Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.02 «Информационные системы и технологии»
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
BI-технологии в анализе данных Федеральной контрактной системы

УДК 004.89:316.3:005.83

Студент

Группа	ФИО	Подпись	Дата
8ИМ6А	Чебоксаров Владимир Александрович		

Руководитель

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	А.В. Кудинов	к.т.н.		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН	Е.В. Старикова	к.филос.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОКД	И.С. Король	к.хим.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ	Н.Г. Марков	д.т.н.		

Томск – 2018 г.

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.02 «Информационные системы и технологии»
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Н. Г. Марков
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ИМ6А	Чебоксарову Владимиру Александровичу

Тема работы:

BI-технологии в анализе данных Федеральной контрактной системы	
Утверждена приказом директора (дата, номер)	

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Задание, полученное от научного руководителя
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов,*	1. Анализ предметной области. 2. Выгрузка и обработка данных с портала Федеральной контрактной системы. 3. Обучение предиктивной модели на основе наивного байесовского алгоритма. 4. Обучение предиктивной модели с помощью искусственных нейронных сетей с различными методами обучения.

подлежащих разработке; заключение по работе).	
Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Схема ИНС 2. Схема ИНС с ЭА 3. Статистика выгруженных данных

Консультанты по разделам выпускной квалификационной работы

(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Е.В. Старикова
Социальная ответственность	И.С. Король
Раздел на иностранном языке	О.В. Комиссарова

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Анализ предметной области (Subject area analysis)

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	А.В. Кудинов	к. т. н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ6А	Чебоксаров Владимир Александрович		

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа информационных технологий и робототехники
Направление подготовки 09.04.02 «Информационные системы и технологии»
Уровень образования магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения (осенний / весенний семестр 2017/2018 учебного года)

Форма представления работы:

магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
18.02.2018	Анализ предметной области	10
28.02.2018	Извлечение и подготовка данных	15
10.03.2018	Обучение модели с помощью НБА	20
20.05.2018	Обучение модели с помощью ИНС	20
23.05.2018	Финансовый менеджмент, ресурсоэффективности и ресурсосбережение	10
14.05.2018	Социальная ответственность	10
28.05.2018	Обязательное приложение на иностранном языке	10
04.06.2018	Оформление пояснительной записки	5

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	А.В. Кудинов	к.т.н.		

СОГЛАСОВАНО:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ	Н.Г. Марков	д.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ИМ6А	Чебоксарову Владимиру Александровичу

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистр	Направление подготовки	09.04.02.Информационные системы и технологии

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. Стоимость ресурсов научного исследования (НИ): человеческих; 2. Используемая система налогообложения, ставки налогов, отчислений;	
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения;	1. Потенциальные потребители результатов исследования; 2. Определение возможных альтернатив проведения научных исследований.
2. Планирование научно-исследовательских работ;	1. Структура работ в рамках научного исследования; 2. Определение трудоемкости выполненных работ; 3. Разработка графика проведения научного исследования; 4. Формирование бюджета научно-технического исследования.
3. Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования	Расчет интегральных показателей разработки и сравнительной эффективности вариантов исполнения.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН	Старикова Екатерина Васильевна	к.филос.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ6А	Чебоксаров Владимир Александрович		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ИМ6А	Чебоксарову Владимиру Александровичу

Институт		Кафедра	
Уровень образования	Магистр	Направление/специальность	09.04.02 Информационные системы и технологии

Исходные данные к разделу «Социальная ответственность»:	
Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Объект исследования: предиктивная модель вычисления вероятности успешного завершения проекта на основе его начальных показателей. Методика: создание алгоритма классификации проектов с использованием наивного байесовского алгоритма, искусственных нейронных сетей и эволюционного алгоритма. Область применения: анализ коммерческих проектов. Рабочая зона: исследовательская лаборатория.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Производственная безопасность 1.1. Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения. 1.2. Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения.	1.1. Анализ вредных и опасных факторов, которые может создать объект исследования. 1.2. Анализ вредных и опасных факторов, которые могут возникнуть на рабочем месте при проведении исследований. 1.3. Обоснование мероприятий по защите исследователя от действия опасных и вредных факторов.
2. Экологическая безопасность	2.1. Анализ возможного влияния объекта исследования на окружающую среду. 2.2. Анализ влияния процесса исследования на окружающую среду. 2.3. Обоснование мероприятий по защите

	окружающей среды.
3. Безопасность в чрезвычайных ситуациях: 3.1.Выбор наиболее типичной ЧС; 3.2.Разработка превентивных мер по предупреждению ЧС;	3.1. Анализ вероятных ЧС, которые может инициировать объект исследований. 3.2. Анализ вероятных ЧС, которые могут возникнуть на рабочем месте при проведении исследований. 3.3. Обоснование мероприятий по предотвращению ЧС и разработка порядка действия в случае возникновения ЧС.
4. Правовые и организационные вопросы обеспечения безопасности: 4.1. Организационные мероприятия при компоновке рабочей зоны.	4.1. Специальные (характерные для рабочей зоны исследователя) правовые нормы трудового законодательства. 4.2. Организационные мероприятия при компоновке рабочей зоны исследователя.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОКД	Король Ирина Степановна	к.хим.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ6А	Чебоксаров Владимир Александрович		

РЕФЕРАТ

Выпускная квалификационная работа 91 с., в том числе 15 рис., 15 табл., 25 источников, 2 прил.

Ключевые слова: Business Intelligence, BI-технологии, Федеральная контрактная система, наивный байесовский алгоритм, искусственная нейронная сеть, эволюционный алгоритм.

Объектом исследования является возможность создания модели с помощью методов машинного обучения для прогнозирования результатов выполнения проекта на основе текстовых данных о проекте в виде договора на его выполнение, а также его основных показателях.

Цель работы – разработка предиктивной модели для оценки возможного результата выполнения проекта, где выходным параметром модели должна стать вероятность успешного завершения оцениваемого проекта.

В процессе исследования проводились анализ методов для классификации текстовых и векторных данных, анализ методов оценки результатов классификации, а также описываются используемые языки программирования и среда разработки.

В результате исследования было разработано программное приложение, основанное на результатах обучения модели, для оценки возможных результатов завершения проекта.

Область применения: разработанное программное приложение может представлять интерес для компаний, занимающихся проектной деятельности в любых сферах предпринимательства.

Экономическая эффективность/значимость работы: себестоимость разработки составила 100445 руб. 00 коп. Уровень научного эффекта – средний.

В будущем планируется:

- Максимальное повышение точности классификации;
- Возможное профилирование программного продукта для различных сфер деятельности.

ОПРЕДЕЛЕНИЯ И ОБОЗНАЧЕНИЯ

ПО – программное обеспечение.

БД – база данных.

ИАД – интеллектуальный анализ данных.

СУБД – система управления базами данных.

ЯП – язык программирования.

НБА – наивный байесовский алгоритм.

ИНС – искусственная нейронная сеть.

ЭА – эволюционный алгоритм.

SQL – Structured Query Language.

HTML – HyperText Markup Language.

XML – Extensible Markup Language.

DIN – Deutsches Institut für Normung.

NLTK – Natural Language Toolkit.

AGPL – Affero General Public License.

SSMS – SQL Server Management Studio.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Программное обеспечение (ПО) для интеллектуального анализа данных (ИАД) позволяет пользователям применять полуавтоматический и прогнозирующий методы для анализа необработанных данных и поиска новых способов получения информации. Данное ПО обычно применяется к очень большим наборам данных и связанным с ними функциям, или любой набор данных, слишком большой или сложный для человеческого анализа.

Приложения для интеллектуального анализа данных помогают пользователям обнаруживать корреляции и соединения в больших наборах данных. Они часто включают многочисленные записи с несколькими переменными и могут содержать даже смешанные структурированные и неструктурированные данные. Из-за размера и сложности этих наборов данных любые ценные корреляции внутри них оставались бы незамеченными, если бы не неустанный алгоритмический анализ, выполненный с инструментами интеллектуального анализа данных.

Целью данной работы является создание предиктивной модели для оценки возможного результата выполнения проекта, где выходным параметром модели должна стать вероятность успешного завершения оцениваемого проекта.

В ходе выполнения работы выполнялись следующие задачи:

1. Анализ предметной области;
2. Выгрузка и подготовка данных по выполнению договоров согласно федеральному закону №223-ФЗ;
3. Обучение предиктивной модели с помощью наивного байесовского алгоритма на основе текстовых данных о проекте;
4. Обучение предиктивной модели с помощью искусственных нейронных сетей на основе основных показателей проекта;
5. Апробация обученных моделей для предсказания успешности проектов;
6. Оценка полученных результатов.

1 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Интеллектуальный анализ данных в управлении проектами

Интеллектуальный анализ данных (ИАД) – это процесс выявления значимых корреляций, образцов и тенденций в больших объемах данных [1].

Особенно широко методы ИАД применяются в бизнес-приложениях аналитиками и руководителями компаний. Для этих категорий пользователей разрабатываются инструментальные средства высокого уровня, позволяющие решать достаточно сложные практические задачи без специальной математической подготовки. Актуальность использования ИАД в бизнесе связана с жесткой конкуренцией, возникшей вследствие перехода от «рынка продавца» к «рынку покупателя». В этих условиях особенно важно качество и обоснованность принимаемых решений, что требует строгого количественного анализа имеющихся данных. При работе с большими объемами накапливаемой информации необходимо постоянно оперативно отслеживать динамику рынка, а это практически невозможно без автоматизации аналитической деятельности.

К наиболее типовым задачам ИАД в управлении проектами можно отнести:

1. Прогнозирование. Прогнозирование является одной из наиболее используемых задач ИАД. Его применение связано с потребностью использования его результатов при планировании и составлении бюджета, т.к. необходима оценка объема будущих продаж и других параметров с учетом многочисленных факторов, которые могут быть связаны между собой (региональные, сезонные и т.д.).

2. Маркетинговый анализ. Разработка маркетингового плана подразумевает под собой определенных знаний. Например, знаний о влиянии различных показателей продукта (стоимость, затраты на маркетинг) на уровень его продаж. В данном случае нейросетевые модели позволяют составить более точный прогноз для оценки этого влияния.

3. Анализ работы персонала. Множество факторов, таких как: уровень оплаты труда, опыт работы, социальный климат внутри компании,

взаимоотношения с руководством, – влияют на производительность труда работников. Анализ описанных выше факторов позволяет разработать выработать методику повышения производительности труда, а также предложить оптимальную стратегию подбора кадров в будущем.

4. Профилирование клиентов. В организациях с большой базой клиентов использование нейросетевых моделей позволяет производить оценку клиентов с целью составления списка наиболее выгодных клиентов для сотрудничества и получения таким образом портрета «типичного клиента компании». Кроме того, с помощью этих моделей можно выяснить причину неэффективной работы с некоторыми заказчиками и выработать стратегию поиска подходящих клиентов в будущем.

5. Оценка потенциальных клиентов. Оценка вероятности успешного завершения переговоров с клиентом в виде подписания договора или продажи товара позволяет повысить эффективность планирования в организациях и увеличить экономию средств. Для составления данной оценки может быть использован опыт работы с предыдущими клиентами компании, а также анализ характерных особенностей тех заявок от клиентов, которые закончились реальными продажами. Результаты описанного анализа могут быть использованы руководящим персонал организации с целью повышения эффективности ее работы за счет заключения договоренностей с наиболее перспективными клиентами.

Очевидно, что перечисленные виды задач актуальны практически для всех отраслей бизнеса: банковского дела и страхования (выявление злоупотреблений с кредитными карточками, оценка кредитных рисков, оценка закладных, выявление профилей пользователей, оценка эффективности региональных отделений, вероятность подачи заявки на выплату страховки и др.), финансовых рынков (прогнозирование, анализ портфелей, моделирование индексов), производства (прогнозирование спроса, контроль качества, оценка дизайна продукции), торговли и т. д.

1.2 Методика оценки успешности выполнения проекта

Перед определением методики успешности проекта необходимо обратиться к самому термину «проект». Одно из наиболее полных определений приведено в стандарте DIN 69901 [2], согласно которому проект – это предприятие (или намерение), которое в значительной степени характеризуется неповторимостью условий в их совокупности, например:

- задание цели;
- временные, финансовые, людские и другие ограничения;
- разграничения от других намерений;
- специфическая для проекта организация его осуществления.

Также для оценки может быть использовано так называемое правило «железного треугольника», которое описывает баланс между стоимостью проекта, временем его выполнения и качеством полученного результата (рис. 1).

Рисунок 1 – «Железный треугольник»

Согласно этому определению можно вывести общий критерий успешности проекта — это достижение целей проекта в запланированное время и в рамках запланированных ресурсов.

1.3 Описание федерального закона №223-ФЗ

В качестве источника данных по выполнению проектов была выбрана Федеральная контрактная система РФ, предоставляющая свободный и безвозмездный доступ к полной и достоверной информации по выполнению

проектов в сфере закупок и закупках товаров, работ, услуг отдельными видами юридических лиц согласно федеральному закону №223-ФЗ [3].

Федеральный закон №223-ФЗ регламентирует общие принципы закупок для:

- Организаций с долей участия государства более 50%;
- Компаний, занимающиеся регулируемыми видами деятельности;
- Организаций-субъектов естественных монополий;
- Бюджетных организаций, проводящих закупку за счет внебюджетных средств.

Выбор портала Федеральной контрактной системы по 223-ФЗ был обусловлен следующими факторами:

- Открытый доступ к данным;
- Достоверная информация о проектах;
- Наличие большого количества данных за разные периоды и по разным регионам РФ;
- Данные хранятся в структурированном и удобном для обработки формате XML;
- Данные содержат информацию о заказчике и исполнителе проекта, а также его временные и стоимостные показатели;
- Данные описывают проекты, выполняемые организациями, которые наиболее приближенны к коммерческому типу.

Таким образом, информация о конкурсах, публикуемая в системе госзакупок, является центральным источником актуальных сведений о возможных «государственных» заказах по профилю для множества компаний из самых разных сфер деятельности — от услуг охраны до геофизических изысканий. Наличие подобного источника данных позволяет нам проверить предположение о существовании определенной зависимости успешного завершения проекта от его начальных показателей.

1.4 Методы решения задачи классификации

Важнейшим моментом в решении поставленной задачи является создание эффективного алгоритма классификации, на основании которого будет происходить обучение модели.

В описываемом в данной работе случае в имеющихся данных по проектам отсутствуют какие-либо категориальные признаки, по которым можно было бы определить, к какому именно классу относится проект (успешно завершенному или нет), соответственно, появляется необходимость в поиске неявных зависимостей между показателями проекта и его результатом. Поэтому в данном случае в качестве методов для обучения классификатора было решено использовать методы машинного обучения.

Среди всех методов классификации можно выделить их следующие основные группы:

- Байесовские классификаторы;
- Искусственные нейронные сети (ИНС);
- Линейные разделители;
- Алгоритмические композиции.

В нашем случае ниже будут рассмотрены наивный байесовский алгоритм и некоторые виды искусственных нейронных сетей, как одни из наиболее популярных методов обучения классификаторов, а также кросс-валидация, как метод оценки результатов обучения модели.

1.4.1 Наивный байесовский алгоритм

Наивный байесовский алгоритм (НБА) – это алгоритм классификации, основанный на теореме Байеса с допущением о независимости признаков [4]. Другими словами, НБА предполагает, что наличие какого-либо признака в классе не связано с наличием какого-либо другого признака. Даже если признаки зависят друг от друга или от других признаков, в любом случае они вносят независимый вклад в результат. В связи с таким допущением алгоритм называется «наивным».

Модели на основе НБА достаточно просты и крайне полезны при работе с очень большими наборами данных. При своей простоте НБА способен превзойти даже некоторые сложные алгоритмы классификации.

Теорема Байеса позволяет рассчитать апостериорную вероятность P(c|x) на основе P(c), P(x) и P(x|c) (рисунок 2).

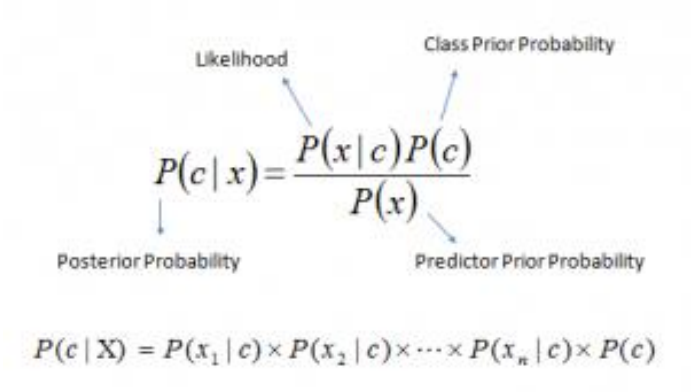

Рисунок 2 – Формула теоремы Байеса

На рисунке выше:

- P(c|x) – апостериорная вероятность данного класса c (т.е. данного значения целевой переменной) при данном значении признака x.
- P(c) – априорная вероятность данного класса.
- P(x|c) – правдоподобие, т.е. вероятность данного значения признака при данном классе.
- P(x) – априорная вероятность данного значения признака.

