

The background features a dark blue gradient with abstract geometric shapes. A large blue triangle is on the left, and a green triangle is positioned below it. In the bottom left, there is a circular inset showing a detailed view of a circuit board. The top right corner has a pattern of grey, stepped rectangular blocks.

Чемпионат - Ярославская область

Прогнозирование риска развития сердечно-сосудистого заболевания пациента

Соколов Д.И.



Задача

Данные содержат опрос реальных пациентов и их диагнозы, включая такие не прямые показатели, как образование, этнос, вид работы и многие другие.

В рамках чемпионата вам необходимо классифицировать наличие/отсутствие у пациента следующих заболеваний:

- Артериальная гипертензия.
- Острое нарушение мозгового кровообращения
- Стенокардия, ИБС, инфаркт миокарда.
- Сердечная недостаточность.
- Прочие заболевания сердца.

Метрика:

- Recall



Используемое ПО и библиотеки

- Pandas (данные, корреляция)
- scikit-learn (метрики)
- seaborn (визуализация)
- CatBoost (градиентный бустинг)

Анализ данных

- Дисбаланс классов.
- Слабая корреляция между независимыми и целевыми переменными (в основном только с 1-й и 3-й целевой переменными).
- Высокая корреляция целевых переменных между собой.

Артериальная гипертензия	1	0.12	0.3	0.27	0.087
ОИМК	0.12	1	0.015	-0.0021	-0.031
Стенокардия, ИБС, инфаркт миокарда	0.3	0.015	1	0.52	0.05
Сердечная недостаточность	0.27	-0.0021	0.52	1	0.14
Прочие заболевания сердца	0.087	-0.031	0.05	0.14	1
	Артериальная гипертензия	ОИМК	Стенокардия, ИБС, инфаркт миокарда	Сердечная недостаточность	Прочие заболевания сердца

