ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
«РОСТЕЛЕКОМ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»

ООО «РТК ИТ»

УТВЕРЖДАЮ:
Генеральный директор

_____В.В. Ерохин

ПРОГРАММА

Профессиональной переподготовки
(вид дополнительной профессиональной программы)

«Аналитика данных и методы искусственного интеллекта на базе решений ПАО Ростелеком»

Автор-составители:
Сахнюк П.А
к.т.н., ведущий аналитик ООО «РТК ИТ»;
Бочаров М.И.
к.пед.н., эксперт-методолог ООО «РТК ИТ».

Москва - 2024

Программа профессиональной переподготовки
«Аналитика данных и методы искусственного интеллекта»

Общие положения

1. Дополнительная профессиональная программа (программа профессиональной переподготовки) ИТ-профиля « Аналитика данных и методы искусственного интеллекта на базе решений ПАО Ростелеком » (далее – Программа, ДПП ПП) разработана в соответствии с нормами Федерального закона РФ от 29 декабря 2012 года № 273-ФЗ «Об образовании в Российской Федерации», с учетом требований приказа Минобрнауки России от 1 июля 2013 г. № 499 «Об утверждении Порядка организации и осуществления образовательной деятельности по дополнительным профессиональным программам», с изменениями, внесенными приказом Минобрнауки России от 15 ноября 2013 г. № 1244 «О внесении изменений в Порядок организации и осуществления образовательной деятельности по дополнительным профессиональным программам, утвержденный приказом Министерства образования и науки Российской Федерации от 1 июля 2013 г. № 499», приказа Министерства образования и науки РФ от 23 августа 2017 г. N 816 «Об утверждении Порядка применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ»; паспорта федерального проекта «Развитие кадрового потенциала ИТ-отрасли» национальной программы «Цифровая экономика Российской Федерации»; постановления Правительства Российской Федерации от 13 мая 2021 г. № 729 «О мерах по реализации программы стратегического лидерства «Приоритет-2030» (в редакции постановления Правительства Российской Федерации от 14 марта 2022 г. № 357 «О внесении изменений в постановление Правительства Российской Федерации от 13 мая 2021 г. № 729»);приказа Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации от 28 февраля 2022 г. № 143 «Об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации» и признании утратившими силу некоторых приказов Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации об утверждении методик расчета показателей федеральных проектов национальной программы «Цифровая экономика Российской Федерации» (далее – приказ Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации № 143); федерального государственного образовательного стандарта высшего образования по направлению подготовки 09.03.02 Информационные системы и технологии (уровень бакалавриата), утвержденного приказом Минобрнауки России от 19сентября 2017 г. (С изменениями и дополнениями от:26 ноября 2020 г., 8 февраля 2021 г.), (далее вместе – ФГОС ВО)), а также профессионального стандарта 06.042 «Специалист по большим данным», утвержденного приказом Министерства труда и социальной защиты РФ от 6 июля 2020 г. № 405н.

2. Профессиональная переподготовка заинтересованных лиц (далее – Слушатели), осуществляемая в соответствии с Программой (далее – Подготовка),

имеющей отраслевую направленность[1]«Информационно-коммуникационные технологии», проводится в ООО «РТК ИТ» в соответствии с учебным планом в очной/заочной форме обучения[2].

3. Разделы, включенные в учебный план Программы, используются для последующей разработки календарного учебного графика, учебно-тематического плана, рабочей программы, оценочных и методических материалов. Перечисленные документы разрабатываются образовательной организацией самостоятельно, с учетом актуальных положений законодательства об образовании, законодательства в области информационных технологий и смежных областей знаний ФГОС ВО и профессионального стандарта 06.042 «Специалист по большим данным».

Цель

Цель программы профессиональной переподготовки: Целью подготовки слушателей по Программе является получение компетенций, необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий: большие данные, сбор, обработка и анализ больших данных в организации, анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры, анализ больших данных в проектах под контролем опытных специалистов, приобретение новой квалификации «Специалист по большим данным».

Характеристика новой квалификации и связанных с ней видов профессиональной деятельности, трудовых функций и (или) уровней квалификации

Виды профессиональной деятельности, трудовая функция, указанные в профессиональном стандарте по соответствующей должности «Аналитик», «Исследователь данных», «Руководитель (специалист) отдела по информационным технологиям», представлены в таблице 1:

[1] Варианты отраслевой направленности: «Городское хозяйство»; «Финансовые услуги»; «Строительство»; «Добывающая промышленность»; «Обрабатывающая промышленность»; «Транспортная инфраструктура»; «Здравоохранение»; «Энергетическая инфраструктура»; «Образование»; «Сельское хозяйство и агропромышленный комплекс»; «Информационно-коммуникационные технологии»; «Искусство и культура»

[2] При реализации Программы допускается использовать сетевую форму обучения с организациями реального сектора экономики субъекта Российской Федерации

Таблица 1

Характеристика новой квалификации, связанной с видом профессиональной деятельности и трудовыми функциями в соответствии с профессиональным стандартом «06.042 «Специалист по большим данным».

Область профессиональной деятельности	**Тип задач профессиональной деятельности**	**Код и наименование профессиональной компетенции**	**Трудовые действия**	**Трудовая функция**	**Обобщенная трудовая функция**	**Вид профессиональной деятельности**
06 Связь, информационные и коммуникационные технологии (в сфере исследования, разработки, внедрения и сопровождения информационных технологий и систем); 40 Сквозные виды профессиональной деятельности в промышленности (в сфере организации и проведения научно-исследовательских и опытно-конструкторских работ в области	Создание информационных технологий нового поколения, обеспечивающих экономически эффективное извлечение полезной информации из больших объемов разнообразных данных путем высокой скорости их сбора, обработки и анализа, и применение этих технологий в информационно аналитической деятельности, в системах управления и принятия	**ОПК-1 (2)** Способность понимать принципы работы современных информационных технологий **ОПК-2 (3)** Способность решать стандартные задачи профессиональной деятельности с применением информационно-коммуникационных технологий **ОПК-3 (7)** Способность осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации	Выявление требований заказчика к результатам анализа, определение возможностей применения анализа больших данных в предметной области и конкретных задачах заказчика. Консультирование заказчика по возможностям имеющейся методологической и технологической инфраструктуры анализа больших данных и результатам	A/01.6 Выявление, формирование и согласование требований к результатам аналитических работ с применением технологий больших данных. A/02.6Планирование и организация аналитических работ с использованием технологий больших данных. A/03.6 Подготовка данных для проведения	Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры исследования с применением технологий больших данных в соответствии с требованиями заказчика	*Специалист по большим данным*

информатики и вычислительной техники)	решений, а также для разработки на их основе новых продуктов и услуг	информационных систем **ПК -1** Способность анализировать большие данные **ПК-2** Способность оценивать возможности применения искусственного интеллекта и машинного обучения **ПК-3** Способность применять Искусственный интеллект и машинное обучение **ПК-4** Способность использовать программные и технические средства для визуализации больших данных	применения технологий больших данных к аналогичным задачам. Согласование с заказчиком и утверждение требований к результатам аналитического исследования	аналитических работ по исследованию больших данных. A/04.6 Проведение аналитического		

Таблица 2

Характеристика новой и развиваемой цифровой компетенции в ИТ-сфере, связанной с уровнем формирования и развития в результате освоения Программы[3] «Аналитика данных и методы искусственного интеллекта»

Наименование сферы	**Код и наименование профессиональной компетенции**	**Примеры инструментов**	**Базовый уровень развития компетенций.**	**Продвинутый уровень развития компетенций.**	**Экспертный уровень развития компетенций.**
Большие данные	**ОПК-1 (2)** Способность понимать принципы работы современных информационных технологий	Python SQL Google Looker Studio Google Colaboratory Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib Matplotlib, Seaborn, Altair, Plotly Express Pandas Profiling, RT.DataLake RapidMiner RT.DataVision Loginom Yandex DataLens			
	ОПК-2 (3) Способность решать стандартные задачи профессиональной деятельности с применением информационно-коммуникационных технологий				
	ОПК-3 (7) Способность осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации				

[3] На основании Модели цифровых компетенций, указанной в Приложении 2

	информационных систем				
	ПК -1 Способность анализировать большие данные		Анализирует большие данные в проектах под контролем опытных специалистов	Выполняет проекты по анализу больших данных: создания эффективных и масштабируемых программ для обработки и анализа больших объемов данных, использование различных алгоритмов машинного обучения и статистических методов для анализа и интерпретации больших объемов данных, опыт работы с более сложными методами анализа, такими как глубокое обучение, рекомендательные системы и т.д. работает с инструментами и технологиями для работы с большими данными включая выбор и настройку инструментов и технологий для	На экспертном уровне контролирует проекты по большим данным. Оценивает и применяет новые аналитические системы и инструменты, способен дать оценку сильных и слабых сторон новых технологических решений и обоснованно сравнить свободно распространяемые и коммерческие решения. Обучает других

				обеспечения потребностей проекта	
Искусственный интеллект и машинное обучение	**ПК-2** Способность оценивать возможности применения искусственного интеллекта и машинного обучения		Оценивает возможности применения искусственного интеллекта и машинного обучения на уровне включения искусственного интеллекта в модель бизнес-процесса как компонента, без подробного описания и с общими требованиями, при внешней постановке задачи	Оценивает возможности применения искусственного интеллекта и машинного обучения, эпизодически прибегая к экспертной консультации. Описывает бизнес-требования, требования к данным и перечень применимых алгоритмов искусственного интеллекта и машинного обучения для решения поставленных задач	Оценивает возможности применения искусственного интеллекта и машинного обучения системно, на экспертном уровне, формируя системное решение с описанием бизнес-требований, бизнес-процессов, требований к данным и корпоративным хранилищам, конвейеров данных, перечень применимых алгоритмов искусственного интеллекта и машинного обучения для решения поставленных задач
	ПК-3 Способность применять Искусственный интеллект и машинное обучение		Участвует в проектах применения искусственного интеллекта и машинного обучения под контролем опытных специалистов	Разрабатывает отдельные части проектов по применению искусственного интеллекта и машинного обучения	На экспертном уровне контролирует проекты применения искусственного интеллекта и машинного обучения. Оценивает и применяет новые аналоги искусственного интеллекта и машинного

					обучения. Обучает других
	ПК-4 Способность использовать программные и технические средства для визуализации больших данных		Реализует настройку визуализации на уровне платформ BI с подготовленным набором данных. Способен освоить визуализировать данные с использованием функций и методов библиотек.	Самостоятельно подбирает программные и технические средства для визуализации больших данных и использует их в работе, эпизодически прибегая к экспертной консультации	Подбирает и использует программные и технические средства для визуализации больших данных в зависимости от специфики данных на экспертном уровне, обучает других

Характеристика новых и развиваемых цифровых компетенций, формирующихся в результате освоения программы

В ходе освоения Программы Слушателем приобретаются следующие профессиональные компетенции:

- Анализирует большие данные в проектах под контролем опытных специалистов;
- Проведение аналитического исследования с применением технологий больших данных в соответствии с требованиями заказчика;
- Выявление, формирование и согласование требований к результатам аналитических работ с применением технологий больших данных.

(Код и наименование профессиональной компетенции Таблица 1)

ПК-1 Способность анализировать большие данные

ПК-2 Способность оценивать возможности применения искусственного интеллекта и машинного обучения

ПК-3 Способность применять Искусственный интеллект и машинное обучение

ПК-4 Способность использовать программные и технические средства для визуализации больших данных

В ходе освоения Программы Слушателем совершенствуются следующие профессиональные компетенции:

ОПК-1 (2) Способность понимать принципы работы современных информационных технологий

ОПК-2 (3) Способность решать стандартные задачи профессиональной деятельности с применением информационно-коммуникационных технологий

ОПК-3 (7) Способность осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации информационных систем

(Код и наименование профессиональной компетенции Таблица 2)

ПК -формирования цифровых компетенций в области создания алгоритмов и компьютерных программ, пригодных для практического применения:

ПК-1 применения анализа больших данных в предметной области и конкретных задачах заказчика.

ПК-2 способность использовать методологическую и технологическую инфраструктуру анализа больших данных.

ПК-3 способность применять Искусственный интеллект и машинное обучение

ПК-4 Способность использовать программные и технические средства для визуализации больших данных

Планируемые результаты обучения по ДПП ПП

Результатами подготовки слушателей по Программе является получение компетенции, необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий «Анализ больших

данных» с использованием существующей в организации методологической и технологической инфраструктуры; приобретение новой квалификации «Специалист по большим данным».

В результате освоения Программы слушатель должен:

Знать:

1. Инструменты и методы согласования с заказчиками требований к результатам аналитических исследований с использованием технологий больших данных.

2. Регламенты организации по оформлению требований к результатам аналитических исследований с использованием технологий больших данных.

3. Технологии межличностной и групповой коммуникации в деловом взаимодействии, основы конфликтологии.

4. Технологии подготовки и проведения презентаций.

5. Предметную область анализа больших данных в соответствии с требованиями заказчика.

6. Возможности имеющейся у исполнителя методологической и технологической инфраструктуры анализа больших данных.

7. Современный опыт использования анализа больших данных.

8. Теоретические и прикладные основы анализа данных.

9. Типы анализа больших данных, виды аналитики.

10. Современные методы и инструментальные средства анализа больших данных.

11. Стандарты проведения анализа данных.

12. Источники информации, в том числе информации, необходимой для обеспечения деятельности в предметной области заказчика исследования.

13. Методы интерпретации и визуализации больших данных.

14. Правила деловой переписки

Уметь:

1. Проводить переговоры с целью выявления требований заказчика к результатам анализа, формировать и согласовывать требования к результатам аналитических работ с использованием технологий больших данных.

2. Проводить презентации при консультировании заказчика, согласовании и утверждении требований к результатам аналитических работ с использованием технологий больших данных

3. Подготавливать документы, регламентирующие требования к результатам аналитического исследования с использованием технологий больших данных в соответствии с существующими регламентами организации.

4. Использовать имеющуюся у исполнителя методологическую и технологическую инфраструктуру анализа больших данных для выполнения аналитических работ.

5. Проводить сравнительный анализ методов и инструментальных средств анализа больших данных.

6. Проводить анализ больших данных в соответствии с утвержденными требованиями к результатам аналитического исследования.

Иметь навыки:

1. Выявления требований заказчика к результатам анализа, определение возможностей применения анализа больших данных в предметной области и конкретных задачах заказчика.

2. Консультирования заказчика по возможностям имеющейся методологической и технологической инфраструктуры анализа больших данных и результатам применения технологий больших данных к аналогичным задачам.

3. Согласование с заказчиком и утверждение требований к результатам аналитического исследования.

Организационно-педагогические условия

Реализация Программы должна обеспечить получение компетенции, необходимой для выполнения нового вида профессиональной деятельности в области информационных технологий Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры; приобретение новой квалификации «Специалист по большим данным».

Учебный процесс организуется с применением электронного обучения, дистанционных образовательных технологий, инновационных технологий и методик обучения, способных обеспечить получение слушателями знаний, умений и навыков в области создания и применения технологий больших данных (Код 06.042).

