Контрольные и тестовые вопросы по ПР11

«Анализ и предсказание данных с применением RandomForest и линейной регрессии» по вариантам с ответами

Вариант 1

1. Какую роль играет параметр Ridge в линейной регрессии в RapidMiner?
   A) Устанавливает максимальную глубину деревьев
   B) Определяет минимальный прирост информации при разбиении узлов
   C) Регулирует регуляризацию модели для предотвращения переобучения
   D) Задает минимальное количество листьев в дереве

Ответ: C

2. Какая из метрик оценки качества регрессионных моделей наиболее чувствительна к выбросам данных?
   A) Mean Absolute Error (MAE)
   B) Mean Squared Error (MSE)
   C) Root Mean Squared Error (RMSE)
   D) Correlation

Ответ: C

3. Какое утверждение правильно характеризует принцип работы Random Forest?
   A) Используется один глубокий классификатор для всех данных
   B) Создаются несколько независимых деревьев с подмножеством признаков
   C) Используется только для категориальных данных
   D) Модель всегда строится без предварительной обрезки деревьев

Ответ: B

4. Что означает высокий коэффициент корреляции при анализе регрессионной модели?

A) Высокая точность и небольшие ошибки прогнозов модели
B) Наличие мультиколлинеарности среди независимых признаков
C) Низкая вариативность признаков в выборке
D) Переобучение модели на обучающей выборке

Ответ: A

5. Для чего в RapidMiner применяется предварительная обрезка деревьев (prepruning)?
   A) Для увеличения глубины каждого дерева
   B) Для ускорения вычисления модели путем отбрасывания менее значимых признаков
   C) Для предотвращения переобучения и повышения обобщающей способности модели
   D) Для генерации дополнительных подмножеств данных

Ответ: C

6. Какой параметр оператора Split Data определяет размер обучающей выборки?
   A) proportion
   B) ratio
   C) split size
   D) training volume

Ответ: B

7. Каким образом корреляционная матрица может быть использована при анализе данных?
   A) Для нормализации признаков перед обучением модели
   B) Для определения доли пропущенных значений в данных
   C) Для выявления линейной зависимости между признаками
   D) Для удаления всех сильно скоррелированных признаков

Ответ: C

8. Какие преимущества дает применение Random Forest по сравнению с одиночным решающим деревом при прогнозировании?
   Ответ: Random Forest снижает вероятность переобучения и повышает точность предсказаний за счет ансамбля деревьев, которые строятся на случайных подмножествах признаков и данных.

9. Почему нормализация данных важна перед построением регрессионных моделей?

Ответ: Нормализация позволяет привести данные к единому масштабу, предотвращая доминирование признаков с большими числовыми значениями, что улучшает точность и устойчивость модели.

10. Что отражает метрика RMSE в контексте качества модели?
    Ответ: Метрика RMSE отражает среднюю величину ошибки модели, измеряемую как квадратный корень из среднего квадрата отклонений предсказанных значений от фактических.

Вариант 2

1. Какой эффект достигается за счет применения оператора Apply Model в RapidMiner?
   A) Разделение данных на тренировочные и тестовые наборы
   B) Добавление к данным столбца с предсказанными значениями
   C) Установка ролей признаков в модели
   D) Вычисление метрик точности модели

Ответ: B

2. Что означает коэффициент корреляции, близкий к 0 в модели регрессии?
   A) Высокую точность модели
   B) Отсутствие линейной зависимости между реальными и предсказанными значениями
   C) Наличие мультиколлинеарности между признаками
   D) Значительную ошибку прогноза

Ответ: B

3. Как влияет параметр minimal leaf size в Random Forest на результат обучения?
   A) Регулирует количество признаков, используемых в каждом дереве
   B) Определяет минимальное количество наблюдений в каждом листе дерева
   C) Устанавливает минимальную глубину деревьев
   D) Управляет скоростью сходимости алгоритма

Ответ: B

4. Что определяет параметр minimal gain при построении деревьев в Random Forest?

