ЦМвГН - квиз №1

* Indicates required question

1. Email *

2. Ваше имя и фамилия *

Квиз

3. Сопоставьте описание проблемы с её названием. * 3 points

1) Имеется датасет фотографий, в котором большинство детей сфотографировано на светлом фоне, а большинство взрослых - на темном. Модель, обученная на этом датасете, делает ошибочные предсказания на тесте, ориентируясь не на лицо, а на фон фотографии;

2) Полиномиальная регрессия 10й степени была натренирована на датасете из 100 примеров. Предсказания на тесте намного хуже предсказаний на валидации;

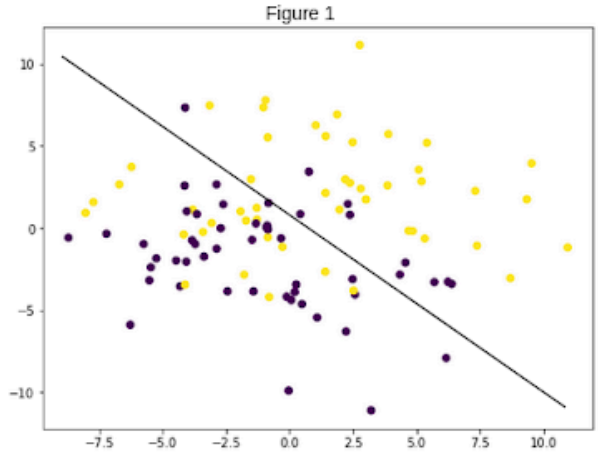
3) Линейная регрессия была применена на датасете из множества переменных. Предсказания на тесте и валидации неудовлетворительные.

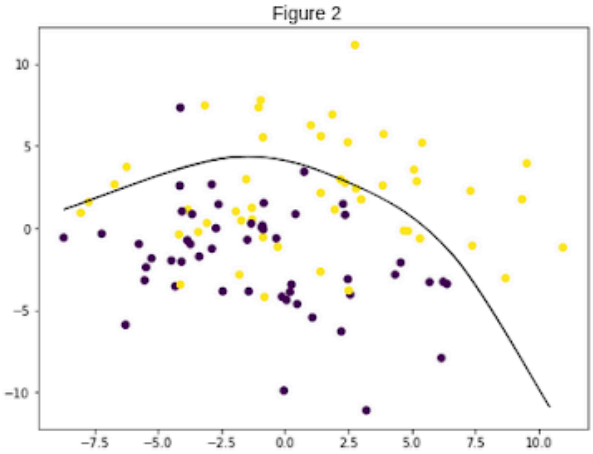
Mark only one oval per row.

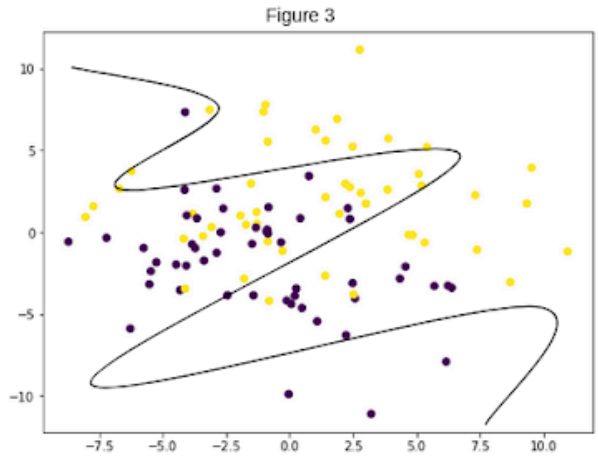
	Переобучение	Недообучение
1	◯	◯
2	◯	◯
3	◯	◯

4. Вы видите графики decision boundary трех разных классификаторов. Сопоставьте картинку и вероятные сведения о bias-variance trade-off классификатора. * 3 points

Внимание! Чтобы ответить на этот вопрос, подумайте, хорошо ли на самом деле работает третий классификатор?







Mark only one oval per row.

	Figure 1	Figure 2	Figure 3
Высокий bias, высокая дисперсия	◯	◯	◯
Высокий bias, низкая дисперсия	◯	◯	◯
Bias и variance сбалансированы	◯	◯	◯

5. Вы моделируете зависимость между ценой квартиры и количеством комнат. Между этими признаками наблюдается линейная зависимость, однако в ваших данных много аутлаеров: маленьких квартир по большой цене либо же больших квартир по скидке. На какую метрику оценки качества регрессии вы обратите внимание в первую очередь? * 1 point

Mark only one oval.