Использование наивного байесовского алгоритма имеет следующие положительные стороны:

- Классификация, в том числе многоклассовая, выполняется легко и быстро.
- Когда допущение о независимости выполняется, НБА превосходит другие алгоритмы, такие как логистическая регрессия, и при этом требует меньший объем обучающих данных.
- НБА лучше работает с категориальными признаками, чем с непрерывными. Для непрерывных признаков предполагается нормальное распределение, что является достаточно сильным допущением.

Однако применение НБА может иметь и отрицательное влияние:

- Если в тестовом наборе данных присутствует некоторое значение категориального признака, которое не встречалось в обучающем наборе данных, тогда модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота». Данную проблему можно решить с помощью сглаживания. Одним из самых простых методов является сглаживание по Лапласу.
- Хотя НБА является хорошим классификатором, значения спрогнозированных вероятностей не всегда являются достаточно точными.
- Еще одним ограничением НБА является допущение о независимости признаков. В реальности наборы полностью независимых признаков встречаются крайне редко.

1.4.2 Искусственные нейронные сети

При решении задач классификации необходимо отнести имеющиеся статические образцы к определенным классам. Возможно несколько способов представления данных. Наиболее распространенным является способ, при котором образец представляется вектором. С обработкой данных, представленных в таком виде, наиболее хорошо справляются искусственные нейронные сети.

1.4.2.1 Персептрон

Персептрон является простейшим представителем ИНС. В основе персептрона лежит математическая модель восприятия информации мозгом [5]. В самом общем своем виде (как его описывал Розенблатт) он представляет систему из элементов трех разных типов: сенсоров, ассоциативных элементов и реагирующих элементов (рис. 3).

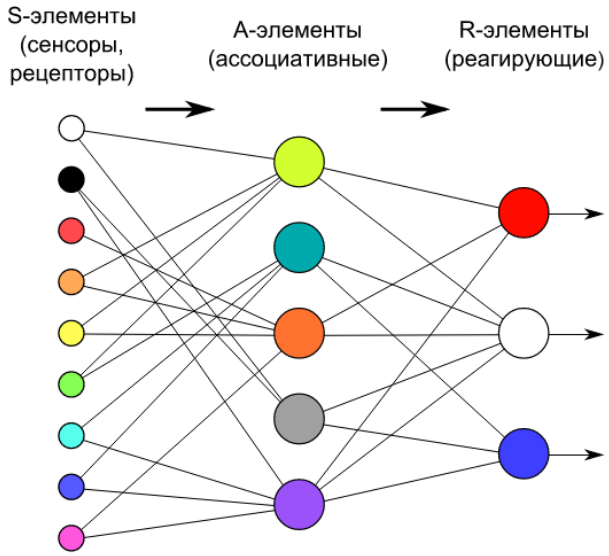

Рисунок 3 – Простейшая модель персептрона

Выделяется три основных вида персептрона:

- Однослойный персептрон;
- Персептрон с одним скрытым слоем;
- Многослойный персептрон.

Выбор определенного вида персептрона варьируется от сложности решаемой задачи и ее предметной областью.

Далее рассмотрим методы обучения ИНС.

1.4.2.2 Метод обратного распространения ошибки

Алгоритм обратного распространения ошибки является одним из методов обучения многослойных нейронных сетей прямого распространения, называемых также многослойными персептронами.

Обучение алгоритмом обратного распространения ошибки предполагает два прохода по всем слоям сети: прямого и обратного. При прямом проходе входной вектор подается на входной слой нейронной сети, после чего распространяется по сети от слоя к слою. В результате генерируется набор выходных сигналов, который и является фактической реакцией сети на данный входной образ. Во время прямого прохода все синаптические веса сети фиксированы. Во время обратного прохода все синаптические веса настраиваются в соответствии с правилом коррекции ошибок, а именно:

фактический выход сети вычитается из желаемого, в результате чего формируется сигнал ошибки. Этот сигнал впоследствии распространяется по сети в направлении, обратном направлению синаптических связей. Отсюда и название – алгоритм обратного распространения ошибки. Синаптические веса настраиваются с целью максимального приближения выходного сигнала сети к желаемому [6].

1.4.2.3 Нейроэволюционный алгоритм

Эволюционные алгоритмы (ЭА) моделируют базовые положения в теории биологической эволюции — процессы отбора, мутации и воспроизводства. Поведение агентов определяется окружающей средой. Множество агентов принято называть популяцией. Такая популяция эволюционирует в соответствии с правилами отбора в соответствии с целевой функцией, задаваемой окружающей средой. Таким образом, каждому агенту (индивидууму) популяции назначается значение его пригодности в окружающей среде. Размножаются только наиболее пригодные виды. Рекомбинация и мутация позволяют изменяться агентам и приспособляться к среде. Такие алгоритмы относятся к адаптивным поисковым механизмам. Схема работы ЭА представлена на рисунке 4.

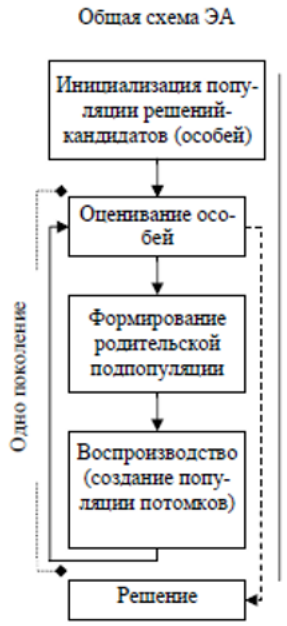

Рисунок 4 – Общая схема эволюционного алгоритма

Одной из отличительных особенностей ЭА являются их адаптивные способности, что дает возможность реализовать подстройку параметров ЭА в процессе его работы для повышения эффективности ЭА и качества результатов. Комбинация ИНС и эволюционных алгоритмов дает возможность совместить гибкость настройки ИНС и адаптивность ЭА, что позволяет реализовать во многом унифицированный подход к решению широкого спектра задач классификации [7]. В нашем случае использование эволюционного подхода позволяет одновременно настраивать веса связей и структуру ИНС.

1.4.3 Кросс-валидация

Кросс-валидация – процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам [8].

Сам процесс кросс-валидации выполняется в несколько этапов. На первом этапе происходит фиксация имеющейся исходной выборки данных на две подвыборки: обучающую и контрольную. Далее для каждой подвыборки выполняется обучение соответствующего алгоритма на полученной обучающей подвыборке. На последнем этапе происходит оценка средней ошибки работы алгоритма на объектах контрольной подвыборки.

Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. Если выборка независима, то средняя ошибка скользящего контроля даёт несмещённую оценку вероятности ошибки. Это выгодно отличает её от средней ошибки на обучающей выборке, которая может оказаться смещённой (оптимистически заниженной) оценкой вероятности ошибки, что связано с явлением переобучения.

Скользящий контроль является стандартной методикой тестирования и сравнения алгоритмов классификации, регрессии и прогнозирования.

При этом для оценки работы алгоритма классификации используются два основных показателя:

- Точность – отношение количества точно определенных классификатором элементов к общему количеству элементов;

• Полнота – отношение количества точно определенных элементов класса к общему количеству элементов этого класса.

Для более простого их описания их описания используем следующие обозначения: TP – истинно-положительные решения классификатора, TN – истинно-отрицательные решения классификатора, FP – ложноположительные решения классификатора, FN – ложноотрицательные решения классификатора.

Тогда точность и полноту можно представить следующими формулами:

$$\text{Точность} = \frac{TP}{TP + FP}$$

$$\text{Полнота} = \frac{TP}{TP + FN}$$

1.5 Описание инструментов разработки

1.5.1 RapidMiner

Написанный на языке программирования Java, этот инструмент предлагает расширенную аналитику с помощью фреймворков на основе шаблонов. Пользователям едва ли нужно писать код. Предлагаемый как услуга, а не часть локального программного обеспечения, этот инструмент занимает верхнюю позицию в списке инструментов интеллектуального анализа данных, находящихся в свободном доступе (рисунок 5).

В дополнение к интеллектуальному анализу данных RapidMiner также предоставляет такие функции, как предварительная обработка данных и визуализация, интеллектуальная аналитика и статистическое моделирование, оценка и развертывание. Также дополнительными возможностями являются предоставляемые им схемы обучения, модели и алгоритмы из сценариев WEKA и R.

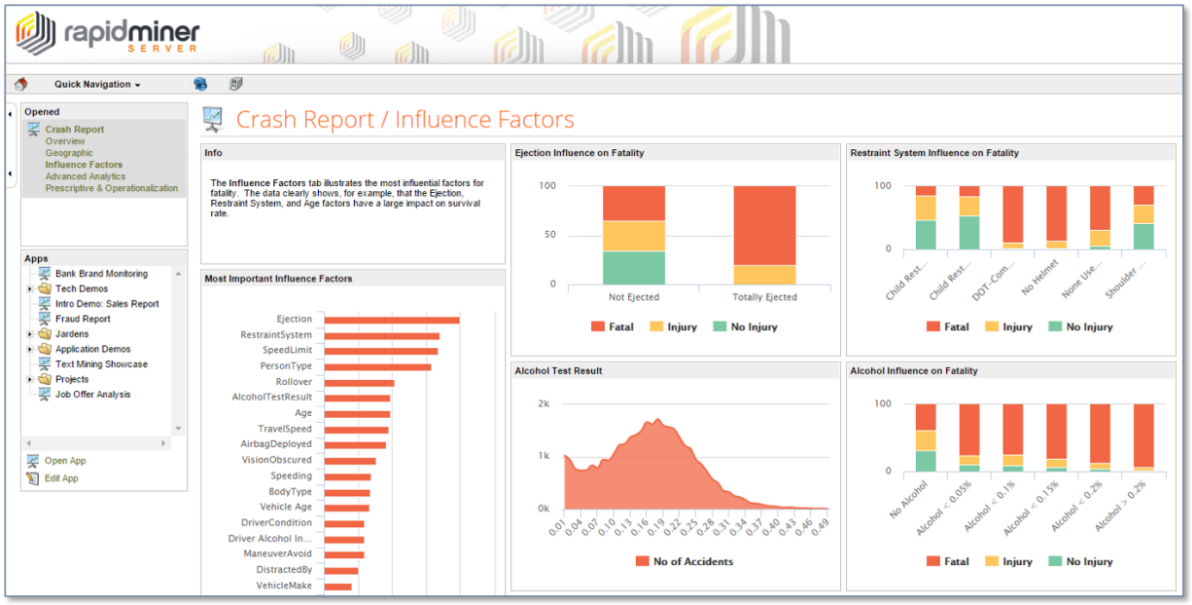

Рисунок 5 – Интерфейс RapidMiner

RapidMiner распространяется под лицензией AGPL с открытым исходным кодом и может быть загружен с SourceForge, где он оценивается как программное обеспечение для бизнес-аналитики номер один [9].

1.5.2 Python

Для реализации модели был выбран язык программирования Python. Выбор был сделан ввиду нескольких причин:

- Простой синтаксис языка;
- Низкий порог вхождения;
- Множество библиотек, помогающих в реализации различного рода задач;
- Наиболее часто используемый язык в анализе данных и машинном обучении;
- Обилие документации.

В результате выполнения описываемой работы с помощью Python должны были быть решены задачи токенизации текстов документов и обучения модели с помощью НБА и ИНС. Для этих целей также использовались следующие библиотеки:

- NLTK – пакет библиотек и программ для символьной и статистической обработки естественного языка [10];
- Scikit-Learn – библиотека, позволяющая реализовать множество из уже имеющих алгоритмов машинного обучения [11].

1.5.3 C#

C# – объектно-ориентированный язык программирования, созданный для разработки приложений для платформы .NET Framework [12].

Выбор данного языка программирования может быть обусловлен следующими факторами:

- Официальная техническая поддержка от Microsoft;
- Хорошая интеграция с СУБД MS SQL;
- Обширная документация;
- Большое количество синтаксических конструкций, направленных на упрощение написания кода;
- Наличие менеджера пакетов NuGet, который позволяет упростить поиск и подключение библиотек к проекту.

1.6 Цели и задачи разработки

Исследование рынка программных продуктов для помощи аналитикам в оценке результатов будущего проекта показало, что данный сегмент рынка пуст, и в целом на данный момент отсутствуют какие-либо программные решение для прогнозирования результатов проекта.

Ввиду того, что на данный момент потребность в программном обеспечении подобного вида не удовлетворена, целью данной работы является создание предиктивной модели для оценки возможного результата выполнения проекта. При этом выходным параметром модели должна стать вероятность успешного завершения оцениваемого проекта.

Так как отрытые источники данных о проектах коммерческих компаний и результатах их выполнения не находятся в открытом доступе, в качестве

источника был выбран портал Федеральной контрактной системы, содержащий данные о государственных контрактах в различных сферах деятельности.

Для повышения качества результатов и увеличения выборки данных в качестве было выбрано два вида исходных данных о проекте:

- Текстовый документ договора на исполнение;
- Основные временные и стоимостные показатели проекта, а также данные о заказчике и поставщике.

Соответственно, для каждого типа исходных данных был выбран собственный метод для обучения модели.

В качестве метода для обработки документов был выбран наивный байесовский алгоритм в силу того, что он является одним из наиболее быстрых, простых и эффективных алгоритмов для работы с документами, имеющими схожую структуру.

Второй вариант обработки данных проектов связан с нахождением корреляции между группой показателей проекта и возможностью его успешного завершения. Так как в данном случае в качестве входные данные представляют собой вектор, содержащий основные показатели проекта, то наиболее правильным будет выбор ИНС в качестве метода обучения классификатора.

Для повышения вероятности успешного обучения также было два метода для корректировки параметров ИНС: метод обратного распространения ошибки и метод, использующий эволюционный алгоритм для подбора структуры ИНС и весов ее синаптических связей.

В ходе выполнения работы были поставлены следующие задачи:

1. Анализ предметной области;
2. Выгрузка по выполнению договоров согласно федеральному закону №223-ФЗ;
3. Подготовка данных и загрузка их в БД;
4. Дополнительная выгрузка документов договоров для проектов;
5. Токенизация документов для применения НБА;

6. Обучение предиктивной модели с помощью НБА на основе текстовых данных о проекте;

7. Выделение и дополнение основных показателей проекта;

8. Обучение предиктивной модели с помощью ИНС методом обратного распространения ошибки на основе основных показателей проекта;

9. Обучение предиктивной модели с помощью ИНС с использованием эволюционного алгоритма на основе основных показателей проекта;

10. Апробация созданных моделей для предсказания успешности проектов;

11. Оценка полученных результатов.

2 ИЗВЛЕЧЕНИЕ И ПОДГОТОВКА ДАННЫХ

Для выполнения практической части использовались данные по государственным закупкам согласно федеральному закону №223-ФЗ по всей территории Российской Федерации, начиная с 2013 года, выгруженным с официального портала государственных закупок.

Данное решение может быть обосновано следующими факторами:

- Свободный доступ к данным;
- Обширная выборка данных;
- Достоверность данных;
- Структурированность данных (хранятся в формате XML).

Извлечение данных по указанному федеральному закону будет осуществляться с FTP-сервера портала Федеральной контрактной системы. Так как в нашем случае количество параметров проектов довольно мало (подробнее описание использованных параметров представлено в главе 4), то для увеличения точно работы классификационного алгоритма было принято решения дополнительно использовать тексты договоров соответствующих проектов. Результатом работы моделей должно стать значение в диапазоне [0, 1], которое будет интерпретироваться как вероятность успешного завершения проекта. Используемые количество данных для обучения и проверки работы модели, а также пороговое значение для оценки успешности проекта будут установлены в соответствии с используемым методом классификации и сложностью обрабатываемых данных.

2.1 Структура данных контрактов по 223-ФЗ

В соответствии с вышеописанной последовательностью выполнения работы, первым этапом является изучение структуры хранения данных на FTP-сервере федеральной контрактной системы.

Данные хранятся на FTP-сервере портала в свободном доступе, однако имеют довольно сложную структуру хранения.

Изначально списки файлов на сервере структурируются по папкам по всем регионам РФ, далее происходит их разграничение в соответствии с назначением описываемого документа (рис. 6).

