ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет «Санкт-Петербургская школа экономики и менеджмента»

Подразделение (департамент, лаборатория, кафедра) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Применение "Random Forest Algorithm" в задачах корпоративного управления**

КУРСОВАЯ РАБОТА

по направлению подготовки 38.03.01 «Экономика»

Образовательная Программа «Экономика»

Санкт-Петербург

2019 г.

Работа сдана в ОСУП

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

Защита состоялась

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_ г.

Состав комиссии и подписи:

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Оценка за работу \_\_\_\_\_\_\_\_\_ (\_\_\_)**

оценка по пяти и (десяти) бальной

системе

Студента (-ки) группы № \_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия И.О., подпись)

Руководитель

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(должность, степень, Фамилия

И.О., подпись)

Содержание

[Введение 3](#_Toc10055910)

[Глава 1. Организационные вопросы управления рисками 5](#_Toc10055911)

[Принципы управления рисками на базе концепции ALARP 5](#_Toc10055912)

[Стратегия управления риском с позиции концепции ALHARP 8](#_Toc10055913)

[Роль внутреннего контроля при анализе ERM 9](#_Toc10055914)

[Глава 2. Построение Случайного леса. 15](#_Toc10055915)

[Метод случайных подпространств 15](#_Toc10055916)

[Методы выбора и агрегирования 15](#_Toc10055917)

[Бутстрэп 15](#_Toc10055918)

[Бэггинг 16](#_Toc10055919)

[Алгоритм случайного леса 17](#_Toc10055920)

[Реализация 18](#_Toc10055921)

[Заключение 26](#_Toc10055922)

[Литература 27](#_Toc10055923)

[Приложения 29](#_Toc10055924)

[Приложение 1 29](#_Toc10055925)

[Приложение 2. 31](#_Toc10055926)

# Введение

Современная экономика характеризуется сильной волатильностью денежных потоков, связанных с оценкой активов и будущих прибылей.

В большинстве экономических взаимоотношений лежит ожидание того, какой доход эти взаимоотношения принесут, и какая прибыль в дальнейшем будет выделена на развитие. При этом, пропорции распределения прибыли в большинстве своём связанны с принятием грамотных управленческих решений и принятых на себя ранее обязательств.

Фирма, как правило, при ведении бизнеса, составляет производственные и финансовые планы на несколько лет вперед, однако при их реализации, ожидаемый результат зачастую бывает не достигнут.

Волатильность будущих денежных потоков связана с рядом эндогенных и экзогенных факторов, которые провоцируют не достижение таких результатов. В качестве агента, способствующего достижению, компания выбирает и использует разного вида консалтинговые услуги, автоматические системы управления, системы планирования, при этом затраты, связанные с привлечением агентов, должны быть как минимум соизмеримы с получаемой добавленной стоимостью.

Превентивные меры по предотвращению получения ожидаемых доходов (как спекулятивного характера, так и в области чистых рисков) заложены в современные системы корпоративного управления рискам фирмы и базируются на целом ряде ключевых правил, основными из которых являются: принцип ALARP, принцип AHARP, правильная организация управления, ориентация на акционерную стоимость и интеграция с внутрифирменным контролем. В настоящей главе раскрыто основное содержание перечисленных принципов.

Целью настоящей работы является изучение возможности применения случайных деревьев при оценке риск-менеджмента. Для этого потребовалось решить следующие задачи:

1. Изучили теоретические и практические основы корпоративного риск-менеджмента.
2. Провести оценку эффективности корпоративного риск-менеджмента методом линейной регрессии (ANOVA) и классификационное дерево (CHAID).
3. Сравнить результаты полученные обычными эконометрическим методом и с помощью метода случайного леса.

# Глава 1. Организационные вопросы управления рисками

# Принципы управления рисками на базе концепции ALARP

Ключевой концепцией в управлении рисками является принцип ALARP (снижать настолько, насколько это является целесообразным). По сути, этот принцип заключается в том, что риски должны быть снижены до такого уровня, при котором расходы на мероприятия по снижению риска уравновешивают пользу (рис. 1.).

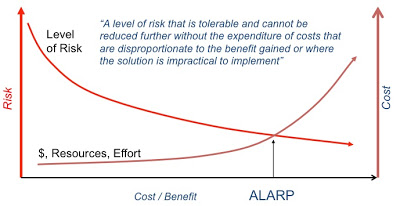
[](http://3.bp.blogspot.com/-EAfVfZJSWfY/TsNb_bU0yhI/AAAAAAAAAVo/BGjZ8z30HF8/s1600/ALARP+1.jpg)

Рис. 1. Принцип управления риском

На практике соблюдать этот принцип намного труднее:

* 1. нужна база данных за период как минимум два года, предшествующих мероприятию по управлению рисками, которая и будет базой для проведения сравнительного анализа эффективности процессов управления риском.
  2. большинство затрат, направленных на управление рисками, относится к неявным затратам. И исчислить экономическую эффективность процесса в таком случае можно с большой долей условности.
  3. Не все мероприятия по управлению рисками дают положительный эффект сразу же после внедрения: некоторые процессы не окупаются в течение первых двух-трех лет, что ставит под сомнение возможность незамедлительно отконтролировать целесообразность внедряемых мероприятий.

Определение области ALARP- отправная точка в построении любой системы показателей.

Основными вопросами при разработке такой системы, отвечающей требованиям бенчмаркинга рисков является:

*Область применения* - Какова сфера применения разрабатываемой системы показателей? Система может быть разработана для всей организации или для подразделения, все ли программы по управлению рисками учитываются или только часть из них.

*Показатели* – какие конкретно показатели нужно включать: основанные на ретроспективных или плановых данных, достаточно ли информации и какую еще информацию нужно собрать?

*Система измерений* – каким образом измеряется эффективность.

*Длительность* – какой период будет подвержен исследованию? Нужно ли использовать ретроспективные данные для проведения анализа?

