Столыпинский вестник

Научная статья

Original article

УДК: 004

КРЕДИТНЫЙ СКОРИНГ, РЕАЛИЗОВАННЫЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

CREDIT SCORING IMPLEMENTED USING MACHINE LEARNING

Воронин Сергей Михайлович, студент, МГТУ им Н.Э.Баумана, Россия, г. Москва, volkodav.sergey00@mail.ru

Совертека Захар Константинович, студент, МГТУ им Н.Э.Баумана, Россия, г. Москва, zsoverteka@mail.ru

Березин Андрей Дмитриевич, студент, МГТУ им Н.Э.Баумана, Россия, г. Москва, a.d.berezin@yandex.ru

Ларин Алексей Игоревич, студент, МГТУ им Н.Э.Баумана, Россия, г. Москва msokol99@mail.ru

Voronin Sergey Mikhailovich, student, Bauman Moscow State Technical University Russia, Moscow, volkodav.sergey00@mail.ru

Soverteka Zakhar Konstantinovich, student, Bauman Moscow State Technical University, Russia, Moscow, zsoverteka@mail.ru

Berezin Andrey Dmitrievich, student, Bauman Moscow State Technical University Russia, Moscow, a.d.berezin@yandex.ru

Larin Alexey Igorevich, student, Bauman Moscow State Technical University Russia, Moscow, msokol99@mail.ru

Аннотация

Нейронные сети вошли в практику везде, где нужно решать задачи прогнозирования, классификации или управления. Одним из примеров их использования является анализ кредитоспособности заемщика в банковской сфере. Актуальность данной темы обуславливается сложностью перебора огромной базы данных и расчета параметра, на основе которого производится оценка кредитоспособности заемщика.

В данной работе будет проанализирован набор данных кредитного скоринга, на основании которого будет определено, стоит ли выдавать определенному лицу кредит или нет.

Annotation

Neural networks have entered the practice wherever it is necessary to solve problems of forecasting, classification or control. One example of their use is the analysis of the creditworthiness of a borrower in the banking sector. The relevance of this topic is due to the complexity of sorting through a huge database and calculating the parameter on the basis of which the borrower's creditworthiness is assessed.

In this paper, a set of credit scoring data will be analyzed, on the basis of which it will be determined whether it is worth issuing a loan to a certain person or not.

Схема построения модели машинного обучения

Построение модели машинного обучения всегда сводится к определенной схеме. Для полного анализа необходимо пройтись по каждому ее шагу и разобрать получившийся результат модели кредитного скоринга.

