УДК: 330.43

Использование машинного обучения в инвестиционной деятельности

Казанская А.А. a.a.nersisian@gmail.com
Канд. экон. наук, доцент **Мишура Л.Г.** mishuralg@rambler.ru
Университет ИТМО
197101, Россия, Санкт Петербург, ул. Чайковского, д. 11/2

Этап развития современного общества можно охарактеризовать, как наиболее значимый и определяющий период для информационных технологий. Период, когда вычисления перешли от крупных электронных вычислительных машин к машинам и роботам с автоматическим управлением, с возможностью хранить данные в облаке. Но что делает этот период по-настоящему захватывающим, так это демократизация различных инструментов и методов, которые развивались одновременно с развитием вычислительной техники. Обработка данных, которая когда-то занимала несколько дней, сегодня занимает всего несколько минут, и все это благодаря развитию машинного обучения. Это причина, по которой Data Science получает колоссальные инвестиции каждый год, что увеличивает спрос на сертификаты в данной области. Российская венчурная компания и партнёр инвестиционной компании iTech Capital Алексей Соловьев представили исследование «Венчурный барометр» о российском венчурном рынке. К участию данного исследования было приглашено 312 представителей венчурного рынка, из которых 83 ответили на вопросы анкеты. Барометр ежегодно задает инвесторам вопрос о наиболее привлекательных сегментах рынка, и, в 2019 году, как и в прошлые 2 года, рейтинг возглавляет искусственный интеллект и машинное обучение. За данную нишу проголосовали 82% инвесторов, как за наиболее перспективную [15]. Данные и методы: Задачей данного исследования является построение оптимального алгоритма, прогнозирующего значение рейтинга инвестиционного проекта при заданном наборе данных. Рассмотрены подробно наиболее популярные алгоритмы машинного обучения, выполнен подробный анализ имеющихся данных, построена модель на основе алгоритма случайного леса по набору тренировочного набора данных, а также проведено тестирование на основе тестовых данных. Данная исследовательская работа выполнена на языке программирования R, с помощью которого осуществлен анализ данных, построение различных таблиц и графиков, а также построена модель и проведена оценка результатов, на основе сравнения с ранее построенными алгоритмами: линейная регрессия, логистическая регрессия. Анализ результатов: В ходе исследования было выявлено, что наиболее оптимальным алгоритмом для прогнозирования рейтинга инвестиционного проекта на основе имеющихся данных является алгоритм случайного леса, точность данного алгоритма на 3.7% выше точности алгоритма линейной регрессии на основе наиболее значимого набора показателей проекта.

Ключевые слова: Информационные технологии, инвестиционная деятельность, инновации, машинное обучение, алгоритм случайного леса.

DOI: 10.17586/2310-1172-2020-13-2-23-34

Using machine learning in investment activity

Kazanskaya A.A. a.a.nersisian@gmail.com
Ph.D. **Mishura L.G.** mishuralg@rambler.ru
ITMO University
197101, Russia, St. Petersburg, 11/2 Tchaikovsky St.

We probably live in the most defining time for information technology. The period when computing switched from large electronic computers to automatic control machines and robots, with the ability to store data in the cloud. But what makes this period truly exciting, is the democratization of various tools and methods that developed simultaneously with the development of computer technology. Data processing, which once took several days, today takes only a few minutes, and all this thanks to the development of machine learning. This is the reason why Data Science receives huge investments every year, which increases the demand for certificates in this area. Russian venture capital company and partner of the investment company iTech Capital Alexey Soloviev presented the study "Venture Barometer" on the Russian venture market. 312 representatives of the venture capital market were invited to participate in this study, of which 83 answered the questionnaire. The barometer annually asks investors about the

most attractive market segments, and in 2019, as in the past 2 years, the ranking is headed by artificial intelligence and machine learning. 82% of investors voted for this field as the most promising [15]. Data and Methods: The objective of this study is to build an optimal algorithm that predicts the rating value of an investment project for a given data set. The most popular machine learning algorithms are examined in detail, a detailed analysis of the available data is carried out, a model based on a random forest algorithm for a training data set is constructed, and testing is conducted based on test data. This research work was performed in the programming language R, with the help of which the data were analyzed, various tables and graphs were constructed, a model was built and the results evaluated based on a comparison with previously constructed algorithms: linear regression, logistic regression. Analysis of Results: The study revealed that the most optimal algorithm for predicting the rating of an investment project based on available data is the random forest algorithm, the accuracy of this algorithm is 3.7% higher than the accuracy of the linear regression algorithm based on the most significant set of project indicators.
Keywords: Information technology, investment activities, innovation, machine learning, random forest algorithm.

Введение

Интерес к машинному обучению особенно увеличился за годы, прошедшие после публикации в Harvard Business Review статьи «Data Scientist» [16]. В настоящее время существуют различные алгоритмы машинного обучения, разработанные для решения сложных реальных задач. Эти алгоритмы являются высокоавтоматизированными и самоизменяющимися, поскольку они продолжают совершенствоваться с увеличением объема данных при минимальном вмешательстве человека.