Реализация Программы обеспечивается научно-педагогическими кадрами Университета, допустимо привлечение к образовательному процессу высококвалифицированных специалистов ИТ-сферы и/или дополнительного профессионального образования в части, касающейся профессиональных компетенций в области создания алгоритмов и программ, пригодных для практического применения, с обязательным участием представителей профильных организаций-работодателей. Возможно привлечение региональных руководителей цифровой трансформации (отраслевых ведомственных и/или корпоративных) к проведению итоговой аттестации, привлечение работников организаций реального сектора экономики субъектов Российской Федерации.

ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ
«РОСТЕЛЕКОМ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ»
ООО «РТК ИТ»

УТВЕРЖДАЮ:
Генеральный директор

_____В.В. Ерохин

УЧЕБНЫЙ ПЛАН
программы профессиональной переподготовки
«Аналитика данных и методы искусственного интеллекта
на базе решений ПАО Ростелеком»

Требования к уровню образования слушателей	лица, имеющие среднее профессиональное или высшее образование; лица, получающие среднее профессиональное или высшее образование
Категория слушателей	
Срок обучения	256 часов
Форма обучения	Очно-заочная, с применением дистанционных образовательных технологий и электронного обучения
Режим занятий	4–8 часов в день

№ раздела	Наименование дисциплины	Трудоемкость		В том числе				Форма контроля
				Аудиторные занятия[4]			Самостоятельная работа	
		В зачетных единицах	В часах	Всего, часов	из них			
					Лекции	Практические занятия		
1	**Модуль 1. Введение в - бизнес-аналитику и искусственный интеллект с применением Python для анализа данных**		**60**	**44**	**10**	**34**	**16**	**Зачет**
2.	**Модуль 2. Методы искусственного интеллекта для анализа табличных данных**		**70**	**52**	**16**	**36**	**18**	**Зачет**

[*] С возможным применением дистанционных образовательных технологий и электронного обучения

3.	**Модуль 3. Современные озера и хранилища данных, аналитика больших данных и методы искусственного интеллекта**		**36**	**28**	**8**	**20**	**8**	**Зачет**
4.	**Модуль 4. Платформы науки о данных и машинного обучения и бизнес аналитики**		**80**	**64**	**16**	**48**	**16**	**Зачет**
	Всего		**246**	**188**	**50**	**138**	**58**	
	Итоговая аттестация		**10**	**10**		**10**		**Выполнение практической работы**
	Общая трудоемкость программы:	**7**	**256**	**198**	**50**	**148**	**58**	

«_____»_____2024 г.

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН

программы профессиональной переподготовки
«Аналитика данных и методы искусственного интеллекта
на базе решений ПАО Ростелеком»»

№ раздела	Наименование дисциплины, модуля	Трудоемкость		В том числе				Форма контроля
				Аудиторные занятия[5]			самостоятельная работа	
					из них			
		В зачетных единицах	В часах	Всего, часов	Лекции	Практические занятия		
М.1	**Модуль 1. Введение в бизнес-аналитику. Python для анализа данных**		**60**	**44**	**10**	**34**	**16**	**Зачет**
1.1.	Введение в Google-таблицы, сводные таблицы Excel		6	4	2	2	2	Тестирование
1.2.	Применение сводных таблиц для маркетинговой сегментации		4	2	-	2	2	Тестирование
1.3.	Создание отчетов в Google Looker Studio		8	6	2	4	2	Решение практических задач
1.4	Применение машинного обучение к данным в Google Таблицах		6	4	-	4	2	Решение практических задач
1.5	Обзор типов данных Pandas.		6	4	-	4	2	Тестирование
1.6.	Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express		8	6	2	4	2	Решение практических задач
1.7	Исследовательский анализ данных (EDA) с использованием pandas		8	6	2	4	2	Решение практических задач
1.8	Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)		12	10	2	8	2	Решение практических задач
	Промежуточная аттестация		2	2		2		Зачет
М.2	**Модуль 2. Методы искусственного интеллекта**		**70**	**52**	**16**	**36**	**18**	**Зачет**

С возможным применением дистанционных образовательных технологий и электронного обучения

	для анализа табличных данных							
2.1.	Машинное обучение для решения задач Data Mining. Линейные модели и градиентный спуск в машинном обучении		8	6	2	4	2	Решение практических задач
2.2.	Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees		8	6	2	4	2	Решение практических задач
2.3.	Бустинг. AdaBoost и градиентный бустинг над решающими деревьями		8	6	2	4	2	Решение практических задач
2.4.	Фреймворки машинного обучения		8	6	2	4	2	Решение практических задач
2.5.	Кластерный анализ, алгоритм k-means, поиск ассоциативных правил		8	6	2	4	2	Решение практических задач
2.6.	Введение в нейронные сети		5	3	1	2	2	Решение практических задач
2.7.	Глубокие нейронные сети		5	3	1	2	2	Решение практических задач
2.8.	Анализ временных рядов		8	6	2	4	2	Решение практических задач
2.9.	Автоматическое машинное обучение (AutoML)		8	6	2	4	2	Решение практических задач
	Промежуточная аттестация		2	2		2		Зачет
М.3	**Модуль 3. Современные озера и хранилища данных, аналитика больших данных и методы искусственного интеллекта**		**36**	**28**	**8**	**20**	**8**	**Зачет**
3.1.	Облачные технологии обработки больших данных		**6**	4	2	2	2	Решение практических задач
3.2.	RT.DataLake		**6**	4	2	2	2	Решение практических задач
3.3.	Маркетинговая аналитика в RT.Warehouse		**10**	8	2	6	2	Решение практических задач
3.4	Решение задач Data Mining в корпоративных хранилищах данных		**12**	10	2	8	2	Решение практических задач
	Промежуточная аттестация		**2**	2		2		Зачет
М.4	**Модуль 4. Платформы науки о данных и машинного обучения и бизнес аналитики**		**80**	**64**	**16**	**48**	**16**	**Зачет**
4.1.	Платформа H2O.ai		**6**	4	2	2	2	Решение практических задач
4.2	Платформа RapidMiner		**8**	6	2	4	2	Решение практических задач

4.3.	Аналитические технологии отечественной платформы Loginom		**12**	10	2	8	2	Решение практических задач
4.4.	Платформа Knime		**8**	6	2	4	2	Решение практических задач,
4.5.	Исследование и визуализация данных в RT.DataVision		**12**	10	2	8	2	Решение практических задач
4.6	Создание интерактивной отчетности в Tableau		**12**	10	2	8	2	Решение практических задач
4.7	Аналитические технологии Power BI		**12**	10	2	8	2	Решение практических задач
4.8	Визуализация данных в Yandex DataLens		**8**	6	2	4	2	Решение практических задач
	Промежуточная аттестация		**2**	2		2		Зачет
М.5	**Итоговая аттестация**		**12**	**12**		**12**		**Выполнение практического задания**
	ИТОГО	**7**	**256**	**198**	**50**	**148**	**58**	

Рабочие программы модулей учебного курса
«Аналитика данных и методы искусственного интеллекта
на базе решений ПАО Ростелеком»

Модуль 1. Введение в бизнес-аналитику. Python для анализа данных.

Цель модуля: приобретение слушателями компетенций, необходимых для понимания и эффективной работы в области анализа больших данных, инструментов и технологий, позволяющих анализировать результаты внутренних процессов организации с помощью Google таблиц и Data Studio, а также инструментами языка Python.

Формируемые компетенции:

способность осуществлять сбор, анализ и обработку данных, необходимых для информационно-аналитического сопровождения деятельности организации;

способность использовать современные информационные технологии в своей деятельности.

умение выявлять бизнес-проблемы или бизнес-возможности.

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН

№№ п/п	Наименование разделов, модулей	Всего		В том числе			Самостоятельная работа	Форма контроля
				Аудиторные занятия				
		В зачетных единицах	В часах	Всего, часов	из них			
					Лекции	Практич занятия		
1	2	3	4	5	6	7	8	9
М.1	**Модуль 1. Введение в бизнес-аналитику. Python для анализа данных**		**60**	**44**	**10**	**34**	**16**	**Зачет**
1.1.	Введение в Google-таблицы, сводные таблицы Excel		6	4	2	2	2	Тестирование
1.2.	Применение сводных таблиц для маркетинговой сегментации		4	2	-	2	2	Тестирование
1.3.	Создание отчетов в Google Looker Studio		8	6	2	4	2	Решение практических задач
1.4	Применение машинного обучение к данным в Google Таблицах		6	4	-	4	2	Решение практических задач
1.5	Обзор типов данных Pandas		6	4	-	4	2	Тестирование

1.6.	Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express		8	6	2	4	2	Решение практических задач
1.7	Исследовательский анализ данных (EDA) с использованием pandas		8	6	2	4	2	Решение практических задач
1.8	Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)		12	10	2	8	2	Решение практических задач
	Промежуточная аттестация		2	2		2		Зачет
	Всего:		**60**	**44**	**10**	**34**	**16**	

Тема 1. Введение в Google-таблицы, сводные таблицы Excel

Назначение Google таблиц и их особенности. Возможности и преимущества Google таблиц для анализа данных, сравнение со возможностями Excel. Создание сводной таблицы данных с помощью автоматических рекомендаций в Таблицах. Сводные таблицы для систематизации данных, выявления закономерностей и упорядочивания информации.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.1. Введение в Google-таблицы, сводные таблицы Excel	Создание Google-таблицы	Использование Google-таблицы	Описание Google-таблицы

Тема 2. Применение сводных таблиц для маркетинговой сегментации

Применение сводных таблиц и их особенности. Возможности и преимущества сводных таблиц для анализа данных. Примеры создание сводных таблиц данных для проведения маркетинговой сегментации:

обобщение больших наборов данных;

проведения анализа больших наборов данных

изучение полученной аналитики данных

представление полученных выводов в удобном для понимания формате

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.2. Применение сводных	Создание сводных таблиц для	Использование сводных таблиц для	Описание маркетинговой сегментации

	таблиц для маркетинговой сегментации	маркетинговой сегментации	маркетинговой сегментации	

Тема 3. Создание отчетов в Google Looker Studio

Многомерное представление данных. Создание источника данных, подключение к внутренним и внешним источникам данных, консолидация источников данных, классификации источников данных по типу данных.

Визуализация данных в Google Looker Studio, как и зачем делать визуализацию данных, загрузка данных в Google Looker Studio, выбор периода визуализации, добавление фильтров, доступы к отчетам, использование готового отчета в качестве шаблона, выводы по использованию Google Looker Studio.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.3. Создание отчетов в Google Looker Studio	Исследование Google Looker Studio	Создание визуализации в Looker Studio	Описание Looker Studio

Тема 4. Применение машинного обучение к данным в Google Таблицах

Применение Tensorflow.js в Google Apps скрипте для проведения машинного обучения в Google Таблицах. Практическое применение методов машинного обучение в Google Таблицах к данным на примере набора данных Boston Housing Prices.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.4. Применение машинного обучение к данным в Google Таблицах	Использование возможностей машинного обучения в Google Таблицах	Создание моделей машинного обучения в Google Таблицах	Описание моделей машинного обучения в Google Таблицах

Тема 5. Обзор типов данных Pandas

Основы программирования на языке Python. История создания и особенности программирования на языке Python. Изучение инструментария языка программирования Python, описание синтаксиса. Базовые типы данных и циклы. Функции и классы. Массивы, множества, словари.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.5. Обзор типов данных Pandas	Использование возможностей Python в анализе данных	Использование языка программирования Python	Описание базовых типов данных

Тема 6. Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express.

Многоплатформенная библиотека визуализации данных, построенная на массивах NumPy и предназначенная для работы с более широким стеком SciPy: основы Matplotlib, структура рисунка, специальные элементы рисунка. Возможность использование в Matplotlib других библиотек. Seaborn - Python-библиотека на основе Matplotlib с предобработкой данных, благодаря тесной интеграции с библиотекой Pandas. Использование библиотеки Altair для создания множества разных статических и интерактивных графиков за несколько строк кода.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1	Тема 1.6. Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express	Исследование статических и интерактивных графиков	Использование возможностей Python-библиотек Matplotlib, Seaborn, Altair	Описание популярных Python-библиотек визуализации.

Тема 7. Исследовательский анализ данных (EDA) с использованием pandas

Многоплатформенная библиотека визуализации данных, построенная на массивах NumPy и предназначенная для работы с более широким стеком SciPy: основы Matplotlib, структура рисунка, специальные элементы рисунка. Возможность использование в Matplotlib других библиотек. Seaborn - Python-библиотека на основе Matplotlib с предобработкой данных, благодаря тесной интеграции с библиотекой Pandas. Использование библиотеки Altair для создания множества разных статических и интерактивных графиков за несколько строк кода.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1.	Тема 7. Исследовательский анализ данных (EDA) с использованием pandas	Исследование статических и интерактивных графиков	Использование возможностей Python-библиотек Matplotlib, Seaborn, Altair	Описание популярных Python-библиотек визуализации.

Тема 8. Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)

Рассмотрим разведочный анализ данных с использованием инструментов реализации анализа, библиотек автоматизации EDA Python (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib). В последние годы появилось несколько мощных библиотек python с низким уровнем кода, которые значительно ускоряют и упрощают этап исследования данных и анализа проектов. Примеры методов EDA в соответствии с ситуацией и доступными типами данных с применением библиотек автоматизации EDA.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
1.	Тема 8. Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)	Исследование инструментов реализации анализа EDA	Использование возможностей библиотек автоматизации EDA Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib	Описание библиотек автоматизации EDA

Содержание самостоятельной работы слушателей

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

№ темы	Наименование (содержание) темы, по которой предусмотрена самостоятельная работа	Формы и методы проведения
Тема 1	Введение в Google-таблицы, сводные таблицы Excel	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 2	Применение сводных таблиц для маркетинговой сегментации	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 3	Создание отчетов в Google Looker Studio	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 4	Применение машинного обучение к данным в Google Таблицах	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 5	Обзор типов данных Pandas	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 6	Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 7	Исследовательский анализ данных (EDA) с использованием pandas	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 8	Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)	Изучение основной и дополнительной литературы по программе; разбор примеров

Форма контроля

В процессе обучения осуществляется текущий и промежуточный контроль знаний. Текущий в виде решения типовых задач, промежуточный – выполнение практического задания по модулю.

Модуль 2. Машинное обучение на Python

Цель модуля: приобретение слушателями компетенций, необходимых для эффективной работы в области анализа больших данных машинного обучения; изучение инструментов и технологий создания, обучения, оценки и развертывания моделей машинного обучения на Python.

Формируемые компетенции:

способность осуществлять сбор, анализ и обработку данных, необходимых для информационно-аналитического сопровождения деятельности организации;

способность осуществлять сбор информации о бизнес-проблемах и бизнес-возможностях;

умение выявлять бизнес-проблемы или бизнес-возможности;

умение - анализировать, обосновывать и выбирать решение.