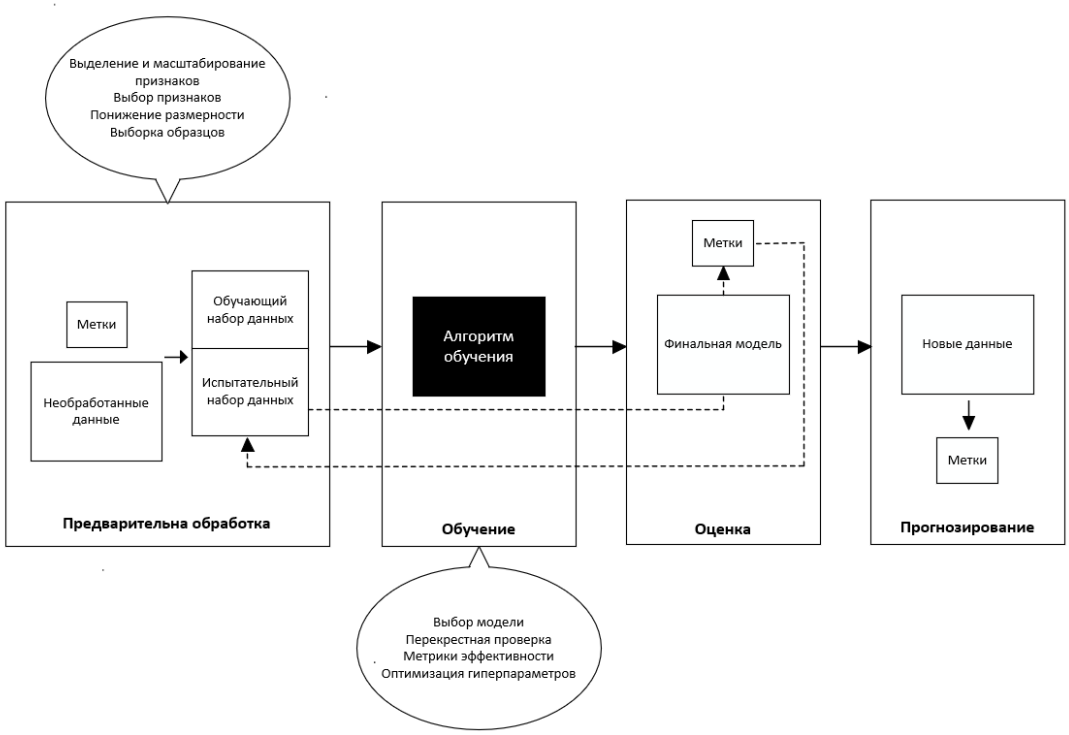

Рисунок 1 – Схема построения модели машинного обучения

Предварительная обработка

Необработанные данные редко поступают в виде, необходимом для обеспечения оптимальной эффективности алгоритмов машинного обучения.

	client_id	app_date	education	sex	age	car	car_type	decline_app_cnt	good_work	score_bki	bki_request_cnt	region_rating	home_address	work
0	25905	01FEB2014	SCH	M	62	Y	Y	0	0	-2.008753	1	50	1	
1	63161	12MAR2014	SCH	F	59	N	N	0	0	-1.532276	3	50	2	
2	25887	01FEB2014	SCH	M	25	Y	N	2	0	-1.408142	1	80	1	
3	16222	23JAN2014	SCH	F	53	N	N	0	0	-2.057471	2	50	2	
4	101655	18APR2014	GRD	M	48	N	N	0	1	-1.244723	1	60	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
73794	54887	04MAR2014	GRD	F	45	N	N	0	0	-1.792064	3	50	1	
73795	76821	24MAR2014	SCH	M	41	Y	Y	0	0	-2.058029	1	50	2	
73796	103695	22APR2014	SCH	M	31	N	N	0	0	-1.512635	4	80	2	
73797	861	04JAN2014	SCH	F	29	N	N	0	1	-1.479334	3	50	1	
73798	15796	23JAN2014	GRD	M	34	N	N	0	0	-1.764711	2	50	2	

73799 rows × 19 columns

Рисунок 2 – Входные необработанные данные

В представленном случае в банке может вестись своя классификация клиентов по определенным признакам. Из всего этого вытекает важная задача в правильной избирательности данных. Не все признаки полезны в решении определенно поставленной задачи, для чего нужно отсеять все ненужное [3, с.43]. Если этого не сделать, то эффективность итоговой модели снизится.

В первую очередь нужно убрать данные, которые ничего не покажут. В данном случае можно смело убрать данные по признаку «ID клиента», ведь ID не является определяющим признаком заемщика.

Для определения значимых переменных можно использовать гипотезы, на основании которых выясняется, как определенный признак влияет на целевую переменную. В нашем случае целевая переменная это флаг default.

Для этого придумывают гипотезы, чем больше их будет придумано, тем лучше. Таким образом наиболее актуальными гипотезами будут следующие:

- Возраст "хороших" заемщиков больше, по сравнению с "плохими": распределения возраста в зависимости от флага default смещено в большую сторону при default=0
- Уровень образования зависит от возраста, что влияет и на возврат кредита, также люди с высшим образованием чаще являются "хорошими" заемщиками
- При good_work = 0 увеличивается риск невозврата кредита
- Доход "хороших" заемщиков больше по сравнению с "плохими": распределения дохода в зависимости от флага default смещено в большую сторону при default=0
- score_bki напрямую взаимосвязан с default, чем он меньше тем выше вероятность клиента выплатить кредит банку