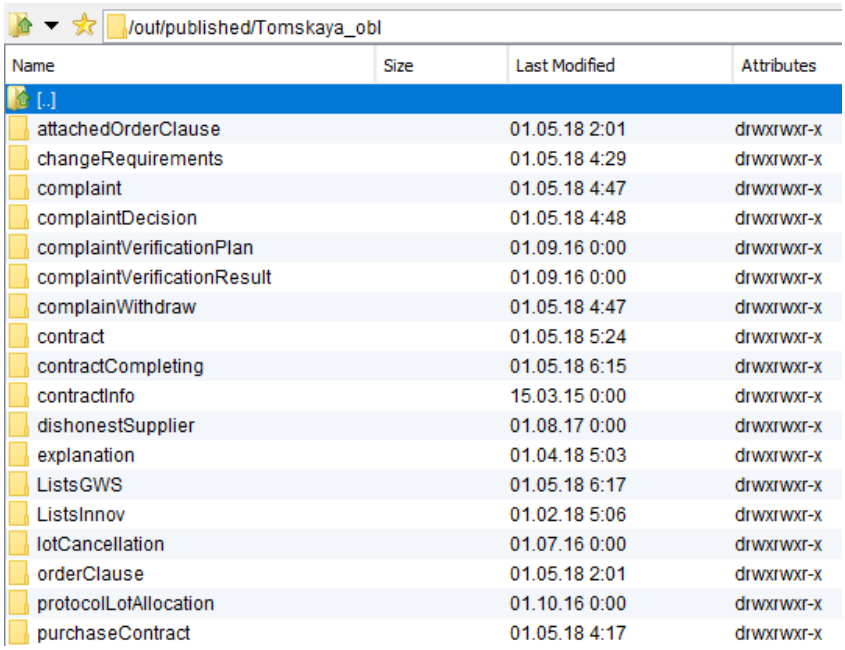

Рисунок 6 – Структура хранения документов согласно их назначению

По представленному выше рисунку можно отметить, что по назначению документы делятся на следующие основные типы:

- Договор;
- Данные об окончании выполнения,
- Жалобы;
- Отмена лота на аукционе;
- Данные о выплатах;
- Изменения требований к выполнению;
- Данные об отмене выплат.

В соответствии с описанной структурой хранения данных был определен следующий список показателей успешности выполнения проекта:

- Отсутствие изменений стоимости проекта в большую сторону;
- Отсутствие жалоб заказчика по выполнению проекта;
- Лот на аукционе не был отменен;

- Отсутствие изменений в датах фактического начала и окончания выполнения проекта;
- Отсутствие изменений длительности выполнения проекта в большую сторону.

Результирующий показатель успешности будет формироваться в соответствии со следующим правилом: проект считается успешным только про одновременном выполнении всех описанных выше показателей.

Пример внутреннего строения XML-файла договора представлен в приложении А.

2.2 Выгрузка и первичная обработка данных

С точки зрения процесса работы с большими данными данный этап работы представляет собой полный цикл процесса ETL [13].

ETL (от англ. Extract, Transform, Load) – один из основных процессов в управлении хранилищами данных, который включает в себя:

1. Извлечение данных из внешних источников;
2. Трансформация и очистка данных с целью приведения их к виду, соответствующему потребностям бизнес-модели;
3. Загрузка данных в соответствующее хранилище данных.

С точки зрения процесса ETL, архитектуру хранилища данных можно представить в виде трёх компонентов:

1. Источник данных: содержит структурированные данные в виде таблиц, совокупности таблиц или просто файла (данные в котором разделены символами-разделителями);
2. Промежуточная область: содержит вспомогательные таблицы, создаваемые временно, и, исключительно для организации процесса выгрузки.
3. Получатель данных: хранилище данных или база данных, в которую должны быть помещены извлечённые данные.

Касательно рассматриваемого случая первый этап процесса занял наибольшее количество временных ресурсов ввиду следующих факторов:

- Большой объем выгружаемых данных;
- Низкая скорость взаимодействия с FTP-сервером портала госзакупок;
- Необходимость в дополнительном поиске документов договоров в силу того, что в наибольшем количестве хранимых на портале документов отсутствовали прямые ссылки на соответствующие им тесты договоров.

Так как в нашем случае отсутствовала необходимость в скачивании абсолютно всех файлов с портала, а полная выгрузка документов привела бы к излишним затратам, было принято решение о написании отдельной программы для извлечения и дальнейшей обработки данных с целью помещения их в базу данных.

Программа была написана на языке программирования C# с использованием основных классов для работы с данными на FTP-серверах (например, FtpWebRequest и FtpWebResponse) и классов для обработки XML-документов (например, XmlDocument и XmlNode).

К примеру, код, позволяющий получить список всех папок и файлов по адресу, имеет следующий вид:

```
var response = (FtpWebResponse)request.GetResponse();

    var responseStream = response.GetResponseStream();
    if (responseStream == null)
    {
        return;
    }

    var reader = new StreamReader(responseStream);
    var folders =
        reader.ReadToEnd()
            .Split(new[] { "\r\n" }, StringSplitOptions.None)
            .ToList();
```

На следующем этапе программа выполняла обработку всех скачанных файлов в 3 шага:

1. Чтение данных из XML-файла;

2. Валидация полученных значений (проверка на пустые и некорректные значения);

3. Запись полученных данных в БД, либо обновление существующих значений.

Однако отдельным этапом в реализации программы стало добавление функционала по скачиванию документов договоров. Причиной этому стал тот факт, что не все файлы, которые должны были содержать ссылки на соответствующие им документы, действительно содержали их. Поэтому был разработан дополнительный модуль для обработки веб-страниц. Функционал данного модуля был предназначен для обработки страниц портала Федеральной контрактной системы с целью поиска ссылок на документы договоров, их дальнейшего скачивания и обработки. Схематическое представление процесса обработки документа в описываемом модуле представлено на рисунке 7.

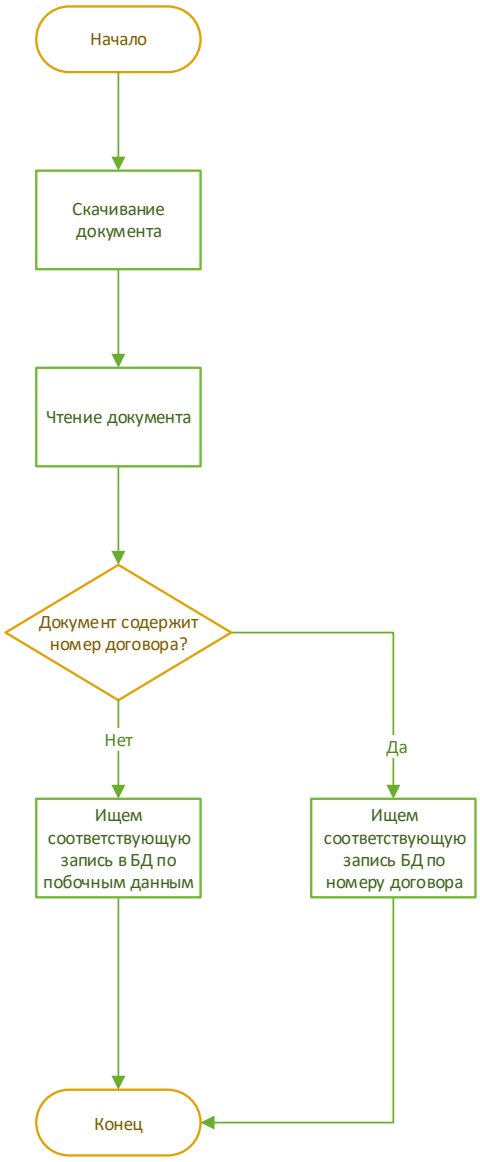

Рисунок 7 – Блок-схема обработки документа

По представленной выше схеме можно заметить, что для установления соответствия между документом и записью в БД используется полнотекстовый поиск по документу, что приводит к замедлению самого процесса обработки данных.

Так как для выгрузки данных был выбран язык программирования C#, то наиболее легким и очевидным решением в выборе СУБД стало решение от компании Microsoft – MS SQL Server 2014. Данная СУБД является системой управления реляционными базами данных, что хорошо согласуется с объектно-ориентированной парадигмой разработки веб-приложения.

Для управления и администрирования SQL Server использовалась утилита SQL Server Management Studio (SSMS), которая включает в себя скриптовый редактор и графическую программу, которая работает с объектами и настройками сервера.

Структура БД в соответствии с описанными данными была реализована в довольно просто виде, всего две таблицы: одна для хранения данных о договорах и другая для хранения данных о компаниях (как заказчиках, так и поставщиках). Схема физической модели БД представлена на рисунке 8.

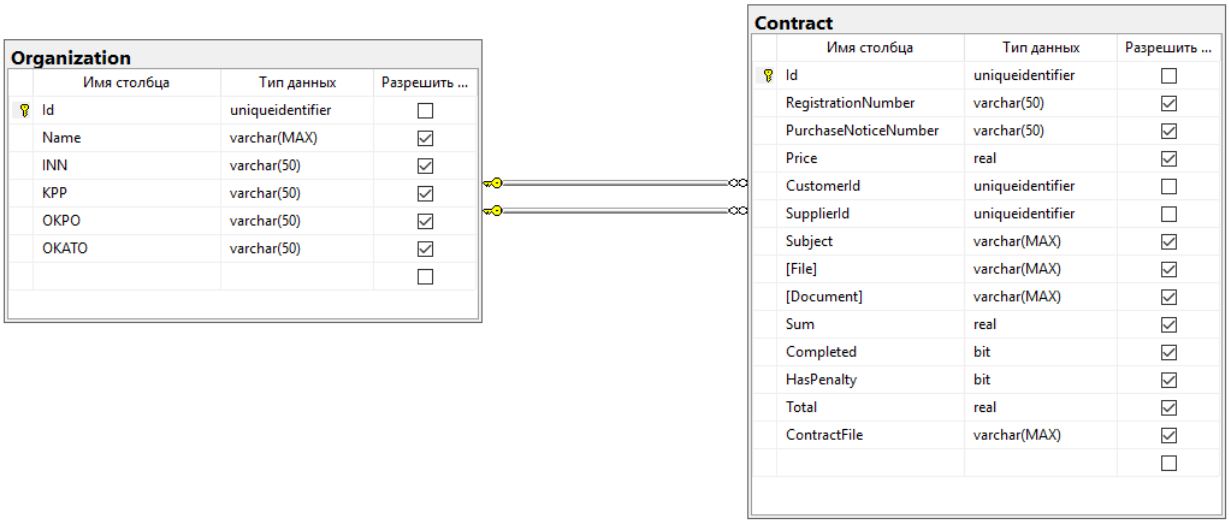

Рисунок 8 – Схема физической модели БД

3 ОБУЧЕНИЕ МОДЕЛИ С ПОМОЩЬЮ НБА

В качестве алгоритма обучения модели на основе текстовых документов договоров был выбран наивный байесовский алгоритм. Модели на основе НБА достаточно просты и крайне полезны при работе с очень большими наборами данных. При этом при своей простоте НБА способен превзойти даже некоторые сложные алгоритмы классификации. Однако так как в рассматриваемом случае производится классификация текстовых данных, то первоначально эти данные должны быть соответствующим образом подготовлены для обработки, а именно разбиты по словам для последующего определения значимости каждого слова в определении результата классификации. Этот процесс носит название токенизации.

В соответствии с вышеописанным процесс обучения был разделен на следующие этапы:

1. Токенизация классифицируемых документов;
2. Реализация НБА;
3. Добавление модификаций для НБА с целью улучшения его работы;
4. Обработка результатов.

3.1 Токенизация

Перед реализацией НБА тексты документов также должны быть обработаны соответствующим образом, а именно разделены на токены (т.е. элементарные единицы текста). Далее найденные токены приводятся к структуре типа размеченный список, в котором каждому токену соответствует целочисленное значение, равное количеству повторений данного токена внутри документа. Данный процесс называется токенизацией.

Токенизация – это процесс разбиения текста на отдельно значимые единицы для его дальнейшей обработки. К токенам относятся как слова, так и знаки пунктуации [14].

В нашем случае процесс токенизации необходим для структуризации содержимого документов. Процесс структуризации представлен на рисунке 9.

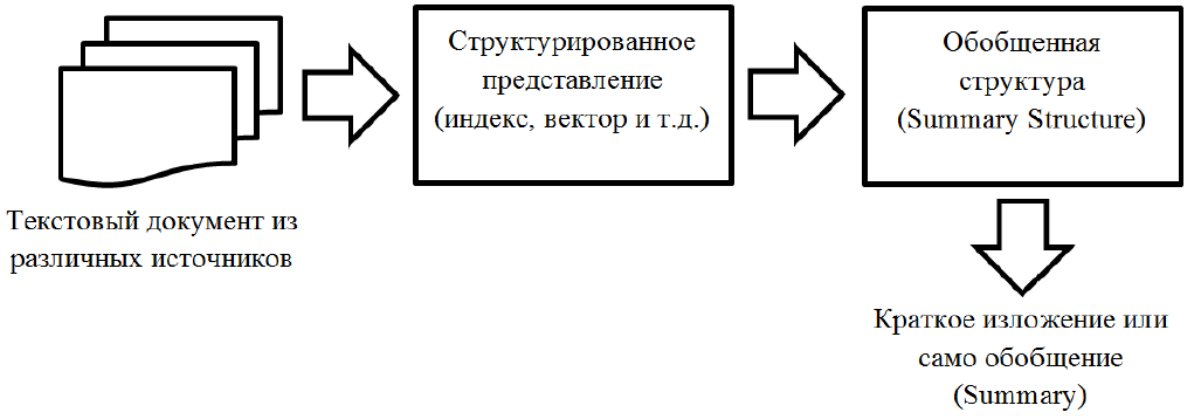

Рисунок 9 – Процесс обработки документов

Для обработки текстов, как наиболее простой вариант, был выбран язык Python с библиотекой NLTK, позволяющей упростить обработку естественного языка и имеющую множество встроенных для этого функций.

В процессе обработки документа в качестве токенов может быть воспринято множество незначимых слов (таких как предлоги, союзы, междометия и т.д.), символов и знаков препинания. Для избавления от ненужной нам для обработки информации был создан соответствующий словарь из слов и символов для его использования в процессе токенизации.

Листинг кода самой функции токенизации представлен ниже:

```
def tokenize(file_text):
    # применение функции токенизации библиотеки NLTK
    tokens = nltk.word_tokenize(file_text)

    # удаление символов пунктуации
    tokens = [i for i in tokens if ( i not in string.punctuation )]

    # удаление ненужных слов
    stop_words = stopwords.words('russian')
    stop_words.extend(exsessiveWOrds)
    tokens = [i for i in tokens if ( i not in stop_words )]

    # очищение токенов
    tokens = [i.replace("«", "").replace("»", "") for i in tokens]

    return tokens
```

Так как количество найденных в документе токенов напрямую зависит от размера самого документа, полученные структуры документов являются

непостоянными и не могут храниться в виде таблицы. В качестве решения для хранения полученной информации был выбран формат JSON.

3.2 Реализация наивного байесовского алгоритма

При создании классификаторов полнотекстовых документов можно увидеть довольно неблагоприятную статистику, говорящую о том, что с ростом размера самих документов процесс классификации заметно усложняется и его точность неизменно падает [15]. В соответствии с этими данными было принято решение внести корректировки и в текущий процесс обучения. В силу наблюдаемой зависимости успешности обучения от размера имело место расширение списка запрещенных слов. Блок-схема улучшенного алгоритма представлена на рисунке 10.

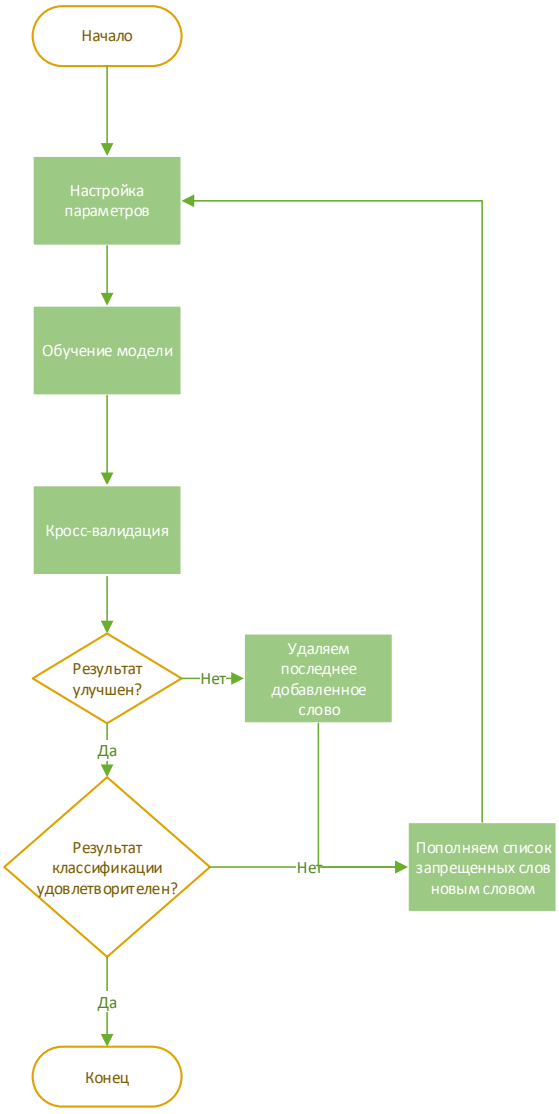

Рисунок 10 – Блок-схема улучшенного НБА

В нашем случае для создания и обучения модели использовался язык программирования Python и библиотека Scikit-Learn, позволяющая максимально упростить этот процесс.

