Вопросы к Итоговому контролю по курсу Анализ данных

Уровень «Знание»

1. Что такое аналитика данных?
2. Чем Data Science отличается от Business Intelligence?
3. Чем аналитика данных отличается от классической статистики?
4. Какие основные роли входят в команду DS-проекта?
5. Что называют внутренними источниками данных?
6. Что называют внешними источниками данных?
7. В чём различие между структурированными, полуструктурированными и неструктурированными данными?
8. Что такое качество данных?
9. Что означает полнота данных?
10. Что означает согласованность данных?
11. Что такое утечка данных (data leakage)?
12. Что такое воспроизводимость аналитического эксперимента?
13. Что такое Data Governance?
14. Что такое метаданные?
15. Перечислите основные фазы методологии CRISP-DM.
16. Что обозначает аббревиатура SEMMA?
17. Что такое условная вероятность?
18. Что такое исследовательский анализ данных (EDA)?
19. Что такое линейная регрессия?
20. Что такое машина опорных векторов (SVM)?

Уровень «Понимание»

21. Почему аналитика данных не сводится только к построению модели?
22. Почему BI обычно отвечает на вопрос «что произошло», а DS — на вопрос «что делать дальше»?
23. Почему качество данных оценивается относительно конкретной задачи, а не вообще?
24. Почему пропущенные значения могут исказить результаты анализа?
25. Почему дубликаты опасны для аналитики и моделирования?
26. Почему воспроизводимость важна для доверия к аналитическому результату?
27. Почему Data Governance повышает надёжность аналитики в организации?

28. Почему для аналитики часто используют OLAP-модели, а не только OLTP-структуры?
29. Почему методология CRISP-DM считается итеративной, а не линейной?
30. Почему формула Байеса важна для задач оценки риска?
31. Почему медиана часто предпочтительнее среднего при скошенном распределении?
32. Почему EDA не является финальным доказательством гипотезы?
33. Почему выбор графика влияет на корректность интерпретации данных?
34. Почему перед обучением модели полезно задать baseline?
35. Почему мультиколлинеарность затрудняет интерпретацию линейной регрессии?
36. Почему логистическая регрессия возвращает вероятность, а не только метку класса?
37. Почему выбор порога классификации зависит от стоимости ошибок?
38. Почему тестовую выборку нельзя использовать для подбора гиперпараметров?
39. Почему одиночное дерево решений часто менее устойчиво, чем ансамбль?
40. Почему масштабирование признаков особенно важно для SVM?

Уровень «Применение»

41. Интернет-магазин хочет автоматически предлагать товары покупателям. Какая цель аналитики здесь формулируется точнее: отчётность BI или DS-решение? Обоснуйте выбор.
42. У вас есть три источника: таблица продаж в SQL, JSON-логи сайта и архив изображений товаров. К какому типу данных относится каждый источник?
43. Компания анализирует рынок труда. Какие из следующих данных вероятнее будут внешними источниками: CRM, вакансии конкурентов, бухгалтерская система, государственная статистика?
44. В столбце email 1200 строк, из них 96 пустые. Как оценить полноту поля в процентах?
45. В таблице клиентов обнаружены пропуски возраста, дубликаты записей и экстремально большие значения дохода. Какие три шага предобработки нужно выполнить в первую очередь?
46. Банк обучает модель дефолта и использует признак «факт просрочки через 30 дней после выдачи кредита». Почему это пример утечки данных?

47. Распределите задачи между ролями: построение ETL-конвейера, визуализация результатов для руководства, развёртывание модели в сервисе, обучение модели классификации.
48. Для дашборда продаж по категориям и месяцам нужно быстрое чтение и агрегирование. Что уместнее выбрать: нормализованную OLTP-схему или витрину/звезду?
49. Команда уже согласовала бизнес-цель, но ещё не изучала сами данные. На какой фазе CRISP-DM она находится?
50. Из 1000 транзакций 50 оказались мошенническими. Какова эмпирическая вероятность мошеннической транзакции?
51. Для анализа доходов сотрудников, где есть несколько очень высоких зарплат, какую меру центра разумнее взять как основную: среднее или медиану?
52. Какой график лучше выбрать для показа распределения одной числовой переменной: гистограмму, boxplot или матрицу ошибок?
53. Определите тип задачи: предсказать цену квартиры, определить «спам/не спам», оценить месячный спрос, распознать «ушёл клиент/не ушёл».
54. В линейной регрессии коэффициент при признаке стаж работы положителен. Как интерпретировать знак этого коэффициента?
55. Для модели обнаружения мошенничества пропуск мошенника намного дороже лишней ручной проверки честной операции. На какую сторону нужно смещать выбор порога — в пользу higher recall или higher precision?
56. Для прогноза продаж по неделям за 3 года какой способ валидации уместнее: случайный shuffle split или временное разбиение?
57. Для задачи, где нужна высокая интерпретируемость правил, что уместнее выбрать первым кандидатом: дерево решений или градиентный бустинг?
58. В датасете 95% объектов класса 0 и 5% класса 1. Какую проблему нужно учесть перед обучением классификатора?
59. В каком случае SVM может быть разумным выбором: очень маленький и плотный табличный датасет или огромный потоковый датасет на десятки миллионов строк?
60. Модель предсказывает вероятность дефолта 0.8, но по историческим данным такие объекты дефолтят лишь в 55% случаев. Как называется проблема?