A) Минимальный прирост информации, необходимый для дальнейшего разделения узлов
B) Максимальное количество признаков в каждом дереве
C) Глубину каждого дерева
D) Процент отбрасываемых признаков

Ответ: A

5. Какова основная причина использования метода Split Data при обучении моделей?
   A) Удаление выбросов
   B) Предотвращение переобучения путем тестирования модели на независимых данных
   C) Определение значимости признаков
   D) Нормализация значений атрибутов

Ответ: B

6. Для чего применяется корреляционная матрица в RapidMiner?
   A) Для оценки качества модели на новых данных
   B) Для проверки нормальности распределения данных
   C) Для выявления мультиколлинеарности признаков
   D) Для генерации дополнительных признаков

Ответ: C

7. Какое преимущество дает использование ансамблевых методов машинного обучения, таких как Random Forest?
   A) Повышает интерпретируемость модели
   B) Уменьшает вычислительные затраты на обучение модели
   C) Снижает риск переобучения и увеличивает точность прогнозов
   D) Упрощает визуализацию результатов

Ответ: C

8. В каких случаях может возникнуть проблема мультиколлинеарности в данных?
   Ответ: Проблема мультиколлинеарности возникает, когда несколько независимых переменных в данных сильно коррелируют друг с другом, что усложняет интерпретацию модели и снижает устойчивость оценки коэффициентов.

9. Почему важно оценивать модель по метрикам RMSE и Correlation одновременно?

Ответ: Совместная оценка RMSE и Correlation позволяет понять, насколько модель точна и насколько хорошо она описывает линейную зависимость между признаками и целевой переменной.

10. Как предварительная обрезка деревьев (prepruning) помогает предотвратить переобучение в Random Forest?
Ответ: Предварительная обрезка деревьев ограничивает сложность деревьев, предотвращая избыточное подстраивание под тренировочные данные и улучшая их обобщающую способность.

Вариант 3

1. Какова основная цель регуляризации (ridge-параметра) в линейной регрессии?
A) Ускорение процесса обучения модели
B) Предотвращение переобучения путем снижения весов признаков
C) Улучшение интерпретируемости признаков
D) Повышение чувствительности модели к шумам

Ответ: B

2. Как интерпретировать низкий RMSE и высокий коэффициент корреляции у регрессионной модели?
A) Модель имеет низкую точность и плохую предсказательную способность
B) Модель склонна к переобучению и нестабильна на новых данных
C) Модель показывает хорошую точность и высокую степень линейной зависимости
D) Признаки в модели мультиколлинеарны

Ответ: C

3. В чем преимущество метода случайных лесов (Random Forest) над линейной регрессией при работе с нелинейными зависимостями?
A) Лучшее описание линейных зависимостей
B) Возможность работы с категориальными признаками без преобразования
C) Способность захватывать сложные нелинейные зависимости в данных
D) Высокая чувствительность к выбросам данных

Ответ: C

4. Чем определяется глубина деревьев (maximal depth) в алгоритме Random Forest?
   A) Максимальным количеством признаков в дереве
   B) Количеством листьев в дереве
   C) Максимальным количеством уровней разделения узлов
   D) Минимальным приростом информации при разделении узла

Ответ: C

5. Какое последствие может иметь слишком большое значение минимального прироста (minimal gain) в Random Forest?
   A) Увеличение точности модели
   B) Недостаточная сложность модели и потеря значимых деталей данных
   C) Повышение чувствительности к выбросам
   D) Усиление мультиколлинеарности признаков

Ответ: B

6. Какой тип данных наиболее чувствителен к масштабированию перед применением регрессионных моделей?
   A) Категориальные данные
   B) Бинарные данные
   C) Числовые данные с широким диапазоном значений
   D) Номинальные данные с малым числом классов

Ответ: C

7. Что такое ансамбль моделей в машинном обучении?
   A) Единая глубокая модель с несколькими выходами
   B) Несколько простых моделей, объединенных для повышения общей точности
   C) Модель, работающая с разными типами данных одновременно
   D) Случайный выбор параметров модели для оценки их эффективности

Ответ: B

8. Почему корреляционная матрица важна при анализе данных для построения моделей?
   Ответ: Корреляционная матрица позволяет выявить сильные взаимосвязи между признаками и целевой переменной, а также

обнаружить мультиколлинеарность, что необходимо учитывать при создании стабильной модели.