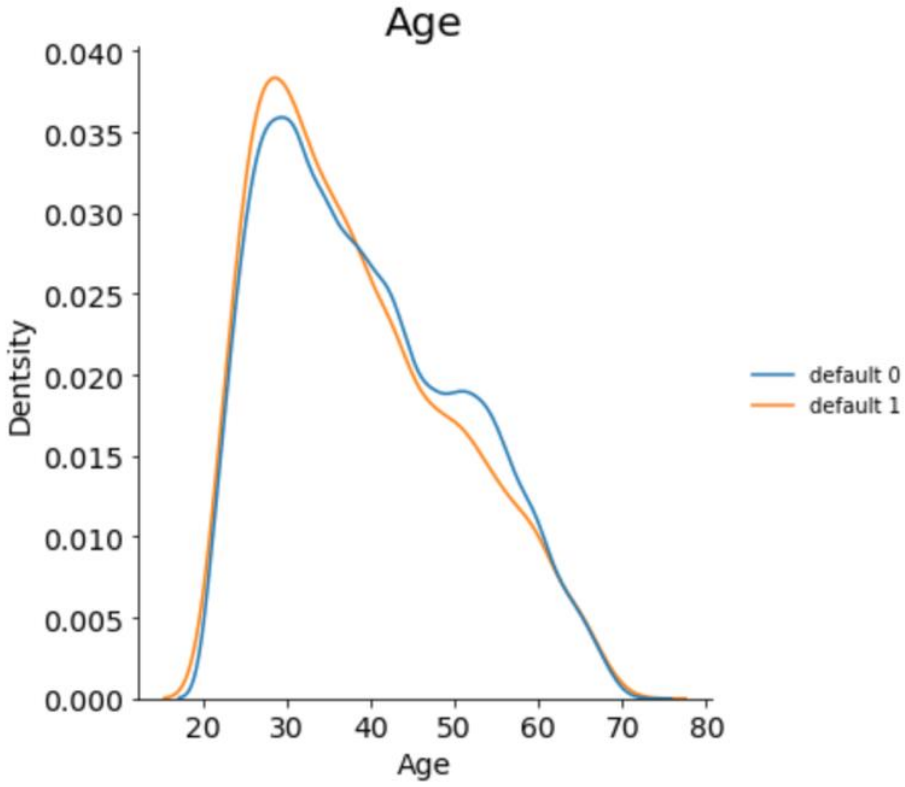

Рисунок 3 – График зависимости возраста и флага по default

Для анализа признаков «зарплата» и «возраст» возьмем средние значения по default = 0 и default = 1.

Таблица 1– Средние значения

default	Зарплата	Возраст
0	41799.71	38
1	36288.11	36

Проанализировав график зависимости возраста и флага по default и средние значения признаков, был сделан вывод, что данные столбцы будут влиять на итоговую модель, а также зарплата будет более показательным параметром чем возраст, из-за большего разброса в значениях при default = 0 и default = 1.

Также возможны случаи, когда есть слишком связанные между собой данные. Если удалить одни из них, они не повлияют на конечный результат, однако если их оставить, то время работы программы увеличится, так как модель будет обучаться с большим количеством переменных. Это называется понижение размерности пространства признаков.