*Производительность* – какие именно показатели производительности организации изучаются применительно к бенчмаркингу. Каким образом риск менеджмент повлиял на них?

*Качество* - Что будет служить сигналом того, что риск менеджмент выстроен качественно?

Следующим принципом, не менее важным в ERM (Enterprise Risk Managment) является принцип AHARP (As High As Reasonably Practicable – завышать настолько, насколько это является целесообразным) относится к позитивным рискам, в отличие от принципа ALARP.

|  |
| --- |
| [Описание: http://4.bp.blogspot.com/-vMzvtecZw3c/TpgxuSEKbxI/AAAAAAAAAQM/MQgjCXWlvYw/s400/Slide01.jpg](http://4.bp.blogspot.com/-vMzvtecZw3c/TpgxuSEKbxI/AAAAAAAAAQM/MQgjCXWlvYw/s1600/Slide01.jpg) |
| Рисунок 2. AHLARP модель |

Совокупное влияние двух концепций мы видим на рис. 2. Точка RTP (risk tipping point) означает переломный момент в управлении риском, в которой положительный эффект от риска начинает перевешивать отрицательный эффект. Эта точка и означает экономическую эффективность процесса управления риском. Именно с этого момента нужно начинать осуществлять какие-то затраты по управлению риском, если мы стремимся превратить корпоративный риск менеджмент в инструмент по созданию стоимости, а не в механизм покрытия убытков. Иными словами если риск отрицательный, то дальнейшее вливание средств в его управление не принесет пользы, если риск положительный, то экономическая эффективность риска будет предельно положительной до того момента, когда темп роста затрат на управление им не превысит темп роста самого риска.

Разница лишь в том, что при отрицательном риске невозможен хоть какой-либо положительный эффект в случае его реализации и при управлении такими рисками оптимальным решением при управлении им будет стратегия минимизации убытков, положительный риск – это непредусмотренные в текущей системе планирования компанией скрытые возможности, которые при возникновении соответствующей ситуации способны создать добавленную стоимость компании.[20]

# Стратегия управления риском с позиции концепции ALHARP

Стратегия управления риском при помощи модели ALHARP сводится к уменьшению влияния неопределенности в отношении того или иного события.

Рис. 3 и 4. отчетливо демонстрирует при помощи концепции ALHARP как направить влияние вероятных результатов того или иного события к позитивному исходу.

|  |
| --- |
| [Описание: http://3.bp.blogspot.com/-pXrdLAA5vYQ/Tpgx16sBknI/AAAAAAAAAQ0/gRUZNocHM64/s400/Slide06.jpg](http://3.bp.blogspot.com/-pXrdLAA5vYQ/Tpgx16sBknI/AAAAAAAAAQ0/gRUZNocHM64/s1600/Slide06.jpg) |
| Рис. 3. Недостаточные ресурсы стимулируют рост вероятности негативных последствий |

|  |
| --- |
| [Описание: http://4.bp.blogspot.com/-X1SAU_wo4FY/Tpgx3aDZS1I/AAAAAAAAAQ8/orvFH5dcs0o/s400/Slide07.jpg](http://4.bp.blogspot.com/-X1SAU_wo4FY/Tpgx3aDZS1I/AAAAAAAAAQ8/orvFH5dcs0o/s1600/Slide07.jpg) |
| Рис 4. Своевременное и достаточное применение ресурсов риск-менеджмента |
|  |

Необходимо отметить, что размер затрат (инвестиций) в управление риском не определяется общей формулой экономической эффективности и в каждом конкретном случае нужно определять индивидуально, однако общий принцип управления рисками заключается в следующем: риски повышенной волатильности требуют большего вовлечения ресурсов, в то время как к простым типовым рискам применяются стандартные методы управления.

# Роль внутреннего контроля при анализе ERM

Внутренний контроль и его организация являются неотъемлемым атрибутом системы ERM - практически все рекомендации в области риск-менеджмента не обходят вниманием организацию внутреннего контроля. На практике, компаниями, соблюдающими требования корпоративных стандартов ведения бизнеса наиболее распространена и рекомендована трехуровневая модель организации управления, в которой органично объединены и функции риск-менеджмента и функции внутреннего контроля: 3LOD (3 level of defense) model.

В данной модели риск менеджмент и комплаенс входят во вторую ступень защиты, реализуемой на фоне безупречно поставленного операционного менеджмента, в котором бизнес-подразделения отвечают за генерируемые в процессе их же деятельности риски, при этом, служба риск менеджмента обеспечивает подразделения необходимым инструментарием для управления принимаемыми рисками( структура лимитов, ключевые индикаторы риска, т.д. ). [21]

Таблица 1. Организационная модель «Three lines of defense»

|  |  |  |
| --- | --- | --- |
| **Уровни управления** | **Какими отделами представлены** | **Функции** |
| 1 | бизнес подразделения | формирование правильного понимания у бизнес подразделений всех рисков, присущих их непосредственной деятельности |
| 2 | интегрированная система активного риск-менеджмента. | формирование соответствующих инструментов по управлению рисками для первой линии защиты, а также внутренних нормативных документов по управлению рисками.  Регламентация взаимодействия структурных подразделений при реализации различных бизнес-процессов |
| 3 | Внутренний аудит | Анализ исполнения структурными подразделениями своих обязанностей, выявление нарушений, зафиксированных в процессе деятельности |

Основной целью при применении этой модели является необходимость избегать дублирования функции риск менеджмента и внутреннего контроля, а так же необходимость избегать конфликта интересов между менеджментом и стейкхолдерами. Анализ теории и практики организации интегрированных систем управления на корпоративном уровне позволил выделить следующие спорные моменты:

1. Внедрение полноценного внутреннего контроля и риск-менеджмента достаточно затратно.[14]
2. Потенциальные стейкхолдеры оценивают компанию по результатам, отраженным в финансовой отчетности компании [15], которая зачастую подвержена различным манипуляциям [16].

