Курсовая работа

Машинное обучение

Состав курсовой работы

1. Титульный лист
   - Тема
   - Вариант
   - ФИО
2. Содержание
3. Описание задачи
4. Основные шаги выполнения
   1. EDA
   2. Подготовка данных
   3. Генерирование признаков
   4. Разделение на выборки
   5. Построение модели
      1. AutoML (baseline)
      2. Моделирование
   6. Лучшая модель
   7. Интерпретация лучшей модели
      1. Глобальная интерпретация
      2. Локальная интерпретация
5. Выводы
6. Приложение 1. Код в формате ipynb

Общие рекомендации

- Помним, что вы пишете курсовую, поэтому необходимо следовать академическому стилю изложения - https://bibloid.ru/kursach/text/academic-writing.html

- Не забывайте про нумерацию страниц!

- Форматирование текста - https://bibloid.ru/kursach/gost/formatting.html

- Орфография и пунктуация. Прежде чем сдавать ваш отчет еще раз проверьте ваш текст, например, при помощи https://languagetool.org/ru https://text.ru/spelling и других программ.

Титульный лист. Тема\Вариант\ФИО

Курсовая работа оформляется как текстовая работа с титульным листом, содержанием и текстом по правилам оформления текстовых документов.

Курсовая работа распечатывается и подписывается студентом и преподавателем.

Законченная курсовая работа защищается – студент отвечает на вопросы по теории, применённой в работе, и по написанному коду.

Оценивается:

- Решение задачи
- Правильность оформления
- Ответы на вопросы при защите

i Note

Тема курсовой. В рамках вашей курсовой вы занимаетесь прогнозированием/построением модели прогнозирования, но никак не оптимизируете или рассчитываете эффективность. Это разные задачи!

Описание задачи

В разделе «Описание задачи» студент кратко на 1 страницу описывает постановку задачи с точки зрения машинного обучения, описывает тип задачи машинного обучения, моделируемую величины и основные критерии успеха моделирования (метрики качества модели).

Note

Разделы "**Описание задачи**" и "**Выводы**" должны биться между собой.
В **Описании** вы пишите про то, какую задачу решаете (с точки зрения машинного обучения - какая ваша целевая переменная, какие данные у вас есть, какие метрики качества планируете использовать), описываете какие шаги вам нужно предпринять для достижения результата (по факту это определение вашей цели и задач https://bibloid.ru/kursach/start/research-goal.html).
В выводах же вы должны писать какие задачи были выполнены вами, какие результаты получены (лучшая модель, фичи, метрики качества) и что можно изменить для улучшения прогнозов.

Основные шаги выполнения. EDA

В разделе «EDA» студент кратко на 2-3 страницы описывает результаты EDA и приводит до 10 графиков с выводами по результатам EDA.
При этом:

- Осознанно стройте корреляционные матрицы. Не мешаете числовые и качественные данные. Указывайте способ оценки корреляции
- Не выводите график ради графики, по каждому должен быть вывод. Например, демонстрация необходимости нормирования данных, необходимость выделения временного интервала и т.п.

i Note

Ваш отчет должен состоять не только из описания действий, которые вы совершили над вашими данными.
Необходимо описывать:

- что вы видите на графиках, которые построили;
- что обнаружили при построения профилирования (pandas profiling);
- по какой причине решили удалить фичи из дата сета и т.д.

Основные шаги выполнения. Подготовка данных

В разделе «Подготовка данных» студент кратко на 1-3 страницы описывает выполненные шаги по подготовке данных.
Это могут быть:

- Преобразования типов
- Кодирование признаков
- Вычисляемые признаки
- Удаление дубликатов
- Заполнение пропусков
- Обогащение данных
- Устранение дисбаланса классов
- Нормализация значений
- И т.д.

В конце раздела приводится описание витрины, которая будет использоваться для моделирования в виде таблицы.

Основные шаги выполнения. Разделение на выборки

В этом разделе студент описывает разделение на обучающую и тестовую выборку.
Какое соотношение, какой метод разделения и почему.