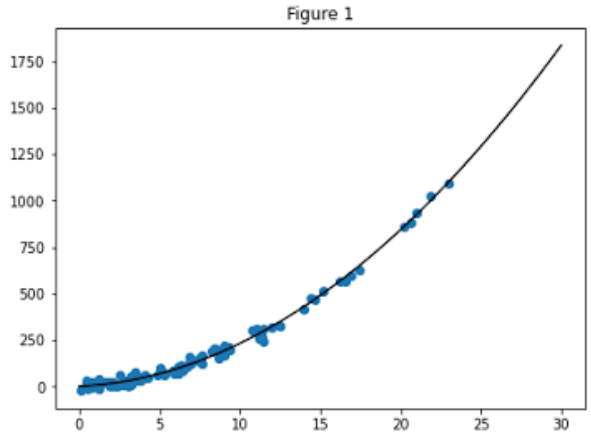
- Средняя абсолютная ошибка
- Среднеквадратичная ошибка
- p-value

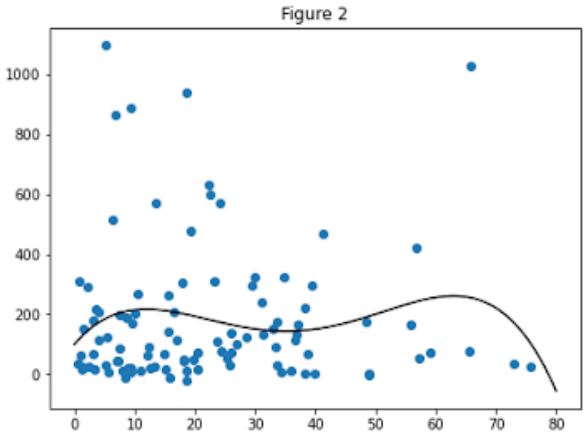
6. Вы обучили регрессию на большом количестве признаков (не менее 20). Какую метрику оценки качества вы примените в первую очередь? * 1 point

Mark only one oval.

- R^2
- Adjusted R^2
- Посмотрю на значения функции потерь

7. К какому типу относятся регрессии на картинках? * 1 point





Mark only one oval.

- 1 - линейная, 2 - полиномиальная
- 1 - логистическая, 2 - полиномиальная
- Обе полиномиальные

8. Имеется датасет со следующими колонками: * 1 point

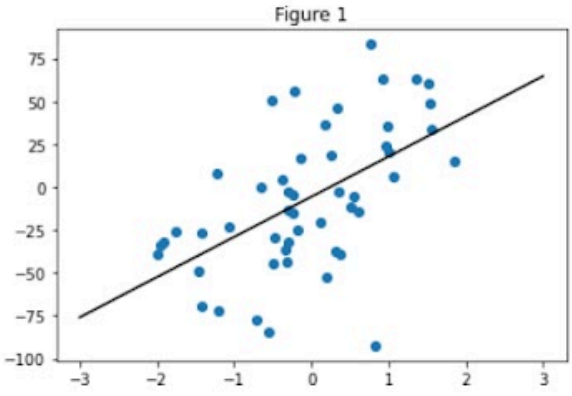
1) Балл студента за ЕГЭ по русскому языку (от 1 до 100);
2) Оценка за русский язык в аттестате (от 3 до 5);
3) Неокругленный итоговый балл студента за предмет "Общее языкознание" (от 0 до 10, возможны дробные оценки).

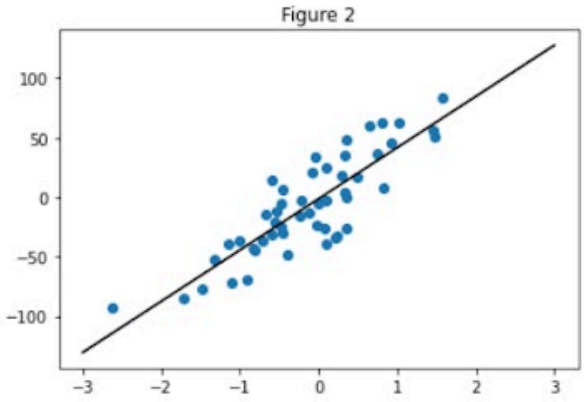
Вы хотите предсказать колонку 3 по колонкам 1 и 2. Какие шаги по предобработке данных вы предпримете? Отметьте все подходящие варианты.