Уровень «Анализ»

61. Сравните BI, классическую статистику и Data Science по цели, результату и степени влияния на действие.
62. Проанализируйте ситуацию: модель оттока имеет высокий recall, но низкий precision. Почему бизнес может остаться недоволен даже при «хорошем» техническом результате?
63. Сравните внутренние и внешние источники данных по контролируемости, стоимости и риску смещения.
64. В системе заказ отмечен как «оплачен» в CRM, но как «ожидает оплаты» в бухгалтерии. Какое измерение качества данных нарушено?
65. Проанализируйте риск для организации, если у неё нет каталога данных и единого глоссария терминов.
66. Сравните 3НФ и схему «звезда» с точки зрения аналитических запросов, JOIN-ов и удобства отчётности.
67. Сопоставьте CRISP-DM и SEMMA: какая методология сильнее привязана к бизнес-контексту, а какая — к потоку моделирования?
68. Проанализируйте утверждение: «Большой объём данных автоматически гарантирует правильный вывод». Почему это неверно?
69. На диаграмме рассеяния видно сильную положительную корреляцию между временем на сайте и суммой покупки. Почему этого недостаточно для вывода о причинности?
70. В резидуальном графике линейной регрессии разброс ошибок растёт вместе с прогнозом. Какое предположение модели нарушается?
71. Сравните MAE и RMSE: какая метрика сильнее штрафует большие ошибки и почему?
72. Модель классификации показала accuracy 96% на сильно несбалансированных классах. Почему этого недостаточно для вывода о качестве?
73. Проанализируйте разрыв: на train качество очень высокое, на validation заметно ниже. Какой тип проблемы вероятнее всего наблюдается?
74. Сравните holdout, k-fold cross-validation и TimeSeriesSplit по области применения.
75. На reliability diagram модель систематически предсказывает вероятность выше фактической частоты события. Как это интерпретировать?
76. Сравните дерево решений, Random Forest и GBDT по интерпретируемости и ожидаемой устойчивости.

77. Проанализируйте, почему выбор порога по умолчанию 0.5 может быть невыгоден для бизнеса.
78. Сравните линейную и нелинейную разделяющую границу для задачи классификации. В какой ситуации линейная модель будет явно недостаточной?
79. Проанализируйте, как отсутствие масштабирования признаков влияет на модели, чувствительные к расстоянию или геометрии зазора.
80. В проекте после хорошего EDA команда сразу вывела модель в прод без контрольного теста и мониторинга. Какие два-три методологические сбоя здесь видны?

Уровень «Синтез и выводы»

81. Сформулируйте бизнес-вопрос и одну измеримую KPI-метрику для проекта по прогнозу оттока клиентов мобильного оператора.
82. Сконструируйте краткий «паспорт источника данных» для таблицы транзакций интернет-магазина: какие поля в нём должны быть обязательно?
83. Предложите минимальный чек-лист первичной валидации данных перед началом анализа нового датасета.
84. Спроектируйте короткий pipeline предобработки для набора данных абитуриентов, где есть пропуски, категориальные признаки и выбросы в доходе семьи.
85. Предложите набор ролей и правил доступа для небольшой организации, которая только начинает внедрять Data Governance.
86. Выберите между CRISP-DM и SEMMA для проекта прогнозирования дефолтов в банке и обоснуйте выбор.
87. Составьте план минимального EDA для датасета успеваемости студентов: какие шаги нужно выполнить в правильной последовательности?
88. Предложите два графика для показа распределения и один график для показа взаимосвязи двух признаков в отчёте для руководства.
89. Спроектируйте базовый план оценки регрессионной модели спроса: baseline, разбиение данных и метрики.
90. Предложите правило выбора порога для модели фрод-мониторинга, если цена ложно-негативной ошибки в 10 раз выше ложно-позитивной.
91. Сконструируйте корректную схему подбора гиперпараметров и финальной оценки модели без утечки теста.
92. Оцените утверждение: «Если ROC-AUC высокий, модель уже готова к внедрению». Какие дополнительные проверки обязательны?

93. Спроектируйте эксперимент сравнения дерева решений, Random Forest и GBDT на одном датасете. Какие условия должны быть одинаковыми для честного сравнения?
94. Предложите стратегию работы с дисбалансом классов для задачи обнаружения редкого события.
95. Сформулируйте критерии, по которым вы будете решать, подходит ли SVM для конкретной задачи классификации.
96. Оцените утверждение: «Accuracy — главная метрика для любой классификации». В каких случаях это решение методологически слабое?
97. Разработайте краткую рубрику выбора стратегии разбиения данных: когда использовать holdout, k-fold, group split и time split?
98. Предложите минимальный набор метрик мониторинга модели после внедрения в продакшен.
99. Спроектируйте мини-проект для студентов Beginner level по полному циклу аналитики данных: от постановки вопроса до интерпретации результата.
100. Оцените этические и правовые риски проекта по аналитике успеваемости студентов и предложите меры снижения этих рисков.

Итоговая контрольная работа оценивается в 50 баллов (максимум). В каждом билете содержится 5 вопроса. Баллы за каждый вопрос распределяются следующим образом:

- 1-й вопрос (легкий - Уровень «Знание») – 5 баллов
- 2-й вопрос (легкий - Уровень «Понимание») – 10 баллов
- 3-й вопрос (средний - Уровень «Применение») – 10 балов
- 4-й вопрос (средно-сложный - Уровень «Анализ») – 10 баллов
- 5-й вопрос (практическое задание - Уровень «Синтез / Оценка») – 15 баллов.

Старший преподаватель кафедры

конвергенции цифровых технологий **Р.Гаипназаров**