В качестве данных для обучения были использованы данные полученные после токенизации документов в виде самих токенов и количества их вхождений в документе.

Листинг кода метода настройки параметров для обучения представлен ниже:

```
def train(samples):
    classes, freq = defaultdict(lambda:0), defaultdict(lambda:0)
     # подсчет частот
    for feats, label in samples:
        classes[label] += 1

        for feat in feats:
            freq[label, feat] += 1

     # нормализация частот
    for label, feat in freq:
        freq[label, feat] /= classes[label]

    for c in classes:
        classes[c] /= len(samples)

    return classes, freq
```

Для проверки работы модели использовался метод кросс-валидации. Данный метод подразумевает разбиение исходной выборки на две части: обучающую и контрольную. Для проверки результатов обучения было принято решение использовать приблизительно 10% данных всей выборки, остальные же данные использовались для обучения модели. Полученные в результате обучения результаты представлены в главе 5.

4 ОБУЧЕНИЕ МОДЕЛИ С ПОМОЩЬЮ ИНС

Искусственные нейронные сети, используя способность обучения на множестве примеров, способны решать задачи, в которых неизвестны закономерности развития ситуации и зависимости между входными и выходными данными [16]. Данное свойство ИНС во много обуславливает их использование применимо к нашему случаю.

В силу того, что ИНС получают на вход вектор параметров для обработки и работают только с численными значениями, на начальном этапе необходимо подобрать подходящие для обработки параметры. В качестве входных параметров для ИНС было решено использовать как полученные значения из обработанных файлов, так и синтетически выведенные значения:

- Запланированная дата начала выполнения;
- Запланированная дата окончания выполнения;
- Запланированная продолжительность проекта;
- Запланированная стоимость проекта;
- Фактическая дата начала выполнения;
- Фактическая дата окончания выполнения;
- Фактическая продолжительность проекта;
- Фактическая стоимость проекта;
- Количество уже имеющихся выполненных проектов у заказчика;
- Процент успешно завершенных проектов у заказчика;
- Количество уже имеющихся выполненных проектов у поставщика;
- Процент успешно завершенных проектов у поставщика;
- Суммарное отклонение по стоимости проектов поставщика;
- Суммарное отклонение по длительности проектов поставщика.

В итоге можно заметить, что входной слой ИНС будем получать на вход вектор, состоящий из 14 параметров. Так как даты не являются численными параметрами, то было решено заменить их значения, на номер дня в году, который соответствует каждой дате. На основании полученных параметров

разрабатываемая модель должна быть обучена с целью расчета вероятности успешного выполнения проекта с высокой точностью.

В качестве методов обучения ИНС были выбраны одни из наиболее популярных и эффективных методов обучения с учителем:

- Метод обратного распространения ошибки;
- Метод с использованием эволюционного алгоритма.

4.1 ИНС с методом обратного распространения ошибки

Как и в случае с большинством нейронных сетей, наша цель состоит в обучении сети таким образом, чтобы достичь баланса между способностью сети давать верный отклик на входные данные, использовавшиеся в процессе обучения (запоминания), и способностью выдавать правильные результаты в ответ на входные данные, схожие, но неидентичные тем, что были использованы при обучении.

Обучение сети методом обратного распространения ошибки включает в себя три этапа:

1. Подача на вход данных, с последующим распространением данных в направлении выходов;

2. Вычисление и обратное распространение соответствующей ошибки;

3. Корректировка весов.

После обучения предполагается лишь подача на вход сети данных и распространение их в направлении выходов. При этом если обучение сети может являться довольно длительным процессом, то непосредственное вычисление результатов обученной сетью происходит очень быстро [17].

После изучения структуры вектора входных сигналов была подобрана наиболее удачная структура ИНС, представленная на рисунке 11.

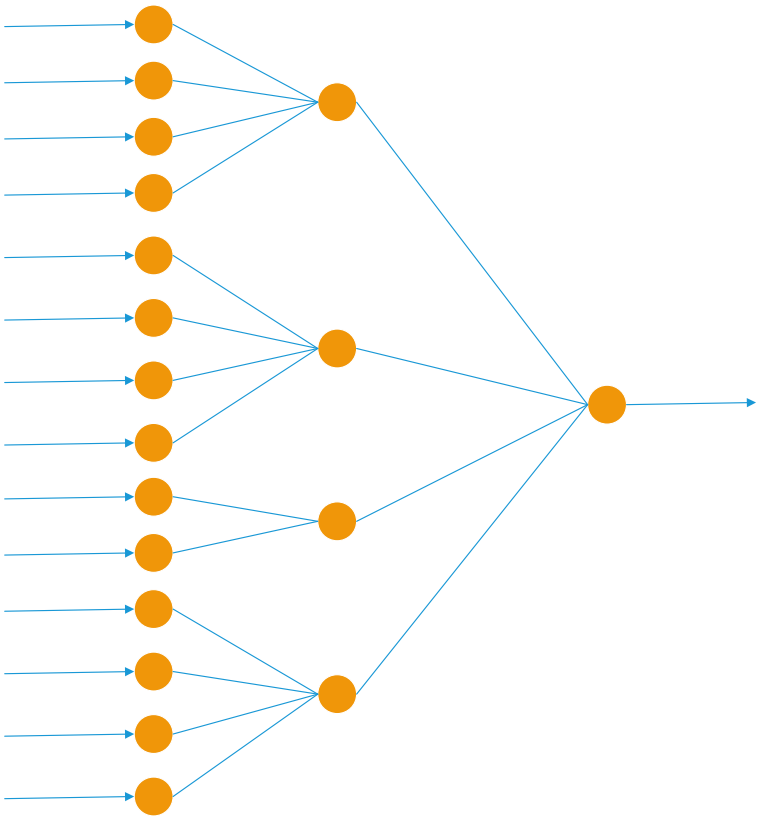

Рисунок 11 – Строение ИНС первого типа

Использование данной структуры приводит к разделению показателей на четыре подгруппы на скрытом уровне ИНС:

- Запланированные временные показатели;
- Фактические временные показатели;
- Показатели работы заказчика;
- Показатели работы поставщика.

Реализация и обучение ИНС происходило на языке Python с помощью класса BackpropTrainer из библиотеки PyBrain [18]. Пример листинга кода метода обучения представлен ниже:

```
def train(self):
    from pybrain.tools.shortcuts import buildNetwork
    from pybrain.structure import SigmoidLayer
    FNN = buildNetwork(DS.indim, 20, DS.outdim, bias=True,
recurrent=True, hiddenclass=SigmoidLayer)
    from pybrain.supervised.trainers import BackpropTrainer
    TRAINER = BackpropTrainer(FNN, dataset=DS, learningrate = 0.0001,
        momentum=0.1, verbose=True)
    for i in range(max):
        print "Test " + str(i)
        TRAINER.train()
```

Проверка результатов обучения также происходила с помощью метода кросс-валидации. Так как на выходе ИНС выдавалось число, которое интерпретировалось как вероятность успешного завершения проекта, и его значения находились в диапазоне [0, 1], было принято решение, считать, что проект считается удачным в случае, если значение выходного сигнала превышает 0,9. Результаты апробации полученной модели представлены в главе 5.

4.2 ИНС с использованием эволюционного алгоритма

Метод построения искусственной нейронной сети с помощью эволюционного алгоритма подразумевает наличие популяции хромосом, отвечающей за значения синаптических связей в ИНС, а также наличие одной главной хромосомы, которая содержит в себе данные о структуре ИНС.

Применение данного подхода дает нам следующие преимущества при обучении ИНС:

- Независимость от структуры ИНС и характеристик функций активации нейронов;
- Возможность автоматического поиска топологии ИНС и получения более точной нейросетевой модели.

Однако данный подход также обладает и некоторыми недостатками, например:

- Сложность точной настройки весов связей на поздних этапах эволюционного поиска;
- Сложности организации поиска топологии ИНС [19].

Для упрощения работы данного алгоритма и уменьшения времени обучения модели были введены следующие ограничения:

- Число скрытых нейронов не более 10;
- Количество связей не более 30;
- Количество скрытых слоев ИНС не более 3.

Реализация ИНС с помощью эволюционной настройки была сделана на языке программирования C#. Были реализованы все этапы эволюционного алгоритма: селекция, скрещивание и мутация, – а также классы нейросети, нейрона, популяции и двух видов хромосом. Пример листинга класса хромосомы представлен ниже:

```
public class Chromosome
{
    public List<double> Weights { get; set; }
    public double Result { get; set; }
    public double Error { get; set; }

    public Tuple<Chromosome, Chromosome> Crossover(Chromosome parent2)
    {
        var rnd = new Random();
        var point = rnd.Next(4);

        var child1 = InnerCrossover(this, parent2, point);
        var child2 = InnerCrossover(parent2, this, 3 - point);

        return new Tuple<Chromosome, Chromosome>(child1, child2);
    }

    private Chromosome InnerCrossover(Chromosome parent1, Chromosome
parent2, int pointToCross);
}
```

После выполнения всех выше указанных процедур и обучения ИНС, ее структура имела вид, представленный на рисунке 12.

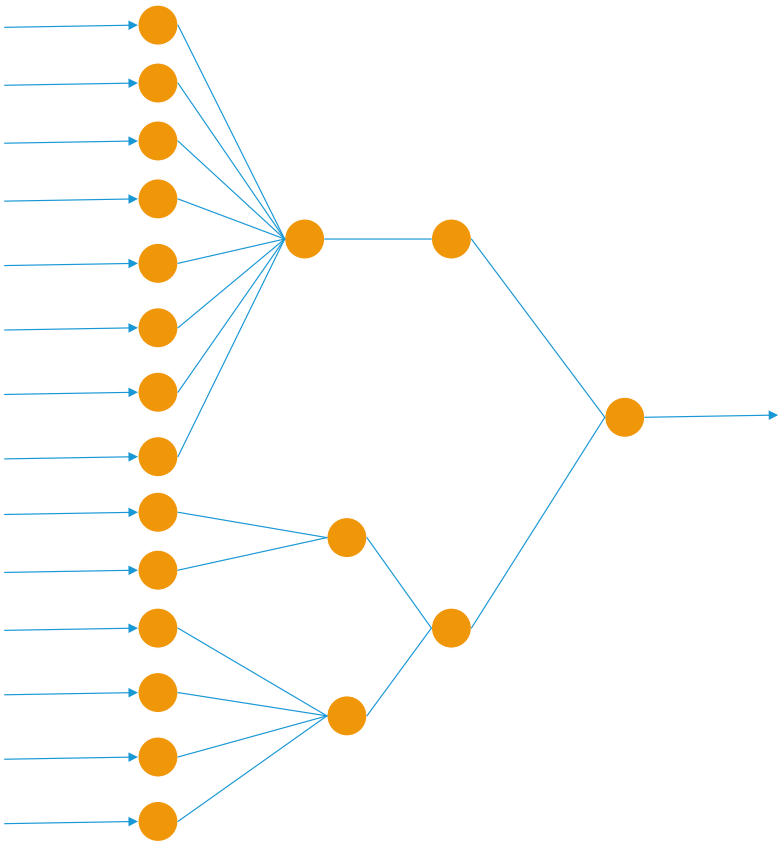

Рисунок 12 – Структура ИНС второго типа

Проверка работы модели проводилась по той же методике, что и для ИНС, обученной методом обратного распространения ошибки. Результаты апробации полученной модели представлены в главе 5.

5 АПРОБИРОВАНИЕ РАЗРАБОТАННЫХ МЕТОДОВ

5.1 Анализ результатов выгрузки данных

После завершения выгрузки данных с FTP-сервера портала Федеральной контрактной системы количество выгруженных файлов составило более 500 000 элементов.

Однако в результате обработки файлов около 70% из них были признаны неполными в силу отсутствия в них каких-либо важных данных, либо ввиду некорректности этих данных. В результате после завершения процесса заполнения БД данными количество записей в БД составило:

- Более 110 000 записей в таблице договоров;
- Почти 25 000 записей в таблице организаций.

Но в дальнейшем неполнота данных и сложная структура их хранения сыграли свою роль, и в итоге среди всех записей в таблице договоров менее половины имели полностью заполненные поля для проведения аналитики (приблизительно 42 000 записей). А количество записей, для которых были найдены соответствующие тексты договоров, не превысило и 10% от всего количества и составило около 6 000 записей.

Рисунок 13 – Статистика выгруженных данных

Полученные данные могут сказать нам о том, что несмотря на крупные вложения государственных средств в виде денежных и трудовых затрат в разработку портала Федеральной контрактной системы, в его работе

наблюдаются довольно большое количество нарушений, связанных с некорректным или неполным заполнением данных и о проектах и процессе их выполнения.

5.2 Анализ результатов обучения модели

Для проверки работы всех обученных моделей использовался метод кросс-валидации. Данный метод подразумевает разделение выборки на две подвыборки: обучающую и контрольную. Для описанных методов обучения были выбраны следующие значения:

- Для НБА в качестве контрольной выборки использовались 10% всех данных;
- Для ИНС в качестве контрольной выборки использовались 5% всех данных.

При рассмотрении результатов для НБА после выгрузки у нас в наличии имелся 6141 текстовый документ договора. Таким образом, для оценки результатов обучения использовались 500 документов в качестве контрольной выборки, остальные же использовались для обучения модели. В качестве порога для успешности проекта было выбрано значение вероятности равное 0,8. Использование данного значения для оценки результата может быть обосновано использованием НБА для обработки текстовых данных, характеризующихся большими объемами и низкой вероятностью повторения слов, что приводит к понижению точности работы классификации.

В результате проверки полученных результатов точность работы полученной модели составила 79%, а ее полнота – 83%, соответственно результаты выполнения 79% проектов были предсказаны точно, и из всего количества успешных проектов были предсказаны 83%. Полученные результаты показывают высокую эффективность работы классификатора и позволяют нам говорить о том, что модель показывает довольно высокую точность результатов, однако может быть улучшена еще больше в процессе дальнейшей модернизации.

Проверка результатов обучения ИНС также происходила с помощью метода кросс-валидации. Так как на выходе ИНС выдавалось число, которое интерпретировалось как вероятность успешного завершения проекта, и его значения находились в диапазоне [0, 1], было принято решение, считать, что проект считается удачным в случае, если значение выходного сигнала превышает 0,9.

В результате обучения ИНС с помощью метода обратного распространения ошибки точность выходного сигнала составила 93%, а его полнота – 96,5%, что может свидетельствовать о довольно точной работе ИНС и ее значительным преимуществом над моделью, обученной с помощью НБА.

В результате обучения ИНС с настройкой структуры и синаптических весов с помощью эволюционного алгоритма точность выходного сигнала и его полнота составили соответственно 87% и 90%, что говорит нам о значительном превосходстве ИНС, обученной с помощью метода обратного распространения ошибки, и подтверждает один из главных недостатков ИНС с использованием ЭА в отсутствии возможности более точного подбора значений синаптических весов.

Гистограмма с результатами работы моделей представлена на рисунке 14.