Теоретические основы

Алгоритмы машинного обучения строят математическую модель на основе выборочных данных, известных как «набор данных для обучения», для того, чтобы делать прогнозы или принимать решения для определенных типов задач без явного программирования [1]. Алгоритмы машинного обучения используются в решении самых разнообразных задач, таких как фильтрация электронной почты и машинное зрение, где сложно или невозможно разработать традиционный алгоритм для эффективного выполнения задачи. Процесс обучения в простой модели машинного обучения делится на два этапа: обучение и тестирование. В процессе обучения в качестве входных данных берутся объекты обучающих данных, которые изучаются с помощью алгоритма и строят модель обучения [2].

Алгоритмы машинного обучения можно разделить на несколько различных типов:

- обучение с учителем (Supervised Learning)

В обучении с учителем есть известный набор входных параметров (признаков) и выходных параметров (меток). Обычно их обозначают X и Y. Цель алгоритма – изучить функцию отображения, которая отображает входные данные в выходные данные. Так что, когда даны новые примеры X, машина может правильно предсказать соответствующие метки Y [3].

В рамках данного типа обучения рассматриваются задачи классификации, регрессии и прогнозирования [4]. Наиболее широко используемыми алгоритмами обучения с учителем являются:

1. линейная регрессия
2. логистическая регрессия
3. дерево решений
4. ансамбли моделей (бэггинг, бустинг, случайный лес)
5. наивный байесовский классификатор
6. метод k-ближайших соседей
7. метод опорных векторов

- обучение без учителя (Unsupervised Learning)

В обучении без учителя имеется лишь набор входных данных (X) и нет соответствующих меток (Y). Цель алгоритма – найти ранее неизвестные закономерности в данных. Довольно часто эти алгоритмы используются для нахождения значимых кластеров схожих выборок X, фактически ищут категории, свойственные данным [3], [5].

Основные применения обучения без учителя - задачи кластеризация и уменьшение размерности. К широко используемым алгоритмам относятся:

1. метод k-средних
2. метод нелинейного снижения размерности и визуализации многомерных переменных (t-Distributed Stochastic Neighbor Embedding)
3. метод главных компонент
4. ассоциативное правило

- обучение с частичным привлечением учителя (Semi-supervised Learning)

В предыдущих двух типах либо отсутствуют метки для всех наблюдений в наборе данных, либо присутствуют. Обучение с частичным привлечением учителя является объединением обучения с учителем и без учителя и включает в себя функции обоих типов обучения [5], [6].

- обучение с подкреплением (Reinforcement Learning)

В обучении с подкреплением испытуемая система (агент) обучается оптимальным действиям взаимодействуя со средой методом проб и ошибок, используя обратную связь от своих собственных действий и опыта. Алгоритм определяет следующее действие, изучая поведение, основанное на его текущем состоянии и максимизирующее вознаграждение в будущем. Обучение с подкреплением является типом машинного обучения и, также, отраслью искусственного интеллекта. Это позволяет машинам и программным агентам автоматически определять идеальное поведение в определенном контексте, чтобы улучшить его производительность [7].

На самом деле, обучение с подкреплением определяется конкретным типом задач, и все решения классифицируются как алгоритмы обучения с подкреплением. В задаче агент должен решить, какое действие лучше выбрать, исходя из его текущего состояния. Когда этот шаг повторяется, задача известна как Марковский процесс принятия решений. Наиболее активно используемыми алгоритмами являются:

1. Q-Learning
2. Temporal difference (TD)
3. Deep Adversarial Networks

Рассмотрим подробнее некоторые алгоритмы машинного обучения с учителем.

1. Логистическая регрессия

Логистическая регрессия является подходящим регрессионным анализом, который необходимо проводить, когда зависимая переменная является дихотомической (двоичной): наборы данных, где y = 0 или 1. Например, при прогнозировании того, произойдет ли событие или нет, есть только две возможности: что оно происходит (которое обозначается как 1) или нет (0).[12], [8].

Логистическая регрессия похожа на линейную тем, что в ней тоже требуется найти значения коэффициентов для входных переменных. Разница заключается в том, что выходное значение преобразуется с помощью нелинейной или логистической функции. Как и все регрессионные анализы, логистическая регрессия является прогностическим анализом. Логистическая регрессия используется для описания данных и нахождения взаимосвязи между одной зависимой двоичной переменной и одной или несколькими номинальными, порядковыми или интервальными независимыми переменными.

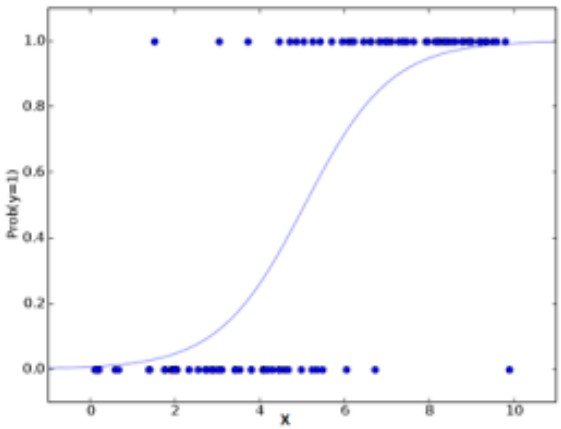

Рис. 1 Модель логистической регрессии

2. Дерево принятия решений

Дерево принятия решений - это ориентированный граф, который начинается с одного узла и продолжается на множество конечных узлов, отображающих категории, которые дерево может классифицировать. Здесь все возможные результаты решения показаны с использованием методологии ветвления дерева. Каждый узел дерева - это проверка для различных условий по определенной переменной, ветви дерева - это результат проверки, а конечные узлы - это решение, принятое после вычисления всех атрибутов [7].