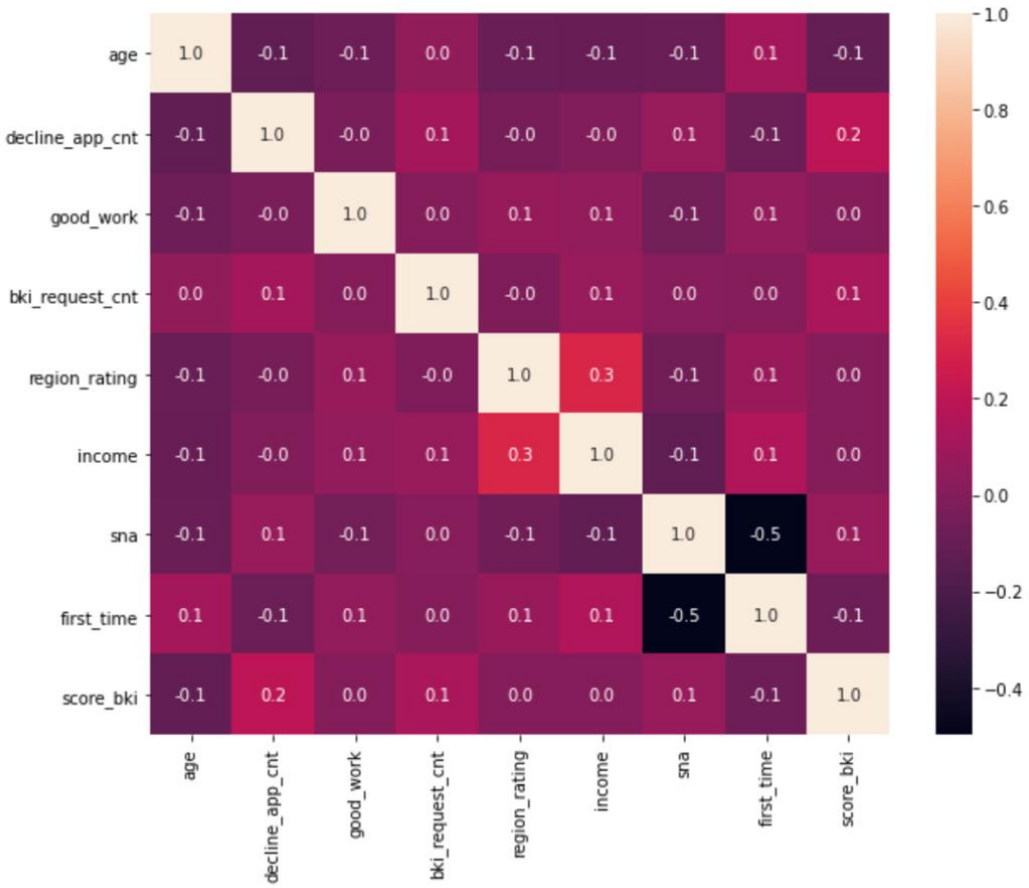

Рисунок 4 – График корреляции признаков

Как только данные будут обработаны, необходимо случайным образом разделить их на обучающие и испытательные наборы. Обучающий набор применяется для обучения и оптимизации модели машинного обучения, тогда как испытательный набор хранится до самого конца и предназначен для оценки финальной модели [2, с.43].

Обучение и выбор прогнозируемой модели

На данном шаге нужно определится с прогнозируемой моделью. Чтобы понять, какой моделью пользоваться, нужно попробовать несколько из них, ведь все модели находятся в равных условиях, только если не были сделаны какие-либо допущения для определенного случая [1, с.26]. В представленной задаче воспользуемся логической регрессией, так как она лучше всего подходит для бинарной классификации. На выходе выводятся величины, показывающие эффективность нашей модели, а именно метрики.

