Прогнозирование риска развития сердечно-сосудистого заболевания пациента

Модель машинного обучения

Решение задачи

Лисовский Дмитрий Александрович

Коротко о себе

Лисовский Дмитрий Александрович, 29 лет

Образование: Врач (Лечебное дело, 2016 год)

Специальности:
- ➔ Хирургия
- ➔ Организация здравоохранения и общественное здоровье

Дополнительное образование:
Аналитик BI (Нетология)
Data Scientist (Skillfactory)

Предыдущие места работы:

НИИ ОЗИММ ДЗМ - Аналитик

Сеченовский Университет - Директор Департамента развития регионального здравоохранения

Актуальность

Умершие в РФ по классам причин смерти 2021 год

Причина смерти	на 100 000	%
Болезни системы кровообращения	638.9	38,3 %
COVID-19	318.4	19,0 %
Новообразования	193.7	11,6 %
Внешние причины смертности	95.1	5,7 %
Другое	424.0	25,4 %

Росстат, gks.ru

Формулировка задачи

Задача — разработать модель машинного обучения, цель которой — предсказание наличия **сердечно-сосудистых заболеваний**.

Перечень сердечно-сосудистых заболеваний для предсказания:

1. Артериальная гипертензия
2. Ишемическая болезнь сердца
3. Сердечная недостаточность
4. Острое нарушение мозгового кровообращения
5. Другие заболевания сердца

Описание данных

Всего признаков (столбцов) - **39** (в том числе **5 целевых**):

Величина тренировочного датасета - **955 строк**

Величина тестового датасета - **638 строк**

Данные являются результатом заполнения опросника

Количество опрошенных у которых есть заболевания (всего 955 человек)				
АГ	ОНМК	ИБС	СН	Другое
509	**41**	**117**	**96**	**86**

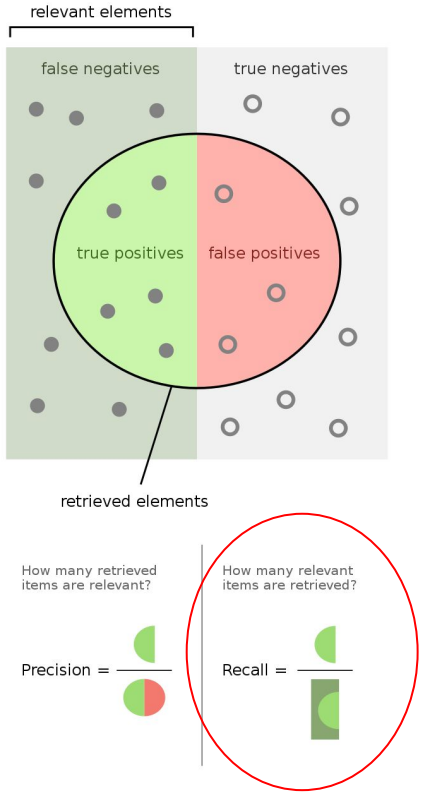
Уточнение задачи

В связи с малым количеством наблюдений и **малым** числом опрошенных, которые имеют заболевания такие как ИБС, СН, ОНМК и “Другие”:

Задача - выявление **высокого риска** наличия/развития конкретных заболеваний

Таким образом, результат работы модели следует трактовать как **“высокий риск развития заболевания”**, а не наличие самого заболевания.