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН

№№ п/п	Наименование разделов, модулей	Всего		В том числе			Самостоятельная работа	Форма контроля
				Аудиторные занятия				
					из них			
		В зачетных единицах	В часах	Всего, часов	Лекции	Практич занятия		
1	2	3	4	5	6	7	8	9
М.2	**Модуль 2. Методы искусственного интеллекта для анализа табличных данных**		**70**	**52**	**16**	**36**	**18**	**Зачет**
2.1.	Машинное обучение для решения задач Data Mining. Линейные модели и градиентный спуск в машинном обучении		8	6	2	4	2	Решение практических задач
2.2.	Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees		8	6	2	4	2	Решение практических задач
2.3.	Бустинг. AdaBoost и градиентный бустинг над решающими деревьями		8	6	2	4	2	Решение практических задач
2.4.	Фреймворки машинного обучения		8	6	2	4	2	Решение практических задач
2.5.	Кластерный анализ, алгоритм k-means, поиск ассоциативных правил		8	6	2	4	2	Решение практических задач
2.6.	Введение в нейронные сети		5	3	1	2	2	Решение

								практических задач
2.7.	Глубокие нейронные сети		5	3	1	2	2	Решение практических задач
2.8.	Анализ временных рядов		8	6	2	4	2	Решение практических задач
2.9.	Автоматическое машинное обучение (AutoML)		8	6	2	4	2	Решение практических задач
	Промежуточная аттестация		2	2		2		Зачет
	Всего:		**70**	**52**	**16**	**36**	**18**	

Тема 1. Машинное обучение для решения задач Data Mining. Линейные модели и градиентный спуск в машинном обучении

Основная задача машинного обучения. Приложения на основе машинного обучения. Жизненный цикл машинного обучения. Виды и алгоритмы обучения. Оценки моделей, метрики. Обучающие, тестовые и валидационные множества, кросс-валидация. Исследовательский анализ данных, классовый дисбаланс, очистка данных и масштабирование данных. Градиентный метод в машинном обучении. Обучение и функция потерь. Минимизация потерь: итерационный подход. Градиентный спуск, стохастический градиентный спуск. Градиентный спуск с линейной регрессией. Регуляризация.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2	Тема 2.1. Машинное обучение для решения задач Data Mining. Линейные модели и градиентный спуск в машинном обучении	Оценка нескольких моделей. Решение регрессионной задачи с помощью библиотеки scikit-learn	Создание, обучение и оценка нескольких моделей. Интерпретация моделей. Инжиниринг признаков. Импорт и загрузка датасета. Очистка набор данных с помощью Pandas, создание моделей машинного обучения с помощью scikit-learn. Визуализация выходных данных с помощью Matplotlib	Методика CRISM-DM, основные алгоритм машинного обучения с учителем. Градиентный спуск с линейной регрессией. Регуляризация. Метрики качества решения регрессионной задачи, сравнение различных регрессионных моделей.

Тема 2. Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees

Алгоритмы построения деревьев решений, являющихся одним из наиболее эффективных инструментов интеллектуального анализа данных и предсказательной

аналитики, которые позволяют решать задачи классификации и регрессии, критерии разделения: прирост информации, Джини. Алгоритм C4.5. Алгоритм CART. Обработка пропущенных значений, стрижка, регуляризация. Сильные и слабые стороны деревьев решений. Общая идея разложения ошибки на смещение и разброс. Композиции алгоритмов. Бэггинг и метод случайных подпространств. Random Forest, Extremely randomized trees. Сильные и слабые стороны Random Forest.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.2. Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees	Исследование алгоритмов построения деревьев решений. Исследование стратегии бэггинга для решения задач Data Mining	Импорт и загрузка набора данных. Создание, обучение и сравнение ансамблей моделей машинного обучения с помощью scikit-learn. Решение задач машинного с использованием алгоритма Random Forest	Метрики качества решения задачи бинарной классификации, сравнение моделей, основанных на стратегии деревьях решений. Описание метода разложения ошибки на смещение и разброс. Сильные и слабые стороны Random Forest.

Тема 3. Бустинг. AdaBoost и градиентный бустинг над решающими деревьями

Применение бустинга для уменьшения смещения. Семейство алгоритмов машинного обучения, преобразующих слабые обучающие алгоритмы к сильным. AdaBoost и градиентный бустинг над решающими деревьями. Различные имплементации градиентного бустинга. Стратегии стекинга.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.3 Бустинг. AdaBoost и градиентный бустинг над решающими деревьями	Исследование алгоритмов бустинг	Использование алгоритмов градиентного бустинга для решения задач классификации и регрессии	Описание алгоритмов бустинга. Настраиваемые параметры градиентного бустинга в библиотеке scikit-learn для решения задач классификации и регрессии.

Тема 4. Фреймворки машинного обучения.

Фреймворки машинного обучения: XGBoost, LightGBM, CatBoost, h2o.ai, scikit-learn, TensorFlow

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.4. Фреймворки машинного обучения	Сравнение фреймворков машинного обучения.	Использование фреймворков машинного обучения для решения задач Data Mining.	Главные (основные) фреймворки машинного обучения, используемые в продакшене.

Тема 5. Кластерный анализ, алгоритм k-means, поиск ассоциативных правил

Введение в кластерный анализ, алгоритм k-means. Самоорганизующиеся сети Кохонена, алгоритм функционирования самообучающихся карт. Рассмотрим применение метода поиска ассоциативных правил, с помощью трех типов алгоритмов.

1. Apriori.
2. Eclat.
3. FP Growth.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.5. Кластерный анализ, алгоритм k-means, поиск ассоциативных правил	Исследование кластерного анализа и метода поиска ассоциативных правил	Использование кластерного анализа для решения задач Data Mining и метода поиска ассоциативных правил	Описание алгоритма кластерного анализа k-means и его модификаций, описание алгоритмов Apriori, Eclat, FP Growth

Тема 6. Введение в нейронные сети.

Введение в нейронные сети. Искусственный нейрон, функции активации, многослойный персептрон. Метод обратного распространения ошибки. Метод обратного распространения ошибки. Примеры применения нейронных сетей для решения задач с неизвестным алгоритмом решения

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.6. Введение в нейронные сети	Исследование многослойного	Использование многослойного	Описание нейронных сетей, метод обратного

		персептрона	персептрона для решения задач классификации и регрессии	распространения ошибки.

Тема 7. Глубокие нейронные сети

Глубокие нейронные сети: сверточные и рекуррентные нейронные сети, Long-Short-Term-Memory (LSTM), Transformers.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.7. Глубокие нейронные сети.	Исследование глубоких нейронных сетей.	Использование глубоких нейронных сетей для решения задача распознавания образов, работа с временными рядами и последовательностями	Описание глубоких нейронных сетей, виды архитектур глубоких нейронных сетей и классы решаемых ими задач

Тема 8. Анализ временных рядов.

Рассмотрим некоторые основные понятия в теории анализа временных рядов, классические статистические алгоритмы прогнозирования, а также рассмотрим применение моделей глубоких нейросетей для решения таких задач. Применение нейронных сетей для предсказания временных рядов.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
2.	Тема 2.8 Анализ временных рядов	Исследование анализа временных рядов	Использование анализа временных рядов для решения задач Data Mining.	Описание анализа временных рядов

.

Тема 9. Автоматическое машинное обучение (AutoML)

Автоматическое машинное обучение (AutoML): использование элементов Auto Sklearn, Tree-Based Pipeline Optimization Tool (TPOT), Auto Keras, AutoGluon. Решение задач классификации и регрессии на структурированных данных. Решение задач на неструктурированных данных: Computer Vision, Natural Language Processing

Содержание практических занятий

№ модул	Наименование темы (раздела)	Тема практического	Содержание практического	Вопросы к практическому занятию

я	дисциплины	занятия	занятия	
9.	Тема 2.9. Автоматическое машинное обучение (AutoML)	Основные фреймворки машинного обучения с функцией AutoML.	Использование и сравнение фреймворков машинного обучения с фикцией AutoML.	Использование Auto Sklearn, Tree-Based Pipeline Optimization Tool (TPOT), Auto Keras, AutoGluon. Решение задач классификации и регрессии на структурированных данных.

Содержание самостоятельной работы слушателей

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

№ темы	**Наименование (содержание) темы, по которой предусмотрена самостоятельная работа**	**Формы и методы проведения**
Тема 1	Машинное обучение для решения задач Data Mining. Линейные модели и градиентный спуск в машинном обучении	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 2	Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 3	Бустинг. AdaBoost и градиентный бустинг над решающими деревьями	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 4	Фреймворки машинного обучения	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 5	Кластерный анализ, алгоритм k-means, поиск ассоциативных правил	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 6	Введение в нейронные сети	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 7	Глубокие нейронные сети	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 8	Анализ временных рядов	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 9	Автоматическое машинное обучение (AutoML)	Изучение основной и дополнительной литературы по программе; разбор примеров

Форма контроля

В процессе обучения осуществляется текущий и промежуточный контроль знаний. Текущий в виде решения типовых задач, промежуточный – выполнение практического задания по модулю.

Модуль 3. Современные озера и хранилища данных, аналитика больших данных и методы искусственного интеллекта

Цель модуля: приобретение слушателями компетенций в области локальных корпоративных хранилищ данных, современных технологий Big Data; языка программирования SQL для аналитики больших данных и облачных технологий обработки больших данных.

Формируемые компетенции:

способность осуществлять сбор, анализ и обработку данных, необходимых для информационно-аналитического сопровождения деятельности организации;

умение использовать современные информационные технологии в своей деятельности;

умение выявлять бизнес-проблемы или бизнес-возможности;

умение обосновывать решения.

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН

№№ п/п	Наименование разделов, модулей	Всего		В том числе			Самостоятельная работа	Форма контроля
				Аудиторные занятия				
		В зачетных единицах	В часах	Всего, часов	из них			
					Лекции	Практич занятия		
1	2	3	4	5	6	7	8	9
М.3	**Модуль 3. Современные озера и хранилища данных, аналитика больших данных и методы искусственного интеллекта**		**36**	**28**	**8**	**20**	**8**	**Зачет**
3.1.	Облачные технологии обработки больших данных		**6**	4	2	2	2	Решение практических задач
3.2.	RT.DataLake		**6**	4	2	2	2	Решение практических задач
3.3.	Маркетинговая аналитика в RT.Warehouse		**10**	8	2	6	2	Решение практических задач
3.4	Решение задач Data Mining в корпоративных хранилищах данных		**12**	10	2	8	2	Решение практических задач
	Промежуточная аттестация		**2**	2		2		Зачет
	Всего:		**36**	**28**	**8**	**20**	**8**	

Тема 1. Облачные технологии обработки больших данных.

Облачные технологии обработки больших данных. Озера данных и современные хранилища данных. Типы данных, функции и операторы SQL

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
3.	Тема 3.1. Облачные технологии обработки больших данных.	Исследование облачных технологий	Использование простых запросов в SQL	Решение задач анализа больших данных с помощью технологий облачных хранилищ данных

Тема 2. RT.DataLake

Выполнение аналитических запросов и трансформацию данных в RT.DataLake с помощью реализованных механизмов MapReduce, Spark, TEZ. Подготовка данных для использования в моделях машинного обучения и для исследования данных, профилирования или построения аналитических отчетов.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
3.	Тема 3.2. RT.DataLake	Исследование RT.DataLake	Использование механизмов MapReduce, Spark, TEZ	Подготовка данных для использования в моделях машинного обучения в RT.DataLake

Тема 3. Маркетинговая аналитика в RT.Warehouse.

Машинное обучение в RT.Warehouse, выполнение аналитических запросов с помощью подключения к внешним источникам без перегрузки данных в хранилище. Поддерживается интеграция с Oracle, Postgres, MS SQL, MySQL, MongoDB, SAP HANA, Hadoop. Использование RT.Warehouse для проведения BI-аналитики и Data Science.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
3	Тема 3.3. Маркетинговая аналитика в RT.Warehouse	Решение задач Data Mining с помощью машинного обучения в RT.Warehouse	Разработка регрессионной модели, модели классификации и кластеризации в RT.Warehouse	Выполнение аналитических запросов с помощью подключения к внешним источникам

Тема 4. Решение задач Data Mining в корпоративных хранилищах данных.

Единая облачная платформа для крупномасштабного проектирования данных и совместной работы. Запросы озер данных с помощью SQL, оптимизированная среда машинного обучения с открытым кодом, управление жизненным циклом машинного обучения с помощью MLFlow. Интеграция популярных фреймворков машинного обучения

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
3	Тема 3.4. Решение задач Data Mining в корпоративных хранилищах данных	Исследование Data Mining в корпоративных хранилищах данных	Решение задач ETL и машинного обучения в корпоративных хранилищах данных	Задачи машинного обучения, решаемые в облачных хранилищах данных

Содержание самостоятельной работы слушателей

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

№ темы	**Наименование (содержание) темы, по которой предусмотрена самостоятельная работа**	**Формы и методы проведения**
Тема 1	Облачные технологии обработки больших данных	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 2	RT.DataLake	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 3	Маркетинговая аналитика в RT.Warehouse	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 4	Решение задач Data Mining в корпоративных хранилищах данных	Изучение основной и дополнительной литературы по программе; разбор примеров

Форма контроля

В процессе обучения осуществляется текущий и промежуточный контроль знаний. Текущий в виде решения типовых задач, промежуточный – прохождение тестирования по модулю.

Модуль 4. Платформы науки о данных и машинного обучения и бизнес аналитики

Цель модуля: приобретения слушателями компетенций для эффективного применения различных платформ машинного обучения и искусственного интеллекта при работе с анализом больших данных и выработки возможных решений посредством сбора и анализа больших данных и элементами информации бизнес-анализа.

Формируемые компетенции:

способность выявлять бизнес-проблемы или бизнес-возможности;
умение - обосновывать решения
способность использовать современные информационные технологии в своей деятельности;
умение выявлять бизнес-проблемы или бизнес-возможности.
.

УЧЕБНО-ТЕМАТИЧЕСКИЙ ПЛАН

№№ п/п	Наименование разделов, модулей	Всего		В том числе				Форма контроля
				Аудиторные занятия			Самостоятельная работа	
		В зачетных единицах	В часах	Всего, часов	из них			
					Лекции	Практич занятия		
1	2	3	4	5	6	7	8	9
М.4	**Модуль 4. Платформы науки о данных и машинного обучения и бизнес аналитики**		**80**	**64**	**16**	**48**	**16**	**Зачет**
4.1.	Платформа H2O.ai		**6**	4	2	2	2	Решение практических задач
4.2	Платформа RapidMiner		**8**	6	2	4	2	Решение практических задач
4.3.	Аналитические технологии отечественной платформы Loginom		**12**	10	2	8	2	Решение практических задач
4.4.	Платформа Knime		**8**	6	2	4	2	Решение практических задач,
4.5.	Исследование и визуализация данных в RT.DataVision		**12**	10	2	8	2	Решение практических задач
4.6	Создание интерактивной отчетности в Tableau		**12**	10	2	8	2	Решение практических задач
4.7	Аналитические технологии		**12**	10	2	8	2	Решение

	Power BI							практических задач
4.8	Визуализация данных в Yandex DataLens		**8**	6	2	4	2	Решение практических задач
	Промежуточная аттестация		**2**	2		2		Зачет
	Всего:		**80**	**64**	**16**	**48**	**16**	

Тема 1. Платформа H2O.ai.

