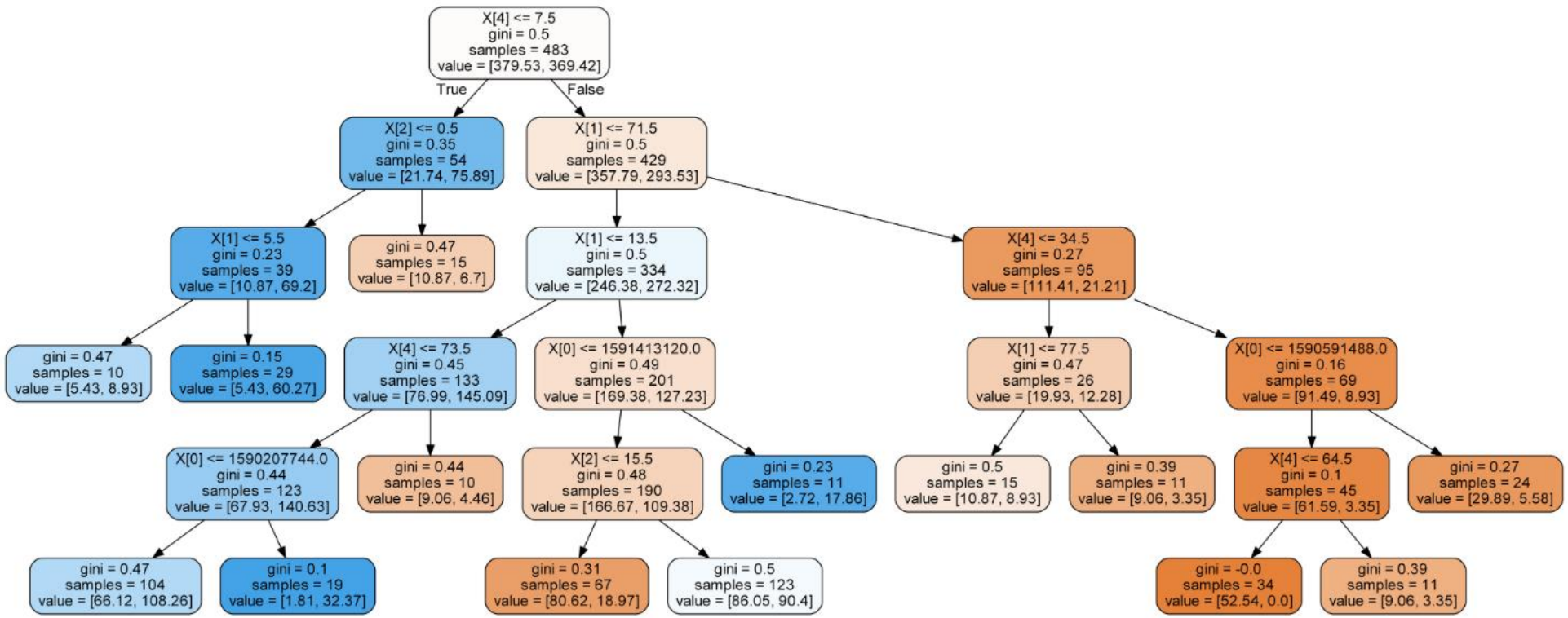


АНАЛИЗ РЕЗУЛЬТАТОВ

04

Визуализация

roc на test 0.795

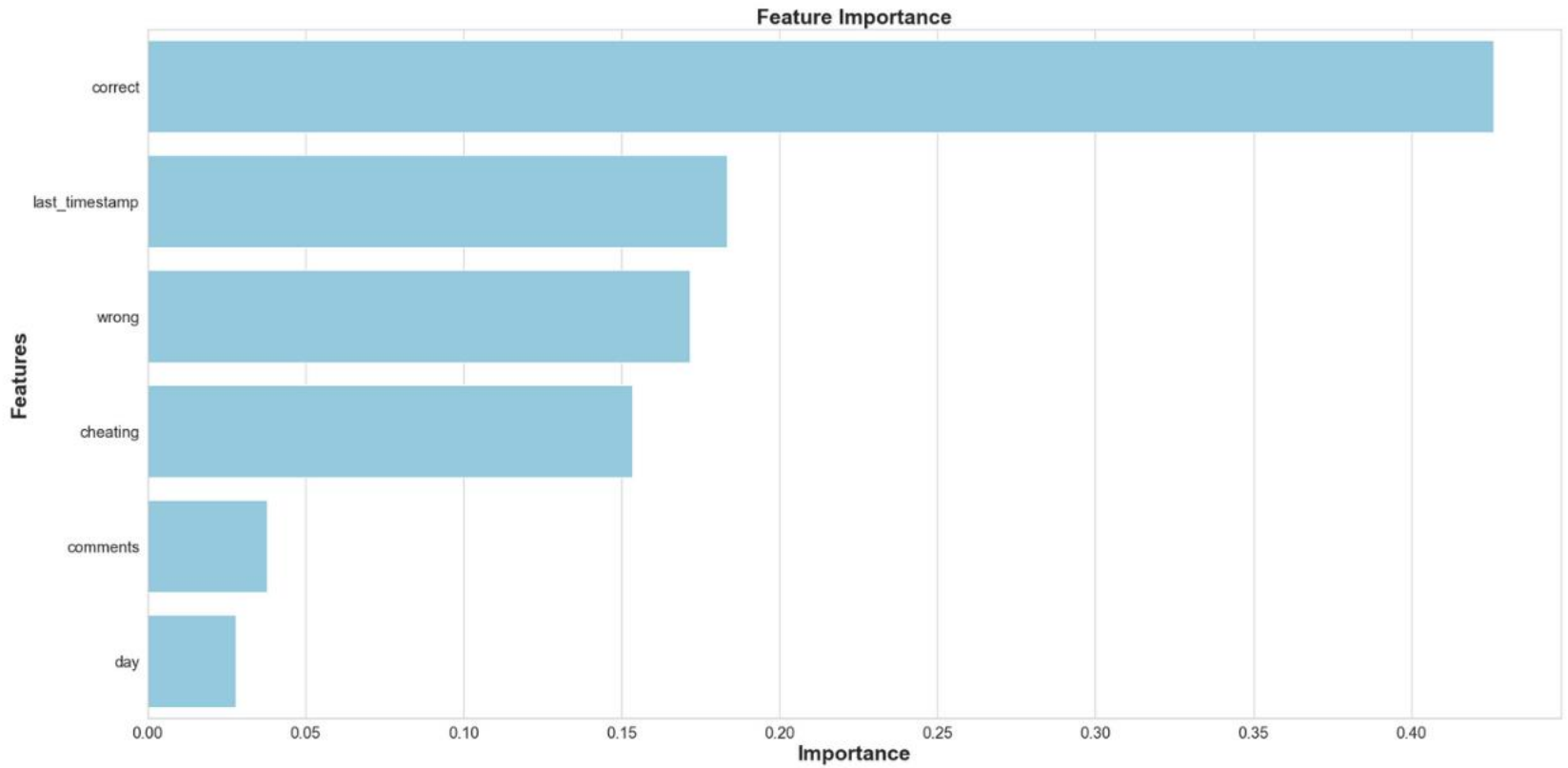


АНАЛИЗ РЕЗУЛЬТАТОВ

04

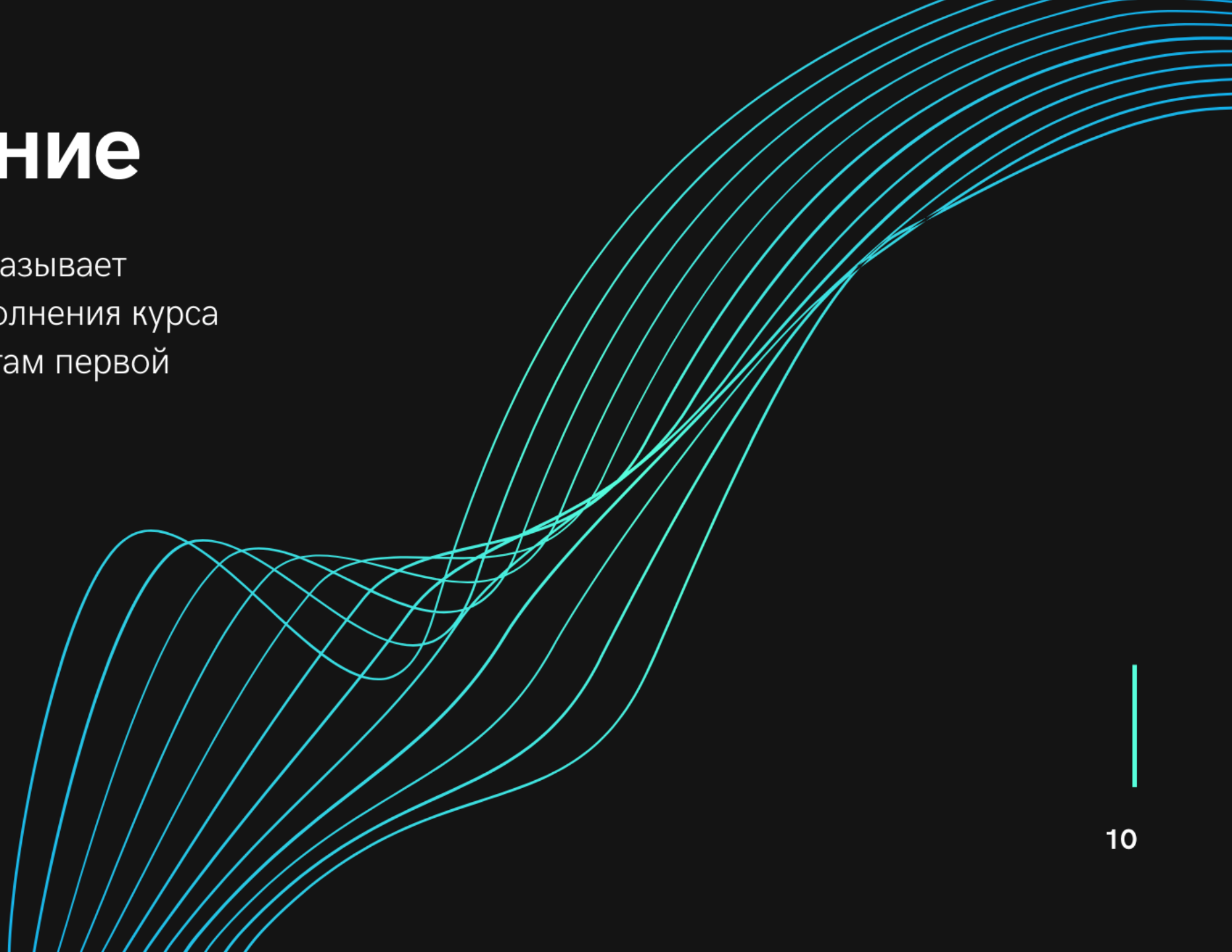
Важные параметры

	Features	Gini-Importance
0	correct	0.426048
1	last_timestamp	0.183236
2	wrong	0.171722
3	cheating	0.153422
4	comments	0.037799
5	day	0.027773



Заключение

Данная модель предсказывает результативность выполнения курса учеником по результатам первой недели работы.



Источники

КУРС ВВЕДЕНИЕ В DATA SCIENCE И МАШИННОЕ ОБУЧЕНИЕ

https://stepik.org/course/4852/syllabus

ДОКУМЕНТАЦИЯ RANDOMFOREST

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

ССЫЛКА НА GITHUB С ИСХОДНЫМ КОДОМ

https://github.com/deisof/BV2021