В данном случае оправдано использование метрики среднего **recall**



https://en.wikipedia.org/

Список признаков

'ID'	***'id'***	идентификатор опрошенного
'Пол'	***'sex'***	пол
'Семья'	***'family'***	семейное положение
'Этнос'	***'ethnos'***	этнос
'Национальность'	***'nationality'***	национальность
'Религия'	***'religion'***	отношение к религиозной группе
'Образование'	***"education"***	уровень образования
'Профессия'	***"profession"***	профессия
'Вы работаете?'	***'job'***	наличие работы в данный момент
'Выход на пенсию'	***'retire'***	является ли опрошенный пенсионером
'Прекращение работы по болезни	***'stop_work_due_disease'***	связано ли прекращение работы с болезнью
'Сахарный диабет'	***'diabetes'***	наличие сахарного диабета (любого типа)
'Гепатит'	***'hepatitis'***	наличие гепатита
'Онкология'	***'oncology'***	наличие онкологических заболеваний
'Хроническое заболевание легких'	***'chronic_lung_disease'***	наличие хронического заболевания легких
'Бронжиальная астма'	***'bronchial_asthma'***	наличие бронхиальной астмы
'Туберкулез легких '	***'tuberculosis'***	наличие туберкулеза
'ВИЧ/СПИД'	***'hiv/aids'***	наличие ВИЧ/СПИД
'Регулярный прием лекарственных средств'	***'intake_medicines'***	факт регулярного приема лекарственных средств
'Травмы за год'	***'trauma_last_year'***	факт получения травм за последний год
'Переломы'	***'fractures'***	наличие переломов
'Статус Курения'	***'smoking'***	статус курения
'Возраст курения'	***'smoking_duration'***	длительность курения

'Сигарет в день'	***'ciggaretes_per_day'***	количество сигарет в день
'Пассивное курение'	***'passive_smoking'***	наличие курящих в окружении
'Частота пасс кур'	***'passive_smoking_frequency'***	частота нахождения среди курящих
'Алкоголь'	***'alcohol'***	отношение к приему алкоголя
'Возраст алког'	***'alcohol_duration'***	длительность употребления алкоголя (годы)
'Время засыпания'	***'time_sleep_onset'***	время засыпания
'Время пробуждения'	***'time_sleep_upset'***	время пробуждения
'Сон после обеда'	***'midday_sleep'***	сон в середине дня (после обеда)
'Спорт, клубы'	***'sport'***	факт регулярного занятия спортом(фитнесом)
'Религия, клубы'	***'religion_clubs'***	факт регулярного посещения религиозных мероприятий, групп
'ID_y'	***'id_y'***	идентификатор (связь с таблицей диагнозов)
'Артериальная гипертензия'	***'arterial_hypertension'***	наличие артериальной гипертензии
'ОНМК'	***'stroke'***	наличие острого нарушения мозгового кровообращения (инсульта)
'Стенокардия, ИБС, инфаркт миокарда'	***'IHD'***	наличие ишемической болезни сердца (все формы)
'Сердечная недостаточность'	***'heart_failure'***	наличие сердечной недостаточности
'Прочие заболевания сердца'	***'other_cardio_diseases'***	наличие других сердечных заболеваний

Дополнительные признаки

unknown_id_feature	Последняя цифра id (всего возможных цифр 4) представленная как порядковый признак
'id_1', 'id_2', 'id_3', 'id_4'	Последняя цифры id представленная как 4 отдельных колонки (закодированы как 1/0)
'comorbid_count'	Количество болезней (сумма по столбцам diabetes, chronic_lung_diseases и др.)
'is_comorbid'	Наличие хотя бы одного заболевания (в столбцах diabetes? chronic_lung_diseases и др.)
'fracture_last_year'	Признак о переломе кости за последний год (предположительно)
'trauma_on_retire'	Признак о наличии травмы за последний год будучи в пенсионном возрасте
'fracture_on_retire'	Признак о переломе кости за последний год (предположительном) будучи на пенсии
'smoker_score'	Рейтинг курильщика (0 - отсутствие курение, 1 - курение в прошлом, 2 - курение в настоящее время)
'smoking_duration_score'	Рейтинг длительности курения (каждые 5 лет курения - 1 балл, максимальный балл - 9)
'ciggaretes_per_day_log'	Прологарифмированный признак количества сигарет в день
'ciggaretes_per_day_score'	Рейтинг количества сигарет в день (каждые 10 сигарет - 1 балл, максимальный балл - 5)
'smoking_score_int_log'	Варианты рейтинга курильщика (комбинации рейтинга курильщика, рейтинга длительности курения и количества сигарет в день
'smoking_score_int_score'	
'smoking_score_int_score_2'	
'smoking_score_int_score_3'	
'smoking_score_int_score_log'	