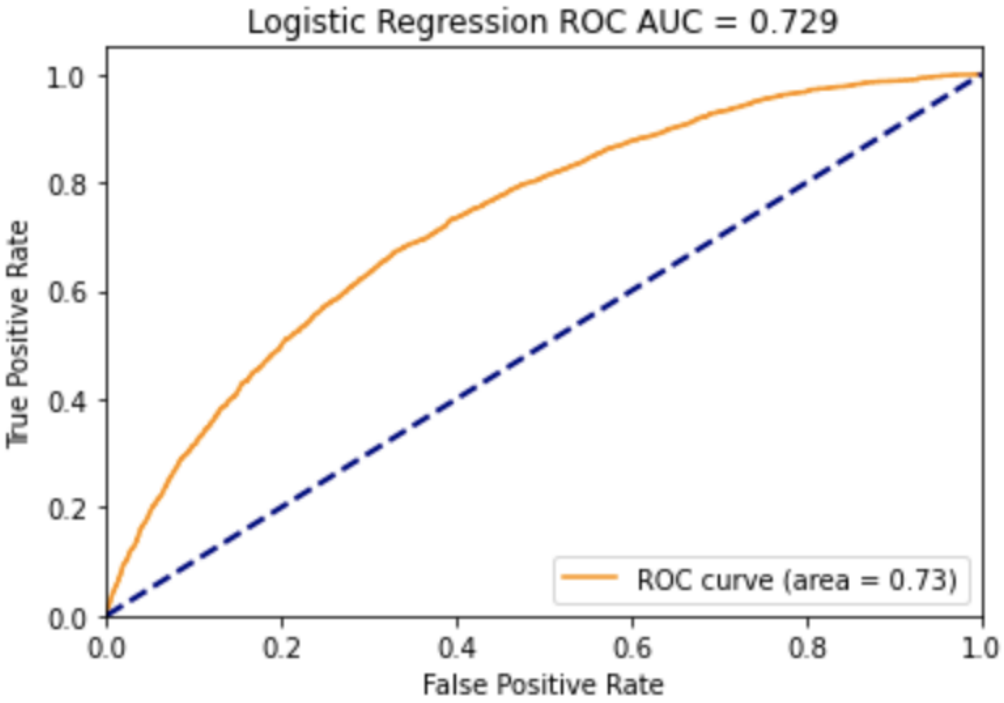

Рисунок 5 – Кривая ошибок

Часто результат работы алгоритма на фиксированной тестовой выборке визуализируют с помощь ROC-кривой или «кривой ошибок», а качество оценивают как площадь под этой кривой – AUC. Величина AUC ROC равна доле пар объектов вида, которые алгоритм правильно упорядочил [5, с.861].

По полученному значению ROC AUC = 0,729 можно сделать вывод, что модель хорошо обучилась, так как она смогла, верно, классифицировать большую часть данных.

Оценка моделей и прогнозирование испытательным набором

После выбора модели, подогнанной к обучающему набору данных, можно использовать испытательный набор данных для оценки точности работы модели на ранее не встретившихся данных.