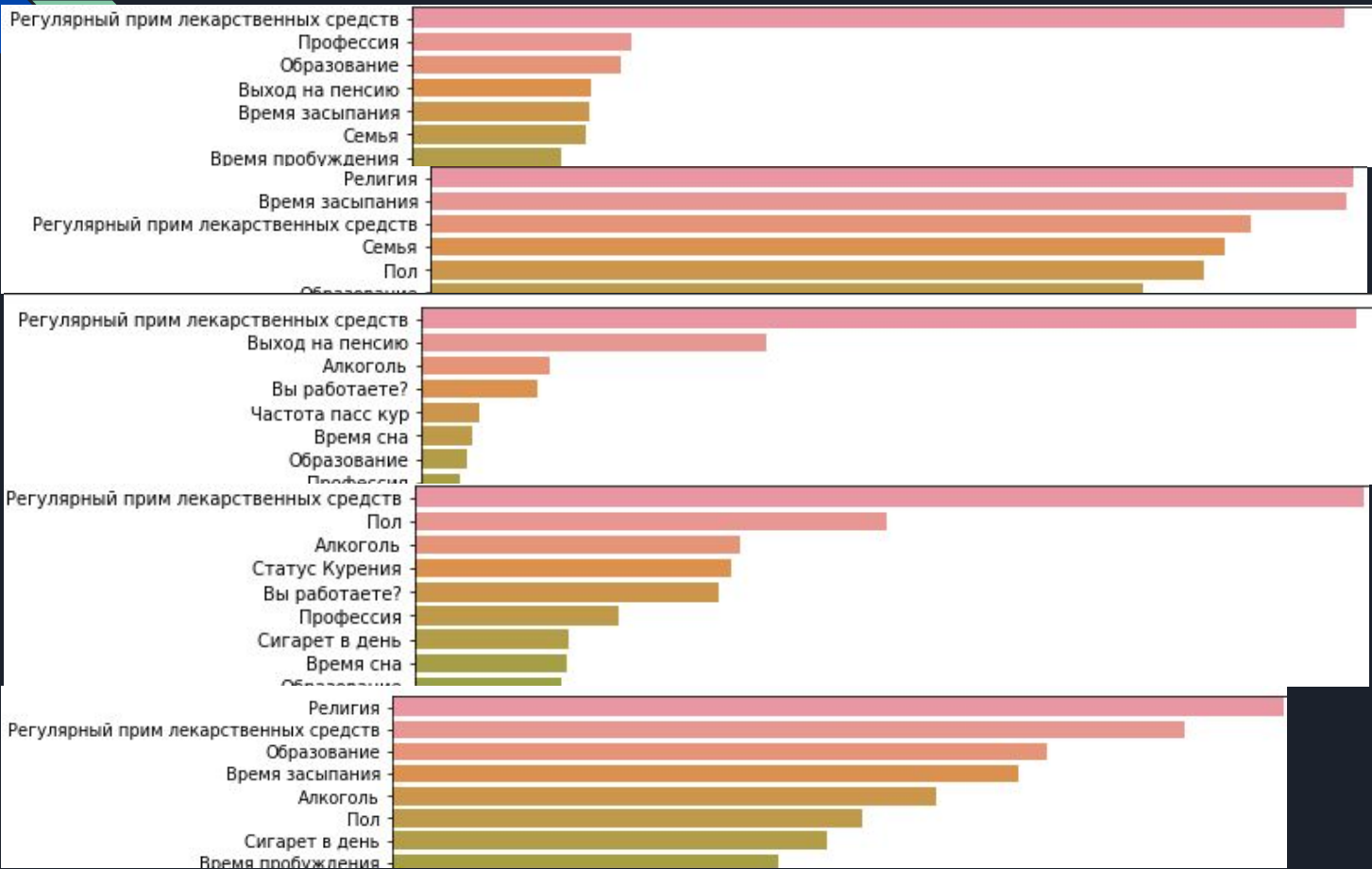


Решение

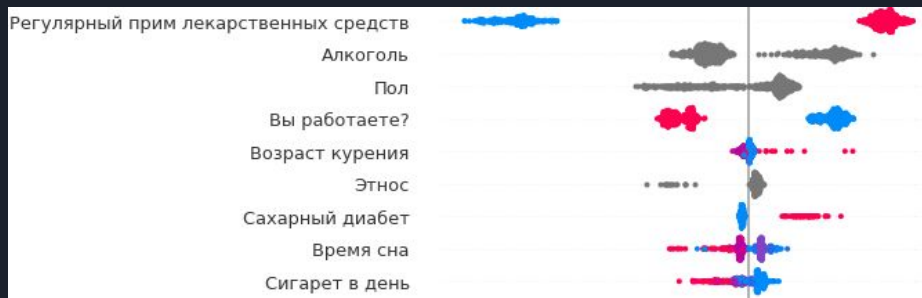
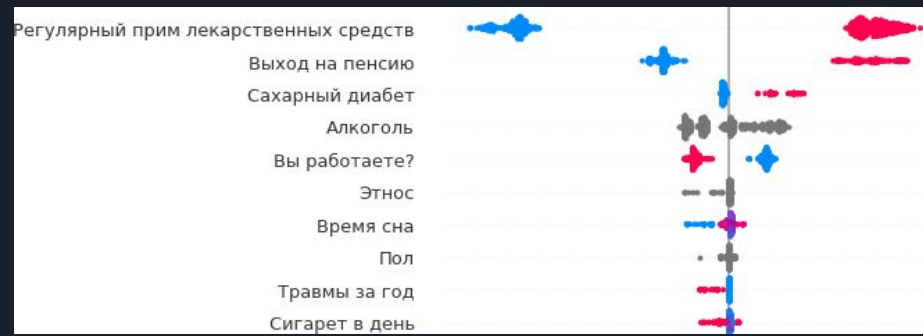
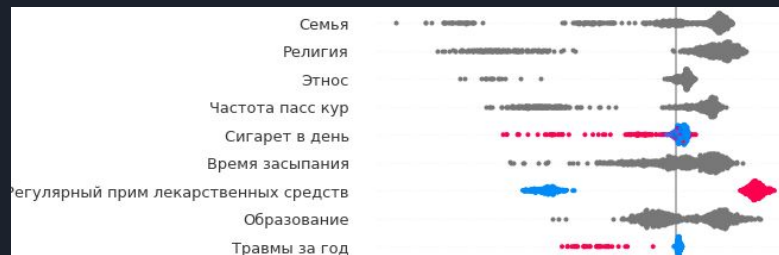
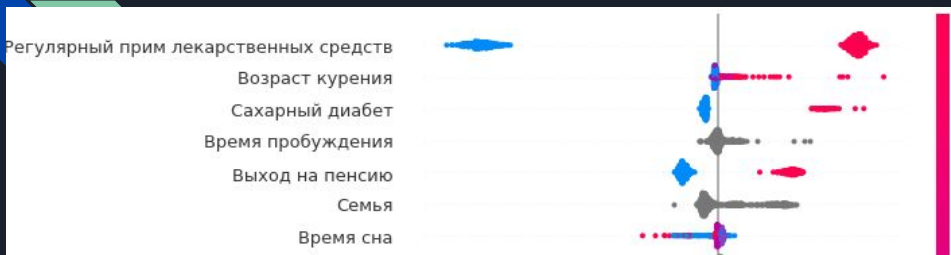
- Добавление признака “время сна”.
- Замена NaN's.
- Разделение трейн - тест (0.3-0.1).
- Подбор нескольких параметров модели вручную.
- Прогнозирование переменных по отдельности (Public score ~ 0.66).
- Локальная валидация не очень стабильная из-за небольшого размера датасета и сильно расходится с Public score