Мотивы, которыми могут руководствоваться представители компании, решившейся на искажение отчетности, могут быть разными. Среди основных причин, толкающих на применение нелегальные способов воздействия на финансовый результат, можно назвать:

* Желание повлиять на решения стейкхолдеров, например, может привести к завышению или занижению результатов прибыли, в зависимости от того, какие цели стоят перед компанией в данный конкретный момент.
* Конфликт между интересами стейкхолдеров и компании в целом. Акционеры, особенно миноритарные, чаще всего заинтересованы в получении максимально возможного размера дивидендов, в то время как компания стремится к конкурентному преимуществу, финансовой независимости, росту кредитоспособности и платежеспособности, увеличению доли рынка и т.д. Для достижения целей компании необходимо реинвестирование прибыли, что может вызвать недовольство акционеров, поэтому руководство компании может занизить финансовые результаты, скрывая часть прибыли в резервах и других статьях отчетности.
* Стремление «приукрасить» финансовые результаты для привлечения инвесторов, например, накануне дополнительной эмиссии акций или в преддверии крупного проекта, требующего инвестиций.
* Попытка занизить прибыть для уменьшения налоговых выплат (стремление представить минимально возможное значение финансового результата во имя снижения налогового бремени). [13]
* Получение государственных субсидий [8]
* Выполнение требований контрактных обязательств, например, условий кредитных договоров, где часто прописан минимальный размер активов компании.[12]
* Желание выставить компанию в наиболее выгодном свете в преддверии сделок слияния и поглощения (может объяснить как завышение, так и занижение тех или иных показателей).
* Обоснование стоимости предоставляемых услуг / предлагаемых товаров компаниями-монополистами. В условиях монополии высокие показатели прибыли однозначно свидетельствуют о том, что компания получает сверхприбыль, что не может не вызвать интереса антимонопольной службы. Во избежание привлечения интереса государства и общественных организаций компания может стремиться искусственно занижать прибыль.[7]
* Стремление менеджеров получить планируемое вознаграждение, которое часто базируется именно на достижении компанией определённых показателей и т.д.

Вне зависимости от мотивов, обуславливающих такие действия, искажение отчетности является недопустимым, поскольку чтение финансовой отчетности, расчёт экономических показателей на её основе, оценка результатов деятельности компании, а также построение прогнозов на будущее теряет смысл, если данные, представленные в отчетности, не соответствуют реальности.

Все перечисленное мероприятия, связанные с стремлением менеджеров максимизировать свою прибыль никак не отражаются в финансовой отчетности, однако при отсутствии должных стимулов, ориентирующих менеджмент на эффективное управление, фирма начинает разрушать свою стоимость.

В первую очередь это реализуется через ориентацию на краткосрочные цели в ущерб долгосрочным. Поведенческая теория максимизации прибыли гласит, что при возникновении агентского конфликта для менеджмента равнозначно, что максимизировать – прибыль или управленческие расходы, поэтому можно считать, что неэффективное управление будет отражено ростом краткосрочных долгов, при непосредственном влиянии операционной прибыли и управленческих расходов.

Таким образом, в качестве объясняющих переменных нами было выбрано:

* + - 1. Объем продаж, как свидетельство масштаба производства.
      2. Операционная прибыль (EBIT) – прибыль от основного вида деятельности.
      3. Бинарная переменная характеризующая инвестирование в нематериальные активы.
      4. Административные расходы.
      5. Долгосрочные обязательства
      6. Рентабельность основных средств.

В качестве объясняемой переменной взяты краткосрочные обязательства.

Применение статистических методов постройки деревьев не дало нужного результата – редко случающиеся признаки долгосрочной заинтересованности менеджмента не достаточно значимы для постройки модели (рис5). (Приложение 1).



Рисунок 5. Классификационное дерево, построенное методом CHAID в СПСС

Поэтому мы прибегли к Алгоритму случайного леса и реализовали его на языке программирования Python.

# Глава 2. Построение Случайного леса.

# Метод случайных подпространств

В методе случайных подпространств (Random subspace method) базовые алгоритмы обучаются на различных подмножествах признакового описания[[1]](#footnote-1), которые также выделяются случайным образом. Данный метод эффективен в задачах с большим числом признаков и относительно небольшим числом объектов, а также при наличии избыточных неинформативных признаков. В этих случаях алгоритмы, построенные по части признакового описания, могут обладать лучшей обобщающей способностью по сравнению с алгоритмами, построенными по всем признакам. RSM позволяет снизить коррелированность между деревьями и избежать переобучения.[17]

Известным алгоритмом, который использует одновременно RSM и бэггинг, является Random Forest. Разбиение объектов в вершине случайного леса ищется среди случайного подмножества признаков, а обучение каждого дерева в композиции происходит на выборке, полученной с помощью операции бутстрэпа.[9]

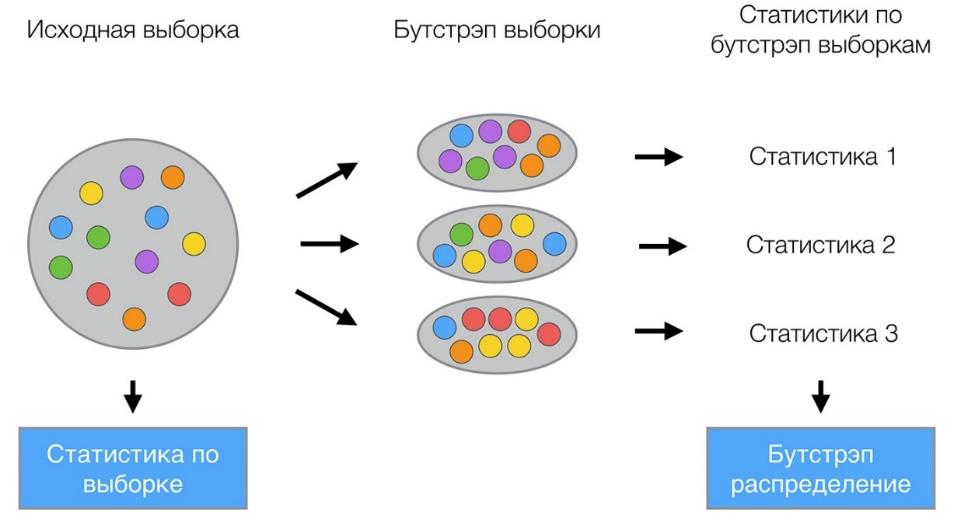
# Методы выбора и агрегирования

# Бутстрэп

Данный метод подробно описан по <https://habr.com/en/company/ods/blog/324402/#butstrep>. Схематично метод изображен на рисунке 6.