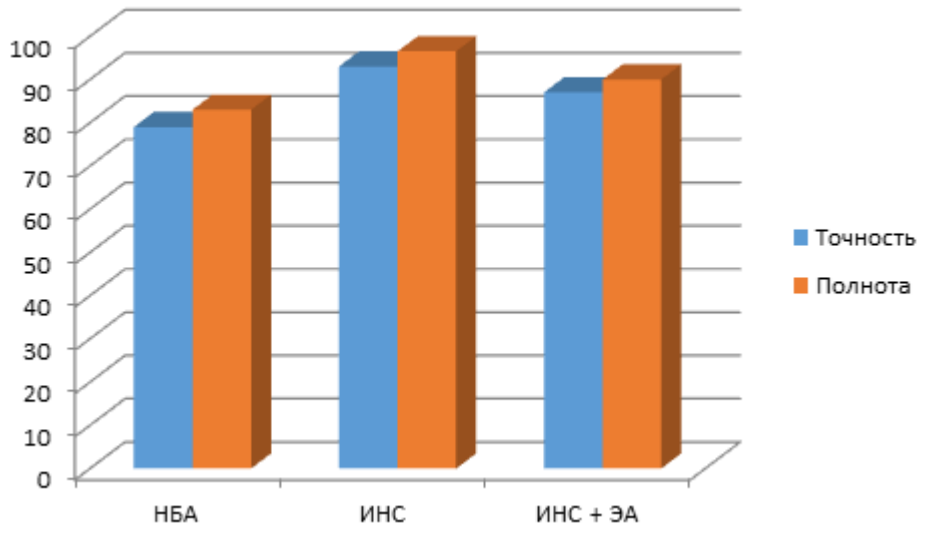

Рисунок 14 – Результаты работы разработаны моделей

Полученные результаты свидетельствуют о полном превосходстве обучения модели с помощью ИНС методом обратного распространения ошибки.

Полученные результаты обучения модели свидетельствуют о высокой точности ее работы и способности предсказывать результаты исполнения проектов.

5.3 Перспективы использования результатов

Разработанная модель в дальнейшем может быть использована в нескольких направлениях с целью извлечения выгоды из ее работы. К таким направления относятся:

- Разработка специализированного ПО с целью его коммерческой реализации;
- Использование модели для первичной оценки проектов в Федеральной контрактной системе;
- Использование модели организациями в собственных целях с использованием дополнительного обучения на результатах выполнения собственных проектов с целью более точной ее настройки под корпоративные нужды.

В качестве показателя полезности возможного использования разработанной модели в будущем может послужить тот факт, что в случае ее использования при работе с обработанными в процессе обучения проектами в прошлом могли быть сэкономлены до полумиллиарда рублей из федеральных источников.

6 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

При разработке программного обеспечения важно провести экономическое обоснование создания программной разработки, а именно: изучить экономическую выгодность разрабатываемого продукта, выявить преимущества и недостатки разработки, провести анализ и рассчитать экономические показатели создаваемого проекта, оценить затраты на проект и его результаты. Провести такого рода анализ необходимо самому разработчику для понимания того, что стоит ждать от проекта, какие перспективы у данной разработки.

Работа по технико-экономическому обоснованию в процессе проектирования преследует одно из основных требований – это подтвердить техническую и экономическую целесообразность реализации разработки, для которой сформирован проект.

Целью выполнения данного раздела является анализ эффективности создания предиктивной модели для оценки вероятности успешного выполнения проекта на основе его начальных показателей. В данном разделе выявлены потенциальные потребители результатов исследования, произведено планирование научно-исследовательских работ, сформирован бюджет затрат научно-исследовательского проекта.

6.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

6.1.1 Потенциальные потребители результатов исследования

Анализ потенциальных потребителей результатов исследования необходим для оценки предпочтений целевой аудитории в отношении конкретной технологии или программного продукта. Такого рода исследование проводится для определения нужности конкретным пользователям данной разработки.

Потенциальными потребителями данного разработанной предиктивной модели для оценки вероятности успешного выполнения проекта на основе его

начальных показателей являются организации, занимающиеся выполнением проектных работ как для коммерческих, так и для некоммерческих организаций. Разработанная модель позволит пользователям провести более точный анализ возможного проекта, оценить его риски и подобрать наиболее выгодные условия для выполнения с целью повышения вероятности успешного выполнения работ. Таким образом, применение данного программного продукта позволит заметно упростить и улучшить аналитическую часть работы с проектом, тем самым увеличивая прибыль предприятий.

Целевым рынком для данной разработки является рынок предприятий, направленных на проектную деятельность в любой сфере производства.

Таким образом, основным критерием сегментации является специализация потенциального потребителя, а также размер предприятия.

Сегментация целевого рынка для данной разработки по виду потребителей:

- предприятия, занимающиеся выполнением проектов для коммерческих организаций;
- предприятия, занимающиеся выполнением проектов для некоммерческих организаций.

Сегментация потребителей по размеру:

- крупные предприятия;
- средние предприятия;
- малые предприятия.

Карта сегментации рынка на основании наиболее значимых критериев для рынка представлена в таблице 1.

Таблица 1 – Карта сегментирования

		Размер предприятия		
		Крупные	**Средние**	**Малые**
Специализация предприятий	Выполнение проектов по строительству, разработке ПО, изготовлению и доставке оборудования, мебели и одежды, для коммерческих организаций			
	Выполнение проектов по строительству, разработке ПО, изготовлению и доставке оборудования, мебели и одежды, для некоммерческих организаций			

Исходя из вышеприведенных данных, можно сделать выводы, определяющие результаты сегментирования рынка:

- сегменты, на которые необходимо ориентироваться: разработка программного продукта для крупных и средних по размеру предприятий;
- сегменты рынка, которые могут быть привлекательны для развития разработки в будущем: адаптация приложения для малых по размеру предприятий.

6.1.2 Диаграмма Исикавы

Диаграмма Исикавы, отображающая причинно-следственные связи приведена на рисунке 15. Созданная диаграмма позволяет выявить факторы и причины, характеризующие необходимость в разработанной модели.

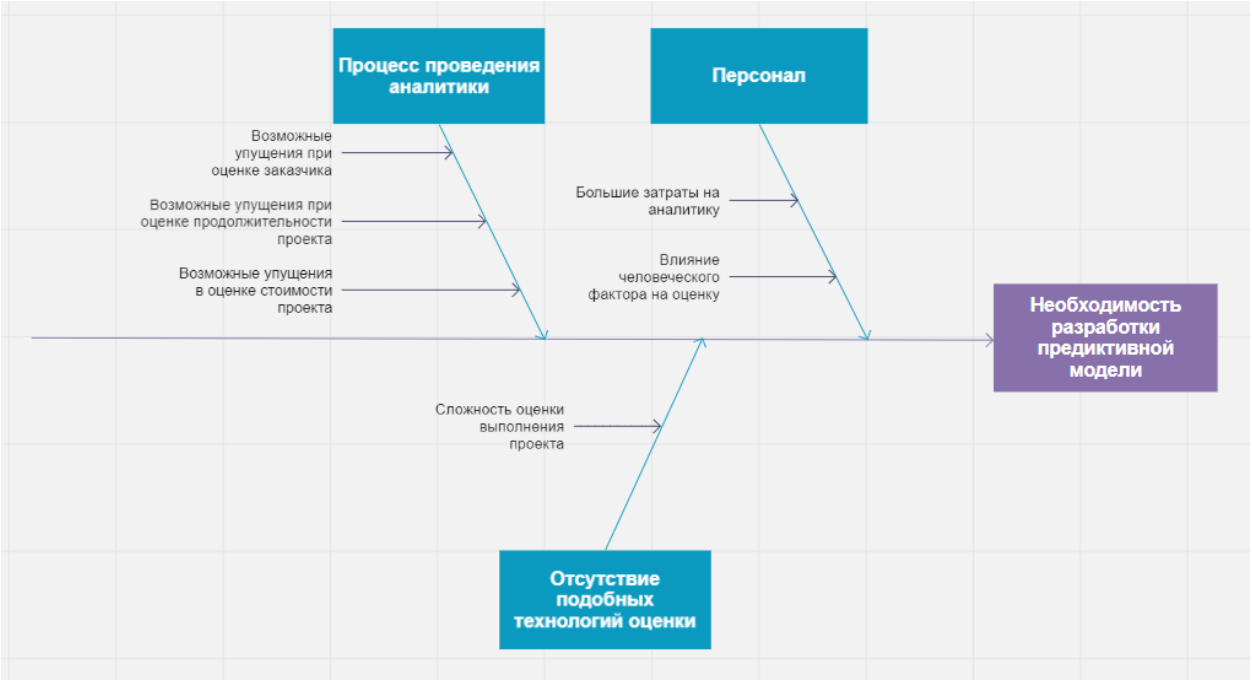

Рисунок 15 – Диаграмма Исикавы

6.1.3 SWOT-анализ

В ходе проведения SWOT-анализа была составлена итоговая матрица (таблица 2), содержащая описание сильных и слабых сторон проекта, выявление возможностей и угроз, а также их корреляцию.

Таблица 2 – Матрица SWOT-анализа

	Сильные стороны научно-исследовательского проекта: С1. Высокая точность оценки. С2. Высокая скорость обработки данных. С3. Возможность автоматизированной оценки без влияния человеческого фактора. С4. Низкие требования к используемым ресурсам.	**Слабые стороны научно-исследовательского проекта:** Сл1. Отсутствие выборки данных по коммерческих организация для обучения модели системы. Сл2. Сложность внедрения на рынке. Сл3. Сложность изменения существующего решения.
Возможности: В1. Появление дополнительного спроса на новый продукт. В2. Повышение точности	В1В2С1 – привлечение целевой аудитории за счёт высокой точности работы модели.	В1В4Сл1 – необходимость поиска дополнительных данные для обучения модели.

работы модели. В3. Повышение уровня благосостояния пользователей. В4. Повышение интереса организацию к прогнозированию результатов работ.	В1В4С3 – повышение заинтересованности в проекте со стороны организаций, занимающихся проектной деятельностью.	
Угрозы: У1. Отсутствие спроса на разработанный продукт. У2. Появление конкурентов с улучшенными технологиями. У3. DDoS-атаки на сервера, с которыми работает модель. У4. Закрытие портала с открытыми данными по проектам.	У2С1 – появление конкурентов с улучшенными технологиями может обнулить преимущества разработки. У1С1С2С3 – отсутствие спроса на продукт может обнулить преимущества разработки.	У1Сл2 – в случае прекращения исследований по данной теме, разработка не сможет раскрыть своего потенциала и улучшить показатели. У3Сл3 – замедление работы модели, которое может привести к падению спроса у потребителей. У4Сл3 – падение точности и спроса на модель.

Таким образом, можно сделать вывод, что проект необходимо развивать в направлении наибольшей универсальности относительно выборки данных по типам проектов, а также улучшения методов обработки данных. При этом следует внимательно следить и по возможности применять новейшие разработки в общей теории алгоритмов классификации данных.

6.2 Определение возможных альтернатив проведения научных исследований

В данном разделе описаны методы, которые позволяют выявить и предложить возможные альтернативы проведения исследования и доработки результатов. Морфологическая матрица для предиктивной модели представлена в таблице 3.

Таблица 3 – Морфологическая матрица для предиктивной модели

	1	2	3
А. СУБД	MS SQL Server	Oracle	MySQL
Б. Язык программирования	C++	Python	C#
В. Метод обучения модели	Искусственная нейронная сеть	Наивный байесовский алгоритм	ИНС + ЭА

Наиболее желательное функциональное решение – А1Б3В1. Возможные варианты решения технической задачи – А1Б2В2, А1Б2В3.

6.3 Планирование научно-исследовательских работ

6.3.1Структура работ в рамках научного исследования

На начальном этапе создания проекта необходимо провести планирование научно-исследовательских работ. Планирование комплекса предполагаемых работ включает в себя определение структуры работ в рамках научного исследования, определение участников каждой работы, установление продолжительности работ, построение графика проведения научных исследований.

В данном разделе составлен перечень этапов и работ в рамках проведения научного исследования, проведено распределение исполнителей по видам работ. Порядок этапов и работ, распределение исполнителей по данным видам работ приведен в таблице 4.

Таблица 4 – Перечень этапов и работ, распределение исполнителей

Основные этапы	№ раб	Содержание работ	Должность исполнителя
Постановка задачи	1	Определение цели и задач создания разработки	Исполнитель Руководитель
Анализ предметной области	2	Изучение функциональных возможностей предиктивных моделей	Исполнитель
	3	Изучение аналогов модели	

	4	Написание технического задания (ТЗ)	Исполнитель Руководитель
Проектирование	5	Проектирование архитектуры модели	Исполнитель Руководитель
	6	Проектирование модификаций для НБА	
	7	Проектирование искусственной нейронной сети	
Программная реализация	8	Разработка необходимого функционала модели	Исполнитель Руководитель
Тестирование	9	Тестирование модели	Исполнитель Руководитель
	10	Внесение изменений	
Подготовка документации	11	Оформление документации	Исполнитель Руководитель
	12	Утверждение документации	

6.3.2Определение трудоемкости работ

В данном разделе необходимо определить трудоемкость выполнения работ. Трудовые затраты в большинстве случаев образуют основную часть стоимости разработки, поэтому являются важным моментом. Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, т.к. зависит от множества трудно учитываемых факторов. Для определения ожидаемого значения трудоемкости $t_{ож}$ используется следующая формула:

$$t_{ожс} = \frac{3 \cdot t_{\min} + 2 \cdot t_{\max}}{5},$$

где t_{min} – минимальная трудоемкость i-ой работы, чел/дн.; t_{max} – максимальная трудоемкость i-ой работы, чел/дн.

Результат расчёта трудоёмкости работ представлен в таблице 5.

Таблица 5 – Трудоемкость выполнения работ

Этап	**Продолжительность, дни**								
	t_{min}			**t_{max}**			**$t_{ож}$**		
	Исп1	**Исп2**	**Исп3**	**Исп1**	**Исп2**	**Исп3**	**Исп1**	**Исп2**	**Исп3**
Постановка задачи	2	3	3	4	5	5	2,8	3,8	3,8
Изучение функциональных возможностей модели	3	3	3	5	5	5	3,8	3,8	3,8

Обзор аналогов системы	6	6	6	10	10	10	7,6	7,6	7,6
Написание технического задания (ТЗ)	2	3	3	3	4	4	2,4	3,4	3,4
Проектирование архитектуры модели	2	3	3	3	4	4	2,4	3,4	3,4
Проектирование модификаций для НБА	2	3	3	3	4	4	2,4	3,4	3,4
Проектирование искусственной нейронной сети	2	2	2	5	5	5	3,2	3,2	3,2
Программная реализация	60	80	70	70	90	80	64	84	74
Тестирование	3	3	3	4	5	5	3,4	3,8	3,8
Оформление документации	14	14	14	17	17	17	15,2	15,2	15,2
Итого	96	120	110	124	149	139	107,2	131,2	121,6

6.3.3Разработка графика проведения научного исследования

При разработке программного обеспечения необходимым является составление плана проведения работ с датами начала каждого этапа и продолжительностью этапов. Данная информация представлена в таблице 6.

Таблица 6 – Длительность этапов разработки

Основные этапы		**Дата начала**	**Длительность, дни**
Постановка задачи		08.01.18	3
Анализ предметной области	Изучение функциональных возможностей модели	11.01.18	5
	Обзор аналогов модели	18.01.18	10
	Написание технического задания (ТЗ)	01.02.18	2
Проектирование	Проектирование архитектуры модели	05.02.18	3
	Проектирование модификаций для НБА	08.02.18	3
	Проектирование искусственной нейронной сети	13.02.18	4

Программная реализация	17.02.18	60
Тестирование	17.04.18	3
Оформление документации	23.04.18	21

Для удобства построения графика, длительность каждого из этапов работ из рабочих дней необходимо перевести в календарные дни. Для этого необходимо воспользоваться следующей формулой:

$$T_{ki} = T_{pi} \cdot k_{кал}$$

где T_{ki}– продолжительность выполнения i-й работы в календарных днях; T_{pi} – продолжительность выполнения i-й работы в рабочих днях; $k_{кал}$– коэффициент календарности.

Коэффициент календарности определяется по следующей формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}}$$

где $T_{кал}$ – количество календарных дней в году; $T_{кал}$ = 365. $T_{вых}$ – количество выходных дней в году; $T_{вых}$ = 118. $T_{пр}$ – количество праздничных дней в году; $T_{пр}$ = 14.

Все рассчитанные значения представлены в таблице 7.