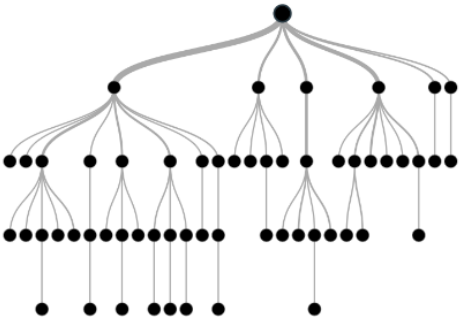

Рис. 2 Модель дерева решений

3. Наивный Байесовский классификатор

При необходимости классифицировать текстовые данные, такие как веб-страницы, документы или электронные письма, используется наивный алгоритм классификатора Байеса. Данный алгоритм опирается на теорему Байеса и классифицирует каждое значение элемента среди совокупности одной из доступных классов/категорий на основе заданного набора признаков, используя вероятности [7], [9].

где, y - переменная класса, а X - вектор зависимых объектов (размером n), где:

Алгоритм называется наивным, потому что предполагает, что каждая входная переменная независимая.

4. Метод k-ближайших соседей

Алгоритм использует весь набор данных в качестве обучающей выборки, а не разделяет данные на набор данных для обучения и теста.

Когда для нового набора данных требуется определить результат, алгоритм проходит весь набор данных, чтобы найти k-ближайших соседей для нового экземпляра, то есть k экземпляров, наиболее похожих на новую точку, а затем решает, к какой группе эта точка относится. Сходство между экземплярами рассчитывается с использованием таких мер, как евклидово расстояние и расстояние Хемминга.

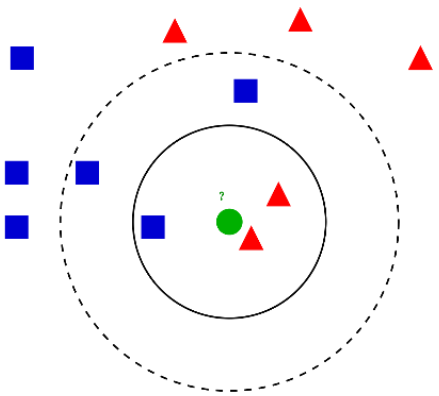

Рис. 3 Метод k-ближайших соседей.

5. Метод опорных векторов

Данный метод используется для задач классификации или регрессии. При этом данные делятся на разные классы путем нахождения конкретной линии (гиперплоскости), которая разделяет набор данных на несколько

классов. Алгоритм ищет лучшую или оптимальную гиперплоскость, разделяющую два класса, - это линия с наибольшей разницей, то есть с наибольшим расстоянием между гиперплоскостью и ближайшими элементами каждой из групп [13].

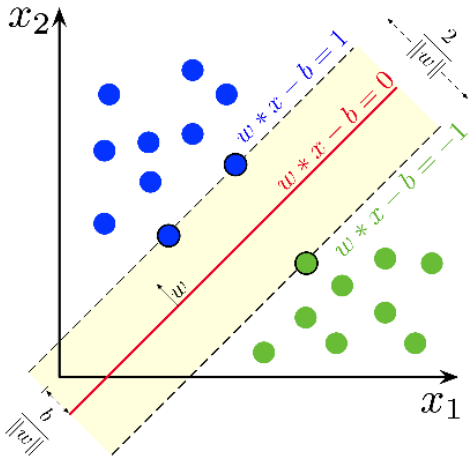

Рис. 4 Метод опорных векторов

6. Случайный лес

Случайный лес - алгоритм машинного обучения, который относится к задачам обучения с учителем. Данный алгоритм может быть использован для задач классификации, регрессии и кластеризации. Он основан на концепции обучения ансамбля, которая представляет собой процесс объединения нескольких классификаторов для решения сложной задачи и повышения производительности модели.

Как следует из названия, случайный лес - это классификатор, который содержит несколько деревьев решений в различных подмножествах данного набора данных и использует среднее значение для повышения точности прогнозирования этого набора данных. Вместо того, чтобы полагаться на одно дерево решений, случайный лес берет прогноз от каждого дерева и основывается на большинстве голосов прогнозов, и далее предсказывает окончательный результат [10].

Большее количество деревьев приводит к более высокой точности и предотвращает проблему переобучения. Приведенный ниже рисунок 5 отражает работу данного алгоритма:

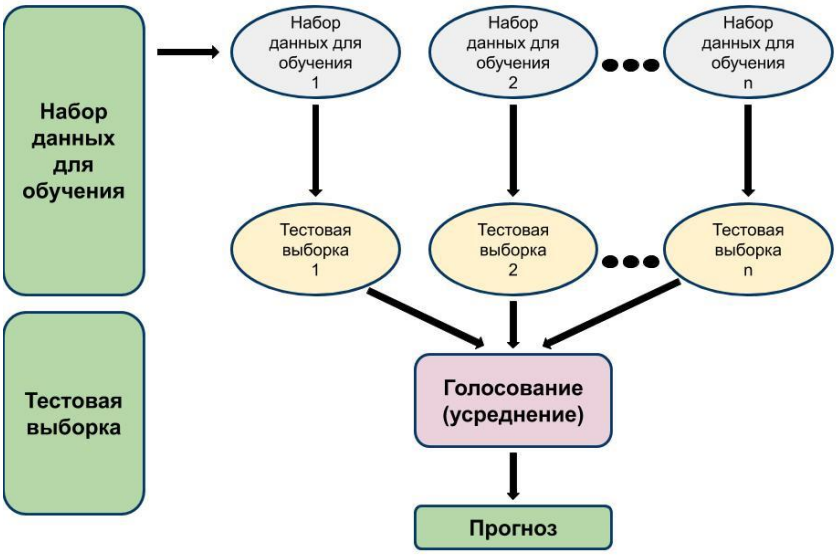

Рис. 5 Алгоритм случайного леса

Случайный лес работает в два этапа:

1. создается случайный лес путем объединения N деревьев решений
2. для каждого дерева, созданного на первом этапе, выполняется прогноз

Рассмотрим подробнее рабочий процесс алгоритма [17]:

1. выбрать случайные K выборок из заданного набора данных
2. далее алгоритм построит дерево решений для каждой выборки
3. получим результат прогнозирования из каждого дерева решений
4. на этом этапе голосование будет выполняться для каждого прогнозируемого результата
5. выбрать результат прогноза с наибольшим количеством голосов в качестве окончательного результата прогноза

Анализ данных

Прежде чем приступить к тестированию различных алгоритмов и отбору оптимальной модели прогнозирования, необходимо исследовать данные и понять, какие существуют предположения о взаимосвязях между переменными.

В статистике разведочный анализ данных (англ. exploratory data analysis, EDA) - это подход к анализу наборов данных для обобщения их основных характеристик, часто с помощью визуальных методов. Статистическая модель может использоваться или нет, но прежде всего EDA предназначена для того, чтобы увидеть, что данные могут сказать, помимо формальной задачи моделирования или проверки гипотез [11]. Джон Тьюки продвигал исследовательский анализ данных, чтобы побудить статистиков исследовать данные и, возможно, сформулировать гипотезы, которые могли бы привести к сбору новых данных и экспериментам.

Имеющаяся база данных тестируемых инвестиционных проектов имеет вид «объект-ответ», что предполагает использование обучения с учителем. Существует некоторая зависимость между объектами и ответами, которую требуется выяснить на основе обучающей выборки. То есть необходимо построить алгоритм, способный в дальнейшем выдать достаточно точный ответ для любого объекта.

В используемой базе данных информация об инвестиционных проектах компании задана с помощью значений следующих параметров, которые подробно описаны в статье [14]:

- стоимость проекта (Financing)
- экологичность проекта (Ecology)
- инновационность проекта (Innovativeness)
- технологичность проекта (Manufacturability)
- безопасность проекта (Norms)
- рейтинг проекта (Rating)

Задачей данного исследования является выбор оптимального алгоритма для прогнозирования зависимой переменной рейтинга проекта от набора объясняющих переменных. Количество наблюдений (проектов), используемых для тренировочного и тестового набора данных, N = 283. Для построения модели набор был разделен в соотношении 60:40.

В первую очередь, требуется подгрузить данные и вывести обобщенную информацию об объекте. Следует заметить, что переменные Norms, Manufacturability, Ecology, Innovativeness, Rating являются категориальными переменными, значения которых расположены в диапазоне [0, 3].

На результатах вывода программы, представленных в табл. 1, можно заметить, что переменная Financing имеет наибольшее среднее значение (Mean SD) при Rating = 3, что может означать, что наиболее дорогостоящий проект с большей вероятностью имел наивысший рейтинг. Также, следует отметить, что значение переменной Norms = 3, то есть проекты с наибольшим значением безопасности были приняты с наивысшим рейтингом. Проекты с переменной Ecology = 0 имели наименьший процент принятия, наряду с этим, проектов со значением Innovativeness = 0 большинство (83.4%), что показывает, что количество инновационных проектов имело небольшое количество, и лишь 15.2% из них были приняты с наивысшим рейтингом. И, при значении переменной Manufacturability = 1, то есть проекты, которые улучшают технологию, стоит наивысший рейтинг принятия проекта.

Таблица 1

Данные о количестве проектов по переменной Rating в зависимости от значения объясняющих переменных

	Rating			
	1 (N=77)	2 (N=81)	3 (N=125)	Overall (N=283)
Norms				
1	57 (74.0%)	74 (91.4%)	70 (56.0%)	201 (71.0%)
2	20 (26.0%)	7 (8.6%)	0 (0%)	27 (9.5%)
3	0 (0%)	0 (0%)	55 (44.0%)	55 (19.4%)
Manufacturability				
0	77 (100%)	81 (100%)	100 (80.0%)	258 (91.2%)
1	0 (0%)	0 (0%)	25 (20.0%)	25 (8.8%)
Ecology				
0	9 (11.7%)	7 (8.6%)	2 (1.6%)	18 (6.4%)
1	68 (88.3%)	74 (91.4%)	123 (98.4%)	265 (93.6%)
Innovativeness				
0	57 (74.0%)	73 (90.1%)	106 (84.8%)	236 (83.4%)
1	20 (26.0%)	8 (9.9%)	19 (15.2%)	47 (16.6%)
Financing				
Mean (SD)	1880 (4540)	2520 (6520)	11400 (39800)	6280 (27100)
Median [Min, Max]	320 [28.0, 27000]	320 [33.9, 34800]	653 [9.00, 314000]	493 [9.00, 314000]

Визуализируем данные для выявления зависимостей с помощью графического метода. Для начала рассмотрим распределение переменной Financing, однако для лучшего понимания возьмем логарифм от данной переменной.