Статистика (для выборки) - измеримая числовая функция от выборки, не зависящая от неизвестных параметров распределения элементов выборки.

Рис. 6 Схематичное изображение бутстрэпа



# 

# Бэггинг

Метод бэггинга предложил Л. Брейман в 1996 году. Из исходной обучающей выборки длины формируются различные обучающие подвыборки той же длины с помощью метода бутстрэпа, описанного ранее. При этом некоторые объекты попадают в под-выборку по нескольку раз, некоторые — ни разу. Базовые алгоритмы, обученные по под-выборкам, объединяются в композицию с помощью простого голосования.[17] Эффективность бэггинга можно объяснить двумя обстоятельствами:

1. Благодаря различности базовых алгоритмов, их ошибки взаимно компенсируются при голосовании. [17]

2. Объекты-выбросы могут не попадать в некоторые обучающие под-выборки. [17]

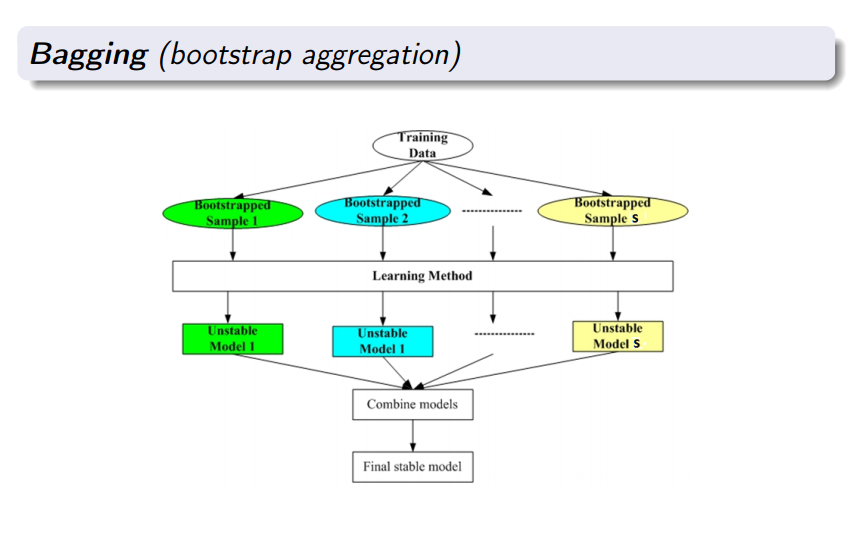


Рисунок 7 схематичное изображение бэггинга.

# Алгоритм случайного леса

**Алгоритм случайного леса** — алгоритм машинного обучения, сочетающий в себе метод бэггинга (комбинация моделей обучения увеличивает общий результат) и метод случайных подпространств. [11]

Алгоритм применяется для задач классификации, регрессии и кластеризации.

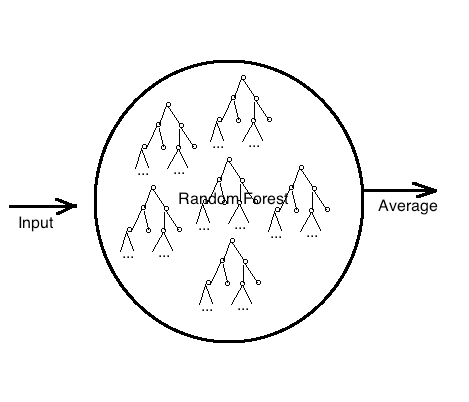


Рисунок 10. Схема случайного леса.

# Реализация

Реализация алгоритма случайного леса будет проведена на языке программирования Python, в среде разработки Jupyter, так как в данном языке уже присутствуют все необходимые модули для упрощения нашей задачи (pandas, skitlearn и т.д.). Главный для нас skitlearn, так как он отвечает за машинное обучение и, в нашем случае, Random forest(from sklearn.ensemble import RandomForestRegressor).

В качестве данных мы берем выборку компаний металлургического сектора в период с 2013 по 2017 годы, количество наблюдений 1280.

import pandas as pd – подключаем модуль pandas (простой инструмент для обработки и Анализа данных в Python).

features = pd.read\_csv(fa.csv') – считываем наш файл с данными

features.head(5) – выводим первые 5 строк (рис. 11)

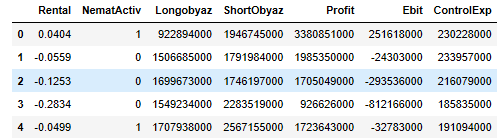


Рисунок 11 Первые 5 строк наших данных.

print('The shape of our features is:', features.shape) – кол-во строк и столбцов

features = pd.get\_dummies(features) – избавляемся от дамми-переменны (если они есть), так как Python с ними плохо работает.