Таблица 7 – Длительность работ в рабочих и календарных днях

Основные этапы		**Длительность работ в рабочих днях, Tpi**			**Длительность работ в календарных днях, Tki**		
		Исп1	**Исп2**	**Исп3**	**Исп1**	**Исп2**	**Исп3**
Постановка задачи		1,4	1,9	1,9	2,1	2,85	2,85
Анализ предметной области	Изучение функциональных возможностей модели	3,8	3,8	3,8	5,7	5,7	5,7
	Обзор аналогов модели	7,6	7,6	7,6	11,4	11,4	11,4
	Написание технического задания (ТЗ)	1,2	1,7	1,7	1,8	2,55	2,55
Проектирование	Проектирование архитектуры модели	1,2	1,7	1,7	1,8	2,55	2,55
	Проектирование	1,2	1,7	1,7	1,8	2,55	2,55

	модификаций для НБА						
	Проектирование искусственной нейронной сети	1,6	1,6	1,6	2,4	2,4	2,4
Программная реализация		32	42	37	48	63	55,5
Тестирование		1,7	1,9	1,9	2,55	2,85	2,85
Оформление документации		7,6	7,6	7,6	11,4	11,4	11,4
Итого		59,3	71,5	66,5	88,95	107,25	99,75

6.3.4 Бюджет научно-технического исследования (НТИ)

Состав бюджета выполнения работ по разработке предиктивной модели включает в себя стоимость всех расходов, необходимых для их выполнения. При формировании бюджета используется группировка затрат по следующим статьям:

- заработная плата;
- отчисления во внебюджетные фонды.

6.3.4.1 Основная заработная плата исполнителей

Данная статья расходов включает заработную плату двух исполнителей. Расчет основной заработной платы выполняется на основе трудоёмкости выполнения каждого этапа и величины месячного оклада исполнителя.

Основная заработная плата (**$З_{осн}$**) рассчитывается по следующей формуле

$$З_{осн} = З_{дн} \cdot T_p,$$

где $З_{осн}$ – основная заработная плата одного работника; $T_р$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.; $З_{дн}$ – среднедневная заработная плата работника, руб.

Для расчета среднедневной заработной платы необходимо воспользоваться формулой:

$$З_{дн} = \frac{З_м \cdot М}{F_д},$$

где $З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года: при отпуске в 24 раб. дня М =11,2 месяца, 5-дневная неделя; при отпуске в 48 раб. дней М=10,4 месяца, 6-дневная неделя; Fд – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн.

Месячный должностной оклад работника рассчитывается по формуле:

$$З_м = З_{тс} \cdot (1 + k_{пр} + k_д) \cdot k_р,$$

где $З_{тс}$ – заработная плата по тарифной ставке, руб.; $k_{пр}$ – премиальный коэффициент, равный 0,3 (т.е. 30% от $З_{тс}$); $k_д$ – коэффициент доплат и надбавок составляет примерно 0,2 – 0,5; $k_р$ – районный коэффициент, равный 1,3 (для Томска).

Рассчитанные значения представлены в таблице 8.

Таблица 8 – Основная заработная плата исполнителей

Исполнители	**$З_{тс}$, руб**	**$k_р$**	**$З_{дн}$, руб**	**$Т_{раб}$, дн**			**$З_{осн}$, руб**		
				Исп1	**Исп2**	**Исп3**	**Исп1**	**Исп2**	**Исп3**
Руководитель	27500	1,3	1247	21	23	23	26187	28681	28681
Исполнитель	9489	1,3	400	107	131	121	42800	52400	48400
Итого							68987	81081	77081

6.3.4.2 Дополнительная заработная плата исполнителей

Расчет дополнительной заработной платы ведется по следующей формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (0,12 – 0,15).

В таблице 9 представлены результаты расчёта дополнительной заработной платы.

Таблица 9 – Дополнительная заработная плата исполнителей

Исполнители	**Основная заработная плата, руб.**			**Коэффициент дополнительной заработной платы**	**Дополнительная заработная плата, руб.**		
Руководитель	26187	28681	28681	0,12	3142,5	3441,7	3441,7
Исполнитель	42800	52400	48400		5136	6288	5808
Итого					8278,5	9729,7	9249,7

6.3.4.3 Отчисления во внебюджетные фонды

Величина отчислений во внебюджетные фонды определяется исходя из формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}),$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды.

В таблице 10 представлены результаты расчёта отчислений во внебюджетные фонды.

Таблица 10 – Отчисления во внебюджетные фонды

Исполнитель	**Основная заработная плата, руб.**			**Дополнительная заработная плата, руб.**		
	Исп.1	Исп.2	Исп.3	Исп.1	Исп.2	Исп.3
Руководитель проекта	26187	28681	28681	3142,5	3441,7	3441,7
Исполнитель	42800	52400	48400	5136	6288	5808
$k_{внеб}$	0,3					
Итого						
Исполнение 1	23179,5					
Исполнение 2	27243,2					
Исполнение 3	25899,2					

6.3.4.4 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанная величина затрат научно-исследовательской работы является основой для формирования бюджета затрат проекта, который при

формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

Определение бюджета затрат на научно-исследовательский проект по каждому варианту исполнения приведен в таблице 11.

Таблица 11 – Расчёт бюджета затрат НТИ

Наименование статьи	Сумма, руб		
	Исп1	Исп2	Исп3
Затраты по основной заработной плате исполнителей темы	68987	81081	77081
Затраты по дополнительной заработной плате исполнителей темы	8278,5	9729,7	9249,7
Отчисления во внебюджетные фонды	23179,5	27243,2	25899,2
Бюджет затрат НТИ	100445	118054	112230

6.4 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

Интегральный финансовый показатель разработки определяется как:

$$I_{\text{финр}}^{\text{исп.i}} = \frac{\Phi_{\text{pi}}}{\Phi_{\text{max}}},$$

где $I_{\text{финр}}$ – интегральный финансовый показатель разработки; Φ_{pi} – стоимость i-го варианта исполнения; Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта.

Интегральные финансовые показатели разработки для вариантов исполнения:

$I_{\text{финр1}} = 100445 / 118054 = 0.85$

$I_{\text{финр2}} = 118054 / 118054 = 1$

$I_{\text{финр3}} = 112230 / 118054 = 0.95$

В таблице 12 представлена сравнительная оценка характеристик вариантов исполнения.

Таблица 12 – Сравнительная оценка характеристик вариантов исполнения проекта

Объект исследования Критерии	Весовой коэффициент параметра	Исп.1	Исп.2	Исп.3
Надежность	0,2	4	4	4
Рациональное использование ресурсов	0,1	5	3	3
Удобство интерфейса	0,3	4	4	3
Возможность использования на мобильных устройствах	0,4	5	5	5
ИТОГО	1			

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum a_i \cdot b_i ,$$

где I_{pi} – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки; a_i – – весовой коэффициент i-го варианта исполнения разработки; b_i – бальная оценка i-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания;

$I_{р\text{-}исп1}$ = 4*0,2 + 5*0,1 + 4*0,3 + 5*0,4 = 4,5;

$I_{р\text{-}исп2}$ = 4*0,2 + 3*0,1 + 4*0,3 + 5*0,4 = 4,3;

$I_{р\text{-}исп3}$ = 4*0,2 + 3*0,1 + 3*0,3 + 5*0,4 = 4;

Интегральные показатели эффективности вариантов исполнения разработки определяются на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{исп.1} = \frac{I_{р-исп1}}{I_{финр}^{исп.1}}, \quad I_{исп.2} = \frac{I_{р-исп2}}{I_{финр}^{исп.2}}$$

$I_{исп1}$ = 4,5 / 0,85 = 5,29

$I_{исп2}$ = 4,3 / 1 = 4,3

$I_{исп3}$ = 4,0 / 0,95 = 4,21

Сравнение интегрального показателя эффективности вариантов исполнения разработки позволит определить сравнительную эффективность проекта и выбрать наиболее целесообразный вариант из предложенных.

Сравнительная эффективность проекта:

$$Э_{ср} = \frac{I_{исп.1}}{I_{исп.2}}$$

$Э_{ср1}$ = 5,29 / 5,29 = 0,93

$Э_{ср2}$ = 4,3 / 5,29 = 0,81

$Э_{ср3}$ = 4,21 / 5,29 = 0,796

В таблице 13 представлена сравнительная эффективность разработки.

Таблица 13 – Сравнительная эффективность разработки

№ п/п	Показатели	Исп.1	Исп.2	Исп.3
1	Интегральный финансовый показатель разработки	0,85	1	0,95
2	Интегральный показатель ресурсоэффективности разработки	4,5	4,3	4
3	Интегральный показатель эффективности	5,29	4,3	4,21
4	Сравнительная эффективность вариантов исполнения	0,93	0,81	0,796

Таким образом, исходя из полученных результатов, можно сделать вывод, что более эффективным вариантом решения поставленной в магистерской диссертации технической задачи с позиции финансовой и ресурсной эффективности является 1 вариант.

7 Социальная ответственность

В данном разделе рассмотрены вопросы обеспечения производственной и экологической безопасности выполняемых работ, а также безопасности в чрезвычайных ситуациях и организационные вопросы обеспечения безопасности.

Во время разработки предиктивной модели вычисления вероятности успешного завершения проекта на основе начальных его показателей выполнялись работы, связанные со сбором, анализом и структуризацией требований, проектированием архитектуры и интерфейсов приложения, и реализацией. Весь объем представленных работ непосредственно связан с вычислительной техникой: персональным компьютером, периферийными устройствами, устройствами ввода и вывода информации. Данное взаимодействие соответственно связывает человека с дополнительным, вредным воздействием группы факторов. В таких условиях необходимым является снижение неблагоприятного воздействия вредных факторов, присутствующих при работе с вычислительной техникой.

7.1 Производственная безопасность

Был проведен анализ вредных и опасных факторов, которые могут возникать при разработке предиктивной модели.

Перечень опасных и вредных факторов, характерных для проектируемой производственной среды, представлен в таблице 14.

Таблица 14 – Опасные и вредные факторы при реализации предиктивной модели вычисления вероятности успешного завершения проекта на основе начальных его показателей

Источник фактора, наименование видов работ	Факторы (по ГОСТ 12.0.003-74)		Нормативные документы
	Вредные	Опасные	
Разработка	1. Электромагнитное	1.Электрический	СанПиН 2.2.2/2.4.1340-03. Гигиенические

подсистемы	излучение;	ток. 2.Пожаро-опасность	требования к персональным электронно-вычислительным машинам и организации работы [20]. ГОСТ 12.1.038–82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов [21]. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты [22].

7.1.1Анализ выявленных вредных факторов при разработке и эксплуатации проектируемого решения

7.1.1.1 Электромагнитное излучение

Электромагнитное излучение представляет собой электромагнитные волны, возбуждаемые различными излучающими объектами, – заряженными частицами, атомами, молекулами.

Компьютер является одним из наиболее распространенных источников влияния электромагнитных излучений на рабочем месте. В качестве источников излучения компьютер имеет монитор и системный блок. Проблема электромагнитного излучения является достаточно важной, так как пользователь может проводить перед компьютером очень длительное время, а значит и время воздействия электромагнитного поля велико.

Электромагнитные излучения наибольшее влияние оказывают на иммунную, нервную, эндокринную систему.

Согласно СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам (ПЭВМ) и организации работы [20] временные допустимые уровни электромагнитных полей (ЭМП), создаваемых персональными компьютерами, не должны превышать значений, представленных в Таблице 15.

Таблица 15 – Временные допустимые уровни электромагнитного поля, создаваемых персональными компьютерами на рабочих местах

Наименование параметров		ВДУ
Напряженность электрического поля	в диапазоне частот 5 Гц–2 кГц	25 В/м
	в диапазоне частот 2 кГц–400 кГц	2,5 В/м
Плотность магнитного потока	в диапазоне частот 5 Гц–2 кГц	250 нТл
	в диапазоне частот 2 кГц–400 кГц	25 нТл
Электростатический потенциал экрана видеомонитора		500 В

Для снижения негативного влияния электромагнитного излучения от монитора, желательно использовать жидкокристаллический монитор. Также приветствуется использование специальных защитных экранов. К рекомендациям можно отнести то, что монитор должен стоять не ближе, чем на расстоянии вытянутой руки.

Чтобы свести к минимуму негативное влияние электромагнитного излучения от монитора, необходимо придерживаться простых правил:

- Выбирая монитор, лучше отдать предпочтение жидкокристаллическому варианту. Излучение мониторов с электроннолучевой трубкой намного сильнее, чем у жидкокристаллических аналогов;
- постараться расположить монитор в углу. Стены будут поглощать электромагнитное излучение, которое испускают боковые и задние стенки;
- не забывать выключать монитор, если отходите ненадолго от рабочего стола;
- монитор должен стоять от стула не ближе, чем на расстоянии вытянутой руки. Не нужно придвигать его слишком близко к лицу и наклоняться к экрану.

7.1.2Анализ выявленных опасных факторов при разработке и эксплуатации проектируемого решения

7.1.2.1 Электрический ток

В связи с наличием электрооборудования для данного производственного объекта характерным является возможность поражения электрическим током.

Основными причинами поражения человека электрическим током может являться:

- удар электрическим током при использовании неисправного электрооборудования;
- касание неизолированных частей электроустановки (контакты, провода, зажимы).

Для снижения данного риска необходимо соблюдать нормы электробезопасности.

Электробезопасность – это система организационных и технических мероприятий и средств, обеспечивающих защиту людей от вредного и опасного для жизни воздействия электрического тока, электрической дуги, электромагнитного поля и статического электричества.

Предельно допустимые уровни напряжений прикосновения и токов регламентируются ГОСТ 12.1.038–82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов [21]. Вопросы требований к защите от поражения электрическим током освещены в ГОСТ Р 12.1.019-2009 ССБТ [22].

Опасность поражения электрическим током, в отличие от прочих опасностей, усугубляется тем, что человек не в состоянии без специальных приборов обнаружить напряжение дистанционно. Опасность обнаруживается слишком поздно – когда человек уже поражён.

Персональный компьютер питается от сети 220 В переменного тока с частотой 50 Гц. Это напряжение опасно для жизни, поэтому обязательны следующие меры предосторожности:

- перед началом работы нужно убедиться, что выключатели и розетка закреплены и не имеют оголённых токоведущих частей;
- при обнаружении неисправности оборудования и приборов необходимо, не делая никаких самостоятельных исправлений, сообщить человеку, ответственному за оборудование.

К мероприятиям по предотвращению возможности поражения электрическим током следует отнести:

- с целью защиты от поражения электрическим током, возникающим между корпусом приборов и инструментом при пробое сетевого напряжения на корпус, корпуса приборов и инструментов должны быть заземлены;
- при включенном сетевом напряжении работы на задней панели должны быть запрещены;
- все работы по устранению неисправностей должен производить квалифицированный персонал;
- необходимо постоянно следить за исправностью электропроводки;
- не оставлять включенные электрические устройства без наблюдения и не допускать к ним посторонних лиц.

7.1.2.2 Опасность возникновения пожара

В рабочих помещениях с персональными компьютерами повышен риск возникновения пожара. Возможными причинами возникновения пожара может быть неисправность электрооборудования, неправильная их эксплуатация, неудовлетворительный надзор за производственным оборудованием и пожарными устройствами.

Пожарная безопасность включает в себя комплекс организационных и технических мероприятий, направленных на обеспечение безопасности людей, предотвращения пожара, ограничение его распространения, а также создание условий для успешного тушения пожара.

Избежать дополнительной пожарной опасности поможет соблюдение соответствующих мер пожарной профилактики: проверка исправности

электрооборудования, наличия и состояния первичных средств пожаротушения, противопожарного состояния электрооборудования, работоспособности системы вентиляции, состояния эвакуационных выходов, проверка пожарной сигнализации. Также с сотрудниками должен проводиться инструктаж по действиям при возникновении данной чрезвычайной ситуации. Во всех служебных помещениях должен присутствовать план эвакуации людей.

Для предотвращения пожара рабочее помещение должно быть оборудовано устройствами, предназначенными для локализации и ликвидации возгорания на начальной стадии – первичными средствами пожаротушения.

7.2 Экологическая безопасность

В данном разделе рассматривается воздействие на окружающую среду деятельности по разработке предиктивной модели, а также самого продукта в результате его реализации.

В ходе выполнения магистерской диссертации и дальнейшем использовании результата разработки отсутствуют выбросы каких-либо вредных веществ в атмосферу и гидросферу, следовательно, загрязнение воздуха и воды не происходит.

7.3 Безопасность в чрезвычайных ситуациях

По характеру источников возникновения [23] чрезвычайные ситуации классифицируют на следующие группы:

- природные (землетрясения, наводнения, ураганы);
- техногенные (взрывы, аварии, пожары, транспортные катастрофы);
- экологические (загрязнения, опустынивание, кислотные дожди);
- биологического происхождения (эпидемии);
- антропогенные (терроризм, войны).

При работе с компьютерной техникой наиболее вероятной чрезвычайной ситуацией является пожар.

7.3.1Наиболее типичная ЧС - пожар

Возникновение пожара в помещениях может обуславливаться следующими факторами:

- возникновением короткого замыкания в электропроводке вследствие неисправности самой проводки;
- возгоранием устройств вычислительной аппаратуры вследствие нарушения изоляции или неисправности самой аппаратуры;
- возгоранием мебели по причине нарушения правил пожарной безопасности, а также неправильного использования электроприборов и электроустановок.