Рис. 6 отображает информацию о распределении логарифма стоимости проектов, данное распределение имеет большой разброс, как по количеству, так и по логарифму стоимости. Построим рис. 7, который описывает распределение логарифма стоимости проекта по переменной Rating. Для построения данного графика создадим новую базу данных FRating, в которой определим тип переменной Rating как дискретный фактор. Центр распределения графика со значением Rating = 3 находится правее, это означает, что с увеличением стоимости проекта вероятность принятия более высокого рейтинга возрастает, как было предположено ранее.

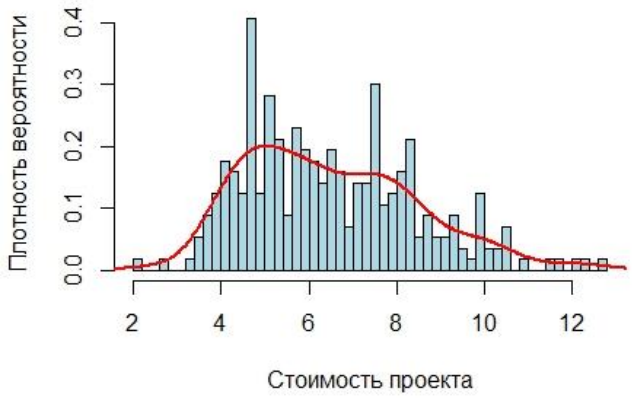

Рис. 6. Гистограмма логарифма стоимости проекта, совмещенная с кривой плотности распределения

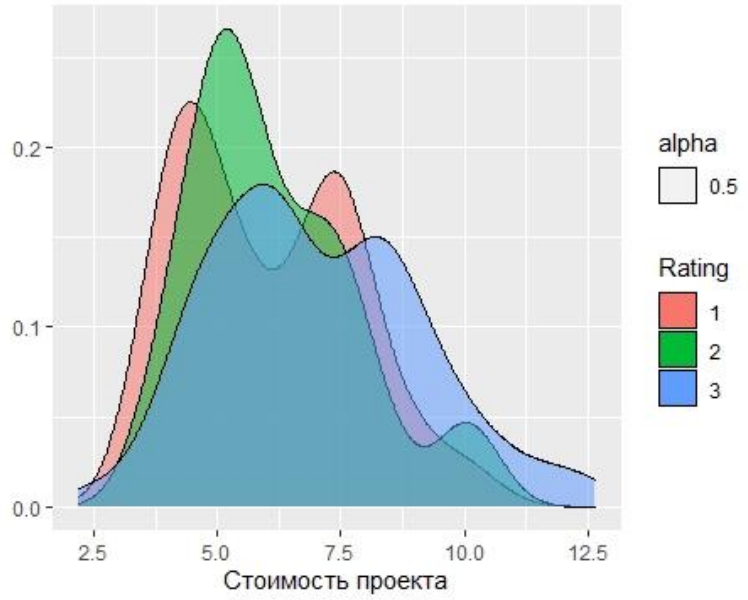

Рис. 7. График распределения логарифма стоимости проекта по переменной Rating

Ниже представлен рис. 8 распределения логарифма стоимости проекта по переменным Rating, Norms, Ecology, где сверху находится разделение графиков по переменной Ecology, а справа по переменной Norms. На графике видно, что все проекты с наивысшей оценкой безопасности проекта являются экологичными и имеют наивысший рейтинг принятия. Также видно, что площадь подфункции графика со значением Rating = 3 при значении переменной Norms = 1 больше у экологичных проектов.

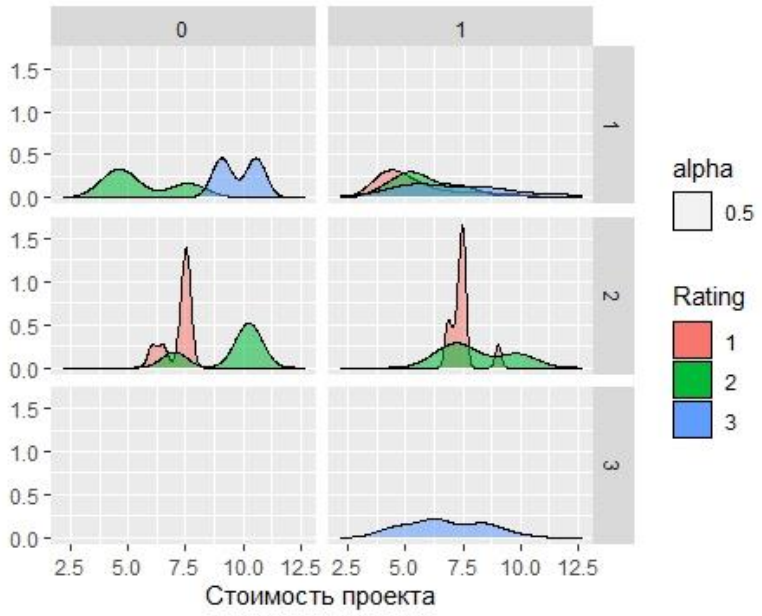

Рис. 8. График распределения логарифма стоимости проекта по переменным Rating, Norms, Ecology

Рассмотрим далее рис. 9 распределения логарифма стоимости проекта по переменным Rating, Ecology, Innovativeness, где сверху отображается деление по переменной Innovativeness, а справа по переменной Ecology. Можно заметить, что все инновационные проекты являются экологичными. Также следует отметить, что площади подфункции графиков со значениями переменных Ecology = 1 и Innovativeness = 1 при значениях Rating = 1 и Rating = 3 практически совпадают, что отображает предположение о том, что данные переменные не оказывают сильного влияния на значение рейтинга проекта.

