



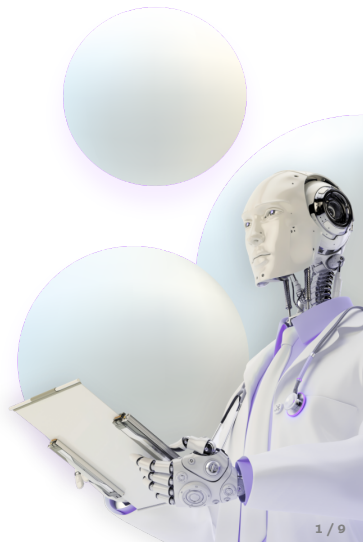
Digital Health Hackathon

LSTM Team

Moscow, 2018

Предсказание диагноза по МКБ

- Более 2300 диагнозов
- В train 62000 записей
- В test 30000 записей
- 48 часов на решение



Признаки

- ID Пациента
- Возраст
- Жалобы
- Источник рекламы
- Номер клиники
- Пол
- Услуга (тип приема)

	Id_Пациента	Возраст	Жалобы	Источник_рекламы	Клиника	Пол	Услуга
Id_Записи							
0	115819	54	на повышение ад утром до 140/90 мм.рт.ст., пер...	Другое	5	2	Прием врача-кардиолога повторный, амбулаторный
1	399973	32	На наличие опухоли в левой молочной железе	Другое	3	2	Прием врача-онколога (маммолога), повторный, а...
2	427563	72	Активных жалоб нет.	Интернет	6	2	Прием первичный врача-пульмонолога
3	257197	55	на сохраняющиеся боли в спине и пояснице, сков...	Другое	3	1	Прием врача-невролога повторный, амбулаторный
4	281066	28	на дискомфорт в горле, слабое першение, слабость	Другое	3	2	Прием врача-оториноларинголога повторный, амбу...
5	341445	46	на боли в 3 пальце правой кисти	Другое	3	2	Прием врача-хирурга первичный, амбулаторный
6	416352	29	Не изменились с момента первого приема	Интернет	2	2	Прием врача-невролога повторный, амбулаторный
7	251280	38	На боли в молочных железах по циклу	Другое	3	2	Прием врача-онколога (маммолога), повторный, а...
8	208376	32	на боли в правой пахово-подвздошной области.	Интернет	5	2	Прием врача-уролога повторный, амбулаторный
9	598841	43	на момент осмотра не предъявляет	Рекомендации знакомых	5	2	Прием врача-акушера-гинеколога повторный, амбу...

Метрика качества - Ассурасу

Интересные наблюдения

- ID пациента очень важен
- Есть жалобы, которые очень часто повторяются:

Не изменились с момента первого приема

2699

на момент осмотра не предъявляет

1472

активно не предъявляет. Явка профилактическая по беременности.

1297

не изменились с момента первого приема

1186

прежние, динамики в состоянии не отмечает

965

прежние

883

активно не предъявляет

- Признак услуги можно рассматривать как категориальный

Генерация признаков












До того как мы определились с моделью, для тестирования признаков использовался случайный лес с небольшим количеством и глубиной деревьев.

- TF-IDF с биграммами на жалобах.
- Категориальный возраст
- Целочисленное деление длины жалобы на 19.
- Агрегация схожих по смыслу услуг.
- Тип приема. (повторный, амбулаторный, профилактический итд.)
- Наличие беременности. (можно извлечь из услуги)
- Стемминг, лемматизация, удаление стоп-слов не помогли.

Обучение модели

- Сервера Google Cloud с 8 ядрами ЦП и 100гб оперативной памяти.
- Нейросети и бустинги нам не подошли.
- Из за длительности обучения и неограниченного числа посылок валидировались по лидерборду.
- Обучение случайного леса со 100 деревьями и глубиной, ограниченной 80, занимало несколько часов.
- После генерации признаков точность предсказания случайным лесом удалось улучшить с 32% до 38-39%.
- При переходе со случайного леса на линейный SVM (перед этим все признаки были бинаризованы), точность резко возросла до 44.5%.
- После тюнинга гиперпараметров точность улучшилась до 46%.

Финальные результаты

#	Δ pub	Team Name	Kernel	Team Members	Score ?	Entries
1	—	[ods.ai] LSTM		  	0.46514	85
2	—	╰(ツ)╯		   	0.44661	48
3	—	[ods.ai] DenisVorotyntsev			0.44100	25
4	—	[ods.ai] Spider ANN			0.43719	7
5	—	[ods.ai] ML SWAT		 	0.39938	44

Что еще можно было сделать?

- В каждой клинике жалобы записывались по-разному, поэтому для каждой клиники можно было ввести персонализированные жалобы, добавив к каждому слову обычной жалобы номер клиники. От этого точность бы однозначно выросла.
- Можно было почистить выбросы, например, встречались записи в которых женщина записалась к урологу. Хотя выбросов было немного.
- Возможно, после удаления очень редких диагнозов точность бы возрасла.

Решение других участников

- LightGBM с удалением редких диагнозов.
- Нейросети, многоуровневые модели.

Спасибо за внимание!