7.3.2Меры по предотвращению ЧС

Для предупреждения возникновения пожара необходимо соблюдать следующие правила пожарной безопасности:

- исключение образования горючей среды (герметизация оборудования, контроль воздушной среды, рабочей и аварийной вентиляции);
- применение при строительстве и отделке зданий несгораемых или трудно сгораемых материалов.

Необходимо в офисе проводить следующие пожарно-профилактические мероприятия:

- противопожарный инструктаж персонала;
- обучение персонала правилам техники безопасности;
- издание инструкций, плакатов, планов эвакуации.

Эксплуатационные мероприятия:

- соблюдение эксплуатационных норм оборудования;
- обеспечение свободного подхода к оборудованию;
- содержание в исправности изоляции токоведущих проводников.

Технические мероприятия:

- соблюдение противопожарных мероприятий при устройстве электропроводок, оборудования, систем отопления, вентиляции и освещения;

- профилактический осмотр, ремонт и испытание оборудования.

Также устройства для локализации и ликвидации возгораний должны быть в рабочем состоянии и должен быть обеспечен свободный подход к этим устройствам в случае возникновения чрезвычайной ситуации.

В помещениях с компьютерной техникой, недопустимо применение воды и пены ввиду опасности повреждения или полного выхода из строя дорогостоящего электронного оборудования.

Для тушения пожаров необходимо применять углекислотные и порошковые огнетушители, которые обладают высокой скоростью тушения, большим временем действия, возможностью тушения электроустановок, высокой эффективностью борьбы с огнем. Воду разрешено применять только во вспомогательных помещениях [24].

7.4 Правовые и организационные вопросы обеспечения безопасности

7.4.1Требования к рабочему помещению для работы с ПЭВМ

Требования к рабочему помещению для работы с ПЭВМ регламентируются СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы [20].

Рабочее помещение, в котором присутствуют персональные компьютеры, должно удовлетворять эргономическим требованиям:

- Помещение должно иметь естественное и искусственное освещение. Окна в помещениях, где эксплуатируется вычислительная техника, преимущественно должны быть ориентированы на север и северо-восток;
- площадь на одно рабочее место пользователя персонального компьютера на базе электроннолучевой трубки должна составлять не менее 6 $м^2$, на базе плоских дискретных экранов (жидкокристаллические, плазменные) - 4,5 $м^2$;

- помещения, где размещаются рабочие места с ПЭВМ, должны быть оборудованы защитным заземлением в соответствии с техническими требованиями по эксплуатации;
- не следует размещать рабочие места с ПЭВМ вблизи силовых кабелей и вводов, высоковольтных трансформаторов, технологического оборудования, создающего помехи в работе ПЭВМ.

7.4.2Требования к рабочему месту с ПЭВМ

Работа с компьютером характеризуется умственным напряжением и высокой напряженностью зрительной работы, поэтому большое значение имеет расположение элементов рабочего места для поддержания оптимальной рабочей позы человека.

Основными элементами рабочего места программиста являются: рабочий стол, рабочий стул (кресло), дисплей, клавиатура, мышь; вспомогательными - пюпитр, подставка для ног [25].

Требования к рабочему месту с ПЭВМ описаны в ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения [25].

Рабочее место должно удовлетворять следующим требованиям:

- Конструкция рабочего стола должна обеспечивать возможность размещения на рабочей поверхности необходимого комплекта оборудования и документов с учетом характера выполняемой работы;
- размеры рабочей поверхности стола должны быть: глубина - не менее 600 мм, ширина - не менее 1200 мм;
- рабочий стол должен иметь пространство для ног высотой не менее 600 мм, шириной - не менее 500 мм, глубиной на уровне колен - не менее 450 мм и на уровне вытянутых ног - не менее 650 мм;
- рабочая поверхность стола не должна иметь острых углов и краев. Покрытие рабочей поверхности стола должно быть из диффузно отражающего материала с коэффициентом отражения 0,45-0,50;

- рабочий стул должен обеспечивать поддержание физиологически рациональной рабочей позы оператора в процессе трудовой деятельности. Рабочий стул должен быть подъемно-поворотным и регулируемым по высоте. Поверхность сиденья должна иметь ширину и глубину не менее 400 мм. Должна быть предусмотрена возможность изменения угла наклона поверхности сиденья от 15° вперед до 5° назад. Высота поверхности сиденья должна регулироваться в пределах от 400 до 550 мм;

- дисплей на рабочем месте должен располагаться так, чтобы изображение в любой его части было различимо без необходимости поднять или опустить голову. Дисплей должен быть установлен ниже уровня глаз оператора. Угол наблюдения экрана оператором относительно горизонтальной линии взгляда не должен превышать 60°;

- монитор должен находиться от глаз пользователя на оптимальном расстоянии 600-700 мм, но не ближе 500 мм;

клавиатура на рабочем месте должна располагаться так, чтобы обеспечивалась оптимальная видимость экрана. Клавиатура должна иметь возможность свободного перемещения. Клавиатуру следует располагать на поверхности стола на расстоянии от 100 до 300 мм от переднего края, обращенного к оператору, или на специальной регулируемой по высоте рабочей поверхности, отделенной от основной столешницы [25].

ЗАКЛЮЧЕНИЕ

По результатам выполнения выпускной квалификационной работы было разработано программное приложение, позволяющее прогнозировать результаты выполнения проекта на основе его основных показателей и текстовой информации о нем в виде заключенного договора на исполнение.

В ходе выполнения работы были выполнены следующие задачи:

1. Было разработано программное приложение для выгрузки и обработки более полумиллиона файлов в формате XML с FTP-сервера Федеральной контрактной системы, а также нескольких тысяч договоров с ее официального портала;

2. На основе полученных данных о договорах по проектам была обучена модель с помощью наивного байесовского алгоритма с некоторыми модификациями для улучшения показателей обучения;

3. На основе полученных из данных и отдельно выведенных показателей проектов были обучены модели с помощью искусственных нейронных сетей двух видов.

Полученные результаты показали высокую точность всех методов обучения.

Стоит отметить, что НБА как один из наиболее простых методов классификации показал довольно высокую точность и полноту обучения (79% и 83% процента соответственно), что может свидетельствовать о его удачной модификации в процессе разработки.

Однако неоспоримым лидером среди всех методов обучения выступила ИНС, обученная методом обратного распространение ошибки, показавшая в результате точность в 93% и полноту равную 96,5%.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. НОУ ИНТУИТ | Лекция | Заключение [Электронный ресурс] - URL: https://www.intuit.ru/studies/courses/1001/297/lecture/7426 (Дата обращения: 28.04.2018).

2. DIN 66901 [Электронный ресурс] - URL: https://pmpractice.ru/knowledgebase/normative/projectstandarts/din-69901 (Дата обращения: 28.04.2018).

3. Федеральный закон "О закупках товаров, работ, услуг отдельными видами юридических лиц" от 18.07.2011 N 223-ФЗ (последняя редакция) / КонсультантПлюс [Электронный ресурс] - URL: http://www.consultant.ru/document/cons_doc_LAW_116964 (Дата обращения: 28.04.2018).

4. 6 простых шагов для освоения наивного байесовского алгоритма (с примером кода на Python) — DataReview.info [Электронный ресурс] - URL: http://datareview.info/article/6-prostyih-shagov-dlya-osvoeniya-naivnogo-bayesovskogo-algoritma-s-primerom-koda-na-python (Дата обращения: 28.04.2018).

5. Персептроны – Нейроныне сети [Электронный ресурс] - URL: https://neuralnet.info/chapter/персептроны (Дата обращения: 28.04.2018).

6. С. Хайкин Нейронные сети: полный курс, 2-е издание. : Пер. с англ. – М: Издательский дом «Вильямс», 2006 – 219 с (Дата обращения: 28.04.2018).

7. Л.А. Зинченко, В.М. Курейчик, В.Г. Редько Бионические информационные системы и их практические применения. – М: ФИЗМАТЛИТ, 2011 – 129 с (Дата обращения: 28.04.2018).

8. Скользящий контроль [Электронный ресурс] - URL: http://www.machinelearning.ru/wiki/index.php?title=CV (Дата обращения: 28.04.2018).

9. Lightning Fast Data Science Platform | RapidMiner [Электронный ресурс] - URL: https://rapidminer.com (Дата обращения: 28.04.2018).

10. Natural Language Toolkit — NLTK 3.3 documentation [Электронный ресурс] - URL: https://www.nltk.org (Дата обращения: 28.04.2018).

11. Documentation scikit-learn: machine learning in Python — scikit-learn 0.19.1 documentation [Электронный ресурс] - URL: http://scikit-learn.org/stable/documentation.html (Дата обращения: 28.04.2018).

12. Руководство по языку C# | Microsoft Docs [Электронный ресурс] - URL: https://docs.microsoft.com/ru-ru/dotnet/csharp (Дата обращения: 28.04.2018)

13. Методы ETL в анализе промышленных данных [Электронный ресурс] - URL: http://statistica.ru/local-portals/quality-control/metody-etl-v-analize-promyshlennykh-dannykh (Дата обращения: 28.04.2018).

14. Гречанин В.А. К вопросу о токенизации текста – Международный научно-исследовательский журнал, 2016 – 25 с (Дата обращения: 28.04.2018).

15. Text Classification for Sentiment Analysis – Precision and Recall | StreamHacker [Электронный ресурс] - URL: https://streamhacker.com/2010/05/17/text-classification-sentiment-analysis-precision-recall (Дата обращения: 28.04.2018).

16. Преимущества нейронных сетей [Электронный ресурс] - URL: http://www.aiportal.ru/articles/neural-networks/advantages.html (Дата обращения: 28.04.2018).

17. Алгоритм обучения многослойной нейронной сети методом обратного распространения ошибки (Backpropagation) / Хабр [Электронный ресурс] - URL: https://habr.com/post/198268 (Дата обращения: 28.04.2018).

18. PyBrain [Электронный ресурс] - URL: http://pybrain.org (Дата обращения: 28.04.2018).

19. Введение в нейроэволюционный подход [Электронный ресурс] - URL: http://masters.donntu.org/2012/iii/kishinskiy/library/article3.htm (Дата обращения: 28.04.2018).

20. СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы //

Электронный фонд правовой и нормативно-технической документации. URL: http://docs.cntd.ru/document/901865498 (Дата обращения: 28.04.2018).

21. ГОСТ 12.1.038–82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов // Электронный фонд правовой и нормативно-технической документации. URL: http://docs.cntd.ru/document/5200313 (Дата обращения: 28.04.2018).

22. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты // Электронный фонд правовой и нормативно-технической документации. URL: http://docs.cntd.ru/document/1200080203 (Дата обращения: 28.04.2018).

23. Чрезвычайная ситуация // Википедия. Свободная энциклопедия. URL: https://ru.wikipedia.org/wiki/Чрезвычайная_ситуация (Дата обращения: 29.04.2018).

24. Чрезвычайные ситуации при работе с ПЭВМ // Студопедия. Ваша энциклопедия. URL: https://studopedia.ru/8_107307_osveshchenie-pomeshcheniy-vichislitelnih-tsentrov.html (Дата обращения: 29.04.2018).

25. ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения // Электронный фонд правовой и нормативно-технической документации. URL: http://docs.cntd.ru/document/gost-r-50923-96 (Дата обращения: 28.04.2018)

ПРИЛОЖЕНИЕ А

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<ns2:contract
xsi:schemaLocation="http://zakupki.gov.ru/223/integration/schema/TFF-3.5
http://zakupki.gov.ru/223/integration/schema/TFF-3.5/Types.xsd"
xmlns="http://zakupki.gov.ru/223fz/types/1">
<header>
<guid>7c100d28-7b17-44dc-9358-c7f7e20929e8</guid>
<createDateTime>2015-05-05T11:12:28</createDateTime>
</header>
<ns2:body>
<ns2:item>
<guid>74673657-da50-48b5-9d03-fb92e2fa69e8</guid>
<ns2:contractData>
      <ns2:guid>54d9e8e7-b1a6-465a-b8c2-ea8fd2dfe5d0</ns2:guid>
      <ns2:urlOOS>https://zakupki.gov.ru/223/contract/private/contract/view/gene
ral-information.html?id=274138&amp;viewMode=VSRI</ns2:urlOOS>
      <ns2:createDateTime>2015-05-05T10:30:07</ns2:createDateTime>
      <ns2:customer>
            <mainInfo>
                  <fullName>ОТКРЫТОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "ГАЗПРОМ
ГАЗОРАСПРЕДЕЛЕНИЕ МАЙКОП"</fullName>
                  <shortName>ОАО "ГАЗПРОМ ГАЗОРАСПРЕДЕЛЕНИЕ МАЙКОП"</shortName>
                  <inn>0105018196</inn>
                  <kpp>010501001</kpp>
                  <ogrn>1020100707318</ogrn>
                  <legalAddress>385003, Адыгея, Майкоп, Апшеронская, дом
4</legalAddress>
                  <postalAddress>385003, Адыгея, Майкоп, Апшеронская, дом
4</postalAddress>
                  <phone>7-8772-523180</phone>
                  <fax>7-8772-560255</fax>
                  <email>kolesnichenko@adyggaz.ru</email>
                  <okato>79401000000</okato>
                  <okopf>47</okopf>
                  <okopfName>Открытые акционерные общества</okopfName>
                  <customerRegistrationDate>1995-03-
30T00:00:00</customerRegistrationDate>
            </mainInfo>
      </ns2:customer>
      <ns2:publicationDate>2015-05-05T11:12:28</ns2:publicationDate>
      <ns2:status>P</ns2:status>
      <ns2:version>1</ns2:version>
      <ns2:digitalPurchase>true</ns2:digitalPurchase>
      <ns2:digitalPurchaseCode>400001</ns2:digitalPurchaseCode>
      <ns2:provider>false</ns2:provider>
      <ns2:changeContract>false</ns2:changeContract>
```

```
      <ns2:attachments/>
      <ns2:contractRegNumber>5010501819615000055000</ns2:contractRegNumber>
      <ns2:name>3488</ns2:name>
      <ns2:contractDate>2015-05-05T00:00:00</ns2:contractDate>
      <ns2:purchaseNoticeInfo>
            <ns2:guid>798a235d-9c60-4ea9-b695-3fffe851ecdb</ns2:guid>
            <ns2:purchaseNoticeNumber>31502213762</ns2:purchaseNoticeNumber>
            <ns2:publicationDateTime>2015-04-
03T08:45:06</ns2:publicationDateTime>
            <ns2:name>оборудование светотехническое</ns2:name>
      </ns2:purchaseNoticeInfo>
      <ns2:lotGuid>af003169-b7f9-4a32-9da0-30d21c26ea5d</ns2:lotGuid>
      <ns2:subjectContract>оборудование светотехническое</ns2:subjectContract>
      <ns2:purchaseTypeInfo>
            <ns2:code>40000</ns2:code>
            <ns2:name>Иной способ закупки, предусмотренный правовым актом
заказчика, указанным в части 1 статьи 2 Федерального закона</ns2:name>
      </ns2:purchaseTypeInfo>
      <ns2:resumeDate>2015-04-23T00:00:00</ns2:resumeDate>
      <ns2:supplierInfo>
            <ns2:name>ООО "Ростовдонконтракт"</ns2:name>
            <ns2:inn>6163133512</ns2:inn>
            <ns2:kpp>616301001</ns2:kpp>
            <ns2:okpo>24191275</ns2:okpo>
            <ns2:type>L</ns2:type>
            <ns2:provider>false</ns2:provider>
            <ns2:subcontractor>false</ns2:subcontractor>
            <ns2:nonResident>false</ns2:nonResident>
            <ns2:registrationDate>2005-01-01T00:00:00</ns2:registrationDate>
            <ns2:tax>true</ns2:tax>
            <ns2:okopf>19000</ns2:okopf>
            <ns2:okopfName>Прочие юридические лица, являющиеся коммерческими
организациями</ns2:okopfName>
            <ns2:address>
                  <ns2:country>
                        <name>РОССИЯ</name>
                        <digitalCode>643</digitalCode>
                  </ns2:country>
                  <ns2:postCode>344000</ns2:postCode>
                  <ns2:oktmo>60701000</ns2:oktmo>
                  <ns2:region>
                        <ns2:kladerCode>61000000000</ns2:kladerCode>
                        <ns2:fullName>Ростовская   обл</ns2:fullName>
                  </ns2:region>
            </ns2:address>
      </ns2:supplierInfo>
      <ns2:hasSubcontractor>false</ns2:hasSubcontractor>
      <ns2:price>76091.12</ns2:price>
```