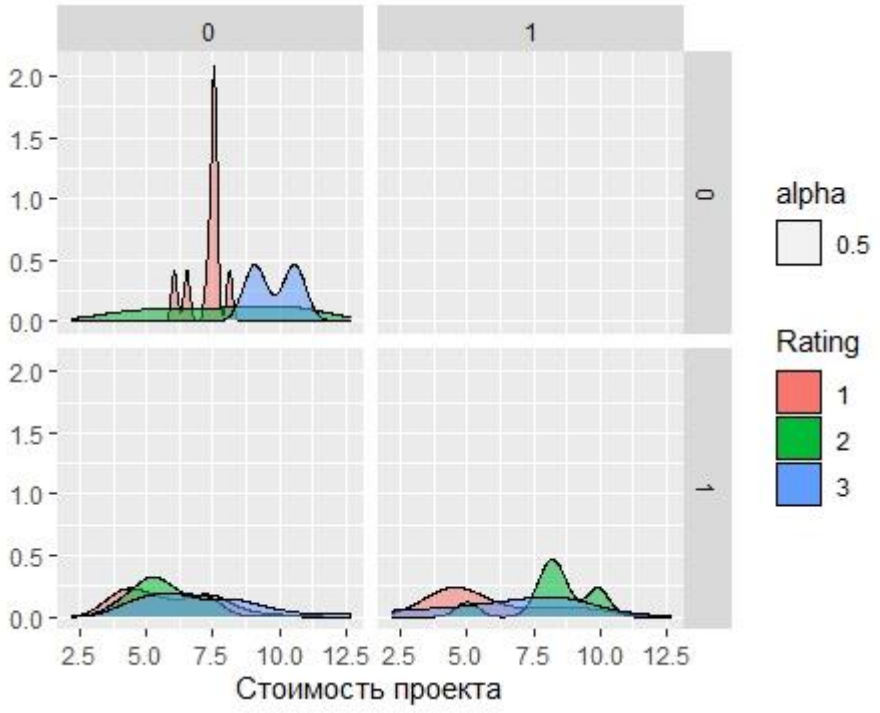

Рис. 9. График распределения логарифма стоимости проекта по переменным Rating, Ecology, Innovativeness

При предположении влияния переменных Norms и Manufacturability на значение рейтинга построим рис. 10 распределения логарифма стоимости проектов по данным переменным, где сверху расположено разделение по переменной Manufacturability, а справа по переменной Norms. Действительно, на графике видно, что все проекты с наивысшим значением безопасности проекта имеют наибольшую оценку рейтинга, что доказывает предположение о том, что предпочтение было отдано именно степени соответствия нормам и правилам проекта. Данная переменная имеет наибольшее влияние на составление рейтинга проекта.

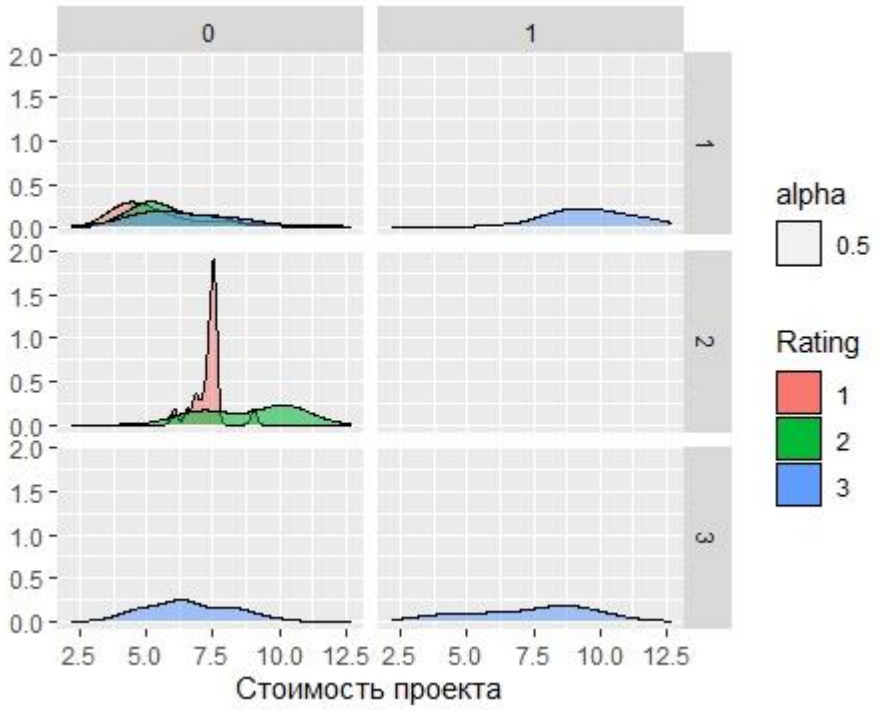

Рис. 10. График распределения логарифма стоимости проекта по переменным Rating, Norms, Manufacturability