Необходимо разделить данные на функции и цели. Целями называются прогнозируемые значения, а функциями – все столбцы, используемые моделью для прогнозирования.

import numpy as np – подключаем модуль NumPy (данный модуль добавляет поддержку больших многомерных массивов и матриц, а также большую библиотеку высокоуровневых и очень быстрых математических функций для операций с этими массивами)

labels = np.array(features[‘ShortObyaz’]) – записываем данные которые мы хотим прогнозировать

features= features.drop(‘ShortObyaz’, axis = 1) – удаляем цели из наших данных, оставляя только функции

feature\_list = list(features.columns) – сохраняем названия функций

features = np.array(features) – конвертируем в массив формата NumPy

Разделяем данные на обучающие и тестовые наборы. Во время обучения мы позволяем модели сопоставлять ответы, в нашем случае – короткосрочные долговые обзательства, чтобы она могла узнать, как предсказать их на основе функций. Предполагаем, что имеется взаимосвязь между всеми функциями и целевыми значениями, и задача нашей модели - изучить эти соотношения в процессе обучения. Далее, когда наступает момент оценки модели, мы просим ее сделать прогноз на тестовом наборе данных, в этом случае предоставлен доступ только к функциям (к ответам доступа нет). Так как есть ответы для набора тестов, можно сравнить полученные прогнозы с истинными значениями, чтобы оценить точность модели. Как правило, при обучении модели мы произвольно разбиваем данные на обучающие и тестовые наборы, чтобы получить представление всех точек данных. Для отладки кода в качестве random\_state будет установлено значение 42 (разбиение всегда будет одинаковым), если нам это не нужно то просто убираем этот параметр. Мы будем тестировать 75% процентов выборки на 25% выборки. Для этого разобьем данные на тестовые и тренировочные.

from sklearn.model\_selection import train\_test\_split подключаем библиотеку для разбиения данных на тестовые и тренировочные

train\_features, test\_features, train\_labels, test\_labels = train\_test\_split(features, labels, test\_size = 0.25, random\_state = 42)

Желательно посмотреть на форму всех данных, чтобы убедиться, что все сделано правильно. Ожидается, что количество функций обучения будет соответствовать количеству столбцов и количеству строк для соответствующих функций обучения, тестирования и меток:

print('Training Features Shape:', train\_features.shape)

print('Training Labels Shape:', train\_labels.shape)

print('Testing Features Shape:', test\_features.shape)

print('Testing Labels Shape:', test\_labels.shape)

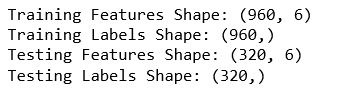


Рисунок 12. Форма данных.

Подготовя данные, нам теперь необходимо построить наш Random Forest.

from sklearn.ensemble import RandomForestRegressor – включаем модель Random Forest в нашу программу.

rf = RandomForestRegressor(n\_estimators = 10, random\_state = 42, max\_depth = 5) – создаем лес на 10 деревьев (чем больше тем, точнее прогноз, но нам необходимо считать в ручную), глубиной 5.

rf.fit(train\_features, train\_labels); – тренируем модель.

Теперь необходимо выяснить насколько наша модель точна (не зря мы всё это затеяли?). Мы выясним как далёк наш средний прогноз от фактического значения.(абсолютную ошибку считаем)

predictions = rf.predict(test\_features) – предсказанные значения на тестовой выборке

errors = abs(predictions - test\_labels) – считаем отклонения

print('Mean Absolute Error:', round(np.mean(errors), 2)) – печатаем нашу ошибку (рис. 13)



Рисунок 13 вывод ошибки, для нашего ограниченного маленького леса.

Далее построим все наши деревья графически, чтобы спрогнозировать значения.

from sklearn.tree import export\_graphviz

import pydot – модули необходимые для визуализации.

tree1 = rf.estimators\_[0] - достаем первое дерево

tree2 = rf.estimators\_[1] - достаем второе дерево

tree3 = rf.estimators\_[2] - достаем третье дерево

tree4 = rf.estimators\_[3] - достаем четвертое дерево

tree5 = rf.estimators\_[4] - достаем пятое дерево

tree6 = rf.estimators\_[5] - достаем шестое дерево

tree7 = rf.estimators\_[6] - достаем седьмое дерево

tree8 = rf.estimators\_[7] - достаем восьмое дерево

tree9 = rf.estimators\_[8] - достаем девятое дерево

tree10 = rf.estimators\_[9] - достаем десятое дерево

export\_graphviz(tree1, out\_file = 'tree1.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем первое дерево в .dot формат

export\_graphviz(tree2, out\_file = 'tree2.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем второе дерево в .dot формат

export\_graphviz(tree1, out\_file = 'tree1.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем первое дерево в .dot формат

export\_graphviz(tree2, out\_file = 'tree2.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем второе дерево в .dot формат

export\_graphviz(tree3, out\_file = 'tree3.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем третье дерево в .dot формат

export\_graphviz(tree4, out\_file = 'tree4.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем четвертое дерево в .dot формат

export\_graphviz(tree5, out\_file = 'tree5.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем пятое дерево в .dot формат

export\_graphviz(tree6, out\_file = 'tree6.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем шестое дерево в .dot формат

export\_graphviz(tree7, out\_file = 'tree7.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем седьмое дерево в .dot формат

export\_graphviz(tree8, out\_file = 'tree8.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем восьмое дерево в .dot формат

export\_graphviz(tree9, out\_file = 'tree9.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем девятое дерево в .dot формат

export\_graphviz(tree10, out\_file = 'tree10.dot', feature\_names = feature\_list, rounded = True, precision = 1) – записываем десятое дерево в .dot формат

Далее, через онлайн конвертер “https://dreampuf.github.io/GraphvizOnline/” конвертируем наши полученные файлы в .png формат в виде деревьев и считаем. (Так как у меня почему-то “Jupyter” не видел мои файлы в .dot формате).

Примеры выгруженных деревьев приведены в приложении ‘Примеры деревьев случайного леса’.

По идее, можно использовать простые команды: (graph, ) = pydot.graph\_from\_dot\_file('tree1.dot'); graph.write\_png('tree1.png'), но у меня пишет, что .dot файл не обнаружен.