9. Какие признаки в регрессионной модели могут создавать проблему мультиколлинеарности?
   Ответ: Признаки, имеющие высокие коэффициенты корреляции друг с другом, могут создавать проблему мультиколлинеарности, которая ухудшает устойчивость оценок и интерпретацию модели.

10. В чем особенность использования метрики RMSE при сравнении моделей?
    Ответ: RMSE квадратично увеличивает влияние крупных ошибок предсказаний, поэтому особенно полезна для задач, где критично избегать больших отклонений в прогнозах.

Вариант 4

1. Какое влияние оказывает увеличение количества деревьев в модели Random Forest?
   A) Повышает вероятность переобучения
   B) Уменьшает вычислительные затраты на обучение
   C) Увеличивает точность и стабильность прогнозов
   D) Снижает способность модели к обобщению

Ответ: C

2. Что произойдет, если параметр минимального размера листьев (minimal leaf size) в Random Forest задан слишком малым?
   A) Повышается скорость обучения модели
   B) Возникает риск переобучения из-за чрезмерного усложнения модели
   C) Увеличивается интерпретируемость модели
   D) Снижается точность прогнозов на обучающей выборке

Ответ: B

3. Какие признаки требуют нормализации перед использованием модели линейной регрессии?
   A) Только категориальные признаки
   B) Только числовые признаки с большим диапазоном значений
   C) Только признаки с нормальным распределением
   D) Признаки, не содержащие выбросов

Ответ: B

4. Чем обусловлено использование коэффициента корреляции при оценке качества регрессионной модели?
   A) Проверкой чувствительности модели к выбросам
   B) Определением наличия линейной зависимости между предсказаниями и реальными значениями
   C) Выявлением нелинейных зависимостей между признаками
   D) Определением размера тренировочной выборки

Ответ: B

5. Что происходит при использовании метода предварительного обрезания деревьев (prepruning)?
   A) Увеличивается сложность каждого дерева
   B) Снижается скорость обучения
   C) Ограничивается глубина и сложность деревьев для улучшения обобщающей способности
   D) Повышается чувствительность модели к шумам данных

Ответ: C

6. В каком случае RMSE будет предпочтительнее MAE (Mean Absolute Error) для оценки регрессионных моделей?
   A) Когда требуется минимизировать влияние мелких ошибок
   B) Когда необходимо подчеркнуть большие ошибки прогнозов
   C) При наличии большого числа категориальных признаков
   D) В случае мультиколлинеарности признаков

Ответ: B

7. Каким образом Split Data помогает предотвратить переобучение модели?
   A) Убирает выбросы из данных
   B) Делает модель менее чувствительной к шумам
   C) Оценивает модель на независимой тестовой выборке
   D) Генерирует дополнительные признаки

Ответ: C

8. Почему параметр "guess subset ratio" важен при работе алгоритма Random Forest?
   Ответ: Параметр "guess subset ratio" позволяет случайным образом

выбирать подмножества признаков для каждого дерева, что повышает разнообразие моделей в ансамбле и улучшает качество прогнозов.

9. Чем опасно наличие высокой корреляции между независимыми признаками при построении моделей?
   Ответ: Высокая корреляция между независимыми признаками ведет к проблеме мультиколлинеарности, которая усложняет оценку и интерпретацию параметров модели и снижает ее устойчивость.

10. Как интерпретировать результаты модели, если RMSE низкий, а коэффициент корреляции высокий?
    Ответ: Такие результаты означают, что модель обладает высокой точностью и успешно выявляет линейную зависимость между прогнозируемыми и фактическими значениями, что подтверждает её хорошую предсказательную способность.