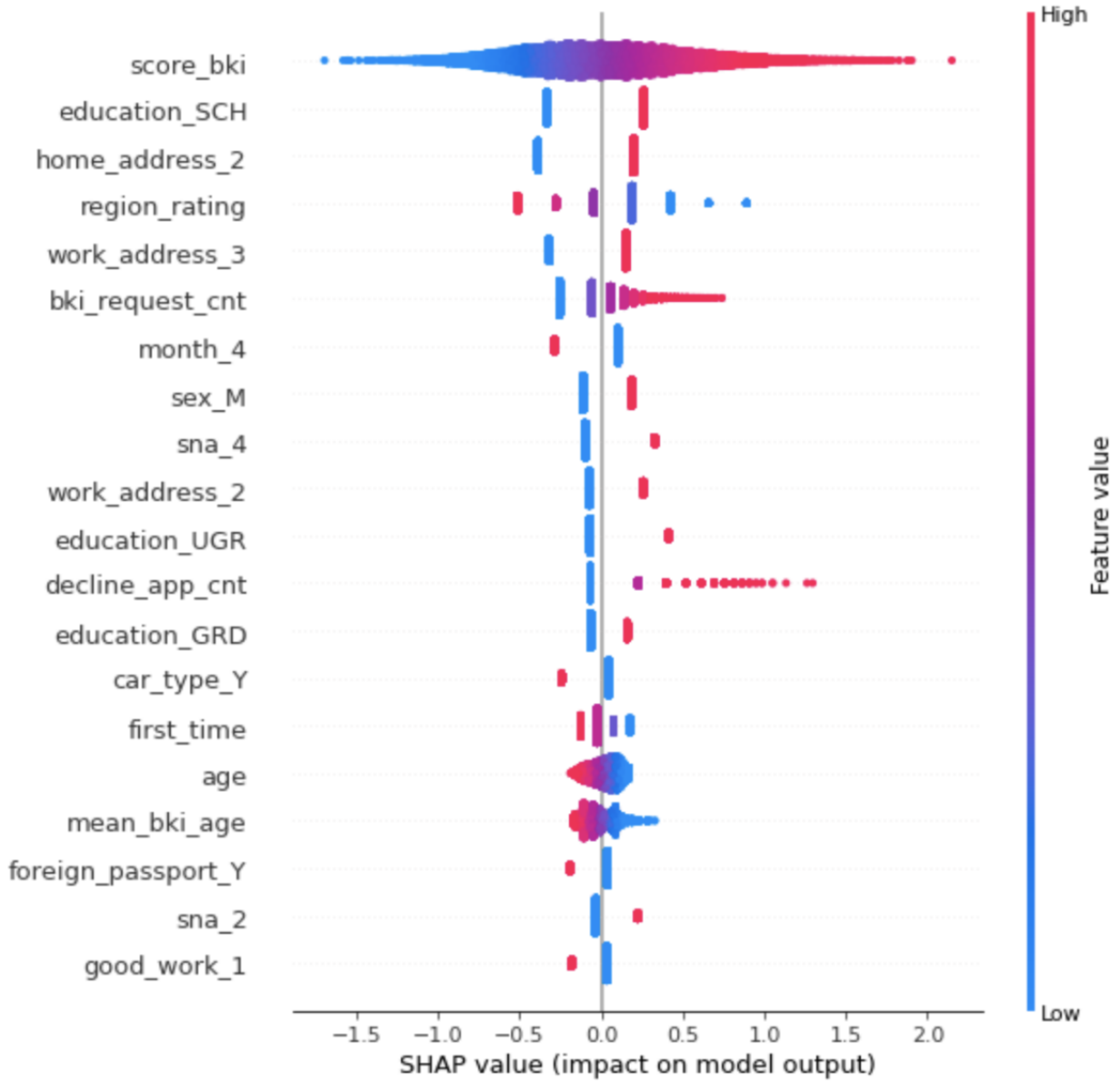

Рисунок 6 – Прогнозная модель

Значение Shap value показывает параметр нашей целевой функции и находится на оси OX, чем больше его значение по сравнению с 0, тем больше значение целевой функции стремится к default = 1 и показывает, что заемщик «плохой». Чем ниже 0, тем больше оно стремится к default = 0 и показывает, что заемщик «хороший».

Признаки расположены по степени их важности по оси OY слева от графика. Каждая точка является отдельно взятым наблюдением, в то время как, цветом обозначены значения соответствующего признака: красный – высокое значение, синий – низкое значение.

Анализ эффективности применения «кредитного скоринга» с помощью алгоритмов машинного обучения

Скоринг представляет собой статистическую модель, с помощью которой на основе кредитной истории «прошлых» клиентов банк пытается определить, насколько велика вероятность, что конкретный потенциальный заемщик вернет кредит в срок. Доступность заемных средств для бизнеса и населения дает им возможность повышать инвестиционную и потребительскую активность, что увеличивает масштабы производства [4, с.6].

С помощью алгоритмов машинного обучения можно автоматизировать процесс «кредитного скоринга». Данное решение позволит:

- Сократить время, требуемое для «кредитного скоринга»
- Уменьшить вероятность ошибки принятого решения
- Работать с большим количеством данных заемщиков

Список литературы

1. Дж. Грас – Data Science. Наука о данных с нуля, 2015.
2. Рашка Себастьян – Python и машинное обучение, 2017.
3. Педро Домингос – Верховный алгоритм, 2015.
4. Ковальчук В.М. – Роль банковского кредитования в развитии экономики страны и его проблемы, 2017.
5. T. Fawcett – An introduction to ROC analysis, 2006.

List of literature

1. J. Grasse – Data Science. Data Science from Scratch, 2015.
2. Rashka Sebastian – Python and Machine Learning, 2017.
3. Pedro Domingos – Supreme Algorithm, 2015.
4. Kovalchuk V.M. – The role of bank lending in the development of the country's economy and its problems, 2017.
5. T. Fawcett – An introduction to ROC analysis, 2006.

Для цитирования: Воронин С.М., Совертека З.К., Березин А.Д., Ларин А.И. КРЕДИТНЫЙ СКОРИНГ, РЕАЛИЗОВАННЫЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ// Научный сетевой журнал «Столыпинский вестник» №10/2022.