Основные шаги выполнения.
Генерирование признаков

В это разделе студент описывает:

1. Исходное количество признаков, дает оценку предсказательной силе данных (с точки зрения бизнеса или статистического анализа, предварительного моделирования)
2. Указывает необходимость расширение количества признаков
3. Обосновывает выбор способа генерации признаков
4. Описать сгенерированные признаки и их отличия от изначальных

Основные шаги выполнения. Построение модели.

Раздел «Построение» модели должен содержать 2 подраздела.

Первый «Построение модели с помощью библиотеки AutoML (baseline)»
В этом разделе студент описывает выбранную библиотеку AutoML и результаты построения модели с помощью этой библиотеки, включая полученные метрики качества (зависят от задачи).

По данному разделу необходимо уметь отвечать на вопросы:

- Какую библиотеку AutoML вы использовали?
- Какие результаты получены?
- Какие параметры для запуска моделирования вы использовали\почему\что они означают?

Основные шаги выполнения. Построение модели.

Второй подраздел «Моделирование».
В этом разделе студент коротко перечисляет использованные им алгоритмы для моделирования и описывает полученные результаты (метрики качества модели).

По данному разделу необходимо уметь отвечать на вопросы:

- Какую алгоритмы вы использовали?
- Какие результаты получены?
- Какие параметры для запуска моделирования вы использовали\почему\что они означают?
- Почему именно этот алгоритм подходит лучше всего?
- Как работают выбранные вами алгоритмы?

Основные шаги выполнения. Лучшая модель

В разделе «Лучшая модель» приводятся параметры вашей лучшей модели, указываются её метрики качества и насколько они лучше, чем результаты AutoML.

Основные шаги выполнения.
Интерпретация лучшей модели

Этот раздел состоит из двух подразделов:

1. Глобальная интерпретация
2. Локальная интерпретация

В подразделе «Глобальная интерпретация» студент проводит интерпретацию модели с помощью PD или SHAP (графики), показывает наиболее значимые признаки.
Если ваш алгоритм поддерживает возможность показать важность переменных – сделайте это.
Если ваш лучший алгоритм – дерево решений, то необходимо привести визуализацию дерева решений.

По данному разделу необходимо уметь отвечать на вопросы:

- Какие признаки (факторы) наиболее важны для модели?

Основные шаги выполнения.
Интерпретация лучшей модели

Этот раздел состоит из двух подразделов:

1. Глобальная интерпретация
2. Локальная интерпретация

В подразделе «Локальная интерпретация» студент выбирает 3 интересных наблюдения из тестовой выборки и проводит их интерпретацию модели с помощью LIME или SHAP (графики), показывает наиболее значимые признаки для этого наблюдения для получившего прогноза.

По данному разделу необходимо уметь отвечать на вопросы:

- Какие признаки (факторы) наиболее важны в конкретном случае?
- Какие признаки влияют положительно на прогноз, а какие – отрицательно?

Выводы

В этом разделе студент коротко описывает выводы по работе.

Например,
Какая лучшая модель, какие метрики качества, какие факторы наиболее значимы для прогноза, хорошая ли получилась модель, какие проблемы есть в данных, что ещё можно попробовать для улучшения прогноза.

Приложение 1. Код в формате ipynb

К работе студент прикладывает блокнот с кодом, который можно запустить и получить результаты, отражённые в пояснительной записке.

В коде должно быть указаны версии используемых вами библиотек.

ДОПОЛНИТЕЛЬНЫЕ БАЛЛЫ

Для получения более высокой оценки студент **может**:

- Придумать и обосновать собственную метрику оценки качества
- Использовать данные, не представленные в исходном наборе (взять данные о погоде, стоимости, географии и т.д.)
- Создать приложение для прогноза на основе созданной модели (например, с использованием Streamlit https://streamlit.io/gallery)
- Использовать ансамбли из нескольких моделей для моделирования
- Добавить кластеризацию и отдельные модели по каждым кластерам, если это необходимо
- Использовать алгоритмы по поиску и устранению аномалий в данных
- Использовать нейронные сети
- Любое расширение задачи и её обоснованное усложнение
- Самостоятельно сформулированная тема работы **(необходимо сообщить до конца апреля)**

ВАРИАНТЫ ЗАДАНИЙ

Всего представлено 10 вариантов заданий.