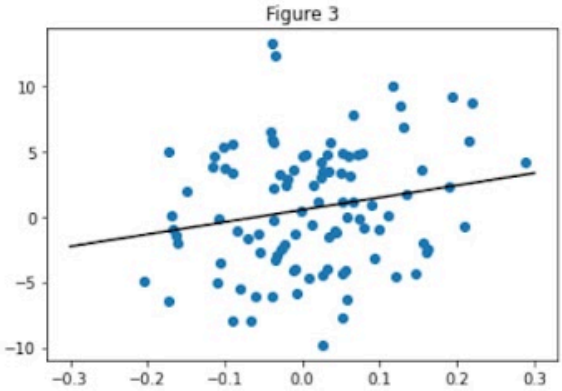
Check all that apply.

- [ ] Шкалирование колонок 1 и 2
- [ ] Кодирование колонки 2 при помощи OrdinalEncoder
- [ ] Удаление аутлаеров
- [ ] Кодирование зависимой переменной

9. Сопоставьте график и вероятное значение коэффициента детерминации (R^2) изображенной модели: * 3 points







Mark only one oval per row.

	0.73	0.04	0.30
Figure 1	◯	◯	◯
Figure 2	◯	◯	◯
Figure 3	◯	◯	◯

10. Сопоставьте задачу и тип классификации: * 3 points

Задача 1: разделите тексты на позитивные, негативные и нейтральные, учитывая, что у текста может быть только одна тональность;

Задача 2: определите авторство текста, выбрав между двумя авторами (они никогда не работали вместе);

Задача 3: определите тему текста, учитывая, что текст может относиться к нескольким темам одновременно.

Mark only one oval per row.

	Мультилейбл	Мультикласс	Бинарная классификация
Задача 1	◯	◯	◯
Задача 2	◯	◯	◯
Задача 3	◯	◯	◯

11. Что можно сказать об этом классификаторе? * 1 point

	precision	recall	f1-score	support
pos	0.59	0.40	0.48	50
neg	0.55	0.72	0.62	50
accuracy			0.56	100
macro avg	0.57	0.56	0.55	100
weighted avg	0.57	0.56	0.55	100

Check all that apply.

- [ ] Он хорошо находит примеры негативного класса среди всех, но хуже определяет, что пример именно негативный
- [ ] Он хорошо находит примеры позитивного класса среди всех, но хуже определяет, что пример именно позитивный
- [ ] В целом, негативный класс определяется хуже позитивного
- [ ] В целом, позитивный класс определяется хуже негативного

12. Как вы определите, какие признаки имеют наибольшее значение для логистической регрессии при принятии решений об отнесении к какому-либо классу? * 1 point

Mark only one oval.

- Посмотрю на confidence values
- Вызову функцию predict_proba() на тестовых данных
- Посмотрю на коэффициенты (атрибут coef_)

13. Вы читаете статью, в которой написано: "Мы использовали метод k ближайших соседей (kNN) для определения тематики текстов". Что вы думаете о корпусе, который использовался для статьи? Выберите наиболее вероятный вариант * 1 point

Mark only one oval.

- У авторов был большой корпус текстов, размеченный по темам
- Скорее всего, у авторов был неаннотированный корпус текстов
- Авторы разметили небольшой кусок неанноторированных данных

14. AUC вашего бинарного классификатора равна 0.42. Что это означает? * 1 point

Mark only one oval.

- Что классификатор угадывает половину отрицательных примеров и половину положительных
- Что классификатор чаще путает позитивные примеры с негативными, чем предсказывает правильно
- Что значения TPR в среднем намного превосходят значения FPR

15. Сопоставьте конкретную задачу и класс задач: * 4 points

Задача 1: предскажите среднее время чтения текста;

Задача 2: предскажите пол автора текста;

Задача 3: визуализируйте эмбеддинги ключевых слов вашего текста (вы уже нашли ключевые слова);

Задача 4: объедините способности нескольких моделей для решения одной и той же проблемы.

Mark only one oval per row.

	Классификация	Ансамблевое обучение	Снижение размерности	Регрессия
Задача 1	⬭	⬭	⬭	⬭
Задача 2	⬭	⬭	⬭	⬭
Задача 3	⬭	⬭	⬭	⬭
Задача 4	⬭	⬭	⬭	⬭