```
      <ns2:currency>
            <code>RUB</code>
            <digitalCode>643</digitalCode>
            <name>Российский рубль</name>
      </ns2:currency>
      <ns2:startExecutionDate>2015-05-05T00:00:00</ns2:startExecutionDate>
      <ns2:endExecutionDate>2015-12-31T00:00:00</ns2:endExecutionDate>
      <ns2:contractPositions>
            <ns2:contractPosition>
                  <ns2:guid>3c7ee112-957d-463c-9860-51dcc9edcbaf</ns2:guid>
                  <ns2:ordinalNumber>1</ns2:ordinalNumber>
                  <ns2:okdp>
                        <code>3190290</code>
                        <name>Прочее электрооборудование, не включенное в другие
группировки</name>
                  </ns2:okdp>

      <ns2:impossibleToDetermineAttr>false</ns2:impossibleToDetermineAttr>
                  <ns2:okei>
                        <code>796</code>
                        <name>Штука</name>
                  </ns2:okei>
                  <ns2:qty>184</ns2:qty>
            </ns2:contractPosition>
      </ns2:contractPositions>
</ns2:contractData>
</ns2:item>
</ns2:body>
</ns2:contract>
```

ПРИЛОЖЕНИЕ Б

Chapter 1
Subject area analysis

Студент:

Группа	ФИО	Подпись	Дата
8ИМ6А	Чебоксаров Владимир Александрович		

Руководитель ВКР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	А.В. Кудинов	к.т.н.		

Консультант школы отделения (НОЦ) ОИТ:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Е.А. Мирошниченко	к.т.н.		

Консультант – лингвист отделения (НОЦ) школы ОИЯ:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ	О.В. Комиссарова	к.филол.н.		

1 Subject area analysis

1.1 Data mining in project management

Data mining (DM) is a process of identifying significant correlations, patterns and trends in large amounts of data [1].

Data mining is widely used in business applications by analysts and managers of companies. There are many high-level tools which are developed for these categories of users that allow solving rather complicated practical problems without special mathematical preparation. The usage of data mining in business is connected to the fierce competition that arose from transition from the “seller’s market” to there “buyer’s market”. In these conditions, the quality and validity of decisions are particularly important, which requires a strict quantitative analysis of the available data. When working with large volumes of accumulated information, it is necessary to constantly monitor the dynamics of the market, and this is practically impossible without automation of analytical activities.

There are a number of typical tasks for data mining in project management:

1. Forecasting. It is one of the most common tasks of DM. It is necessary to forecast sales volumes and other parameters taking into account the many interrelated factors: seasonal, regional, general economic, etc.

2. Marketing analysis. In order to develop an effective marketing plan, you need to know how the sales level is affected by factors such as the cost of the product, the cost of product promotion and advertising. Models based on neural network allow managers and analysts to predict such an impact.

3. Analysis of the work of staff. The productivity of employees depends on the level of training, on wages, work experience, relationships with management, etc. Analysis of the impact of these factors allows us develop a methodology for increasing labor productivity, as well as to suggest an optimal strategy for recruitment in the future.

4. Customers profiling. With the help of models based on neural network it is possible among the numerous clients of the company to choose those whose cooperation is most beneficial – to get a portrait of the “typical client of the

company". In addition, it's possible to find out why work with some of the customers has become ineffective, and develop a strategy for finding the right customers in the future.

5. Evaluation of potential customers. When planning preliminary negotiations, it makes sense to determine with what percentage of probability they will end with the conclusion of the contract (or the sale of products). The analysis of experience with customers allows us to identify the characteristics of those applications that ended in real sales. Using the results of this analysis, managers can focus on more promising client applications.

Obviously, the listed types of tasks are relevant for virtually all business sectors: banking and insurance, financial markets, production, trade, etc.

1.2 Methodology for assessing the success of the project

Before determining the methodology for assessing the success of a project, it is necessary to refer to the very term "project". One of the most comprehensive definitions is given in DIN 69901 standard [2], according to which the project is an enterprise (or intention), which is largely characterized by the uniqueness of the conditions in their totality, for example:

- Goals;
- Temporary, financial, human and other restrictions;
- Differentiation from other intentions;
- Specific for the project organization of its implementation.

Also, the so-called "iron triangle" rule can be used for the evaluation, which describes the balance between the project cost, the time of its execution and the quality of the result (Fig. 1).

Figure 1 – "Iron triangle"

According to this definition, it is possible to derive a general criterion for the success of the project - this is the achievement of the project's objectives at the planned time and within the planned resources.

1.3 Description of the federal law №223-FL

The federal contract system of the Russian Federation was chosen as a source of data about contracts, because it provides free and unrestricted access to full and reliable information about the implementation of contracts in procurement and procurement of goods, work, services by certain types of legal entities under Federal Law No. 223-FL [3].

Federal Law No. 223-FL regulates the general principles of procurement for:

- Organizations with a state stake of more than 50%;
- Companies engaged in regulated activities;
- Organizations-subjects of natural monopolies;
- Budget organizations that purchase from extrabudgetary funds.

Thus, information on competitions published in the public procurement system is the central source of up-to-date information on possible “state” orders for a profile for a multitude of companies from a wide range of activities - from security services to geophysical surveys.

1.4 Methods for solving the classification problem

The most important point in solving this problem is the creation of an effective classification algorithm, on which basis the model will be trained.

The following main groups can be distinguished among all classification methods:

- Bayesian classifiers;
- Artificial neural networks (ANN);
- Linear separators;
- Algorithmic compositions.

In our case, we will consider below a naive Bayesian algorithm and some types of artificial neural networks, as one of the most popular methods of classifier training, as well as cross-validation as a method for evaluating the learning outcomes of a model.

1.4.1 Naive Bayesian algorithm

A naive Bayesian algorithm (NBA) is a classification algorithm based on the Bayes theorem with the assumption of independence of features [4]. In other words, the NBA assumes that the presence of any feature in the class is not related to the presence of any other sign. Even if the signs depend on each other or on other signs, in any case they contribute independently to the result. In connection with this assumption, the algorithm is called “naive”.

Models based on the NBA are fairly simple and extremely useful when working with very large sets of data. With its simplicity, the NBA is able to surpass even some complex classification algorithms.

The use of the naive Bayesian algorithm has the following positive aspects:

- Classification, including multi-class, is easy and fast;
- When the independence assumption is fulfilled, the NBA surpasses other algorithms, such as logistic regression, while requiring less training data;
- NBA works better with categorical features than with continuous ones. For continuous features, a normal distribution is assumed, which is a fairly strong assumption.

However, the use of NBA can have a negative impact:

- If there is some value of a category characteristic in the test data set that was not found in the training data set, then the model will assign a zero probability to this value and will not be able to make a prediction. This phenomenon is known as the “zero frequency”. This problem can be solved with the help of smoothing. One of the simplest methods is Laplace smoothing.
- Although the NBA is a good classifier, the values of the predicted probabilities are not always accurate enough.

- Another limitation of the NBA is the assumption of the independence of signs. In reality, sets of completely independent signs are extremely rare.

1.4.2 Artificial neural networks

In a process of solving classification problems, it is necessary to classify existing static samples as belonging to certain classes. There are several possible ways of presenting data. The most common method is when the sample is represented by a vector. With the processing of data presented in this form, artificial neural networks do the best.

1.4.2.1 Perceptron

Perceptron is the simplest representative of ANN. The perceptron is based on a mathematical model of information perception by the brain [5]. In its most general form, it represents a system of elements of three different types: sensors, associative elements, and reactive elements (Figure 2).

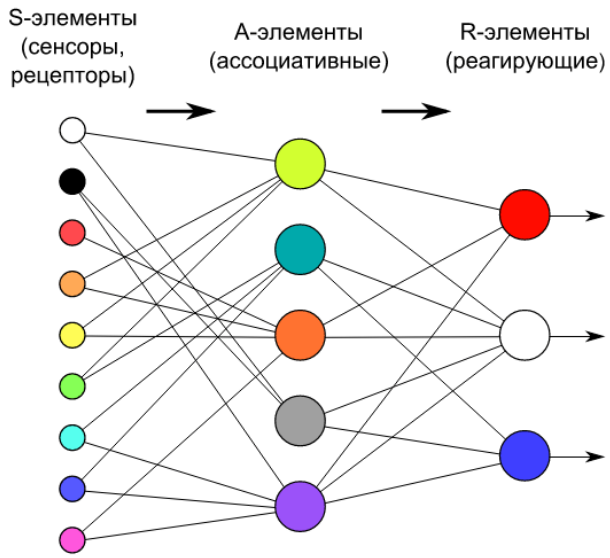

Figure 2 - The simplest model of the perceptron

There are three main types of perceptron:

- Single-layer perceptron;
- Perceptron with one hidden layer;
- Multilayer perceptron.

The choice of a certain type of perceptron varies depending on the complexity of the problem being solved and its subject area.

Next, consider the methods of teaching ANN.

1.4.2.2 Backpropagation

Backpropagation algorithm is one of the methods for learning multi-layer neural networks of direct propagation, also called multilayer perceptrons.

This algorithm assumes two passes on all layers of the network: forward and backward. With a forward pass, the input vector is fed to the input layer of the neural network, and then propagates through the network from layer to layer. As a result, a set of output signals is generated, which is the actual response of the network to this input image. During the forward passage, all the synaptic weights of the network are fixed. During the backward pass, all synaptic weights are adjusted according to the error correction rule, namely: the actual network output is subtracted from the desired one, resulting in an error signal. This signal subsequently spreads through the network in the direction opposite to the direction of the synaptic connections. Synaptic weights are adjusted in order to maximize the output of the network to the desired [6].

1.4.2.3 Neuroevolutionary algorithm

Evolutionary algorithms (EA) model the basic positions in the theory of biological evolution - the processes of selection, mutation and reproduction. The behavior of agents is determined by the environment. A lot of agents are usually called a population. Such a population evolves in accordance with the selection rules in accordance with the objective function assigned by the environment. Thus, each agent (individual) of the population is assigned the value of its suitability in the environment. Only the most suitable species reproduce. Recombination and mutation allow agents to change and adapt to the environment. Such algorithms refer to adaptive search engines. The scheme of the EA is presented in Figure 3.

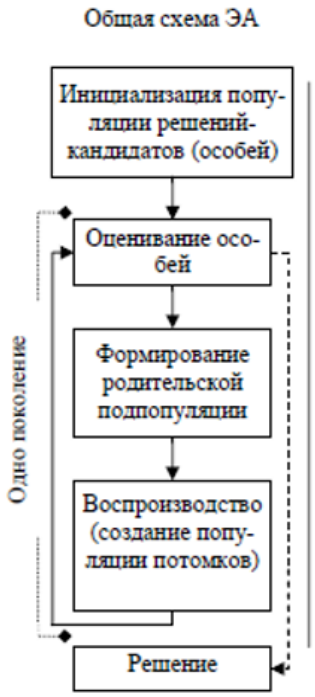

Figure 3 - General scheme of the evolutionary algorithm

One of the distinctive features of EA is their adaptive abilities, which makes it possible to realize the adjustment of the parameters of the EA in the process of its operation to improve the efficiency of EA and the quality of the results. The combination of ANN and evolutionary algorithms makes it possible to combine the flexibility of the ANN setting and the adaptability of EA, which allows implementing a largely unified approach to solving a wide range of classification problems [7]. In our case, the use of the evolutionary approach allows us to simultaneously adjust the weight of the links and the structure of the ANN.

1.4.3 Cross-validation

Cross-validation is a procedure of empirical estimation of the generalizing ability of algorithms trained by use of precedents [8].

In the process of cross-validation a certain number of partitions of the initial sample are fixed into two subsamples: teaching and control. For each partition, the algorithm for the training subsample is tuned, then its mean error is estimated at the objects of the control subsample. The estimation of the sliding control is the mean for all partitions of the error value in the control subsamples.

If the sample is independent, the average error of the sliding control gives an unbiased estimate of the error probability. This distinguishes it from the average error in the training sample, which may turn out to be biased (optimistically low) estimate of the probability of error, which is associated with the phenomenon of retraining.

When cross-validating for each partition, the algorithm for the training subsample is tuned, then its mean error is estimated at the objects of the control subsample. The estimation of the sliding control is the mean for all partitions of the error value in the control subsamples.

Sliding control is a standard technique for testing and comparing classification, regression and prediction algorithms.

1.5 Development tools

1.5.1 RapidMiner

This tool offers advanced analytics using templates based on templates. Users hardly need to write code. Offered as a service, not part of the local software, this tool takes the top position in the list of data mining tools that are freely available (Figure 4).

In addition to intelligent data analysis, RapidMiner also provides functions such as data pre-processing and visualization, intelligent analytics and statistical modeling, evaluation and deployment. Also, additional opportunities are provided by the training schemes, models and algorithms from the scenarios WEKA and R.

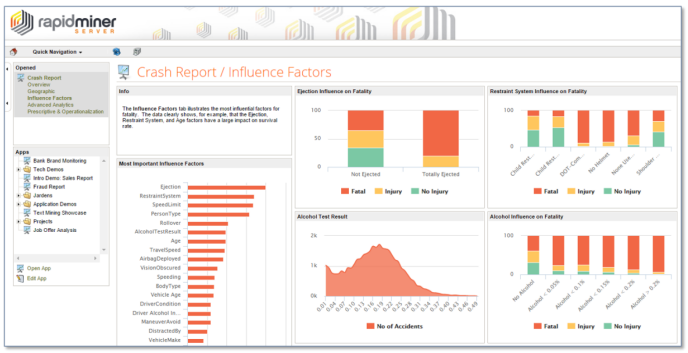

Figure 4 - RapidMiner user interface

RapidMiner is distributed under the AGPL license with open source and can be downloaded from SourceForge, where it is ranked as the number one business intelligence software [9].

1.5.2 Python

The Python programming language was chosen to implement the model. The choice was made for several reasons:

- Simple language syntax;
- Low entry threshold;
- Lots of libraries that help in the implementation of various tasks;
- The most commonly used language in data analysis and machine learning;
- Abundance of documentation.

As a result of the described work with the help of Python, the tasks of tokenization of the texts of documents and learning the model with the help of the NBA and INS were to be solved. For these purposes, the following libraries were also used:

- NLTK - a package of libraries and programs for symbolic and statistical processing of natural language [10];
- Scikit-Learn - a library that allows you to implement many of the already existing algorithms of machine learning [11].

1.5.3 C#

C# is an object-oriented programming language designed to develop applications for the .NET Framework [12].

The choice of this programming language can be caused by the following factors:

- Official technical support from Microsoft;
- Good integration with MS SQL;
- Extensive documentation;
- A large number of syntactic constructions designed to simplify the writing of code;
- The presence of the NuGet package manager, which makes it easier to find and connect libraries to the project.

1.6 Goals and objectives of the development

Market research of software products for evaluating the results of the future project has shown that this market segment is empty, and as a whole, at the moment there are no software solutions for forecasting the results of the project.

In view of the fact that at the moment the need for software of this kind is not satisfied, the purpose of this work is to create software to assess the possible outcome of the project.

Since the open sources of data on the projects of commercial companies and the results of their implementation are not in the public domain, the portal of the Federal Contract System containing data on state contracts in various fields of activity was chosen as the source.

To improve the quality of the results and increase the sample of data, two types of initial data on the project were selected in quality:

- Text document of the performance contract;
- Basic time and cost indicators of the project, as well as data on the customer and the supplier.

Accordingly, for each type of source data, we chose our own method for learning the model.

A naive Bayesian algorithm was chosen as a method for processing documents, because it is one of the fastest, simplest and most effective algorithms for working with documents that have a similar structure.

The second variant of data processing projects is related to the correlation between the group of project indicators and the possibility of its successful completion. Since in this case, as the input data represent a vector containing the main indicators of the project, the most appropriate choice is the INS as the method of classifier training.

There were also two methods for adjusting the INS parameters in order to increase the likelihood of successful learning: the method of back propagation of the error and a method using an evolutionary algorithm to match the structure of the ANN and the weights of its synaptic connections.

In the course of the work, the following tasks were set:

1. Analysis of the subject area;

2. Unloading of the implementation of contracts under Federal Law No. 223-FL;

3. Preparing data and loading it into the database;

4. Additional unloading of contract documents for projects;

5. Tokenization of documents for NBA application;

6. Training of the predictive model with the help of the NBA on the basis of text data about the project;

7. Allocation and addition of key project indicators;

8. Training of the predictive model using the ANN using the method of back propagation of the error based on the main project indicators;

9. Training of a predictive model with the help of ANN using an evolutionary algorithm based on the main project indicators;

10. Approbation of the created models for predicting the success of projects;

11. Evaluation of the results.