Возможности платформы H2O.ai. Ключевые особенности: ведущие алгоритмы, доступ из python, R, Flow. AutoML, распределенная обработка в памяти, бесшовное развертывание модели. Sparkling Water - усовершенствованное машинное обучение для Spark. H2O Driverless AI: автоматические проектирование признаков и визуализаций, интерпретируемость машинного обучения, NLP с TensorFlow, конвейер машинного обучения, временные ряды.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.1. Платформа H2O.ai.	Возможности платформы H2O.ai	Разработка, обучение и развертывание моделей машинного обучения с технологиями H2O.ai	Алгоритмы и линейка продуктов H2O.ai

Тема 2. Платформа RapidMiner.

Технология RapidMiner Turbo Prep для очистки и подготовки данных, AutoML для автоматического проектирования признаков, автоматического выбора модели, настройки гиперпараметров и интерпретация результатов машинного обучения.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.2. Платформа RapidMiner.	Исследование возможностей платформы RapidMiner	Разработка сценариев рабочих процессов решений задач в RapidMiner	Применение RapidMiner Turbo Prep и AutoML для принятия взвешенных бизнес-решений по предоставленным данным

Тема 3. Аналитические технологии отечественной платформы Loginom

Существующие программные решения для OLAP-моделирования. Многомерный анализ – OLAP-кубы в платформе Loginom. Применение OLAP при решении аналитических задач: разведочный анализ, исследование данных, аналитическая отчетность, финансовый анализ и др. в платформе Loginom. Введение в маркетинговую аналитику. KPI и метрики. Системы аналитики и сбор данных в

платформе Loginom. Методы сегментации клиентов и целевой аудитории. Сценарии выполнения ABC, XYZ, ABC-XYZ, RFF анализа в платформе Loginom.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.3. Аналитические технологии Loginom	Исследование технологий Loginom. Применение платформы Loginom в клиентской и маркетинговой аналитике	Построение и анализ многомерных кубов в платформе Loginom. Разработка сценариев ABC, XYZ, ABC-XYZ, RFF анализа в платформе Loginom	Применение OLAP при решении аналитических задач Методы ABC-XYZ, RFF- анализа. Виды обработчиков (мастеров) в платформе Loginom

Тема 4. Платформа Knime.

Анализа, интеграция данных и подготовки отчётности в платформе с открытым исходным кодом Knime. Объединение различных компонентов для машинного обучения и интеллектуального анализа данных с помощью концепции модульной конвейерной обработки данных «Lego of Analytics». Создание сквозных рабочих процессов в науке о данных, смешивание разных инструментов, очистка и предобработка данных, использование машинного обучения и искусственного интеллекта, совместная работа, интеграция с фреймворками машинного обучения с открытым кодом.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.4 Платформа Knime	Исследование платформы Knime	Разработка сценариев рабочих процессов решений задач в Knime	Решение задачи комплексной аналитики в десктопной платформе Knime Analytics

Тема 5. Исследование и визуализация данных в RT.DataVision

Принципы работы, функциональные возможности и основные компоненты RT.DataVision. RT.DataVision – BI-решение на базе Apache Superset. Анализ, интеграция данных и подготовки отчётности в платформе с RT.DataVision. Подключение к данным, преобразование и формирование данных, создание модели, визуализаций и отчетов, информационных панелей мониторинга, совместная работа в RT.DataVision.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.5 Исследование и визуализация данных в RT.DataVision	Исследование платформы RT.DataVision	Разработка визуализаций и отчетов, информационных панелей мониторинга в RT.DataVision.	Создание модели данных Формирование интерактивных отчётов и информационных панелей мониторинга в RT.DataVision.

Тема 6. Создание интерактивной отчетности в Tableau

Язык визуальных запросов VizQL. Технология Data Engine компании Tableau. Технология Hyper: генерация динамического кода и методы параллелизма для достижения высокой производительности при создании экстрактов и выполнении запросов. Ключевые преимущества Tableau и функционал Tableau. Источники данных и подключения. Визуальный анализ и вычисления. Использование параметров. Создание дашбордов и форматирование.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема 4.6. Создание интерактивной отчетности в Tableau	Исследование технологий Tableau	Разработка аналитических приложений бизнес-пользователями с применением Tableau	Подключение к источникам данных, очистка и трансформация данных с применением технологий Tableau

Тема 7. Аналитические технологии Power BI

Назначение и функциональные блоки в Power BI Desktop. Назначение облачного сервиса аналитики PowerBI.com. Ключевые отличия и преимущества Power BI, возможности Power BI по обработке больших данных. Подключение к данным, преобразование и формирование данных, создание модели, визуализаций и отчетов, информационных панелей мониторинга, совместная работа в Power BI. DAX и язык M. Обработка естественного языка, технология вопрос и ответов Q&A в Power BI.

Содержание практических занятий

№ моду	Наименование темы (раздела)	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию

ля	дисциплины			
4	Тема 4.7. Технологии Power BI	Исследование технологий Power BI	Разработка аналитических приложений конечными пользователями на базе платформы Power BI	Консолидация данных, создание модели данных, создание новых мер – ключевых показателей эффективности, формирование интерактивных отчётов и информационных панелей мониторинга.

Тема 8. Визуализация данных – Yandex DataLens

Визуализация данных и бизнес-аналитика: создание в несколько кликов графиков, чтобы быстро проверить гипотезу, разработка полноценного дашборда для мониторинга ключевых бизнес-метрик. Работа с разными источниками данных: подключение к своим облачным и локальным базам данных, сервисам и плоским файлам, комбинирование данных из разных источников в одном дашборде. Геоаналитика на Яндекс.Картах: использование возможностей Яндекс.Карт для корпоративной аналитики, подключение своего ключа API Яндекс карт для продвинутых возможностей геокодинга. Добавление учётных записей команды или даже внешних партнёров для совместной работы.

Содержание практических занятий

№ модуля	Наименование темы (раздела) дисциплины	Тема практического занятия	Содержание практического занятия	Вопросы к практическому занятию
4.	Тема .4.8. Визуализация данных – Yandex DataLens	Исследование технологий Yandex DataLens	Разработка аналитических приложений бизнес-пользователями с применением Yandex DataLens	Подключение к источникам данных, виды визуализаций и методика создания дашбордов в Yandex DataLens

Содержание самостоятельной работы слушателей

Основная цель самостоятельной работы слушателей – закрепление знаний, полученных в ходе лекционных и практических занятий.

№ темы	**Наименование (содержание) темы, по которой предусмотрена самостоятельная работа**	**Формы и методы проведения**
Тема 1	Платформа H2O.ai	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 2	Платформа RapidMiner	Изучение основной и дополнительной литературы по программе; разбор примеров
Тема 3	Платформа Knime	Изучение основной и дополнительной литературы по программе; разбор примеров

Тема 4	Платформа Trifacta	Изучение основной и дополнительной литературы по программе; разбор примеров

Форма контроля

В процессе обучения осуществляется текущий и промежуточный контроль знаний. Текущий в виде решения типовых задач, промежуточный – прохождение тестирования по модулю.

Список литературы учебного курса
«Аналитик данных»

Основная литература

1. Data Science. Наука о данных с нуля. / Билл Фрэнкс.; пер. с англ. Евстигнеева И.В. – М.: Издательство «Альпина Паблишер». – 2018. – 320 с.
2. Набатова Д. С. Математические и инструментальные методы поддержки принятия решений: учебник и практикум для бакалавриата и магистратуры / Д.С. Набатова. – Москва: Юрайт, 2016. – 292 с. – То же [Электронный ресурс]. –2018.– Режим доступа: **https://urait.ru/book/matematicheskie-i-instrumentalnye-metody-podderzhki-prinyatiya-resheniy-469195.**
3. Курносов Ю.В. «Азбука аналитики», Издательство «Концептуал», 2018 -240 с.
4. Б. Марр «Ключевые инструменты бизнес-аналитики»/ пер с англ. Егоров В. Н., Издательство «Лаборатория знаний», 2018 – 339 с.
5. де Прадо М. «Машинное обучение: алгоритмы для бизнеса», Санкт Петербург: Издательский дом «Питер». – 2019. – 432 с.
6. Плас вандер Д. «Python для сложных задач: наука о данных и машинное обучение», Санкт Петербург: Издательский дом «Питер». – 2018. – 576 с.
7. Лакшманан В., Тайджани Д. «Google BigQuery. Всё о хранилищах данных, аналитике и машинном обучении», Санкт Петербург: Издательский дом «Питер». – 2021. – 496с.

Дополнительная литература

1. Глубокое обучение на Python. – СПб.: Питер, 2018. – 400 с.: ил. – (Серия «Библиотека программиста»).
2. Основы Data Science и Big Data. Python и наука о данных. / Силен Д., Мейсман А., Али М.; пер. с англ. – Санкт Петербург: Издательский дом «Питер». – 2018. – 336 с.
3. Радченко И.А, Николаев И.Н. Технологии и инфраструктура Big Data. – СПб: Университет ИТМО, 2018. – 52 с.
4. Data Science. Наука о данных с нуля. / Джоэл Грас.; пер. с англ. Логунов А.В. – Санкт Петербург: Издательство «БХВ-Петербург». – 2018. – 336 с.
5. Google BigQuery. Всё о хранилищах данных, аналитике и машинном обучении. – СПб.: Питер, 2021. – 496 с.: ил.
6. К. Андерсон «Аналитическая культура»/Издательство. Манн, Иванов, Фербер, 2017 – 332 с.
7. Д. Битти К. Вигерс «Разработка требований к программному обеспечению», Издательство BHV, 2019 -737 с.

Описание системы оценки качества освоения программы

Результаты входного тестирования, выполнения кейсов и практико-ориентированных заданий, тестирования в рамках текущего контроля успеваемости, промежуточной и итоговой аттестации являются показателями цифрового следа в уровне сформированности общепрофессиональных и профессиональных компетенций по программе.

Контроль результатов освоения программы профессиональной переподготовки осуществляется в ходе текущего контроля успеваемости, промежуточной аттестации и итоговой аттестации.

1. **Текущий контроль успеваемости** осуществляется в процессе изучения слушателями учебного материала в форме выполнения практических заданий и разбора практических ситуаций по каждой теме в личном кабинете слушателя.

2. Промежуточная аттестация проводится в форме зачета в виде тестирования или в виде выполнения практической работы с выставлением оценки– «зачтено»; «незачтено».

Критерии оценивания промежуточной аттестации в форме зачета.

Порядок проведения тестирования: тестирование проводится с личного компьютера слушателя, 20 тестовых вопросов по отдельным модулям, 60 мин., количество попыток – 2 по каждому модулю.

Критерии выставления оценки за промежуточное тестирование приведены в таблице:

Количество правильных ответов при тестировании	**Оценка за промежуточное тестирование по дисциплине**
≥50%	незачтено
<50%	зачтено

Порядок выполнения практической работы: практические работы по модулям курса выполняются слушателями на личных компьютерах, в требуемом программном обеспечении и выкладываются слушателями в личном кабинете СДО в виде ссылок.

Критерии оценивания: критерии выставления оценки за выполнение практической работы приведены в таблице:

Критерии к выставлению оценки	**Оценка за выполнение практической работы по модулям**
оценка выставляется при наличии серьезных ошибок и пробелов в знаниях при выполнении практической работы	незачтено
оценка выставляется при выполнении практической работы в соответствующем программном обеспечении в полном соответствии со всеми пунктами задания; а также ответы имеющие некоторые неточности при выполнении и описании работы	зачтено

3. Итоговая аттестация

После успешного освоения всех модулей программы и успешного прохождения промежуточной аттестации, для слушателей, завершающих обучение обязательной является итоговая аттестация.

Проведение итоговой аттестации. Итоговая аттестация проводится в решения и защиты кейса.

Итоговая аттестация состоит из выбора индивидуального кейса (набор данных) и решения заданий, которые охватывают все практические методы, подходы в соответствующих программных продуктах, применяемые в машинном обучении и анализе больших данных рассмотренные и изученные в соответствующих темах в рамках учебной программы. Решение кейса включает в себя три задания – два обязательных задания и одно дополнительное (на выбор) задание. В ходе решения

В результирующую оценку по итоговой аттестации входит оценка уровня сформированности у слушателя универсальных и профессиональных компетенций, а также оценки собственно результата/продукта, полученного в ходе выполнения и защиты кейса.

Порядок проведения итоговой аттестации: размещение письменного ответа на задания кейса в СДО с последующим выступлением и защитой кейса (в виде видео или онлайн защиты) с выставлением оценки по 4 балльной шкале: "неудовлетворительно"; "удовлетворительно", "хорошо", "отлично".

Критерии оценивания: для выставления оценки по итоговой аттестации необходимо пользоваться следующими критериями, приведенными в таблице.

Критерии оценки итоговой аттестационной работы.	**Критерии к выставлению оценки / Оценка за итоговую аттестационную работу по программе**
оценка выставляется при наличии серьезных ошибок и пробелов в знаниях в практической работе	неудовлетворительно
оценка выставляется при наличии отдельных неточностей в ответах, при неполных ответах на задания, частичном выполнении заданий или при наличии замечаний к практической работе непринципиального характера (описки, случайные ошибки арифметического характера, грамматические ошибки)	удовлетворительно
оценка выставляется при наличии верных ответов на все задания в ходе выполнения практической работы, при грамотном выполнении всех заданий, но при отсутствии отличительных признаков, как, например: детальных выкладок или пояснений, качественного оформления работы	хорошо
оценка выставляется при четком достижении цели и выполнении всех заданий практической работы, то есть при наличии полных (с детальными пояснениями выкладок и выводов), оригинальных и правильных решений, а также при полных и развернутых ответах на комментарии преподавателя и итоговой аттестационной комиссии	отлично

Оценочные материалы:

Тестовые вопросы для промежуточной и итоговой аттестации, практико-ориентированные задания и кейсы по модулям:

Входное тестирование.

1. Правильная последовательность в Business Intelligence:
   a) данные-информация-знания-принятие решения
   b) информация-данные-знания-принятие решения
   c) принятие решения-информация-данные-знания

2. В платформе для бизнес-анализа должны быть реализованы:
   a) 10 ключевых возможностей
   b) 15 ключевых возможностей
   c) 20 ключевых возможностей

3. Перечислите правильную последовательность этапов Knowledge Discovery in Databases – процесса обнаружения знаний в базах данных
   a) трансформация, интерпретация результатов, выборка, очистка, построение моделей.
   b) построение моделей, выборка, очистка, трансформация, интерпретация результатов.
   c) выборка, очистка, трансформация, построение моделей, интерпретация результатов.

4. OLAP-системы это:
   a) информационные системы многомерного анализа данных в реальном времени.
   b) информационные системы автоматической обработки данных.
   c) информационные системы алгоритмической обработки данных.