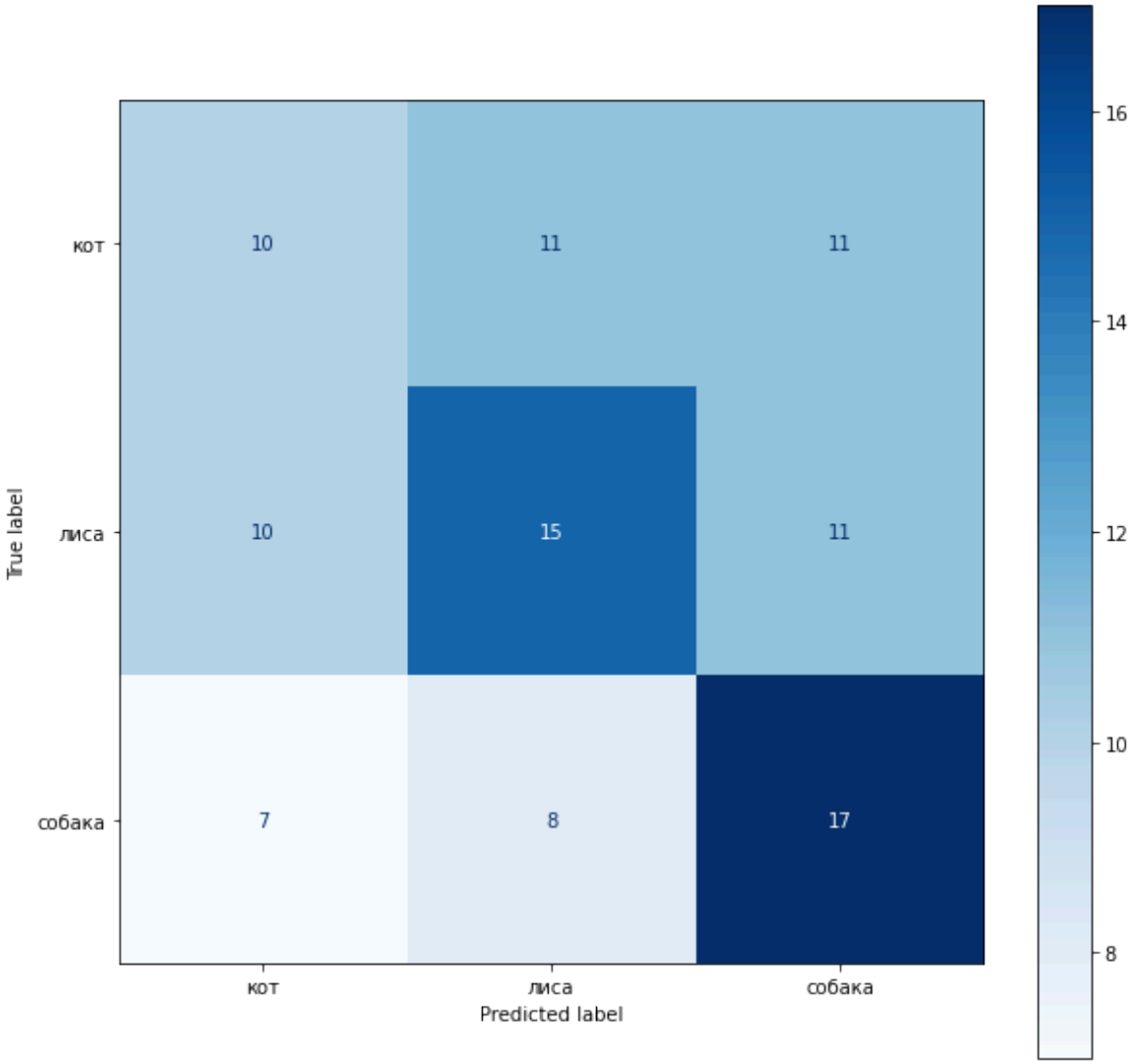
16. Рассмотрите псевдокод на картинке. Считайте, что X - корпус текстов. Зависимая переменная включает только те классы, которые вы видите на экране. Какую ошибку вы заметили? * 1 point

```
X = [.....]
y = [0, 1, 0, 0.....]


tf_idf = TfidfVectorizer()
X_transformed = tf_idf.fit_transform(X)

X_train, y_train, X_test, y_test = train_test_split(X_transformed, y)

lr = LogisticRegression()
lr.fit(X_train)
```

Check all that apply.

- [ ] Нужно было закодировать лейблы
- [ ] Строка 6: следовало применять метод fit векторайзера только на тренировочной выборке
- [ ] Строка 10: нужно выбрать другую модель

17. Посмотрите на приведенную матрицу путаницы и выберите ВСЕ ВЕРНЫЕ утверждения. Не забудьте: истинные лейблы классов находятся на оси Y, предсказанные моделью - на оси X. * 3 points



Check all that apply.

- [ ] Модель лучше всего определяет, что на картинке собака
- [ ] Модель чаще принимает лис за собак, чем за котов
- [ ] Модель чаще принимает лис за собак, чем собак за лис
- [ ] Модель реже принимает котов за лис, чем за собак