Задание на курсовую работу

**Оп**исан**ие исходных данных**

goal\_cipher.csv – целевые метки

feature\_data\_new\_mini.csv – массив признаков (закодированные названиями фруктов событий)

Варианты.xlsx – файл с указанием, какую целевую метку из goal\_cipher.csv использовать

**Порядок выполнения**

1. **Загрузка массива целей**

Определите целевой массив исходя из вашего варианта

Изображение выглядит как текст, стол

Автоматически созданное описание

Столбцы файла Варианты.xlsx (fail\_reason, fail\_place, fail\_type) есть также в feature\_data\_new\_mini.csv.

* 1. Загрузите массив целей
  2. Сформируете новый массив из feature\_data\_new\_mini.csv по соответствующему столбцу и названию фрукта
  3. Проверьте размер массива

Пример кода для пунктов 2-3:

goal = pd.read\_csv('goal\_cipher.csv')

my\_goal = goal[goal['fail\_reason'] == 'Барбадосская вишня']

len(my\_goal.index)

* 1. Если количество строчек меньше 500 свяжитесь с лектором.

1. **Объединения массива целей с массивом признаков**

Ключи для объединения: ser\_loc, zns\_loc, prs\_loc

1. **Выполнить обработку пропусков после объединения. Обосновать замену или удаления пропусков.**
   1. Совет 1. Пропущенный day означает отсутствие отказа в текущем месяце. Придумайте чем заменить.
   2. Вы можете удалить строки с пропущенными значениями. В этом случае обоснуйте что будет предсказывать ваша модель
2. **Выбрать метод обучения: регрессия или классификация.**
3. **Исходя из выбранного методы преобразовать целевые метки.**
   1. Если в качестве метода обучения выбирается регрессия, то целевой меткой является признак day из массива целей.
   2. Если в качестве метода обучения выбирается классификатор, то возможны следующие стратегии дискретизации:
      1. 0 – не было отказа, 1 – был отказ.
      2. 1 – отказ в первой половине месяца, 0 – во второй.

Можете предложить свою стратегию.

1. **Инженерия данных**
   1. Определите необходимость нормирования данных. Имейте в виду, что у вас в работе 255 столбцов. Не все встроенные методы это осилят.
   2. Не забудьте что ser\_loc, zns\_loc, prs\_loc – столбцы идентификаторы, а не признаки. Лучше направить их в индексы.
   3. Проведите балансировку данных при необходимости.
   4. Разделите ваш массив на обучающую и тестовую выборку.
2. **Обучение моделей**
3. **Проверка качества**
   1. Напишите собственную функцию для проверки качества.
   2. Сравните встроенные функции проверки качества с вашими.