Для визуализации большого количества качественных переменных используем функцию mosaic, которая представляет собой мозаичный график. Построим данный рис. 11 по переменным Rating, Norms и Manufacturability, для установления связей между значениями переменных. Поделим сначала все проекты по переменной Norms, видно, что проектов с наименьшим значением данной переменной больше, также можно заметить, пропорции при переменной Norms = 1 примерно постоянные, но, при этом значение Manufacturability не оказывает значительного влияния. С учетом предположений, сделанных ранее, можно сказать, что на значение рейтинга проекта оказывает большое влияние его стоимость. Площадь на данном графике пропорциональна соответственно количеству наблюдений, которые попадают в эту категорию. Распределение цветов происходит по автоматической проверке компьютером гипотезы о независимости признаков и, соответственно, наличие различных цветов показывает наличие взаимосвязи между переменными.

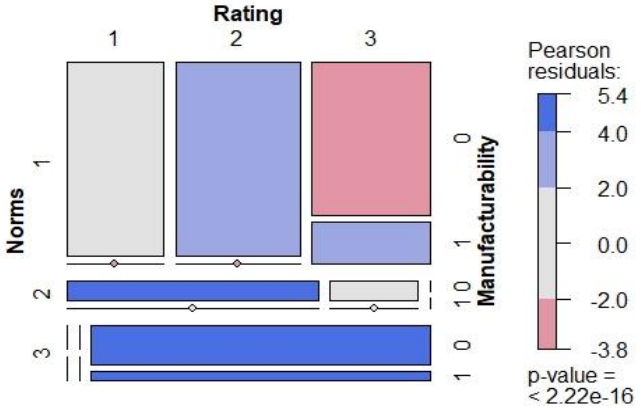

Рис. 11. Мозаичный график по переменным Rating, Norms, Manufacturability

После изучения выборочных наблюдений, было выявлено значительное влияние объясняющих переменных Norms, Manufacturability, Financing на переменную Rating, модель линейной регрессии, подробно описанная в статье [14], построенная на основании значимых объясняющих переменных обеспечила результативность прогнозирования на уровне 68%, построение модели на основе алгоритма случайного леса повысило результат прогнозирования до 71.7%.

Построение модели

Для построения модели на основе алгоритма случайного леса необходимо разделить набор данных на тренировочный и тестовый, в работе данные были разделены в соотношении 60:40. Далее была построена модель на основе всего набора данных на тренировочных данных, проведено прогнозирование на тестовых данных и выведен рис. 12 прогнозов:

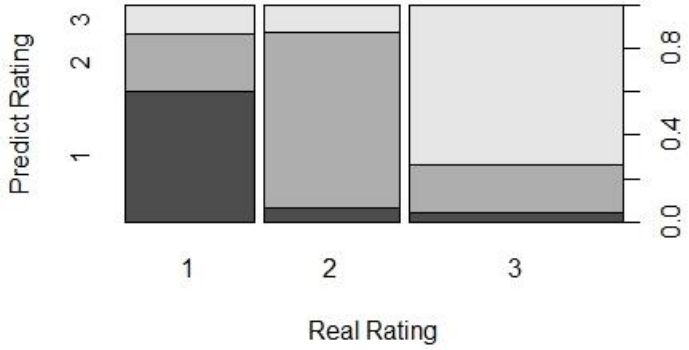

Рис. 12. Результаты прогнозирования алгоритма

На рис.12 можно заметить, что большая часть из предсказанных значений совпадает с изначальными данными, точность прогноза алгоритма случайного леса на полном наборе факторов составила 71.7%, что на 3.7% больше модели линейной регрессии, построенной на основе значимых регрессоров, а именно: Norms, Manufacturability, Financing.

Заключение

Финансовая индустрия подвержена различным рискам, особенно при инвестировании. Технологии искусственного интеллекта могут помочь принять обоснованное решение об инвестициях и предсказать возможные риски, используя алгоритмы анализа данных, технологию глубокого обучения и машинного обучения. Некоторые из них существуют как аналитические платформы, которые применяют анализ данных или другие решения. Например, корейское приложение KOSHO [18], для частных лиц анализирует три ключевых рыночных фактора (индекс рыночной волатильности, индекс рынка ценных бумаг и инфляция) с использованием глубокого обучения и исторических данных финансового рынка.

Следует отметить, что в России также активно развивается сфера искусственного интеллекта, было открыто множество новых программ в университетах, онлайн-курсов и стажировок, направленных на развитие данного направления во всех регионах страны, так, Сбербанк запустил бесплатную образовательную программу «Машинное обучение в финансах» в 2019 году.

В заключении, необходимо отметить, что обоснованность выбора инвестиционного проекта зависит от качества и количества критериев, характеризующих проект, задач машинного обучения и выбранного алгоритма, который должен обеспечить наиболее точный прогноз. На данный момент исследования алгоритм случайного леса, используемый для оценки инвестиционной привлекательности проектов с рассмотренным набором критериев, является наиболее оптимальным алгоритмом машинного обучения с учителем.