'alcohol_score'	Рейтинг употребления алкоголя (0 - никога не употреблял, 1 - употреблял в прошлом, 2 - употрябляю на стоящее время)
'alcohol_duration_score'	Рейтинг длительности приема алкоголя, где каждые 5 лет - 1 балл, максимальный балл - 9
'alcohol_duration_log'	Прологарифмированный признак длительности употребления алкоголя
'alcohol_int_score_1'	Варианты рейтинга употребления алкоголя (комбинации рейтинга употребления и длительности)
'alcohol_int_score_2'	
'alcohol_int_score_3'	
'sleep_onset_early_22'	Признаки указывающие н аконкретный час засыпания
'sleep_onset_22'	
'sleep_onset_23'	
'sleep_onset_later_after_0'	
'early_onset'	Раннее засыпание (ранее 23 часов)
'lately_onset'	Позднее засыпание (после 23 часов)
'sleep_upset_early_6'	Признаки указывающие на конкретный час просыпания
'sleep_upset_6'	
'sleep_upset_7'	
'sleep_upset_8'	
'sleep_upset_after_9'	
'early_upset'	Ранний подъем (6 - 7 часов утра)
'lately_upset'	Поздний подъем (после 8 утра)
'sleeptime'	Длительность сна
'high_amout_sleep'	Признак указыающий на длинную продолжительность сна (более 9 часов)
'low_amount_sleep'	Признак указывающий на короткую продолжительность сна (менее 6 часов)
'cardio_score'	Рейтинг сердечно сосудистого риска (Модифицированный Jakarta Score)