5. OLTP-системы это:
   a) информационные системы оперативной транзакционной обработки данных
   b) информационные системы оперативного анализа данных
   c) информационные системы автоматической обработки данных

6. Если для реализации многомерной модели используют реляционные базы данных, то способ реализации гиперкуба называется
   a) MOLAP
   b) ROLAP
   c) HOLAP

7. Если для реализации многомерной модели используют и многомерные, и реляционные базы данных, то способ реализации гиперкуба называется
   a) MOLAP
   b) ROLAP

c) HOLAP

8. Информационные хранилища созданы для удобства …
a) руководителей всех уровней для принятия решений
b) предметных приложений
c) редактирования данных

9. Большинство методов Data Mining были разработаны в рамках…
a) Теории искусственного интеллекта
b) Классического анализа данных
c) Теории баз данных

10. Классификация – …
a) разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов
b) высокоуровневые средства отражения информационной модели и описания структуры данных
c) это установление зависимости дискретной выходной переменной от входных переменных

Практико-ориентированные задания и кейсы по модулям.

Модуль 1. Введение в бизнес-аналитику и искусственный интеллект с применением Python для анализа данных

1.1. Введение в Google-таблицы, сводные таблицы Excel

По индивидуальному заданию (датасет) провести быструю аналитика в таблицах Google:

– найдите закономерности в данных с помощью автоматически формируемых диаграмм.

1.2. Применение сводных таблиц для маркетинговой сегментации

По индивидуальному заданию (датасет)

– создайте сводные таблиц на основе своих данных;

– визуализируйте подмножества данных, выбирая только нужные столбцы или ячейки.

1.3. Создание отчетов в Google Looker Studio

По индивидуальному заданию (датасет) создать интегративный отчет (дашборд):

– создать новый пустой отчет, добавить источник данных в отчет, добавить таблицу (с помощью панели инструментов), добавить график временных рядов (с помощью меню), настроить стиль отчета, добавить баннер и добавить заголовок к отчету;

– добавить диаграмму и настроить различные диаграммы, гистограммы и карты;

– настроить режим просмотра отчета и поделиться им.

1.4. Применение машинного обучение к данным в Google Таблицах

1.5. Обзор типов данных Pandas.

Изучить особенности языка Python для аналитики, работа с окружением Python и основными требуемыми инструментами, и библиотеками; работу с базовыми концепциями, понятиями, принципами и возможностями Python.

Рассмотреть основные структуры данных в Pandas классы Series и DataFrame. Импорт и чтение данных, обобщенная информация и краткая статистка, изменение типов и сортировка данных, индексация и извлечение данных, сводные таблицы, группирование данных, преобразование датафреймов, визуализации.

1.6. Библиотеки визуализации данных Matplotlib, Seaborn, Altair, Plotly Express

По индивидуальному датасету:

– применить библиотеки Matplotlib для создания графиков, гистограмм, спектров мощности, круговых диаграмм и др.

– применить библиотеки Seaborn и изучить особенности данной библиотеки, применяемой для визуализации данных.

– применить декларативную библиотеку Altair для создания эффективных визуализаций с минимальным количеством кода.

1.7. Исследовательский анализ данных (EDA) с использованием pandas

По индивидуальному заданию (датасет) выполнить исследовательский анализ данных (EDA) с помощью библиотеки pandas:

– изучить распределения данных в виде статистик и визуализаций;

– обработать отсутствующие значений набора данных (наиболее частая проблема с каждым набором данных);

– найти и обработать выбросы и аномалии;

– найти и удалить повторяющиеся данных (устранить дубликаты);

– провести кодирование категориальных переменных;

– выполнить нормализацию и масштабирование числовых признаков.

1.8. Разведочный анализ данных с использованием библиотек автоматизации EDA (Pandas Profiling, Sweetviz, Dataprep, D-Tale, Mitosheet, Bamboolib)

По индивидуальному заданию (датасет) выполнить исследовательский анализ данных (EDA) с помощью любой библиотеки с автоматизацией EDA:

– изучить распределения данных в виде статистик и визуализаций;

– обработать отсутствующие значений набора данных (наиболее частая проблема с каждым набором данных);

– найти и обработать выбросы и аномалии;

– найти и удалить повторяющиеся данных (устранить дубликаты);

– провести кодирование категориальных переменных;

– выполнить нормализацию и масштабирование числовых признаков.

Модуль 2. Методы искусственного интеллекта для анализа табличных данных

2.1. Машинное обучение для решения задач Data Mining. Градиентный спуск в машинном обучении

Изучение данных NASA по климату и их визуализация с помощью библиотеки Matplotlib. Выполнение линейной регрессии с помощью NumPy. Выполнение линейной регрессии с помощью scikit-learn, визуализация результатов с использованием библиотеки Seaborn и их анализ.

2.2. Алгоритмы построения деревьев решений, критерии разделения. Бэггинг, Random Forest, Extremely randomized trees

Использование алгоритма Random Forest для решения задачи бинарной классификации на датастете Titanic: импорт и чтение данных, обобщенная информация и краткая статистка, изменение типов данных и заполнение пропущенных значений. Выбор признаков, создание и обучение модели. Оценка качества модели. Настройка гиперпараметров.

2.3. Бустинг. AdaBoost и градиентный бустинг над решающими деревьями Предсказание цены недвижимости на датасете Boston, используя алгоритм градиентного бустинга библиотеки scikit-learn, сравнение результатов с алгоритмом Random Forest. Оценка качества моделей. Настройка гиперпараметров.

2.4. Фреймворки машинного обучения

Сравнение результатов машинного обучения (метрики качества моделей) для задач классификации и регрессии в популярных фреймворках: LightGBM, XGBoost, CatBoost. Особенности работы.

2.5. Кластерный анализ, алгоритм k-meansи и поиск ассоциативных правил

Решение задачи бинарной классификации на датасете Titanic, используя алгоритм кластерного анализа k-means. Выбор признаков, создание и обучение модели. Оценка качества модели, сравнение с результатами алгоритма Random Forest и Gradient Boosting for classification библиотеке scikit-learn.

2.6. Введение в нейронные сети

Решение задач классификации и регрессии с помощью многослойного персептрона. Сравнение с результатами алгоритма Random Forest и Gradient Boosting for classification библиотеке scikit-learn.

2.7. Глубокие нейронные сети

Решение задач классификации изображений, обнаружения объектов, анализ настроений с помощью глубоких нейронных сетей реализуемых во фреймворках AutoGluon.

2.8. Анализ временных рядов

Применение технологий автоматического машинного обучения во фреймворках Auto Sklearn, Tree-Based Pipeline Optimization Tool (TPOT), Auto Keras, h2o.ai. для решения задач классификации и регрессии на структурированных данных.

2.9. Автоматическое машинное обучение (AutoML)

Применение технологий автоматического машинного обучения во фреймворках Auto Sklearn, Tree-Based Pipeline Optimization Tool (TPOT), Auto Keras, h2o.ai. для решения задач классификации и регрессии на структурированных данных.

Модуль 3. Современные озера и хранилища данных, аналитика больших данных и методы

3.1. Облачные технологии обработки больших данных

Типы данных, функции и операторы SQL в Google BigQuery:

Числовые типы и функции. Математические функции. Стандартное вещественное деление. Сравнение. Условные выражения. Строковые функции. Парсинг и форматирование отметок времени. Арифметические операции с временными метками. Функции для работы с географическими Координатами.

3.2. RT.DataLake

Знакомство с платформой RT.DataLake, с архитектурой, основными частями и пользовательским интерфейсом, доступ к наборам данных RT.DataLake для их изучения. Визуализация данных RT.DataLake в записной книжке Jupyter, преимущества использования RT.DataLake для решения задач по науке о данных.

3.3. Маркетинговая аналитика в RT.Warehouse

Машинное обучение в RT.Warehouse. Ограничения RT.Warehouse на виды моделей: Линейная регрессия (LINEAR_REG). Логистическая регрессия (LOGISTIC_REG). KMEANS. TENSORFLOW, решение задач Data Minig.

3.4. Решение задач Data Mining в корпоративных хранилищах данных.

Загрузка данных, импорт данных из файла. Доступ к импортированным данным. Подготовка и очистка данных, исследовательский анализ данных (EDA), корреляционный анализ, создание визуализаций. Расширенный исследовательский анализ данных: создание простой базовой модели, OneHotEncoder и масштабирование функций, уменьшение размерности и изучение важности признаков. Создание озера данных и управление им: импорт необработанных данных и сохранение их в таблице Delta Lake (Bronze), сохранение подготовленных и очищенных данных в Silver table in Delta Lake, создание для предоставления чистых и надежных данных для конкретного бизнес-подразделения или варианта использования.

Модуль 4. Платформы науки о данных и машинного обучения и бизнес аналитики

4.1. Платформа H2O.ai

Использование веб интерфейса H2O.ai Flow для создания, обучения и развертывания моделей машинного обучения без написания кода.

4.2. Платформа RapidMiner

Используя технологии RapidMiner Turbo Prep и AutoML создать, обучить, оценить и развернуть модели классификации и регрессии в RapidMiner.

4.3. Аналитические технологии Loginom

Представление данных в виде многомерных кубов (OLAP-кубов). Интерфейс OLAP-куба в Loginom: область свободных полей; область измерений в строках; область фактов; область измерений в колонках; область фильтрации по измерениям; панель инструментов куба. Манипулирование данными «на лету», отображение в виде кросс-таблиц и кросс-диаграмм, возможность

углубления в данные, Ad-hoc запросы, технологии drill-down, drill-up. Применение OLAP при решении многих аналитических задач: разведочный анализ, исследование данных, аналитическая отчетность, финансовый анализ, бюджетирование и прочее.

4.4. Платформа Knime.

Интеграция визуального анализа и технологий машинного обучения: для выявления скрытых закономерностей в данных: создать, обучить и развернуть модель машинного обучения в платформе KNIME, используя интеграцию с технологиями H2O.ai и фреймворком XGBoost.

4.5. Исследование и визуализация данных в RT.DataVision

По индивидуальному заданию (кейс) создание физического и виртуального датасетов и настройка их параметров, загрузка данных, построение дашбордов в программном продукте RT.DataVision, выявление инсайтов, оформление историй (Story). Встраивание чартов и дашбордов разработанных RT DataVision во внешние сервисы.

4.6. Создание интерактивной отчетности в Tableau Desktop

По индивидуальному заданию (кейс):

1. создание KPI дашбордов
2. создание Top to Bottom дашбордов
3. создание. Bottom to Top дашбордов
4. создание. Q&A дашборд (Вопрос-Ответ)
5. создание. Single Viz дашборд (Один график)
6. создание визуализаций сравнений
7. создание дашбордов сравнений во времени

4.7. Аналитические технологии Power BI Desktop.

По индивидуальным данным.

1. создание простой модели данных, объединяющую четыре таблицы
2. использование вычисляемых столбцов для новых способов группировки данных
3. создание круговой диаграммы для сравнения
4. создание дашбордов с KPI и диаграммами для анализа
5. создание древовидной карты в качестве среза и столбчатой диаграммы заданных данных.

4.8. Визуализация данных – Yandex DataLens

По индивидуальному заданию (кейс) разработка полноценного дашборда для мониторинга ключевых бизнес-метрик, выявление инсайтов, геоаналитика на Яндекс. Картах. Добавление учётных записей команды для совместной работы.

Примерные тестовые вопросы для аттестации.

1. Правильная последовательность в Business Intelligence:

a) данные-информация-знания-принятие решения
b) информация-данные-знания-принятие решения
c) принятие решения-информация-данные-знания

2. В платформе для бизнес-анализа должны быть реализованы:
a) 10 ключевых возможностей
b) 15 ключевых возможностей
c) 20 ключевых возможностей

3. Перечислите правильную последовательность этапов Knowledge Discovery in Databases –процесса обнаружения знаний в базах данных
a) трансформация, интерпретация результатов, выборка, очистка, построение моделей.
b) построение моделей, выборка, очистка, трансформация, интерпретация результатов.
c) выборка, очистка, трансформация, построение моделей, интерпретация результатов.

4. OLAP-системы это:
a) информационные системы многомерного анализа данных в реальном времени.
b) информационные системы автоматической обработки данных.
c) информационные системы алгоритмической обработки данных.

5. OLTP-системы это:
a) информационные системы оперативной транзакционной обработки данных
b) информационные системы оперативного анализа данных
c) информационные системы автоматической обработки данных

6. Если для реализации многомерной модели используют реляционные базы данных, то способ реализации гиперкуба называется
a) MOLAP
b) ROLAP
c) HOLAP

7. Если для реализации многомерной модели используют и многомерные, и реляционные базы данных, то способ реализации гиперкуба называется
a) MOLAP
b) ROLAP
c) HOLAP
8. Информационные хранилища созданы для удобства …
a) руководителей всех уровней для принятия решений
b) предметных приложений
c) редактирования данных

9. Большинство методов Data Mining были разработаны в рамках…
a) Теории искусственного интеллекта
b) Классического анализа данных

c) Теории баз данных

10. Классификация – …

a) разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов

b) высокоуровневые средства отражения информационной модели и описания структуры данных

c) это установление зависимости дискретной выходной переменной от входных переменных

11. Выражения анализа данных DAX применяется в платформе бизнес-аналитики

a) Qlik Sense
b) Tableau
c) Power BI

12. Какой языки программирования не поддерживает Power BI Desktop:

a) R
b) Julia
c) Python

13. Каждые два года объем данных увеличивается приблизительно:

a) на 10 зеттабайтов информации
b) в 2 раза
c) в 6 раз

14. Алгоритмы машинного обучения на больших данных реализуются с помощью:

a) Spark
b) Pig
c) Hive

15. Самым крупным облачным провайдером (по модели «инфраструктура как услуга») является:

a) Google
b) Microsoft
c) Amazon

16. Централизованным хранилищем, позволяющим хранить все структурированные и неструктурированные данные в любом масштабе, является:

a) Хранилище
b) База данных
c) Озеро данных

17. Простой визуальной средой разработки моделей машинного обучения, в которой можно перетаскивать элементы прямо в браузере и не нужно писать код, является:

a) Databricks
b) студия машинного обучения Azure
c) TensorFlow

18. Классификация относится к стратегии:
a) Обучения с учителем
b) Обучения без учителя
c) Оба ответа неверны

19. Обработка данных в современных платформах бизнес-аналитики происходит
a) в оперативной памяти
b) на жестком диске
c) в сети пользователей

20. Лист дерева решений является:
a) конечным узлом
b) узлом проверки
c) узлом решения

Задания к итоговой аттестации

Задание:

1. Выбрать индивидуальное задание: можно использовать “свои” данные или использовать датасеты с Kaggle https://www.kaggle.com/datasets?search=customer&fileType=csv, импортировать кейс в RT DWH
2. Подключившись к DWH выполнить ABC-XYZ, RFM-анализ в DBeaver, Jupyter Notebook, KNIME Analytics Platform, Loginom (в любом инструменте на выбор)
3. Продвинутый уровень (не обязательно): построить модель оттока клиентов в любом инструменте на выбор
4. Обогащенный аналитикой кейс импортировать RT.Warehouse
5. Подключившись RT DWH из RT.DataVision создать несколько дашбордов по выборном кейсу, обогащенному аналитикой (учебный курс по RT.DataVision доступен на youtube https://www.youtube.com/watch?v=cs_81LKRDNM&list=PLAy-0-KaZdI1KiVL64Bd7n5u2Z48_5rQg)

Методические рекомендации:

Поскольку RT.Warehouse поддерживает сложные запросы, обрабатывающие большие объемы данных, в том числе сложные аналитические функции, она эффективно может использоваться для построения корпоративного хранилища данных, BI-аналитики, AD-HOC запросов и Data Science. Типовые кейсы использования:

- Построение рекомендательных моделей продаж на базе искусственного интеллекта в сегментах B2B, B2C
- Построение моделей оттока на базе искусственного интеллекта в сегментах B2B, B2C

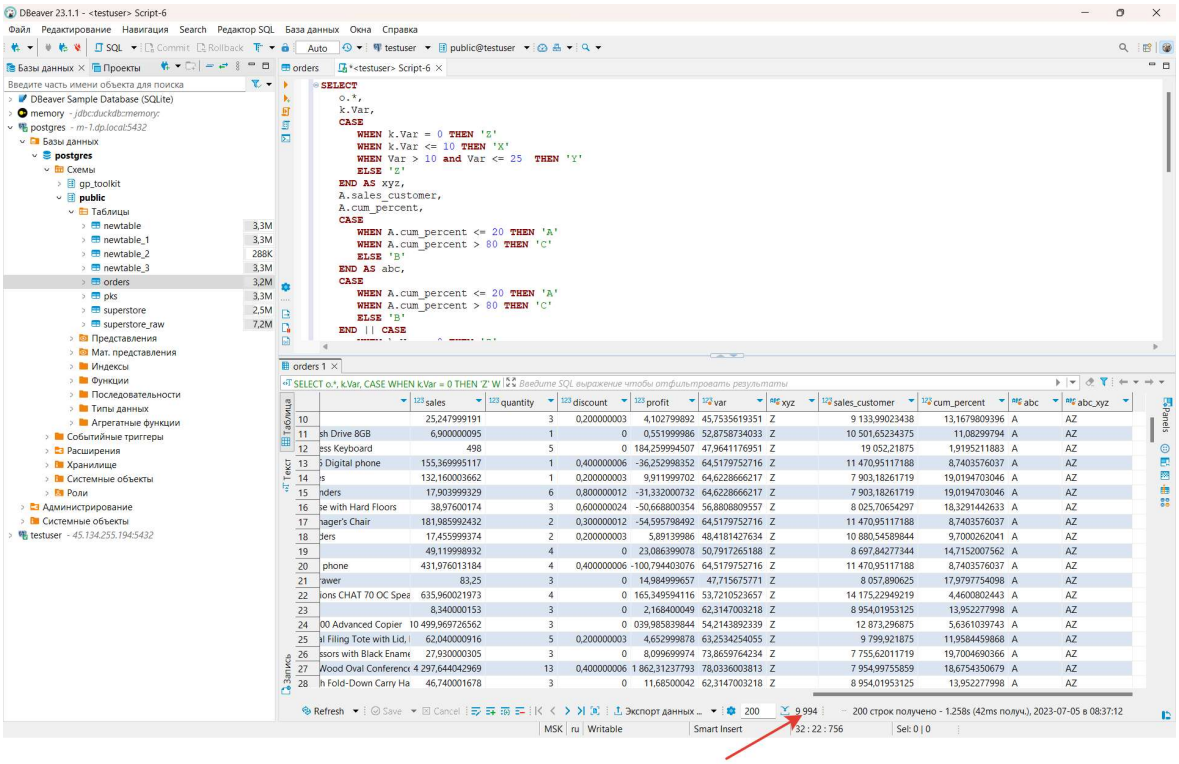
ример выполнения ABC-XYZ, RFM-анализа в DBeaver

```
SELECT
   o.*,
   k.Var,
   CASE
      WHEN k.Var = 0 THEN 'Z'
      WHEN k.Var <= 10 THEN 'X'
      WHEN Var > 10 and Var <= 25  THEN 'Y'
      ELSE 'Z'
   END AS xyz,
   A.sales_customer,
   A.cum_percent,
   CASE
      WHEN A.cum_percent <= 20 THEN 'A'
      WHEN A.cum_percent > 80 THEN 'C'
      ELSE 'B'
   END AS abc,
   CASE
      WHEN A.cum_percent <= 20 THEN 'A'
      WHEN A.cum_percent > 80 THEN 'C'
      ELSE 'B'
   END || CASE
      WHEN k.Var = 0 THEN 'Z'
      WHEN k.Var <= 10 THEN 'X'
      WHEN Var > 10 and Var <= 25  THEN 'Y'
      ELSE 'Z'
   END as abc_xyz
FROM public.orders o
INNER JOIN (
   SELECT
      customer_id,
      SQRT(VARIANCE(quantity))*100/AVG(quantity) as Var
   FROM public.orders
   GROUP BY customer_id
) as k ON o.customer_id = k.customer_id
INNER JOIN (
   SELECT
      customer_id,
      m.sales_customer,
      m.percent,
      SUM(m.percent) OVER (
         ORDER BY m.percent DESC
         ROWS UNBOUNDED PRECEDING
      ) AS cum_percent
   FROM (
      SELECT
         customer_id,
         k.sales_customer,
         k.sales_customer *100/ SUM(k.sales_customer) OVER () AS percent
      FROM (
         SELECT
            customer_id,
            SUM(sales) sales_customer
         FROM public.orders
         GROUP BY
            customer_id
      ) AS k
      ORDER BY
         sales_customer DESC
   ) m
   ORDER BY m.sales_customer DESC
) AS A ON o.customer_id = A.customer_id
ORDER BY abc_xyz
```

```
SELECT
  o.*,
  k.Var,
  CASE
    WHEN k.Var = 0 THEN 'Z'
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END AS xyz,
  A.sales_customer,
  A.cum_percent,
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END AS abc,
```

```
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END || CASE
    WHEN k.Var = 0 THEN 'Z'
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END as abc_xyz
FROM Demo.Orders o
INNER JOIN (
  SELECT
    customer_id,
    SQRT(VARIANCE(quantity))*100/AVG(quantity) as Var
  FROM Demo.Orders
  GROUP BY customer_id
) as k ON o.customer_id = k.customer_id
INNER JOIN (
  SELECT
    customer_id,
    m.sales_customer,
    m.percent,
    SUM(m.percent) OVER (
      ORDER BY m.percent DESC
      ROWS UNBOUNDED PRECEDING
    ) AS cum_percent
  FROM (
    SELECT
      customer_id,
      k.sales_customer,
      k.sales_customer *100/ SUM(k.sales_customer) OVER () AS percent
    FROM (
      SELECT
        customer_id,
        SUM(sales) sales_customer
      FROM Demo.Orders
      GROUP BY
        customer_id
    ) AS k
    ORDER BY
      sales_customer DESC
  ) m
  ORDER BY m.sales_customer DESC
) AS A ON o.customer_id = A.customer_id
ORDER BY abc_xyz
```

Результат выполнения запроса:



Разберем запрос более подробно:

```
SELECT
  k.customer_id,
  k.Var,
  CASE
    WHEN k.Var = 0 THEN 'Z'
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END AS xyz,
  A.sales_customer,
  A.cum_percent,
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END AS abc,
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END || CASE
    WHEN k.Var = 0 THEN 'Z'
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END as abc_xyz
```

Здесь мы выбираем несколько полей, результаты вычисления которых были получены как результат выполнения подзапросов:
- `k.customer_id`: идентификатор клиента, полученный из подзапроса, который вычисляет дисперсию и среднее значение количества товаров, купленных клиентом (`VARIANCE` и `AVG`).
- `k.Var`: стандартное отклонение количества товаров, купленных клиентом (`SQRT` от дисперсии), выраженное в процентах от среднего значения количества товаров, которые купил клиент.
- `CASE ... END AS xyz`: поле, которое вычисляет категорию `xyz` для каждого клиента.
- `A.sales_customer`: общее количество продаж (необязательное поле, оставлено для контекста).
- `A.cum_percent`: кумулятивный процент от общего объема продаж каждого клиента в порядке убывания, вычисляемый в подзапросе.
- `CASE ... END AS abc`: поле, которое вычисляет категорию `abc` для каждого клиента.
- `CASE ... END AS abc_xyz`: поле, которое объединяет значения `abc` и `xyz` в одно значение.

Следующая часть запроса:

```
FROM (
  SELECT
    customer_id,
    SQRT(VARIANCE(quantity))*100/AVG(quantity) as Var
  FROM public.orders
  GROUP BY customer_id
) as k
```

Мы выбираем данные из подзапроса, который вычисляет стандартное отклонение количества товаров, купленных каждым клиентом, и группирует результаты по идентификатору клиента.

Следующая часть - добавляем новый `INNER JOIN` с подзапросом, который вычисляет общую сумму продаж для каждого клиента, а также вычисляет долю этих продаж (в процентах), которые составляет каждый клиент относительно общей суммы продаж. Затем мы объединяем результаты этого подзапроса с результатами первого подзапроса (которые мы назвали `k`) используя поле `customer_id`:

```
INNER JOIN (
  SELECT
    customer_id,
    m.sales_customer,
    m.percent,
    SUM(m.percent) OVER (
      ORDER BY m.percent DESC
      ROWS UNBOUNDED PRECEDING
    ) AS cum_percent
  FROM (
    SELECT
```

```
      customer_id,
      k.sales_customer,
      k.sales_customer *100/ SUM(k.sales_customer) OVER () AS percent
    FROM (
      SELECT
        customer_id,
        SUM(sales) sales_customer
      FROM public.orders
      GROUP BY
        customer_id
    ) AS k
    ORDER BY
      sales_customer DESC
  ) m
  ORDER BY m.sales_customer DESC
) AS A ON k.customer_id = A.customer_id
```

Сортируем результаты по полю `abc_xyz` в порядке возрастания.

```
ORDER BY abc_xyz
```

Добавляем новый `INNER JOIN` с таблицей `public.orders`. Далее, чтобы связать результаты из этой таблицы с результатами наших первых двух `INNER JOIN``ов, мы используем поле `customer_id`.

```
FROM public.orders o
```

Следующая часть запроса - выбираем все поля из таблицы `public.orders` и дополнительные поля, которые были вычислены в наших двух `INNER JOIN``ax:

```
SELECT
  o.*,
  k.Var,
  CASE
    WHEN k.Var = 0 THEN 'Z'
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END AS xyz,
  A.sales_customer,
  A.cum_percent,
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END AS abc,
  CASE
    WHEN A.cum_percent <= 20 THEN 'A'
    WHEN A.cum_percent > 80 THEN 'C'
    ELSE 'B'
  END || CASE
    WHEN k.Var = 0 THEN 'Z'
```

```
    WHEN k.Var <= 10 THEN 'X'
    WHEN Var > 10 and Var <= 25  THEN 'Y'
    ELSE 'Z'
  END as abc_xyz
```

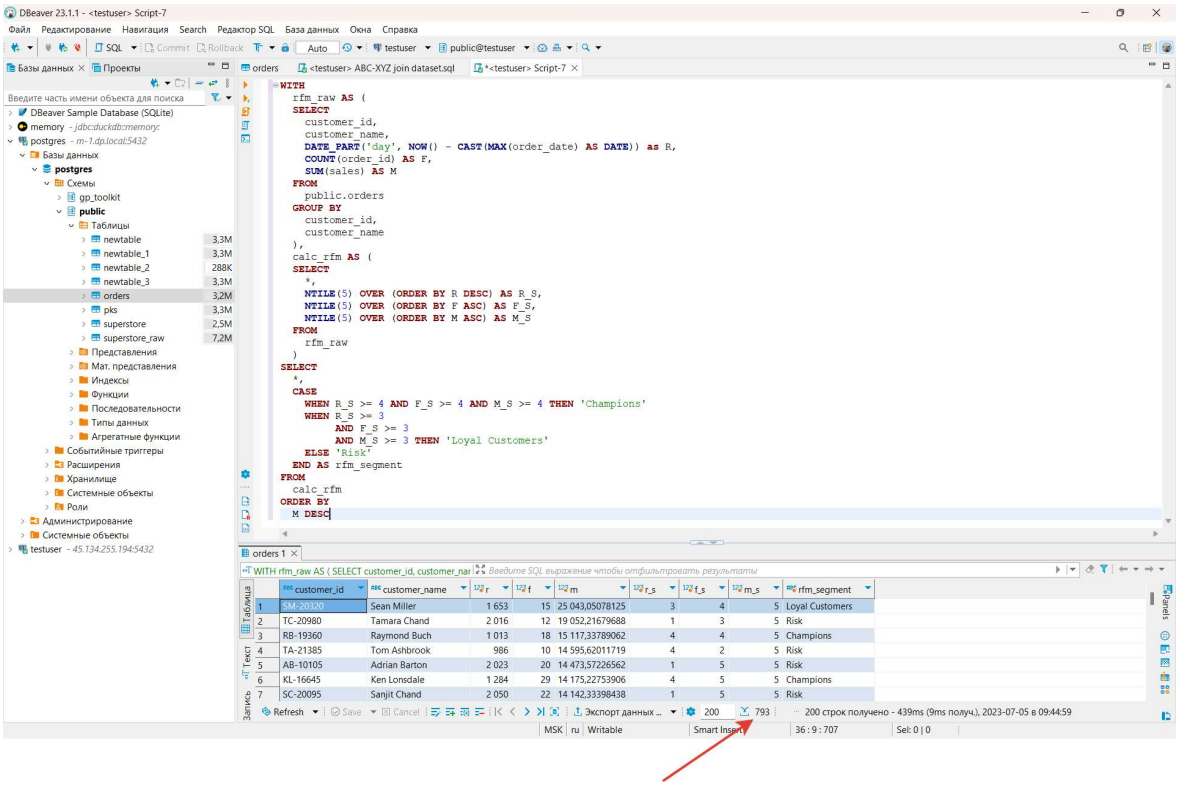
И сортируем результаты по полю `abc_xyz` в порядке убывания.

```
ORDER BY abc_xyz DESC
```

RFM-анализ:

```
WITH
  rfm_raw AS (
  SELECT
    customer_id,
    customer_name,
    DATE_PART('day', NOW() - CAST(MAX(order_date) AS DATE)) as R,
    COUNT(order_id) AS F,
    SUM(sales) AS M
  FROM
    public.orders
  GROUP BY
    customer_id,
    customer_name
  ),
  calc_rfm AS (
  SELECT
    *,
    NTILE(5) OVER (ORDER BY R DESC) AS R_S,
    NTILE(5) OVER (ORDER BY F ASC) AS F_S,
    NTILE(5) OVER (ORDER BY M ASC) AS M_S
  FROM
    rfm_raw
  )
SELECT
  *,
  CASE
    WHEN R_S >= 4 AND F_S >= 4 AND M_S >= 4 THEN 'Champions'
    WHEN R_S >= 3
         AND F_S >= 3
         AND M_S >= 3 THEN 'Loyal Customers'
    ELSE 'Risk'
  END AS rfm_segment
FROM
  calc_rfm
ORDER BY
  M DESC
```

Результат выполнения запроса:



Этот код представляет собой запрос на анализ данных (RFM-анализ) для определения категории клиентов на основе трех показателей: длительности последнего заказа (R - Recency), количества заказов (F - Frequency) и общей суммы заказов (M - Monetary). Запрос состоит из нескольких частей:

1. Сначала создается общий запрос с использованием оператора WITH и определяются два временных подзапроса (rfm_raw и calc_rfm), которые затем будут использоваться в основном запросе.

