Блок 1: Что джун сделал не так?

1. Предобработка данных

Ошибка: Джун игнорировал пропуски и аномальные значения, считая их несущественными, либо просто удалял строки с такими данными.

Правильный подход: Необходимо тщательно обработать пропущенные и аномальные данные. Пропуски могут быть заполнены средними значениями, медианными значениями или интерполяцией. Аномальные значения должны быть выявлены и обработаны методами нормализации или устранения выбросов.

Пример заполнения пропусков медианными значениями

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='median')

X\_train\_imputed = imputer.fit\_transform(X\_train)

#### 2. Выбор модели

Ошибка: Джун выбрал модель без должной подготовки и настройки параметров, полагаясь лишь на звучащее название ("случайный лес").

Правильный подход: Важно выбрать подходящую модель, основываясь на характере данных и задачах. Гиперпараметры следует подбирать через перекрестную проверку (cross-validation), чтобы избежать переобучения и улучшить общую производительность модели.

# Пример подбора гиперпараметров для Random Forest

from sklearn.model\_selection import GridSearchCV

from sklearn.ensemble import RandomForestClassifier

param\_grid = {

'n\_estimators': [100, 200],

'max\_depth': [None, 10, 20]

grid\_search = GridSearchCV(RandomForestClassifier(), param\_grid, cv=5)

grid\_search.fit(X\_train, y\_train)

best\_params = grid\_search.best\_params\_

print("Best parameters:", best\_params)

#### 3. Анализ корреляций

Ошибка: Джун провел поверхностный анализ корреляционных связей между признаками, не углубляясь в детальный разбор взаимосвязей.

Правильный подход: Анализ корреляционной матрицы позволяет выявить признаки, сильно коррелирующие друг с другом, что может негативно сказаться на производительности модели. Также важно исследовать причинно-следственные связи между переменными.

import seaborn as sns

import matplotlib.pyplot as plt

corr\_matrix = df.corr()

sns.heatmap(corr\_matrix, annot=True)

plt.show()

Блок 2: Восстановление ситуации

Теперь, когда мы знаем, что было сделано неправильно, давай посмотрим, как можно восстановить ситуацию и правильно подойти к решению задачи.

1. Предобработка данных

- Проверь наличие пропусков и выбросов.

- Заполни пропуски подходящим способом (среднее значение, медиана, интерполяция).

- Нормализуй или устрани выбросы.

2. Выбор модели

- Подбери несколько моделей, подходящих под задачу (например, логистическая регрессия, дерево решений, градиентный бустинг).

- Используй кросс-валидацию для подбора оптимальных гиперпараметров каждой модели.

3. Анализ корреляций

- Построить корреляционную матрицу и проанализировать её.

- Удали сильно коррелированные признаки, оставив наиболее информативные.

Заключение

Тимлид всегда знает, что нужно делать! Теперь, когда ты исправил ошибки джуна, можно двигаться дальше и создавать качественную модель для прогнозирования оттока клиентов.