Сделав графическое отображение, теперь мы можем перейти к расчетной части. Для этого переделаем наш лес в полностью случайный и увеличим количество деревьев до 100000 (можно и больше, но моему процессору уже тяжело 100000 построить).

train\_features, test\_features, train\_labels, test\_labels = train\_test\_split(features, labels, test\_size = 0.25)

rf = RandomForestRegressor(n\_estimators = 100000)

rf.fit(train\_features, train\_labels);

predictions = rf.predict(test\_features)

errors = abs(predictions - test\_labels)

print('Mean Absolute Error:', round(np.mean(errors), 2)) – Рисунок 14.



Рисунок 14. Ошибка нашего полного леса.

features1 = pd.read\_csv('fina.csv') – загружаем регрессоры, с помощью которых будем прогнозировать

features1 = np.array(features1) – переводим их в формат NumPy

pred = rf.predict(features1) – делаем предсказания

i = 0

while i < 11:

print(pred[i]) – выводим их. Рисунок 15.

i=i+1

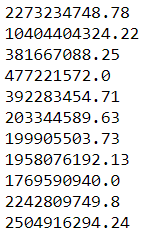
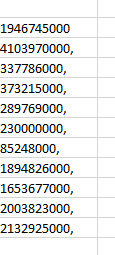


Рисунок 15. Наши предсказания, для нашего набора тестовых данных и истинные значения (слева).

importances = list(rf.feature\_importances\_)

feature\_importances = [(feature, round(importance, 5)) for feature, importance in zip(feature\_list, importances)]

feature\_importances = sorted(feature\_importances, key = lambda x: x[1], reverse = True)

[print('Variable: {:20} Importance: {}'.format(\*pair)) for pair in feature\_importances];

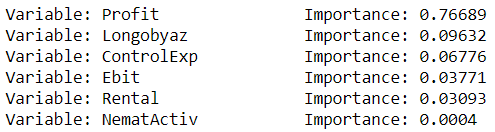


Рисунок 16. Важность переменных для регрессии.

Вывод: метод случайных деревьев позволил включить в анализ факторы, которые редко встречаются в выборке. Факт присутствия долгосрочных займов, факт присутствия нематериальных активов.

# Заключение

Таким образом в нашей работе мы получили подтверждение высокой эффективности алгоритма случайных деревьев с данными с низкой частотой отслеживания. В настоящей работе были решены все поставленные задачи:

1. При изучении основ риск-менеджмента показано, что модель 3LOD является наиболее эффективной формой организации управления рисками, позволяющей снизить степень агентского конфликта и максимизирующую добавленную стоимость. Тем не менее, при работе с реальными финансовыми данными менеджеры при соблюдении норм комплаенса вынуждены манипулировать отчетностью. Манипуляция происходит между периодами отчетности. При таком подходе эффективность построения моделей для анализа риск-менеджмента на реальных данных достаточно низкая.
2. В работе была произведена оценка эффективности корпоративного риск-менеджмента методом линейной регрессии (ANOVA) и классификационное дерево (CHAID). Показано, что при регрессионном анализе наиболее значимым являются нематериальные активы и управленческие расходы, тогда как при построении организационных алгоритмов (дерево CHAID) частоты данных параметров недостаточно.
3. В результате алгоритма случайного леса полученные стандартная ошибка оказалась меньше (2 млрд. < 11 млрд) и значимость переменных выросла (рисунок 16).

Таким образом все поставленные задачи выполнены, цели достигнуты. Также стоит отметить, что при построении случайного леса нет необходимости каким-либо образом стандартизировать данные для получение нормального распределения.

# Литература

1. Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес. URL: <https://habr.com/ru/company/ods/blog/324402/#ansambli>
2. Skurichina M., Duin R. P. W. Limited bagging, boosting and the random subspace method for linear classifiers // Pattern Analysis & Applications. 2002. Pp. 123–126.
3. Breiman, Leo. Bagging predictors // Machine learning 24.2 (1996): 123-140.
4. Will Kehrsen. Random Forest in Python URL: <https://towardsdatascience.com/random-forest-in-python-24d0893d51c0>
5. Е.А. Соколов ФКН ВШЭ: Лекция 3. Решающие деревья URL: <https://www.hse.ru/mirror/pubs/share/215285956>
6. Артиков М. Э. Анализ эффективности применения методов классификации URL: https://moluch.ru/th/8/archive/62/2632/
7. Бычкова, С.М; Итыгилова, Е.Ю. Риски искажения учетной информации: выявление в процессе аудита // Аудиторские ведомости, 2010, №12. URL: http://www.lawmix.ru/bux/476
8. Макарова В. А. Эффективность риск-менеджмента: оценка и ее влияние на инвестиционную привлекательность бизнеса // Управление финансовыми рисками. 2015. № 04(44). с. 270- 287.
9. Breiman L. Random forests // Machine Learning. — 2001. — Vol. 45, no. 1. — Pp. 5–32
10. Попков, М.И. Автоматическая система классификации текстов для базы знаний предприятия, МГУ имени Ломоносова, 2014, Магистерская диссертация. URL: <http://master.cmc.msu.ru/files/master2014_1_popkov.pdf> 27-28
11. Шамаева Д.Р. Деревья решения для задач построения рейтинга коммерческих банков // Технические науки: проблемы и перспективы: материалы V Междунар. науч. конф. (г. Санкт-Петербург, июль 2017 г.). — СПб.: Свое издательство. С. 18-22.
12. Вахрушина, М.А. К вопросу о транспарентности финансовой отчётности. URL: http://www.finotchet.ru/article.html?id=468
13. Черкасова, Е.В; Истомин, В.С. Манипулирование финансовой отчётностью. URL: http://www.amursu.ru/attachments/article/9530/N53\_26.pdf
14. Ingley C., Van der Walt Do Board Processes Influence Director and Board Performance // Corporate Governance: An International Review. 2005. Vol. 13. No. 5. P. 632-653.
15. Ittner, C. D. Larcker D. F.*.* Coming up short on nonfinancial performance measurement. Harvard Business Review (November): 88-95.
16. Jaspal S. (2011) *Jaspal S.* Metrics to Measure Risk Management Strategy
17. Воронцов К.В. Лекции по алгоритмическим композициям, 7 октября 2012. C 15-16 URL: <http://www.machinelearning.ru/wiki/images/0/0d/Voron-ML-Compositions.pdf>
18. Scikit Learn URL: [https://scikit-learn.org](https://scikit-learn.org/)
19. Machinelearning.ru URL: Machinelearning.ru
20. Tablot J., Jakeman M. How to Perfomance Benchmark Your Risk Management, 18 November 2016, Pp. 29-31.
21. Potter P., Toburen M.. The 3 Lines of Defense for Good Risk Management, June 1 2016. URL: <http://www.rmmagazine.com/2016/06/01/the-3-lines-of-defense-for-good-risk-management/>