2. Подзапрос rfm_raw вычисляет три переменные для каждого клиента: R (количество дней, прошедших с даты последнего заказа), F (количество заказов) и M (общая сумма заказов). Эти переменные вычисляются с помощью функций DATE_PART для вычисления длительности заказа, COUNT для подсчета числа заказов и SUM для подсчета суммы заказов. Затем эти данные сгруппированы по идентификатору клиента (customer_id) и имени клиента (customer_name).

3. Подзапрос calc_rfm использует оператор NTILE для создания категорий R_S, F_S и M_S на основе переменных R, F и M в каждом заказе. Этот подзапрос использует результаты из rfm_raw в качестве входных данных и вычисляет категории для каждого значения R, F и M.

4. Основной запрос SELECT объединяет результаты из calc_rfm и добавляет новую переменную rfm_segment. В этой переменной используется оператор CASE для определения категории клиента на основе трех переменных R_s, F_s и M_s. В зависимости от значений этих переменных клиенты классифицируются как "Champions" (если все три значения превышают 4), "Loyal Customers" (если все три значения больше или равны 3) или "Risk" (если это не так). Наконец, запрос сортируется в порядке убывания по переменной M, чтобы вывести клиентов с наибольшей суммой заказов.

В результате этот код вычисляет основные показатели для каждого клиента и классифицирует их на основе этих показателей в одну из трех категорий, чтобы более эффективно управлять отношениями с клиентами.

Подключение в Jupyter Notebook

```
In [1]: import psycopg2
        import pandas as pd
        print (pd.__version__)

        1.3.5

In [2]: conn = psycopg2.connect(
           host="m-1.dp.local", # адрес сервера базы данных
           database="postgres", # имя базы данных
           user="admin", # имя пользователя
           password="admin" # пароль
        )

In [3]: sql = "SELECT * FROM public.superstore_raw"
        df = pd.read_sql(sql, conn)

In [5]: import pandas as pd; import numpy as np
        # Step: Change data type of order_date to Datetime
        df['order_date'] = pd.to_datetime(df['order_date'], infer_datetime_format=True)

        import pandas as pd; import numpy as np
        # Step: Drop duplicates based on ['row_id', 'order_id', 'order_date', 'ship_date', 'ship_mode',
        #'customer_id', 'customer_name', 'segment', 'country', 'city', 'state',
        #'postal_code', 'region', 'product_id', 'category', 'sub_category', 'product_name', 'sales', 'quantity', 'discount', 'profit', '

        df = df.drop_duplicates(keep='first')

        df.head()
```

Out[5]:

	row_id	order_id	order_date	ship_date	ship_mode	customer_id	customer_name	segment	country	city	...	product_id	category	sub_category	p
0	8310	CA-2014-168312	2018-01-03	2018-07-03	Standard Class	GW-14605	Giulietta Weimer	Consumer	United States	Houston	...	OFF-ST-10003692	Office Supplies	Storage	
4	5076	CA-2014-134572	2018-04-20	2018-04-22	Second Class	SV-20365	Seth Vernon	Consumer	United States	Houston	...	FUR-TA-10001705	Furniture	Tables	
8	5077	CA-2014-134572	2018-04-20	2018-04-22	Second Class	SV-20365	Seth Vernon	Consumer	United States	Houston	...	OFF-ST-10004634	Office Supplies	Storage	
9	5077	CA-2014-134572	2018-04-20	2018-04-22	Second Class	SV-20365	Seth Vernon	Consumer	United States	Houston	...	OFF-ST-10004634	Office Supplies	Storage	
12	1759	CA-2014-139017	2018-11-05	2018-05-17	Standard Class	RM-19375	Raymond Messe	Consumer	United States	Houston	...	TEC-AC-10001013	Technology	Accessories	

```
import psycopg2
import pandas as pd
conn = psycopg2.connect(
   host="m-1.dp.local", # адрес сервера базы данных
   database="postgres", # имя базы данных
   user="admin", # имя пользователя
   password="admin" # пароль
)
sql = "SELECT * FROM public.superstore_raw"
df = pd.read_sql(sql, conn)
```

Можно выполнять ABC-XYZ, RFM-анализ на SQL и Python в Jupyter Notebook.
ABC-XYZ анализ:

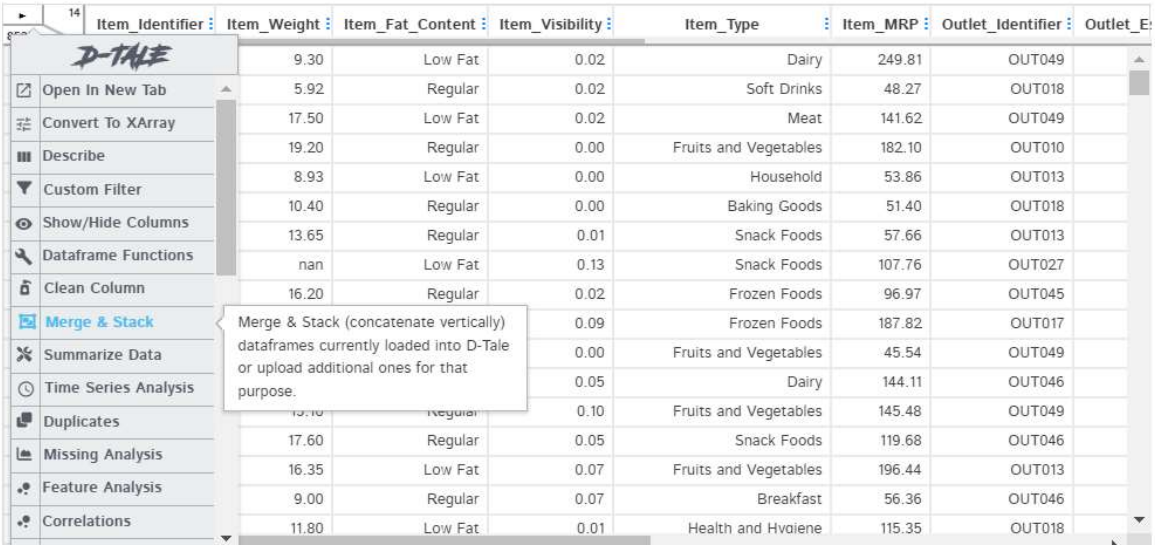
```
sql = "\
SELECT \
   o.*, \
   k.Var, \
   CASE \
      WHEN k.Var = 0 THEN 'Z' \
      WHEN k.Var <= 10 THEN 'X' \
      WHEN Var > 10 and Var <= 25  THEN 'Y' \
      ELSE 'Z' \
   END AS xyz, \
   A.sales_customer, \
   A.cum_percent, \
   CASE \
      WHEN A.cum_percent <= 20 THEN 'A' \
      WHEN A.cum_percent > 80 THEN 'C' \
      ELSE 'B' END AS abc,  \
   CASE \
      WHEN A.cum_percent <= 20 THEN 'A' \
      WHEN A.cum_percent > 80 THEN 'C' \
      ELSE 'B' \
   END || CASE \
      WHEN k.Var = 0 THEN 'Z' \
      WHEN k.Var <= 10 THEN 'X' \
      WHEN Var > 10 and Var <= 25  THEN 'Y' \
      ELSE 'Z' \
   END as abc_xyz FROM public.orders o \
INNER JOIN ( \
   SELECT \
      customer_id, \
      SQRT(VARIANCE(quantity))*100/AVG(quantity) as Var \
   FROM public.orders \
   GROUP BY customer_id \
) as k ON o.customer_id = k.customer_id \
INNER JOIN ( \
   SELECT \
      customer_id, \
      m.sales_customer, \
      m.percent, \
      SUM(m.percent) OVER ( \
         ORDER BY m.percent DESC \
         ROWS UNBOUNDED PRECEDING \
      ) AS cum_percent \
   FROM ( \
      SELECT \
         customer_id, \
         k.sales_customer, \
         k.sales_customer *100/ SUM(k.sales_customer) OVER () AS percent \
      FROM ( \
         SELECT \
            customer_id, \
            SUM(sales) sales_customer \
         FROM public.orders \
         GROUP BY \
            customer_id \
      ) AS k \
      ORDER BY \
         sales_customer DESC \
   ) m \
   ORDER BY m.sales_customer DESC \
) AS A ON o.customer_id = A.customer_id \
ORDER BY abc_xyz"

df = pd.read_sql(sql, conn)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 27 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   row_id          9994 non-null   int64
 1   order_id        9994 non-null   object
 2   order_date      9994 non-null   object
 3   ship_date       9994 non-null   object
 4   ship_mode       9994 non-null   object
 5   customer_id     9994 non-null   object
 6   customer_name   9994 non-null   object
 7   Segment         9994 non-null   object
 8   country         9994 non-null   object
 9   city            9994 non-null   object
 10  state           9994 non-null   object
 11  postal_code     9994 non-null   int64
 12  region          9994 non-null   object
 13  product_id      9994 non-null   object
 14  category        9994 non-null   object
 15  sub_category    9994 non-null   object
 16  product_name    9994 non-null   object
 17  sales           9994 non-null   float64
 18  quantity        9994 non-null   int64
 19  discount        9994 non-null   float64
 20  profit          9994 non-null   float64
 21  var             9989 non-null   float64
 22  xyz             9994 non-null   object
 23  sales_customer  9994 non-null   float64
 24  cum_percent     9994 non-null   float64
 25  abc             9994 non-null   object
 26  abc_xyz         9994 non-null   object
dtypes: float64(6), int64(3), object(18)
memory usage: 2.1+ MB
```

Разведочный анализ данных легко выполнять с помощью библиотеки dtale ([знакомство с интерфейсом и простейшие действия](#)):

```
#!pip install -U dtale
import dtale
dtale.show(df)
```



!pip install -U dtale
import dtale
dtale.show(df)

RFM-анализ:

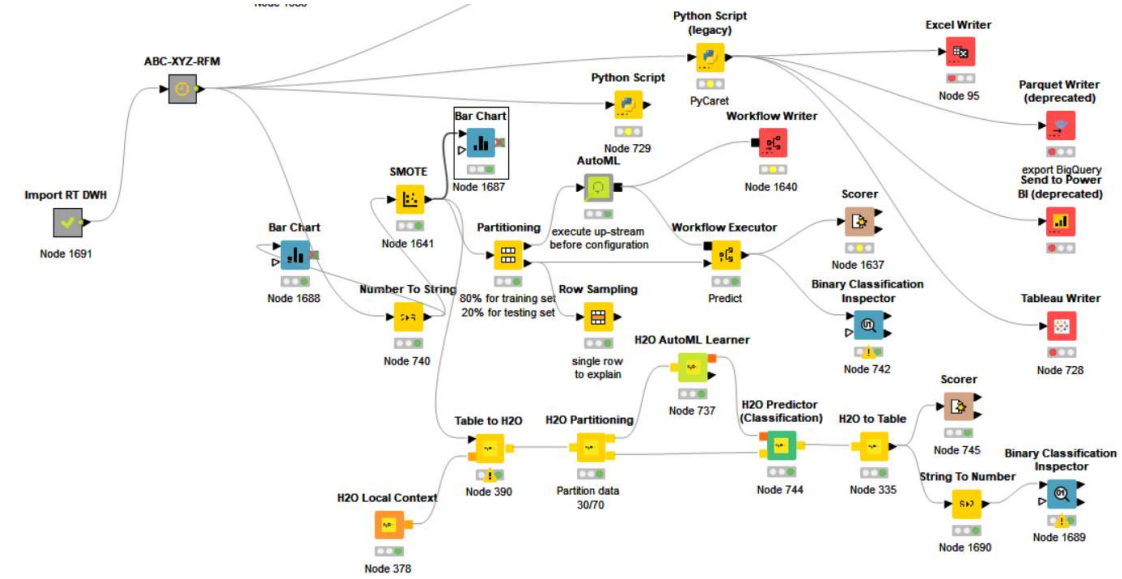
```
sql = "\
WITH \
  rfm_raw AS ( \
  SELECT \
    customer_id, \
    customer_name, \
    DATE_PART('day', NOW() - CAST(MAX(order_date) AS DATE)) as R, \
    COUNT(order_id) AS F, \
    SUM(sales) AS M \
  FROM \
    public.orders \
  GROUP BY \
    customer_id, \
    customer_name \
  ), \
  calc_rfm AS ( \
  SELECT \
    *, \
    NTILE(5) OVER (ORDER BY R DESC) AS R_S, \
    NTILE(5) OVER (ORDER BY F ASC) AS F_S, \
    NTILE(5) OVER (ORDER BY M ASC) AS M_S \
  FROM \
    rfm_raw \
  ) \
SELECT \
  *, \
  CASE \
    WHEN R_S >= 4 AND F_S >= 4 AND M_S >= 4 THEN 'Champions' \
    WHEN R_S >= 3 \
         AND F_S >= 3 \
         AND M_S >= 3 THEN 'Loyal Customers' \
    ELSE 'Risk' \
  END AS rfm_segment \
FROM \
  calc_rfm \
ORDER BY \
  M DESC"

rfm = pd.read_sql(sql, conn)
```

```
import pandas as pd; import numpy as np
# Step: Change data type of order_date to Datetime
df['order_date'] = pd.to_datetime(df['order_date'], infer_datetime_format=True)
# Step: Change data type of ship_date to Datetime
df['ship_date'] = pd.to_datetime(df['ship_date'], infer_datetime_format=True)
# Step: Left Join with rfm where customer_id=customer_id
df = pd.merge(df, rfm, how='left', on=['customer_id'])
# Step: Change data type of r_s to Categorical/Factor
df['r_s'] = df['r_s'].astype('category')
# Step: Change data type of f_s to Categorical/Factor
df['f_s'] = df['f_s'].astype('category')
# Step: Change data type of m_s to Categorical/Factor
df['m_s'] = df['m_s'].astype('category')

Создаем признак оттока клиента
df['churn'] = df['r'].apply(lambda x: 'Yes' if x > 1200 else 'No')
# Step: Change data type of churn to Categorical/Factor
df['churn'] = df['churn'].astype('category')

Формируем датафрейм для создания модели машинного обучения
# Step: Drop columns
df_churn = df.drop(columns=['row_id', 'order_id', 'order_date', 'ship_date', 'ship_mode',
'customer_id', 'customer_name', 'country', 'postal_code', 'product_id', 'var', 'xyz', 'sales',
'cum_percent', 'abc', 'quantity', 'r', 'f', 'm', 'r_s', 'rfm_segment', 'sales_customer'])
df_churn
Для создания модели будем использовать фреймворк Catboost
!pip install catboost
!pip install scikit-learn
!pip install ipywidgets
!jupyter nbextension enable --py widgetsnbextension

from catboost import CatBoostClassifier

#Creating a training set for modeling and validation set to check model performance
X = df_churn.drop(['churn'], axis=1)
y = df_churn.churn
from sklearn.model_selection import train_test_split
X_train, X_validation, y_train, y_validation = train_test_split(X, y, train_size=0.7,
random_state=1234)

X.info()

categorical_features_indices = np.where(X.dtypes != float)[0]

from catboost import CatBoostClassifier, Pool, metrics, cv
from sklearn.metrics import accuracy_score

model = CatBoostClassifier(
    custom_loss=[metrics.Accuracy()],
    random_seed=42,
    logging_level='Silent'
)
```