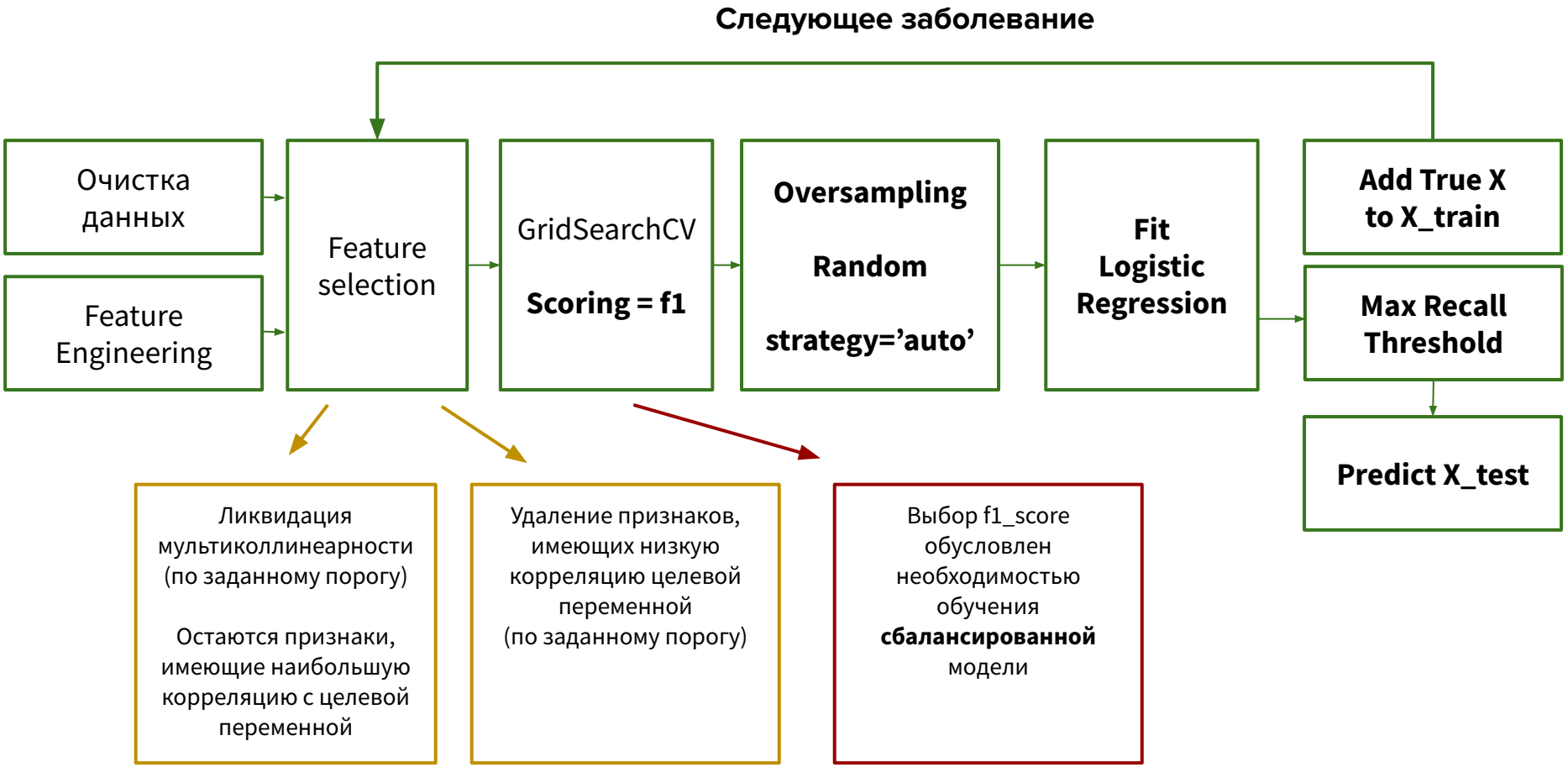
Описание работы модели

Использованные библиотеки: **pandas, sklearn, imblearn**
Использованные модули: **LogisticRegression, GridsearchCV, RandomOversampler**

Основные этапы работы **прототипа**:

1. Очистка данных, перевод признаков формат, воспринимаемый моделью (числа)
2. Генерация дополнительных признаков, таких как: “рейтинг курильщика”, “рейтинг сердечно сосудистого риска” и других из имеющихся данных
3. Выбор признаков на основе наибольшей корреляции с каждым наличием каждого заболевания в отдельности
4. Обучение модели логистической регрессии для определения каждого заболевания в отдельности (всего 5 моделей, которые в последующем могут быть объединены в одну последовательно выполняющуюся программу)
5. Каждая модель вычисляет число от 0 до 1 для каждого нового наблюдения (опроса), что является условной “вероятностью” наличия болезни.
6. Модель автоматически выбирает пороговое значение (от 0 до 1) для определения наличия или отсутствия болезни для максимизации целевой метрики

Схема работы модели



Результаты работы модели

Результаты на hacks-ai.ru (публичный лидерборд) без установленного random_state

submit_stable_figuration_10_no_seed_5.csv	0.733723
submit_stable_figuration_10_no_seed_1.csv	0.725988
submit_stable_figuration_10_no_seed_2.csv	0.711430
submit_stable_figuration_10_no_seed_4.csv	0.719489
submit_stable_figuration_10_no_seed_3.csv	0.722951

Из 638 наблюдений, модель предсказывает наличие болезней в количестве:

Артериальная гипертензия	336
ОНМК	203
Стенокардия, ИБС, инфаркт миокарда	293
Сердечная недостаточность	299
Прочие заболевания сердца	298

dtype: int64

Модель явно предсказывает болезней **больше, чем в реальности** есть у опрошенных, но это хорошо, так как таким образом **формируются группы риска**
(во многом это благодаря выбора метрики recall)

Возможность применения модели

Поскольку задача модели - определить группу риска, целью работы модели может являться **побудить человека обратится в медицинское учреждение для комплексного обследования**

Где такой опросник может быть заполнен?

1. В рамках более широкого опроса, вроде переписи населения и др.
2. С помощью терминала с сенсорным экраном
   a. В социальном учреждении
   b. В парке
   c. др.

https://kiosks.ru/index.php/product/terminal-dlya-polikliniki-atm-med/

Требуется разработка пилотного проекта с исследованием результатов.
Возможно использование модели будет неэффективно

Дополнительные возможности использования модели

Благодаря тому, что модель является интерпретируемой, есть возможность в результате определять не только группу риска, но и выводить некий совет по дальнейшему снижению риска, например:

1. Отказаться от курения (рекомендации + предложение посетить центр здоровья*)
2. Заняться спортом (рекомендации + предложения)
3. Другие рекомендации

* в каждой поликлинике функционирует центр здоровья, в рамках которого работает кабинет по отказу от курения

Приказ Минздравсоцразвития России от 15.05.2012 № 543н
“Об утверждении Положения об организации оказания первичной медико-санитарной помощи взрослому населению”

Контактные данные

Лисовский Дмитрий Александрович

тел. +7-985-999-65-91

e-mail: Lisikux@gmail.com

telegram: @Lisikux