# Приложения

# Приложение 1

Регрессионный анализ исследовательской проблемы

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables Entered/Removeda** | | | |
| Model | Variables Entered | Variables Removed | Method |
| 1 | Нематериальные активы, RUB, Рентабельность активов (ROA), %. Управленческие расходы, RUBb | . | Enter |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | |
| b. All requested variables entered. | | | |

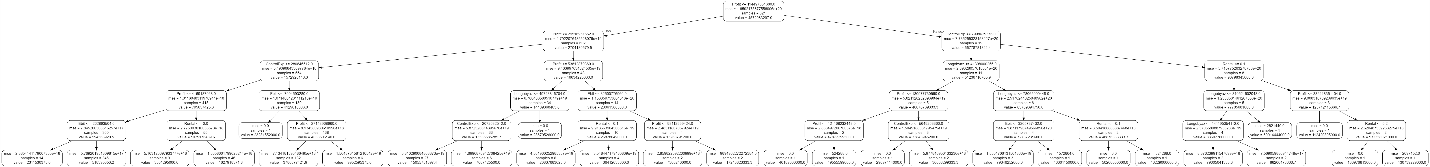
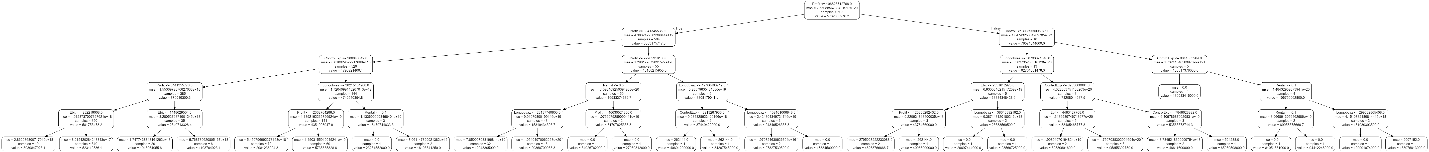
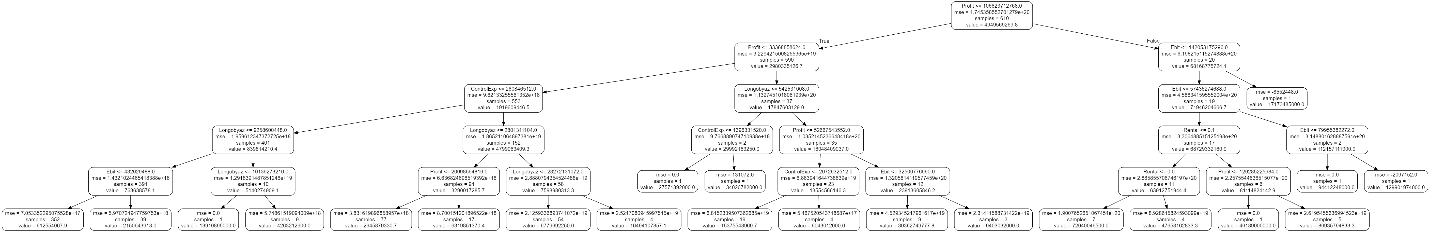
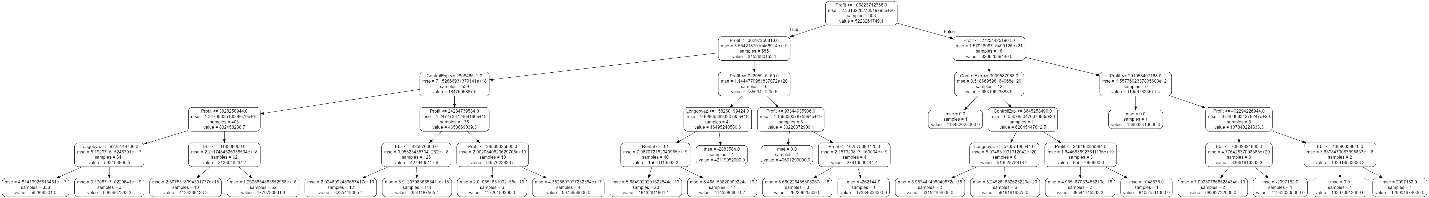
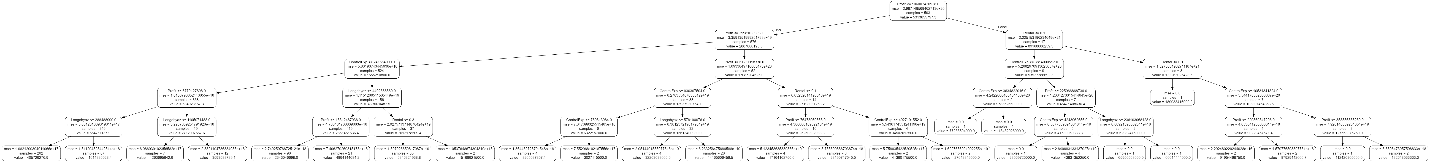
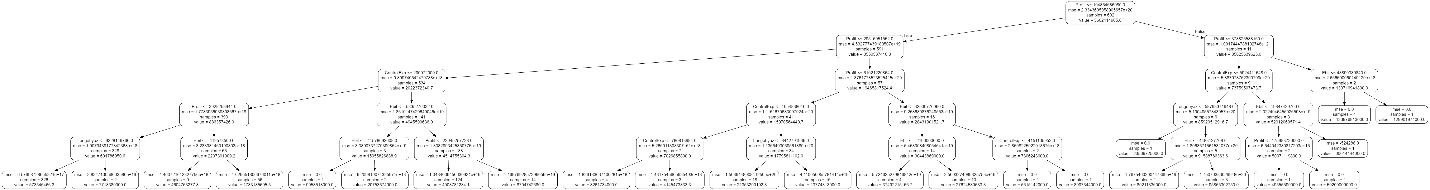
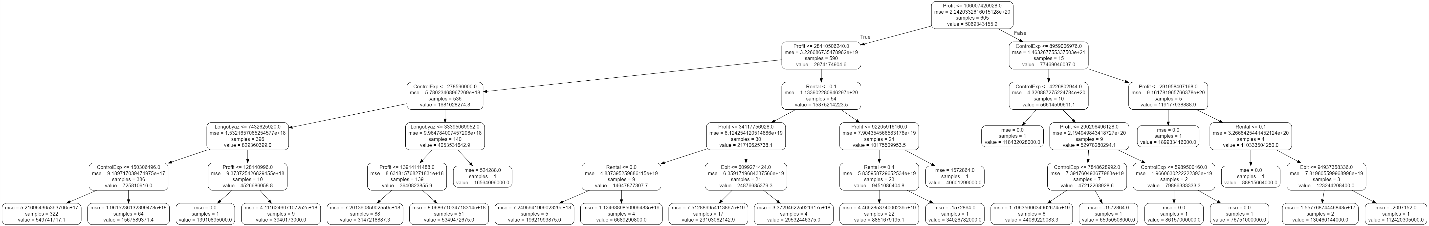
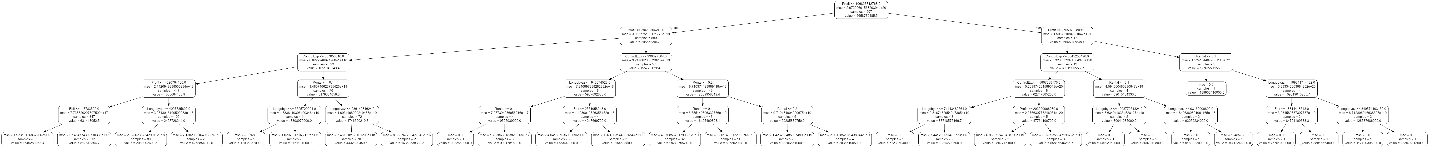
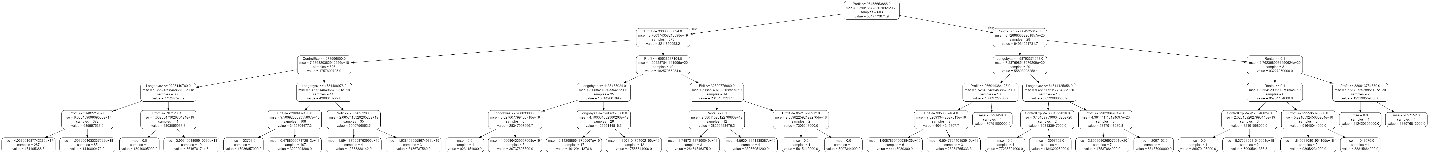
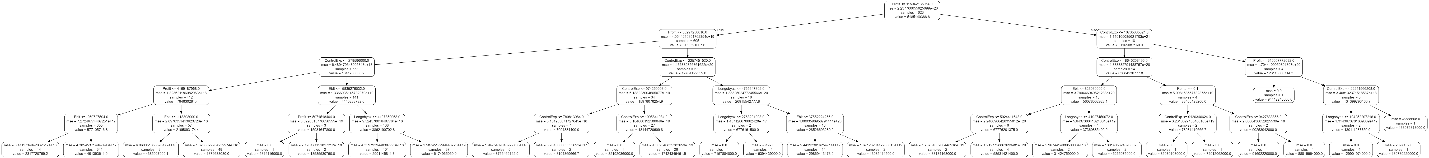
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,864a | ,747 | ,746 | 11060298363,6844220 |
| a. Predictors: (Constant), 2013, Нематериальные активы, RUB, 2013, Рентабельность активов (ROA), %, 2013, Управленческие расходы, RUB | | | | |