Литература

1. *Samuel A.L.* Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 1959. P. 206–226.
2. *Sandhya N. dhage, Charanjeet Kaur Raina.* A review on Machine Learning Techniques. In International Journal on Recent and Innovation Trends in Computing and Communication, Volume 4 Issue 3. 2016.
3. *Alloghani, Mohamed & Al-Jumeily, Dhiya & Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed.* A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. 2020. P. 3–23.
4. Alpaydın, E. Introduction to machine learning. Cambridge, MA: MIT Press. 2014.
5. *Huang, G., Song, S., Gupta, J. N. D., & Wu, C.* Semi-supervised and unsupervised extreme learning machines. IEEE Transactions on Cybernetics, 44(12). 2014. P. 2405–2417.
6. *Liu, N., & Zhao, J.* Semi-supervised online multiple kernel learning algorithm for big data. TELKOMNIKA, 14(2). 2016. P. 638–646.
7. *Nasteski, Vladimir.* An overview of the supervised machine learning methods. HORIZONS.B. Volume 4. 2017. P. 51-62.
8. *Daniel Jurafsky & James H. Martin.* Speech and Language Processing. 2016.
9. *Tom Mitchell, McGraw Hill.* Machine Learning. 2015.
10. *Gerard Biau.* Analysis of a Random Forests Model. Journal of Machine Learning Research 13. 2013. P. 1063-1095.
11. *Yu, Chong Ho.* Exploratory data analysis in the context of data mining and resampling. International Journal of Psychological Research. 3. 2010.
12. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс. 2004. Стр. 322-324.
13. *Воронцов К.В.* Лекция «Линейные методы классификации: метод опорных векторов», курс «Машинное обучение». 2014.
14. *Казанская А.А., Мишура Л.Г.* «Использование линейной регрессии при выборе инвестиционного проекта». Альманах научных работ молодых ученых Университета ИТМО. 2020.
15. Исследование российского рынка венчурных инвестиций на 2019 год, [Электронный ресурс]. – Режим доступа: https://drive.google.com/file/d/1BpXNU_3f8_zEeL8tryyvOpv3AOF5UYY/view
16. Статья «Data Scientist» в Harvard Business Review, [Электронный ресурс]. – Режим доступа: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
17. Машинное обучение с Python – Краткое руководство. 2018. [Электронный ресурс]. – Режим доступа: https://goo.su/0z3q
18. Корейское приложение KOSHO. [Электронный ресурс]. – Режим доступа: https://www.kosho.ai

Reference

1. Samuel A.L. Some Studies in Machine Learning Using the Game of Checkers // *IBM Journal of Research and Development*. 1959. P. 206–226.
2. Sandhya N. dhage, Charanjeet Kaur Raina. A review on Machine Learning Techniques // *In International Journal on Recent and Innovation Trends in Computing and Communication*. Volume 4 Issue 3. 2016.
3. Alloghani, Mohamed & Al-Jumeily, Dhiya & Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. 2020. P. 3–23.
4. Alpaydın, E. Introduction to machine learning. Cambridge, MA: MIT Press. 2014.
5. Huang, G., Song, S., Gupta, J. N. D., & Wu, C. Semi-supervised and unsupervised extreme learning machines // *IEEE Transactions on Cybernetics*. 44(12). 2014. P. 2405–2417.
6. Liu, N., & Zhao, J. Semi-supervised online multiple kernel learning algorithm for big data // *TELKOMNIKA*. 14(2). 2016. P. 638–646.
7. Nasteski, Vladimir. An overview of the supervised machine learning methods // *HORIZONS.B*. Volume 4. 2017. P. 51-62.
8. Daniel Jurafsky & James H. Martin. Speech and Language Processing. 2016.
9. Tom Mitchell, McGraw Hill. Machine Learning. 2015.
10. Gerard Biau. Analysis of a Random Forests Model. Journal of Machine Learning Research 13. 2013. P. 1063-1095.
11. Yu, Chong Ho. Exploratory data analysis in the context of data mining and resampling.// *International Journal of Psychological Research*. 3. 2010.
12. Magnus Ya.R.. Katyshev P.K.. Peresetskiy A.A. Ekonometrika. Nachalnyy kurs. 2004. Str. 322-324.
13. Vorontsov K.V. Lektsiya «Lineynyye metody klassifikatsii: metod opornykh vektorov». kurs «Mashinnoye obucheniye». 2014.
14. Kazanskaya A.A.. Mishura L.G. «Ispolzovaniye lineynoy regressii pri vybore investitsionnogo proyekta». Almanakh nauchnykh rabot molodykh uchenykh Universiteta ITMO. 2020.
15. Issledovaniye rossiyskogo rynka venchurnykh investitsiy na 2019 god. [Elektronnyy resurs]. – Rezhim dostupa: https://drive.google.com/file/d/1BpXNU_3f8_zEeL8tryyvOpv3AOF5UYY/view
16. Statia «Data Scientist» v Harvard Business Review. [Elektronnyy resurs]. – Rezhim dostupa: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
17. Mashinnoye obucheniye s Python – Kratkoye rukovodstvo. 2018. [Elektronnyy resurs]. – Rezhim dostupa: https://goo.su/0z3q
18. Koreyskoye prilozheniye KOSHO. [Elektronnyy resurs]. – Rezhim dostupa: https://www.kosho.ai

Статья поступила в редакцию 12.02.2020 г