Recall on Public: ~0.664

Важность признаков



Важность признаков (SHAP)



Довольно много “важных” признаков выглядят случайными, можно сделать вывод, что они вряд ли имеют зависимость в реальности и модель не очень качественная.





Решение

Решено использовать, факт о высокой корреляции целевых признаков между собой и сделать поэтапное предсказание.

Обучаем несколько моделей: каждая модель предсказывает целевой признак, используя разное количество независимых переменных (каждый раз добавляем новую целевую переменную к независимым переменным).

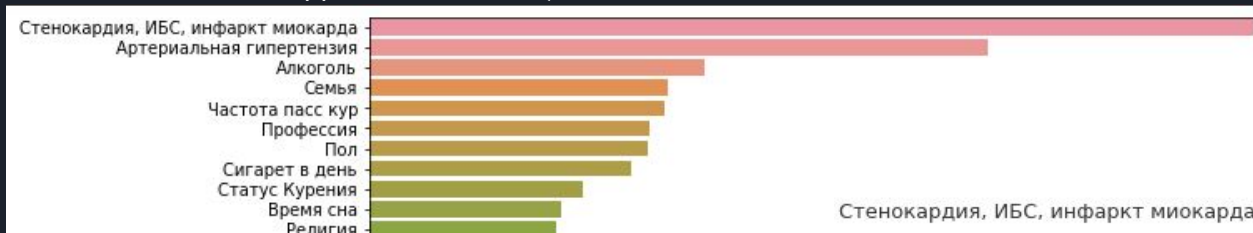
На этапе предсказания на каждом шаге предсказываем новую целевую переменную и добавляем предсказание к независимым переменным, для модели на следующий шаг.

Объединяем предсказания 5-ти моделей.

Recall on Public: ~0.705

Важность признаков


Как видим, модель успешно использует ранее предсказанные признаки, для обучения предсказания новых (4-й шаг, обучение для предсказания признака 'Сердечная недостаточность').






Что не сработало

- Multiclass, MultiLogLoss (CatBoost) для Multilabel предсказания - низкая точность (~ 0.53), медленно обучается, не все метрики доступны, нельзя установить веса для балансировки классов. Нужен пре/пост процессинг данных для Multiclass.
- Ручной подбор порогов предсказанных вероятностей.
- RidgeClassifier, RandomForest, ExtraTrees. (Что-то предсказывают только для 1-й целевой переменной).
- В OneVsRestClassifier нельзя передать список категориальных переменных для CatBoost.



Планы на будущее (не хватило времени)

- Кросс-валидация.
- Предсказание целевых переменных в разном порядке.
- Усреднение предсказаний.
- Стекинг разных моделей.
- XGBoost, LightGBM.



Спасибо за внимание!

Telegram: @dimkoss11