| b. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 271863372887974620000000,000 | 3 | 90621124295991550000000,000 | 740,791 | ,000b |
| Residual | 92114640519971390000000,000 | 753 | 122330199893720300000,000 |  |  |
| Total | 363978013407946000000000,000 | 756 |  |  |  |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | | | |
| b. Predictors: (Constant), 2013, Нематериальные активы, RUB, 2013, Рентабельность активов (ROA), %, 2013, Управленческие расходы, RUB | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 306307495,260 | 445184912,991 |  | ,688 | ,492 |  |  |
| Рентабельность активов (ROA), % | -117637167,650 | 224453939,532 | -,010 | -,524 | ,600 | ,996 | 1,004 |
| Управленческие расходы, RUB | 7,991 | ,196 | ,935 | 40,763 | ,000 | ,639 | 1,565 |
| Нематериальные активы, RUB | -6,011 | 1,071 | -,129 | -5,610 | ,000 | ,639 | 1,564 |
| a. Dependent Variable: 2013, Краткосрочные обязательства, RUB | | | | | | | | |

# Приложение 2.

Примеры деревьев случайного леса.



1. признаковое описание - вектор, составленный из значений, не обязательно числовых, соответствующих некоторому набору признаков для данного объекта.[19] [↑](#footnote-ref-1)