СЕССИЯ 2

1) Data.zip (Результаты предыдущей сессии)
2) Машинное обучение – С2.pdf (Инструкция к 2 сессии)

Результаты работы:

1) Data.zip (Предобработанные данные)
2) Report_C2.html, Report_C2.ipynb (Отчет о проделанной работе)
3) Readme.txt (Дополнительные комментарии)

ВВЕДЕНИЕ

В этой сессии вы продолжаете работать с данными, подготовленными в предыдущей сессии. Предстоит провести разведочный анализ для изучения имеющихся данных (EDA) и заняться построением признаков.

Какая-либо работа, обусловленная задачами предыдущей сессии, выполненная в ходе текущей, оцениваться не будет, поэтому проделывайте её только в случае необходимости.

ЗАДАНИЕ

2.1 Визуальный анализ данных

Используя программные средства, визуализируйте зависимости атрибутов в наборе данных. Визуализация должна отражать влияние атрибутов на целевую переменную. Приведите интерпретацию полученным результатам.

2.2 Конструирование признаков (Feature Engineering)

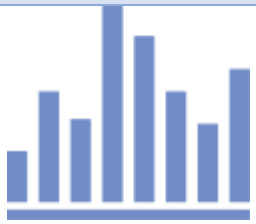
Создайте такой признак, который будет представлять собой число вида ГГГГММ от той даты, когда проходила сессия. Например, 201407 - 2014 год и 7 месяц. Таким образом, мы будем учитывать помесячный линейный тренд за весь период предоставленных данных. Добавьте новые признаки, которые на ваш взгляд позволят улучшить качество выбранной модели. Напишите функцию для создания новых признаков и примените ее к исходным данным. Опишите приемы генерации новых данных и результаты. Проведите выбор признаков, т.е. удалите часть признаков, чтобы помочь модели лучше обобщать новые данные ради повышения её точности.

Проведите визуальное исследование полученных признаков и сделайте вывод об их значимости.

2.3 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Отчет должен включать следующие пункты:

- Выбор способов визуализации
- Результаты визуализации
- Выбор и обоснование дополнения выборки новыми признаками



– Анализ визуального исследования новых признаков

Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.