Номер вашего варианта определяется порядковым номеров в вашей группе по следующим правилам:

- Порядковый номер в группе с 1 по 10 – вариант с 1 по 10
- Порядковый номер в группе с 11 по 20 – вариант с 10 по 1
- Порядковый номер в группе с 21 по 30 – вариант с 1 по 10

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 1.

Набор данных: Dataset 1
Данные об отправке грузов

Дополнительные условия:
Возьмите только тип парка «Транзитный». (Поле «Тип парка» содержит «Транзитный»)

Моделируемая величина:
«Срок доставки факт»

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 2.

Набор данных: Dataset 1
Данные об отправке грузов

Дополнительные условия:
Возьмите только тип парка «Местный». (Поле «Тип парка» содержит «Местный»)

Моделируемая величина:
«Срок доставки факт»

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 3.

Набор данных: Dataset 1
Данные об отправке грузов

Дополнительные условия:
Возьмите только тип парка «Порожний». (Поле «Тип парка» содержит «Порожний»)

Моделируемая величина:
«Срок доставки факт»

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 4.

Набор данных: Dataset 1
Данные об отправке грузов

Дополнительные условия:
Возьмите только оператора АО "ПГК" . (Поле «Оператор» - АО "ПГК")

Моделируемая величина:
«Срок доставки факт»

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 5.

Набор данных: Dataset 1
Данные об отправке грузов

Дополнительные условия:
Возьмите всех операторов, кроме АО "ПГК" . (Поле «Оператор» - **НЕ** АО "ПГК")

Моделируемая величина:
«Срок доставки факт»

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 6.

Набор данных: Dataset 2
Комплексное исследование по поведению во время ежедневных поездок. Целью этого проекта является прогнозирование предпочтительного способа передвижения (например, автомобиль, велосипед, общественный транспорт) на основе личных и бытовых характеристик, включая местоположение рабочего места и дома.

Дополнительные условия:
KAFSTV(Distance class trip) до 10 км

Моделируемая величина:
'KHVM' (Main transport mode class trip)

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 7.

Набор данных: Dataset 2
Комплексное исследование по поведению во время ежедневных поездок. Целью этого проекта является прогнозирование предпочтительного способа передвижения (например, автомобиль, велосипед, общественный транспорт) на основе личных и бытовых характеристик, включая местоположение рабочего места и дома.

Дополнительные условия:
KAFSTV(Distance class trip) больше 3.5 км

Моделируемая величина:
'KHVM' (Main transport mode class trip)

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 8.

Набор данных: Dataset 2
Комплексное исследование по поведению во время ежедневных поездок. Целью этого проекта является прогнозирование предпочтительного способа передвижения (например, автомобиль, велосипед, общественный транспорт) на основе личных и бытовых характеристик, включая местоположение рабочего места и дома.

Дополнительные условия:
DAGSOORT(type of day: weekday, weekend)=1 (weekday). (Путешествия по рабочим дням)

Моделируемая величина:
'KHVM' (Main transport mode class trip)

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 9.

Набор данных: Dataset 2
Комплексное исследование по поведению во время ежедневных поездок. Целью этого проекта является прогнозирование предпочтительного способа передвижения (например, автомобиль, велосипед, общественный транспорт) на основе личных и бытовых характеристик, включая местоположение рабочего места и дома.

Дополнительные условия:
VPLDAGNR in 0,1,2 (not 3)

Моделируемая величина:
'KHVM' (Main transport mode class trip)

ВАРИАНТЫ ЗАДАНИЙ. ВАРИАНТ 10.

Набор данных: Dataset 2
Комплексное исследование по поведению во время ежедневных поездок. Целью этого проекта является прогнозирование предпочтительного способа передвижения (например, автомобиль, велосипед, общественный транспорт) на основе личных и бытовых характеристик, включая местоположение рабочего места и дома.

Дополнительные условия:
TOER=0

Моделируемая величина:
'KHVM' (Main transport mode class trip)