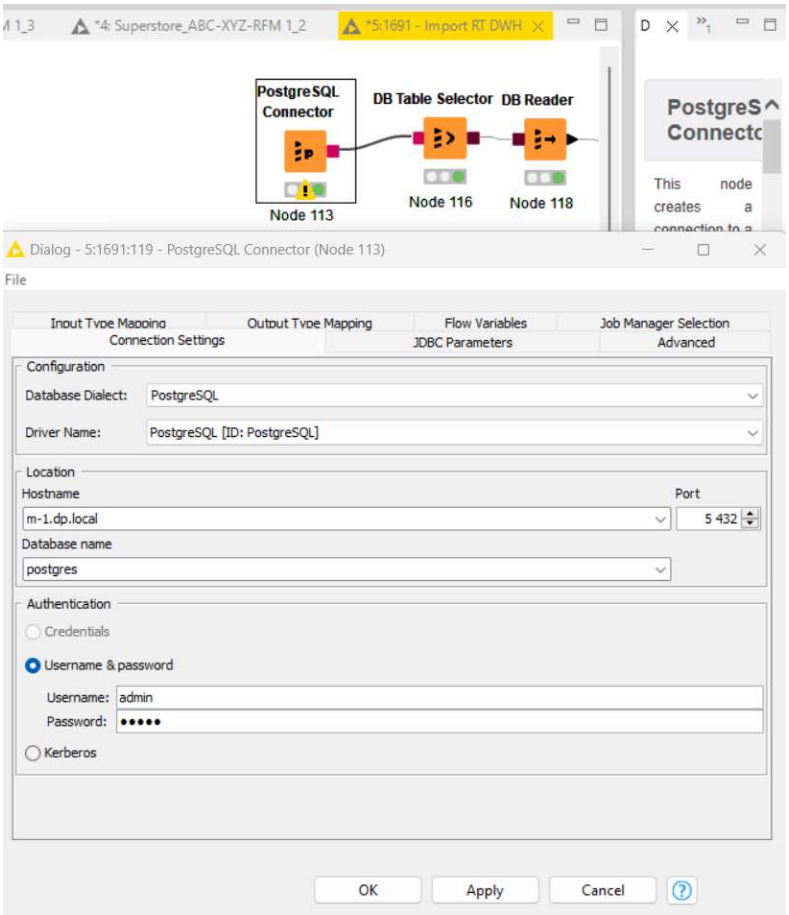
```
model.fit(
    X_train, y_train,
    cat_features=categorical_features_indices,
    eval_set=(X_validation, y_validation),
#     logging_level='Verbose',  # you can uncomment this for text output
    plot=True
)

cv_params = model.get_params()
cv_params.update({
    'loss_function': metrics.Logloss()
})
cv_data = cv(
    Pool(X, y, cat_features=categorical_features_indices),
    cv_params,
    plot=True
)

print('Best validation accuracy score: {:.2f}±{:.2f} on step {}'.format(
    np.max(cv_data['test-Accuracy-mean']),
    cv_data['test-Accuracy-std'][np.argmax(cv_data['test-Accuracy-mean'])],
    np.argmax(cv_data['test-Accuracy-mean'])
))

print('Precise validation accuracy score: {}'.format(np.max(cv_data['test-Accuracy-mean'])))

train_pool = Pool(X_train, y_train, cat_features=categorical_features_indices)
validate_pool = Pool(X_validation, y_validation, cat_features=categorical_features_indices)

model = CatBoostClassifier(iterations=50, random_seed=42,
logging_level='Silent').fit(train_pool)
feature_importances = model.get_feature_importance(train_pool)
feature_names = X_train.columns
for score, name in sorted(zip(feature_importances, feature_names), reverse=True):
    print('{}: {}'.format(name, score))

predictions_test = model.predict(X_validation)
predictions_probs_test = model.predict_proba(X_validation)
print(predictions_test[:10])
print(predictions_probs_test[:10])

predictions = model.predict(X)
predictions_probs = model.predict_proba(X)
print(predictions[:10])
print(predictions_probs[:10])
```

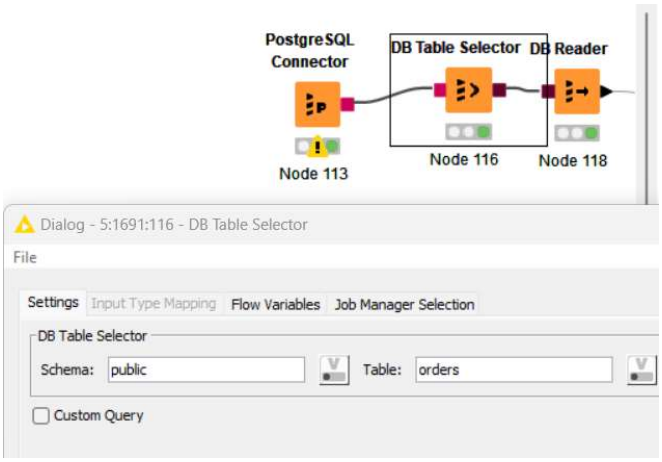
https://disk.yandex.ru/i/4fkYCnrYP74tQw

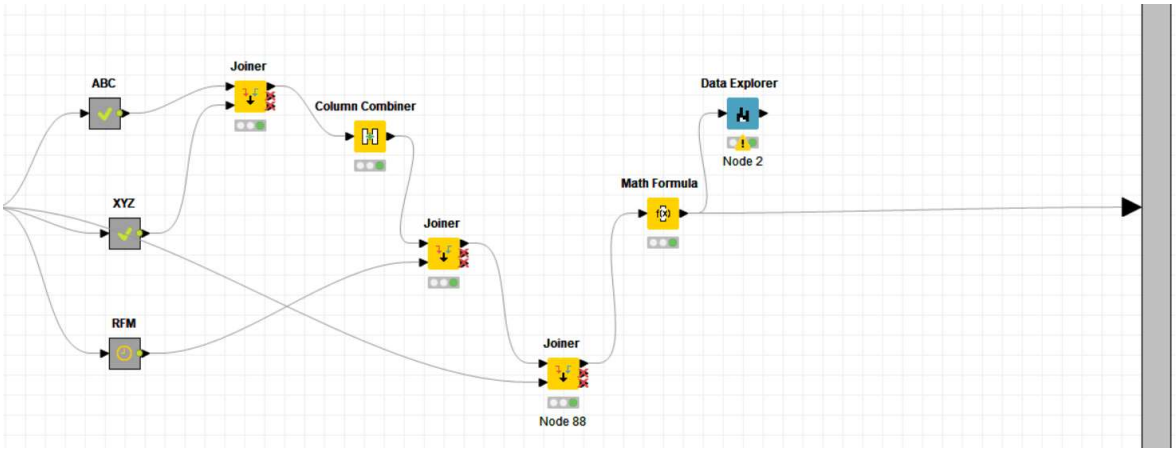
Решение задачи предсказания оттока клиента с использованием ABC-XYZ, RFM-анализа и библиотек python AutoML в KNIME Analytics Platform



Подключение к RT DWH:

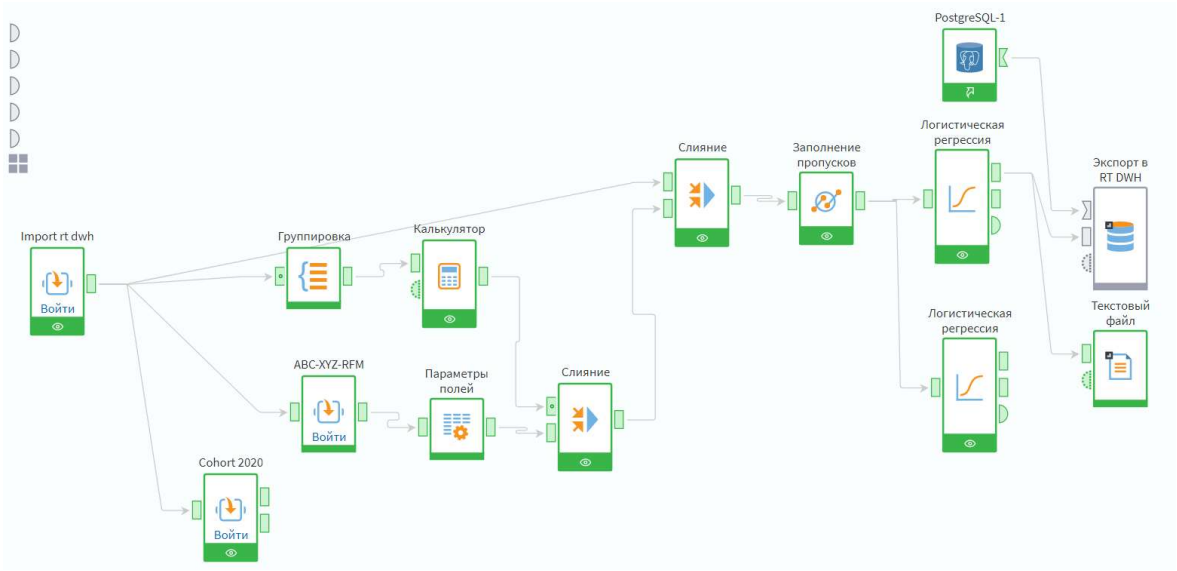


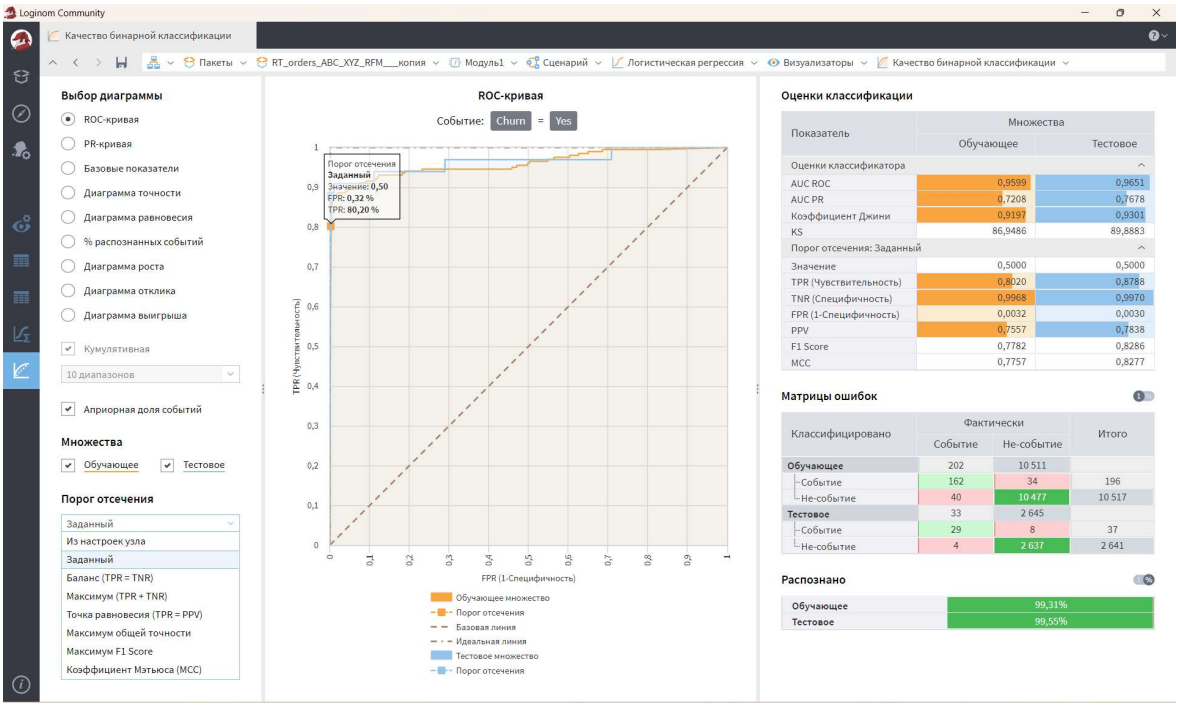




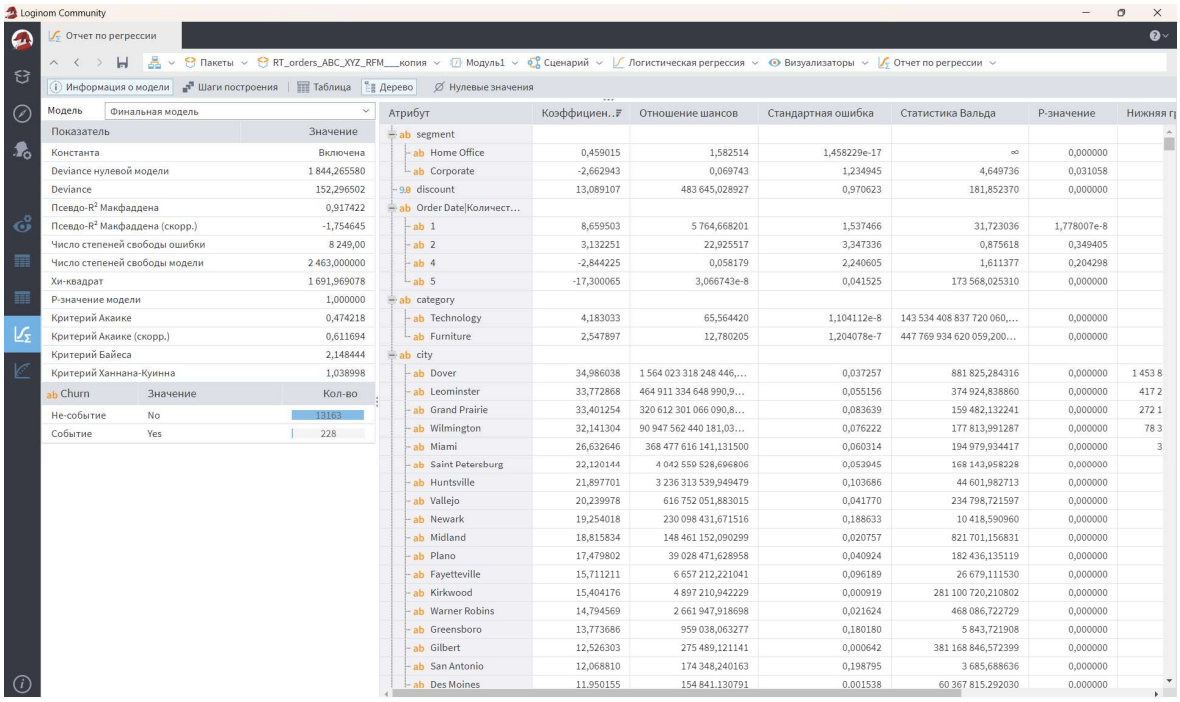
Оценка качества модели:

Решение задачи предсказания оттока клиента с использованием ABC-XYZ, RFM-анализа и библиотек python AutoML в Loginom





Качество бинарной классификации



Отчет по бинарной регрессии